UNDERSTANDING GENOME STRUCTURE AND RESPONSE TO PERTURBATION

by

Ron Ammar

A thesis submitted in conformity with the requirements
for the degree of Doctorate of Philosophy

Department of Molecular Genetics
University of Toronto

Understanding genome structure and response to perturbation

Ron Ammar

Doctor of Philosophy

Department of Molecular Genetics, University of Toronto, 2013

Abstract

The past few decades have witnessed an array of advances in DNA science including the introduction of genomics and bioinformatics. The quest for complete genome sequences has driven the development of microarray and massively parallel sequencing technologies at a rapid pace, yielding numerous scientific discoveries. My thesis applies several of these genome-scale technologies to understand genomic response to perturbation as well as chromatin structure, and it is divided into three major studies. The first study describes a method I developed to identify drug targets by overexpressing human genes in yeast. This chemical genomic assay makes use of the human ORFeome collection and oligonucleotide microarrays to identify potential novel human drug targets. My second study applies genome resequencing of yeast that have evolved resistance to antifungal drug combinations. Using massively parallel genomic sequencing, I identified novel genomic variations that were responsible for this resistance and it was confirmed *in vivo*. Lastly, I report the characterization of chromatin structure in a non-eukaryotic species, an archaeon. The conservation of the nucleosomal landscape in archaea suggests that chromatin is not solely a hallmark of eukaryotes, and that its role in transcriptional regulation is ancient. Together, these 3 studies illustrate how maturation of genomic technology for research applications has great utility for the identification of potential human and antifungal drug targets and offers an encompassing glance at the structure of genomes.

Acknowledgments

In my doctorate work, I have been very lucky to be supervised by Corey Nislow and Gary Bader. Their direction has guided me through challenges and exciting successes, and I would like to express my gratitude for their unwavering support, critical feedback and valuable counsel. Corey's passion to "talk science" and use large-scale genetic methods has helped drive my research forward and shape my execution of the scientific method. Gary's vast experience with diverse genomic and bioinformatic work has been invaluable to me as I straddled the wet-lab bench and computational methods, serving as a great model for my work.

I would also like to thank my committee members Guri Giaever and Leah Cowen for the great advice and feedback they have given me on my research. In particular, I would like to thank Guri, who rigorously applies the scientific method to all of her projects. Working with her has ensured that my work is held to a high standard.

As a member of both the Bader and Giaever/Nislow labs, I have been privileged to work alongside many intelligent and collaborative scientists. I would like to thank all members of my labs past and present for help with everything ranging from scientific discussions to lab cleanup. The members include: Simon Alfred, Mohammed Alshalalfa, Anthony Arnoldo, Anatasia Baryshnikova, Nick Berbenetz, Florence Cavalli, Kahlin Cheung-Ong, Elke Ericson, Max Franz, Marinella Gebbia, Deena Gendoo, David Gfeller, Larry Heisler, Shirley Hui, Ruth Isserlin, Shobhit Jain, Pegah Khosravi, Jing Kittanakom, Melissa Landon, Brian Law, Anna Lee, Siyang Li, Elena Lissina, Christian Lopes, Daniele Merico, Magali Michaut, Jason Montojo, Marina Olhovsky, Vuk Pavlovic, Michael Proctor, Jüri Reimand, Igor Rodchenkov, Harold Rodriguez, Laure Sambourg, Daniel Shabtai, Xiaojian Shao, Tanvi Shekhar-Guturja, David Shih, Andrew Smith, Kevin Song, Oliver Stueker, Anu Surendra, Chris Tan, Nikko Torres, Kyle Tsui, Malene

Urbanus, Andrea Utrecht, Veronique Voisin, Omar Wagih, Iain Wallace, Zhun Yan, Khalid Zhuberi and Scott Zuyderduyn.

As well, I would like to thank my excellent collaborators who have worked with me on some of my projects, including Jason Moffatt, Jessica Hill and Leah Cowen. In addition, much of my work would not be possible without the help of the technical staff at the Donnelly Sequencing Centre and the Banting computer cluster: Thanks to Tanja Durbic, Danica Leung, Dax Torti and Jeff Liu.

Ontario Graduate Scholarships have funded my work for a large portion of my PhD for which I am grateful. I would also like to express my gratitude to the University of Toronto and the Donnelly Centre for Cellular and Biomolecular Research for the excellent research environment.

Finally I would like to thank my parents, sister, family and friends for their support over my many years of academic scholarship. And an especially big thank you to my fiancée Helen.

If I have seen further it is by standing on the shoulders of giants.

For Helen

Contents

Figures

Tables

Abbreviations

aa	amino acid
bp	base pair
cDNA	complementary DNA
CNV	Copy Number Variant
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
GC	guanine or cytosine nucleotide
gDNA	genomic DNA
HIP	HaploInsufficiency Profiling
HOP	Homozygous Deletion Profiling
MGC	Mammalian Gene Collection
MNase	micrococcoal nuclease
MSP	Multi-copy Suppression Profiling
nt	nucleotide
OD	Optical Density
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
RNA	ribonucleic acid
SGD	Saccharomyces Genome Database
SNV	Single Nucleotide Variant
TSS	Transcription Start Site
TTS	Transcription Termination Site
YKO	Yeast KnockOut deletion collection

1 Introduction to genomic technology

When the human genome sequencing project was being debated in the 1980s, it was compared to sending a man to the moon, and, at an estimated $3 billion, it would cost about the same (Brenner 2013). It is an oft-quoted fact that the span of time from the dawn of flight, when the Wright brothers built the first flying machine, to the Apollo 11 moon landing is just over 50 years. During this short period of time, scientists and engineers overcame many challenges to advance technology at an expeditious pace accomplishing manned space flight and lunar exploration. Equally astonishing is the short time period between Watson and Crick's discovery of the structure of DNA (Watson and Crick 1953) and the completion of the draft human genome sequence (Lander et al. 2001; Venter et al. 2001), spanning roughly 50 years as well. Just as advances in rocket science yielded great innovation leading to modern satellite systems and planetary exploring robotics, we find ourselves at the zenith in the study of genes and genomes. Application of these genomic tools and technologies promises to expand our comprehension of biological systems and fuel future innovation.

1.1 A brief history of DNA research

In a classic experiment by Griffith, mice were injected with two different strains of *S. pneumoniae*, a smooth type (Type III-S) and a rough type (Type II-R). The rough strain was nonvirulent, while injection of the virulent smooth strain resulted in mouse death. Griffith heat-killed the smooth strain and found that this sample was nonvirulent in mice. However, when he mixed the nonvirulent rough strain with the heat-killed smooth strain in a test tube and then injected the mixture into mice, the mixture killed the host (Griffith 1928). Based on this

observation a “transforming principle” was formulated stating that molecules from a dead bacterial strain could be transformed into a live strain to confer its genetic properties.

To identify which biomolecules were responsible for the transformation principle, Avery and colleagues heat-killed the smooth bacteria, deproteinized the bacteria in solution and enzymatically digested the polysaccharide bacterial capsule. They then precipitated the “active material” by alcohol fractionation and found that this active fraction was capable of inducing the transformation of rough strains into virulent smooth strains (Avery et al. 1944). This precipitated material, identified as nucleic acid, was studied by Levene and found to be composed of four nitrogenous bases (now known to be adenine, cytosine, guanine and thymine) linked to a five-carbon sugar and phosphate groups, and each base-sugar-phosphate unit was termed a nucleotide (Levene 1919). Nucleic acid was further solidified as the genetic information of the cell by the Hershey-Chase experiment. In this classic experiment, T2 bacteriophage were radiolabelled with either ^{32}P or ^{35}S. Since sulphur was primarily incorporated into peptides and not DNA, and phosphorus was incorporated into DNA, they were able to determine which labelled material was transferred to the bacteria after phage infection (Hershey and Chase 1952). After processing in a Waring blender and centrifuging to separate phage, they observed that high levels of extracellular ^{35}S and low levels of ^{32}P, suggesting that the viral protein coats remained outside the bacteria. Freezing-induced lysis of the bacteria revealed high levels of intracellular ^{32}P, suggesting that the phage DNA, rather than protein, was the genetic material of the cell.

In 1953, Watson and Crick, using X-ray diffraction data from Wilkins and Franklin, published a predicted molecular structure of double-stranded DNA postulating that “this structure has two helical chains each coiled round the same axis.” (Watson and Crick 1953). This model described the bases on the inside of the helical chains and the phosphates as outer groups, a “radically” different structure from the three chain structures proposed by Pauling and others (Pauling and Corey 1953). As research in the field of nucleic acids grew, Crick and others continued to investigate the concept of DNA as the universal genetic material of organisms. In a review, he claimed, “It is thus clear that the synthesis of proteins must be radically different from the synthesis of polysaccharides, lipids, co-enzymes and other small molecules; that it must be relatively simple, and to a considerable extent uniform throughout Nature; that it must be highly

specific, making few mistakes; and that in all probability it must be controlled at not too many removes by the genetic material of the organism." (Crick 1958). He continued to theorize that nucleic acids controlled the synthesis of protein.

Crick, Brenner and colleagues soon published a groundbreaking study demonstrating that DNA encoded proteins using triplets of bases (codons), and this code was likely degenerate, such that each amino acid could be encoded by more than one codon (Crick et al. 1961). Nirenberg and Matthaei subsequently showed that a polyuridylate RNA sequence directed the synthesis of a phenylalanine in a cell-free amino acid incorporating system, and concluded that the codon UUU encoded the amino acid phenylalanine (Matthaei et al. 1962). Subsequent work by Nirenberg, Ochoa and Khorana elucidated the genetic code describing the amino acids corresponding to each of the possible 64 triplet codons (Nirenberg et al. 1965; Khorana 1968).

Almost a decade later, extracts of *H. influenza* were found to possess endonuclease activity (Kelly and Smith 1970; Smith and Wilcox 1970). This "restriction" endonuclease was found to have no activity against the DNA of its source organism (it would "restrict"/degrade only foreign DNA) and was able to digest T7 phage DNA. Recombinant DNA technology was now possible when combined with the discoveries of bacterial genetic transformation using plasmid DNA and recognition site specificity of restriction endonuclease RI (Cohen et al. 1972; Hedgpeth et al. 1972). However, this technology was limited because scientists lacked a method to detect specific DNA sequences in recombinant clones.

Southern provided a solution with his eponymous blotting assay (Southern 1975). In the experiment, he performed agarose gel electrophoresis of EcoRI-digested DNA fragments and transferred these onto nitrocellulose filters. These fragments were subsequently hybridized to radiolabelled ribosomal RNAs (rRNA) from *E. coli* and *X. laevis*, demonstrating that 1) EcoRI cleaved each of 18S rRNA and 28S rRNA only once and 2) identifying conservation in this digestion pattern across five mammalian species. The Southern blot assay became the method of choice for screening recombinant clones and was soon adapted to large-scale analyses.

Southern blots were configured into "dot blots", a technique where samples were applied to a nitrocellulose filter in a specific pattern, facilitating analysis of multiple hybridization targets in

parallel (Kafatos et al. 1979). As well, researchers were starting to apply blotting methodology to larger biological screens. For example, in attempt to isolate eukaryotic genes, a method was designed to screen recombinant clones by growing and lysing cells on nitrocellulose filters, and probing these with radiolabelled yeast rRNA (Benton and Davis 1977). Arrays were also constructed by depositing ~380 *E. coli* clones transformed with mouse colon carcinoma cDNA (complementary DNA) onto nitrocellulose and probing with labelled cDNA from normal or tumor mouse tissues (Augenlicht and Kobrin 1982). Using this method, the expression of each cDNA could be determined for a given tissue type. This technology, employed to tools of computational biology, allowed this same assay to be expanded to systematically analyze ~4000 human cDNA probes and compute the relative expression of each cDNA for a set of human colonic carcinoma cells (Augenlicht et al. 1987). Blotting approaches were also used to screen for structural features of DNA by electrophoresing plasmid DNA, transferring it to a nitrocellulose filter and probing it with a monoclonal antibody (Hoheisel and Pohl 1987).

Until the late 1980s, most probes consisted of biologically derived material, and only a few studies had shown that synthetic oligonucleotides could be used as hybridization probes to detect base pair mismatching or multiple alleles (Wallace et al. 1979; Conner et al. 1983). Building upon these designs, Southern and colleagues explored a new blotting framework using synthetic oligonucleotides arrayed on a non-porous support (Maskos and Southern 1992). Allele-specific probes were synthesized *in situ* with solid phase chemical methods to a length of 19bp and used to detect three different alleles of the β-globin sickle cell locus (Maskos and Southern 1993). At the same time, microarrays were developed by preparing a small library of sequenced cDNAs or Expressed Sequence Tags (ESTs) and depositing these onto glass slides using robotic printers. The first such array facilitated detection of differential expression between wild-type and transgenic *A. thaliana* lines using two-colour fluorescence hybridization (Schena et al. 1995). Thus, with the development of these two technologies, the oligonucleotide microarray and the cDNA microarray, the modern microarray was born.

A key difference between microarrays and dot blots was that dot blot target sequences were typically arrayed on a support and hybridized to a single labelled probe. On microarrays, probes were arrayed on a membrane support and the target DNA would be applied to the entire array,

allowing a scientist to query many loci of a genomic sample in a single experiment. Critical to the development of microarrays was a change in the support material. Blotting techniques used porous supports, such as nitrocellulose, to provide a large binding surface area, however, the boundaries of the spots were diffuse and the oligonucleotides could not be deposited with precision.

As genomic sequence data for various organisms became available, the need for large-scale hybridization-based analysis became apparent, and it was not possible to further reduce the size of spots on porous supports (Southern 2001). The introduction of impermeable supports offered several advantages, specifically the ability to synthesize probes *in situ*, a uniform density of nucleotides, the lack of array feature diffusion and quicker interaction with target nucleotides. The best supports identified were glass and silicon (Shchepinov et al. 1997).

The deposition of probes onto microarrays was accomplished by multiple modes of fabrication, which eventually branched out to form different platform technologies. The simplest form was spotting presynthesized probes, such as PCR-amplified cDNAs, onto glass slides, but this was superseded by *in situ* synthesis of probes. The main methods for *in situ* synthesis used either ink-jet printers, flow channel irrigation or semiconductor-based photolithographic methods. The ink-jet methods were based on bubble-jet or ink-jet colour printing technologies, designed to propel microdroplets of ink at paper. These were adapted to propel solutions of nucleotide reagents onto glass supports; instead of four colours of ink, four different precursors for different bases could be printed to achieve solid-phase synthesis of oligonucleotides (Blanchard 1996). Flow channels or flow cells used a simpler approach, wherein precursors for the four bases were laid down in four strips on a square surface. Within each strip, the four bases were deposited again, and this process was iterated several times, the array was turned 90 degrees, and the process was repeated again to yield all possible oligonucleotides of a predetermined length with feature sizes as small as 10μm (Southern et al. 1992; Maskos and Southern 1993). Finally, photolithographic (light-directed) fabrication was another important method developed for oligonucleotide synthesis (Fodor et al. 1991). This method made use of photocleavable protecting groups, with base additions across the entire array requiring a set of patterned photolithographic masks. Each position of the predetermined length of oligonucleotide required four masks, one

for each potential base. For each base addition, the glass surface would be irradiated to remove the protecting group of the exposed nucleotide that was not covered by the mask. The glass surface would be subsequently coated with a coupling agent for the new base addition, and the process would be repeated such that for an array with 25bp probes, fabrication would require 100 masks (25 positions × 4 different nucleotides = 100 masks) (Southern 2001). The establishment of these methods opened up the possibility of conducting massively parallel analysis of biological targets. However, in order to accurately construct the oligonucleotide arrays, scientists had to know the DNA sequences of the loci that were being queried.

Well before the revolution in array technology has been reduced to practice, DNA sequencing technology was being pioneered. A first key advance was described in 1965 by Sanger *et al*. who used two-dimensional fractionation of radiolabelled nucleotides to identify di-, tri- and tetranucleotides of *E. coli* 16S and 23S rRNA based on a series of degradation products (modelled after early amino acid sequencing techniques) (Sanger et al. 1965). This method was used to determine the RNA nucleotide sequence of the RNA-based bacteriophage MS2 coat protein, the first complete gene sequenced (Min Jou et al. 1972). However, two-dimensional electrophoresis of digested fragments was so inefficient that the method yielded multiple small sequence fragments, and, to create the entire MS2 coat protein sequence, the authors based the nucleotide sequence on the known amino acid sequence of the protein. Their sequence data confirmed that RNA was translated based on the Ochoa-Nirenberg-Khorana codon table (Nirenberg and Matthaei 1961; Kellogg et al. 1966). The sequencing of the MS2 coat protein gene laid the groundwork for one of the landmarks in genome biology, decoding the complete sequence of the bacteriophage MS2 (Fiers et al. 1976). The MS2 RNA was determined to be 3569 nucleotides long.

Another major milestone in genome biology came when Sanger and colleagues determined the nucleotide sequenced of the single-stranded DNA bacteriophage ΦX174, the first complete DNA genome sequence (Sanger et al. 1977a). In addition to establishing a genome's sequence, this study also found that coding regions could overlap one another in DNA. This genome sequence was created using a novel "Plus and Minus" DNA sequencing method invented by Sanger and Coulson two years earlier which used *E. coli* DNA polymerase I or T4 DNA polymerase and a

primer to generate complementary strands of DNA on a single-stranded DNA template of interest (Sanger and Coulson 1975). The "Minus system" was based on the premise that in the absence of a single nucleotide, of a mix of all four nucleotides, *E. coli* DNA polymerase I would cease to extend the complementary DNA along the template when that base was required. So, for example, "-A" would contain cytosine (C), guanine (G), thymine (T), but not adenine (A), and the polymerase would synthesize until it encountered a thymine on the template strand. The "Plus system" worked on the premise that in the presence of a single nucleotide, T4 DNA polymerase would degrade double-stranded DNA from its 3' end until it encountered the nucleotide that was present. So, "+A", where only dATP (deoxyadenosine triphosphate) was present, would only produce DNA chains that terminated in an adenine. Polyacrylamide gel electrophoresis (PAGE) would resolve the fragment sizes for all four minus and four plus reactions to produce DNA sequence reads up to ~85bp in length (Sanger and Coulson 1975).

At the same time that Sanger *et al.* published the complete sequence of ΦX174, Maxam and Gilbert described a simpler method to sequence double-stranded DNA based on terminal labelling of chemically fragmented DNA (Maxam and Gilbert 1977). The technique used distinct reactions to cleave specifically at the end of each of the four nucleotides, radiolabel the 5' ends of the fragments, and visualize the fragment lengths by PAGE. Maxam-Gilbert sequencing became the method of choice for most scientists because it used double-stranded DNA, which could be easily purified, compared to Sanger and Coulson's plus and minus method, which required each sequence read to be cloned to form single-stranded DNA.

However, a novel "chain termination" method, published by Sanger and colleagues, superseded Maxam-Gilbert sequencing, and due to its efficiency, accuracy and lack of toxic chemicals such as hydrazine, it eventually set the standard for DNA sequencing experiments. Sanger sequencing was a variant of the plus and minus method, requiring single-stranded template DNA, *E. coli* DNA polymerase I, deoxynucleotriphosphates (dNTPs or nucleotides) and, the additional vital component, chain-terminating dideoxynucleotriphosphates (ddNTPs) (Sanger et al. 1977b). These ddNTPs were identical in structure to dNTPs except for the absence of a 3' hydroxyl group, the result of which prevented further DNA elongation when incorporated into a complementary strand by polymerase. Initially, sequence reads were typically up to 50bp in

length, with the first ~15bp being unreadable due to primer annealing (Sanger et al. 1977b). It was not until single-stranded phage vectors were introduced that Sanger's technique could be broadly applied (Gronenborn and Messing 1978; Sanger 1981).

These two sequencing methods, became the most predominant DNA sequencing methods used in research for the next two decades, with Maxam-Gilbert sequencing often referred to as "chemical sequencing" and Sanger sequencing often referred to as "dideoxy sequencing". Maxam-Gilbert sequencing represented a chemical cleavage paradigm akin to amino acid sequencing, while Sanger's chain-termination method offered a more versatile "sequencing by synthesis" paradigm.

With new sequencing methods came novel sequencing strategies. Sanger and colleagues pioneered a technique called "shotgun sequencing" which involved fragmenting DNA with restriction endonucleases and cloning these into single-stranded DNA bacteriophage to obtain templates for dideoxy sequencing (Sanger et al. 1980). The goal was to obtain a sufficient number of sequence reads such that the entire genome sequence could be reconstituted from the fragments as a contiguous sequence or "contig". In concert, a novel computational pipeline was created for the storage of DNA reads from DNA gel data and the overlapping of fragments to form a genomic contig (Staden 1980). This strategy was used to successfully sequence the viral genome of bacteriophage λ, a DNA genome 48502bp in length (Sanger et al. 1982).

