Notes on data curation, including notes on apparently erroneous sequences in NCBI GenBank

Search strings for download (29/03/2012)

Basic taxon-ID complement: (txid3514[Organism:exp] OR txid3520[Organism:exp] OR txid60415[Organism:exp])

A) Marker used in Li & al. (2004)

Cp- and mt-genes: … AND atpB [gene] // AND matK [gene] // AND matR [gene] // AND rbcL [gene]

*trn*L-UAA intron, and spacer: … AND (trnL [gene] or trnL-trnF) NOT atpB-rbcL

18S rRNA gene: … AND 18S NOT (5.8S OR [25S AND 26S AND 28S])

B) Additional nuclear marker

GBSSI (exons 4 and 5): … AND GBSSI [gene]

*rps*16: … rps16 [gene]

ITS region: … AND 5.8S

C) Additional plastid marker

*rpl*16 (exons 1 and 2): … AND rpl16 [gene]

*atp*B/*rbc*L region: …. AND (atpB [gene] OR atpB-rbcL OR rbcL-atpB OR rbcL [gene]) NOT GBSSI [gene]

*trn*H-*psb*A spacer: … AND (psbA-trnH OR trnH-psbA)

NIA gene and intron: … AND “nitrate reductase”

Alignment notes

General note: All auto-alignments were visually inspected; sequence ends were trimmed where necessary; 5’ and 3’ ends only available for single genera and deviating from other genera were cut.

18S rDNA, nuclear

General note: GenBank has mostly old 18S data without voucher information; there is little variation within Betulaceae, and to *Ticodendron* and *Casuarina* except at pos. 1327–1335 and 1615–1631. However, this would need to be verified with newer sequence data.

Curation: The following ITS accessions were omitted from the matrix: X68133–X68139; we also replaced apparent internal sequencing/editing artefacts in *Alnus glutinosa*, X54984, *Betula papyrifera* L00971 (e.g. pos. 157–161 CGTTT instead CCCGTT in all other sequences; pos. 1022 TT instead T i.a.o.s.; pos. 1331–1339 GCACGGA instead GGCCACGGAA i.a.o.s.), and *Corylus avellana*, AY263895 (pos. 498f CCC instead CC i.a.o.s.; pos. 522 CC instead C i.a.o.s.) by missing (“?”). Other possible sequencing artefacts (e.g. CTG instead GTC at pos. 821–823; extra T at pos. 864 in X54984; see note on old and newer sequences of *Fagus sylvatica* in Grimm, 2003) were left unmodified.

*atp*B-*rbc*L spacer/*rbc*L gene, plastid

Problems: There is no sequence overlap between the *atp*B and *atp*B-*rbc*L accessions, and overlap between *atp*B-*rbc*L and *rbc*L accessions is insufficient for auto-alignment; each group forms a separate sub-alignment; the two sub-alignments are then arbitrarily joined by the alignment programme and have to be aligned with each other by hand.

Curation: Omitted short *Alnus glutinosa* sequence AY165749. All *atp*B sequences were deleted from the alignment. Block with *atp*B-*rcb*L sub-alignment was moved to the 5' part of the *rbc*L alignment; sites without data were eliminated. One of the *atp*B-*rbc*L sequences in GenBank was oriented 3'-5' (*Carpinus rankanensis*, AY014606). The reverse complement of the sequence was re-aligned using the pairwise alignment option in MESQUITE.

*atp*B gene, plastid

General note: The alignment comprises only limited taxon sampling; no sequence younger than 5 years old is available, and most variation appears random and restricted to a single accession.

GBSSI gene (intron mainly), nuclear

General note: The central part (pos. 470–635) is affected by prominent intergeneric length-polymorphism, the longest sequence variants are those of *Betula*. Compared to these, in *Alnus* the upstream part is deleted (29 nt), in *Carpinus, Ostrya,* and *Ostryopsis* the downstream part (81–93 bp). Positions of the deletions are straightforwardly defined by high sequence similarity, and the automatic aligner therefore could recognize them.

