We are resubmitting a white paper entitled, "Resequencing multiple strains of *Saccharomyces cerevisiae*". Both the Comparative Genome Evolution (CGE) working group and the Coordinating Committee (CC) raised a number of important issues regarding this project. Below we have summarized these concerns as well as our response.

1. *Number of strains*. Both the CGE and CC pointed out that the genome sequence of 5 or 20 strains may be just as valuable as 100 strains. We agree that many population and quantitative genetics questions could be answered with 5 or 20 strains. However, the key difference between 100 and even 50 strains (see Figure 3) lies in their value for association studies. As outlined in the white paper and summarized below, *S. cerevisiae* provides an excellent system in which to develop and experimentally validate novel methods of association that combine whole-genome genotype data, comparative genome sequence data, expression and linkage data.

2. *Relevance to human health and biology*. The CC raised the question of whether yeast is really a good model for population genetics that could provide broader lessons for human studies, either conceptually or methodologically. We agree that many aspects of yeast population and quantitative genetics are not relevant to understanding human health and biology. However, yeast does provide an excellent system to develop and test computational methods and statistical models for distinguishing functional from nonfunctional polymorphism. Specifically, the yeast data would make it possible to:

i. Test the extent to which sequencing conservation across species can be used to identify function polymorphism within a species. There are many possible strategies to estimating the probability a polymorphism is functional. The development and experimental validation of such methods would be of direct value to identifying functional polymorphism in humans.
ii. Develop and test novel methods of association. These may combine both direct and indirect methods of association along with linkage and expression data. Development of these methods would be valuable to understanding why both yeast and humans show high rates of genetic variation in gene expression and what downstream phenotypic consequences there might be.

3. *Community value*. The CGE noted that the community of yeast biologists working on population and quantitative genetics is small. This is true. However, the same was true of comparative genomics before the genome sequences of multiple *Saccharomyces* species became available. The generation of this sequence data combined with high-throughput, 96-well phenotype data would pose the broadly relevant question to the computational and statistical genomics community: to what extent can we association a genotype with a phenotype given whole-genome genotype data?

4. *Use of the data*. The CC noted that it was not clear how the data would be used. We have now outlined the specific uses of the data, which are summarized as follows:

i. Association studies: combine sequence data with high-throughput phenotyping to create a resource for testing novel methods of linking phenotype to genotype.
ii. Genetic variation in gene expression: provide a resource for elucidating the relationship between genotype, gene expression and phenotype.
iii. Quantitative traits: assess the frequency of functional polymorphism and combine

association studies with linkage data to identify mutations responsible for a variety of model quantitative traits, such as sporulation efficiency and gene expression levels.

iv. Comparative genomics: develop and experimentally validate computational methods for predicting and characterizing functional population genetic variation.

v. Yeast biology and genome annotation: identify differences in open reading frames and other features annotated in the reference genome in comparison to other yeast strains.

5. *Sequencing strategy*. The CGE suggested that 454 technology or sequencing by hybridization would make the project much cheaper. We agree. As outlined in the proposal, we propose to generate the majority of sequence data (88%) from 100 runs on a 454 sequencing machine.

A White Paper for

Sequencing Multiple Strains of *Saccharomyces cerevisiae*

Submitted by

Justin Fay, Washington University and
Leonid Kruglyak, Princeton University

In collaboration with
Washington University's Genome and Sequencing Center
February 3, 2006

Contact: Justin Fay, Department of Genetics and Center for Genome Sciences,
Washington University School of Medicine, 4444 Forest Park Ave, St. Louis, MO 63108
Phone: 314-747-1808, Fax: 314-362-2156, Email: jfay@genetics.wustl.edu

Summary

The fields of quantitative genetics, population genetics and human genetics all seek a comprehensive understanding of DNA sequence variation segregating in natural populations. Multiple genome sequences of the same species provide a key resource to achieving this goal. We propose to sequence 100 strains of *S. cerevisiae*. The generation of this sequence data combined with phenotype data, such as gene expression level and rate of drug metabolism, linkage data, and cross-species genome sequence data would make it possible to use computational methods and statistical models to address the broadly relevant question: to what extent can we association a genotype with a phenotype given whole-genome genotype data? Specifically, this data would:

i. address the extent to which sequence conservation across species can be used to identify functional polymorphism within a species, e.g. how best to define sequence conservation and what types of changes within conserved sequences are most likely functional.
ii. make it possible to empirically test the value of whole-genome genotype data in association studies, e.g. genome-wide scans of direct rather than indirect tests of association and coalescence-based association scans that combine data from linked markers,
iii. enable computational or statistical methods that identify functional polymorphism to be experimentally validated using yeast genetic and genome technologies.

More broadly, the data would be valuable to similar projects in other organisms, including humans, serving as both a guide as well as catalyzing the development of novel algorithms and methods that make use of whole-genome genotype data.