Numerous other sequencing strategies emerged as well. The shotgun sequencing strategy was classified as a random strategy because of the random selection of subclones from which the genomic sequence was generated. Various insert lengths of genomic DNA would be cloned into different host vectors, including yeast artificial chromosomes (YACs, 100-1000kbp), cosmids (30-45kbp) and λ phage (100bp-2.5kbp). Smaller subclones of these large clones would be generated in single-stranded vector systems to generate a nested DNA template while maintaining a physical map (Hunkapiller et al. 1991). Shotgun sequencing was effective in determining the bulk of any large sequence, such as a YAC, but the number of clones required to completely sequence the entire YAC or cosmid was prohibitively high, due to the random nature of subclone generation. To solve this problem, shotgun sequencing was used to determine up to 90% of the large fragment sequence, and directed sequencing strategies would be used to fill in

the gaps. Directed methods permitted sequential sequencing of DNA fragments such that the reads were related and overlapped with one another. The simplest variant, the "walking method", entailed repetitive cycles of synthesis of new primers based on the previously acquired read data (Studier 1989). Other directed methods included construction of nested deletion clones using exonucleases or clones that contained randomly distributed transposons (Poncz et al. 1982; Adachi et al. 1987). Computational methods were also developed to overlap abundant sequence reads to form contigs for both random and directed strategies (Rice et al. 1991).

DNA sequence analysis was now a multi-step process, divided into several stages: 1) DNA fragmentation and cloning, 2) sequence determination and 3) data analysis. Sequence read lengths had peaked at ~500bp using PAGE due to the limits on fragment migration as a function of the log of fragment length (Birren et al. 1990). Furthermore, as researchers pushed to increase the number of samples loaded on a gel (up to 96) the ability to distinguish one sample from another, or lane tracking, became more difficult. The best way to obtain more sequence data was via automation of sequencing reactions and omitting the sample loading step entirely. This was accomplished using gel-filled capillaries (Swerdlow and Gesteland 1990). The capillaries were filled with linear polyacrylamide as opposed to crosslinked polyacrylamide allowing for a significant decrease in electrophoresis time and a significant increase in read resolution, up to 1kbp, when compared to the polyacrylamide slab gels. Two sequencing platforms emerged that made use of automated capillary sequencing. The first system used a single label, four lane approach, containing a specific ddNTP in each of the four lanes, as in the original PAGE dideoxy sequencing scheme. Four reactions were performed in four distinct capillaries, and a computer would read the sequence by scanning the resultant fragments in the capillaries (Ansorge et al. 1987). The ddNTPs were fluorescently labelled, for simple excitation and detection by automated machines. The second system used a four label, single lane approach, where four distinct colours were used for four fluorescent ddNTPs. These four ddNTPs were combined in a single reaction and could be resolved within a single capillary because each base corresponded to a specific colour, which could be detected by a computer (Connell et al. 1987; Prober et al. 1987).

By the late 1980s, DNA sequencing technology had matured sufficiently to facilitate the sequencing of free-living organisms larger than phage, and the ultimate goal of sequencing the

human genome was on its way to becoming possible. Sequences of many human genes were determined in individual labs and then on a more "industrial scale" when Venter and colleagues partially sequenced randomly selected cDNA clones from human brain tissue (Adams et al. 1991; Adams et al. 1993). These partially sequenced cDNA fragments were known as Expressed Sequence Tags (ESTs) (Putney et al. 1983). With estimates for the completion of the human genome sequence over a decade away, these ESTs offered scientists a manner to identify human gene sequences before the entire human genome sequence was completed. It also allowed researchers to generate microarrays of these sequences to measure gene expression profiles of human brains in different conditions. The first major genome project initiated was the *S. cerevisiae* (yeast) genome. This was built upon the high resolution physical map of *S. cerevisiae* total nuclear DNA constructed using a global restriction mapping approach (Olson et al. 1986; Link and Olson 1991). After determining the size of double-digested restriction fragments from random clones they were computationally paired to generate contigs which were assembled to determine the genomic organization of yeast. In 1989, an international consortium was created to sequence the 12Mbp genome of the S288c strain of *S. cerevisiae* (Levy 1994). Individual chromosomes were sequenced by over 600 researchers from over 100 labs using a variety of libraries using both shotgun (random) and directed approaches, and chromosome sequences were published independently over the span of several years (SGD ; Oliver et al. 1992; Feldmann et al. 1994; Bussey et al. 1995; Jacq et al. 1997). However, less than a year before the yeast genome sequence was completed, the distinction of first complete genome sequence of a free-living organism went to *Haemophilus influenzae* when Venter and colleagues published its 1.8Mbp genome (Fleischmann et al. 1995). In contrast to the yeast genome initiative, Venter and colleagues opted to circumvent nested directed and random strategies to avoid time-consuming mapping of clones. Instead, they adopted a workflow that took advantage of assembly software, the TIGR assembler (TIGR, The Institute for Genomic Research), developed for sequencing ESTs, to assemble thousands of sequence reads into megabase-sized contigs. This allowed them to sequence the entire *H. influenzae* genome accurately and cost-effectively using only a shotgun strategy, with coverage estimates based on a statistical model developed by Lander and Waterman (Lander and Waterman 1988). The yeast genome sequence was completed a year later, and became the first eukaryotic genome sequence (Goffeau et al. 1996; Goffeau 1997). It

catalogued ~6000 genes on 16 chromosomes assembled from ~300000 sequence reads (Mewes et al. 1997).

With a wealth of DNA sequences being published, public access to sequence data became a major consideration. Nucleotide sequence databases from the Los Alamos National Laboratory and the European Molecular Biology Laboratory were organized into a single database, as part of a community project called GenBank (Kanehisa et al. 1984; Burks et al. 1985). Maintenance of the database was later undertaken by the National Center for Biotechnology Information at the National Institutes of Health. There was now sufficient biological data for researchers to compare sequences within or across species and identify regions of homology and conservation. Modern sequence analysis algorithms were created to compare nucleotide or protein sequences, including the Smith-Waterman algorithm which performed local sequence alignment (Smith and Waterman 1981). This algorithm was optimized to align regions of sequences that exhibited high similarity, while ignoring dissimilar regions. However, the algorithm exhibited great precision at the expense of time when querying a sequence against an entire database of sequences, so faster algorithms were developed. FASTA performed these pairwise comparisons with greater efficiency by using regions of similarity to restrict the alignment search space, and popularized the now standard FASTA format for biological data (Lipman and Pearson 1985). Algorithms were made even faster, with almost equivalent accuracy, and in 1990, the Basic Local Alignment Search Tool (BLAST) was created (Altschul et al. 1990). It offered extremely rapid database querying with accurate results, and it became a standard tool to statistically identify homologous biological sequences in GenBank. In addition to pairwise comparisons, algorithms were also created to compute multiple sequence alignments (MSAs), such as CLUSTAL, which aided in the identification of sequence conservation and the construction of phylogenetic trees (Higgins and Sharp 1988). Along with sequence analysis tools, the field of computational biology expanded rapidly to accommodate desired analyses including protein 3D visualizations and gene expression heatmaps among others.

DNA research entered the mainstream with the completion of the Human Genome Project (HGP). The HGP was officially founded in 1990, with the goals of producing a genetic linkage map (fingerprinted DNA libraries), improving the efficiency of DNA sequencing technology and

the completion of a human haploid genomic sequence (Barnhart 1989). The construction of a human genetic linkage map was conceived of a decade earlier by Botstein, Davis and colleagues (Botstein et al. 1980). The authors proposed that restriction fragment length polymorphisms (RFLPs), a newly developed molecular tool to establish genetic linkage, could be used to create a linkage map for the human genome. These maps were vital to establish the relative position of sequenced loci (Donis-Keller et al. 1987; Lander and Green 1987). The bulk of the human genome was sequenced using capillary electrophoresis with the one label, four lane approach and the instruments achieved read lengths of ~600bp with 24-hour automated operation yielding 115kbp per day (Mardis 2011). In 2001, the 3Gbp draft haploid human genome sequences were released, one from the HGP, now known as the RefSeq, and one from a private firm, Celera Genomics (Lander et al. 2001; Venter et al. 2001).

The HGP drove the development of automated DNA sequencing technologies and high-density microarray platforms and inspired the formation of many companies that sought to commercialize and benefit from the growing interest in genomic science. With the demand for personalized human genomes growing and the requirement of many research facilities to study a cornucopia of strains, species and cell lines, multiple important microarray and massively parallel sequencing platforms emerged at the turn of the millennium. These platforms represent the current state-of-the-art for DNA research.

1.2 Microarrays

For most research applications, microarrays are being supplanted by massively parallel sequencing technologies. However, a few microarray technologies are still widely-used in research and diagnostic settings, and there are a several system platforms for these purposes.

One of the most popular platforms is the Affymetrix GeneChip DNA microarray system. The chips contains 25bp oligonucleotides immobilized onto a quartz wafer synthesized *in situ* using the aforementioned photolithography manufacturing process to form 5μm features (Dalma-Weiszhausz et al. 2006). Sample DNA is sheared, biotin-labelled, hybridized to the arrays and

stained by binding biotin to streptavidin-conjugated phycoerythrin. Arrays are scanned, and images of the features are processed to yield probe hybridization intensity data.

Agilent produces another widely-used DNA microarray platform using ink-jet based (5 inks: 4 nucleotides and a catalyst) *in situ* printed oligonucleotides 60bp in length to form 30μm features (Wolber et al. 2006). Samples are analyzed by generating cDNA with incorporated Cy3- or Cy5-labelled cytosine. The cDNA is hybridized to the microarray probes and intensities are detected with a scanner at a 2μm resolution. Since both Cy3 and Cy5 dyes can be used, the Agilent platform is capable of two-channel expression analysis where two samples can be assessed relative to one another.

The newest microarray platform is the Illumina BeadArray system. Synthetic oligonucleotides are immobilized onto 3μm silica beads that self-assemble onto arrays. Originally, these beads were assembled onto fiber optic bundles, where each bundle consisted of 50000 fibers. Each fiber contained a micro-etched well to house a single bead, and light could pass through the fiber uninterrupted to determine signal intensity (Oliphant et al. 2002). Currently, the system uses etched silicon wafers as an array of microwells for the random silica bead self-assembly. Once the beads are randomly positioned in the wells, the system identifies the specific sequences on each bead using a decoding process (Gunderson et al. 2004). Sample cDNA oligonucleotides are hybridized to the bead array with either Cy3- or Cy5-labelled common primers. The fluorescence signal is detected by a reader to determine hybridization intensity, and the system is capable of two-channel detection.

1.3 Algorithm development for microarrays

Because probes must uniquely identify the many species of DNA within a sample, there are several criteria for probe design that must be considered. In the case of expression arrays, where cDNA is hybridized to an array, probes must uniquely identify transcripts. A single probe is often insufficient to detect an entire transcript, so multiple probes are used as a probe set. During the array design phase, these DNA probes are selected from and then screened against known genomic sequence to ensure that they are unique, and this can be done using an algorithm like

BLAST (Altschul et al. 1990). More advanced probe design software ensures that for a desired probe length, there is optimal hybridization with minimal secondary structure, optimal GC composition and desired melting temperature (Rouillard et al. 2003).

Microarrays offer a massively parallel approach to probing biological material for specific sequences. With arrays often querying in excess of a million probes per experiment, the resulting output must be processed by computer software. Typically, microarray scanners use a laser to excite a fluorescent dye that is bound to the sample DNA, and the emitted signal is scanned and processed. Before hybridization intensities can be interpreted, several analyses are performed. First, the features are identified on the array using gridding and segmentation algorithms, to extract specific feature intensities (Dalma-Weiszhausz et al. 2006; Wolber et al. 2006). After background correction, which removes fluorescence due to non-specific binding of the sample to the array surface, the feature intensities are extracted. With feature intensities computed, the relative abundance of sample bound to each probe can be determined. Typically, if more than one microarray is being considered, the data are normalized to allow comparison between samples.

To determine the relative abundance of specific mRNAs, algorithms are required to summarize the probe intensities of all probes in a specific probe set. Affymetrix arrays, a popular array platform, used an initial design consisting of perfect match (PM) and mismatch (MM) probes, where MM probes contained a non-reference nucleotide at the centre position. Because sample DNA would not hybridize as effectively to the MM probes, the microarray suite 5.0 (MAS5) algorithm could use the mean $\log_2$ differences between PM and MM probes to determine overall probe set abundance. However, having a MM probe for every PM probe was an expensive and inefficient use of microarray real estate, occupying 50% of the microarray surface. Robust multi-array analysis (RMA), a log scale linear additive model, was created to summarize probe intensities and it did so without considering MM probes and achieved accurate results (Irizarry et al. 2003). RMA is currently the standard for probeset summarization on microarrays. Its model uses only PM probes and is based on sample material available to bind features, scanner measurement error and probe affinity (Okoniewski and Miller 2008).

Because RMA does not require MM probes, current Affymetrix array designs use a set of ~25000 control probes with varying GC composition that are designed not to match the reference sequence (genome, transcriptome, ORFeome, etc.). For each PM probe, a detection above background (DABG) analysis is performed using background probes with the same GC content to determine probe intensity significance, and probes that are not significant are rejected from subsequent analysis (Okoniewski and Miller 2008).

After microarray data has been extracted from any of the aforementioned platforms, it can be subjected to many analyses to normalize, classify and cluster gene expression patterns (Berrar et al. 2003). In particular, a visualization commonly associated with gene expression analyses is the expression heatmap, which allows researchers to observe trends in gene expression across multiple conditions. These analyses have been extended to massively parallel sequencing, which is capable of performing similar sample detection.

Microarray data are commonly archived in public databases, primarily the NCBI Gene Expression Omnibus using microarray standards such as the Minimum Information About a Microarray Experiment (MIAME) guidelines (Brazma et al. 2001; Edgar et al. 2002; Brazma 2009).

1.4 Applications of microarray technology

Despite the emergence of massively parallel sequencing, microarray assays are still available. A key limitation of microarray technology is that arrays can only be designed if the DNA sequence of genes or genomes has been previously determined, so that probes can be appropriately designed to query biological samples.

One of the first applications of DNA microarrays was to determine gene expression, and this remains a primary use of both cDNA and oligonucleotide array technology (Schena et al. 1995; Lockhart et al. 1996; Wodicka et al. 1997). Gene expression profiling is a vital application that allows scientists to determine transcript abundance of genes to identify coregulated genes by expression correlation and determining differential gene and pathway regulation. Many critical findings of the last decade have been based on expression profiling, including the identification

of induced pluripotency factors, which was accomplished by studying genes upregulated in embryonic stem cells and tumors (Takahashi and Yamanaka 2006).

Chromatin immunoprecipitation (ChIP), on a microarray, assays the specificity of DNA-binding proteins such as transcription factors and histones. ChIP-chip (a microarray-based ChIP) works by crosslinking the proteins to DNA, immunoprecipitating them, de-crosslinking and hybridizing the DNA fragments. To effectively perform ChIP experiments, a microarray must cover the entire genome so tiling arrays are typically used for ChIP-chip experiments. Unlike most microarrays, tiling arrays contain probes designed across the entire genome at a specific interval, such that the entire genomic locus is queried by the array. The first such experiment was described for cohesin binding to chromosome 3 of *S. cerevisiae* (Blat and Kleckner 1999).

A related method is also widely-used to determine chromatin architecture via nucleosome occupancy. Histones are crosslinked to DNA and unbound DNA is preferentially digested using micrococcal nuclease (MNase). The remaining fragments represent histone-bound genomic DNA, and these are hybridized to a microarray (Bernstein et al. 2004; Lee et al. 2007).

With the ability to query millions of probes in a single experiment, the opportunity to count molecular barcodes from pooled experiments is another useful application of microarrays. This was initially pioneered to count the relative abundance of yeast knockout strains, but can be applied for any counting assay (Shoemaker et al. 1996). Oligonucleotides can also be synthesized to create k-mer arrays, where, for example, every possible DNA 10-mer can be synthesized on an array. This approach has been used to identify binding specificities of transcription factors (Berger et al. 2006).

Another key application of microarrays is to determine polymorphisms or mutations within a sample. This is achieved using Single Nucleotide Polymorphism (SNP) arrays, which often contain multiple tiled probes for each known SNP. In this manner, an individual can be genotyped for thousands of SNPs in a single experiment, as was originally demonstrated for human genome polymorphisms (Wang et al. 1998). In the Affymetrix design, each SNP requires two probes, one with the polymorphism and one with the wildtype allele. The biological sample will bind more efficiently to the probe to which it is 100% complementary, and the SNP is

detected based on hybridization intensity of one probe compared to the other (LaFramboise 2009). The Illumina SNP array hybridizes sample DNA to the probe with the query position at the end of the probe. This is extended by a single base, and the specific nucleotide incorporated is detected to determine the genotype (LaFramboise 2009).

In addition to single nucleotide variation, microarrays are capable of assessing larger-scale variation such as genomic copy number variants (CNVs) using comparative genomic hybridization (CGH). Sample DNA from a reference and a test are hybridized to either a single microarray using different labels, or to two different microarrays. A profile is subsequently created to identify genomic loci that have increased, decreased or identical copy number abundance in the test sample relative to the reference. Array CGH was instrumental in the discovery of CNVs in the human genome (Iafrate et al. 2004; Sebat et al. 2004; Redon et al. 2006).

1.5 Next-generation sequencing

The bulk of innovation in DNA sequencing technology was motivated by the HGP. While one of the primary goals of the HGP was technology development, each innovation did not revolutionize DNA sequencing. It was not until 2004, when the National Human Genome Research Institute announced the goal of the $1000 genome that breakthrough technologies were introduced. A key focus for advanced sequencers was to increase throughput to increase the quantity of sequence data output per run.

Dideoxy sequencing, using Sanger's sequencing by synthesis biochemistry, is generally considered to be the first generation of sequencing technology. These systems typically consist of 96 or 384 reactions analyzed simultaneously via an array of gel-filled capillaries. The second generation of sequencers, referred to as "next-generation sequencers" are based on the sequencing-by-synthesis and sequencing-by-ligation paradigms. This refers to the polymerase or ligase driven extension of primed templates for data acquisition (Shendure and Ji 2008). To achieve higher throughput data acquisition, bases are resolved by "cyclic array sequencing" (Shendure et al. 2008). A crucial distinction between first and second generation sequencing is the library preparation. Instead of cloning DNA fragments into bacteria and isolating the subsequent DNA, second generation methods randomly fragment DNA and ligate the genomic

fragments *in vitro* to synthetic DNA sequences called adapters. By avoiding the requirement to amplify DNA in bacteria, complications with uncloneable fragments are avoided.

All cyclic array sequencing platforms simultaneously resolve an array with millions to billions of distinct sequence features and are referred to as "massively parallel" sequencers. The individual features are clonal such that each resolvable unit contains identical copies of DNA (a genomic fragment ligated to an adapter) that has been amplified to form ~1000 copies. The features are immobilized onto an array surface so that reagents can be applied to all features in parallel. For each cycle of the system, an enzymatic step (polymerization or ligation) interrogates the identity of a single nucleotide position for each feature. The enzymatic step emits light, and a camera records images of the array for each successive nucleotide incorporation event. At the end of each cycle, sequence reads are generated by a computational base-calling analysis of the image series.

There are three principle commercial cyclic array sequencing platforms: Roche 454, Illumina and Life Technologies SOLiD. These differ from one another by having distinct biochemistry at the cyclic step and by their unique feature generation methods. All three use polymerase-based amplification of a library of sequence templates to form locally clustered features sometimes referred to as polymerase colonies (Mitra and Church 1999). Due to the error-prone activity of DNA polymerases under these conditions, this amplification step is responsible for the majority of sequencing errors by these systems (Mardis 2011). For a comparative summary of these sequencers, see the *Field guide to next-generation DNA sequencers* (Glenn 2011).

