Published in final edited form as:

Adv Genet. 2007 ; 57: 49–96. doi:10.1016/S0065-2660(06)57002-6.

Enabling a Community to Dissect an Organism: Overview of the Neurospora Functional Genomics Project

Jay C. Dunlap^{*}, Katherine A. Borkovich[†], Matthew R. Henn[‡], Gloria E. Turner[§], Matthew S. Sachs[¶], N. Louise Glass^{**}, Kevin McCluskey^{††}, Michael Plamann^{††}, James E. Galagan[‡], Bruce W. Birren[‡], Richard L. Weiss[§], Jeffrey P. Townsend^{‡‡}, Jennifer J. Loros^{*}, Mary Anne Nelson^{§§}, Randy Lambregts^{*}, Hildur V. Colot^{*}, Gyungsoon Park[†], Patrick Collopy^{*}, Carol Ringelberg^{*}, Christopher Crew[†], Liubov Litvinkova[†], Dave DeCaprio[‡], Heather M. Hood[¶], Susan Curilla^{*}, Mi Shi^{*}, Matthew Crawford[‡], Michael Koerhsen[‡], Phil Montgomery[‡], Lisa Larson[‡], Matthew Pearson[‡], Takao Kasuga^{**}, Chaoguang Tian^{**}, Meray Baştürkmen[¶], Lorena Altamirano[†], and Junhuan Xu^{§§}

^{*}Department of Genetics, Dartmouth Medical School, Hanover New Hampshire 03755

[†]Department of Plant Pathology, University of California Riverside, California 92521

[‡]Broad Institute of MIT & Harvard, Cambridge, Massachusetts 02142

[§]Department of Chemistry and Biochemistry, University of California Los Angeles, California 90095

[¶]Oregon Health & Science University, Beaverton, Oregon 97006

^{**}Department of Plant and Microbial Biology, University of California Berkeley, California 94720

^{††}Fungal Genetics Stock Center, School of Biological Sciences University of Missouri, Kansas City, Missouri 64110

^{‡‡}Department of Ecology and Evolutionary Biology, Yale University New Haven, Connecticut 06520

^{§§}Department of Biology, University of New Mexico, Albuquerque New Mexico 87131

Abstract

A consortium of investigators is engaged in a functional genomics project centered on the filamentous fungus *Neurospora*, with an eye to opening up the functional genomic analysis of all the filamentous fungi. The overall goal of the four interdependent projects in this effort is to accomplish functional genomics, annotation, and expression analyses of *Neurospora crassa*, a filamentous fungus that is an established model for the assemblage of over 250,000 species of nonyeast fungi. Building from the completely sequenced 43-Mb *Neurospora* genome, Project 1 is pursuing the systematic disruption of genes through targeted gene replacements, phenotypic analysis of mutant strains, and their distribution to the scientific community at large. Project 2, through a primary focus in Annotation and Bioinformatics, has developed a platform for electronically capturing community feedback and data about the existing annotation, while building and maintaining a database to capture and display information about phenotypes. Oligonucleotide-based microarrays created in Project 3 are being used to collect baseline expression data for the nearly 11,000 distinguishable transcripts in *Neurospora* under various conditions of growth and development, and eventually to begin to analyze the global effects of loss of novel genes in strains created by Project 1. cDNA libraries generated in Project 4 document the overall complexity of expressed sequences in *Neurospora*, including alternative splicing alternative promoters and antisense transcripts. In addition, these studies have driven the assembly of an SNP map presently populated by nearly 300 markers that will greatly accelerate the positional cloning of genes.

I. INTRODUCTION

The availability of whole genomic sequences has vastly accelerated the pace of research in eukaryotic model systems. However, to exploit this resource, research communities must (1) annotate the genome to extract the relevant information, (2) systematically disrupt the function of the identified genes, (3) examine the regulation of the genes in different biological contexts, and finally (4) communicate this information to the scientific community at large, particularly to those studying similar problems in other systems. We may then integrate all these aspects of phenotype and regulation into a comprehensive portrait describing the biology of organisms. It is a tautology to state that the simplest organisms are the easiest to dissect but may reveal the least, and that the most complicated organisms, while the most information rich, may be beyond the scope of current efforts. Yet it is apparent from the extant genomic comparisons that conservation of important biological processes is the rule, and that simple models can inform more complex systems. The desirability of rich biology coupled with the realistic need for approachable genetics recommends the filamentous fungi in general, and prominent among the established model organisms are *Neurospora* and *Aspergillus*.

A. Why study fungi?

Fungi, plants, and animals represent the three phylogenetic kingdoms within the eukaryotes. Within the ~250,000 different species of fungi, about 75% belong to the Ascomycetes (90% being filamentous fungi, the remainder being yeasts) and 25% are Basidiomycetes (those that form fruiting bodies such as mushrooms). The fungi have an enormous impact on the United States and world economies. Because fungi constitute the Kingdom most closely related to Animalia, and yet are exceptionally tractable experimentally, fungi are universally used as model organisms for understanding all aspects of basic cellular regulation. These regulatory networks include cell cycle progression, gene expression, circadian timing, light sensing, recombination, secretion, and multicellular development. Additionally, mycorrhizal fungi (those that grow interdependently with the roots of plants) are essential symbionts without which most trees and many grasses cannot live; fungi also carry out most biomass turnover. The filamentous fungi also include serious plant and human pathogens; the latter have particular impact on the health of those that are immunocompromised. In addition, because they are eukaryotes, treatment of opportunistic infections by fungi poses special risks and challenges not encountered with bacteria. Filamentous fungi are furthermore metabolically gifted with the ability to produce secondary metabolites, many of which are now recognized as both important pharmaceuticals and salient environmental toxins. Pharmaceutical manufacture using fungi constitutes a multi-billion dollar per year industry. Penicillin and similar β -lactams, all produced by fungi, are the world's largest-selling antibiotics. At the same time estimates are that 10–35% of the world's food supply is lost each year due to fungal contamination, a loss of over \$200 billion per year. Contamination of grains with aflatoxin, produced by *Aspergillus flavus*, is considered the chief cause of cancers in developing countries. Within the United States alone, fungicide sales constitute more than a \$1 billion per year industry. Industrial production of chemicals by filamentous fungi constitutes a greater than \$35 billion per year industry. The United States is a net importer of some of these chemicals (such as citrate), representing in excess of a \$1 billion annually. Industrial production of enzymes, largely by filamentous fungi, constitutes a \$1.5 billion per year industry.

B. Why *Neurospora*?

Neurospora crassa has been studied for decades and is perhaps the best understood filamentous fungus. It is a saprophyte that displays both asexual and sexual life cycles.

Neurospora exists vegetatively as an incompletely septate syncytium, growing equally well in simple liquid or on solid media of known composition. It is nonpathogenic, although very closely phylogenetically related to pathogens. Both asexual development and sexual differentiation are highly influenced by environmental factors including nutrient, light, and temperature. Neurospora is typically haploid undergoing only a very transient diploid stage immediately prior to meiosis. Unlike the yeasts, *N. crassa* elaborates at least 28 distinct cell types that contribute to a wonderfully complex life cycle (Fig. 2.1) (Bistis *et al.*, 2003; Borkovich *et al.*, 2004). Because of its interesting and diverse biology, ease of culture, facile genetics, and rapid growth rate, Neurospora remains a widely used model that sustains a wide community.

The reference-quality genomic sequence (ca. 16-fold coverage) is complete and automated annotation (<http://www-genome.wi.mit.edu/annotation/fungi/neurospora/index.html>) predicts about 10,000 genes (Galagan *et al.*, 2003). The genetics of Neurospora is unparalleled within the filamentous fungi—most identified genes, densest and most accurate genetic map—a legacy of 70 years effort that began concurrently with *Drosophila* and now supports a large research community. The field of biochemical genetics arose from work in Neurospora (Beadle and Tatum, 1945) and the ease of culture, rapid growth (mass doubling time of 140 min), and ease of harvest continue to aid research. Crosses are technically trivial and the genetic generation time (from progeny to progeny) is about 3 weeks. Sexual spores are stable for years at 4°C and asexual spores or mycelia can be stored for decades. Molecular tools are comfortably advanced and being steadily improved as is typical in a vibrant research community. Neurospora was the first filamentous fungus to be transformed (soon after yeast; Davis, 2000), and transformation is routine at frequencies up to many thousands of transformants per microgram of DNA. A variety of selectable markers exist, and several regulatable promoters are routinely used to control expression of transgenes (Aronson *et al.*, 1994a). As in animal cells, transformation in completely wild-type cells is typically the result of ectopic insertion, although targeted disruption by reciprocal homologous integration is widely used to generate knockouts of known sequences via insertion of selectable markers (Aronson *et al.*, 1994b) at frequencies now routinely approaching 100% of transformants depending on the recipient strain, construct, and the locus (Colot *et al.*, 2006; Ninomiya *et al.*, 2004). Alternative methods for knockouts or knockdowns are also used with success: first is RIP (repeat-induced point mutation; Selker, 1990), a rapid method whereby duplicated genes are detected in a parental strain during a sexual cross and mutated prior to meiosis. “Knockdowns” are also made through quelling, a form of RNAi/cosuppression found in Neurospora (Cogoni and Macino, 1997).

Several independent estimates have suggested a surprising degree of genetic novelty in the Neurospora genome—repeated estimates suggest that over a third of Neurospora genes have no homologues or orthologues in GenBank (perhaps because of the 13 million sequences in GenBank, only a small fraction comes from Ascomycetes). These are either novel genes with novel functions or novel genes representing different ways of carrying out known functions; in either case, they will be of great interest. The complexity of the Neurospora genome approaches that of *Drosophila*. This degree of complexity—approximately twice that of yeasts—appears typical for filamentous fungi. Moreover, because of mechanisms in Neurospora that target and eliminate duplications (e.g., RIP), there are few gene families so that nearly all of the added sequence complexity reflects actual diversity. Neurospora shares gene sequences with a variety of taxonomic groups, so information from Neurospora informs projects covering the breadth of biology, not just within the fungi. Although both nonpathogenic and easy to manipulate, Neurospora is phylogenetically very closely allied and genetically syntenic, both with important animal and plant pathogens and with agriculturally and industrially important production strains (such as *Cochliobolus*, *Fusarium*, *Magnaporthe*, and *Trichoderma reesei*). Strong parallels have been noted in signaling

pathways, photobiology, developmental regulation, and many aspects of metabolism including secondary metabolism, to name a few. Information from *Neurospora*, especially in terms of functional genomics and regulation, is readily transferable. An understanding of the functional genomics of *Neurospora* provides a gateway to the genomes of the fungi.

C. Overview of the functional genomics effort

The overall effort to develop the functional genomics of *Neurospora* has been divided into four subprojects, each of which is partially dependent on and informs the others.

The first project is concerned with the systematic mutation of genes in the organism through targeted knockouts and is centered at Dartmouth and UC Riverside. The initial phenotypic characterization of knockout strains generated at these two sites is coordinated by investigators at UC Los Angeles (UCLA). This group shares a major ongoing commitment to minority undergraduate education, so this effort has enlisted significant involvement of minority undergraduates in the MBRS/IMSD, MARC, and CAMP programs at UCLA. In addition, they have seeded smaller efforts elsewhere to characterize additional genes. In addition to the phenotypic characterization, strains are archived at the Fungal Genetics Stock Center (FGSC) at the University of Missouri from which they are distributed to the research community at large. The generation of knockout constructs in Project 1 is informed by the annotation completed in Project 2 using the Expressed Sequence Tag (EST) sequences generated in Project 4, and knockouts created in Project 1 are being analyzed via microarrays in Project 3.

Annotation and Genomics Core work has been the focus of Project 2 centered at the Broad Institute, Massachusetts Institute of Technology, and Harvard (formerly the Whitehead Institute Center for Genome Research; WICGR) and at Oregon Health & Science University (OHSU). The sequence of the *Neurospora* genome was completed at WICGR (Galagan *et al.*, 2003), and automated annotation is ongoing there. OHSU has provided liaison with the *Neurospora* community in terms of linking the genomics with the genetic map and the existing *Neurospora* compendium of genes and phenotypes. Aims of Project 2 have been to (1) build a platform for electronically capturing community feedback and data about the existing annotation (gene calls), (2) build and maintain a database to capture and display information about phenotypes resulting from gene knockouts and disruptions, and (3) utilize data from ongoing EST analyses in Project 4 (as described below) to refine the gene structures.

Project 3, transcriptional profiling, began with the creation of DNA microarrays corresponding to the ~10,000 uniquely distinguishable transcripts in *Neurospora*. Centered at UC Berkeley, this project has provided a baseline analysis of gene expression for *N. crassa* under a variety of growth conditions and environmental stresses, and has begun to analyze the global effects of loss of novel genes in strains created by Project 1. These results provide the foundation for the effects of mutations on particular pathways of gene expression, and the availability of affordable microarrays has nucleated this technology in the community.