Introduction

The genome sequence of a model organism is invaluable to its scientific community. However, the full value of a genome sequence depends on the annotation of genes and their cis-

regulatory sequences. The genome sequences of multiple closely related species have resulted in significant progress in achieving this goal. In this proposal, we argue that the genome sequences of multiple organisms of the same species are just as valuable to understanding the genetic basis and evolution of phenotypic variation present in natural populations.

The current cost of sequencing prohibits extensive re-sequencing in humans and most model systems. However, as the cost of sequencing drops, our ability to link an organisms phenotype to its genotype will instead be limited by sample size and our ability to extract biologically relevant polymorphism from the much larger class of non-functional polymorphism. This will require the development of computational and statistical methods that maximize the value of whole-genome genotype data. For example, direct association studies could be performed on a much smaller subset of polymorphism in the human genome if *the vast majority of functional polymorphism resides within sequences that are conserved across species*.

The justification for conducting a whole-genome polymorphism survey in *Saccharomyces cerevisiae* is two-fold. First, it will catalyze the development of computational methods for the analysis of whole-genome polymorphism studies and will provide a complete set of genotype data for testing novel whole-genome association methods. The potential is that yeast will be just as valuable to future whole-genome polymorphism surveys as it was for cross-species comparative genome sequencing projects. Second, the power of yeast genetics combined with genome technologies makes it the best system to functionally validate computational and statistical methods of identifying functional polymorphism and associating it with a phenotype.

In this proposal, we first describe data that would be generated and how it would be used. We then describe its relevance to human health and biology. Finally, we describe preliminary data, materials and methods, and community support.

Whole-genome polymorphism data

Whole-genome polymorphism data enumerates all functional polymorphism responsible for differences among organisms within the same species as well as the much larger class of non-functional polymorphism. A variety of classes of polymorphism are known to segregate within natural populations.

Mutational classes of polymorphism:

i. Rearrangements, e.g. translocations, inversions.
ii. Copy number polymorphism, e.g. insertions, deletions.
iii. Single-nucleotide polymorphisms (SNPs).

The value of whole-genome sequencing data depends on our ability to identify functional polymorphism and accurately associate it with a downstream phenotype. For some classes of polymorphism, the functional impact is easy to infer, e.g. frame-shifts. In other cases, the function, if any, is very difficult to infer, e.g. SNPs in noncoding sequences.

Functional classes of polymorphism data:

i. Gene content, e.g. insertions or deletions of exons or entire genes.
ii. Gene structure, e.g. polymorphism in start, stop codons or splice sites.
iii. Gene regulation, e.g. gain, loss or modulation of cis-regulatory sequences.
iv. Gene function, e.g. fame-shift or missense mutations.

Regardless of how easy it is to distinguish functional from non-functional polymorphism, whole-genome sequencing provides a high-throughput method of identifying population genetic variation. To conduct a whole-genome polymorphism survey in yeast we propose to:

i. Sequence 90 strains of *S. cerevisiae* at low coverage (2-3x) and 10 strains at draft coverage (~6x). Low coverage sequencing will identify SNPs and small insertion/deletion polymorphism. At 2-3x coverage, ~80% of each strain's genome will be sequenced and consequently every base will be sequenced in an average of ~80 different strains. As described in more detail later, an effective sample size 80 will make it possible to associate a putatively functional polymorphism with a phenotype. Draft coverage will identify large copy number polymorphism and rearrangements, which may represent a significant class of functional polymorphism (Tuzun *et al*., 2005).
ii. Use a combination of capillary and 454 sequencing technology. With 454 sequencing 2-3x coverage can be obtained from a single sequencing run, assuming 25 Mbp of sequence data or more per run. To obtain draft coverage (6x) for ten of the strains, capillary sequencing will be used to obtain 3-4x with 4 kb plasmids and 0.1x with 40 kb fosmids in addition to the 2-3x obtained from 454 sequencing. Assuming a genome size of 13.5 Mbp, 12 Mbp from the finished S288C genome plus 1.5 Mbp of rDNA repeats, the entire project would entail 285x coverage of the yeast genome and 3.85 Gbp of raw sequence data, ~88% of the data being obtained from 100 454 sequencing runs.

Use of whole-genome polymorphism data

Whole-genome polymorphism data is of direct value to population and quantitative genetics. While the community of yeast biologists interested in polymorphism data is small, significant achievements have been made in understanding genetic variation in gene expression (Brem *et al*., 2002; Yvert *et al*., 2003; Ronald *et al*., 2005; Brem *et al*. 2005; Storey *et al*. 2005; Brem and Kruglyak, 2005) and in dissecting quantitative traits to the nucleotide level (Steinmetz *et al*., 2002; Deutschbauer and Davis, 2005). A whole-genome polymorphism would provide additional resources to these projects; furthermore, it would motivate a large variety of computational and statistical methods that would be useful to whole-genome polymorphism surveys in other organisms. The large body of comparative genomics methods developed after the sequencing of multiple yeast species provides an excellent example of the potential value of a whole-genome polymorphism survey in yeast. Although a polymorphism survey in other organisms would also be valuable, yeast provides the unique opportunity to rapidly validate computational and statistical methods using existing genetic and genomic technologies.