The 454 sequencer was the first commercially available next-generation sequencing platform, released in 2005, and it uses a technique called pyrosequencing (Ronaghi 2001; Margulies et al. 2005). Library preparation requires short DNA fragments to be ligated to adapters that are bound to 26μm beads such that each bead is bound to a single species of DNA. Clonal features are generated on individual beads by emulsion PCR (Dressman et al. 2003). The emulsion is then broken and the solution is enriched for amplicon-bound beads. A sequencing primer is annealed to the common priming site on the adapters, polymerase is incubated with the beads and the solution is deposited onto an array of picoliter-scale 29μm wells accommodating one bead per well. Smaller beads coated in sulfurylase and luciferase are flooded into the wells, to detect released pyrophosphate. The key step in pyrosequencing occurs when a single nucleotide is

incorporated onto the template which results in pyrophosphate release leading to luciferase-based light emission (Ronaghi et al. 1996). Each of the four nucleotides are flowed across the microwell array one at a time and a camera coupled to a computer analyzes the emitted light to determine sequence reads. Pyrosequencing uses unmodified nucleotides which do not contain a terminating moiety. Accordingly, when consecutive runs of a single base are encountered (homopolymers such as CCC or AAAA), multiple nucleotides are incorporated in series, and the number of incorporation events must be determined from increases in light intensity. Consequently, the main error made by 454 sequencers are typically insertions or deletions at homopolymeric sites rather than substitutions. At the time of writing, a key feature of the 454 system is that it offers read lengths of up to 1kbp.

454 features paired-end sequencing, which outputs read information from both ends of genomic inserts. Paired-end architecture enables researchers to know the relative position of two reads to one another, a useful factor for algorithms when aligning reads to genomes with repeat regions. As well, it orients contigs during short-read-based *de novo* genome assembly.

Illumina sequencing is based on a technology called reversible termination (Fedurco et al. 2006; Turcatti et al. 2008). As with 454, libraries are prepared by first ligating adapters onto genomic fragments, but rather than binding these to beads, the library of single molecules is hybridized to a lane within a flow cell. The flow cell is a glass slide consisting of multiple lanes which are coated in a "lawn" of adapter-complementary oligonucleotides, as described initially by Southern *et al.* (Southern et al. 1992; Maskos and Southern 1993). Library fragments are hybridized to the flow cell in a limiting manner to balance high density and optimal cluster resolution. Bound DNA molecules are converted into visible features by bridge amplification which generates clusters of ~1000 amplicon molecules (Adessi et al. 2000; Fedurco et al. 2006). The clusters are linearized to ssDNA and a sequencing primer is annealed to the universal priming site on the adapters. During each Illumina cycle, all four nucleotides are flowed over the clusters simultaneously and a single incorporation event occurs. These nucleotides, the reversible terminators, have a 3' hydroxyl moiety that permits only a single incorporation event on the template. This is analogous to Sanger sequencing, except that the Illumina terminators can be cleaved so that termination is reversible. Each of the four reversible terminators has a distinct fluorescent label.

The incorporated bases are interrogated across the entire flow cell, and the 3' hydroxyl group is cleaved to allow further polymerase-based extension. The main errors made by Illumina sequencers are typically substitutions rather than insertions or deletions. At the time of writing, Illumina systems offer read lengths of up to 250bp in paired or unpaired formats.

SOLiD (Sequencing by Oligonucleotide Ligation and Detection) cyclic array sequencing systems use DNA ligase to interrogate base positions (Shendure et al. 2005; McKernan et al. 2006). Libraries are prepared similarly to 454 sequencing by ligating adapters to genomic fragments and hybridizing these to microbeads. After emulsion PCR, template-bearing beads are covalently immobilized on a slide by 3' modification and terminal deoxytransferase (Nislow, *pers. comm.*). DNA synthesis is achieved by ligation of fluorescently labelled di-base probes to primed templates. Each di-base probe is an octamer with a unique label corresponding to the first and second bases. After ligation and imaging, the labeled portion (final three nucleotides) are cleaved, and a new probe is ligated. The process is repeated at evenly spaced intervals interrogating the sequence at those positions. Denaturation of the extended complementary fragment then resets the system, a primer with a different offset is annealed (e.g. 1bp shorter) and a new set of discontiguous bases are determined. Since each label corresponds to two nucleotides, each base position must be interrogated twice. Alignment algorithms must operate in "colour space" to determine the final read sequence, and algorithms that naively decode colour reads can fail when single colours are erroneously reported. This introduces further complexity in data analysis, which can require alignment to a "colour reference", but also yields higher confidence read data with low error rates, since each base is interrogated twice. As of 2013, SOLiD systems offer read lengths of up to 50bp in paired format or 75bp unpaired.

These second generation sequencers have been updated and refined to increase throughput and read quality, up to 100-fold since their introduction. Their foundational technology remains unchanged since platform inception. These methods achieve high-throughput detection of DNA using PCR to generate clonal features, whose fluorescence is detectable via CCD cameras. However, PCR is error-prone and is known to exhibit sequence bias in amplification, altering the relative abundance of clonal features. So-called third generation sequencers have recently been developed and they offer diverse approaches to determine DNA sequence. Some instruments

operate independently of fluorescence imaging, while others determine sequence from individual DNA molecules without the need for PCR amplification.

The Helicos single molecule sequencing system was the first sequencer to forgo PCR-based feature amplification and determine the DNA sequence of individual templates (Braslavsky et al. 2003; Harris et al. 2008). Helicos technology is similar to that of Illumina, but detects fluorescence without the use of bridge-amplified clusters. The workflow begins with DNA fragmentation followed by ligation to poly(A) oligonucleotide adapters. These fragments are hybridized to poly(T) oligonucleotides that are immobilized on a flow cell. A polymerase then incorporates a labelled "virtual terminator" nucleotide containing a fluorophore that prevents further template-directed extension. A laser-based microscopic imager illuminates the labels, records an image and the fluorophore is cleaved making room for the next base incorporation. These sequencers are capable of read lengths up to 55nt and average 25nt. Because single molecules are interrogated, a misincorporation or failure to incorporate a nucleotide causes a permanent loss of that reaction. Unfortunately, although this was a promising technology, delays in achieving reliable performance in a competitive field forced Helicos Biosciences to cease production of their sequencers and file for Chapter 11 in November 2012.

Another breakthrough in single molecule sequencing came from Pacific Biosciences with their real-time single molecule readout (Levene et al. 2003; Lundquist et al. 2008; Eid et al. 2009). In this workflow, template-based extension is driven by Φ29 DNA polymerase immobilized in the detection zone of a nanophotonic well structure called a zero-mode waveguide (ZMW) which provides excitation confinement on the zeptoliter scale (10^{-21}L) (Eid et al. 2009). Using confocal fluorescence microscopy, the machine measures the release of phospholinked fluorophores from labelled nucleotide incorporation by polymerase-catalyzed phosphodiester bond formation. Nucleotides are incorporated at ~6bp/s per ZMW with each cell containing 150000 ZMWs, and the system is capable of reading 75000 at a time. At the time of writing, Pacific Biosciences reports read lengths of up to 20kbp with base quality similar to Illumina output (Carneiro et al. 2012).

Another major advance was made when the Ion Torrent semiconductor-based sequencer was unveiled (Rothberg et al. 2011). While this sequencer still uses clonally amplified features,

detection of incorporation events is non-optical, greatly simplifying the apparatus and the computer software required to determine final read sequence. The workflow starts with DNA fragments ligated to adapters that are clonally amplified on 2μm acrylamide beads. Sequence primers and DNA polymerase are bound to templates on the beads and loaded onto a semiconductor chip of 3.5μm diameter wells accommodating one bead per well. Based on CMOS imaging technology, the chip detects pH changes within wells in real time as nucleotides are incorporated. The hydrolysis reaction of template-driven extension releases a hydrogen ion (H^+), changing the pH within its respective well. Each species of nucleotide is flowed in one at a time to distinguish incorporation events from one another. At the time of writing, Ion Torrent sequencers are capable of producing reads up to 400nt in length, with average reads of 200nt. Because the nucleotides are unmodified, their incorporation is very fast, but as with 454 pyrosequencing, homopolymers must be detected by relative changes in signal intensity. Currently, the Ion Torrent vendor reports the ability to distinguish up to 8bp homopolymers, which can be a source of error for this platform.

On of the most anticipated single molecule sequencing technologies is that of Oxford Nanopore. This system promises to be relatively inexpensive because it does not require an elaborate tabletop system to generate read data. Nanopore sequencing uses engineered protein channels to rapidly detect ssDNA in real time and requires no library preparation. These nanopores were initially described by measuring the passage of ssDNA molecules through *Staphylococcus aureus* α-hemolysin, a 2.6nm diameter ion channel, detected as a decrease of ionic current proportional to fragment length (Kasianowicz et al. 1996). However, DNA moved through the channel too rapidly for detection, so, after much trial and error, researchers found that variants of Φ29 DNA polymerase could slow the passage of DNA (Cherf et al. 2012). Independently, another team modified *Mycobacterium smegmatis* porin A (MspA) to facilitate DNA inflow, and with the addition of Φ29 polymerase they achieved an inflow rate of ~1bp/28ms (Derrington et al. 2010; Manrao et al. 2012). Using this system, the authors were able to demonstrate pore-based sequencing of DNA fragments up to 53bp in length, the first time DNA has been decoded by nanopores. Oxford Nanopore is using a similar system with their own proprietary motor protein to manage DNA inflow (Pennisi 2012). Oxford Nanopore claims to be able to sequence 5.4kbp of viral genomic DNA in a single pass by sequencing triplets rather than single nucleotides (Hayden

2012). A major drawback of nanopore sequencing is its high error rate of 4% which occurs due to DNA moving forward or backward an extra base while passing through the pore. However, this will likely be resolved with multiple rounds of sequencing the same molecule and its complementary strand (Pennisi 2012). While currently in alpha testing, when the nanopore sequencing system is available it may revolutionize affordable access to sequencing.

1.6 Algorithm development for massively parallel sequencing

Computational methods designed specifically for massively parallel sequencing data are used for several main tasks: base calling and sequence quality scoring, alignment of reads to a reference, *de novo* assembly, visualization or browsing and data archiving.

For the most part, massively parallel sequencing has continued using the Phred (*Q*) base-call quality scores established for Sanger sequencing (Ewing and Green 1998; Ewing et al. 1998; Richterich 1998). Nonetheless, additional metrics are required to measure technical reproducibility and errors associated with specific sequencing biochemistries (e.g. 454 and Illumina have very different error profiles for homopolymer detection) to facilitate better multi-platform analyses (Nekrutenko and Taylor 2012). The current data format standard for recording sequence reads and corresponding quality scores is the FASTQ format (Cock et al. 2010). Unmapped reads are commonly archived in public databases, primarily the NCBI Short Read Archive (SRA), which is a GenBank-like standard repository for large raw short read data sets (Wheeler et al. 2008b).

The most active developments in this field consist of updates and variations to alignment and assembly methods. The accuracy of read alignment combined with short read lengths is not optimal using traditional alignment algorithms like BLAST, SSAHA and BLAT, which were popular with capillary sequence reads (Altschul et al. 1990; Ning et al. 2001; Kent 2002). Other algorithms have been created and optimized for short read data, and are typically based on hash table or suffix tree indexing paradigms. Hash table indexing was initially used by BLAST which used a seed-and-extend algorithm (Altschul et al. 1990). Eland, one of the first short read aligners used a similar approach by indexing short reads and using spaced seeding to align them (Cox, unpublished). SOAP used a similar approach with an indexed genome instead of reads, and

MAQ used a similar approach to Eland, but allowed *k*-mismatches (Li et al. 2008a; Li et al. 2008b). These methods were soon improved upon with spaced seed methods which allowed gaps within seed sequences. As well, seed extension was greatly improved by acceleration of the Smith-Waterman algorithm with vectorization, which was implemented in algorithms such as NovoAlign (www.novocraft.com) and SHRiMP (Farrar 2007; Rumble et al. 2009). These short read aligners offered significant increases in performance and precision with the ability to align gapped reads. The next major class of short read aligners is based on suffix trees, often offering improved memory footprints using an FM-index (Ferragina and Manzini 2000). With these indices, the footprint of the entire human genome ranges from 2-8GB of memory (Li and Homer 2010). Due to this small memory footprint, the FM-index has been quite popular for short read alignment algorithms including Bowtie, Bowtie 2, SOAP2, BWA and BWA-SW (Langmead et al. 2009; Li and Durbin 2009; Li et al. 2009b; Li and Durbin 2010; Langmead and Salzberg 2012). Using short read aligners that allow for gapped alignment greatly reduces the number of false positive variant calls, as does the use of base quality information and read pairs such as paired-end and mate-pair reads (Li and Homer 2010). Specialized read aligners also exist for specific purposes; For example, TopHat aligns spliced reads derived from transcribed sequences that omit intronic and intergenic sequences (Trapnell et al. 2009).

Assembly strategies for short read technology have also varied significantly from early contig assembly, driven mostly by the abundance of reads generated by massively parallel sequencers and their reduced length. Current *de novo* assembly methods are based on three different approaches: greedy, overlap-layout-consensus (OLC) and de Bruijn graph strategies (Miller et al. 2010). The greedy graph strategy iteratively merges pairs of reads with the greatest overlaps to extend contigs and is applied in assemblers like SSAKE (Warren et al. 2007). However, it does not necessarily find the optimal solution and can depend on the merging order of equally scored pairs. The OLC strategy involves three steps: construct an overlap graph of the reads (pairwise comparisons of all reads), identify the optimal paths traversing the graph (called a Hamiltonian circuit) and determine the consensus sequence by multiple sequence alignment. This approach is used in assemblers such as Newbler, specifically designed to handle 454 homopolymer length ambiguity, and Shorty, which estimates the distance between contigs by using seeds that are 300-500bp long (Margulies et al. 2005; Hossain et al. 2009). OLC methods are computationally

intensive because they involve all-by-all pairwise sequence comparisons and must find a Hamiltonian circuit, which can only effectively be accomplished by approximation (Paszkiewicz and Studholme 2010). These problems are avoided by using de Bruijn graphs which eliminate the all-by-all comparisons. Instead, the reads are broken into smaller DNA words and used to construct a de Bruijn graph, where the optimal path is simpler to identify. Currently, the most popular assemblers use de Bruijn graph strategies and include ABySS, ALLPATHS, SOAPdenovo and Velvet (Butler et al. 2008; Zerbino and Birney 2008; Simpson et al. 2009; Li et al. 2010b).

Mapped read data is most commonly stored in Sequence Alignment/Map (SAM) format or the compressed equivalent, BAM format (Li et al. 2009a). These files can be passed to other software packages to call genetic variants, such as SAMTools or the Genome Analysis Toolkit (GATK), which realign reads to accurately call single nucleotide variants (SNVs) and short insertions or deletions (indels) (Li et al. 2009a; McKenna et al. 2010). Software also exists to comparatively use sequencing data to determine copy number variants (CNVs) (Xie and Tammi 2009). These mapped data can also be visualized using genome browsers such as the Integrative Genomics Viewer (IGV) and the Savant Genome Browser which display coverage and variant information across samples (Fiume et al. 2010; Robinson et al. 2011; Thorvaldsdottir et al. 2012). Also, graphics can be generated to display genome-wide trends using tools such as the popular Circos circular layout visualization software (Krzywinski et al. 2009). For an overview of many bioinformatic tools, consult *The Elements of Bioinformatics* (http://elements.eaglegenomics.com/).

1.7 Applications of massively parallel sequencing

Many of the applications of massively parallel sequencing are derivations of microarray applications albeit with higher throughput and more accurate (i.e. single base pair) observations. Next-generation sequencers were, however, primarily designed for the purpose of genome sequencing. To this end, a key application of massively parallel sequencing is *de novo* genome assembly. As described earlier, current sequencing platforms are capable of generating an abundance of short reads that can be used in single read or paired read architectures to assemble whole organismal genomes. Since the early sequencers produced reads that were more abundant

but significantly shorter than Sanger sequencing reads, the first massively parallel sequenced genomes were of megabase-scale bacterial genomes (Margulies et al. 2005). As throughput increased, eukaryotic genomes were assembled as well, including the 2.46Gb genome of the giant panda (Li et al. 2010a; Nowrousian et al. 2010).

The assembly of genomes is a crucial first step in genome sequence analysis, but once the reference genomes are established, the task of genome sequence determination becomes much simpler and is known as genome resequencing. This analysis is typically performed to identify genomic variation within a population or patient cohort. Alignment of reads to a reference genome is much simpler than assembly and was accomplished earlier than next-generation assembly using existing Sanger sequencing-based reference assemblies (Bentley et al. 2008; Wang et al. 2008; Wheeler et al. 2008a; McKernan et al. 2009; Pushkarev et al. 2009). Nonetheless, as alignment to references becomes more routine researchers are seeking associations with rare variants.

RNA sequencing (RNA-seq) is a major sequencing application that is has roots in gene expression microarrays. It allows researchers to sequence cDNAs at great depth to determine both transcript abundance and transcript boundaries. It offers much higher depth and throughput than EST sequencing, and in contrast to microarrays it offers improved sensitivity, dynamic range and limits of detection with less background. In addition, microarrays can only detect transcripts if corresponding probes have been designed, while RNA-seq requires no prior knowledge or annotation. The first RNA-seq studies were performed in *S. cerevisiae* and mammalian cell lines, but the method is replacing gene expression quantitation by microarrays (Cloonan et al. 2008; Lister et al. 2008; Mortazavi et al. 2008; Nagalakshmi et al. 2008; Wilhelm et al. 2008).

Chromatin immunoprecipitation methods have also benefitted from the transition from microarray-based ChIP-chip to the sequencing-based ChIP-seq. ChIP-seq works by crosslinking DNA-binding proteins to DNA, immunoprecipitating them, de-crosslinking and sequencing the DNA fragments. The sequencing adaptation of this protocol delivers an increase in coverage and simplicity, since a microarray must cover the entire genome, or all genomic sequences, to effectively perform ChIP experiments. ChIP-seq was first performed with *S. cerevisiae* and

mammalian cell lines (Barski et al. 2007; Johnson et al. 2007; Mikkelsen et al. 2007; Robertson et al. 2007). A variant of ChIP-seq, sometimes called MNase-seq, is a widely-used method to determine nucleosome occupancy in a genome of interest. MNase-seq crosslinks proteins to DNA and subsequently preferentially digests unbound DNA using micrococcal nuclease (MNase). The remaining fragments represent histone-bound genomic DNA (Albert et al. 2007).

Another application of massively parallel sequencing adapted from microarrays is the counting of molecular barcodes by sequencing (Bar-seq). Sequencing circumvents the need for custom barcode microarrays, which are costly to design and manufacture, and it offers numeric barcode counts rather than relative abundance data calculated from hybridization intensity. This was first demonstrated using the yeast deletion collection in chemical genomic assays (Smith et al. 2009).

At the time of writing, it is still prohibitively expensive for many clinical studies to perform whole genome resequencing of multiple human individuals. Since the coding portion of the human genome accounts for under 2% of the genomic DNA, a targeted sequencing strategy capturing only the exons called exome sequencing has become popular. Exome-seq uses capture probes, similar to microarray oligonucleotide probes, but longer, to specifically isolate and sequence the coding regions of the human genome. This has proved to be an effective strategy to discover variants underlying Mendelian diseases or phenotypes. Exome-seq is now widely-used and has identified mutations associated with brain malformations and Kabuki syndrome as well as clinical diagnoses (Choi et al. 2009; Ng et al. 2009; Bilguvar et al. 2010; Ng et al. 2010; Worthey et al. 2011).

In addition to these popular methods, there are many other custom applications for massively parallel sequencing. Recently, synthesized DNA has been used to store binary data in an information dense manner exceeding any other available technology (Church et al. 2012). Due to the incredibly high-throughput of current technologies, if a biological question can be answered via sequencing-based experimentation, application of the technology typically yields the most cost-effective results (Lander 2012).

1.8 Thesis objectives

My PhD study aims to develop new ways to apply genome-scale technologies to biological problems. In this thesis, I will describe three major studies where I used some of the aforementioned techniques to answer questions in molecular biology. My first data chapter uses genome resequencing of antifungal-resistant yeast strains to identify genetic factors underlying the development of drug resistance. This analysis involved the use of massively parallel genomic DNA sequencing to identify novel genomic variation. In my second data chapter I report the discovery of conserved chromatin architecture in a non-eukaryotic species, an archaeon. This study made use of several massively parallel sequencing assays including MNase-seq and RNA-seq. In my final data chapter, I describe a method I developed to identify human drug targets by overexpression of human genes in yeast. At the time, DNA microarrays were best suited to this problem, and I aided in the design of an array for this purpose.

My thesis research applies DNA microarray and sequencing technologies to develop novel methods for drug target identification and to study the conservation of chromatin structure.

