# Step I: White Paper Application

## Application Guidelines

1. *The application should be submitted electronically per requirements via the web site of any of the NIAID Genomic Sequencing Centers for Infectious Diseases. Include all attachments, if any, to the application.*
2. *There are no submission deadlines; white papers can be submitted at anytime.*
3. *GSC personnel at any of the three Centers can assist / guide you in preparing the white paper.*
4. *Investigators can expect to receive a response within 4-6 weeks after submission.*
5. *Upon approval of the white paper, the NIAID Project Officer will assign the project to a NIAID GSC to develop a management plan in conjunction with the participating scientists.*

# White Paper Application

**Project Title:** Using next-generation sequencing to define strain variation and virulence in clinical isolates of *Candida albicans*
OR Identification of Virulence Factors by Genomic Analysis of *Candida albicans* Strains

**Authors:**

**Primary Investigator Contact:**

| | |
|---|---|
| Name | Christina Cuomo |
| Position | Group Leader, Fungal Genome Sequencing & Analysis |
| Institution | Broad Institute |
| Address | 301 Binney St, Cambridge |
| State | MA |
| ZIP Code | 02142 |
| Telephone | (617) 714-7904 |
| Fax | (617) 714-8931 |
| E-Mail | cuomo@broadinstitute.org |

| | |
|---|---|
| Name | Richard Bennett |
| Position | Assistant Professor |
| Institution | Brown University |
| Address | 171 Meeting St |
| State | RI |
| ZIP Code | 02912 |
| Telephone | (401) 863-6341 |
| Fax | (401) 863-2925 |
| E-Mail | Richard_Bennett@brown.edu |


## 1. Executive Summary *(Please limit to 500 words.)*

*Candida albicans* is normally a harmless component of the human microbiota, commonly found inhabiting the skin, oral mucosa, and gastrointestinal tract. However, it is also a highly successful pathogen and is responsible for both debilitating mucosal infections and disseminated systemic infections, with the latter resulting in death in up to 50% of cases. *C. albicans* also displays a remarkable ability to adapt to changing conditions, and is able to colonize and infect almost any organ in the human body. Despite its medical importance, the characterization of *C. albicans* virulence factors has lagged behind other pathogenic microbes, in part due to the lack of a complete sexual cycle and the inability to carry out conventional genetic studies.

Greater understanding of mechanisms of fungal colonization and infection are required if we are to develop new lines of antifungal drugs and combat the growing problem of drug resistant strains. To this end, we propose to carry out genomic analysis of 30 clinical isolates of *C. albicans* and transcriptional profiling of 6 of these isolates. These isolates have been carefully chosen to represent a wide cross-section of *C. albicans* biology. First, the majority of these natural strains have been chosen as they exhibit significant differences in virulence in a mammalian model of systemic infection. Second, these strains are different cell mating types (**a**/**a**, α/α, and **a**/α strains), and differences at the

mating locus have been shown to result in altered virulence. Third, a set of strains have been chosen as they are mucosal isolates which display lower virulence in mice; other strains of *C. albicans*, such as the reference SC5314, are not as efficient colonizers of mucosa. Taken together, we believe that extensive profiling of this set of *C. albicans* strains will provide a detailed picture of the strain variations that are the basis for differences in systemic and mucosal infections, as well as differences in drug susceptibility.

The proposed studies will build on existing expertise at the Broad Institute, which has been a pioneer in fungal genomics through the Fungal Genome Initiative (FGI). As part of the FGI, the Broad has sequenced one isolate (strain WO-1) of *Candida albicans*, as well as isolates of four closely related *Candida* species [1]. Completion of the work outlined in this proposal would build on the expertise of the FGI, and would provide a comprehensive genomic analysis of the most commonly isolated human fungal pathogen, *C. albicans*. In particular, it is expected that novel genes regulating fungal virulence will be identified, including those affecting both systemic and mucosal infections. The detailed analysis of multiple independent isolates of *C. albicans* will also prove to be an invaluable resource to other researchers in the field, and will provide the foundation for identification of new drug targets for treating mucosal and disseminated infections.

## 2. Justification

Relevance to infectious disease:

*Candida albicans* is the most commonly isolated human fungal pathogen. In healthy individuals infections are often limited to mucosal layers, while in the immunocompromised host infections can target almost any organ in the human body. Systemic infections are associated with high rates of mortality and morbidity.