Progress in Project 4 at Dartmouth and the University of New Mexico has focused on EST analyses and creation of a Single Nucleotide Polymorphism (SNP) map, which has greatly facilitated the cloning of genes and analysis of complex genetic traits. To make maximum use of the products of this EST-sequencing effort, the Mauriceville strain of *Neurospora* is being used for construction of the cDNA libraries at OHSU as well as at Dartmouth and the University of New Mexico. This is an alternative wild-type strain that is fully interfertile with the Oak Ridge strain (sequenced to 16-fold coverage by the Broad), but that displays a sufficient number of nucleotide variants such that it has been used as the crossing strain for

an existing *N. crassa* restriction fragment length polymorphism (RFLP) map (Metzenberg and Grotelueschen, 1995). The SNPs arising from comparison of EST sequences from Mauriceville and Oak Ridge have provided the basis for constructing the SNP map. Additionally, unlike the apparent case in yeasts, alternative splicing and use of alternative promoters appear to contribute widely to the overall complexity of expressed sequences in *Neurospora* (Colot *et al.*, 2005), and there are known cases of long antisense transcripts (Kramer *et al.*, 2003). The Broad Institute has carried out limited EST sequencing of cDNA libraries and begun to use these data to establish the prevalence of alternative splicing and antisense transcripts. Project 4 has relied on resources from Project 2 and provided data essential for the annotation in Project 2 which, in turn, informed the creation of knockouts in Project 1. Additionally, EST sequences generated in Project 4 continue to inform the choice of probes used for the creation of microarrays in Project 3.

The interplay and interrelatedness among projects are shown schematically in Fig. 2.2. This serves as a useful scaffold for information flow as the efforts and progress of each project are considered in more detail below.

II. PROJECT 1: SYSTEMATIC GENE KNOCKOUTS

The goal of this project is straightforward. *Neurospora* has ~10,000 genes, a number that continues to fluctuate as annotation continues. Phenotypes are now associated with about 15% of these, and roughly a third of the genes have no strong sequence homologues in other organisms; thus, there are no clues to their functions. Furthermore, it is likely that functions of some genes sharing sequence similarity with genes in other organisms will be modified slightly or greatly, reflecting the novel biology of filamentous fungi. Our long-term goal is to create gene knockouts in all of these genes as a first step to determining each gene's function(s), and the immediate goal of Project 1 is to facilitate this. Following on the remarkable success and utility of the complete set of knockouts made in the yeast *Saccharomyces cerevisiae* (Martin and Drubin, 2003; Ooi *et al.*, 2006; Winzeler *et al.*, 1999), this approach scarcely requires defense.

There are several differences between this effort and the one originally carried out in yeast. First, this work follows on and benefits from that in yeast, and we tried to make good use of rationales, approaches, and lessons learned from that project. Second, many *Neurospora* genes are also found in yeast where they likely perform similar functions, so there may be less pressing need to examine them all. There are, however, many more genes in filamentous fungi as compared to yeasts, and in some cases the biology of a filamentous fungus may influence the functional significance of even known genes in unknown ways, so the magnitude of the endeavor here is somewhat larger than that faced by the yeast effort. Third, there exist several possible different ways in which gene function can be eliminated in *Neurospora*. One is RIP, which utilizes a *Neurospora*-specific phenomenon in which duplicated sequences, when passed through meiosis, undergo frequent C to T transitions thus usually resulting in loss of function (Selker, 1997). Additionally, genes can be targeted for “knockdowns” through RNAi or through “Quelling” (Cogoni and Macino, 1997). Finally and importantly, gene disruption yielding unambiguous nulls can be achieved by replacement through standard double homologous recombination. In this regard, unengineered *Neurospora* is more typically eukaryotic than yeast in that the frequency of homologous recombination rarely approaches 100% but lies instead between a few percent and 30% depending on the gene and the length of the homologous DNA flanking the gene in the construct.

A. Creation of gene knockouts in *Neurospora*

We chose to knock out genes in *Neurospora* by targeted gene replacement through homologous recombination. To achieve this, we needed to make a knockout cassette carrying a selectable marker for each gene to be replaced, to transform these cassettes into *Neurospora*, and to identify among the selected transformants the ones carrying the replacement but not other extraneous copies of the selectable marker. The knockout project can thus be thought of in three parts: construction of the cassettes, transformation, and examination of the transformants. At the outset of this effort we have devoted considerable time and energy to optimizing steps in this process, and in this section go over some of the rationale behind these choices as well as describing the final design of the work. This provides both greater detail and more insight into the process than is available in the initial description of this work (Colot *et al.*, 2006).

1. Creation of the knockout cassettes

a. The selectable marker: We chose the hygromycin phosphotransferase gene (Gritz and Davies, 1983) encoding resistance to hygromycin because (1) it is a very widely used marker in the filamentous fungal community; (2) it can be selected in *Escherichia coli*, yeast, and *Neurospora*; and (3) it is a dominant selectable marker, so the recessive auxotrophic mutations that are commonly used in genetic analysis do not have to be sacrificed. It is best when this bacterial gene is driven by a promoter from another fungus that contains no *Neurospora* sequences that could be mutated by RIP (Selker, 1997). We tried two different promoters for use in driving expression of the selectable marker: (1) the *Ashbya translation elongation factor* promoter p-TEF [driving *hph* (Goldstein and McCusker, 1999), total size of cassette is ~2 kb]; this is a weaker promoter but one that works in *E. coli*, yeast, and *Neurospora*; and (2) the *A. nidulans trpC* promoter (Pandit and Russo, 1992). In all cases the flanks of the cassette end in the recognition sequence for the restriction enzyme MmeI (TCCpuAC) for reasons that will be explained further below.

The knockout cassettes are generated by assembly of PCR-generated DNA fragments in yeast (Oldenburg *et al.*, 1997; Raymond *et al.*, 1999) (Fig. 2.3). The entire procedure through production of the final *Neurospora* knockout cassette was designed for high throughput to facilitate use of a 96-well format. Homologous recombination in yeast was, obviously, a cornerstone of the methodologies used for the yeast knockout project, where it was noted that there is a trade-off between the efficiency of recombination and the length of the oligonucleotide-derived overlap. Regarding the length of oligonucleotides used for gene replacement constructs, 45 base pairs (bp) of homology yielded 80% correct disruptants whereas 30 bp yielded only about 50% (Mark Johnston, personal communication). As a rule they used 45-mers. However, we used the overlap only to assemble fragments and determined that lower efficiency of recombination resulted in only a marginally lower yield of the correctly assembled fragments. In addition, we ultimately amplified the final linear knockout construct from crude yeast DNA preparations using PCR, thus eliminating any background due to incorrect or incomplete plasmids in the DNA used to transform *Neurospora*. Following this logic, we settled on overlapping homologous regions of 29 bp in the DNA fragments comprising the initial gene replacement “knockout cassette” used to transform yeast. This afforded considerable financial savings in the oligonucleotide syntheses necessary for the genome-wide knockout strategy.

b. The DNA template for PCR generating the fragments and design of primers:

Initially, we had believed that it would be necessary to use cosmid or BAC DNA as the template for the PCRs generating the to-be-assembled fragments. However, we determined that genomic DNA worked fine: 49-bp primers (having 20 bp of overlap with the appropriate genome sequence) and *Neurospora* genomic DNA plus polymerase in the

appropriate buffer yield products of sufficient mass that they can be directly transformed into yeast with no cleanup. This results in a considerable savings in time and supplies.

For primers to be gene-specific, the left and right flanks for each cassette must be tailored to individual genes. Thus, the choice of primers is a computationally intense problem. We used a custom-built application (written by John Jones, UCR) to retrieve regions adjacent to each Open Reading Frame (ORF) and pass them to Primer 3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi), which would automatically select a list of candidate primers with the following parameters: “PRIMER_GC_CLAMP 2,” “PRIMER_OPT_TM 56.0,” “PRIMER_MIN_TM 50.0,” “PRIMER_MAX_TM = 63.0,” “PRIMER_MIN_GC = 50.0,” “PRIMER_MAX_GC = 65.0.” Each selected primer pair was checked for uniqueness in the genome.

c. Molecular “bar coding” to individually identify knockout strains: In the yeast knockout project, each gene knockout was constructed so that it was marked by unique 20-bp sequences that provided novel sequence tags (“bar codes”) that could be used to identify the mutant. These short sequences were incorporated into the knockout cassettes in separate PCR steps, and thereafter into the genome of each mutant; they can be detected by PCR or by hybridization to a membrane containing the sequence of the tag. The construction to include bar codes required an extra PCR step (half again more primers and more labor) or longer primers, but because the cost of retrofitting such tags would be prohibitive they were included even though no experiments were planned for their use at that time. Subsequently, of course, they have proven to be quite useful. They provide unambiguous identifiers for each strain, and allow large numbers of deletion strains to be pooled and analyzed in parallel following selective growth. In yeast they can greatly reduce the cost and increase the efficiency of identifying phenotypes associated with deletion strains.

Generally, the utility of the tags in yeast has often hinged on the ability of yeast to grow without fusing, whereas *Neurospora* hyphae in culture rapidly anastomose to form local heterokaryons, so that rare recessive mutations are quickly complemented; this suggested that they might not be useful in *Neurospora*. However, even in yeast the present uses were not anticipated or even fully appreciated at the time the tags were first incorporated, but later developed as a result of their availability. We expected the same would be true in *Neurospora*, but the cost of the extra primers and PCR step to incorporate the yeast sequence tags was prohibitive. As a work-around, we developed what we hope will be an economical way in which to incorporate such molecular bar codes. The restriction enzyme *MmeI* has found use chiefly for SAGE analysis in which use is made of the fact that *MmeI* has a 5-bp recognition sequence but makes a 2-bp staggered cut 20-bp 3' from the recognition sequence. We incorporated *MmeI* sites at the right- and left-hand ends of the *hph* cassette (Figs. 2.3 and 2.4). Oligo 3f primes from the left-hand side of the rightward piece of genomic DNA flanking the *hph* gene in the knockout cassette and oligo 5r from the right side of the left flank. In the construction of knockout cassettes for each new gene, both oligos 3f and 5r must be individually synthesized since they must contain the gene-specific 20 bp as well as the portion overlapping the *hph* cassette; that is, they are gene specific and unique in the genome and thus fulfilled the prime requirement for a bar code and no extra cost (Fig. 2.4). When the knockout cassette is fully assembled, the *MmeI* sites will lie next to the “bar code.” When genomic DNA is digested with *MmeI*, many fragments will be generated, but only two will contain the right- and left-hand margins of the *hph* gene, respectively, and it should be possible to amplify the bar codes through ligation-mediated PCR, just as if they were SAGE tags (Fig. 2.4).

d. The construction method: In this first series of PCR reactions, three fragments are produced: (1) The “selectable marker” (*hph*) was amplified using a simple pair of primers,

generating a generic fragment that can be used for all knockouts. (2) Depending on the desired construct, 1–3 kb of sequence immediately upstream of the target ORF was amplified by a pair of 49-mers that use 20 bp to prime on the genomic template and add 29 bp of DNA homologous to left side of the vector gap (5f) and the left portion of the *hph* cassette (5r), respectively. (3) Similarly, the downstream-flanking sequence was amplified by a pair of 49-mers that use 20 bp to prime on the genomic template and add 29 bp of DNA homologous to right side of the vector gap (3r), and the right margin of the *hph* cassette (3f). By design, the last (5′) 6 bp of the 29-bp *hph*-related part of the 49-mers were the *Mme*I sites used as a part of our molecular bar-coding scheme.

In the next step, these fragments were transformed into yeast along with the gapped yeast shuttle vector. As originally implemented, the vector contained a cycloheximide sensitivity gene (*CYH2*) that provided a positive selection for recombination by being displaced on recombination of the left side fragment. However, our selection for circle closure and replication serves the same purpose and we found both selections redundant. We used lithium acetate transformation of yeast with success. After brief recovery, yeast cells were cultured in medium lacking uracil to stationary phase. The product of this step is a yeast strain harboring a shuttle plasmid containing the three fragments recombined in the proper order: left (5′) flank, *hph* cassette, and right (3′) flank. When driving *hph* with TEF, the resulting plasmid confers hygromycin resistance to the recipient yeast, but this was not the case using the *trpC* promoter.

This method of assembly is extremely robust and has only rarely failed, having been successfully used by five different investigators in two different laboratories. In later experiments, there seemed no reason to continue to monitor individual plasmids so PCR was simply used to generate the correctly assembled cassette from the mixed yeast on plates, and in these cases the correctly assembled cassette was recoverable by PCR in nearly all of the constructions.

PCR using LA Taq (TaKaRa), a genomic DNA template, and the primers as described above consistently gave very high yields of product such that a few microliters (ca. 0.5 μ g) of each PCR reaction can be combined directly with 0.5 μ g of the standard *hph* fragment and 100 ng of the gapped vector in the PEG/lithium acetate transformation. DNA from the transformed yeast culture was prepared by a streamlined protocol (Colot *et al.*, 2006) and was then used directly for the production of the final *Neurospora* knockout cassettes. The knockout cassettes were amplified using LA Taq and the same 5f and 3r primers that flank the ends of the *Neurospora* DNA in the yeast vector. The resulting linear DNA fragments were subjected to a PCR cleanup protocol and then frozen until ready for transformation into *Neurospora* (see below). The production of knockout cassettes was successfully automated using a Biomek NX robot (Beckman Coulter, Fullerton, CA) so that production was between 400 and 600 per week, and the cassettes for nearly the entire genome were completed within a year.