2 Genetic and Genomic Architecture of the Evolution of Resistance to Antifungal Drug Combinations

Fungal infections are a major source of human morbidity and mortality, a phenomenon that is not widely recognized. These infections are difficult to treat partly due to the limited number of antifungal drugs, whose effectiveness is compromised by the emergence of drug resistance. Combination therapy has emerged as a powerful strategy to abolish this resistance, but the impact of drug combinations on the evolution of resistance remains largely unexplored. Here we provide the first analysis of the genetic and genomic changes that underpin the evolution of resistance to antifungal drug combinations in the leading human fungal pathogen, *Candida albicans*, and model yeast, *Saccharomyces cerevisiae*. Since inhibiting the molecular chaperone Hsp90 or its downstream effector calcineurin inhibits fungal stress responses, experimental populations were evolved in the presence of combinations including inhibitors of Hsp90 or calcineurin and the most widely used antifungal in the clinic, the azoles, which inhibit ergosterol biosynthesis. Using whole genome sequencing, diverse resistance mutations were identified among the 14 of 290 lineages that evolved resistance to the drug combination. These included mutations in genes encoding the drug targets, a transcriptional regulator of multidrug transporters, a transcriptional repressor of ergosterol biosynthesis enzymes and a regulator of sphingolipid biosynthesis. Aneuploidies in several *C. albicans* lineages were also found. This study reveals multiple mechanisms by which resistance to drug combination can evolve, suggesting novel strategies to combat drug resistance.

Portions of this chapter have been adapted from the following manuscript:

Hill, J. A., Ammar, R., Torti, D., Nislow, C. & Cowen, L. E. Genetic and Genomic Architecture of the Evolution of Resistance to Antifungal Drug Combinations. 2013. *PLoS Genetics*. 2013. 9(4): e1003390.

Author contributions:

JAH: Conception and design, prepared biological material, performed molecular analyses, interpretation of results, drafting and revising the article

RA: Conception and design, acquisition of data, performed computational analyses, interpretation of results, drafting and revising article

DT: Performed the sequencing

CN: Conception and design, interpretation of results, drafting and revising the article

LEC: Directed the research, conception and design, interpretation of results, drafting and revising the article

2.1 Pathogenic fungi, antifungal drugs and resistance mechanisms

Mycoses have a dramatically negative effect on human health affecting billions annually, with mortality rates exceeding 50% for invasive fungal infections (Brown et al. 2012a). Yet, the disease burden of these pathogenic fungi is widely unappreciated because they often become life threatening in immunocompromised individuals who are already burdened by other diseases. These include AIDS patients and organ transplant patients treated with immunosuppressive therapy to prevent organ rejection. Even with current treatment options, fungal pathogens kill as many people as tuberculosis or malaria (Pfaller and Diekema 2010; Brown et al. 2012b).

In general, fungal infections can be subdivided into two major classes: superficial and invasive (Brown et al. 2012a). Superficial infections are very common, often occurring when dermatophytes colonize the skin or nails. These also encompass mucosal infections of the oral and genital tracts, which are very common but seldom life-threatening, such as thrush or vulvovaginitis (Sobel 2007). Invasive infections are much less common as the immune system of healthy individuals is capable of prevention, although these infections are associated with high mortality. The majority of all mortality due to invasive fungal infections is due to infection with species from the genera *Aspergillus*, *Candida*, *Cryptococcus* and *Pneumocystis* (Shapiro et al. 2011; Brown et al. 2012a). *Cryptococcus* species are typically found in soil or on trees and infect individuals who inhale the fungi, after which it can spread within immunocompromised individuals. *Cryptococcus* frequently infects the central nervous system leading to fungal meningoencephalitis and is most commonly found in AIDS patients (Sionov et al. 2010). Both *Aspergillus* and *Pneumocystis* species are airborne opportunistic pathogens that can cause lethal respiratory infections in immunocompromised individuals. *Candida* species are the most common commensal fungi and live on the human epithelium as a natural member of the mucosal microbiota of healthy humans, but can cause life-threatening illness in immunocompromised individuals (Horn et al. 2009; Marr 2010; Pfaller and Diekema 2010). In particular, *Candida albicans*, a budding yeast, is the most pervasive fungal pathogen, the leading

cause of death due to fungal infection (Pfaller and Diekema 2007) and the fourth leading cause of hospital-acquired infectious disease (Pfaller and Diekema 2007; Pfaller and Diekema 2010).

Often fungal infections progress unimpeded due to delays in diagnostic testing and limited therapeutic options (Brown et al. 2012a). With the widespread deployment of antimicrobial agents in both clinical and environmental settings, the rate at which resistance evolves in pathogen populations outpaces the rate of drug development (Bush et al. 2011; Chopra 2012). The evolution of resistance to antifungal drugs is of particular concern given the increasing incidence of life-threatening invasive fungal infections, and the limited number of antifungal drugs with distinct targets (Shapiro et al. 2011). Unlike for antibacterials, fungal-specific drug targets are limited, in part due to the close evolutionary relationships of these eukaryotic pathogens with their human hosts, rendering most treatments toxic to the host or ineffective in combating infections (Cowen 2008). Thus, there is a pressing need to develop new strategies to enhance the efficacy of antifungal drugs and to minimize the emergence of drug resistance.

A powerful strategy to extend the life of current antimicrobial agents is drug combination therapy (Torella et al. 2010). Combination therapy has the potential to minimize the evolution of drug resistance by more effectively eradicating pathogen populations and by requiring multiple mutations to confer drug resistance (zur Wiesch et al. 2011). Great success has been achieved with combination therapy in the treatment of HIV (Hogg et al. 1998; Palella et al. 1998; Nakagawa et al. 2012), and it is currently the recommended strategy for treatment of tuberculosis and malaria (WHO 2011; WHO 2012). Combination therapies have been less well explored in the clinic for fungal pathogens, however, targeting cellular regulators of fungal stress responses has emerged as a promising strategy to enhance the efficacy of antifungal drugs and to abrogate drug resistance (Steinbach et al. 2007; Cowen 2008). Two key cellular regulators that are critical for orchestrating cellular responses to drug-induced stress are Hsp90 and calcineurin. The molecular chaperone Hsp90 regulates the stability and function of diverse client proteins (Cowen and Lindquist 2005; Taipale et al. 2010), and controls stress responses required for drug resistance by stabilizing the protein phosphatase calcineurin (Imai and Yahara 2000; Cruz et al. 2002; Cowen and Lindquist 2005; Cowen et al. 2006; Singh et al. 2009). Compromise of Hsp90 or calcineurin function transforms antifungals from fungistatic to fungicidal and enhances the

efficacy of antifungals in mammalian models of systemic and biofilm fungal infections (Steinbach et al. 2007; Uppuluri et al. 2008; Cowen et al. 2009; Robbins et al. 2011), suggesting that combination therapy with azoles and inhibitors of Hsp90 or calcineurin may provide a powerful strategy to treat life-threatening fungal infections.

Targeting fungal stress response regulators holds particular therapeutic promise for enhancing the efficacy of the azoles, which are the class of antifungal drug that has been used most widely in the clinic for decades. Azoles block the production of ergosterol, the major sterol of fungal cell membranes, by inhibition of lanosterol demethylase, Erg11, resulting in a depletion of ergosterol and the accumulation of the toxic sterol intermediate, 14-α-methyl-3,6-diol, produced by Erg3 (Sanglard 2002). The azoles are generally fungistatic, causing inhibition of growth rather than cell death, and thus impose strong selection for resistance on the surviving fungal population (Anderson 2005). As a consequence, resistance is frequently encountered in the clinic (White et al. 1998). Azole resistance mechanisms fall into two broad classes: those that block the effect of the drug on the fungal cell and those that allow the cell to tolerate the drug by minimizing its toxicity (Cowen 2008). The former class of resistance mechanisms includes upregulation of drug efflux pumps (Balzi et al. 1994), or mutation of the azole target that prevents azole binding (Favre et al. 1999). The latter class includes loss-of-function mutations in *ERG3*, which encodes a Δ-5,6-desaturase in the ergosterol biosynthesis pathway. Erg3 loss-of-function blocks the accumulation of a toxic sterol intermediate, conferring azole resistance that is contingent on cellular stress responses (Anderson et al. 2003; Cowen and Lindquist 2005). Azole resistance acquired by loss of function of Erg3 or by many other mutations is dependent on Hsp90 and calcineurin (Cowen and Lindquist 2005). Inhibition of these stress response regulators enhances azole sensitivity of diverse clinical isolates, and compromises azole resistance of isolates that evolved resistance in a human host (Marchetti et al. 2000; Cruz et al. 2002; Cowen and Lindquist 2005; Robbins et al. 2011). Inhibition of Hsp90 or calcineurin with molecules that are well tolerated in humans can impair the evolution of azole resistance (Cowen and Lindquist 2005; Cowen et al. 2006), though the potential for evolution of resistance to the drug combinations remains unknown.

Azole resistance mechanisms have been studied most extensively in *C. albicans* and the model yeast *Saccharomyces cerevisiae*. Drug resistance can readily evolve in *C. albicans* in the laboratory

and the clinic, and molecular studies have revealed a diversity of resistance mechanisms (Pfaller 2012). Molecular studies with *C. albicans* are hindered by its obligate diploid state, lack of meiotic cycle, unusual codon usage, and inability to maintain plasmids (Berman and Sudbery 2002), thus complementary experiments are often performed with its genetically tractable relative, *S. cerevisiae*, with which it often shares drug resistance phenotypes and underlying molecular mechanisms (Cowen and Steinbach 2008). For both species, inhibition of Hsp90 or calcineurin reduces azole resistance acquired by diverse mutations (Cruz et al. 2002; Onyewu et al. 2003; Cowen and Lindquist 2005; Cowen et al. 2009). With short generation times and relatively small genomes, these organisms provide tractable and complementary systems to explore the dynamics and mechanisms underpinning the evolution of resistance to drug combinations.

In collaboration with Jessica Hill and Leah Cowen, I provide the first analysis of the genetic and genomic architecture of the evolution of resistance to drug combinations in fungi. To recapitulate a clinical context where Hsp90 or calcineurin inhibitors could be used in combination with azoles to render azole-resistant fungal pathogens responsive to treatment, we initiated an evolution experiment with strains that are resistant to azoles in a manner that depends on Hsp90 and calcineurin. Populations of *S. cerevisiae* and *C. albicans* that were resistant to azoles due to loss of function of Erg3 were evolved with a combination of an azole and geldanamycin, an inhibitor of Hsp90, or FK506, an inhibitor of calcineurin, to identify the mechanisms by which resistance evolves to the drug combinations. Of 290 lineages initiated, most went extinct, yet 14 evolved resistance. We identified mechanisms of resistance in the evolved lineages using a hypothesis-driven approach based on cross-resistance profiling and a complementary unbiased approach using whole genome sequencing. Resistance mutations in the drug target of FK506 or geldanamycin were identified and validated in five lineages. Non-synonymous substitutions conferring resistance were identified in a transcriptional activator of drug efflux pumps, Pdr1, and in a regulator of sphingolipid biosynthesis, Lcb1. Resistance also arose by premature stop codons in the catalytic subunit of calcineurin and in a repressor of ergosterol biosynthesis genes, Mot3. Several of the mutations conferred resistance to geldanamycin or FK506, while other mutations transformed azole resistance from dependent on calcineurin to independent of this stress response regulator. Genome analysis also identified

extensive aneuploidy in four of six *C. albicans* lineages. Thus, we illuminate the molecular basis for the transition of azole resistance from calcineurin dependence to independence, and establish numerous mechanisms by which resistance to drug combinations can evolve, providing a foundation for predicting and preventing the evolution of drug resistance.

2.2 Results

2.2.1 Experimental evolution yields resistance to drug combinations

Inhibition of Hsp90 or calcineurin has emerged as promising strategy to enhance the efficacy of azoles against resistant fungal pathogens, motivating this study to monitor the evolution of resistance to the drug combinations in azole-resistant populations. To do so, J. Hill used an experimental evolution approach starting with *C. albicans* and *S. cerevisiae* strains that harbour *erg3* loss-of-function mutations or deletions, rendering them resistant to azoles in a manner that depends on the stress response regulators Hsp90 and calcineurin (Cowen 2008). Propagation of these strains in the presence of azole and the Hsp90 inhibitor geldanamycin or azole and the calcineurin inhibitor FK506 at concentrations that exert selection pressure for resistance to the drug combination could lead to the evolution of resistance to geldanamycin or FK506, or the evolution of an azole resistance mechanism that is independent of Hsp90 or calcineurin among extant lineages (Fig. 2-1a). Lineages were propagated by serial transfer of between 33 and 100 generations until robust growth of extant lineages was observed in the presence of the drug combination (Fig. 2-1b). The effective population size per lineage was ~4.6 x 10^6, given that cultures reached saturation (~10^7 cells/ml) between transfers. Of the 290 lineages initiated, the majority went extinct. 14 lineages evolved resistance to the combination of azole and inhibitor of Hsp90 or calcineurin (Fig. 2-1c); seven of these lineages are *C. albicans* and seven are *S. cerevisiae* (Table 2-1). Six *C. albicans* lineages evolved resistance to azole and FK506 (Ca-F lineages), and only one evolved resistance to azole and geldanamycin (Ca-G lineage). Four *S. cerevisiae* lineages evolved resistance to azole and geldanamycin (Sc-G lineages) and three evolved resistance to azole and FK506 (Sc-F lineages).

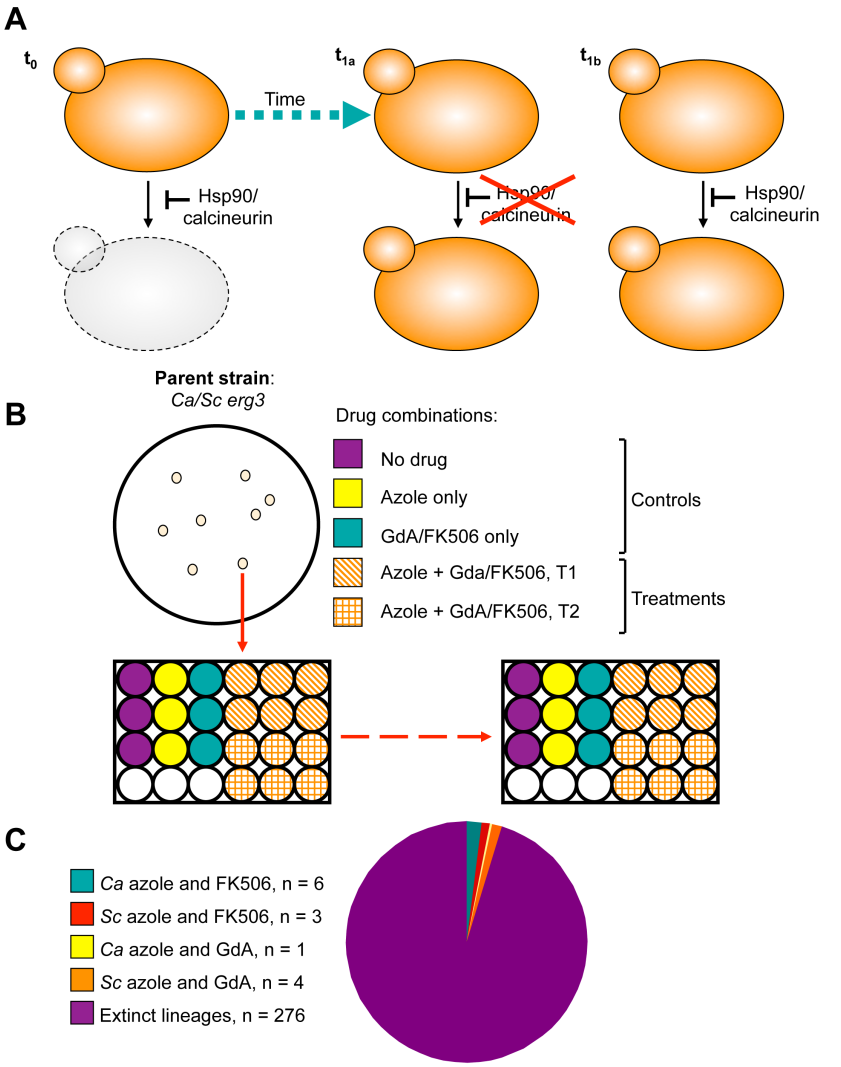


Figure 2-1. Design and outcome of the experimental evolution of resistance to drug combinations. (A) Experimental populations were initiated with *S. cerevisiae* and *C. albicans* strains resistant to azoles due to *erg3* loss of function. Resistance is contingent on Hsp90 and calcineurin, such that inhibition of either of these cellular stress response regulators results in cell death (t_0). Propagation in the presence of azole and the Hsp90 inhibitor geldanamycin or azole and the calcineurin inhibitor FK506 at concentrations that exert selection pressure for resistance to the drug combination results in the evolution of resistance to geldanamycin or FK506 (t_{1a}) or the evolution of an azole resistance mechanism that is independent of Hsp90 or calcineurin (t_{1b}) among extant lineages. **(B)** Single colony founders were used to initiate evolution experiments in 24- or 96-well plates containing control and treatment wells. Controls consisted of: no drug, azole, geldanamycin or FK506, where drug concentrations were not inhibitory. Treatment wells consisted of combinations of azole and geldanamycin or FK506, selected based on dose response matrices. **(C)** Experimental evolution of resistance to azole and geldanamycin or azole and FK506 yielded 14 resistant lineages out of 290 initiated. *Ca = Candida albicans; Sc = Saccharomyces cerevisiae*.

Strain name	*Ancestor*	*Drug combination evolved in*	*Fluconazole (FL) or miconazole (M) concentration evolved in (µg/ml)*	*FK506 or geldanamycin (GdA) concentration evolved in (µM)*	*Number of transfers (generations)*	*Number of wells in plate evolved in*
Sc-F-1	ScLC7	FL and FK506	32	2.5	13 (~ 86)	24
Sc-F-2	ScLC7	FL and FK506	64	0.03	13 (~ 86)	24
Sc-F-3	ScLC7	M and FK506	75	0.06	6 (~ 40)	24
Ca-F-4	CaLC660	FL and FK506	256	20	9 (~ 60)	96
Ca-F-5	CaLC660	FL and FK506	256	20	9 (~ 60)	96
Ca-F-6	CaLC660	FL and FK506	256	1.2	13 (~ 86)	24
Ca-F-7	CaLC660	FL and FK506	4	2	5 (~ 33)	24
Ca-F-8	CaLC660	FL and FK506	4	2	5 (~ 33)	24
Ca-F-9	CaLC660	M and FK506	64	1.2	5 (~ 33)	24
Ca-F-10	CaLC660	FL and GdA	0.1875	0.16	13 (~ 86)	24
Sc-G-11	ScLC10	FL and GdA	256	0.6	13 (~ 86)	24
Sc-G-12	ScLC7	FL and GdA	256	0.6	13 (~ 86)	24
Sc-G-13	ScLC10	FL and GdA	16	2.5	5 (~ 33)	24
Sc-G-14	ScLC10	FL and GdA	16	2.5	5 (~ 33)	24

Table 2-1. Evolution experiment treatments and conditions.

Resistance levels to the drug combinations of all fourteen evolved lineages were evaluated by performing minimum inhibitory concentration (MIC) assays in the presence of the inhibitors with which they were evolved, azole and FK506 (Fig. 2-2a,b) or azole and geldanamycin (Fig. 2-2c-e). Because the azole resistance phenotypes of the starting strains were abrogated by geldanamycin or FK506, resistance of the evolved lineages was monitored with a fixed concentration of azole and a gradient of concentrations of geldanamycin or FK506. Resistance was monitored for a population of cells from each archived lineage, and for four clones isolated from the evolved population. In all cases, the clones reflected the resistant phenotype of the population, suggestive of strong selective sweeps as mutations were rapidly fixed in the population. For each population, a clone was archived and further analyses were performed on that strain. The lineages evolved distinct levels of resistance to the drug combinations (Fig. 2-2), indicating that they acquired different mutations conferring resistance.

J. Hill assessed cross-resistance profiles to determine the mechanisms of drug combination resistance. Assays were performed in the presence of a fixed concentration of an azole and a gradient of concentrations of the structurally dissimilar counterpart to the Hsp90 or calcineurin inhibitor with which the population was evolved (native inhibitor), as well as with an azole and an inhibitor of the other stress response regulator not targeted in the evolution experiment (naïve inhibitor; i.e. Hsp90 inhibitor if the population was evolved with a calcineurin inhibitor). Cross-resistance profiles can be used to predict candidate resistance mechanisms based on an understanding of how these inhibitors bind to and inhibit their targets (Fig. 2-3).

While successful, these hypothesis driven approaches did not uncover candidate resistance mutations for the all evolved lineages. We therefore turned to whole genome sequencing to provide an unbiased approach to identify mutations that accompany the evolution of resistance to the drug combinations on a genomic scale.