Existing Data:

The FGI at the Broad Institute has previously sequenced several clinically important *Candida* species, including one isolate of *C. albicans* (strain WO-1, an $\alpha/\alpha$ strain). In addition, the genome of the laboratory strain of *C. albicans* (SC5314, an **a**/$\alpha$ strain) was sequenced at Stanford together with the Biotechnology Research Institute of the National Research Council Canada. Three additional related species were sequenced at the Welcome Trust Sanger Institute and at Genoscope. However, there has been no large-scale sequencing or profiling studies carried out on clinical isolates of *C. albicans*. We have therefore identified a representative set of strains that exhibit widely varying abilities to infect the mammalian host.

Utility of Data Created:

The primary goal of the proposed study is to generate genome sequences for multiple clinical isolates of *C. albicans*, and to provide transcriptional profiles for these strains during growth under laboratory conditions as well as in models of infection (systemic and vaginal infection). The study will be the first to yield high-resolution information about natural variation and species diversity amongst *C. albicans* strains, as well as revealing how genomic differences impact transcriptional circuits. There is already evidence that different *C. albicans* isolates exhibit widely varying properties, including differences in colonization and infection, and that these differences correlate with differences at the

genomic level [2, 3]. The proposed study therefore aims to identify those traits that are most closely associated with host invasion and pathogenesis. In fact, it was recently suggested that sequencing of multiple *C. albicans* isolates from different clades is the best approach to uncover polymorphisms associated with disease in pathogenic strains [3].

In addition, we note that comparative analysis of *C. albicans* genomes with isolates of other sequenced *Candida* species will further reveal how increased pathogenicity evolved in the *C. albicans* lineage in comparison with related *Candida* species.

## 3. Rationale for Strain Selection

In this study, we propose to generate diploid genome assemblies and transcriptional profiles for 30 clinical isolates of *Candida albicans*. The majority of these strains (20 isolates) have been selected due to their differing abilities to cause disease in a mouse tail-vein model of systemic infection [4]. Using the same inoculum of infection ($10^6$ cells), the most virulent isolates resulted in death of all of the infected animals within the first week following injection, while the least virulent isolates killed only 20% of infected animals within 22 days of injection (see Figure 1). In total, we have divided the 20 strains into 3 categories representing highly virulent strains (5 strains), moderately virulent strains (6 strains) and strains with low virulence (9 strains). We note that each of these strains has undergone only minimal passaging after original isolation [4].

In addition, the set of 20 strains have been chosen as being representative of the major clades of *C. albicans*. The Soll lab has designated five major groups of *C. albicans* strains: I, II, III, SA (South Africa) and E (European) [5]. The prospective set of strains includes members from each clade: clade I (8 strains), clade II (2 strains), clade III (2 strains), clade SA (5 strains) and clade E (1 strain). Most strains are from clade I as this is the predominant clade of *C. albicans* associated with both commensalism and infection [2]. The strains were originally isolated from different niches in the human body. For example, the majority of the strains (13 isolates) were from bloodstream infections, 5 strains were oral isolates, and 2 strains were vaginal isolates.

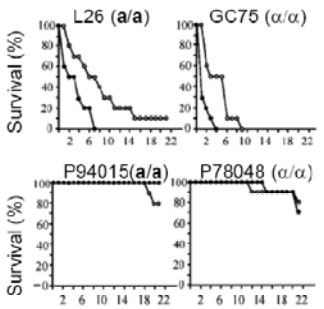



**Figure 1. Differences in virulence in natural isolates of *C. albicans*.** Strains were injected into the tail vein of female ND4 mice using $10^6$ *C. albicans* cells. The top 2 panels indicate highly virulent strains, while the bottom 2 panels indicate strains of low virulence. Adapted from Wu *et al.* (2007).

The set of strains also include those with different mating types as they include **a**/α (10 strains), **a**/**a** (6 strains), and α/α (4 strains) genotypes. This is significant as the *MTL* (mating-type like) locus has been associated with differences in virulence, although the mechanism by which genes at this locus affect virulence is unknown [4, 6]. By including genes of different mating types we therefore expect to establish the role of *MTL* genes during infection of the host. The list of clinical isolates is provided in Supplemental Table 1.