Association studies: Whole-genome polymorphism data from 100 strains of yeast combined with high-throughput phenotyping would create a valuable resource for testing novel methods of linking phenotype to genotype (Figure 1). These methods could then combine this data with expression data, linkage analysis and comparative genomics methods to empirically test their ability to identify functional polymorphism underlying a diverse array of phenotypes. Rather than trying to enumerate a list of novel methods or concepts that could be tested in yeast, we next outline a single illustrative example.

A reverse genetics approach to association studies: Currently, whole-genome association studies face a huge multiple hypothesis testing problem. This is because a large number of neutral markers must be tested for an indirect association with a phenotype through linkage disequilibrium with the functional polymorphism. This problem could be ameliorated by

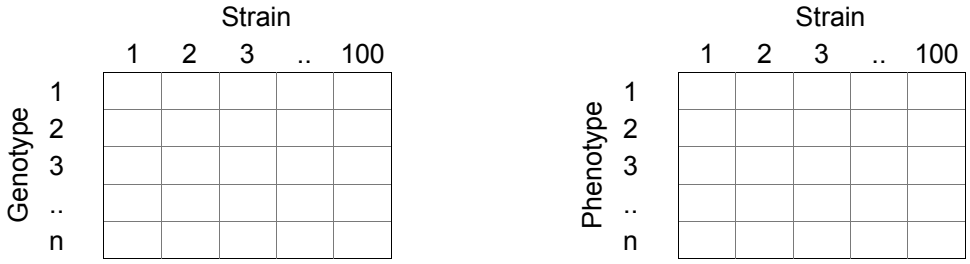

Figure 1. A whole-genome polymorphism survey (left) in combination with high-throughput phenotyping (right) provides a valuable resource for association studies.

direct tests of association between a phenotype and all potentially functional polymorphism. Potentially functional polymorphisms can be identified using comparative genomics methods. However, there are likely thousands of potentially functional polymorphisms. Thus, a whole-genome polymorphism survey is particularly well suited to a reverse genetics approach.

A reverse genetics association study:

i. identify candidate functional polymorphism that disrupts annotated genes or regulatory sequences of interest and/or occurs in coding or noncoding sequences conservation across species, and
ii. associate the candidate polymorphism with a molecular or organismal phenotype.

This approach could vastly reduce the problem of multiple tests faced by whole-genome association studies by testing a candidate polymorphism for association with a small array of relevant molecular or organismal phenotypes. Thus, candidate polymorphism in genes and their regulatory sequences could be associated with molecular phenotypes, such as expression level, metabolite level, protein activity or the activity of a known pathway. Although this limits the identification of new genes that may contribute to a phenotype of interest, other combinations of linkage and association data may not be limited in this manner.

This reverse genetics association approach is commonly used in humans. What is not known is the extent to which this approach would be more powerful if i) *all* polymorphism at a locus were identified, and ii) conservation across species, which could be defined in a variety of ways, could be used to conduct direct tests of association on a small number of candidates rather than using LD based methods.

Genetic variation in gene expression: High rates of genetic variation in gene expression have been found in all organisms surveyed, including both yeast (Brem *et al*., 2002; Cavalieri *et al*. 2000; Townsend *et al*., 2003; Fay *et al*., 2004; Yvert *et al*. 2003;) and humans (Enard *et al*. 2002; Rockman and Wray 2002; Cowles *et al*. 2002; Bray *et al*. 2003; Lo *et al*. 2003; Whitney *et al*. 2003; Pastinen *et al*. 2004; Morley *et al*. 2004). Dissecting the genetic basis and downstream consequence of this variation is important to understanding its contribution to phenotypic variation and and human genetic diseases. A whole-genome polymorphism survey would provide a valuable resource to studies aimed at elucidating the relationship between genotype, gene expression and phenotype (Cavalieri *et al*., 2000; Brem *et al*., 2002; Yvert *et al*., 2003; Fay *et al*., 2004). Specifically, cis-regulatory polymorphism in known or predicted transcription factor binding sites could be directly associated with differentially expressed genes.

Analysis of quantitative traits: Yeast provides the capability of rapidly identifying the

molecular basis of quantitative traits. Two recent studies have identified naturally occurring mutations affecting a high temperature growth phenotype (Steinmetz *et al*., 2002) and sporulation efficiency (Deutschbauer and Davis, 2005). Both cases support a long-held view in quantitative genetics research that major effect QTL can be caused by multiple linked mutations rather than a single mutation of large effect. The later study also showed that defects in sporulation efficiency were caused by a combination of relatively rare mutations in the lab strain S288C. A whole-genome polymorphism survey would make it possible to quickly assess the frequency of functional polymorphism as well as make it possible to combine association studies with linkage data to identify other mutations responsible for defects in sporulation found in other yeast strains.