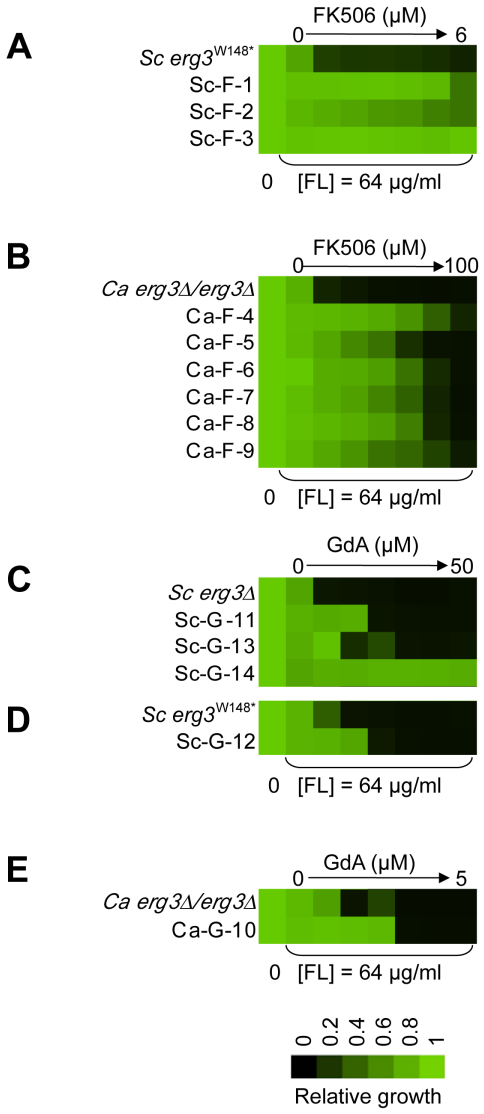
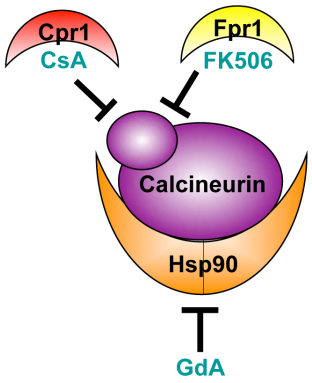


Figure 2-2. The populations evolved distinct resistance profiles. Levels of resistance to azole and FK506 **(A, B)** or azole and geldanamycin **(C – E)** of evolved strains of *S. cerevisiae* **(A, C, D)** and *C. albicans* **(B, E)**, relative to their ancestors. Resistance was measured with a constant concentration of azole and a gradient of geldanamycin or FK506 in YPD at 30˚C for 2 days **(B)** or 3 days **(A, C- E)**. Optical densities were averaged for duplicate measurements and normalized relative to drug-free controls (see colour bar). GdA = geldanamycin and FL = fluconazole.

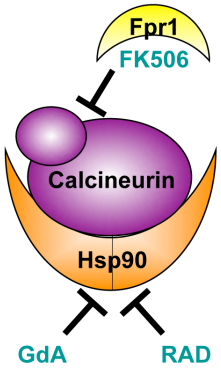
A



B

Cross-resistance profile	Candidate resistance mechanism
Azole and FK506	Mutation in *FPR1*
Azole and FK506 Azole and CsA	Mutation in calcineurin, or Calcineurin-independent resistance
Azole and FK506 Azole and CsA Azole and GdA	Up-regulation of azole efflux pumps, or Mutation in *ERG11*, or Overexpression of *ERG11*

C



D

Cross-resistance profile	Candidate resistance mechanism
Azole and GdA	Mutation in *HSP90*, or GdA-specific metabolism or efflux
Azole and GdA Azole and RAD	Mutation in *HSP90*
Azole and GdA Azole and RAD Azole and FK506	Up-regulation of azole efflux pumps, or Mutation in *ERG11*, or Overexpression of *ERG11*

Figure 2-3. Cross-resistance profiles provide a strategy to predict resistance mechanisms. (A) Strains evolved in azole and FK506 were tested for cross-resistance to azole and the calcineurin inhibitor cyclosporin A as well as azole and the Hsp90 inhibitor geldanamycin. **(B)** Candidate resistance mechanisms based on specific cross-resistance profiles of strains evolved with azole and FK506. **(C)** Strains evolved in azole and geldanamycin were tested for cross-resistance to azole and the Hsp90 inhibitor radicicol as well as azole and the calcineurin inhibitor FK506. **(D)** Candidate resistance mechanisms based on specific cross-resistance profiles of strains evolved with azole and geldanamycin. GdA = geldanamycin; RAD = radicicol; CsA = cyclosporin A; and FL = fluconazole.

2.2.2 Sequence analysis workflow

Sequence reads generated from the *S. cerevisiae* strains were aligned to the S288c genome, a high fidelity sequence derived from an individual yeast colony from F. Dietrich's lab at Duke University; It is the Saccharomyces Genome Database reference genome as of February 2011 (Engel et al. 2010). Reads from the *C. albicans* strains were aligned to the SC5314 genome from the Candida Genome Database (Skrzypek et al. 2010). While *C. albicans* is an obligate diploid, the current build of the genome, assembly 21, is a haploid genome, and is an improvement over the original diploid genome, assembly 19 (Jones et al. 2004; van het Hoog et al. 2007). The diploid assembly was not used because it features 412 supercontigs with non-obvious heterozygosity, whereas the haploid assembly has been curated and organized into 8 contiguous chromosomes (van het Hoog et al. 2007).

The sequence analysis workflow used several popular short read aligners and variant calling software to achieve whole genome resequencing of the yeast strains. Illumina single-end and paired-end reads were trimmed from the 5' and 3' ends for low quality basecalls and subsequently aligned to the reference assemblies with Bowtie 2. This aligner is capable of aligning reads of variable lengths with gaps, which yields single nucleotide variant (SNV) calls with a lower false positive rate. The alignments were output in the standard Sequence Alignment/Map (SAM) file format and processed with Picard (http://picard.sourceforge.net.) to sort and compress the SAM files into binary SAM (BAM) files and create BAM indices to quickly access the data (Li et al. 2009a). Genome coverage was computed using the Genome Analysis Toolkit (GATK), a framework for processing aligned short read data (McKenna et al. 2010; DePristo et al. 2011). Single nucleotide variants for *S. cerevisiae* were called using the UnifiedGenotyper package from the GATK and parental SNVs were subtracted from evolved strain lists to yield a list of novel SNVs. Since *C. albicans* is obligate diploid, I processed those strains with a probabilistic tool called JointSNVMix which uses paired parental and evolved strain sequence data to determine significant novel variants (Roth et al. 2012). Indels were called for both organisms with the UnifiedGenotyper. Thresholds for high confidence variants were set to minimize false positive SNV calls (see Materials and Methods). Finally, copy number variants (CNVs) were detected

using CNV-seq (Xie and Tammi 2009). All short read alignments were visualized using the Savant Genome Browser (Fiume et al. 2010).

2.2.3 Whole genome sequencing identifies candidate resistance mutations

Whole genome sequencing provided an unbiased approach to identify mutations that accompany the evolution of resistance to the drug combinations on a genomic scale. For example, *S. cerevisiae* Sc-F-1 was evolved with azole and FK506 and demonstrated robust resistance to the combination of azole and FK506 as well as azole and cyclosporin A (Fig. 2-4a). This resistance profile suggested a possible mechanism of resistance involving alteration of calcineurin that prevents the binding of both protein-drug immunophilin complexes, or the emergence of a calcineurin-independent azole resistance mechanism. Calcineurin is encoded by the redundant catalytic subunits *CNA1* and *CNA2* and the regulatory subunit *CNB1* in *S. cerevisiae* (Cyert et al. 1991; Hemenway and Heitman 1999). Sequencing of *CNA1, CNA2* and *CNB1* did not reveal any mutations. Intriguingly, abrogating calcineurin function by deletion of *CNB1* did not reduce resistance to azole and FK506 in Sc-F-1, indicating a calcineurin-independent mechanism of resistance had evolved (Fig. 2-4b). Whole genome sequencing at high coverage (Table 2-2) identified two non-synonymous mutations (Table 2-3), as well as 58 mutations that were synonymous or in non-coding regions (Table 2-4); the best candidate for a mutation for affecting resistance was a mutation in *MOT3*, a transcriptional repressor of ergosterol biosynthesis genes (Hongay et al. 2002). The non-synonymous substitution in *MOT3* resulted in a premature stop codon near the middle of the coding sequence, $MOT3^{G265*}$, suggesting that this might be a loss-of-function allele. Deletion of *MOT3* in the background of the ancestral strain or in Sc-F-1 phenocopied the level of resistance of Sc-F-1, which is consistent with $MOT3^{G265*}$ being a loss-of-function allele that confers resistance in Sc-F-1 (Fig. 2-4c).

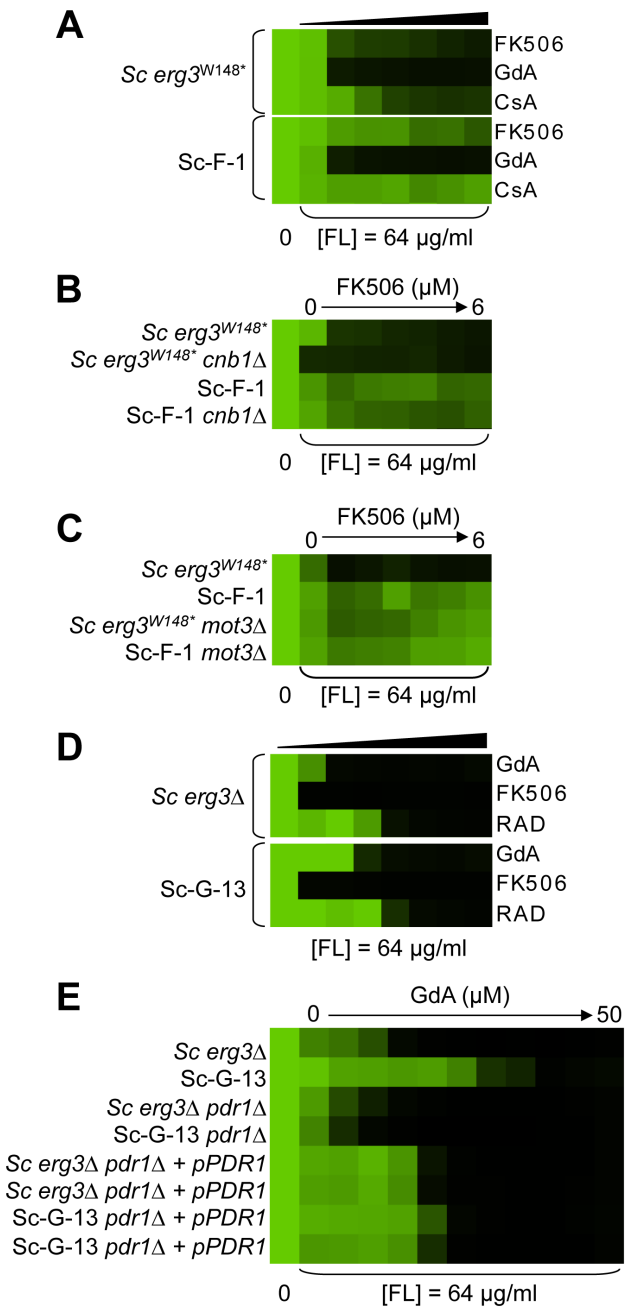


Figure 2-4. Whole genome sequencing identifies mutations that confer resistance to azole and FK506, as well as azole and geldanamycin. (A) Sc-F-1 is resistant to azole and FK506 and cross-resistant to azole and cyclosporin A. **(B)** Resistance of Sc-F-1 is calcineurin-independent. Deletion of *CNB1*, which encodes the regulatory subunit of calcineurin required for its activation does not affect resistance of Sc-F-1. **(C)** Deletion of *MOT3* in the ancestral strain confers resistance to azole and FK506 equivalent to Sc-F-1, which is consistent with the $MOT3^{G265*}$ allele of Sc-F-1 conferring resistance to azole and FK506. **(D)** Sc-G-13 is slightly resistant to azole and geldanamycin. **(E)** Resistance to azole and geldanamycin in Sc-G-13 is reduced when PDR^{R865P} is deleted and *PDR1* is expressed on a plasmid. Resistance assays were performed and analyzed as in Figure 2-2, with incubation for 2 days at 30˚C in YPD **(A – D)** or SD **(E)**. CsA = cyclosporin A; GdA = geldanamycin; RAD = radicicol; and FL = fluconazole.

S. cerevisiae lineage Sc-G-13 was evolved with azole and geldanamycin and demonstrates only a small increase in resistance to this combination, with no cross-resistance to either azole and FK506 or azole and radicicol (Fig. 2-4d). This resistance profile is consistent with a mutation in *HSC82* or *HSP82* that partially reduces binding of geldanamycin, however, no mutations were identified upon sequencing *HSC82* and *HSP82*. Genome sequencing of Sc-G-13 identified five non-synonymous mutations, as well as 130 that were synonymous or in non-coding regions (Table 2-3); the best candidate for a mutation affecting resistance was a C2593G mutation in *PDR1*, which encodes a transcription factor that regulates the expression of numerous multidrug transporters such as *PDR5*. Gain-of-function mutations in *PDR1* are a well-established mechanism of azole resistance that is independent of Hsp90 and calcineurin (Kolaczkowska and Goffeau 1999; Anderson et al. 2003; Cowen and Lindquist 2005). The mild resistance phenotype of Sc-G-13 suggested that the $PDR1^{P865R}$ allele in Sc-G-13 confers only a slight increase in drug efflux pump expression. Cross-resistance to azole and FK506 was not observed, likely because FK506 inhibits Pdr5-mediated efflux (Hendrych et al. 2009). To evaluate the importance of the $PDR1^{P865R}$ allele in resistance to azole and geldanamycin, *PDR1* was deleted from the ancestral strain and the evolved Sc-G-13 lineage and the ancestral *PDR1* allele was introduced on a plasmid driven by the *GPD1* promoter. Replacing the $PDR1^{P865R}$ allele of Sc-G-13 with the ancestral *PDR1* allele reduced resistance of Sc-G-13 (Fig. 2-4e). Resistance remained slightly increased relative to the ancestral strain, likely due to higher expression of *PDR1* from the *GPD1* promoter relative to the native promoter; consistent with this possibility, simply replacing the ancestral *PDR1* allele in the ancestor with the same allele on the plasmid conferred a small increase in resistance (Fig. 2-4e). Since there was no difference in resistance phenotype between the ancestral and evolved strains when the plasmid provided the only allele of *PDR1*, there are likely no other mutations conferring resistance in Sc-G-13.

Strain	*Mean Coverage*
Sc-F-1	102
Sc-G-13	71
Ca-F-4	75
Ca-F-5	176
Ca-F-6	163
Ca-F-7	179
Ca-F-8	165
Ca-F-9	204

Table 2-2. Mean coverage whole-genome sequenced strains.

Strain	*Gene ID*	*Gene Name*	*GO Biological Process*	*Nucleotide Change*	*Non-synonymous Change*
Sc-F-1	YMR070W	MOT3	cellular hyperosmotic response; negative regulation of ergosterol biosynthetic process; negative regulation of transcription from RNA polymerase II reporter; positive regulation of transcription from RNA polymerase II promoter	C792T	Q265*
Sc-F-1	YBL096C		unknown	T140G	N47K
Sc-G-13	YGL013C	PDR1	positive regulation of cellular response to drug; positive regulation of transcription from RNA polymerase II promoter	C2593G	P865R
Sc-G-13	YLR162W-A	RRT15	unknown	T81C	S28P
Sc-G-13	YGR090W	UTP22	maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA); rRNA processing; tRNA export from nucleus	G3648A	E1217K
Sc-G-13	YJR035W	RAD26	nucleotide-excision repair; transcription-coupled nucleotide-excision repair	G1590T	V531F
Sc-G-13	YFL017W-A	SMX2	mRNA splicing, via spliceosome	A148C	D50A

Table 2-3. Non-synonymous *S. cerevisiae* single nucleotide variants.

For the six *C. albicans* lineages evolved with fluconazole and FK506 (Ca-F-4, Ca-F-5, Ca-F-6, Ca-F-7, Ca-F-8, and Ca-F-9), candidate resistance mutations were not identified by hypotheses-based cross-resistance profiles. These lineages shared the same cross-resistance profile of resistance to high concentrations of FK506 and increased resistance to cyclosporin A in the presence of azole (Fig. 2-5). This profile suggested that either a mutation in calcineurin preventing binding of both drug-immunophilin complexes occurred or a calcineurin-independent mechanism of resistance to azoles evolved. We sequenced the genomes of all six lineages of this resistance class.

Genome analysis revealed aneuploidies in four of these evolved lineages. For Ca-F-4, I identified extensive aneuploidies in the absence of any non-synonymous mutations (Fig. 2-6). This lineage exhibited increased copy number of chromosomes 4, 6 and 7 as well as an increase in copy number of the right arm of chromosome 5. Since approximately half the genome of Ca-F-4 had elevated copy number, resistance might be conferred by a combination of mechanisms including overexpression of the many relevant genes that were amplified including the gene encoding the drug transporter Mdr1, genes encoding ergosterol biosynthetic enzymes, the gene encoding the calcineurin regulatory subunit *CNB1*, or those encoding regulators of many other cellular pathways. I also identified increased copy number of chromosome 4 in three of the lineages, Ca-F-5, Ca-F-6 and Ca-F-7, as observed in Ca-F-4 (Fig. 2-6). Ca-F-5 also had an increased copy number of chromosome 7. The remaining two lineages, Ca-F-8 and Ca-F-9, had no copy number variation other than variation in chromosome R, which was observed in all of the *C. albicans* lineages sequenced. Chromosome R contains the genes coding for rDNA, and extensive variation in size of the rDNA array has been observed in experimental populations of *C. albicans* (Cowen et al. 2000), likely as a consequence of the highly repetitive nature of the genomic context.

Two non-synonymous mutations were identified in *C. albicans* lineage Ca-F-9 (Table 2-5), and 7 mutations that were synonymous or in non-coding regions. The best candidate for a resistance mutation is the C1201A mutation in *CNA1*, the gene encoding the catalytic subunit of calcineurin; this mutation leads to a premature stop codon, S401*. Truncation of *C. albicans* Cna1 at position 499 removes the autoinhibitory domain, resulting in a constitutively activated form of calcineurin (Sanglard et al. 2003).

Strain	*# of SNVs*
Sc-F-1	60
Sc-G-13	135
Ca-F-4	169
Ca-F-5	23
Ca-F-6	23
Ca-F-7	20
Ca-F-8	21
Ca-F-9	9

Table 2-4. Number of high confidence single nucleotide variants (SNVs) (coding and non-coding).

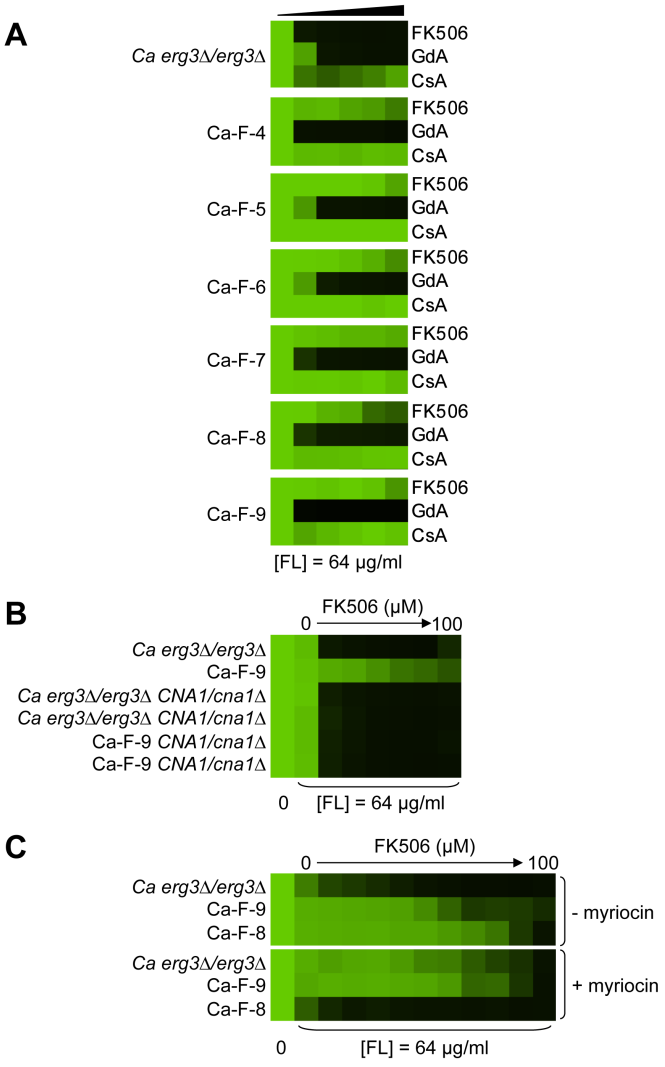


Figure 2-5. Six *C. albicans* lineages evolved with azole and FK506 share the same cross-resistance profile, and a mutation in *CNA1* and *LCB1* confer resistance. (A) Each *C. albicans* lineage is resistant to high concentrations of FK506 or cyclosporin A in the presence of azole. **(B)** *CNA1*^{S401*} confers resistance to azole and calcineurin inhibitors in Ca-F-9. The C1201A mutation in *CNA1*, the gene encoding the catalytic subunit of calcineurin leads to a premature stop codon, and removal of the autoinhibitory domain. Deletion of *CNA1*^{S401*} in Ca-F-9 abrogates resistance, while deletion of one allele of *CNA1* in the parental strain has no impact on sensitivity. **(C)** The A1169T mutation in orf19.6438 resulting in non-synonymous substitution (L390F) in this ortholog of *S. cerevisiae LCB1* likely confers resistance in Ca-F-8. Lcb1 encodes a component of serine palmitoyltransferase that is responsible for the first committed step in sphingolipid biosynthesis, along with Lcb2. Inhibition of Lcb1 and Lcb2 with myriocin (900 nM) abrogates resistance of Ca-F-8. Resistance assays were performed and analyzed is in Figure 2-2, with incubation for 2 days at 30˚C in YPD. GdA = geldanamycin; CsA = cyclosporin A; and FL = fluconazole.