**Table 1. Survival after intravenous inoculation with 2.5 x $10^5$ CFU of *C. albicans* mucosal clinical isolates in a murine model of systemic candidiasis.**

| | Mean Survival (days) | Median Survival (days) | % Survived to Day 31 |
|---|---|---|---|
| SC5314 | 9.5 | 7 | 0 |
| 3153A | 10.2 | 8 | 17 |
| A48 | 31 | 31 | 100 |
| A176 | 31 | 31 | 100 |
| A123 | 31 | 31 | 100 |
| A92 | 31 | 31 | 100 |
| A67 | 31 | 31 | 100 |
| A46 | 29.2 | 31 | 83 |
| A155 | 27.8 | 31 | 83 |
| A20 | 29.7 | 31 | 83 |
| A203 | 27.8 | 31 | 83 |
| A84 | 28.8 | 31 | 83 |

In addition, we will sequence a small number of *C. albicans* strains isolated from AIDS patients with mucosal infections [7]. Clinical isolates that displayed significantly reduced virulence in a mouse model of systemic infection were chosen as sequencing targets (Table 1, Supplemental Table 1). Furthermore, these strains display increased ability to colonize mucosal tissues in a murine model of vaginitis compared with the reference strain SC5314 (Figure 2). It has been previously reported that although SC5314 is virulent systemically, it weakly colonizes mucosal tissues, including the oral and vaginal mucosa, [8] and the GI tract (Figure 3). Therefore, it is likely that these clinical isolates are adapted for mucosal growth and have genomic and/or transcriptional differences with systemic infection isolates. This proposal would analyze these strains by genome and transcriptome analysis to identify potential factors that make these strains successful at colonization and infection of this niche.

**Figure 2. *C. albicans* fungal burden from the vaginal tract of mice inoculated with mucosal clinical isolates.** CBA/J mice were given injected SQ with estrogen 3 d prior to inoculation to induce pseudoestrus. Mice were inoculated by pipetting 5 x $10^5$ CFU of *C. albicans* on day 0 and vaginas were lavaged on day 4 post-inoculation. Lavage samples were homogenized in sterile water and plated onto SDB plates for enumeration of CFU.

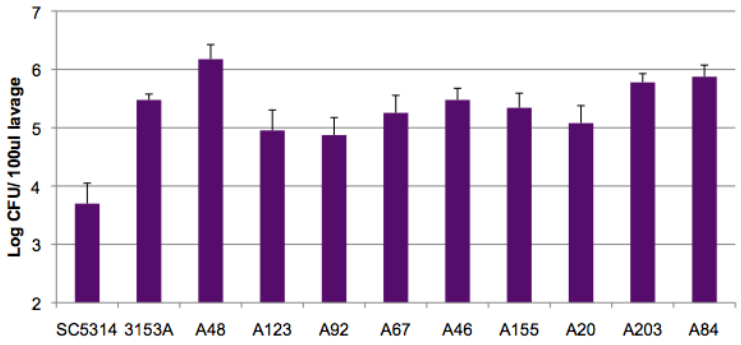
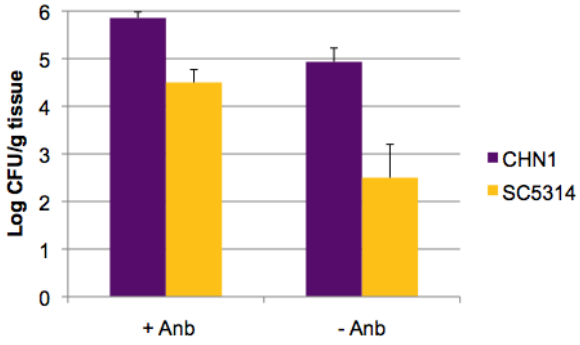



**Figure 3. *C. albicans* fungal burden from the GI tract of mice in the presence and absence of antibiotics (anb).** C57BL/6 mice were given oral cefoperazone (0.5 mg/ml) for 5 days (d -4 through d 0). Mice were inoculated by gavage with $10^7$ CFU of *C. albicans* strain CHN1 or SC5314 on d 0 and GI tissues (small intestine, cecum, and colon) were harvested on day 14. Tissues were homogenized in sterile water and plated onto SDB plates for enumeration of CFU.




Overall, we emphasize the diversity inherent in the set of *C. albicans* strains to be analyzed and predict that this will increase their general applicability to researchers in the community, as discussed below.