2. Transformation of *Neurospora*—Although we had limited data for transformation using circular DNA, it appeared that linearization of the gene replacement vector and/or gel purification of the insert prior to transformation reduced the incidence of ectopic integration events in transformants with the desired homologous recombination event. Also, higher rates of homologous recombination seemed generally correlated with larger flanks; for wild-type strains 3-kb flanks are the desired size. The logic of the full size cassette was the same as in the original yeast knockout procedure. A selectable marker (here *hph*) is flanked by DNA homologous to chromosomal DNA flanking a gene-to-be-deleted. On transformation, reciprocal homologous recombination on both sides of the targeted gene serves to replace the targeted gene with the selectable marker. Replacement transformants in which the

targeted gene was knocked out were identified based on the selectable marker, and confirmed by Southern analysis. The method worked dependably in that there has never been a gene that we could not delete in this way. However, it has the drawback that the entire selectable marker is incorporated into the knockout cassette, so for organisms in which ectopic insertion happens in a finite proportion of the total transformations (such as *Neurospora*), ectopic insertion events constitute a background to the desired single gene replacements. This problem is exacerbated by the fact that both an ectopic and a targeted replacement in a strain might provide as much as twice the drug resistance of a single replacement, so there is a mild selection in favor of ectopic events. One solution to this problem is to split the selectable marker into pieces (the split-marker technique), and another solution on which we settled was to use the stronger *A. nidulans trpC* promoter to drive hygromycin resistance so that there would be less marginal selection for ectopic transformations.

In split-marker transformation, the whole disruption cassette is supplied in the transformation cocktail as two separate pieces of DNA; neither piece on its own encodes *hph* but recombination recreates the complete gene. A successful gene replacement, then, requires recombination not only in the flanks but also within the overlapping central part of the selectable marker. This has the effect of lowering transformation efficiency. However, there is no longer any selective advantage to ectopic insertions, so that a higher proportion of the hygromycin-resistant strains contain only the desired replacements. The split-marker technique has been used in yeast (Fairhead *et al.*, 1996) and in *Cochliobolus* (Catlett *et al.*, 2003), and in our hands yielded gene replacements in 44% of transformants for which 68% were free from ectopic insertions (Colot *et al.*, 2006). These results were clearly adequate, and we were prepared to use a split-marker strategy to replace all the genes until a significant improvement appeared.

Work in yeast (Ooi *et al.*, 2001) has suggested that the Ku70 and Ku80 proteins are important for the nonhomologous end-joining process that gives rise to ectopic insertions. On the basis of this we originally proposed in our application for funding to delete the corresponding genes in *Neurospora* in hopes of improving the efficiency of homologous replacements; however, non-filamentous fungal-based members of the study section recommended that this aspect should be deleted. Fortunately, Inoue and colleagues independently came up with the idea, performed the experiments in *Neurospora*, and found that the resulting mutant strains were remarkably efficient at gene replacement (Ninomiya *et al.*, 2004). Because Inoue and colleagues used *hph* (the selectable marker in our knockout cassettes) to generate the $\Delta mus-51$ and $\Delta mus-52$ (homologues of *ku70* and *ku80*) mutants, we reengineered these mutants using *bar* as a selectable marker (confers resistance to phosphinothricin) (Colot *et al.*, 2006). In our hands, with the analysis of over 600 independent transformants by Southern analysis, we found that over 98% showed clean gene replacements inserting the knockout cassette in place of the resident gene with no accompanying ectopic insertion (Colot *et al.*, 2006). The great success of this method also meant that the relatively long 3-kb flanks were no longer needed and gave rise to the final plan using 1-kb flanks in the knockout constructs. This led to more efficient amplification of the shorter fragments using PCR, thus further streamlining our protocol.

3. High-throughput production of gene replacements in *Neurospora* to generate unambiguous functional knockouts—Knockouts are being generated in the Oak Ridge wild-type genetic background that was used for the 16-fold coverage genome sequencing at the Broad Institute (Galagan *et al.*, 2003). Since *Neurospora* naturally and rapidly forms heterokaryons, and because the conidia that will serve as transformation recipients have, on average, 2.5 nuclei per spore, primary transformants are heterokaryons that serve to shelter loss of essential genes in the transformed nucleus. We carry out the

transformation using the *Neurospora* knockout cassettes described above, and select for resistance to hygromycin. Primary transformants are then crossed to a closely related but opposite mating type Oak Ridge wild-type strain. Phosphinothricin-sensitive progeny (lacking the *bar* marker and the corresponding *mus* mutation) are then screened from the hygromycin-resistant progeny to select for strains in which a single gene knockout resides in an otherwise wild-type genetic background.

Final molecular verification of the knockout is provided by Southern analysis of hygromycin-resistant and phosphinothricin-sensitive progeny. We use a custom software program (written by John Jones) to both predict the best restriction enzymes to use for Southern analysis of a given gene and to provide the sizes of hybridizing fragments from the knockout and wild-type alleles. All Southern blots are probed using the full-length knockout cassette, thus allowing detection of any rare ectopic integrations. In addition to confirming the presence of knockout mutations and absence of ectopic integrations, this step verifies strain integrity. In practice, the software program has performed very well, greatly facilitating an extremely tedious manual procedure.

As soon as the confirmed homokaryons are generated in one or both mating types, they are shipped to the FGSC for distribution to the research community at large. Since some genes being disrupted prove to be essential to life, it is not always possible to recover hygromycin-resistant homokaryotic knockout strains from crosses. In such cases, we assume the gene is essential (either for life or for meiosis/spore viability) and the original heterokaryotic transformant is checked using Southern analysis (should contain both wild-type and knockout-hybridizing fragments) and then sent to the FGSC. To date nearly 2300 strains have been deposited and the fulfillment of requests to the FGSC makes the products of the knockout collection a heavily used resource within the FGSC's operations.

An up-to-date listing of knockout strains produced or in progress as well as protocols and methods for making or requesting knockouts can be found at [http://www.dartmouth.edu/~neurosporgenome/1_s1.html]. All strains, procedures, and reagents used for knockouts have been deposited with the FGSC and are available there.

a. Schedule and throughput: In production mode, we try to adhere to a schedule wherein a new palte of 96 strains begins every 3 weeks at both Dartmouth and UCR. Because it takes closer to 10–12 weeks to complete a full knockout construction and verification by Southern, this means that an investigator will have several overlapping cohorts ongoing at one time at varying stages of completion. To monitor progress, track strains, and house the primer design and Southern restriction enzyme software, we have implemented a Laboratory Information Management System (LIMS; written by John Jones; <http://www.borkovichlims.ucr.edu>). The current LIMS tracks the more than 30 steps in the entire knockout procedure. Plates and products at various stages of the protocol are labeled with physical bar codes that can be scanned by a hand-held bar code scanner. The progress of a strain or set of 96 strains can then be tracked through the entire procedure, culminating in assignment of an FGSC number to the final homokaryotic or heterokaryotic mutant strains. These bar codes can be recognized by personnel at the FGSC and UCLA using their own bar code scanners by logging in to the UCR database.

By maintaining overlapping cohorts of strains, one person can handle between eight and nine batches of knockouts ($\sim 8.5 \times 96 = 816$ genes) per year. In this scenario, two laboratories each with two people at the bench will produce about 3200 knockouts per year. If 65% of these are clean knockouts that are easy to characterize, this yields 2100 genes per year, with a total of just over 7000 by 2009. If 95% are clean knockouts, 3100 genes can be mutated each year, with just over 10,000 completed by 2009.

These perspectives raise the issue of what can and does go wrong. So far cassette construction (as noted above) has worked extremely well, with the only problems being rare operator errors, mistakes with the oligo supplier, and a very few problem sequences that may not be compatible with the yeast system. Clean ectopic-free gene replacements occur at a dependable rate of about 98%, so there are always a few strains that need to be redone, but this is not onerous. Verification by Southern that strains do indeed contain gene replacements and are free of ectopic insertions has been the greatest remaining bottleneck. This step requires growth of the strain, preparation of sufficiently pure genomic DNA to allow restriction digestion, followed by electrophoresis, blotting, probing, and interpretation of the results. As mentioned above, our custom-written software for choosing an appropriate enzyme and predicting the correct Southern pattern has immensely helped in this endeavor. We also have a robust nonradioactive method for generating probes and detecting hybridizing fragments. However, isolation of genomic DNA from recalcitrant strains that grow poorly and/or are hard to lyse has presented technical hurdles and required making modifications to the standard protocol in some instances. Other than this, another challenge is the continual process of genome annotation that periodically adds or deletes many hundreds of genes based on new molecular evidence from ESTs or comparative studies. Each change immediately sends that gene back to square one, and these will undoubtedly comprise the major source of uncompleted genes when the project winds down.

b. Choosing genes to disrupt and coordinating the effort: The most recent annotation version 3 of assembly 7 of the *Neurospora* genome has revealed 9826 potential protein-coding “genes,” although this number continues to fluctuate as a result of the manual annotation in Project 2 and EST analyses ongoing in Project 4. Although in the best of all possible worlds it may be possible to knock out most of these genes, there must be priorities assigned to which ones to do first. Our initial effort went into a proof-of-principle experiment where just over 100 genes encoding transcription factors were deleted and the phenotypes analyzed (Colot *et al.*, 2006). Subsequent efforts have been organized on the principle of generating the strains that will be of the greatest utility to the community. These have been chosen in two different ways.

First anyone can request that knockout strains be constructed and such requests immediately go into the queue. Such requests are made to a knockout-specific (knockouts@dartmouth.edu) and care is taken to be sure that all orders are handled anonymously so that a particular request is never associated with the name of the requestor. To date we have generated several hundred strains requested by the community at large, including some sizable blocks of genes such as >100 products associated with hyphal growth.

Second, we have chosen for disruption groups of genes that can form the basis of research projects or that are commonly among those that are needed for dissection of biological processes. These groups have included, for instance, the genes encoding the remaining ~100 transcription factors, genes encoding protein kinases and phosphatases, and chromatin-remodeling enzymes.

More than a third of the genes in *Neurospora* have no homologues outside of the filamentous fungi, so the elucidation of the phenotypes of such genes promises to inform much ongoing work among these organisms. With this in mind, as we move beyond the analysis of known genes, some general rules have been established:

1. Is it novel? If yes do it; if not, go on to 2.
2. Is it already associated with a phenotype deriving from a known mutation in *Neurospora*? If yes, skip it; if no, go on.

3. Is it in *Saccharomyces cerevisiae* or *Schizosaccharomyces pombe*? If no, do it; if yes, go on.
4. Is it essential in either or both yeasts? If yes, skip it; if no, go on.
5. Is it associated with an obvious housekeeping function in yeast, such as intermediary metabolism, such that the phenotype of a null is easily predictable? If yes, skip it; if no, do it.

The products of this high-throughput knockout effort are strains bearing a molecularly verified gene replacement marked by hygromycin resistance. Each tube is physically bar coded, this time with a bar code corresponding to the FGSC number. Each laboratory generating the knockouts retains a tube for each mutant, and a replicate bar-coded tube is sent to the FGSC for distribution to the community and, if needed, to UCLA for phenotypic analysis.

B. Basic phenotypic characterization of mutants

It is not feasible to carry out a comprehensive analysis of every phenotype caused by each knockout mutation. However, a preliminary characterization of each knockout strain—to provide enough basic phenotypic information to enable other researchers to productively use the set of mutants—has enormous added value. Because the techniques associated with this work are relatively straight-forward, and because *Neurospora* is absolutely nonpathogenic, we determined early on that it would be desirable to use the characterization of these strains as a vehicle to introduce undergraduate students, and in particular where possible underrepresented minority students, to *Neurospora* and microbiology in general. This was the origin of the very successful *Neurospora* Genetics and Genomics Summer Research Institute (NGGSRI) at UCLA where phenotypic analysis protocols for beginning science students have been developed. These methodologies are now being implemented at several other sites worldwide.

Novel knockout strains arrive from the FGSC, Dartmouth, or UCR. Two replicates are generated and serve as a backup and a student set. Using the student set, each student generates a separate stock for their individual knockout mutant analysis. Multiple students perform the analyses on an individual strain so that collectively each strain is examined in quadruplicate. This redundancy, along with supervision of the final data set, ensures quality control of the data.

Five assays are performed for each mutant analysis, which examine morphology, asexual development, growth, and sexual development using wild-type strains as a reference.