Curation: In *Corylus* 107 nt of the downstream part were deleted. The autoalignment placed the last block of *Corylus* at the 5' end, causing numerous differences to all other genera. This block of 17–18 nt was moved to the 3' end because of its high sequence similarity with the 3' portion of *Betula* (and *Alnus*), a region deleted in *Carpinus, Ostrya,* and *Ostryopsis.*At pos. 659ff, a thymidine-rich portion shows prominent length-polymorphism in *Betula*. Nucleotides of shorter variants were scattered by the auto-aligner and were right-aligned for better visibility.

ITS region including ITS1 and ITS2 spacers, and 5.8S rDNA, nuclear

General note: The ITS region of *Casuarina,* for which so far only limited data are available, is significantly longer than that of *Ticodendron* and the Betulaceae, inflicting a number of gapped portions in the latter subalignment (comparable to the situation between *Fagus* and the other Fagaceae). Furthermore, the general sequence structure of xx is strikingly different from zz, and only the most conserved ITS1 parts can be straightforwardly homologised. ITS2 can be reliably aligned, except for the 5' and 3' parts, which are often length-polymorphic. Because of the lack of representative data for *Casuarina* and the homology issues, all *Casuarina* sequences were excluded from the ITS alignment. /richtig?/

Sequences deviating from the rule:
Alnus formosana (AJ251678) — ITS1, 5.8S rDNA, and ITS2 show some mutations indicating starting sequence degradation (pseudogeny). Overall, the sequence type matches that of *A. firma*, not of newer sequences of *A. formosana*
A. hirsuta var. *sibirica* (JF975852): The central part of ITS2 deviates from the *Alnus*-consensus.
A. maritima (AJ251679): Sequence pseudogenous.
A. nitida (AJ251677) : Sequence pseudogenous.
The three sequences JF975852, AJ251679, and AJ251677, all with a pseudogenous tendency, are similar to each other.
Betula neoalaskana (AY761123): A number of sequencing/artefacts (single or two nucleotides instead 2 or 3 consecutive).
Curation: Since no site showed a pseudomutation, and the deviation was systematic, only the number of nucleotides was corrected (indicated by lowercase letters).

Carpinus caroliniana (DQ005985; Kress et al., *PNAS*, 2005): Strongly deviating sequence (but of similar length than in Betulaceae); BLAST revealed that this sequence is mis-labelled and belongs to *Ulmus* sp. **The sequence was deleted from the data set**.
Corylus heterophylla (AF081519): (Pseudo-?)polymorphisms at pos. 191ff were not optimally aligned. **Curation:** Realigned by one position.
C. heterophylla var. *sutchuenensis* (AY006351): Sequence (ITS2 only) with numerous polymorphic base calls at positions otherwise conserved within the genus.
Ostryopsis davidiana (AY006334): Sequence deviates at a few positions from newer sequences showing classical sequencing/editing artefacts (TT instead TTT and GG instead GGG at end of 5.8S, pos. 600ff, CCCA instead CCAA, pos. 676–679 i.a.o.s.; pseudopolymorphisms pos. 155ff)

*trn*K intron including *mat*K gene, plastid

General note: Alignment free of length-polymorphism except for the 5' and 3' parts, which comprise the flanking *trn*K intron, in which the *mat*K gene is embedded.

Omitted (mis-labelled) sequences: *Betula pendula* (AM503811; Li et al., *Cladistics*, 2008): Markedly deviating from the *Betula* consensus, inflict gaps in the *mat*K partition. BLAST revealed that this sequence comes from *Antirrhinum majus* (asterid: Lamiales).