## 4a. Approach to Data Production: **Data Generation**

The proposed studies will generate genome sequence and assemblies of each of the *C. albicans* strains. The Broad Institute has previously sequenced and assembled the WO-1 isolate of *C. albicans* using Sanger technology, but most microbial genomes are now being assembled using Illumina data. We will generate Illumina sequence on the HiSeq platform of the submitted isolates as paired 100b reads from two libraries: size selected 180b fragments and 3kb fragments. We aim to generate a total of 45X of reads from 180b fragments and 45X of reads from 3-5kb fragments. This data will allow *de novo* assembly of each strain using ALLPATHS, based on assemblies of microbial and mammalian genomes [9]. The assemblies will be used to identify regions unique to each strain, and to characterize rearrangement events between strains, which occur most frequently at a major repeat sequence found in one or two copies on most chromosomes. Copy number variation will be detected by examining the assembly coverage across each chromosome, and identifying regions with possible ploidy changes. To identify single nucleotide polymorphisms, sequence reads will be aligned to the SC5314 reference assembly with the BWA aligner [10] in the Broad Picard pipeline. SNPs will be called using SAMtools [11], using parameters shown to perform with high sensitivity and specificity in other diploid fungal genomes, filtering for coverage, quality, and allele balance for heterozygous calls. In previous work, we have shown that SNPs identified in *C. albicans* from Illumina data with these parameters highly reproduce those identified from Sanger data, and that requiring high quality of the assembly SNP base improves the specificity of the calling. Genomic DNA will be generated for each of the *C. albicans* strains (provided by the Bennett and Noverr laboratories) and supplied to the Broad.

Transcriptional profiling of a subset of the *C. albicans* strains will also be performed, initially selecting 6 phenotypically and phylogenetically diverse strains. These will include two strains each from high, mid, and low virulence categories (Supplementary Table 1). We have chosen 6 experimental conditions for which total RNA will be generated for each strain. These conditions are representative of the wide variety of conditions in which *C. albicans* can grow. In particular, they will compare growth in yeast and hyphal (filamentous) forms, both of which are necessary for infection of the mammalian host.

1. *C. albicans* growth in YPD medium at 30°C. YPD (yeast extract, dextrose) is a standard medium for growing *C. albicans* strains during *in vitro* culture, and will be used as a baseline for transcriptional comparisons.
2. Growth in YPD containing 1% serum at 37°C. These growth conditions are used to induce formation of the filamentous form of *C. albicans*, which is also the invasive form of the organism in the host. We note that while 10% serum is commonly used for induction of filamentation, 1% serum provides a more sensitive condition for distinguishing filamentation phenotypes between strains [12].
3. Growth in Spider medium at 30°C. This is a starvation media that induces filamentation in *C. albicans*, although the genetic requirements for filamentation are distinct from those during serum-induced filamentation [12].
4. *C. albicans* cells cultured in YPD medium under embedded conditions that are decreased for oxygen tensions. These conditions can also induce filamentation and are thought to mimic those encountered deep inside tissues in the host. Filamentation during hypoxia is regulated by different signaling pathways than those during starvation or serum-induction [13].

5. Expression profiling of *C. albicans* at low pH. *C. albicans* naturally resides in low pH environments in the gastrointestinal tract and expression of pH-regulated genes has been associated with virulence [3].
6. Recovery of *C. albicans* cells from infected kidneys in a mouse model of systemic infection. This protocol has been performed in the Bennett laboratory for 5 years and is approved by IACUC. Systemic infection will be used to analyze genes expressed during infection of host kidneys, which are the primary organs targeted during disseminated disease. Expression profiling of *C. albicans* cells recovered from this organ has been shown to be successful [14, 15].

We note that we are continuing to characterize the clinical isolates, and if additional *in vitro* growth conditions are identified that correlate with differences in virulence *in vivo*, then RNA will also be isolated from these conditions and analyzed.

For analysis of mucosal *C. albicans*, we will profile two mucosal infecting strains and two control strains (CHN1 and 3153A) using an animal model of mucosal infection (vaginitis). This protocol has been performed in the Fidel laboratory for 15 years and the Noverr laboratory for 3 years and is approved by IACUC. For these studies, CBA/J mice will be used. Estrogen is required to maintain a persistent vaginal *C. albicans* infection in mice. Mice are injected subcutaneously with 0.02mg estradiol valerate dissolved in 0.1mL sesame oil. Estrogen injections continue weekly throughout the course of the experiment. For vaginal inoculation, mice are sedated with isoflurane 4% solution. After monitoring the appropriate depth of the anesthesia, $5x10^5$ *C. albicans* blastospores in 20μL PBS are pipetted into the mouse vagina. Negative control animals receive 20 μL of PBS. At d4 and d7 post-inoculation, separate groups of mice will be sacrificed by cervical dislocation and the vaginas will be lavaged with 100μL of PBS. The lavage fluid will be frozen in RNALater (Ambion) at -80°C. Strains will also be grown in YPD to compare to the infection samples to identify genes induced during infection. RNA will be extracted using the RiboPure Yeast RNA kit (Ambion). The Fidel laboratory has experience isolating high quality *C. albicans* RNA from *in vivo* samples.