First, analysis of colony growth and morphology is performed on solid Vogel's minimal (VM) medium and on minimal medium plus yeast extract (VM + YE) at both 25 and 37°C. Petri dishes inoculated at their centers are cultured under ambient light/dark conditions and digital images recorded at 24 and 48 h [Infinity Camera with a Navitar Zoom 7000 lens (Lumenera Scientific, Ottawa, Canada)]. For each strain, the hyphae at the edge of the colony are photographed after 24 and/or 48 h using an S8 Apo Stereo Zoom microscope mounted with a DFC 280 digital camera (Leica, Wetzlar, Germany). Second, the rate of extension of basal hyphae is measured on race tubes (Dunlap and Loros, 2005) containing 13 ml of VM agar medium at 25°C in ambient light/dark conditions. Growth marks are recorded twice per day (morning and afternoon) over a 72-h period, the data graphed to verify linearity, and the growth rate expressed as mm/day. These procedures identify growth and morphological aberrations.

Third, the production of aerial hyphae and conidia is measured on slants containing VM medium grown at 25°C for 3 days and then put at room temperature for 3–5 days. Conidiation, pigmentation, and aerial hyphae are scored using the wild-type strains as a reference. The extension of aerial hyphae is measured in 13 × 100 mm test tubes containing 2 ml of VM or VM + YE as standing liquid cultures at 25°C. The top edge of the mycelial mat as seen after 24 h is marked on the tubes, the cultures incubated statically for a total of 72 h, and the difference in height in mm recorded. These results reveal mutants in asexual development.

Fourth, the sexual developmental pathway is analyzed in the following increments. Formation of protoperithecia (female sexual structures) are assessed following 7–8 days growth at room temperature (22–25°C) on plates or tubes of synthetic crossing medium (Davis and deSerres, 1970) containing either 0.1% or 1% sucrose. Fifth, plates or tubes are then fertilized with 10⁶ wild-type conidia of the opposite mating type, incubated at room temperature for another 7–8 days, and mature perithecial formation is scored using a stereomicroscope. Plates or tubes are then finally checked for ascospore development 2 weeks after fertilization.

The phenotype data are recorded and, along with the digital images, uploaded to the Broad Institute (Fig. 2.5). Students are given user names and passwords for the Broad and each establishes his or her own database. A summary tool developed by the Broad Institute allows access to all databases for the UCLA curator. The summary tool is organized by NCU No. and contains all images and data for every mutant analysis. Following curation and verification by the UCLA staff, the data are made public in the format that is shown in Project 2 (Fig. 2.8). To date nearly 400 strains have been characterized at UCLA and several additional phenotyping sites will come online in 2007 using the UCLA protocols.

C. Deposition of the strains in the Fungal Genetics Stock Center (FGSC) and their distribution to the scientific community at large

The FGSC receives bar-coded hetero- or homokaryotic strains from UCR or Dartmouth. The complete set of mutants as it is assembled is listed in the FGSC online catalog, and made available to the scientific community by the FGSC through its normal mechanisms. Protocols have now been developed whereby sets of 96 knockouts can be archived, replicated, and sent out in deep-well microtiter plates. Requests for individual knockout strains as well as for the entire emerging collection make this a heavily used resource within the FGSC's portfolio.

D. Summary of gene knockouts

The first systematic mutation of a eukaryotic genome was that of *Saccharomyces*, and the multiple impacts of that work have by all accounts been spectacularly successful. The products of this approach, the assembled mutants potentially addressing the function of each gene in a genome, constitute one of the central cornerstones of a modern research model system. Given the central importance of the filamentous fungi to medicine, to agriculture, and to industry, and considering their surprising genomic complexity and novelty, the *Neurospora* knockout collection is proving to be an invaluable resource.

III. PROJECT 2: GENOME INFORMATICS AND FUNCTIONAL ANNOTATION STUDIES

A. Introduction

To increase the value of the *N. crassa* genome sequence we must continue to refine the genome annotation and also capture and integrate with the genome sequence the wealth of

information within the research community about the genes and genetics of the organism. To this end, the Annotation and Genomics Group of the NIH Neurospora Program Project has the specific aims of (1) generating EST data and using them to both improve gene predictions in the genome and delineate single nucleotide polymorphisms (SNPs) between the Oak Ridge and Mauriceville strains, (2) building a platform for capturing and curating community input about the genome annotation, and (3) building and maintaining a database to capture and display information about phenotypes resulting from gene knockouts and disruptions.

The Annotation and Genomics Group of the Neurospora Program Project is centered at the Broad Institute of MIT and Harvard with an associated community-based annotation effort centered at OHSU. It uses the Broad's Calhoun system as the main infrastructure. Calhoun is an informatics system developed for whole genome annotation and analysis. The system is based on a modular and extensible architecture, and has been applied successfully to the genome annotation and analysis of a multitude of fungi, microbes, and other organisms sequenced through the Broad's Fungal Genome Initiative and Microbial Sequencing Center. The features of Calhoun include (1) an Oracle relational database for the storage and management of genome sequence data, (2) an extensive Application Programming Interface (API) for data insertion, retrieval, and manipulation, (3) a high-throughput sequence analysis pipeline, (4) extensive tools for data mining, and (5) sophisticated user interfaces and client tools. Calhoun's architecture is organized into data storage, data interface, analysis, and presentation layers.

B. Improving automated genome annotation

To realize the full potential of genomic sequence, important genetic features must be identified, and these features must be associated with biologically relevant functional information. High-quality gene annotation is the starting point for harnessing the power of genome sequence. For example, an accurate annotation of the location of genes is essential for the development of microarrays, and errors and revisions to the number, structure, and locations of genes have major impacts on the quality and cost of gene knockout efforts.

The task of identifying all genes in a genome and associating functional information with them is not a completely solved problem. Many tools exist for computationally predicting the location and structure of genes. These include *ab initio* gene prediction tools, homology search tools, and sequence alignment tools. However, these tools do not produce perfect predictions, and their ability to identify genes in a particular organism depends heavily on additional supporting data. Among the different forms of evidence used for gene calling, EST sequences are especially valuable both as training sets for *ab initio* gene predictors and as raw data for use by gene-building systems.

Neurospora EST data from previously characterized libraries (Bell-Pedersen *et al.*, 1996; Nelson *et al.*, 1997; Zhu *et al.*, 2001) have been valuable for training gene predictors, validating automated gene predictions, and constructing microarrays for functional studies. Although the Neurospora gene annotation is highly accurate in detecting gene loci, the fine structure (introns/exon boundaries, start and stop codons, untranslated regions, and alternative splicing) still requires refinement. Additional EST coverage is the surest method for improving gene structure prediction. Currently only ~1/3 of Neurospora genes have EST support. To expand this coverage, the Broad Institute is sequencing ESTs generated for the Mauriceville strain by the Neurospora Program Project and integrating these EST data into the automated annotation of Neurospora genes. The Broad has built and continues to refine an automated gene-calling pipeline that uses multiple gene prediction algorithms and selects the best gene call using multiple forms of evidence including ESTs. ESTs are the highest ranked form of evidence used to call a gene structure from the structures derived by the

prediction algorithms. When EST coverage is available, gene predictions can be evaluated for (1) Spurious Predicted Introns: predicted gene structures with EST alignments completely spanning an intron (without corresponding gaps in the alignment), (2) Missing Exons: EST alignments not in predicted coding regions, and (3) Incorrect Splice Sites: predicted gene structures with EST alignments covering an intron and an adjacent exon. The gene predictions with the most agreements to the aligned ESTs for (1) through (3) are selected (Fig. 2.6A). In addition, the ESTs are used to define untranslated regions (UTR) and alternative transcripts (Fig. 2.6B). The Broad Institute is currently implementing an algorithm that directly use EST data for gene prediction versus only for gene calling.

In addition to using the new ESTs in gene calling, the Annotation and Genomics Core is using the ESTs to identify SNPs that are in turn being used to build a new SNP-based genetic map of *Neurospora* ESTs from the highly polymorphic Mauriceville strain are automatically aligned to the Oak Ridge genome sequence using the Calhoun infrastructure (Section V). To increase the specificity in our SNP identification, SNPs are defined using the neighborhood quality scoring algorithm (NQS) developed at the Broad Institute (Altshuler *et al.*, 2000). A polymorphism is called a SNP if there is a mismatch between two bases in the alignment, each base has a PHRED quality score of at least 25, and that polymorphism is in a neighborhood of bases (defined by five bases upstream and downstream of the mismatch) each with PHRED of at least 20. SNPs delineated using NQS are then validated using both PCR-based and restriction digest-based validation methods at participating laboratories in the *Neurospora* Program Project. Validated SNPs are displayed on the genome sequence and integrated with the *Neurospora* genetic map.

C. Community annotation

Computational methods alone are not adequate to produce the highest quality genome annotation; manual annotation and curation are also necessary. Automated gene prediction systems produce systematic as well as specific errors that are difficult to correct without manual intervention. Furthermore, automated methods remain incapable of tapping into the vast wealth of knowledge about the genes and genetics of an organism that are collectively contained within the research community.

One way to provide for manual annotation and to capture community knowledge is to create an infrastructure that enables researchers in the community to submit and curate gene annotations. Although variations on this theme exist, they are typically referred to as "Community Annotation Projects" (CAP). The Annotation and Genomics Core has developed a robust infrastructure called CAP that uses the automated annotation of the *Neurospora* genome as a scaffold onto which expert knowledge about genes is mapped to the genome. CAP is powered by Calhoun and is accessed through the Broad Institute's web interface for the *Neurospora* genome pages. CAP provides researchers the ability to improve the *Neurospora* annotation by integrating information including: (1) refined gene structures and alternative splices, (2) gene symbols, (3) gene synonyms, (4) gene product names, (5) functional information such as ontology terms, (6) associated literature, and other information directly onto the genome. Community annotations made using CAP are attached to the official annotation and CAP provides tools to make these community annotations part of the official annotation through a curation process (Fig. 2.7). CAP annotations are immediately searchable and retrievable by all users once saved into the Calhoun database. Until made official, CAP annotations are visualized in a summary format at the bottom of a "gene detail page" (Fig. 2.7); clicking on a CAP "ID" opens a detailed view of the CAP annotation. Community annotations can be viewed, edited, as well as commented on by other members of the research community.

More than 60 years of research on *Neurospora* constitutes a vast resource of information contained within the *Neurospora* community for interpreting the genome sequence. Capturing a portion of this information in the genome annotation will provide an invaluable resource for *Neurospora* researchers and the wider scientific community. CAP provides this opportunity. For example, the large amount of data in Alan Radford's electronic version of the *Neurospora* compendium (Perkins *et al.*, 2001) has been incorporated into CAP, as have the new gene names and phenotypes associated with the transcription factor knockouts generated in the initial work of Project 1 (Borkovich *et al.*, 2004; Colot *et al.*, 2006).

Finally, as part of the CAP project, a web-accessible Textpresso database of the *Neurospora* literature has been developed to enable rapid full-text searching of papers that refer to this organism. This resource (www.textpresso.ebs.cgi.edu) is useful for annotation and for providing access to specific information in this literature. The Textpresso application, originally created by researchers at Caltech for curation of the *Caenorhabditis elegans* literature (Muller *et al.*, 2004), has been used by us to build a *Neurospora* database. Textpresso accepts queries to search for specific words alone or in combination, as well as ontology relationships defined by specific categories. It responds by providing the queried terms in the original context of the research paper in which they are found, as well as links to PubMed. Thus, for example, it is possible to search the literature for instances in which specific alleles of genes are used by searching on the allele name or to search for instances in which specific compounds are named, and rapidly find the specific contexts in which the alleles or compounds have been discussed.

D. Annotation of allele-specific phenotypes

The automated annotation of the *Neurospora* genome sequence has revealed 9826 genes, many of which have as yet no clear homologues in other organisms. To provide functional annotation of these genes, the results of large-scale gene disruption analyses of these genes must be mapped to the genome and put in the context of what is already known about *N. crassa* mutant phenotypes. Mutations in *N. crassa* can be considered in terms of a variety of major phenotypic effects. The most fundamental is how they affect viability. Such observations date from the demonstration in the 1940s that single gene mutants in *N. crassa* caused inviability at high temperature, important evidence for showing that the one-gene one-enzyme theory did not apply only to dispensable genes such as those that caused auxotrophy. Major common types of phenotypic effects include alterations in morphology, physiology, and development. More specific effects include alterations to nutritional requirements, sexual fertility, the circadian clock, the capacity to form heterokaryons, and to posttranscriptional gene silencing. Integrating with the genome both existing phenotype data as well as new information obtained from ongoing knockout experiments that are part of the *Neurospora* Program Project will leverage both sequence and phenotype data to enable functional annotation.