Comparative genomics: Comparative genomics methods use sequence conservation between species to identify and characterize functional noncoding seqeunces, particularly those involved in transcriptional regulation. A natural extension of these methods is to use them to identify functional polymorphism based on sequence conservation across species. Although some methods have been developed, many are specific to protein coding sequences, e.g. (Ng and Henikoff, 2001). A whole-genome polymorphism survey would provide the data needed to develop and experimentally validate computational methods for predicting and characterizing functional population genetic variation.

Single base resolution is not needed to identify functional SNPs. In the context of a codon, three bases, or a transcription factor binding site, 4-8 bases, functional SNPs can be identified using their context. We have shown that just three Saccharomyces species can reliably identify functional transcription factor binding sites (Doniger *et al*. 2005), which implies that the same methods could be used to identify functional SNPs that disrupt conserved transcription factor binding sites. With a large number of yeast genome sequences, polymorphism in protein coding sequences may also be reliably classified since proteins can easily be aligned from distantly related species.

Yeast biology and genome annotation: The comparative annotation of multiple strains of *S. cerevisiae* would be valuable to understanding yeast biology and biological variation. Annotation of the *S. cerevisae* genome has been greatly improved through the use of sequence comparisons between species. For example, after comparison with other species, the S288C reference genome was reannotated to have 503 fewer genes, 43 new genes, and 540 genes with a change in the start or stop codon (Kellis *et al*., 2003).

In many instances the S288C reference genome may not be wildtype, in which case the genome sequence of additional strains of yeast could lead to changes in gene annotation. For example, S288C is known to contain a premature stop codon in *FLO8*, resulting in a defect in psuedohyphal growth and flocculation (Liu *et al*., 1996). In most laboratory and industrial strains the aquaporin *AQY2* is split into two open reading frames by an eleven bp deletion and both functional and nonfunctional alleles of the *AQY1* gene have been found in different strains (Laize *et al*., 2000). The *MPR1* and *MPR2* genes, responsible for resistance to a proline analogue, are present in some strains but not S288C (Shichiri *et al*., 2001). The genome sequence of a small number of yeast strains would reveal any other S288C anomalies. A substantial number are likely to exist in the subtelomeric regions which are known to vary in size between different strains of yeast (Carro *et al*., 2003; Dunn *et al*., 2005). Identifying the functional significance of variation in gene content among strains would greatly benefit from multiple complete or nearly complete genome sequences.

Relevance to human health and biology

Association studies play an essential role in identifying the molecular basis of human genetic diseases. The completion of the HapMap project has provided a valuable resource to association studies that aim to, i) identify a causal mutation following a linkage study, ii) test candidate genes, and iii) conduct whole-genome scans. However, association studies face a number of issues relevant to this proposal:

i. Multiple tests of association require low significance thresholds, e.g. $P < 10^{-6}$, and consequently very large sample sizes.
ii. Linkage disequilibrium can limit the resolution of association studies, such as when the causal mutation is in complete association with a number of other SNPs forming a haplotype block.
iii. Causal mutations in noncoding sequences are not as easily identified as those in coding sequences, e.g. only 1% of mutations in the human gene mutation database consist of regulatory mutations (Cooper et al. 1998).

The direct relevance of sequencing 100 strains of *S. cerevisiae* to human health and biology is that it would make it possible to test whether these issues may be partially or completely avoided by, i) direct rather than indirect (LD based) tests of association, in combination with ii) using sequence conservation across species to narrow the list of candidate causal mutations. Specifically, the yeast data would make it possible to combine computational methods, statistical models and empirical validation to:

i. identify what level of sequence conservation should be used to distinguish conserved and unconserved sequences,
ii. identify what types of polymorphism within conserved sequences are likely to disrupt function, i.e. use models of transcription factor binding sites to distinguish functional and non-functional changes similar to distinguishing synonymous, missense and nonsense changes in coding sequences,
iii. test statistical models based on direct rather than indirect tests of association, e.g. coalescence-based association scans that estimate association by combining data from linked markers (Marchini and Donnelly, 2006).

Identifying a complete list of all polymorphism present in the human genome would undoubtedly be valuable to human health and biology. The value of this data, however, depends on our ability to distinguish functional from non-functional polymorphism. The yeast data would make it possible to determine the extent to which various computational and statistical methods may accomplish this goal.

If the whole-genome genotype-phenotype associations are successful in yeast, the sequence of 100 human genomes in combination with comparative mammalian sequence data could be used to identify all potentially functional polymorphism segregating at appreciable frequency (>5%) in the entire human population. Although there are approximately 10 million SNPs in the human genome, only a small unknown fraction (say 1%) may be potentially functional and relevant to direct association studies.