For RNA sequencing, we will construct strand specific libraries from mRNA samples [16]. We aim to generate a total of 25M 76 base paired Illumina HiSeq reads per sample; this should allow for good coverage of expressed genes. RNA sequence reads will be aligned to the genome reference using the TopHat spliced read aligner [17], or by first assembling and then aligning the data using a novel algorithm recently developed at the Broad called Inchworm (http://sourceforge.net/projects/inchworm/). Gene expression levels will be calculated based on the most recent gene set available in the Candida Genome Database (CGD). As *C. albicans* has already been re-annotated using RNA-sequence data [18], the current CGD gene set is expected to detect most ORFs. An expression level for each gene will be calculated using the DEGseq R package [19]. We will also predict transcripts for RNA sequence clusters which align to the genome but do not overlap predicted genes.

To visualize SNPs, coverage levels, and RNA-sequence based transcripts across the genome, all data will be imported into the Broad developed Integrated Genome Viewer (www.broadinstitute.org/igv). This platform can provide access to visual inspection of the data and alignments by collaborating groups.

We emphasize that we have deliberately chosen a diverse set of clinical isolates for these primary studies in order to generate a global data set for *C. albicans* strains. Subsequent studies by multiple investigators are expected to build on this data set and will focus on strains that are more closely related to be able to identify specific genotype/phenotype associations.

## 4b. Approach to Data Production: **Data Analysis**

The completion of this project will result in annotated genome sequences for a diverse set of clinical isolates of *C. albicans*. As described above, we will identify SNPs, copy number variations, and rearrangement differences between the strains. We will compare these features between clinical isolates to identify subsets that are unique to particular strains or groups of strains. Any changes in copy number of genomic regions will be compared to previously characterized copy number variations. SNPs will be further classified as synonymous or nonsynonymous variants within genes or nongenic. The frequency and pattern of polymorphisms across species from different clades will be determined. In particular, it will be revealing to compare genotypes from strains with high, mid, and low virulence (as defined in the animal model of systemic infection). The data will provide the most complete set of polymorphisms with which to identify those traits that are closely associated with virulence in bloodstream infections. As recently noted by the Schmid lab, genomic comparison of multiple strains is predicted to prove the most informative approach for identifying virulence factors in *C. albicans* [3].

Transcriptional profiles of the clinical strains will also be generated for the 6 culture conditions described above. Expression levels for each gene will be compared between conditions and strains, using principle component or other clustering methods to identify similar conditions and strains, and examining significant differences in gene expression. This will prove to be an extremely rich and useful data set and will further examine the correlation between expression of different gene families and strains exhibiting the highest virulence. In particular, attention will be given to gene sets that have been implicated in affecting virulence, including genes involved in the yeast-hyphal switch, genes regulated by pH, and genes whose ORFs contain repetitive elements [3]. This data will also be a tremendous resource for the community, as labs will be able to compare the genomic sequence and expression profile of their gene(s) of choice across multiple clinical isolates from different clades. In addition, it is expected that future experiments will involve generating recombinant strains and confirming potential virulence factors by quantitative trait loci (QTL) analyses.

As a platform for sharing views of the data between the collaborating labs, we will use the Broad's Integrated Genome Viewer. This will allow integrated display of SNPs, genomic coverage, and RNA-sequence alignment data, along with other feature tracks such as annotated genes or repeats.

## **5.** Community Support and Collaborator Roles:

In the U.S. alone there are currently at least 40 NIH funded laboratories investigating *Candida* biology, and many more that are examining related mechanisms of fungal pathogenesis in other species. The strains, data, and analyses generated in this study will

be made readily available to the community and will serve as a resource for investigating strain diversity, population genetics, drug resistance, and mechanisms of pathogenicity. There are currently no studies that have analyzed a large collection of clinical fungal strains to determine how natural variation between isolates can lead to differences in virulence and drug resistance. Lack of a conventional sexual cycle means that a genomic approach is the most appropriate method for analysis of traits affecting *C. albicans* biology.