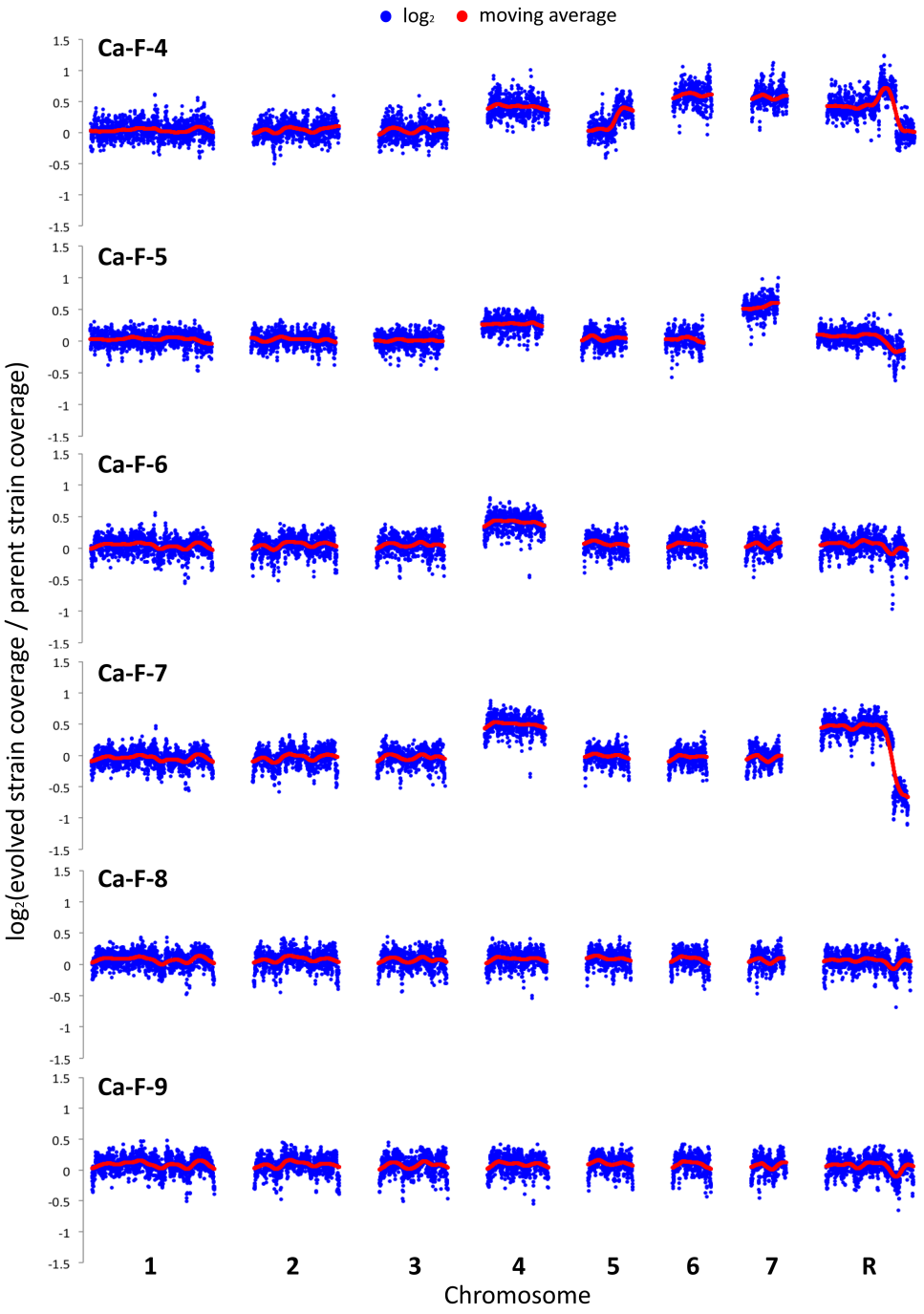


Figure 2-6. Aneuploidies identified in four *C. albicans* lineages that evolved resistance to the combination of azoles and calcineurin inhibitors. The genomes of six evolved strains were sequenced and profiled for copy number variants using CNV-Seq. Four of the strains contain aneuploidies: Ca-F-4, Ca-F-5, Ca-F-6, and Ca-F-7. Notably, chromosome 4 is increased in copy number in all four strains, suggesting that a locus on this chromosome is related to the mechanism of resistance. Blue = $\log_2$ values; Red = moving average values.

Strain	*Gene ID*	*Gene Name*	*GO Biological Process*	*Nucleotide Change*	*Non-synonymous Change*
Ca-F-6	orf19.3041		unknown	G376T	G126V
Ca-F-6	orf19.5015	*MYO2*	actin cytoskeleton reorganization; cell growth mode switching, budding to filamentous; cell morphogenesis; cellular response to heat; establishment of nucleus localization; establishment or maintenance of cell polarity; filamentous growth; filamentous growth of a population of unicellular organisms in response to heat; nucleus organization	G2404C	R802T
Ca-F-7	orf19.3041		unknown	G376T	G126V
Ca-F-8	orf19.2929	*GSC1*	(1,3)-β-D glucan biosynthetic process; fungal-type cell wall organization; pathogenesis	C1152A	H385N
Ca-F-8	orf19.3041		unknown	G376T	G126V
Ca-F-8	orf19.1263	*CFL1*	copper iron import; iron ion transport	G1591T	R531I
Ca-F-8	orf19.6131	*KSR1*	sphingolipid biosynthetic process	A21T	I8F
Ca-F-8	orf19.6438		predicted: sphingolipid biosynthetic process	A1169T	L390F
Ca-G-10	orf19.2174	*RAD57*	predicted: heteroduplex formation; meiotic DNA recombinase activity; reciprocal meiotic recombination; telomere maintenance via recombination	T341G	C114W
Ca-G-10	orf19.4337		monocarboxylic acid transport	C1627A	S543Y
Ca-G-10	orf19.6515	*HSP90*	cellular response to drug; cellular response to heat; filamentous growth; filamentous growth of a population of unicellular organisms; intracellular steroid hormone receptor signaling pathway; negative regulation of filamentous growth of a population of unicellular organisms; pathogenesis; protein folding; regulation of apoptotic process	G270T	D91Y
Ca-G-10	orf19.6693		predicted: proteolysis	A3306C	I1103L
Ca-F-9	orf19.6033	*CNA1*	cellular response to biotic stimulus; cellular response to cation stress; cellular response to starvation; filamentous growth; filamentous growth of a population of unicellular organisms; filamentous growth of a population of unicellular organisms in response to biotic stimulus; filamentous growth of a population of unicellular organisms in response to starvation; fungal-type cell wall organization; hyphal growth; pathogenesis; regulation of apoptotic process	C1201A	S401*
Ca-F-9	orf19.3041		unknown	G376T	G126V

Table 2-5. Non-synonymous *C. albicans* single nucleotide variants.

Consistent with this mutation conferring resistance to the combination of azole and FK506 or azole and cyclosporin A, deletion of the evolved *CNA1* allele in Ca-F-9 abrogates resistance to the combination of azole and calcineurin inhibitor (Fig. 2-5b). Deletion of one allele of *CNA1* in the ancestral strain has no effect on sensitivity to the drug combination. Thus, hyperactivation of calcineurin provides a mechanism by which resistance to azoles and calcineurin inhibitors can evolve.

Five non-synonymous mutations were identified in the *C. albicans* lineage Ca-F-8 (Table 2-5), and 16 mutations that were synonymous or in non-coding regions. The best candidate for a resistance mutation is the A1169T mutation identified in orf19.6438 resulting in non-synonymous substitution, L390F. orf19.6438 remains uncharacterized in *C. albicans* but is an ortholog of *S. cerevisiae LCB1*, which encodes a component of serine palmitoyltransferase that is responsible for the first committed step in sphingolipid biosynthesis, along with Lcb2 (Buede et al. 1991). Sphingolipids are a necessary component of the fungal cell membrane and have known interactions with ergosterol (Dickson and Lester 2002), while inhibitors of sphingolipid biosynthesis can enhance the efficacy of azoles (Spitzer et al. 2011). To test the model that *LCB1*^{L390F} confers resistance to the combination of azole and calcineurin inhibitor, J. Hill used the serine palmitoyltransferase inhibitor myriocin, which inhibits Lcb1 and Lcb2 (Chen et al. 1999). Inhibition of Lcb1 with myriocin abrogated resistance to azole and FK506 of the evolved lineage Ca-F-8 but did not affect resistance of Ca-F-9 (Fig. 2-5c), suggesting that *LCB1*^{L390F} confers resistance to the drug combination. Notably, myriocin caused an increase in resistance of the ancestral strain to azole and FK506 suggesting that resistance phenotypes are exquisitely sensitive to the balance of sphingolipid biosynthesis.

2.3 Discussion

This study provides the first experimental analysis of the evolution of resistance to drug combinations in fungi, illuminating the molecular basis of a transition of drug resistance from dependence on a key stress response regulator to independence, and a diversity of resistance mechanisms that can evolve in response to selection. This work addresses some of the most fundamental questions about the nature of adaptation. One key question is how many mutations

underlie adaptive evolution. For all of the lineages for which we functionally tested the importance of mutations identified, we found that a single mutation was responsible for adaptation, in contrast to other experimental evolution studies with *S. cerevisiae* where multiple adaptive mutations were implicated (Anderson et al. 2010; Dettman et al. 2012). The small number of adaptive mutations identified in our study may reflect the short duration of the evolution experiment and the strength of the selection. Despite the limited number of adaptive mutations, we identified a larger number of total mutations in many lineages than reported in other studies (Dettman et al. 2012). The elevated number of mutations may be specific to the intense drug selection pressure, just as bacterial mutation rates can increase in the presence of antibiotic selection (Blazquez 2003), and antifungals have been associated with the rapid appearance of aneuploidies and genomic instability (Selmecki et al. 2009b). Another central question is how many genetic routes are there to adaptation. Among only 14 evolved lineages, we identified a diversity of adaptive mechanisms including target-based resistance to Hsp90 or calcineurin inhibitors and distinct mutations that render azole-resistance independent of cellular stress response regulators, suggesting a complex adaptive landscape with multiple genotypes leading to high fitness adaptive peaks. Exploring the impact of the adaptive mutations on fitness in different environments, including in the absence of drug, will be key to understanding fitness costs of drug resistance, evolutionary trade-offs, and the limits of adaptation.

By starting the evolution experiment with strains that are resistant to azoles in a manner that depends on Hsp90 and calcineurin, we provide relevance for a clinical context where Hsp90 and calcineurin inhibitors could be deployed in combination with azoles to render azole-resistant isolates responsive to treatment. There is some precedent for the evolution of resistance to these drug combinations, as clinical isolates recovered from an HIV-infected patient over the course of two years evolved increased resistance to the combination of azole and inhibitors of Hsp90 or calcineurin (Cowen and Lindquist 2005). While this patient was not treated with Hsp90 or calcineurin inhibitors, fever may have provided the selection for Hsp90 independence given that febrile temperatures cause global problems in protein folding that can overwhelm Hsp90 function and reduce azole resistance in a manner that phenocopies Hsp90 inhibition (Cowen and Lindquist 2005). In our experimental evolution study, most of the 290 lineages initiated went extinct, while the 14 lineages that evolved resistance to the combination of azole and inhibitor of

Hsp90 or calcineurin acquired a diversity of resistance mechanisms. These resistance mechanisms included mutations that rendered *erg3*-mediated azole resistance independent of the stress response regulator calcineurin, mutations that blocked the effects of the Hsp90 or calcineurin inhibitor, and large-scale aneuploidies. This experimental evolution approach provides a powerful system to predict the mechanisms by which resistance to drug combinations may evolve in the clinic. Consistent with the relevance of our findings, the increased resistance to azole and inhibitor of Hsp90 or calcineurin in isolates that evolved in an HIV-infected patient was accompanied by mutations causing overexpression of multidrug transporters (White 1997; Cowen and Lindquist 2005), as expected for the *PDR1* mutation identified in one of our lineages.

One of the most prevalent mechanisms of resistance identified in our evolved populations was mutation in the target of the drug used in combination with azole during the evolution experiment. For Hsp90 inhibitors, it has been predicted that the probability of target-based resistance would be relatively low given that the amino acid residues in the nucleotide-binding site of Hsp90 family members are highly conserved from bacteria to mammals (Chen et al. 2006), suggesting that mutations that confer resistance would likely inactivate this essential molecular chaperone. This has helped fuel research on Hsp90 as a target for development of anti-cancer drugs, where inhibiting Hsp90 can impair the function of a multitude of oncoproteins (Whitesell and Lindquist 2005; Trepel et al. 2010; Neckers and Workman 2012). Despite the constraints, there is precedent for point mutations in Hsp90 conferring resistance to Hsp90 inhibitors. One study engineered *S. cerevisiae* strains to be hypersensitive to drugs and expressed yeast or human Hsp90 as the sole source of the chaperone; introduction of a single point mutation (A107N for yeast, A121N for human Hsp90α, and A116N for human Hsp90β) conferred resistance to Hsp90 inhibitors (Millson et al. 2010). Further, the fungus that produces radicicol, *Humicola fuscoatra*, harbours an Hsp90 with reduced binding affinity to radicicol but not geldanamycin (Prodromou et al. 2009). Three of our evolved lineages acquired substitutions in Hsp90 that rendered *erg3*-mediated azole resistance more recalcitrant to the effects of Hsp90 inhibitors. For one *S. cerevisiae* lineage (Sc-G-14) and one *C. albicans* lineage (Ca-G-10), the mutations were in the nucleotide-binding domain, consistent with impairing drug binding. For *S. cerevisiae* lineage Sc-G-12, the mutation led to a premature stop codon (K385*); consistent with this *HSC82* mutation causing a loss of function, deletion of *HSC82* in the parental strain phenocopied resistance of Sc-

G-12. Reducing dosage of a drug target often confers hypersensitivity to the drug rather than resistance (Ericson et al. 2010); this may suggest compensatory upregulation of the other *S. cerevisiae* gene encoding Hsp90, *HSP82*, which could confer elevated resistance. Target-based resistance to Hsp90 inhibitors has yet to emerge in Hsp90 inhibitor clinical trials, suggesting that these mutations may be associated with a fitness cost.

Mutations in the drug target also emerged as a mechanism that renders *erg3*-mediated azole resistance recalcitrant to the effects of calcineurin inhibitors in our evolved lineages. Two *S. cerevisiae* lineages acquired mutations in *FPR1*, which encodes the immunophilin that FK506 must bind to in order to form the protein-drug complex that inhibits calcineurin function (Kissinger et al. 1995). A V108F substitution was identified in Sc-F-3 and a nine amino acid duplication near the protein midpoint was identified in Sc-F-2 (dupG53-D61). These alterations likely reduce but do not block FK506 binding, given that deletion of *FPR1* conferred a higher level of FK506 resistance. There is precedent for overexpression or disruption of *FPR1* conferring resistance to FK506 in *S. cerevisiae* (Heitman et al. 1991b), as well as for a W430C amino acid substitution in one of the two redundant calcineurin catalytic subunits Cna2 (Cardenas et al. 1995). One *C. albicans* lineage, Ca-F-9, acquired a mutation in the catalytic subunit of calcineurin, $CNA1^{C1201A}$, which results in a S401* premature stop codon that confers resistance to azole and both FK506 and cyclosporin A (Fig. 2-5), likely due to hyperactivation of calcineurin (Sanglard et al. 2003). Despite the emergence of target-based resistance to calcineurin inhibitors *in vitro*, there may be significant constraints that minimize the emergence of resistance in the human host. FK506 (tacrolimus) and cyclosporin A are front line immunosuppressants broadly used in the clinic to inhibit calcineurin function, thereby blocking T-cell activation in response to antigen presentation and suppressing immune responses that contribute to transplant rejection (Hemenway and Heitman 1999; Gaali et al. 2011). Invasive fungal infections occur in ~40% of transplant recipients including those that receive a calcineurin inhibitor as an immunosuppressant (Paya 1993), however, this immunosuppressive therapy does not select for resistance to calcineurin inhibitors in *C. albicans* or *Cryptococcus neoformans* recovered from these patients (Blankenship et al. 2005; Reedy et al. 2006). That resistance has not been observed in the host suggests that the resistant mutants may have reduced fitness relative to their sensitive counterparts or that other selective constraints alter the evolutionary dynamics.

Several of our evolved lineages took a distinct evolutionary trajectory, and evolved azole resistance mechanisms that are independent of the cellular stress response regulators. *S. cerevisiae* lineage Sc-F-1 evolved cross-resistance to azole and FK506 as well as azole and cyclosporin A (Fig. 2-4). The azole resistance phenotype was independent of calcineurin but dependent on Hsp90 (Fig. 2-4), suggesting a resistance mechanism that is contingent upon distinct Hsp90 downstream effectors, such as Mkc1 (LaFayette et al. 2010). We identified an adaptive mutation in *MOT3* (Table 2-3), a transcriptional repressor of ergosterol biosynthesis genes (Hongay et al. 2002), which resulted in a premature stop codon, G265* and likely a loss-of-function allele (Fig. 2-4c). Loss of function of Mot3 would lead to overexpression of ergosterol biosynthesis genes, which could minimize the impact of azoles on their target or could lead to a change in sterol balance that reduces the dependence of azole resistance on calcineurin. Changes in membrane composition may also explain the resistance of *C. albicans* lineage Ca-F-8 to azoles and calcineurin inhibitors, which was attributed to a mutation in the ortholog of *S. cerevisiae LCB1* (Fig. 2-5), encoding a regulator of sphingolipid biosynthesis. Notably, Mot3 is also a prion protein, which can convert between structurally and functionally distinct states, at least one of which is transmissible (Alberti et al. 2009); changes to Mot3 conformation and activity can modulate phenotypic variation in *S. cerevisiae*, and thus may influence the evolution of drug resistance phenotypes. *S. cerevisiae* lineage Sc-G-13 evolved a small increase in resistance to azole and geldanamycin associated with a mutation in *PDR1*, which encodes a transcription factor that regulates the expression of drug transporters such as *PDR5* (Fig. 2-4). Gain-of-function mutations in *PDR1* are known to confer azole resistance that is independent of Hsp90 and calcineurin (Kolaczkowska and Goffeau 1999; Anderson et al. 2003; Cowen and Lindquist 2005). Cross-resistance to azole and FK506 may not have been observed because FK506 inhibits Pdr5-mediated efflux (Hendrych et al. 2009). The weak resistance phenotype could reflect a small increase in transporter expression, or a fitness cost of the *PDR1* mutation in an *erg3* mutant background (Anderson et al. 2006).

