A provisional UniGene clone set based on ESTs from *Neurospora crassa.*

William H. Dvorachek, Jr., Patricia L. Dolan, Mary Anne Nelson and Donald O. Natvig. Department of Biology, University of New Mexico, Albuquerque, NM

We have constructed a list of *N. crassa* cDNA clones for which partial sequences exist, toward the goal of maximizing the number of genes represented while avoiding redundancy. This effort employed GenBank sequences from the combined *N. crassa* EST projects at the University of New Mexico, the University of Oklahoma and Dartmouth College (27,557 ESTs; Nelson *et al*. 1997 Fungal Genet. Biol. **21**:348-363; Zhu *et al*. 2001 Genetics **157**: 1057-1065). The current list, subject to ongoing revision, includes 2842 clones and is available at the web site of the *Neurospora* Genome Project (NGP) at the University of New Mexico (http://www.unm.edu/~ngp/), along with details of its construction. Each cDNA clone in the list represents a unique gene.

We have also assembled a UniGene set of cDNA clones for that portion of the UniGene set that is represented in libraries constructed by the NGP at UNM. This UniGene library is comprised of 1786 clones distributed in 20 96-well dishes, and it is available through the Fungal Genetics Stock Center.

Method. Our strategy was based on contig assemblies constructed using a parallel version of Phrap (http://www.genome.washington.edu/UWGC/analysistools/phrap.htm) obtained from Southwest Parallel Software, Inc. (SPS Phrap; http://www.spsoft.com/). Phrap output includes contigs, each assembled from two or more overlapping sequences, in addition to singlets, unique sequences which are not included in any contig.

All cDNA sequences analyzed were derived from directional clones, with the result that a given sequence can be classified as originating from the 5' (forward sequence) or 3' (reverse sequence) end of the relevant gene. Relevant to our analysis, when forward and reverse sequences from a single clone fail to assemble into a contig, Phrap reports that the two sequences are "linked" based on clone name. Accordingly, such links can be assessed in the context of contigs. A failure to assemble typically occurs when inserts are of sufficent size that forward and reverse sequences fail to overlap. However, the Phrap output also includes "false links" that result from bookkeeping errors in the assignment of clone names to sequence reads (see below).

In many instances the inclusion of contigs or singlets in the UniGene set was straightforward. Clones for singlets that linked to no other contig or singlet were included, as were clones that generated two linked singlets (one forward and one reverse), but which linked to no other clone or singlet. Contigs that linked to no other contig, including contigs that linked to singlets, were designated isolated contigs and were also included.

The false linkage of one contig to another contig complicates the identification of isolated contigs. In the absence of false links, when inserts are short enough that forward and reverse sequences overlap, the two sequences from a given clone will be included in the same contig. There are two defining characteristics of such contigs derived from short inserts: (1) there are no links to other contigs, and (2) forward and reverse sequences occur in roughly equal numbers. False links are in many cases evident as single reads that link such a contig to another contig. Therefore, when a contig had approximately equal numbers of forward and reverse sequence reads, despite (presumably) false links to other contigs, the contig was designated a *probable* isolated contig and was included in the set.

Another type of gene added to the UniGene set represented a set of linked contigs, a supercontig, wherein each individual contig included a portion of a given gene whose transcript was too large for the forward and reverse sequences from a given clone to assemble. An idealized supercontig consists of two or more Phrap contigs, with one comprised of reverse reads and one or more comprised of forward reads. For construction of supercontigs, each contig was characterized as a forward contig (a possible supercontig component), a reverse contig (another possible supercontig component), or a mixed contig (a likely isolated contig). For each reverse contig, we identified all forward contigs to which it linked, creating a single supercontig.

Among the sequences analyzed, it was not uncommon to encounter intron splice variants derived from the same gene (manuscript in preparation). Typically, Phrap will assemble spliced and unspliced sequences into separate contigs. We identified such contig groups with a parallel version of the program CrossMatch (http://www.genome.washington.edu/uwgc/analysistools/swat.htm), SPS_Cross Match (http://www.spsoft.com/), which applies the Smith-Waterman algorithm, by performing an "all-to-all" comparison with the Phrap contigs. Only one member of any such contig group was selected for the UniGene set.

Clone choice. For singlets there was only one clone present to be chosen as a representative clone. For a given isolated contig, we chose as the representative clone one that had both forward and reverse sequences present in the contig. For supercontigs, we usually chose a clone that was represented in linked forward and reverse contigs.

Acknowledgements: We thank Gary Montry of Southwest Parallel Software, Inc., for generously providing the SPSOFT Phrap package. We thank Joe Fawcett and Norman Doggett of the Life Sciences Division at Los Alamos National Laboratory for help with robotic clone picking, Mark Fleharty for computational assistance, and Gabriel Quinones for technical support. We thank Dan Ebbole, Texas A&M University, for providing an unpublished earlier provisional *N. crassa* UniGene list. We gratefully acknowledge computer support from the Albuquerque High Performance Computing Center at the University of New Mexico. This work was supported by NSF grants MCB-0078306 (D.O.N.) and MCB-9874488 (M.A.N.).