In the future, the generated data set will be used to compare and contrast genes affecting other important aspects of *C. albicans* biology, including commensal growth and biofilm formation. Only a handful of genes have been identified that influence the ability to grow as a commensal in the gastrointestinal tract [20], but the ability to compare colonization by the set of sequenced strains and to analyze gene differences is likely to ignite interest in this area of research. Similarly, biofilm formation is a critical first step in infections arising from medical-device associated infections (e.g. from catheters, synthetic heart valves, etc.), yet the pathways affecting this process are still poorly defined [21]. It is therefore expected that new regulators of commensalism and biofilm formation will be uncovered by comparison of gene polymorphisms among the sequenced set of strains. Finally, the ability to mate strains and generate recombinant organisms using the parasexual mating cycle researched in the Bennett lab [22], will also allow this data set to be used for mapping of *QTLs* that affect *C. albicans* phenotypes including virulence, biofilm formation, and drug resistance.

All the data that are generated from this project will be made available to the community as per the NIAID data release policies. The raw primary sequence data and assemblies will be made available in accordance with the Broad protocol. In addition, as CGD is the primary community database for *C. albicans*, we will make this data available for incorporation into CGD, or other appropriate NIAID data repositories.

Project Collaborators:
The overall project will be performed by the following four research groups according to the organizational plan outlined below:

1. Christina Cuomo, Ph.D.
   Broad Institute.

2. Richard Bennett, PhD.
   Brown University.

3. Mairi Noverr, PhD.
   Louisiana State University.

4. Paul Fidel, Ph.D.
   Louisiana State University.

Genomic DNA and total RNA for each of the clinical isolates of *C. albicans* will be provided by the Bennett and Noverr laboratories. The Broad Institute Sequencing Platform will carry out all sequencing on Illumina HiSeq machines. Project management staff at the Broad will coordinate sample transfer and sequencing work. Christina Cuomo will be

primarily responsible for coordinating the computational analysis of the sequence data by appropriate personnel at the Broad, working with collaborators to interpret the results of the data and engage them in the genomic analysis methods, and ensuring that the goals of the project are met.

Funding sources:
The Bennett lab is funded by NIAID (RO1 AI081704, R21 AI081560 and R56 AI087401), NSF (MCB-1021120) and an investigator in the Pathogenesis of Infectious Disease Award from the Burroughs Wellcome.

The Noverr lab is funded by NIAID (R01 AI072406-01) and a Scientist Development Award from the American Heart Association.

The Fidel lab is funded by NCRR CoBRE grant (P20 020160).

## 6. Availability & Information of Strains:

All of the strains required for this study (Table 1) are available in the laboratories of Bennett, Noverr, and Fidel. The Bennett laboratory is already experienced with the generation of high quality DNA and RNA for microarray analyses, and will provide these samples to the Broad for next-generation sequencing. The Fidel laboratory has experience isolating high quality *C. albicans* RNA from *in vivo* vaginal lavage samples.

Please see attached spreadsheet for details of *C. albicans* strains to be used in this study. All isolates are *C. albicans* strains and most have also been analyzed to determine which clade of the species they belong to, as indicated.

**7. Compliance Requirements:**

**7a. Review NIAID's Reagent, Data & Software Release Policy:**

*NIAID supports rapid data and reagent release to the scientific community for all sequencing and genotyping projects funded by NIAID GSC. It is expected that projects will adhere to the data and reagent release policy described in the following web sites.*

*http://www3.niaid.nih.gov/research/resources/mscs/data.htm*

*http://grants.nih.gov/grants/guide/notice-files/NOT-OD-08-013.html*

*<Each Center to include their website that describes/points to the guidelines>*

*Once a white paper project is approved, NIAID GSC will develop with the collaborators a detailed data and reagent release plan to be reviewed and approved by NIAID.*

Accept ☒ Decline ☐

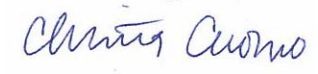
**7b. Public Access to Reagents, Data, Software and Other Materials:**

The *C. albicans* strains sequenced by this project will be deposited in the BEI repository or alternatively in the ATCC. All sequence and assemblies will be made available in compliance with the NIAID data release policy.