Several of the *C. albicans* lineages that evolved resistance to azole and calcineurin inhibitors demonstrated a complex genomic landscape of aneuploidies. The emergence of azole resistance in *C. albicans* has been associated with general aneuploidies as well as the formation of a specific isochromosome composed of two left arms of chromosome 5 (i5L) (Selmecki et al. 2006). The

isochromosome confers azole resistance due to increased dosage of two genes located on the left arm of chromosome 5: *ERG11*, which encodes the target of the azoles; and *TAC1*, which encodes a transcriptional regulator of multidrug efflux pumps (Selmecki et al. 2008). Our lineages were resistant to azoles at the outset of the experiment, suggesting that the aneuploidies emerged in response to stress or were selected as a mechanism of resistance to the drug combination. Ca-F-4, Ca-F-5, Ca-F-6, and Ca-F-7 all had numerous aneuploidies relative to the parental strain (Fig. 2-6). One aneuploidy that was common to all four lineages, was increased copy number of chromosome 4, suggesting that an important resistance determinant might reside on this chromosome. While one might predict that such aneuploidies would be associated with a fitness cost, it is notable that a previous analysis of isolates carrying the i5L isochromosome demonstrated improved fitness in the presence and absence of azoles, relative to their drug-sensitive counterpart (Selmecki et al. 2009b). In contrast, many azole resistance mutations are associated with a fitness cost (Sasse et al. 2012), though this cost can be mitigated with further evolution (Cowen et al. 2001). The prevalence of aneuploidies in the *C. albicans* lineages underscores the remarkable genomic plasticity of this pathogen (Selmecki et al. 2010), and the diversity of genomic alterations that can accompany adaptation.

The landscape of genetic and genomic changes observed in our evolved lineages illuminate possible mechanisms by which resistance to drug combinations might evolve in the human host and suggest candidate targets to minimize the emergence of resistance. Despite optimizing our selection conditions to favour the evolution of resistance to the drug combination, the majority of lineages went extinct (Fig. 2-1). Consistent with constraints that minimize the evolution of resistance to these drug combinations, treatment of organ transplant patients with calcineurin inhibitors has not yielded resistance to these drugs in fungal pathogens recovered from these patients despite the extensive use of these drugs in patient populations (Blankenship et al. 2005; Reedy et al. 2006). While Hsp90 inhibitors remain at the clinical trial stage for cancer and other diseases (Luo et al. 2010; Trepel et al. 2010; Dolgin and Motluk 2011; Neckers and Workman 2012), resistance has yet to emerge in these patient populations. Although there are a multitude of mechanisms that can confer resistance to the drug combinations, they may not be favoured due to fitness costs in the complex host environments.

The mechanisms by which resistance to the drug combinations evolved in our lineages suggest novel targets that could be exploited to block the evolution of drug resistance. Drug interactions have tremendous potential to influence the evolution of drug resistance (Yeh et al. 2009). Elegant studies with antibacterials emphasize that the impact of these interactions are often more complex than anticipated (Chait et al. 2007; Hegreness et al. 2008; Michel et al. 2008; Torella et al. 2010). While synergistic interactions that yield inhibitory effects can maximize the rate at which infection is cleared, antagonistic interactions that yield inhibitory effects smaller than expected can suppress the evolution of multi-drug resistance. Ultimately, a systems biology approach incorporating experimental evolution, genetics and genomics, and clinical samples will be crucial for the development of effective strategies to enhance the efficacy of antimicrobial agents and minimize the evolution of drug resistance.

2.4 Materials and Methods

Strain construction and culture conditions. All *Saccharomyces cerevisiae* and *Candida albicans* strains were archived in 25% glycerol and maintained at -80˚C. Strains were typically grown and maintained in rich medium (YPD: 1% yeast extract, 2% bactopeptone, 2% glucose, with 2% agar for solid medium only), or in synthetic defined medium (SD, 0.67% yeast nitrogen base, 2% glucose, with 2% agar for solid medium only), supplemented with amino acids, as indicated. Strains were transformed using standard protocols. Strains used in this study are listed in Table 2-6.

Plasmid construction. Plasmids were constructed using standard recombinant DNA techniques. All plasmids were sequenced to confirm the absence of spurious non-synonymous mutations.

Evolution experiment. Evolution experiments were initiated with three ancestral strains of *erg3*-mediated azole resistant strains: two haploid *S. cerevisiae* strains (*erg3Δ* and *erg3*^{W148*}) and one *C. albicans* strain (*erg3Δ*/*erg3Δ*; see Table 2-6). A founder colony was established for each ancestral strain and grown overnight in liquid, rich medium (YPD) without drug. Culture was transferred to a YPD plate with combinatorial drug concentrations of azole (fluconazole or miconazole) and geldanamycin, or azole (fluconazole or miconazole) and FK506 (i.e. treatments; see Table 2-1). Geldanamycin and FK506 were selected based on their specificity of target inhibition and their

capacity to abrogate *erg3*-mediated azole resistance (Cowen and Lindquist 2005); fluconazole and miconazole were selected as clinically relevant azoles of the triazole and imidazole class, respectively (Cowen and Steinbach 2008; Shapiro et al. 2011).

Strain name	***Strain number***	***Species***	***Genotype***	***Source***
Sc erg3^{W148*}	ScLC7	*S. cerevisiae*	*ura3*::KAN, *ERG3*^{W148*}	(Cowen and Lindquist 2005)
*Sc erg3*Δ	ScLC10	*S. cerevisiae*	As BY4741 (*his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*); *erg3*::KAN	(Cowen and Lindquist 2005)
*Ca erg3*Δ/*erg3*Δ	CaLC660	*C. albicans*	*arg4/arg4 his1/his1 URA3/ura3*::*imm434 IRO1/iro1*::*imm434 CaTAR::HIS3 erg3::FRT/erg3::FRT*	(Robbins et al.)
Sc-F-1	ScLC1367	*S. cerevisiae*	As ScLC7, *MOT3* (G265*)	This study
Sc-F-2	ScLC1441	*S. cerevisiae*	As ScLC7, *FKS1* (dupG53 – D61)	This study
Sc-F-3	ScLC1443	*S. cerevisiae*	As ScLC7, *FKS1* (V108F)	This study
Ca-F-4	CaLC1370	*C. albicans*	As CaLC660	This study
Ca-F-5	CaLC1371	*C. albicans*	As CaLC660	This study
Ca-F-6	CaLC1372	*C. albicans*	As CaLC660	This study
Ca-F-7	CaLC1373	*C. albicans*	As CaLC660	This study
Ca-F-8	CaLC1374	*C. albicans*	As CaLC660	This study
Ca-F-9	CaLC1503	*C. albicans*	As CaLC660	This study
Ca-G-10	CaLC1486	*C. albicans*	As CaLC660, *HSP90* (D91Y)	This study
Sc-G-11	ScLC1538	*S. cerevisiae*	As ScLC10	This study
Sc-G-12	ScLC1539	*S. cerevisiae*	As ScLC7, *HSC82* (K385*)	This study
Sc-G-13	ScLC1540	*S. cerevisiae*	As ScLC10, *PDR1* (P865R)	This study
Sc-G-14	ScLC1541	*S. cerevisiae*	As ScLC10, *HSC82* (I117N)	This study
*Sc erg3*Δ *pdr1*Δ	ScLC485	*S. cerevisiae*	his3Δ0 leu2Δ0 met15Δ0 ura3Δ0 lys2Δ0 *erg3*::KAN *pdr1*::KAN	
Sc-F-1 *cnb1*Δ	ScLC1437	*S. cerevisiae*	As ScLC1367, *cnb1*::NAT	This study
*Sc erg3 cnb1*Δ	ScLC1439	*S. cerevisiae*	As ScLC7, *cnb1*::NAT	This study
Sc-F-2 *cnb1*Δ	ScLC1454	*S. cerevisiae*	As ScLC1441, *cnb1*::NAT	This study
Sc-F-3 *cnb1*Δ	ScLC1456	*S. cerevisiae*	As ScLC1443, *cnb1*::NAT	This study
Sc-F-3 *fpr1*Δ	ScLC1569	*S. cerevisiae*	As ScLC1443, *fpr1*::HYGB	This study
Sc erg3 ^{W148*} *fpr1*Δ	ScLC1570	*S. cerevisiae*	As ScLC7, *fpr1*::HYGB	This study
Sc-F-3 *fpr1*Δ	ScLC1584	*S. cerevisiae*	As ScLC1569, pLC564 (*URA3*)	This study
Sc-F-3 *fpr1*Δ	ScLC1585	*S. cerevisiae*	As ScLC1570, pLC564 (*URA3*)	This study
Sc-F-3 *fpr1*Δ	ScLC1598	*S. cerevisiae*	As ScLC1569, pLC565 (*URA3*)	This study
Sc-F-3 *fpr1*Δ	ScLC1599	*S. cerevisiae*	As ScLC1570, pLC565 (*URA3*)	This study

Sc-G-12 *hsc82*Δ	ScLC1650	*S. cerevisiae*	As ScLC1539, *hsc82*::HYGB	This study
Sc-G-14 *hsc82*Δ	ScLC1652	*S. cerevisiae*	As ScLC1541, *hsc82*::HYGB	This study
*Sc erg3*Δ *hsc82*Δ	ScLC1658	*S. cerevisiae*	*erg3*::natR *hsc82*::kanR can1::*MFA1*pr-*HIS3 lyp*-1 *leu2*Δ*0 ura3*Δ*0 met15*Δ*0*	(Tong et al. 2004)
Sc-F-2 *fpr1*Δ	ScLC1879	*S. cerevisiae*	As ScLC1441, *fpr1*::HYGB	This study
Sc-G-14 *hsc82*Δ + p*HSC82*	ScLC2024	*S. cerevisiae*	As Sc-G-14, pLC28 (*LEU2*)	This study
Sc-G-14 *hsc82*Δ + p*HSC82*^{I117N}	ScLC2025	*S. cerevisiae*	As ScLC1541, pLC636 (*LEU2*)	This study
*Sc erg3*Δ *hsc82*Δ + p*HSC82*	ScLC2026	*S. cerevisiae*	As ScLC1658, pLC28 (*LEU2*)	This study
*Sc erg3*Δ *hsc82*Δ + p*HSC82*^{I117N}	ScLC2027	*S. cerevisiae*	As ScLC1658, pLC636 (*LEU2*)	This study
Sc-F-2 *fpr1*Δ + p*FPR1*	ScLC2126	*S. cerevisiae*	As ScLC1879, pLC564 (*URA3*)	This study
Sc erg3^{W148*} *fpr1*Δ + p*FPR1*$^{dupG53 - D61}$	ScLC2127	*S. cerevisiae*	As ScLC1570, pLC653 (*URA3*)	This study
Sc-F-2 *fpr1*Δ + p*FPR1*$^{dupG53 - D61}$	ScLC2128	*S. cerevisiae*	As ScLC1879, pLC564 (*URA3*)	This study
Sc-G-13 *pdr1*Δ	ScLC2134	*S. cerevisiae*	As ScLC1540, *pdr1*::HYGB	This study
Sc erg3^{W148*} *hsc82*Δ	ScLC2139	*S. cerevisiae*	As ScLC7, *hsc82*::HYGB	This study
Ca-G-10 *HSP90/HSP90*	CaLC2293	*C. albicans*	As CaLC1486, pLC455	This study
Ca-G-10 *HSP90/HSP90*	CaLC2294	*C. albicans*	As CaLC1486, pLC455	This study
Ca erg3/erg3 HSP90/HSP90^{D91Y}	CaLC2339	*C. albicans*	As CaLC660, pLC700	This study
Ca erg3/erg3 HSP90/HSP90^{D91Y}	CaLC2340	*C. albicans*	As CaLC660, pLC700	This study
Sc erg3^{W148*} *mot3*Δ	ScLC2455	*S. cerevisiae*	As ScLC7, *mot3*::HYGB	This study
Sc-F-1 *mot3*Δ	ScLC2457	*S. cerevisiae*	As ScLC1367, *mot3*::HYGB	This study

Table 2-6. Strains used in this study.

Treatments were selected for the evolution experiment based on growth phenotype in the dose response matrixes, such that strong directional selection for resistance would be applied. Concentrations were also varied to favour the emergence of distinct mechanisms of resistance. Lineages were then propagated in replicate in either 96-well plates (Sarstedt; 48 lineages initiated in this format) or 24-well plates (Becton Dickinson Labware; 242 lineages initiated in this format). The plates were formatted as described in Fig. 2-1b. For propagation in 96-well plates, 1 µl of culture was transferred from the overnight culture to a final volume of 100 µl. Lineages were grown in a Tecan GENios plate reader and incubator at 30°C with constant agitation for three days. Subsequently, 1 µl of culture was transferred to a new plate containing YPD and treatment. Transfers occurred every 3 days to allow slow growing lineages to reach carrying capacity. This process was repeated until robust growth was present in some treatment wells. The experimental design for lineages propagated in 24-well plates was the same with the following adjustments: different drug combinations were selected for treatments; 10 µl of culture was transferred to 990 µl of YPD with treatment; plates were maintained at 30°C with constant agitation in a shaking incubator and transfers occurred every two days. With this dilution factor of 1/100, ~ 6.6 generations occurs between transfers. The effective population size per lineage of ~4.6×10^6 was estimated as described (Wahl and Gerrish 2001), given that cultures reached saturation of ~10^7 cells/ml between transfers. Lineages that demonstrated reproducible resistance to the drug combination in which they were propagated were archived. Lineages unable to grow in the presence of the drug combination, either from when the cultures were initiated or over the course of the evolution experiment, were considered extinct. A summary of treatment concentrations, number of transfers and type of plate evolved in can been found in Table 2-1.

Minimum inhibitory concentration and checkerboard assays. Resistance to drug combinations was assayed in 96-well microtiter plates, as previously described (Cowen and Lindquist 2005; Singh et al. 2009). Minimum inhibitory concentration (MIC) assays were set up to a final volume of 0.2 ml/well. MICs were performed in the absence of fluconazole (Sequoia Research Products) or with a constant concentration of fluconazole or miconazole (Sigma–Aldrich Co.), as indicated in the figures. All gradients were two-fold dilutions per step, with the final well containing no drug. The starting concentration of geldanamycin (Invivogen) gradients was 50 µM for *S. cerevisiae* strains and 5 µM for *C. albicans* strains. The starting concentration of FK506 (A.G.

Scientific) gradients was 6 μM for *S. cerevisiae* strains and 100 μM for *C. albicans* strains. The starting concentration of radicicol (A.G. Scientific) gradients was 25 μM for both *S. cerevisiae* and *C. albicans* strains. The starting concentration of cyclosporin A (Calbiochem) gradients was 50 μM for both *S. cerevisiae* and *C. albicans* strains. The cell densities of overnight cultures were determined and diluted to an inoculation concentration of ~ 10^3 cells/well. Plates were incubated at 30˚C in the dark for the period of time specified in the figure legend. Cultures were resuspended and absorbance at 600 nm was determined using a spectrophotometer (Molecular Devices) and corrected for background of the corresponding medium. OD measurements were standardized to either drug-free or azole-only control wells, as indicated. Data was plotted quantitatively with colour using Java Treeview 1.1.3 (http://jtreeview.sourceforge.net/). Resistance phenotypes were assessed on multiple occasions and in duplicate on each occasion with concordant results, validating that the phenotypes are reproducible and stable.

Dose response matrices, or checkerboard assays, were performed to a final volume 0.2 ml/well in 96-well microtiter plates, as previously described (LaFayette et al. 2010). Two-fold dilutions of fluconazole were titrated along the X-axis from a starting concentration of 256 μg/ml, with the final row containing no fluconazole. Along the Y-axis, either geldanamycin or FK506 was titrated in two-fold dilutions with the final column containing no geldanamycin or FK506. The starting concentration of geldanamycin was 5 μM for checkerboards with either *S. cerevisiae* or *C. albicans* strains. The starting concentration of FK506 was 4 μM for checkerboards with *S. cerevisiae* and 40 μM for checkerboards with *C. albicans* strains. Concentrations were selected to cover a range that spanned from no effect on growth to near complete inhibition of growth. Plates were inoculated and growth assessed as was performed for MIC assays.

Fluconazole was dissolved in sterile ddH_2O. The Hsp90 inhibitors geldanamycin and radicicol and the calcineurin inhibitors FK506 and cyclosporin A were dissolved in DMSO. Myriocin (Sigma) was dissolved in methanol.

Genome Sequencing. C. albicans cell pellets were digested with R-Zymolase for 1 hour (Zymo Research, D2002), prior to genomic DNA extraction with phenol-chloroform (EMD Millipore, EMD6810), and sodium acetate precipitation. Whole genome libraries were prepared using Nextera XT kits (Illumina, FC-131-1096) according to manufacturer's protocol. Libraries were

sequenced on the Illumina HiSeq2000 platform using paired reads (101 bp) and version 3 reagents and chemistry.

The yeast genomes were sequenced in a multiplexed format, where an oligonucleotide index barcode was embedded within adapter sequences that were ligated to genomic DNA fragments (Smith et al. 2010). Only one mismatch per barcode was permitted to prevent contamination across samples. Next, the sequence reads were filtered for low quality base calls trimming all bases from 5' and 3' read ends with Phred scores < Q30. Trimming sequence reads for low quality base calls drastically lowered false positive SNV calls.

De-multiplexed and trimmed reads from the *S. cerevisiae* strains were aligned to the *S288c 2010* genome, a high fidelity sequence from an individual yeast colony (from F. Dietrich's lab at Duke University; it is the SGD reference genome as of February 2011) (Engel et al. 2010). Reads from the *C. albicans* strains were aligned to the SC5314 genome from CGD (Skrzypek et al. 2010). While *C. albicans* is an obligate diploid, the current build of the genome, assembly 21, is a haploid genome, and is more accurate than the original diploid genome, assembly 19 (Jones et al. 2004; van het Hoog et al. 2007). The diploid assembly was not used because it features 412 supercontigs with non-obvious heterozygosity, whereas the haploid assembly has been curated and organized into 8 chromosomes (van het Hoog et al. 2007).

Sequence reads were aligned with Bowtie2, which was chosen over other commonly used short-read aligners such as Illumina's Eland, Maq, SOAP and BWA because it has been reported to be one of the fastest accurate aligners (Li et al. 2008a; Li et al. 2008b; Langmead et al. 2009; Li and Durbin 2009; Langmead and Salzberg 2012). Additionally, it was chosen because it is updated frequently, supports variable read lengths within a single input file, is multi-threaded with a minimal memory and temporary file footprint and supports the standard Sequence Alignment/Map (SAM) file format (Langmead et al. 2009; Li et al. 2009a; Li and Homer 2010; Langmead and Salzberg 2012). Alignments and all subsequent sequence data were visualized using the Savant Genome Browser (Fiume et al. 2010). The average coverage of each genome was calculated and was sufficient for confident variant detection (Table 2-2).

Aligned sequence reads for *S. cerevisiae* in binary SAM (BAM) format were subsequently processed using the UnifiedGenotyper package of the Genome Analysis Toolkit (GATK), which features a comprehensive framework for discovering SNVs and calculating coverage across genomic data (McKenna et al. 2010; DePristo et al. 2011). Variants detected in the *S. cerevisiae* parental strains were subtracted from complete variant lists, yielding a set of novel variants that emerged during strain growth in the presence of drug. Since *C. albicans* is diploid, we processed the reads with a more accurate approach using the probabilistic framework JointSNVMix, which uses paired parental and evolved strain sequence data to determine significant novel variants (Roth et al. 2012). After identifying candidate SNVs, the threshold for homozygous SNV calls for both haploid (*S. cerevisiae*) and diploid (*C. albicans*) systems was set to 85% alternate (non-reference) basecalls at a specific position. In a diploid system, 35% was the threshold set to identify heterozygous SNVs. All variant positions required a minimum coverage of 15x to be considered as a candidate SNV. The total number of high-confidence novel mutations agrees with mutation rates observed previously for *S. cerevisiae* (Liti et al. 2009). To further verify that the sequence data are of high quality, we compared two distinct sequence runs from two different sequence library preparations of the same parent *C. albicans* strain CaLC660. The total number of diploid single nucleotide variants that exist between the parent strain and the reference genome (SC5314) is 3748, therefore there is 99.99% concordance between both sequence replicates.

The software package CNV-seq was used to identify chromosomal regions that varied in copy number between parental strains and evolved lineages (Xie and Tammi 2009). This analysis found no significant CNVs in the *S. cerevisiae* strains, but numerous large variants were observed in *C. albicans*.

Sequence data is publicly available on the NCBI Short Read Archive with accession SRA065341.