To this end, the Annotation and Genomics Group developed, maintains and continues to refine the database that captures and displays information about alleles derived from the phenotypic analysis of mutants carried out at UCLA and other sites as described in Section II.B. This infrastructure is powered by Calhoun and is accessed through a web interface. Like CAP, the genome annotation is used as a scaffold onto which allele phenotypes are mapped. The system has the ability to track multiple alleles associated with a gene, capture defined phenotypic characterizations, and obtain such information for multiple assay conditions (Fig. 2.8). Researchers characterize knockout strains using a defined vocabulary and images of the mutant's morphology and then upload this information using a web-based data-entry form (Section II.B). To maintain consistency in the annotation of phenotypic data and provide a database that is fully searchable, a controlled vocabulary is used to define phenotypes. To ensure this annotation meets the scientific needs of the *Neurospora*

community and is valuable to the wider fungal community, a substantial effort is dedicated to establishing and maintaining the vocabulary for describing abnormal phenotypes. This task is undertaken by an Annotation Steering Committee.

Once characterizations are expertly curated using tools developed by the Broad they are added as a feature in the gene annotation and can be searched and viewed through the Broad Neurospora webpages. Since allele summaries (Fig. 2.8) are stored as a feature of a gene they are accessed directly from a gene's detail page; alternatively they are retrievable by browsing a table of phenotype characterizations and associated genes on the Neurospora alleles homepage.

The Annotation and Informatics Group of the NIH Neurospora Program Project continues to enhance the structural and functional annotation of the Neurospora genome. This is accomplished through the development of community annotation and allele phenotype characterization infrastructures that for the first time provide the research community the ability to link information about genetic features with the *Neurospora* genome, to refine gene structures and to curate all entries in a searchable database. Gene models are also improved using EST sequence that moreover enable the generation of an SNP-based genetic map. All of these activities will provide the high-quality annotation of the Neurospora genome that will empower researchers to new discoveries in eukaryotic biology.

IV. PROJECT 3: PROFILING TRANSCRIPTION IN *NEUROSPORA*

The biology of filamentous fungi remains relatively unexplored, especially when compared to fungal model yeast species, such as *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. As mentioned above, in contrast to yeasts, filamentous fungi are characterized by a complex life cycle, and there are at least 28 distinct cell types in *N. crassa* (Bistis *et al.*, 2003; Borkovich *et al.*, 2004). To help to provide a solid foundation for understanding this interesting and yet tractable level of complexity, and to enhance the technology for inference of gene expression levels in *Neurospora*, we are creating whole-genome-spotted oligonucleotide microarrays. Spotted microarrays allow the measurement of the abundance of transcripts from thousands of genes simultaneously by the competitive hybridization of labeled cDNA transcripts (targets) to immobilized probes cross-linked to the surface of amine-coated microscope slides (Derisi and Iyer, 1999; Eisen and Brown, 1999). Gene expression data generated by microarray technology can be used to expedite the annotation process of the *N. crassa* genome and will aid in assigning putative functions for unique genes.

A. Oligonucleotide design and synthesis

We designed 70-mer oligonucleotide-immobilized probes from 10,526 open-reading frames (ORFs) predicted from the *N. crassa* genome sequence (Broad Institute, http://www.broad.mit.edu/annotation/fungi/neurospora_crassa_7/index.html and MIPS, http://pedant.gsf.de/cgi-bin/wwwfly.pl?Set=Ncrassa_annotations&Page=index) using the bioinformatic tool ArrayOligoSelector (Bozdech *et al.*, 2003). Array OligoSelector identifies a unique 70-bp segment to represent each ORF, avoiding self-annealing structures and repetitive sequences. It preferentially chooses oligonucleotides that are located close to the 3'-terminal region of each gene and that conform to a narrow range of GC content. A total of 10,910 70-mer oligonucleotides were synthesized by Illumina, Inc., San Diego CA. These represent the 10,526 predicted genes as well as an additional 384 70-mer oligonucleotides designed for intergenic or telomeric regions. *Neurospora* full genome microarrays are printed onto γ -aminopropyl silane slides and are available at cost from the FGSC (<http://www.fgsc.net/>). Information on the oligonucleotide gene set is available at the *Neurospora*

Functional Genomics Database (<http://www.yale.edu/townsend/Links/ffdatabase/introduction.htm>).

B. Experimental design for microarray experiments

Two samples of target cDNA are competitively hybridized against the immobilized probe in spotted DNA microarray hybridizations. Thus, primary analysis yields the ratios of gene expression between two samples. However, most experimental designs incorporate multiple developmental, genetic, and/or environmental states. Inference of gene expression level in multiple states from ratiometric measurements requires judicious experimental design. In closed circuit designs (Fig. 2.9), each strain is compared head-to-head with other strains, in a circular or multiple-pairwise fashion. The ability to detect differences is maximized because the comparisons are directly between individual strains or conditions of interest, and the problems of using a common standard (Townsend, 2003) are avoided. The sole disadvantage of this method is that immediate inference from the raw data is not easy because ratios observed across multiple pairwise comparisons do not all compare or contrast in an intuitive way. However, this is a problem that is rapidly solved by accessible methods of analysis (Kerr and Churchill, 2001; Townsend and Hartl, 2002; Wolfinger *et al.*, 2001). For instance, a Bayesian analysis of gene expression levels (Townsend and Hartl, 2002) uses all transitive and direct comparisons from any replicated, interconnected experimental design to infer relative gene expression levels and 95% confidence intervals. The results are reported with gene expression in the sample with lowest expression as one unit; the samples are scaled appropriately. Although the inferred expression levels are of arbitrary unit scale, this scaling has the intuitive appeal that all gene expression level measurements are positive, as they should be. Circuit designs of microarray comparisons have been strongly endorsed by statisticians (Kerr and Churchill, 2001; Yang and Speed, 2002) and have demonstrated dramatically improved resolution with regard to identifying differential gene regulation below the twofold level (Townsend and Hartl, 2002; Vinciotti *et al.*, 2005).

Subsequent hierarchical clustering of inferred levels of gene expression can greatly assist in understanding the function of previously uncharacterized genes as there is a tendency for sets of coexpressed genes to be involved in common cellular functions (Eisen *et al.*, 1998). Data can be analyzed by multiple computational methods, including self-organizing maps (SOMs), k-means clustering, and principal component analysis. Motif searches conducted, using programs such as BioProspector (McGuire *et al.*, 2000), MDscan (Liu *et al.*, 2002), and MEME (Bailey and Elkan, 1994), can be utilized to identify *cis*-regulatory elements attributable to the coregulation of gene clusters using profiling data from *N. crassa* (Kasuga *et al.*, 2005).

C. Technical aspects to transcriptional profiling: RNA extraction, cDNA labeling, image acquisition, and normalization procedures

Protocols for RNA extraction and cDNA labeling are available at the *Neurospora* Functional Genomics Database (<http://web.uconn.edu/townsend/Links/ffdatabase/introduction.htm>). Briefly, total RNA is extracted from samples using TRIzol reagent (Invitrogen Life Technologies, Burlington, ON). A 100- μ g sample of total RNA from each sample is cleaned using RNeasy Mini Protocol for RNA cleanup (Qiagen, Valencia, CA). For cDNA synthesis, 20 μ g of total RNA is mixed with 5 μ g of an anchored 17-mer oligo dT and 3.3 ng of ArrayControl single RNA spike mixture (Ambion, Austin, TX). cDNA is synthesized in a final volume of μ l with 500 μ M each of dATP, dCTP, and dGTP, 300 μ M of dTTP, 200 μ M of aminoallyl-dUTP, 10-mM DTT, and 100 units Stratascript reverse transcriptase (Stratagene, La Jolla, CA). For conjugation to fluorescent dyes, 10 μ l of 0.05-M sodium bicarbonate is added to the monofunctional NHS-esters of Cy3 or Cy5 (CyDye Post-Labeling Reactive Dye, Amersham Bioscience, Piscataway, NJ) and 5 μ l of the dye solution

is added to the cDNA solution. The labeled cDNA is purified with the CyScribe GFX Purification Kit (Amersham) and dried under vacuum.

Slides are prehybridized and labeled cDNA is resuspended in hybridization buffer and pipetted into the space between a microarray slide and a LifterSlip cover glass (Erie Scientific, Portsmouth, NH). An Axon GenePix 4000B scanner (Axon Instruments, CA) is used to acquire images and GenePix Pro 4.1 software is used to quantify hybridization signals.

To evaluate the ratio of mRNA from comparative hybridizations for normalization purposes, control RNA spikes are used as internal standards. The control spikes consist of eight polyadenylated bacterial mRNAs at concentrations ranging from 50 to 1000 pg/ μ l, which are complementary to eight ArrayControl Sense oligonucleotides (Ambion). The ArrayControl oligonucleotides are added as duplicate spots to the *N. crassa* oligonucleotide microarrays. A total of 3.3 ng of each control mRNA spike is added to each of the 20- μ g total RNA samples for each time point.

D. *Neurospora* functional genomics microarray database

We constructed a *Neurospora* microarray database in the public standard minimal information about a microarray experiment (MIAME) format (Brazma, 2001) (<http://web.uconn.edu/townsend/Links/ffdatabase/introduction.htm>). MIAME is a reporting protocol that describes the minimum information required to ensure that microarray data can be easily interpreted and that results derived from its analysis can be independently verified. Using this reporting protocol will facilitate the database deposition into public repositories and enable the development and use of novel data analysis tools. The *Neurospora* microarray database stores raw and normalized expression data from microarray experiments and also provides detailed discussions and information on experimental design, data analysis methods, and web interfaces for scientists to retrieve, analyze, and visualize their data. In the future, this database will be integrated into the *N. crassa* database at the Broad Institute (http://www.broad.mit.edu/annotation/fungi/neurospora_crassa_7/index.html).

E. Proof-of-principle: Transcriptional profiling of conidial germination

Filamentous fungi undergo complex asexual and sexual developmental programs. In addition, their mycelial growth habit differs substantially from that of unicellular organisms such as *Saccharomyces cerevisiae*. However, since both *N. crassa* and *Saccharomyces cerevisiae* are fungi, and extensive profiling experiments have been performed on *Saccharomyces cerevisiae* (see Saccharomyces Genome Database; <http://www.yeastgenome.org/>), a comparison of transcriptional profiles of these two species in response to environmental and nutritional factors, DNA damage, and the cell cycle will be especially informative. Other aspects of the life cycle in *N. crassa* have no obvious counterpart in *Saccharomyces cerevisiae*, such as development of female reproductive structures, formation of asexual spores, creation of the interconnected mycelium, and asexual spore germination (Davis, 2000). As a proof-of-principle experiment for the development of oligonucleotide arrays, we performed transcriptional profiling of conidial germination in *N. crassa* on a partial genome array comprising 3366 immobilized probes for predicted genes (Kasuga *et al.*, 2005). Conidial germination in filamentous fungi is a highly regulated process that is triggered by environmental stimuli. Biochemical and morphological aspects associated with conidial germination have been well documented in *N. crassa* (for review, see Bonnen and Brambl, 1983; Denfert, 1997; Riquelme *et al.*, 1998; Roca *et al.*, 2005; Schmidt and Brody, 1976). However, fundamental genetic mechanisms, such as those

that drive the germination process and underlie the timing of gene expression and metabolic pathway activation, remain obscure.

We chose a closed-circuit experimental design to determine the expression levels of genes relevant to conidial germination (Fig. 2.9) (Kasuga *et al.*, 2005). RNA was isolated from eight time points during the germination process: times 0 and 30 min, and times 1, 2, 4, 8, 12, and 16 hours post inoculation. In this circuit of experimental comparisons, each sample was compared head-to-head with other samples, in a circular, and in some cases, multiple-pairwise fashions. We obtained precise statistical estimates of expression levels for 1287 genes during the process of conidial germination. Estimates of gene expression levels were remarkably consistent with previous data assessing transcript levels of a number of genes and with biochemical processes that have been associated with conidial germination (Kasuga *et al.*, 2005; Sachs and Yanofsky, 1991; Schmidt and Brody, 1976).

Of the 1287 genes for which strong estimates of gene expression level were acquired, 473 have been described as hypothetical, conserved hypothetical, or putative genes. Gene expression data for the remaining 814 genes with functional annotation were evaluated by the MIPS functional catalogue (FunCat) (Kasuga *et al.*, 2005), which is an annotation scheme for the functional description of proteins (Ruepp *et al.*, 2004). Concordance was apparent between predicted function of transcriptionally regulated genes and previously identified biochemical processes associated with conidial germination (Fig. 2.10). For example, a large number of genes belonging to the FunCat category ribosomal biogenesis showed maximum expression between 1 and 4 h after conidia are induced to germinate by inoculation into growth medium, consistent with biochemical data indicating that RNA and protein syntheses are activated soon after the induction of germination (Schmidt and Brody, 1976). Many genes in the FunCat cell cycle and DNA processing showed maximum expression between 30 min and 4 h after inoculation, consistent with biochemical observations that the initiation of DNA replication occurs ~2 hours postinoculation (Schmidt and Brody, 1976). Transcriptional profiles for some of the 473 hypothetical, conserved hypothetical, or putative genes correlated well with biochemical processes associated with conidial germination. Thus, microarray data in *N. crassa* will guide future laboratory experiments with regard to functional annotation of hypothetical genes.