Thus, a major motivation for sequencing multiple strains of *S. cerevisiae* is that it will improve our ability to associate human polymorphism with human disease. While it is not clear

how powerful these methods may become, current advances in sequencing technology suggest that whole-genome polymorphism data will soon be available for humans, e.g. the Cancer Genome Project.

Preliminary projects

Four strains of *S. cerevisiae* have been sequenced besides the S288C reference genome (Table 1). A large number of phenotypic differences exist between these strains. RM11 phenotypes include more than 1500 expression differences in comparison to S288C (Brem *et al*., 2002). YJM789 has the ability to grow in immuno-compromised animals, compared to S288C which cannot (Steinmetz *et al*., 2002). M22 produces hydrogen sulfide, YPS163 is sensitive to copper sulfate, and both show a large number of expression differences with respect to S288C (Fay *et al*. 2004). We have also found a large number of drug dependent growth rate differences from a screen of a 1280 compound drug library (unpublished data).

The identification of functional population genetic variation was the primary motivation behind the sequencing of two yeast strains to ~2.5x coverage (Table 1). This data was collected at Washington University's Genome and Sequencing center as part of an internally funded pilot and feasibility program. Although the analysis of this data is only preliminary, we have found a total of 42k SNP differences, 0.55%/bp, between M22 and S288C and 51k, 0.64%/bp, between YPS163 and S288C. A significant number of these SNPs reside in stretches of noncoding sequences that are conserved across species. For example, we have found 1101 SNPs that occur within 5 bps of conserved noncoding sequence. The statistical models needed to address these preliminary data have yet to be developed. In addition, we do not yet know whether these SNPs are rare, strain-specific SNPs, or not.

In addition to the strains listed in Table 1, we have learned, through communications with Dr. Ed Lewis, that he has begun sequencing a handful of laboratory and natural isolates of *S. cerevisiae* along with a much larger number of *S. paradoxus* strains. It may be possible to combined resources to prevent any overlap between these projects, as was the case with *S. mikatae* and *S. bayanus* (Kellis *et al*. 2003; Cliften *et al*. 2003).

Table 1. Yeast sequencing projects.

Strain	Source	Coverage	Contigs	N50 Contig size (kb)	Investigator(s)
RM11	Vineyard, USA	~8	115		Leonid Kruglyak
YJM789	Clinical, USA	~10	295		L. Steinmetz & R. Davis
M22	Vineyard, Italy	2.6	3946	2.9	Justin Fay
YPS163	Oak, USA	2.8	3330	3.0	Justin Fay
S. paradoxus		7.7	832	51.0	Kellis et al. (2003)
S. mikatae		5.9	1650	20.0	Kellis et al. (2003)
S. bayanus		6.4	1098	25.0	Kellis et al. (2003)
S. kudriavzevii		3.4	2074	4.9	Cliften et al. (2003)
S. castellii		3.8	2050	5.1	Cliften et al. (2003)
S. castellii after prefinishing		3.9	741	14.8	Cliften et al. (2003)

Coverage is the sequencing redundancy of assembled contigs.

Materials and methods

Choice of strains: Strains must be capable of sporulation and should be as diverse as possible, both at the phenotypic and DNA level. Sporulation is a requirement for any subsequent genetic studies. Sporulation also ensures that either haploid or homozygous diploid strains can be created for sequencing, i.e. strains with no heterozygosity. DNA sequence variation has been examined at five unlinked genes (~7kb) in a collection of 81 strains of *S. cerevisiae* (Figure 2), isolated from natural sources, vineyards, immunocompromised individuals and various fermentations (Fay and Benavides, 2005). Divergence is 0.34%/bp between strains, on average. This data indicated that vineyard and sake strains represent two distinct groups of yeast that appear to have been independently domesticated; as such, inclusion of members from both groups may inform the history of these events. Our strategy for strain selection is to: i) select only strains that can sporulate, ii) select strains of interest from the yeast community, iii) select strains showing the greatest divergence at the DNA sequence level, and iv) select strains isolated from a diverse array of sources. Because some of the strains that have been suggested by the yeast community have no preliminary sequence data, each strain would be sequenced at the five genes surveyed by Fay and Benavides (2005). The total list of candidate strains include: 81 strains described in Fay and Benavides (2005), a wine strain from Spain used by a number of laboratories, 7 Nigerian palm wine strains collected by Dr. O. Ezeronye, and 44 strains collected from geographically diverse natural sources by Dr. G. Naumov. These sources include soil from South Africa, wild grape berries from Armenia, tree sap from Sri Lanka, and tree exudate from Japan.