**7c. Research Compliance Requirements**

*Upon project approval, NIAID review of relevant IRB/IACUC documentation is required prior to commencement of work. Please contact the GSC Principal Investigator(s) to ensure necessary documentation are filed for / made available for timely start of the project.*

**Investigator Signature:**



**Investigator Name:** Christina Cuomo

**Date:** 12/10/2010

**Investigator Signature:**

**Investigator Name:** Richard Bennett

**Date:** 12/10/2010

**References**


1. Butler, G., et al., *Evolution of pathogenicity and sexual reproduction in eight Candida genomes.* Nature, 2009. **459**(7247): p. 657-62.
2. Schmid, J., et al., *Evidence for a general-purpose genotype in Candida albicans, highly prevalent in multiple geographical regions, patient types and types of infection.* Microbiology, 1999. **145 ( Pt 9)**: p. 2405-13.
3. Zhang, N., et al., *Distribution of mutations distinguishing the most prevalent disease-causing Candida albicans genotype from other genotypes.* Infect Genet Evol, 2009. **9**(4): p. 493-500.
4. Wu, W., et al., *Heterozygosity of genes on the sex chromosome regulates Candida albicans virulence.* Mol Microbiol, 2007. **64**(6): p. 1587-604.
5. Pujol, C., M. Pfaller, and D.R. Soll, *Ca3 fingerprinting of Candida albicans bloodstream isolates from the United States, Canada, South America, and Europe reveals a European clade.* J Clin Microbiol, 2002. **40**(8): p. 2729-40.
6. Ibrahim, A.S., et al., *Effects of ploidy and mating type on virulence of Candida albicans.* Infect Immun, 2005. **73**(11): p. 7366-74.
7. Taylor, B.N., et al., *In vivo virulence of Candida albicans isolates causing mucosal infections in people infected with the human immunodeficiency virus.* J Infect Dis, 2000. **182**(3): p. 955-9.
8. Rahman, D., et al., *Murine model of concurrent oral and vaginal Candida albicans colonization to study epithelial host-pathogen interactions.* Microbes Infect, 2007. **9**(5): p. 615-22.
9. Maccallum, I., et al., *ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads.* Genome Biol, 2009. **10**(10): p. R103.
10. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
11. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
12. Uhl, M.A., et al., *Haploinsufficiency-based large-scale forward genetic analysis of filamentous growth in the diploid human fungal pathogen C.albicans.* Embo J, 2003. **22**(11): p. 2668-78.
13. Setiadi, E.R., et al., *Transcriptional response of Candida albicans to hypoxia: linkage of oxygen sensing and Efg1p-regulatory networks.* J Mol Biol, 2006. **361**(3): p. 399-411.
14. Walker, L.A., et al., *Genome-wide analysis of Candida albicans gene expression patterns during infection of the mammalian kidney.* Fungal Genet Biol, 2009. **46**(2): p. 210-9.
15. Andes, D., et al., *A simple approach for estimating gene expression in Candida albicans directly from a systemic infection site.* J Infect Dis, 2005. **192**(5): p. 893-900.
16. Parkhomchuk, D., et al., *Transcriptome analysis by strand-specific sequencing of complementary DNA.* Nucleic Acids Res, 2009. **37**(18): p. e123.
17. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
18. Bruno, V.M., et al., *Comprehensive annotation of the transcriptome of the human fungal pathogen Candida albicans using RNA-seq.* Genome Res, 2010.
19. Wang, L., et al., *DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.* Bioinformatics, 2010. **26**(1): p. 136-138.
20. Rosenbach, A., et al., *Adaptations of Candida albicans for growth in the mammalian intestinal tract.* Eukaryot Cell, 2010. **9**(7): p. 1075-86.
21. Nobile, C.J. and A.P. Mitchell, *Genetics and genomics of Candida albicans biofilm formation.* Cell Microbiol, 2006. **8**(9): p. 1382-91.
22. Alby, K. and R.J. Bennett, *Sexual reproduction in the Candida clade: cryptic cycles, diverse mechanisms, and alternative functions.* Cell Mol Life Sci, 2010.

Supplemental Table 1. Characteristics of strains targeted by this project.