F. Future prospects

1. Baseline transcriptional profiling of filamentous fungal colonial growth—

The mycelial colony of filamentous fungi is the hallmark of this group of organisms. We currently use full genome arrays (10,910 oligonucleotides) to obtain baseline transcriptional profiling data growth and reproduction in *N. crassa*. For example, we profiled *N. crassa* vegetative growth by comparing the transcriptomes of ~2-, ~7-, and ~12-h old sections of a colony grown on solid medium (Fig. 2.11). Expression of over 7000 genes was detected at statistically significant levels; 72 genes showed statistically significant differences in expression level between 2- and 12-h old hyphae within a *Neurospora* colony. As an example, expression of *cgc-1* at 12 h was 30-fold higher than in 2-h old hyphal tips (Fig. 2.11B). In contrast, *hex-1* (encoding the structural element of the Woronin body) expression was fivefold higher in the colony periphery versus 12-h old hyphae (Fig. 2.11B). These results are consistent with published Northern RNA blot results for these genes (Tey *et al.*, 2005).

2. Deciphering the transcriptional regulatory network of *Neurospora*—

The *N. crassa* genome encodes ~176 transcription factors with five DNA-binding motif families, that is, basic helix loop helix (HLH), bZIP, zinc finger C2H2 (*z*²C2H2), GATA zinc finger, and Zn(2)-Cys(6) binuclear cluster (Zn2Cys6) (Borkovich *et al.*, 2004). We have performed

phylogenetic analysis of predicted transcription factors of *N. crassa* as compared to other filamentous fungi, including *Saccharomyces cerevisiae*. Many of the predicted transcription factor genes have close orthologs in other filamentous fungi but not to *Saccharomyces cerevisiae*. The vast majority of these transcription factors are completely uncharacterized. Transcriptome analysis of three mutants (NCU00340, a *ste-12* ortholog; NCU03725, an NDT80 homologue; and NCU07392, a zinc binuclear cluster TF) have been conducted. As an example, the NCU00340 mutant showed 129 differentially expressed genes. A putative cis-element (CATCNTCAT) was enriched in the downregulated gene set ($p = 0.00017$, fisher test). The identified cis-element of NCU00340 putative target genes does not show similarity to the *Saccharomyces cerevisiae* Ste12p DNA-binding site (ATGAAAC) (Zeitlinger *et al.*, 2003). These data suggest that at least some transcriptional regulatory networks have diverged in fungi, as previously documented for Rpn4 orthologs (Gasch *et al.*, 2004). Further analysis of transcription factor mutants in *N. crassa* in comparison to wild type will reveal transcriptional regulatory networks that have been conserved among fungi and others that have diverged among filamentous fungi, yeast, and more distantly related eukaryotic species.

V. PROJECT 4: cDNA LIBRARIES AND THE GENERATION OF A HIGH-DENSITY SNP MAP

A. Introduction and rationale for the design of the project

SNPs are the most common genetic variants between individual genomes. Dense SNP maps can be used to precisely and rapidly map single gene mutations as well as polygenic traits and quantitative trait loci (QTL). This approach has been exploited for the identification of genes in the nematode (Wicks *et al.*, 2001), the fruit fly (Berger *et al.*, 2001), and yeast (Winzeler *et al.*, 1999), as well as many other organisms. An eventual future application would be the generation of oligonucleotide microarrays based on the identified SNPs. All SNPs in a strain could then be typed in parallel in a single experiment, by hybridizing under conditions that disallow the formation of stable hybrids if there is a single mismatch (or SNP).

We reasoned that by simply using a wild-collected *N. crassa* strain for the isolation of mRNA, one with many nucleotide differences from the standard laboratory (Oak Ridge) strain, one could generate the sequence information required to construct a detailed SNP map for *Neurospora*. Moreover, if the strain of *N. crassa* was genetically close enough to Oak Ridge, the ESTs generated would serve to bolster the gene calling and annotation aspects of this effort while still providing the genetic variability necessary for an SNP map. The wild-type *Neurospora* strain isolated in Mauriceville Texas is the strain used for RFLP mapping in *Neurospora*; it is fully interfertile with *N. crassa* Oak Ridge, but since it is sufficient for identification of RFLPs we reasoned that it would be distinct enough that SNPs could reliably be found in the same manner that RFLPs have been reliably found for the past 16 years (Metzenberg *et al.*, 1985). This was confirmed by comparing several noncoding regions of the Mauriceville and Oak Ridge genomes: SNPs were detected at a rate of ~2% in our hands, data comparable to published reports showing variable frequencies, from 0.2% to 2.1%, in three *Neurospora* ORFs (Table 2.1).

On the basis of averaged data from the whole genome (Section 2 Galagan *et al.*, 2003), *Neurospora* genes appear about every 3.7 kb along the chromosomes. The average transcript is 1.3 kb in length, and preliminary data suggested that the frequency of SNPs within coding regions is, conservatively, >0.2%. This suggested that the average cDNA would identify several SNPs. Since the rate of recombination in *Neurospora* is low compared to *Saccharomyces*, 1 map unit (1% recombination) corresponds to about 20 kbp in a typical

region in the middle of a chromosome arm (McClung *et al.*, 1989). A 20-kbp region might on average, then, contain 5 genes and 10–40 SNPs at saturation. These data suggested that sequencing of genomic DNA would not be necessary to ensure an adequate distribution of SNPs across the genome unless there are areas that have a paucity of genes or are, mysteriously, devoid of SNPs while still encoding transcripts.

Analysis of EST sequences from Mauriceville should reveal a trove of SNPs that can be used to augment the genetic map and greatly improve the efficiency with which informative mutations can be mapped. An application of RFLP/SNPs using a similar strategy in *C. elegans* demonstrated mapping to chromosome arm, and subsequently to cosmid, within a day after initial progeny were scored for presence/absence of a marker to be mapped (Wicks *et al.*, 2001). Overall then, this project would win in two ways, both by supplying needed EST-based data to bolster gene calling for annotation in Project 2 and in providing the basis for an SNP map.

B. Construction of the map

1. Source of SNPs—A cDNA library was made from FGSC 2225 (*N. crassa* Mauriceville, A). Conidia were harvested and inoculated into liquid culture (2% glucose, 1 × Vogel's salts, 0.5% arginine, and 0.05-μg/ml biotin) and grown at 30°C in light for 4 h. The germinated conidia were harvested, total RNA isolated poly A + RNA from this. One microgram of this Mauriceville mRNA was used to synthesize cDNA (Bell-Pedersen *et al.*, 1996), which was sized on a Sepharose CL-2B column. Two cDNA fractions were ligated with the Uni-ZAP XR (Stratagene; precut with XhoI and EcoRI) vector, and the ligation products packaged *in vitro* (Stratagene Gigapack III kit), yielding a titer of about 3.3×10^4 pfu/ml. Following infection into XL1-blue MRF bacterial cells, phagemids were mass-excised and introduced into an SOLR strain to get a stable plasmid form. 5695 of these plasmids derived from germinated conidia of a wild-type strain of *N. crassa* from Mauriceville, Texas were then sequenced at the Broad Institute (Fig. 2.12A).

2. Identification of SNPs within ESTs—To be maximally useful, the SNP data must be integrated with the sequence and genetic maps of *Neurospora*; this integration will be achieved through collaboration with the Broad Institute (Section III.B). Briefly, traces from EST sequencing are aligned with the assembled genomic sequence (from the Oak Ridge wild-type strain) and the top-scoring hit identified as the corresponding gene from Mauriceville. Due to RIP and other duplication scanning mechanisms (Selker, 1997), few problems with identification of duplicated sequences in the *Neurospora* genome were anticipated, so Mauriceville ESTs would have one and only one corresponding gene in Oak Ridge; this has been the case so far. The sequences are aligned (or multiply aligned where more than one sequence is available due to multiple ESTs being sequenced or to overlap of the traces from each end of an EST) and the probability of a polymorphism determined. Since the error rate in the genomic sequence is well under 1/10,000 (Project 2) equivalent to a PHRED of 40, putative SNPs are called only if they fall within a region of otherwise high PHRED in the EST sequence. Likewise, for insertions and deletions, a SNP is inferred only if the PHRED value on either side of the gain/loss is high, and if the insertion/deletion does not occur within a region prone to compression artifacts. It is a truism that some SNPs will create RFLPs, so we will be able to independently verify our assignment criteria, in addition to amplifying the existing RFLP map and tying it directly to the assembled genomic sequence. Since the SNPs are being generated in the context of an assembled genome, they are already mapped with regard to each other, and can be used directly for fine mapping of novel genes.

Specifically then, sequences arising from the ESTs were aligned with the Oak Ridge genome using BLAT (Jim Kent, UCSC) for SNP detection. A mismatch that meets the NQS (Neighborhood Quality Standard) is called SNP; its PHRED quality score is higher than 25 and the five bases to either side display a PHRED score higher than 20 and are not mismatched themselves. This algorithm was adopted for detecting SNPs in human genome with an accuracy rate around 95% (Altshuler *et al.*, 2000). As predicted, SNPs have appeared among the sequences at a raw frequency of about 1 per kilobase. They are less common among coding sequences than in noncoding transcribed regions or intergenic regions, but there are still plenty of SNPs within genes for purposes of map construction.

3. Detection of SNPs—In all cases, SNPs are detected among cross progeny using PCR-based strategies, but two different methods have been used successfully in our hands. The first is differential amplification (<http://ausubellab.mgh.harvard.edu>) and consists basically of a sensitized PCR screen in which three oligonucleotides are needed to differentiate one SNP with certainty. Oligonucleotides are designed to have an exact match at the 3' end with one or the other SNP, and each is paired with a third common oligonucleotide with which it can produce an amplified product (Drenkard *et al.*, 2000). Ideally, each diagnostic oligo will only form a product with one of the two sequence variations of the SNP; however, in our hands the results have been mixed such that often several different oligo sequences must be tried before a good one can be identified, and even then weak-amplified bands sometimes appear from the “wrong” genotype. Due in part to these uncertainties, for generation of the SNP map we have settled more recently on another simple and efficient PCR-based method, CAPS (cleaved amplified polymorphic sequence), or “snip-SNPs.” CAPS is based on the notion that some SNPs will make or destroy restriction enzyme-recognition sites, and this will facilitate their detection. With CAPS, a region around a (typically single nucleotide) polymorphism is amplified by PCR using two primers designed to invariant regions of sequence. The products are subjected to digestion with a restriction enzyme that acts on one version of the polymorphism but not on the other (Konieczny *et al.*, 1991), thus creating an RFLP (Fig. 2.13).

The real power of the method lies in the ability to analyze groups of progeny for a single SNP in a single step. A single DNA preparation for each individual offspring can be used for screening hundreds of these “snip-SNPs” in parallel. As seen in Fig. 2.13, after a cross between a mutant isolated in an Oak Ridge background and the FGSC 2225 Mauriceville mapping strain, progenies are separated by phenotype and DNA pools made of mutant versus wild types. Aliquots of the pool are used for CAPS mapping with different markers; the PCR products are digested with the diagnostic enzyme and visualized on a gel. A measure of linkage to a given SNP can be estimated as the ratio of cut to uncut DNA for that marker, and unlinked markers ought to be equally represented in each DNA preparation. Going from gross to fine resolution within a chromosomal region is then just a matter of the numbers of SNPs used and their location, just as in classical transmission genetics where enhanced resolution is achieved through analysis of individual allele (here SNP) segregation in more and more individual segregants. In principal, once the SNP map is populated and using only common tools such as PCR and restriction digestion, it should be quite feasible to begin with isolation of DNA from bulked cultures of similar phenotype arising from 200 ascospores and then proceed to 1 map unit resolution (about 20 kb, within a cosmid) within a few days.

C. Validation of the SNP map

Of course, this approach is contingent on the availability of a sufficient number of CAPS markers well spaced over the entire genome. As seen in Fig. 2.12A, the sequence information from the 5695 plasmids yielded about 5500 sequences and nearly 40,000

putative mismatches (SNPs), but most of these were the result of sequencing ambiguities and were screened out immediately. This left about 9500, about half of which were duplicates, leaving 4338 unique SNPs. Of these, for the 12 most commonly used four-cutters that we have examined, 669 created an extra site in Mauriceville, 707 an extra site in Oak Ridge, and 45 extra sites in both.

These are all putative CAPS markers, but two additional criteria must be met for them to be useful. First, the DNA fragments resulting from digestion must be distinguishable from each other and from nonunique bands arising in the same digestion. For instance, creation of a new site in a sea of closely placed identical sites will not yield a useful SNP. Second, the markers must be well spaced through the genome. Two markers lying within a few kilobases of each other will be identifiable at the molecular level but redundant genetically.