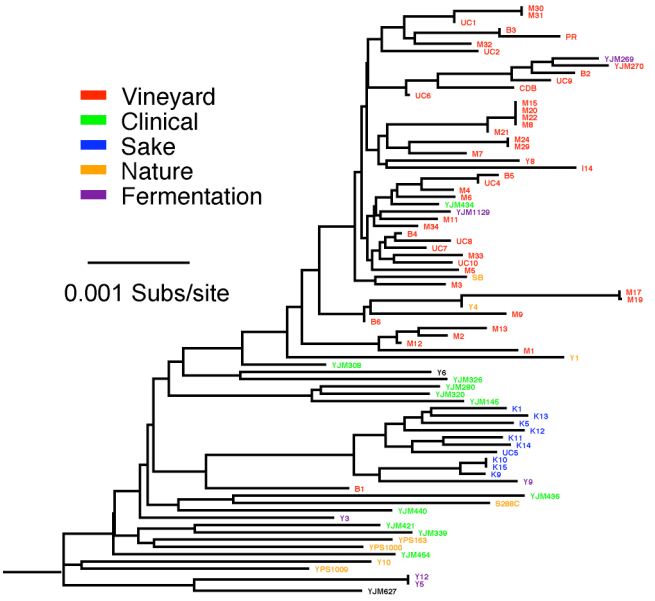

Figure 2. Neighbor-joining tree of 81 strains of *S. cerevisiae* based on 7 kb of sequence from 5 unlinked loci. Strains are colored based on the source from which they were isolated.

Number of strains and coverage: The number of genomes sequenced and their coverage is directly related to the value as well as the cost of the data. We have proposed 10 strains at high coverage (6x) and 90 strains at low coverage (2-3x). High coverage and low coverage sequencing are aimed at identifying different classes of variation:

i. High coverage - rearrangements, large insertions and deletions, polymorphism within repetitive sequences such as recently duplicated genes, transposable elements, subtelomeric sequences, telomeres and centromeres.
ii. Low coverage - small insertions and deletions and SNPs.

Justification for 10 strains at high coverage - Recent work has shown that the human genome is polymorphic for rearrangements, large insertions or deletions and repetitive sequences (Eichler, 2006). A significant fraction of this polymorphism may be functional and is thus

relevant to identifying all functional variation segregating at appreciable frequencies within a species. The rationale for 10 strains at high coverage is that the majority of common variation, i.e. >20%, would be detected. Although a sample size of 10 would be too small to associate these variations with a phenotype, once discovered they could be typed in a much larger number of strains. We next address the question of whether it would be more cost effective to sample and then genotype for SNPs as well.

Justification for 90 strains at low coverage - The power of an association study is directly related to sample size. Typically, association studies require sample sizes much larger than 100. However, direct tests of association require much smaller sample size (Ohashi and Tokunaga, 2001). We have calculated (Sokal and Rohlf, 1995) the power of direct association as a function of the P-value cutoff for a sample size of 100 in comparison to a sample size of 50 (Figure 3). For an effect of a mutation equal to one environmental standard deviation (a relative risk of 2 assuming a threshold character), power is much higher for a sample size of 100 compared to 50.

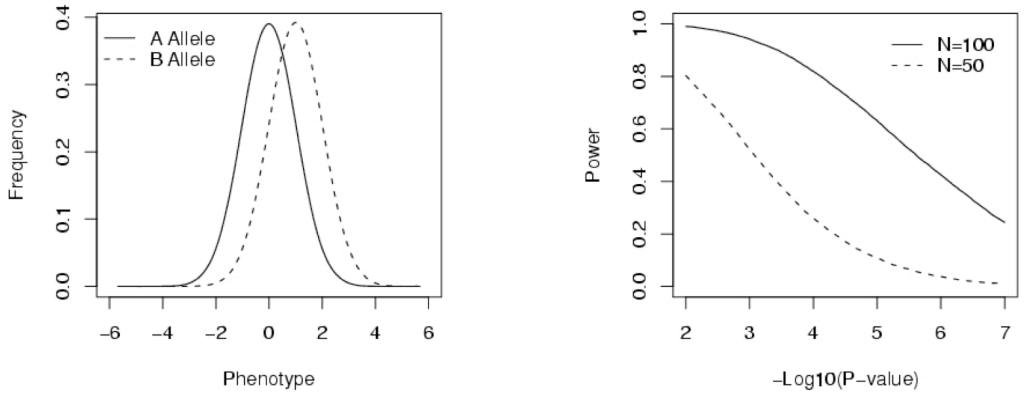

Figure 3. Plot of phenotypic distribution of two allelic classes (Left) and the power of detecting an association between the phenotype and the causal mutation (A–>B) as a function of the P–value cutoff (Right).