*mat*R gene, mitochondrial

General note: The alignment lacks discriminating signal at the intergeneric level. Mutational patterns shared by two or more genera (e.g. pos. 749, 1219) are dissolved into intrageneric variation in genera covered by more than a singleton or appear random (e.g. pos. 1014; T shared by singletons *Ostrya* and *Casuarina*)

NIA gene (intron mostly), nuclear

General note: One long older sequences (labelled as NIA1; *Betula pendula*, acc. no. X54097) is not well aligned by the autoalignment and was removed. All newer sequences come from Li et al., *Syst. Bot*., 2007, and Li, *J. Syst. Evol.*, 2008. and stem from 1–7 cloned sequences. Complex patterns of inter-specific differentiation, associated with high levels of inter-generic divergence involving very prominent length-polymorphic patterns, make alignment of Betuloideae with Coryloideae problematic. The overall data structure hence confirms the Bayesian-inferred root (see main text; matrix supplied in data archive). We opted against including the NIA1 sequences in our concatenated data to avoid long-branching artefacts. Also, two of the genera, *Alnus* and *Ostryopsis* are only represented by a single sequences so far.

Problem: Accession AJ001725 included in the download is a sequence of unknown origin labelled as "NR-gene; promotor" (Strater & Hachtel, *Plant Sci.*, 2000) and was excluded from the alignment. If showing a part of the NIA gene, this portion has no overlap with the other accessions harvested from GenBank.

*trn*H-*psb*A spacer, plastid

General note: The *trn*H-*psb*A spacer shows significant length polymorphism, both on the intra- and intergeneric level. For the most part, length-polymorphic patterns are linked to duplications or, less often, deletions of sequence portions, hence, the auto-alignment performs better than in the case of ITS. Still, the auto-alignment is rather gappy, in particular, in the 3' part (pos. ~430–580), where the general sequence structure varies between species and genera. The alignment of *Casuarina* is problematic.

Problem: A number of GenBank sequences are orientated 3'-5', accordingly the autoaligner produces two sub-alignments that could not be merged.
Curation: Sequences were replaced by their reverse complements and re-aligned using the pairwise alignment tool and a correctly oriented sequence of the same species/genus as template. In the case of the four *Casuarina* sequences, *Alnus acuminata* (FJ011867) and *Ostryopsis nobilis* (JN045685) were

used as template for the initial alignment; then, several blocks were moved by eye to ensure consistency within *Casuarina*. The alignment of *Casuarina* with the Betulaceae may be problematic in general: the spacer is much shorter than in Betulaceae, and in the central part (pos. ~300–600 in the alignment) one could choose between various alternative positions that may be homologous. The sequence structure in this central part differs from the one in the Betulaceae.

Further curation: An AT-dominated length-polymorphic portion in the 5' part of the spacer was re-aligned, in order to minimize substitutions and eliminate inconsistencies within *Alnus* using the longest variant (*A. acuminata*, FJ011867, as template). *Betula* block was moved by 2 bp downstream at pos. 311ff. Auto-alignment failed at the 3' end, which was therefore re-aligned by eye.

Mis-labelled sequences:
Carpinus caroliniana (DQ006158; Kress et al., *PNAS*, 2005). According to BLAST *Ulmus* psbA-trnH (same voucher as DQ005985). **The sequence was deleted from the data set**.
Alnus glutinosa (FN687523; Piredda, uploaded 2010); shows the *Corylus* sequence type. A *Corylus avellana* [FN687522; same author(s)]; shows the *Alnus* type. The high sequence identity to other species indicates that the sequences were mixed up during uploading; sequences were re-labelled as *Alnus* sp. and *Corylus* sp., respectively.

Sequences deviating from the rule:
Alnus formosana (FJ844545) shows a number of mutations not found in any other *Alnus* (incl. a second sequence from this taxon) in the central part of the sequence (331–364); the according part has been replaced by "?"
The two sequences of *Alnus alnobetula* subsp. *sinuata* (FJ844485; FJ844486) show a prominent deletion in the 5' part of the spacer unlike other *Alnus*; the remaining sequence sections are of the *Alnus* type.