A script was developed to select potential CAPS markers by these criteria, and both served to eliminate a number of useless markers. Figure 2.14 shows evidence for clustering of SNPs on contig 1. SNP positions tend to cluster closely together, with large gaps in between clusters (due to the fact that the length of sequenced ESTs is very small compared with their low frequency); hence a naïve-clustering method is sufficient to select among these: simply seed the first cluster in a given contig with the first SNP, and then for each consecutive SNP add it to the current cluster if it is closer to the previous SNP than a cutoff value “minsep,” or create a new cluster to accommodate it if it is not. This gives very similar results for different values of minsep within a reasonable range (Table 2.2). Next, within the set of usable clusters, potential CAPS markers were identified based on whether an enzyme could be found that yielded a unique diagnostic fragment at least 50 bases different in size from any nonunique bands. Applying both criteria, the current set of EST sequences led to the selection and validation of close to 300 confirmed CAPS markers, spanning about 70% of the predefined SNP clusters (at a 10K resolution).

D. Implementation of the SNP map

As SNPs markers are validated they are placed into a database that will be maintained both at Dartmouth and integrated with the online physical maps at the Broad Institute. In this visual context, they can be associated with the maps of each linkage group as shown in Fig. 2.12B.

As the SNP map develops more, the goal is to have markers at least every 2 centimorgans (cM) along each chromosome; current resolution is about 7 cM on average. Searchable databases will list the identity of each CAP as well as its sequence context, the sequence of suitable primers for its amplification, and the restriction enzyme used for its detection.

VI. CONCLUSIONS

We have described here the ongoing progress on the functional genomic analysis of the *Neurospora* genome. As outlined in Section I there are a large number of genes that are unique to the filamentous fungi, and many of these organisms are of great importance as pathogens or for industrial manufacture but are experimentally less tractable than *Neurospora*. Abundant evidence from the comparisons of fungal genomes shows that the repertoire of genes in fungi is in part distinct from other organisms but is held in common among the filamentous fungi. There is thus every reason to believe that the functions that can be ascribed to genes in *Neurospora* will inform the study of a wide variety of related organisms. And this is, after all, the function of a model system.

Acknowledgments

This work was supported by grant P01 GM068087 from the National Institute of General Medical Sciences. We would like to thank John Jones for software design and LIMS implementation, and the following students who participated in the Neurospora Genetics and Genomics Summer Research Institute (NGGSRI) at UCLA for phenotypic analysis: Cynthia Aguirre, Eliana Alcantar, Andrea Cahuantzi, Natalie Cornejo, Zachary W. Cue, Evelyn De Los Santos, Anthony Dualo, Thomas J. Duneheew, Mina El-Mastry, Jonathan Finley, Lizette C. Flores, Christopher Fonseca, Rukhsana A. Khan, Carolyn Kingsley, Juan Lupercio, Criseyda Martinez, Rosaura Ochoa, Olufisayo Oke, Cam M. Phu, Chloe Rivera, Michael Smith, Desree L. Tesada, Tuan D. Tran, and Jackelyn Valladares. The NGGSRI program was supported by NIH/NIGMS 5 R25 GM050067 and NIH/NIGMS 5 R25 GM055052.

References

- Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature*. 2000; 407(6803):513–516. [PubMed: 11029002]
- Aronson BD, Johnson K, Loros JJ, Dunlap JC. Negative feedback defining a circadian clock: Autoregulation in the clock gene *frequency*. *Science*. 1994a; 263:1578–1584. [PubMed: 8128244]
- Aronson BD, Lindgren KM, Dunlap JC, Loros JJ. An efficient method of gene disruption in *Neurospora crassa* with potential for other filamentous fungi. *Mol. Gen. Genet.* 1994b; 242:490–494. [PubMed: 8121407]
- Bailey, TL.; Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*; Menlo Park, CA: AAAI Press; 1994. p. 28-36.
- Beadle GW, Tatum EL. *Neurospora* II. Methods of producing and detecting mutations concerned with nutritional requirements. *Am. J. Bot.* 1945; 32:678–686.
- Bell-Pedersen D, Shinohara M, Loros J, Dunlap JC. Circadian clock-controlled genes isolated from *Neurospora crassa* are late night to early morning specific. *Proc. Nat. Acad. Sci. USA*. 1996; 93:13096–13101. [PubMed: 8917550]
- Berger J, Suzuki T, Senti KA, Stubbs J, Schaffner G, Dickson BJ. Genetic mapping with SNP markers in *Drosophila*. *Nat. Genet.* 2001; 29(4):475–481. [PubMed: 11726933]
- Bistis GN, Perkins DD, Read ND. Different cell types in *Neurospora crassa*. *Fungal Genet. Newslett.* 2003; 50:17–19.
- Bonnen A, Brambl R. Germination physiology of *Neurospora crassa* conidia. *Exp. Mycol.* 1983; 7:197–207.
- Borkovich K, Alex L, Yarden O, Freitag M, Turner G, Read N, Seiler S, Bell-Pedersen D, Paietta J, Plesofsky N, Plamann M, Schulte U, et al. Lessons from the genome sequence of *Neurospora crassa*: Tracing the path from genomic blueprint to multicellular organism. *Microbiol. Mol. Biol. Rev.* 2004; 68:1–108. [PubMed: 15007097]
- Bozdech Z, Zhu J, Joachimiak MP, Cohen FE, Pulliam B, DeRisi JL. Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray. *Genome Biol.* 2003; 4(2):R9. [PubMed: 12620119]
- Brazma A. On the importance of standardisation in life sciences. *Bioinformatics*. 2001; 17(2):113–114. [PubMed: 11238066]
- Catlett NL, Lee B, Yoder OC, Turgeon BG. Split-marker recombination for efficient targeted deletion of fungal genes. *Fungal Genet. Newslett.* 2003; 50:9–11.
- Cogoni C, Macino G. Isolation of quelling-defective (qde) mutants impaired in posttranscriptional transgene-induced gene silencing in *Neurospora crassa*. *Proc. Natl. Acad. Sci. USA*. 1997; 94:10233–10238. [PubMed: 9294193]
- Colot H, Park G, Jones J, Turner G, Borkovich K, Dunlap JC. High throughput knockout of transcription factors in *Neurospora* reveals diverse phenotypes. *Proc. Natl. Acad. Sci. USA*. 2006; 103:10352–10357. [PubMed: 16801547]
- Colot HV, Loros JJ, Dunlap JC. Temperature-modulated alternative splicing and promoter use in the Circadian clock gene *frequency*. *Mol. Biol. Cell*. 2005; 16(12):5563–5571. [PubMed: 16195340]

- Coulter KR, Marzluf GA. Functional analysis of different regions of the positive-acting CYS3 regulatory protein of *Neurospora crassa*. *Curr. Genet.* 1998; 33(6):395–405. [PubMed: 9644202]
- Davis, RH. *Neurospora: Contributions of a Model Organism*. Oxford University Press; Oxford, UK: 2000.
- Davis RL, deSerres D. Genetic and microbial research techniques for *Neurospora crassa*. *Methods Enzymol.* 1970; 27A:79–143.
- Denfert C. Fungal spore germination-insights from the molecular genetics of *Aspergillus nidulans* and *Neurospora crassa*. *Fungal Genet. Biol.* 1997; 21(2):163–172.
- Derisi JL, Iyer VR. Genomics and array technology. *Curr. Opin Oncol.* 1999; 11(1):76–79. [PubMed: 9914882]
- Dillon D, Stadler D. Spontaneous mutation at the mtr locus in *Neurospora*: The molecular spectrum in wild-type and a mutator strain. *Genetics.* 1994; 138(1):61–74. [PubMed: 8001794]
- Drenkard E, Richter BG, Rozen S, Stutius LM, Angell NA, Mindrinos M, Cho RJ, Oefner PJ, Davis RW, Ausubel FM. A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in Arabidopsis. *Plant Physiol.* 2000; 124(4):1483–1492. [PubMed: 11115864]
- Dunlap JC, Loros JJ. Analysis of circadian rhythms in *Neurospora*: Overview of assays and genetic and molecular biological manipulation. *Methods Enzymol.* 2005; 393:3–22. [PubMed: 15817284]
- Eisen MB, Brown PO. DNA arrays for analysis of gene expression. *Methods Enzymol.* 1999; 303:179–205. [PubMed: 10349646]
- Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA.* 1998; 95:14863–14868. [PubMed: 9843981]
- Fairhead C, Llorente B, Denis F, Soler M, Dujon B. New vectors for combinatorial deletions in yeast chromosomes and for gap-repair cloning using “split-marker” recombination. *Yeast.* 1996; 12:1439–1457. [PubMed: 8948099]
- Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, Rehman B, Elkins T, et al. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature.* 2003; 422:859–868. [PubMed: 12712197]
- Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB. Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PLoS Biol.* 2004; 2(12):e398. [PubMed: 15534694]
- Goldstein AL, McCusker JH. Three new dominant drug resistance cassettes for gene disruption in *Saccharomyces cerevisiae*. *Yeast.* 1999; 15:1541–1543. [PubMed: 10514571]
- Gritz L, Davies J. Plasmid-encoded hygromycin B resistance: The sequence of the hygromycin B phosphotransferase gene and its expression in *Escherichia coli* and *Saccharomyces cerevisiae*. *Gene.* 1983; 25:179–188. [PubMed: 6319235]
- Kasuga T, Townsend JP, Tian C, Gilbert LB, Mannhaupt G, Taylor JW, Glass NL. Long-oligomer microarray profiling in *Neurospora crassa* reveals the transcriptional program underlying biochemical and physiological events of conidial germination. *Nucleic Acids Res.* 2005; 33(20): 6469–6485. [PubMed: 16287898]
- Kerr MK, Churchill GA. Statistical design and the analysis of gene expression microarray data. *Genet. Res.* 2001; 77(2):123–128. [PubMed: 11355567]
- Konieczny A, Voytas DF, Cummings MP, Ausubel FM. A superfamily of *Arabidopsis thaliana* retrotransposons. *Genetics.* 1991; 127(4):801–809. [PubMed: 1709409]
- Kramer C, Loros JJ, Dunlap JC, Crosthwaite SK. Role for antisense RNA in regulating circadian clock function in *Neurospora crassa*. *Nature.* 2003; 421:948–952. [PubMed: 12607002]
- Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat. Biotechnol.* 2002; 20(8):835–839. [PubMed: 12101404]
- Martin AC, Drubin DG. Impact of genome-wide functional analyses on cell biology research. *Curr. Opin. Cell Biol.* 2003; 15:6–13. [PubMed: 12517698]
- McClung CR, Fox BA, Dunlap JC. The *Neurospora* clock gene *frequency* shares a sequence element with the *Drosophila* clock gene *period*. *Nature.* 1989; 339:558–562. [PubMed: 2525233]

- McGuire AM, Hughes JD, Church GM. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* 2000; 10(6):744–757. [PubMed: 10854408]
- Metzenberg RL, Grotelueschen J. Restriction polymorphism maps of *Neurospora crassa*: Update. *Fungal Genet. Newslett.* 1995; 42:82–90.
- Metzenberg RL, Stevens JN, Selker EU, Morzycka-Wroblewska E. A restriction fragment length polymorphism map for *Neurospora crassa*. *Proc. Natl. Acad. Sci. USA.* 1985; 82:2067–2071. [PubMed: 3157192]
- Muller HM, Kenny EE, Sternberg PW. Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol.* 2004; 2:e309. [PubMed: 15383839]
- Nelson MA, Kang S, Braun EL, Crawford ME, Dolan PL, Leonard PM, Mitchell J, Armijo AM, Bean L, Blueyes E, Cushing T, Errett A, et al. Expressed sequences from conidial, mycelial, and sexual stages of *Neurospora*. *Fungal Genet. Biol.* 1997; 21:348–363. [PubMed: 9290248]
- Ninomiya Y, Suzuki K, Ishii C, Inoue H. Highly efficient gene replacements in *Neurospora* strains deficient for nonhomologous end-joining. *Proc. Natl. Acad. Sci. USA.* 2004; 101(33):12248–12253. [PubMed: 15299145]
- Oldenburg KR, Vo KT, Michaelis S, Paddon C. Recombination-mediated PCR-directed plasmid construction *in vivo* in yeast. *Nucleic Acids Res.* 1997; 25:451–452. [PubMed: 9016579]
- Ooi SL, Shoemaker DD, Boeke JD. A DNA microarray-based genetic screen for nonhomologous end-joining mutants in *Saccharomyces cerevisiae*. *Science.* 2001; 294:2552–2556. [PubMed: 11701889]
- Ooi SL, Pan X, Peyser BD, Ye P, Meluh PB, Yuan DS, Irizarry RA, Bader JS, Spencer FA, Boeke JD. Global synthetic-lethality analysis and yeast functional profiling. *Trends Genet.* 2006; 22(1):56–63. [PubMed: 16309778]
- Pandit NN, Russo VE. Reversible inactivation of a foreign gene, *hph*, during the asexual cycle in *Neurospora crassa* transformants. *Mol. Gen. Genet.* 1992; 234:412–422. [PubMed: 1383683]
- Perkins, DD.; Radford, A.; Sachs, MS. The *Neurospora* Compendium. Academic Press; San Diego: 2001.
- Raymond CK, Pownder TA, Sexson SL. General method for plasmid construction using homologous recombination. *Biotechniques.* 1999; 26:134–141. [PubMed: 9894602]
- Riquelme M, Reynaga-Peña CG, Gierz G, Bartnicki-García S. What determines growth direction in fungal hyphae? *Fungal Genet. Biol.* 1998; 24(1–2):101–109. [PubMed: 9742196]
- Roca MG, Arlt J, Jeffree CE, Read ND. Cell biology of conidial anastomosis tubes in *Neurospora crassa*. *Eukaryot. Cell.* 2005; 4:911–919. [PubMed: 15879525]
- Ruepp A, Zollner A, Maier D, Albermann K, Hani J, Mokrejs M, Tetko I, Guldener U, Mannhaupt G, Munsterkotter M, Mewes HW. The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.* 2004; 32(18):5539–5545. [PubMed: 15486203]
- Sachs MS, Yanofsky C. Development expression of genes involved in conidiation and amino acid biosynthesis in *Neurospora crassa*. *Dev. Biol.* 1991; 148(1):117–128. [PubMed: 1834495]
- Schmidt JC, Brody S. Biochemical genetics of *Neurospora crassa* conidial germination. *Bacteriol. Rev.* 1976; 40:1–41. [PubMed: 5072]
- Selker EU. Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annu. Rev. Genet.* 1990; 24:579–613. [PubMed: 2150906]
- Selker EU. Epigenetic phenomena in filamentous fungi. *Trends Genet.* 1997; 13:296–301. [PubMed: 9260514]
- Tey WK, North AJ, Reyes JL, Lu YF, Jedd G. Polarized gene expression determines Woronin body formation at the leading edge of the fungal colony. *Mol. Biol. Cell.* 2005; 16(6):2651–2659. [PubMed: 15800068]
- Townsend JP. Multifactorial experimental design and the transitivity of ratios with spotted DNA microarrays. *BMC Genomics.* 2003; 4(1):41. [PubMed: 14525623]
- Townsend JP, Hartl DL. Bayesian analysis of gene expression levels: Statistical quantification of relative mRNA level across multiple strains or treatments. *Genome Biol.* 2002; 3 RESEARCH0071.