Figure 3 shows that the power of *direct* tests of association are sufficient for a whole-genome scan. A P-value cutoff of 10^{-6} corresponds to a Bonferroni correction for 50,000 tests. We have identified about 50,000 SNPs in our preliminary data and so this provides a reasonable approximation for the number of common SNPs that would be tested in a genome-wide scan. A much smaller number of tests, say 5000 or $P < 10^{-5}$, would be needed to test all *potentially functional* polymorphism in the genome. Even fewer tests would be needed if one were specifically interested in sporulation efficiency, the expression level of a single gene or if a reverse genetics approach were taken and a single SNP were tested for association with an array of phenotypes.

The power of association is substantially reduced for a sample size of 50 compared to 100 (Figure 3). Because shotgun sequencing is incomplete, the actual coverage for each SNP is likely to be in between 50 and 100. A strain at 2-3x coverage should result in coverage of approximately 80% of the reference S288C genome. Lander-Waterman would predict 88-95% but our preliminary data is closer to 80% coverage of the S288C genome. Thus, our proposal would provide data for ~80% strains at any given position in the genome.

Genotyping versus sequencing: An alternative strategy would be to identify polymorphism in a much smaller number of strains and genotype 100 strains or more. This strategy only works for indirect tests of association since many functional mutations may reside at a frequency of ~10% in the population. A sample size of 10 would obtain only 65% of SNPs at

10% frequency and fewer SNPs at a lower frequency. In comparison, a sample size of 100 would obtain more than 99.9% of SNPs at 10% frequency and 99.4% of SNPs at 5% frequency in the population.

Methods

DNA: Each strain will be made homozygous by deriving isogenic strains from single haploid spores. Since the mitochondrial genome is ~70 kb and present at roughly 50 copies per cell, it will be removed by isolating rho mutants through treatment with ethidium bromide. DNA will be extracted and sheared for library construction. No DNA contaminants are expected. Preparation of DNA can be done in a short period of time.

Sequencing: We propose whole-genome shotgun sequencing. The use of homozygous strains will greatly facilitate assembly of whole-genome shotgun sequence data. For 6x strain coverage this will include 4 kb plasmids and 40 kb fosmids. The fosmids will make it possible to assemble these genomes independent of the S288C reference genome.

We propose a combination of two shotgun sequencing technologies: 454 pyrosequencing and traditional capillary sequencing. Traditional capillary sequencing is necessary for plasmid and fosmid end-reads. 454 pyrosequencing provides 100-fold increase in capacity and is well suited to whole-genome genotyping.

Facilities: The proposed project can be carried out at Washington University's Genome Sequencing Center in collaboration with Dr. Richard Wilson and Dr. Elaine Mardis, who have both expressed enthusiasm for this project. A number of 454 sequencing projects have already been conducted at Washington University's Genome and Sequencing Center, making it an excellent place to scale up the Center's 454 sequencing capacity.

Support

Given the possibility of sequencing multiple strains of *S. cerevisiae*, we have established community support for a yeast Genome REsequencing Project (GREP). The main goal was to ensure that members of the yeast community were aware of the proposed research and could submit any strains that would be of value to their own work. GREP was met with considerable enthusiasm and numerous strains have been suggested for sequencing. Investigators that have corresponded their support for both sequencing and data analysis or who have contributed strains are listed below.

Investigator	Institution
John McCusker	Duke University
Julian Adams	University of Michigan
Audrey Gasch	University of Wisconsin
Duccio Cavalieri	Harvard University
Jose Perez	University of Valencia
Sakkie Pretorius	Australian Wine Research Institute
Daniel Hartl	Harvard University
Lars Steinmetz	European Molecular Biology Laboratory
Barak Cohen	Washington University
Edward Louis	University of Nottingham
Gianni Liti	University of Leicester
Michael Travisano	University of Houston
Michael Eisen	Lawrence Berkeley National Laboratory
Maitreya Dunham	Princeton University
Obioha Ezeronye	Michael Okpara University of Agriculture
Cletus Kurtzman	National Center for Agricultural Utilization Research
Gennadi Naumov	State Institute for Genetics and Selection of Industrial Microorganisms
Paul Sniegowski	University of Pennsylvania
Mark Johnston	Washington University
Justin Fay	Washington University
Leonid Kruglyak	Princeton University
Vladimir Jiranek	University of Adelaide

References

Bray, N.J., P.R. Buckland, M.J. Owen and M.C. O'Donovan. 2003. Cis-acting variation in the expression of a high proportion of genes in human brain. *Hum Genet* **113**: 149-153.

Brem, R.B., J.D. Storey, J. Whittle and L. Kruglyak. 2005. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* **436**: 701-703.

Brem, R.B., G. Yvert, R. Clinton and L. Kruglyak. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752-755.

Cavalieri, D., J.P. Townsend and D.L. Hartl. 2000. Manifold anomalies in gene expression in a vineyard isolate of Saccharomyces cerevisiae revealed by DNA microarray analysis. *Proc Natl Acad Sci U S A* **97**: 12369-12374.

Cliften, P., P. Sudarsanam, A. Desikan, L. Fulton, B. Fulton, J. Majors, R. Waterston, B.A. Cohen and M. Johnston. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* **301**: 71-76.