*rpl*16 intron, plastid

Problem: A number of *rpl*16 intron sequences (AB237203; AB23723; AB537994; AB537990; AB537999; AB538004; AB538009; AB538014; AB538019; AB538024; AB538034; AB538029) are orientated 3'-5' in GenBank. **Curation:** Sequences were replaced by their reverse complements and re-aligned using the pairwise alignment tool and a correctly oriented sequence of the same species/genus as template.
Further Curation: AAT (pos. 143ff) aligned left (auto-alignment placed a gap before the AAT of most *Carpinus*). For consistency the downstream multiple-A motif was also aligned left, inflicting two transversions compared to *Betula* to gauge the original alignment (inflicting these transversions only in the longest motives)

*trn*L intron/*trn*L-*trn*F spacer, plastid

General remarks: The *trn*L intron is highly conserved across Betulaceae, with only *Alnus* differing from the general consensus. There are also two species-limited duplications of 6 and 5 nt in its 5' portion (one in one of two *Ostrya knowltonii* accessions and the two *O. rehderiana* accessions including this part; the other in *Corylus yunnanensis*). Furthermore, all *Betula* accessions save one (AY147068; another accession from the same species follows the rule) exhibit a deletion of 163 nt in the 3' portion. The *trn*L-*trn*F spacer appears gappy in the auto-alignment, however, most of these gaps are due to straightforwardly defined duplications/insertions, which are restricted to a single or few accessions, hence, contain little phylogenetic information. Several length-polymorphic regions were slightly adjusted by eye. The alignment further comprises a number of single-nt gaps or unique nucleotides in highly conserved exons, which are likely sequencing/editing artefacts (pseudomutations) in most cases. Due to the lack of comparative data, no general changes were applied to the sequences, but singleton deviations in the exons were replaced by "?" (except in the case of the singletons *Casuarina* and *Ticodendron*). We also deleted the 5th C at the end of the 3' trnL exon in a few sequences contrasting the highly conserved sequence in all other accessions (FJ012034; FJ012035; FJ012052; FJ012057; FJ012062; K.O. Yoo & J. Wen, uploaded 2008)

The download included a number of sequences with the wrong locus information ('misc_feature' flag) "*trn*L-*trn*F intergenic spacer region"; the headers reading otherwise indicating the putative loci "*pet*D-*rpo*A…", "*trn*W-*trn*P…", "*psb*C-*trn*S…" , "*trn*T-*trn*L…", "*trn*S-*trn*G…", "*trn*G-*trn*fM intergenic spacer…" (AB237206–08; AB237210–12; AB237218–25; AB237236–38; Iwasaki et al., *J. Plant Res.,* 2006).
Curation: For better comparability and to increase consistency, the length-polymorphic thymidine-dominated part at pos. 389ff was re-aligned. An insertion at pos. 719ff was shifted in one of the *Ostryopsis davidiana* sequences to match its position in the other accessions; placed both insertions (*O. davidiana, Ostrya carpinifolia*) in the same gap region. Pos. 848ff all initial thymidines were right-aligned in *Betula*, where they had been spread out? by the auto-aligner. A microsatellite-like region (pos. 892–948) with numerous polymorphic base-calls (e.g. AW instead AT in all other sequences) in *Ostryopsis davidiana* (AY147071; newer sequences from the same taxon clear) was replaced by "?". At pos. 1047ff several accessions were re-aligned to fit with the position of the shared central motif (KTCT).

Sequence deviating from the rule:
Alnus japonica (AY211427). The *trn*L intron part of this accession deviates from all other *Alnus* and shows a *Betula*-specific deletion; it may be a literal duplicate of *Betula occidentalis* (AY211428), vouchered from the same arboretum and coming from the same study (Yoo & Wen, *Am. J. Bot.*, 2003). The *trn*L-*trn*F spacer, however, is more similar to other *Alnus* than to *Betula.* This likely artificial chimeric sequence was tentatively relabelled to *Betula* x *Alnus*.
Alnus nepalensis (FJ012050) shows an insert within the end of the 3' *trn*L exon, which is a joined duplication of both flanking highly conserved exon sequences (*a/b*) in mixed order: *a*-C-*b*-TGG-*a*-*b*. The insertion could be moved to the start of the spacer, if it wasn't for the additional starting cytosine.