- Vinciotti V, Khanin R, D'Alimonte D, Liu X, Cattini N, Hotchkiss G, Bucca G, de Jesus O, Rasaiyaah J, Smith CP, Kellam P, Wit E. An experimental evaluation of a loop versus a reference design for two-channel microarrays. *Bioinformatics*. 2005; 21(4):492–501. [PubMed: 15374872]
- Wicks S, Yeh R, Gish W, Waterson R, Plasterk H. Rapid gene mapping in *C. elegans* using a high density polymorphism map. *Nat. Genet.* 2001; 28:160–164. [PubMed: 11381264]
- Winzeler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, et al. Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*. 1999; 285(5429):901–906. [PubMed: 10436161]
- Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. *J. Comput. Biol.* 2001; 8(6):625–637. [PubMed: 11747616]
- Yang YH, Speed T. Design issues for cDNA microarray experiments. *Nat. Rev. Genet.* 2002; 3:579–588. [PubMed: 12154381]
- Young JL, Marzluf GA. Molecular comparison of the negative-acting nitrogen control gene, *nmr*, in *Neurospora crassa* and other *Neurospora* and fungal species. *Biochem. Genet.* 1991; 29(9–10): 447–459. [PubMed: 1663340]
- Zeitlinger J, Simon I, Harbison CT, Hannett NM, Volkert TL, Fink GR, Young RA. Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell*. 2003; 113(3):395–404. [PubMed: 12732146]
- Zhu H, Nowrousian M, Kupfer D, Colot H, Berrocal-Tito G, Lai H, Bell-Pedersen D, Roe B, Loros JJ, Dunlap JC. Analysis of expressed sequence tags from two starvation, time-of-day-specific libraries of *Neurospora crassa* reveals novel clock-controlled genes. *Genetics*. 2001; 157:1057–1065. [PubMed: 11238394]

Figure 2.1.

The life cycle of *N. crassa* showing some of the many life stages and cell types seen in the sexual and asexual cycles. Adapted from {www.fgsc.net/Neurospora/sectionB2.htm}.

Figure 2.2.

Goals within and interactions among the four major research groups in the Neurospora functional genomics project.

Figure 2.3.

Strategy for creating deletion cassettes. 3' and 5' flank fragments are PCR'd separately from genomic DNA using primers 3f + 3r and 5f + 5r; primers 3f and 5r incorporate MmeI sites have 5' tails matching the *hph* cassette whereas 3r and 5f match the vector. The two flanks, *hph* cassette and gapped vector are cotransformed into yeast where homologous recombination recreates the circular construct. The final linear deletion cassette is PCR'd from the pooled yeast DNA using primers 3r and 5f. The cassette is constructed so that *hph* is transcribed antisense relative to the target gene.

Figure 2.4.

Detailed view of a primer (3f or 5r) used for assembly of the knockout cassette and one end of the *hph* cassette used for gene deletions, showing the location of the bar code. The circle represents the restriction enzyme Mme1 binding to its recognition sequence, and the attached curved arrow points to the position where it will cut genomic DNA containing a disrupted gene. After cutting with Mme1, the “gene-specific 20-mer” comprises a molecular bar code that may be detected through ligation of a linker (shown as a gray bar on the right) followed by PCR using primers (horizontal thick black arrows) specific for the linker and the *hph* cassette.

Neurospora crassa: Pheno Form 96/05/2006 05:34 PM

BROAD INSTITUTE Home > Data > Genomes > Pheno Form Login

[Home](#) | [News](#) | [Info](#) | [Community](#) | [Download](#) | [BLAST](#) | [Find Features](#) | [Regions](#) | [Gene Index](#) | [Alleles](#) | [Maps](#) | [FAQ](#) | [Contact](#) | [Links](#)

Pheno Form

User: Locus:

	Plates			
	25/VM	25/VMCone	37/VM	37/VMCone
Plate Photo 2dhr	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
Plate Photo 4dhr	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
Edge Photo	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
Growth Pattern	☐ abnormal ☐ normal	☐ abnormal ☐ normal	☐ abnormal ☐ normal	☐ abnormal ☐ normal
Plate Pigmentation	☐ normal ☐ abnormal ☐ intensity diminished ☐ white ☐ other ☐ intensity enhanced ☐ pink ☐ yellow ☐ peach	☐ normal ☐ abnormal ☐ intensity diminished ☐ white ☐ other ☐ intensity enhanced ☐ pink ☐ yellow ☐ peach	☐ normal ☐ abnormal ☐ intensity diminished ☐ white ☐ other ☐ intensity enhanced ☐ pink ☐ yellow ☐ peach	☐ normal ☐ abnormal ☐ intensity diminished ☐ white ☐ other ☐ intensity enhanced ☐ pink ☐ yellow ☐ peach

Slants		Sexual Development	
Slant Pigmentation	☐ abnormal ☐ yellow ☐ intensity enhanced ☐ peach ☐ white ☐ intensity diminished ☐ other ☐ pink ☐ normal	Protophythecium Formation	☐ abnormal ☐ reduced number ☐ reduced size ☐ reduced number & size ☐ not formed ☐ normal
Conidiation	☐ abnormal ☐ reduced formation ☐ not formed ☐ abundant ☐ normal	Perithecium Formation	☐ normal ☐ abnormal ☐ reduced size ☐ not formed ☐ ostiole formation abnormal
Aerial Hyphae Formation	☐ normal ☐ abnormal ☐ reduced formation ☐ not formed ☐ short	Asclepias Formation	☐ normal ☐ abnormal ☐ reduced number ☐ not formed

http://mode015.broad.mit.edu:50846/context/neurospora/PhenoForm.html Page 1 of 2

Figure 2.5.
Screen shot of page 1 of the phenotype data entry form used by undergraduates for web-paged entry of basic phenotypic data.

Figure 2.6.
How EST sequences are used for (A) gene calling and (B) UTR and alternative splice prediction.

Figure 2.7.

Community Annotation Project (CAP) process and Gene Detail page showing community annotation. CAP is an infrastructure developed by the Broad Institute in collaboration with members of the Neurospora Functional Genomics Project that uses the automated genome annotation as a scaffold on which the research community can build a refined manual annotation. CAP provides the research community the ability to annotate the structure of a gene as well as capture a wealth of functional information about a gene through a user-friendly web interface. Community annotations are attached to the official genome annotation, are immediately searchable, and are incorporated into the official annotation through a manual curation process (see <http://www.broad.mit.edu/annotation/genome/neurospora/CASearchAnnotations.html/direct>).

Allele KO1 (knockout) on Locus NCJ08726
Phenotype Summary View

Phenotype Record	Observations			Images
	Physiology	Morphology	Sexual Development	
SCM with 0.1% sucrose:25C UCLA NGGSR1 2005 29-Sep-05	viaible	-	no protoperithecia formed no perithecia formed no ascospores formed	Plate: 24 Hr Plate: 48 Hr edgePhoto
minimal:25C UCLA NGGSR1 2005 29-Sep-05	viaible linear growth:20-25 mm/day aerial hyphae extension:05-10 mm/day	abnormal macroconidia pigmentation (Slants) no conidia formed (Plates) diminished pigmentation (Slants) short aerial hyphae (Slants) no conidia formed (Slants)	-	
minimal:37C UCLA NGGSR1 2005 29-Sep-05	viaible	no conidia formed (Plots)		Plate: 24 Hr Plate: 48 Hr edgePhoto: 24 Hr
supplemented:25C UCLA NGGSR1 2005 29-Sep-05	viaible aerial hyphae extension:00-05 mm/day	-	-	Plate: 24 Hr Plate: 24 Hr Plate: 48 Hr
supplemented:37C UCLA NGGSR1 2005 29-Sep-05	viaible	no conidia formed (Plates) normal pigmentation	-	Plate: 24 Hr Plate: 48 Hr edgePhoto: 24 Hr

Home | Internal Web | Contact Us | Site Map | Search
© 2006 Broad Institute

Figure 2.8.
A typical “alleles phenotype characterization” summary page for a Neurospora gene, taken from <http://www.broad.mit.edu/annotation/genome/neurospora/AlleleDetails.html>.

Figure 2.9.

Schematic representation of the data comparisons executed in a tightly fashioned closed-loop design for microarray data collection. Images (not to scale) represent the shape of cells from the time conidia are inoculated into growth medium (time 0) through the first 16 h of growth. RNA samples collected at each time point are hybridized to multiple microarrays such that each pair of arrows represents one experiment of reference and control (Kasuga *et al.*, 2005).

Figure 2.10.

Changes in expression of different categories of genes over the course of the first 16-h growth in *Neurospora* as elucidated by microarray analysis.

Figure 2.11.
An example of detailed changes in gene expression associated with early development in *Neurospora* as revealed by microarray analysis.

Figure 2.12.

Flow chart for the preliminary SNP generation and mapping project. On the left is shown the progressive stages of data generation and screening, and on the right an example of the final product. This shows a fragment of chromosome VI L including parts of two contigs (green and black). Shown are the optical, physical, and genetic maps. Green dashes EST band indicate regions with EST coverage (at 3-kb resolution), black dashes are unconfirmed SNPs, and validated CAPS are identified by a three-letter enzyme designator and a number in red (see <http://www.broad.mit.edu/annotation/genome/neurospora/maps/ViewMap.html?sp=5>).

Figure 2.13.

A schematic of the data produced from a bulked progeny SNP-mapping experiment.

Figure 2.14. SNP Clustering. (A) Position of SNPs within the 1,800,000 bp of contig 1 in assembly 7 of the *Neurospora* genome. Note the long horizontal stretches that correspond to clusters. (B) Number of SNPs in each region of 50 kb along contig 1.

Table 2.1SNP Frequency in *Neurospora* Open Reading Frames^a

Gene	Number of polymorphisms in Mauriceville with respect to Oak Ridge	Nature of polymorphisms (all are bp substitutions)	References
<i>cys-3</i>	15 in 708-bp ORF	9 silent; 6 substitutions in nonessential regions	Coulter and Marzluf, 1998
<i>mtr</i>	3 in 1472-bp ORF	3 silent	Dillon and Stadler, 1994
<i>nmr</i>	6 in 1464-bp ORF	5 silent; 1 substitution truncates ORF by 3 codons	Young and Marzluf, 1991

^aThe percentage of SNPs in these three ORFs is 0.7%.

Table 2.2

Clustering Results for Different Values of Minsep

Minsep	5K	10K	20K	50K	100K
Cluster width (average $\pm$ standard deviation)	0.6 $\pm$ 1.2	1.1 $\pm$ 2.5	3.2 $\pm$ 7.7	16 $\pm$ 28	43 $\pm$ 73
Maximal cluster width	10	16	64	188	360
Number of clusters	452	424	373	263	163
Number of nonsingleton clusters	309	289	252	179	89
Cluster gap (average $\pm$ standard deviation)	66 $\pm$ 60	71 $\pm$ 60	80 $\pm$ 59	110 $\pm$ 58	160 $\pm$ 55

Cluster width is defined as the difference between highest and lowest position within a cluster, and cluster gap as that between the highest in a cluster and the lowest in the next (both in kilobases).