Cooper, D.N., E.V. Ball and M. Krawczak. 1998. The human gene mutation database. *Nucleic Acids Res* **26**: 285-287.

Cowles, C.R., J.N. Hirschhorn, D. Altshuler, E.S. Lander, P. Green, W. Miller, E.D. Green, I. Ruvinsky and G. Ruvkun. 2002. Detection of regulatory variation in mouse genes. *Nat Genet Nature Development* **32**: 432-437.

Deutschbauer, A.M. and R.W. Davis. 2005. Quantitative trait loci mapped to single-nucleotide resolution in yeast. *Nat Genet* **37**: 1333-1340.

Doniger, S., J. Huh and J.C. Fay. 2005. Identification of functional transcription factor binding sites using closely related Saccharomyces species. *Genome Research* **15**: 701-709.

Eichler, E.E. 2006. Widening the spectrum of human genetic variation. *Nat Genet* **38**: 9-11.

Enard, W., P. Khaitovich, J. Klose, S. Zollner, F. Heissig, P. Giavalisco, K. Nieselt-Struwe, E. Muchmore, A. Varki, R. Ravid, G.M. Doxiadis, R.E. Bontrop and S. Paabo. 2002. Intra- and interspecific variation in primate gene expression patterns. *Science* **296**: 340-343.

Fay, J.C. and J.A. Benavides. 2005. Evidence for domesticated and wild populations of Saccharomyces cerevisiae. *PloS Genetics* **1**: 66-71.

Fay, J.C., H.L. McCullough, P.D. Sniegowski and M.B. Eisen. 2004. Population genetic variation in gene expression is associated with phenotypic variation in Saccharomyces cerevisiae. *Genome Biol* R26.

Kellis, M., N. Patterson, M. Endrizzi, B. Birren and E.S. Lander. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**: 241-254.

Lo, H.S., Z. Wang, Y. Hu, H.H. Yang, S. Gere, K.H. Buetow and M.P. Lee. 2003. Allelic variation in gene expression is common in the human genome. *Genome Res* **13**: 1855-1862.

Marchini, J., S. Myers, G. McVean and P. Donnelly. 2006. A novel bayasian approach to localizing disease genes.

Morley, M., C.M. Molony, T.M. Weber, J.L. Devlin, K.G. Ewens, R.S. Spielman and V.G. Cheung. 2004. Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**: 743-747.

Ng, P.C. and S. Henikoff. 2001. Predicting deleterious amino acid substitutions. *Genome Res* **11**: 863-874.

Ohashi, J. and K. Tokunaga. 2001. The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. *J Hum Genet* **46**: 478-482.

Pastinen, T., R. Sladek, S. Gurd, A. Sammak, B. Ge, P. Lepage, K. Lavergne, A. Villeneuve, T. Gaudin, H. Brandstrom, A. Beck, A. Verner, J. Kingsley, E. Harmsen, D. Labuda, K. Morgan, M.C. Vohl, A.K. Naumova, D. Sinnett and T.J. Hudson. 2004. A survey of genetic and epigenetic variation affecting human gene expression. *Physiol Genomics* **16**: 184-193.

Rockman, M.V. and G.A. Wray. 2002. Abundant raw material for cis-regulatory evolution in humans. *Mol Biol Evol* **19**: 1991-2004.

Ronald, J., R.B. Brem, J. Whittle and L. Kruglyak. 2005. Local Regulatory Variation in Saccharomyces cerevisiae. *PLoS Genet* **1**: e25.

Sokal, R.R. and F.J. Rohlf. 1995. Biometry: the principles and practice of statistics in biological research. 887.

Steinmetz, L.M., H. Sinha, D.R. Richards, J.I. Spiegelman, P.J. Oefner, J.H. McCusker and R.W. Davis. 2002. Dissecting the architecture of a quantitative trait locus in yeast. *Nature* **416**: 326-330.

Storey, J.D., J.M. Akey and L. Kruglyak. 2005. Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol* **3**: e267.

Townsend, J.P., D. Cavalieri and D.L. Hartl. 2003. Population genetic variation in genome-wide gene expression. *Mol Biol Evol* 955-963.

Tuzun, E., A.J. Sharp, J.A. Bailey, R. Kaul, V.A. Morrison, L.M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, M.V. Olson and E.E. Eichler. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37**: 727-732.

Whitney, A.R., M. Diehn, S.J. Popper, A.A. Alizadeh, J.C. Boldrick, D.A. Relman and P.O. Brown. 2003. Individuality and variation in gene expression patterns in human blood. *Proc Natl Acad Sci U S A* **100**: 1896-1901.

Yvert, G., R.B. Brem, J. Whittle, J.M. Akey, E. Foss, E.N. Smith, R. Mackelprang and L. Kruglyak. 2003. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat Genet* **35**: 57-64.