

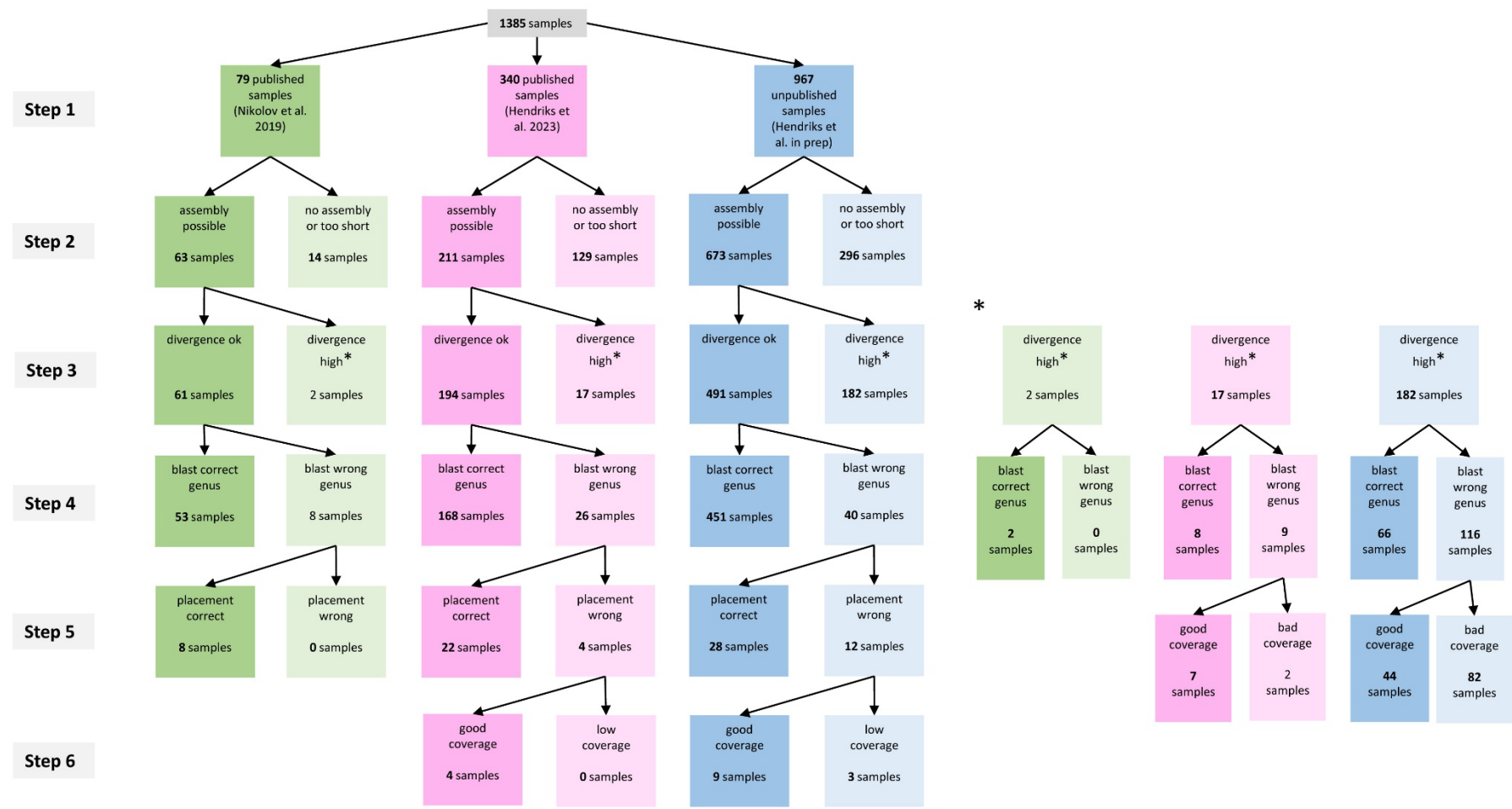
Online resource 3:

Capacity building in crucifer research – from *BrassiBase* to a taxonomic expert network (*BrassiTEN*)

Plant Systematics and Evolution; Kiefer et al.

corresponding author: Marcus A. Koch; marcus.koch@cos.uni-heidelberg.de

Supplementary Figure S2. Workflow for determining high quality ITS sequences generated from NGS data for integration into *BrassiBase*.



In **step 1** Illumina sequencing data of 1,385 samples originating from Nikolov *et al.* (2019), Hendriks *et al.* (2023) and Hendriks *et al.* (in prep) were subjected to *de novo* assembly by SPAdes (Prjibelski *et al.*, 2020). Samples from the Nikolov *et al.* (2019) were also reanalysed and integrated in Hendriks *et al.* (2023). In **step 2** promising contigs representing the ITS region were determined by blast search using a published ITS sequence and filtered for a length >500bp. The resulting sequences were integrated into the ITS reference alignments from BrassiBase and dubious sequences were excluded after visual inspection (**step 3**). To ensure that only good quality sequences were kept, in **step 4** the sequences were blasted against our BrassiBase blast database and genus as well as tribe identity of the blast hit were recorded and compared to the taxon identity given for each sample by the respective authors. In case that the best hit in blast search pointed to a different genus or even tribe the sequence was also phylogenetically placed as phylogenetic placement is more accurate than blast search (**step 5**). As sequences identified in step 5 are potentially originating from wrongly determined material, we introduced an additional check of the coverage in the assembly (**step 6**). If coverage was given as <20 in SPAdes it was qualified as bad coverage. Only samples showing a wrong placement after blast search as well as a bad placement and a good coverage should be considered as originating from wrongly determined material.

High divergence, as identified in step 3, can also be caused by sequences originating from wrongly determined material. Therefore, we also subjected the sequences identified as divergent to blast search against our ITS database in BrassiBase. Also, here we checked whether coverage of the contigs containing the ITS was <20. If yes, these samples could be considered as wrongly determined if blast search is wrong but the coverage >20. However, it should always be considered that there is the possibility of by chance sequencing of ITS paralogues or pseudogenes which can also lead to the introduction of seemingly random SNPs which lead to a bad blast result. So probably our numbers here are an overestimate.

As low coverage can lead to assembly errors, contigs with low coverage or originating from samples with few reads were excluded from analysis (step 3). To ensure that only good quality sequences were kept, in step 5 the sequences were blasted against our BrassiBase blast database and genus as well as tribe identity of the blast hit were recorded and compared to the taxon identity given for each sample by the respective authors. In case that the best hit in blast search pointed to a different genus or even tribe the sequence was also phylogenetically placed as phylogenetic placement is more accurate than blast search (step 6).