| Strain | Genotype | Origin | Geographical Location | Clade | Reference |
|---|---|---|---|---|---|
| 12C# | a/a | VP(or) | Mich. (USA) | I | 2 |
| L26# | a/a | VP(v) | Iowa (USA) | I | 2 |
| P94015 | a/a | BSI | Utah (USA) | I | 5 |
| P87 | a/a | HIV+(or) | South Africa | SA | 1 |
| P37005# | a/a | H(or) | Fla. (USA) | I | 3 |
| P60002 | a/a | BSI | Ariz. (USA) | SA | 5 |
| 19F | α/α | VP(v) | Mich. (USA) | I | 2 |
| GC75 | α/α | H(or) | South Africa | SA | 1 |
| P78048 | α/α | BSI | Manitoba (Canada) | I | 4 |
| P57072 | α/α | BSI | Iowa (USA) | II | 4 |
| P34048 | α/α | BSI | Turkey | ND | 5 |
| P37037# | a/α | H(or) | Wisc. (USA) | I | 3 |
| P37039 | a/α | BSI | NJ (USA) | I | 3 |
| P75010 | a/α | BSI | Belgium | E | 4 |
| P75016 | a/α | BSI | Israel | SA | 4 |
| P57055 | a/α | BSI | Omaha (Neb.) | III | 4 |
| P75063# | a/α | BSI | France | SA | 4 |
| P76055 | a/α | BSI | Iowa (USA) | II | 4 |
| P76067 | a/α | BSI | Ontario (Canada) | nd | 4 |
| P78042# | a/α | BSI | Indianapolis | III | 4 |
| SC5314# | a/α | BSI | Unknown | nd | 6 |
| CHN1# | ND | CI | Unknown | nd | 7 |
| 3153A# | ND | CI | Unknown | nd | 8 |
| A48# | ND | HIV+ (or) | ACTG | nd | 8 |
| A123 | ND | HiV+ (or) | ACTG | nd | 8 |
| A92 | ND | HIV+ (or) | ACTG | nd | 8 |
| A67 | ND | HIV+ (or) | ACTG | nd | 8 |
| A46# | ND | HIV+ (VP) | ACTG | nd | 8 |
| A155 | ND | HIV+ (or) | ACTG | nd | 8 |
| A20 | ND | HIV+ (or) | ACTG | nd | 8 |
| A203 | ND | HIV+ (or) | ACTG | nd | 8 |
| A84 | ND | HIV+ (or) | ACTG | nd | 8 |

Abbreviations in table: BSI, Bloodstream isolate; VP, vaginitis patient; v, vaginal sample; or, oral sample; CI = clinical isolate of unknown origin; ACTG = AIDS clinical trial group; HIV+, HIV-positive; H, healthy individual; nd, not determined.
# Candidate strains for transcriptional profiling.

References for strains:

1. Blignaut, E., Pujol, C., Lockhart, S., Joly, S. and Soll D.R. (2002) Ca3 fingerprinting of *Candida albicans* isolates from human immunodeficiency virus-positive individuals reveals a new clade in South Africa. J Clin Microbiol 40: 826-836.
2. Lockhart, S.R., Reed, B., Pierson, C.L, and Soll, D.R. (1996) Most frequent scenario for recurrent *Candida* vaginitis is strain maintenance with "substrain shuffling": demonstration by sequential DNA fingerprinting with probes Ca3, C1, and CARE2. J Clin Microbiol 34: 767-777.
3. Lockhart, S.R., Pujol, C., Daniels, K., Miller, M., Johnson, A. and Soll D.R. (2002) In *Candida albicans*, white-opaque switchers are homozygous for mating type.
Genetics 162: 737-745.

4. Pujol, C., Pfaller, M. and Soll, D.R. (2002) Ca3 fingerprinting of *Candida albicans* bloodstream isolates from the United States, Canada, South America, and Europe reveals a European clade. J Clin Microbiol 40: 2729-2740.

5. Wu, W., Lockhart, S.R., Pujol, C., Srikantha, T., Soll, D.R. (2007). Heterozygosity of genes on the sex chromosome regulate *Candida albicans* virulence. Molecular Microbiology 64 (6): 1587-1604

6. van het Hoog, M., Rast, T.J., Martchecnk, M., Grindle, S., Dignard, D., Hogues, H., Cuomo, C., Berriman, M., Scherer, S., Magee, B.B., Whiteway, M., Chibana, H., Nantel, A., Magee, P.T. (2007). Assembly of the *Candia albicans* genome into sixteen supercontigs aligned on the eight chromosomes. Genome Biol. 8:R52.

7. Noverr, M.C., Huffnagle, G.B. (2004). Regulation of *Candida albicans* morphogenesis by fatty acid metabolites. Infect. Immun. 72: 6206-6210.

8. Taylor, B.N., Fichtenbaum, C., Saavedra, M., Slavinsky, J., Swoboda, R., Wozniak, K., Arrivas, A., Powderly, W., and Fidel, P. L. (2000). In vivo virulence of *Candida albicans* isolates causing mucosal infections in people infected with the Human Immunodeficiency Virus. Journal of Infectious Disease 182:955-999.