# GenBank harvest, December 13th, 2013

## Search strings

txid22788[Organism:exp] AND …

### Plastome

… (atpB [title] OR "adenosine triphosphatase" [title]) — for *atp*B gene, will not catch two accessions of *Akebia quinata* without info or writing errors in title (L37924; AF209523). A complete accession covering *atp*B and *rbc*L genes for the taxon is in the data harvest

… (matK [title] OR trnK [title]) — for *trn*K/*mat*K region

… rbcL [title] — for *rbc*L gene

… (trnL [title] OR trnL-trnF [title]) — for *trn*L/LF region

### Nucleome

… 18S [title] NOT internal [title] — for 18S rDNA

… internal [Title] — for ITS region

## Pre-analysis data processing

All data were aligned using mafft v. 6.935b; alignment algorithms were selected by the programme (option --auto). All auto-alignment were checked by eye using mesquite v. 2.75, and sequence end were trimmed were necessary. Leading and trailing regions with limited data were excluded. No further manipulation of the auto-alignments was performed beyond the curation procedure outline below.

### Region #1: Nuclear-encoded 18S rRNA gene, partial sequences

**General—**The data comprises only nine old accessions without voucher information. All sequences are highly similar in general, several sequencing gaps are present. Many deviations from the consensus are restricted to a single sequence, and frequently associated to nucleotide-repetitions or sequencing gaps (e.g. GGACCAA at pos. 67–73 in one accession of *Akebia quinata*, L31795,compared to GAACAAA in all other accessions including the second *A. quinata*; see also pos. 542ff; 711ff; 739ff; 782­ff; 825ff; 1184f), which may be indicative of a sequencing/editing bias. Aside from these potential pseudo-apomorphic sequence features, there are very few positions (<10) that may actually be able to discriminate between genera.

Without newer data for comparison, it is difficult to judge which mutation patterns are reliable and which are simply sequencing/editing artefacts.

**Curation —** The gene region has been excluded from further analyses due to the lack of signal and reliability of data.

### Region #2: Nuclear-encoded ITS region

**General —** The ITS1 region of *Sargentodoxa* (*cuneata*; 11 accs)has a markedly different structure than in the remaining Lardizabalaceae, and shows a large gap in the central part also present in the singletons representing *Lardizabala* and *Boquila.* Divergence in ITS1 sequence structure is mirrored to a lesser degree in the ITS2 regions. As known from many data sets, the 5’ and 3’ region of the ITS2 includes some length-polymorphisms, the remainder of the region is generally length-conserved and straightforwardly aligned.

Another strongly unique ITS1 region both upstream and downstream of the putative ITS1 cleavage region is exhibited by *Sinofranchetia* (AY029782); the ITS2 sequence shares similarities with those of *Lardizabala* and *Boquila.* The ITS1 regions of this accession is so unique that it does not result in a significant hit using nblast (blast-search for ‘somewhat similar sequences’).

**Curation —** Sequence blocks at pos. 19–24 and 191–203 in *Sargentodoxa* were shifted by one site. Isolated placed C and G in *Akebia* and *Boquila*, respectively,in the length-polymorphic ITS2 region at pos. 504ff were left-aligned; and the *Decaisnea* sequence re-aligned for better match with all other sequences. The downstream portion of the *Decaisnea* sequence shows signs of sequencing/editing errors (several single-nt indels in a generally length-conserved region) at pos. 662ff; the problematic part was replaced by missing data.

### Region #3: Plastid (*trn*K/)*mat*K region

**General —** Only minor length-polymorphism. Data mostly from the *mat*K gene, only three accessions including data from the *trn*K intron.

**Curation** — Two *trn*K 3’ intron (identified by comparison with the complete plastome sequence of *Nandina domestica* DQ923117) accessions (GQ452040; GQ452039) were not correctly aligned with the remainder of the data due to missing overlap and were excluded.

Gaps were shifted to match codon-position; occasional incomplete codons were replaced by missing data.

### Region #4: Plastid *trn*L/*trn*L-*trn*F region

**General—**Sequences cover the *trn*L intron and *trn*L-*trn*F spacer regions (coding regions identified using the *Nandina* complete plastome as reference). Length-polymorphism restricted to simple indel patterns. Most distinct sequence type is that of *Sargentodoxa.* Several isolated single-nt indels in the central part of sequences.

**Curation—**Sequence blocks shifted relating to indels for *Sargentodoxa* (pos. 174ff, 835ff) in order to minimise number of necessary mutations. Left-/right-aligned pos. 350–376. Uninformative single-nt indels linked to ‘basepair hickups’ (e.g. AA instead of A in all other sequences) in older sequences were cleaned in case comparative data were available. Re-aligned *Decaisnea, Sinofranchetia, Sargentodoxa* at pos.643 ff and 835ff (*Sargentodoxa* only) for consistency with apparent insertion pattern.

### Region #5: Plastid *atp*B gene

**General**—Overall divergence of the nine accessions covering all currently recognised genera, is extremely low. Most of the 39 variable sites show genus-unique mutations from the family consensus, the most unique sequence represented by the *Sargentodoxa* singleton. Only five sites contain information on potential bipartitions of the taxon set (Fig. S1).

**Curation—**Analogously to 18S rDNA, this gene region has been excluded from further analyses due to the lack of signal and doubts about the reliability of data.

Fig. S1. *Full* *Median-network based on* atp*B gene data. Mutations limited to a single accession annotated by stippled lines and numbers.*

### Region #6: Plastid *rbc*L gene

**Note—**The search string caught the only *atp*B-*rbc*L spaceraccession (HE651093) of the family, which was removed.

**General—**As expected for *rbc*L no length-polymorphism is found. Overall, intra-generic divergence is low, but not as low as in *atp*B. However, most site variation distinguishes one genus from all other. Data quality can be assessed since most taxa are represented by more than one accession, including recent re-sequencings (in total 27 accessions covering all genera and 16 spp.)

**Curation—**Corrected (blanked by missing data) for incomplete codons and potential sequencing/ editing artefacts: unique deviations at generally family-conserved positions typically restricted to the oldest accessions and first-sequenced species (Dxxxxx, Lxxxxx; orange background in NEXUS file). Also among newer sequences (of more recently sequenced species), unique site mutations are occasionally found (yellow background). Lacking comparative data, it cannot be judged if this is intrageneric divergence or just pseudomutations.

**Problematic data—**One of the three accession of *Stauntonia hexaphylla* (D85694) differs from the other two by being identical to the other sequences of the (revised) genus (which are highly similar to each other). Available data is does not allow to decide, if this is intra-species variation, and uploading error or due to directed editing, hence, all three sequences were kept.

Sequence AB729099 (*Akebia trifoliata,* partial *rbc*L; red background in NEXUS file) misses the only two mutations, which differentiate the genus *Akebia* (two species, with four accession), but not the former *Archakebia* *apetala* (now included in *Akebia*)from all other genera in the covered fragment. The sequence has been tentatively excluded from analysis.