

Kellogg lab procedures

Outlined below are the general procedures in current use for systematic studies in the Kellogg lab.

Please feel free to cite this page as E. A. Kellogg. 2004. Standard lab procedures, <http://www.umsl.edu/divisions/artscience/biology/Kellogg/Kellogg/home.html>

People who are accustomed to field or herbarium taxonomy are often unaware of the different rhythm of lab studies. In particular, in the lab there is a major up-front investment of time and effort to learn techniques and then to identify a set of phylogenetic markers that will solve the question at hand. The goal is not just to collect data, but to get useful data. It is very common for people to spend three or four months in the lab before finally beginning to collect data that will be used in a publication. Simply learning the necessary techniques requires two to four weeks full time (80-160 hours). Because of this, it's usually not a good idea to start in the lab until you have large blocks of time – at least several days per week – to invest. It's also been my observation that it is impossible to work simultaneously on the herbarium and the lab aspects of a project. Sequential is better. See below.

I. Prior herbarium and literature investigations. Only after these have been completed are you ready to proceed to the lab.

Begin by identifying a monophyletic study group. It's important to have a good hypothesis of a synapomorphy (not a unique combination of characters!) that unites the group. This can be based on a previous molecular study or morphological taxonomy. Know the possible outgroups and be able to identify the next largest monophyletic group and its synapomorphies.

Do enough herbarium work to be familiar with all the morphospecies in the study group. Look at all the species in the herbarium and learn their geographical range and their distinctive characteristics. Read everything that you can find about the taxonomy of the study group and its near relatives. If you are planning a monograph as part of your study, this is a good time to get it started.

Find out about all molecular studies done on the family and related families. Know what genes have been used, how well they resolved particular taxonomic problems, and what their primer sequences are.

Collect as much material as you can. The ideal is to have fresh growing material for at least a few species so you can optimize DNA extraction protocols on that before proceeding to dried material. Second-best is silica dried material that you have collected recently. Try to have 10 g dry weight if at all possible, in case you need to do large-scale DNA extractions to get good quality DNA. Herbarium material is a last resort. It works well for some groups, unpredictably for others, and not at all for some. When you are getting started you won't know if your inability to get DNA out of a herbarium specimen has to do with your lack of technical skill, the inadequacy of the particular extraction protocol, or the general reluctance of dried specimens of your group to surrender their DNA; easiest solution – only use herbarium specimens after you have found a good reliable extraction protocol that works for you.

II. Preliminary sequence studies

It is often challenging to find a set of genes that will answer a particular phylogenetic question. We've found that it is worth spending the time (often several months) to find one or more genes that will resolve the phylogeny, rather than using a gene chosen a priori. In general, our philosophy is that if there isn't enough variation to answer the question, stop sequencing! And find something else that will work. Often you can get a hint of what may or may not be useful from the literature, but it still needs to be tried.

When a student starts a molecular phylogenetic study in the lab, she/he identifies a small number (four to six) of morphologically disparate species in the study group that are hypothesized to represent different evolutionarily lineages. These form a sample for preliminary molecular work. (Preliminary work should be done using fresh or silica-dried leaf material as herbarium samples can be difficult to extract; see above.)

Chloroplast genes. Often for learning how to do PCR, sequencing and alignment, *trnL-F* is a good place to start because it amplifies easily and can be easily sequenced. Most people in our lab start with this one. However, in most cases, sequences of the sample species are almost invariant. Therefore *trnL* is generally abandoned after sequencing the preliminary sample, although when it does work, it's great.

Most people in the lab end up using *ndhF* for a chloroplast marker, although depending on the system, we have used *rps4* or *rbcL*.

Nuclear genes. We routinely use low copy nuclear genes, and in particular nuclear introns, for phylogeny reconstruction. These genes often provide more characters than the standard phylogenetic markers such as *trnL-F* or ITS.

Find as many sets of primers as possible from the freezer and order more if necessary. We currently have primers for genes such as *TPI*, *CHS*, *CHI*, *PRK*, *PEPC*, *waxy* (*GBSSI*), and many others. The primers are placed in exons and span introns. Try all possible primers on the preliminary set of taxa. Find a set of 2 or 3 that work easily on the preliminary taxa. If a pair of primers don't work, it probably isn't worth the time trying to optimize them. On the other hand, if you get weak amplification, it might be worth spending time to see if you can make it stronger by changing the PCR conditions.

Clone the PCR products and sequence them. Assess level of variation and ability to align them. Proceed then with the introns that exhibit the appropriate level of variation – i.e. can be aligned but exhibit appreciable variation.

Use ITS only if all else fails. It can provide reliable phylogenies but needs to be cloned to assess the level of variation and to check for possible pseudogenes. Multiple clones need to be sequenced for each species, so in the long run it often ends up being as much work as (or even more than) the low-copy nuclear genes.

III. Sequencing

Once genes have been identified that are likely to provide phylogenetic information, proceed to sequence everything you can get your hands on.

Cloned template is purified using plasmid prep rather than reamplifying off the clone to avoid adding another polymerase step with its accompanying errors.

We sequence both strands of a single clone and one strand of at least one more clone to estimate sequences. If polymorphism levels are high we sequence more clones. Based on previous work in the lab (Razafimandimbison et al., 2004, Systematic Biology),

we find that a good assessment of ITS variation can occasionally require as many as 30 clones per plant.

We check phred and phrap values on all sequences produced in the lab after January 2004. The goal is for double stranded sequence with all bases on both strands with phred values above 20. We can routinely achieve 90% double-strand coverage at this level, but often end up with phred values below 20 on one strand. We generally accept these if they constitute less than about 10% of the length of the fragment AND are verified by high-quality (phred > 20) sequence on the other strand. Note that this level of sequence quality is still well below the standards accepted by the genome projects. (See Bermuda standards, or Standard Finishing Practices and Annotation of Problem Regions for the Human Genome Project, at the website of the National Human Genome Research Institute.)

Needless to say, be sure every sequence can be traced directly back to a voucher specimen. Be tough enough to throw out a sequence if you can't find the voucher; unvouchered data have no value, so you aren't throwing away valuable data.

IV. Analysis

Analyze the data from time to time to be sure you are still getting useful information. Align the sequences in ClustalX and then adjust the alignment by eye in Se-Al and/or MacClade. If you are working with a coding sequence, check for open reading frames by translating the sequence; premature stop codons could indicate a problem with sequence quality.

For quick searches, do a neighbor-joining tree or a heuristic parsimony search. If you are working with cloned sequences, analyze them frequently as you go along to see if the clones of a single species are all more closely related to each other than to clones of another species. If species share alleles, you may need to sequence more clones of each.

If you are not getting enough variation to resolve a phylogeny for your group, stop sequencing, and change markers. It is a waste of time and money to collect data that do not answer the question.

If you have a sequence that is appearing in an odd place in a phylogeny, go back and verify the voucher, the extraction gel, and the sequence quality. Redo the sequence if at all possible. You may have a) discovered something truly novel, or b) made a routine mistake. Before you publish on (a), you'll need to convince yourself that (b) didn't happen.

When you get near a final data set, have someone else look at the alignment to see if it really is as convincing as you think it is. Then analyze it as extensively as you can – multiple methods of analysis, multiple attempts to assess support, include and exclude outgroups and critical ingroup taxa, force certain topologies to see how much worse they are than the optimal one. Plan to spend several months (3 to 6) on analyzing a reasonable sized data set. After all the work you've put into gathering the data, you owe it a good analysis.