3 Conservation of chromatin architecture from Archaea to Eukarya

The eukaryotic nucleosome is the fundamental unit of chromatin, comprising a protein octamer that wraps approximately147 bp of DNA and has essential roles in DNA compaction, replication and gene expression. Nucleosomes and chromatin have historically been considered to be unique to eukaryotes, yet studies of select archaea have identified homologs of histone proteins that assemble into tetrameric nucleosomes. Here I report the first archaeal genome-wide nucleosome occupancy map, as observed in the halophile *Haloferax volcanii*. Nucleosome occupancy was compared with gene expression by compiling a comprehensive transcriptome of *Hfx. volcanii*. I found that archaeal transcripts possess hallmarks of eukaryotic chromatin structure: nucleosome-depleted regions at transcriptional start sites and conserved −1 and +1 promoter nucleosomes. These observations demonstrate that histones and chromatin architecture evolved before the divergence of Archaea and Eukarya, suggesting that the fundamental role of chromatin in the regulation of gene expression is ancient.

Portions of this chapter have been adapted from the following publication:

Ammar, R., Torti, D., Tsui, K., Gebbia, M., Durbic, T., Bader, G. D., Giaever, G. & Nislow, C. 2012. Chromatin is an ancient innovation conserved between Archaea and Eukarya. *eLife*, 1, e00078.

Author contributions:

RA: Conception and design, acquisition of data, analysis and interpretation of data, drafting and revising the article

DT: Performed the sequencing

KT: Prepared samples for sequencing

MG: Prepared samples for sequencing

TD: Performed the sequencing

GB: Conception and design, analysis and interpretation of data, drafting and revising the article

GG: Interpretation of data, drafting and revising the article

CN: Directed the research, prepared biological material, conception and design, acquisition of data, analysis and interpretation of data, drafting and revising the article

3.1 Chromatin, histones and archaea

The genomic DNA of eukaryotes is wrapped around protein complexes composed of highly conserved core histone proteins (Jiang and Pugh 2009). DNA is packaged into cells by folding and compacting negatively charged double-stranded helices around positively charged histone residues. This compaction yields chromatin, widely recognized as a hallmark of eukaryotes. The fundamental repeating unit of chromatin is the nucleosome, defined as the stretch of DNA that is bound to a histone complex (Fig. 3-1). Nucleosomes occupy the majority of genomic DNA in eukaryotes with linker DNA gaps between them. Crystal structures have revealed that ~150bp of DNA is wrapped by eukaryotic histone complexes (Richmond and Davey 2003).

Histone fold domains are not exclusive to eukaryotes and have been shown to exist in some archaeal proteins as well (Arents and Moudrianakis 1995). These histone proteins are present across the archaeal domain suggesting that histones evolved from a common ancestor to Archaea and Eukarya (Sandman and Reeve 2006; Friedrich-Jahn et al. 2009). Archaeal nucleosome core particles protect ~60 bp of DNA, approximately half that of eukaryotic nucleosomes, as demonstrated by the work of Reeve and colleagues (Pereira et al. 1997). Comparing both eukaryotic and archaeal nucleosomes, the former is an octamer composed of heterodimers of histones H2A, H2B, H3 and H4 whereas the latter histones assemble from homologs of H3 and H4 proteins (Pereira and Reeve 1998; Talbert and Henikoff 2010). Archaeal histones can form both homodimers and heterodimers, as well as homotetramers, whereas eukaryotic histones contain hydrophobic dimerization surfaces that restrict assembly of the octamer from H2A-H2B and H3-H4 heterodimers (Sandman and Reeve 2006; Talbert and Henikoff 2010).

Using single-nucleotide resolution maps of archaeal nucleosome occupancy and gene expression, I demonstrate that the architecture of archaeal chromatin and the occupancy of its nucleosomes along transcription units are conserved. I constructed a nucleosome occupancy map of the halophilic archaeon *Haloferax volcanii*, a member of the phylum euryarchaeota, originally discovered in the highly saline sediment of the Dead Sea (Mullakhanbhai and Larsen 1975). The genome of *Hfx. volcanii* has an average GC content of 65% and a total genome length of 4Mb (Hartman et al. 2010) composed of five circular genetic elements: a 2.8Mb main chromosome,

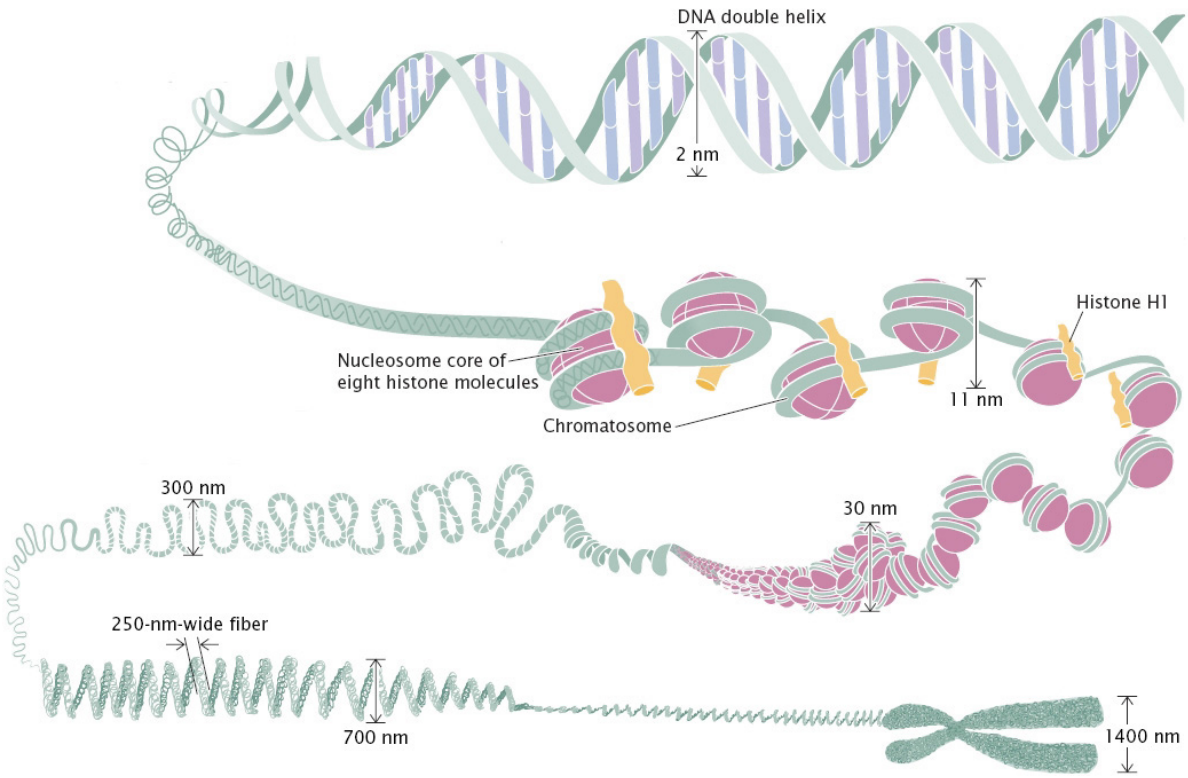


Figure 3-1. Organization of eukaryotic histones and chromatin. In eukaryotes, the double helix of genomic DNA is coiled around histone octamers comprised of histones H2A, H2B, H3 and H4 sometimes accompanied by histone H1. These wrapped nucleosomes form a structure analogous to “beads on a string” that folds into a 30nm fibre conformation. These fibres are subsequently folded in structures of higher order to form a fully compacted and condensed chromosome (Figure modified from Annunziato, 2008).

three smaller chromosomes pHV1, pHV3 and pHV4 and the plasmid pHV2. It is highly polyploid with ~15 genome copies during exponential growth and ~10 during stationary phase (Breuert et al. 2006). The histone protein of *Hfx. volcanii*, hstA (HVO_0520), has a domain architecture containing two distinct histone fold domains within the same peptide that heterodimerize similar to that of the *Methanopyrus kandleri* histone (HMk) (Geer et al. 2002; Talbert and Henikoff 2010; Marchler-Bauer et al. 2011).

3.2 Results

3.2.1 High-throughput sequencing of mononucleosomal DNA

Hfx. volcanii was cultured in rich media containing 2M NaCl (Mullakhanbhai and Larsen 1975). Genomic DNA was cross-linked and digested with micrococcal nuclease (MNase), with cell disruption accomplished by bead-beating (Tsui et al. 2012). Nucleosome-bound cross-linked genomic regions are protected from MNase digestion, in contrast to the linker DNA between nucleosomes. Mononucleosome-sized (50-60bp) DNA fragments were gel purified and libraries were sequenced on an Illumina HiSeq2000 (Fig. 3-2a). Sequence reads were aligned to the published *Hfx. volcanii* DS2 genome (Hartman et al. 2010) to generate a genome-wide nucleosome occupancy map. Controls included crosslinked DNA without MNase digestion as well as MNase treated nucleosome-free genomic DNA. The nucleosome occupancy data was significantly different than the control MNase digest of deproteinized "naked" genomic DNA ($r = 0.071$), indicating that the nucleosome map is unaffected by any potential MNase sequence bias, as expected (Chung et al. 2010; Allan et al. 2012).

To determine nucleosome midpoints, I smoothed the occupancy data using a symmetrical convolution sum with a Gaussian filter (Smith 1997). Extrema were detected in the smoothed signal, and maxima were defined as nucleosome midpoints. In the smoothed signal, the mean peak-to-peak distance for the main chromosome was 68.5bp in genic regions and 76.1bp in non-genic regions. Genic regions were defined as the transcribed region plus 40bp (the average promoter length based on Palmer and Daniels (1995)) upstream of the 5' end (Palmer and Daniels 1995). I observed a greater nucleosome density in *Hfx. volcanii* vs. all eukaryotes likely

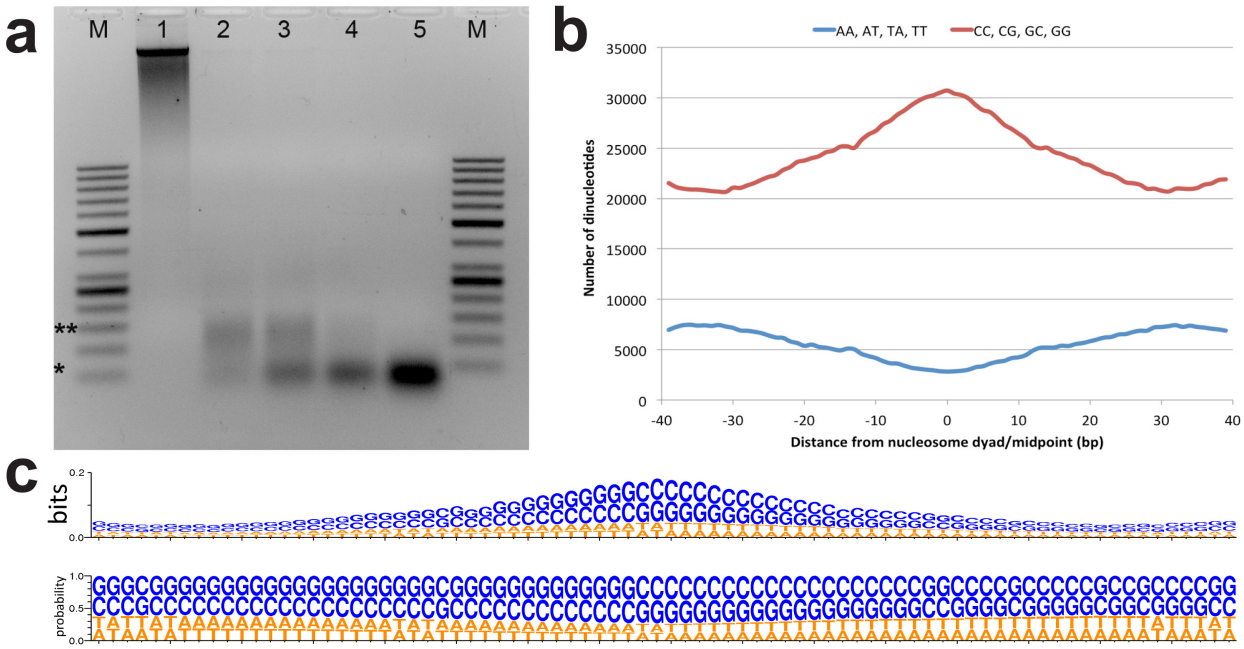


Figure 3-2. Micrococcal nuclease digestion produces nucleosomal fragments from crosslinked *Hfx. volcanii* chromatin. (**A**) Formaldehyde cross-linked chromatin was subjected to MNase digestion with increasing amounts on micrococcal nuclease (from 1 unit to 5 units). De-crosslinked DNAs were separated on a 3% agarose gel and ~60bp and ~120bp mono- and di-nucleosomes were observed. Markers (M) indicate * 50bp and ** 150bp. (**B**) The counts of AA, AT, TA, TT or CC, CG, GC, GG dinucleotides are reported at each position showing an enrichment of G/C nucleotides and a depletion of A/T nucleotides at the dyad relative to the end points of the protected fragment. This differs from the observation of Bailey *et al.* (2000), where GC, AA and TA dinucleotides were repeated at ~10bp intervals in recombinant archaeal histone B from *Methanothermus fervidus* (rHMfB) (Bailey et al. 2000). (**C**) The sequence logo of a nucleosome-binding site in *Hfx. volcanii* centered at the nucleosome midpoint. There is a significant GC enrichment towards the nucleosome midpoint. This is exhibited using both bit score and probability measures.

due to the shorter length of DNA wrapped around the archaeal histone tetramer (Pereira et al. 1997). Based on these data, the *Hfx. volcanii* genome has 14.2 nucleosomes/kilobase compared to 5.2 nucleosomes/kilobase in *Saccharomyces cerevisiae*. The resulting map reveals a periodic pattern similar to that seen in all eukaryotes examined to date; with protected regions appearing as peaks and linker regions as troughs. Sequence analysis of the entire nucleosome map showed that nucleosome midpoints were enriched with G/C nucleotides from 61.4% GC at the edge of the protected fragment to 74.6% GC at the midpoint (dyad). We found an increase of G/C nucleotides and a decrease in A/T nucleotides at the midpoint, as described recently for human cell lines (Fig. 3-2b,c) (Valouev et al. 2011). In contrast to previous studies in eukaryotes, we did not observe a periodicity in dinucleotide frequency relative to the nucleosome midpoint (Satchwell et al. 1986; Bailey et al. 2000; Albert et al. 2007).

3.2.2 Creating a transcriptome by RNA-seq

We next investigated the relationship between nucleosome occupancy and gene expression. The existing genome annotation for *Haloferax* is derived almost exclusively from ORF predictions (Hartman et al. 2010). To augment these predictions, we used deep sequencing to create a high confidence transcriptome of the main chromosome of *Hfx. volcanii*. This map allowed us to define both 5'UTR lengths and transcriptional start sites (TSSs) and transcriptional termination sites (TTSs). Total RNA was extracted from *Hfx. volcanii* cells, repetitive RNA was partially depleted via duplex-specific nuclease (DSN) normalization followed by RNA-seq (see Materials and Methods) (Zhulidov et al. 2004). Transcript sequences were aligned, assembled and quantified using TopHat and the Genome Analysis Toolkit (Trapnell et al. 2009; McKenna et al. 2010) and transcript boundaries were further trimmed based on RNA-seq coverage information, as described previously (Wurtzel et al. 2010). The final set of transcripts were manually curated yielding 3059 transcriptional units in *Hfx. volcanii*, a number that is greater than observed previously in the comparable transcriptome of the sulfur-metabolizing archaeon *Sulfolobus solfataricus* (Wurtzel et al. 2010) but fewer than the 4073 predicted *Hfx. volcanii* genes. It is likely that in the rich media conditions used in this study, not all genes are expressed. Specifically 75% of the predicted transcripts were detectably expressed, and this fraction is consistent with observations obtained for yeast gene expression in rich media (David et al. 2006). 32 novel

transcripts (absent from the predicted sequence annotation) were identified in the RNA-seq data. Most of these 32 transcripts lack significant sequence homologs, and several were classified as transposases with paralogs in *Hfx. volcanii* (Table 3-1). Notably, the gene that was most highly expressed in the transcriptome (NTRANS_0004) was not previously annotated and contains a putative N-Acyltransferase (NAT) superfamily domain. Homology searches revealed that this transcript appears to be restricted to the genomes of other halophilic archaea (Altschul et al. 1990). The architecture of this domain is homologous to chain A of the well-characterized histone acetyltransferases Gcn5, Gna1, Hpa2 in *S. cerevisiae*, suggesting a possible role for this transcript in regulating transcription via histone acetylation(Marchler-Bauer et al. 2011). Additional acyltransferases with a similar architecture have been implicated in bacteriophage-encoded DNA modifiers as well as in cold and ethanol tolerance in yeast (Du and Takagi 2007; Kaminska and Bujnicki 2008). Thus, while histone post-translational modifications have not been observed in archaeal histones (Forbes et al. 2004), our observation suggests that some rudimentary control over chromatin accessibility may occur via the action of ancient NAT family members. Furthermore acetyltransferase and deacetylase orthologs, which appear to have enzymatic activity based on their sensitivity to the histone deacetylase (HDAC) inhibitor trichostatin A have been identified in *Hfx. volcanii* (Altman-Price and Mevarech 2009). In our subsequent analysis, we focused on all genes we empirically determined to be expressed.

ID	Locus	Best BLAST hit (cutoff E-value < 1.00E-03)	E-value	Putative conserved domains
NTRANS_0001	NC_013967.1: 214386-215027	ribonuclease BN [Haloferax volcanii DS2] (YP_003534314)	2.00E-63	N/A
NTRANS_0002	NC_013967.1: 492144-492417	transposase (ISH18) [Haloferax volcanii DS2] (YP_003533477)	8.00E-22	DEDD_Tnp_IS110 super family[cl03258], Transposase
NTRANS_0003	NC_013967.1: 835366-835867	transposase (ISH51) [Haloferax volcanii DS2] (YP_003537056)	4.00E-92	N/A
NTRANS_0004	NC_013967.1: 936059-936368	hypothetical protein NatpeDRAFT_2433 [Natrinema pellirubrum DSM 15624] (ZP_08964227)	6.00E-13	NAT_SF super family[cl00357], N-Acyltransferase superfamily
NTRANS_0005	NC_013967.1: 1080193-1080364	N/A	N/A	N/A
NTRANS_0006	NC_013967.1: 1234552-1234718	N/A	N/A	N/A
NTRANS_0007	NC_013967.1: 1280150-1280358	N/A	N/A	N/A
NTRANS_0008	NC_013967.1: 1299138-1299531	hypothetical protein HVO_1426 [Haloferax volcanii DS2] (YP_003535476)	9.00E-42	DUF3984 super family[cl16124], Protein of unknown function (DUF3984)

NTRANS_0009	NC_013967.1: 1329306-1329496	N/A	N/A	N/A
NTRANS_0010	NC_013967.1: 1446004-1446415	N/A	N/A	N/A
NTRANS_0011	NC_013967.1: 1574463-1574873	hypothetical protein NatgrDRAFT_0350 [Natronobacterium gregoryi SP2] (ZP_08966502)	6.00E-58	PLPDE_III super family[cl00261], Type III Pyridoxal 5-phosphate (PLP)-Dependent Enzymes
NTRANS_0012	NC_013967.1: 1594239-1594693	N/A	N/A	N/A
NTRANS_0013	NC_013967.1: 1665493-1665917	hypothetical protein Hlac_2469 [Halorubrum lacusprofundi ATCC49239] (YP_002567112)	1.00E-07	N/A
NTRANS_0014	NC_013967.1: 1677103-1677246	N/A	N/A	N/A
NTRANS_0015	NC_013967.1: 1693889-1694145	N/A	N/A	N/A
NTRANS_0016	NC_013967.1: 1697867-1698447	transposase (ISH51) [Haloferax volcanii DS2] (YP_003533812)	N/A	N/A
NTRANS_0017	NC_013967.1: 1733429-1733686	N/A	N/A	N/A
NTRANS_0018	NC_013967.1: 1857340-1857746	N/A	N/A	N/A
NTRANS_0019	NC_013967.1: 1927912-1928132	N/A	N/A	N/A
NTRANS_0020	NC_013967.1: 2024079-2024730	hypothetical protein HVO_2159 [Haloferax volcanii DS2] (YP_003536184)	1.00E-19	genX[TIGR00462], EF-P lysine aminoacylase GenX
NTRANS_0021	NC_013967.1: 2070130-2070379	N/A	N/A	N/A
NTRANS_0022	NC_013967.1: 2229138-2229404	N/A	N/A	N/A
NTRANS_0023	NC_013967.1: 2283587-2283738	N/A	N/A	N/A
NTRANS_0024	NC_013967.1: 2316341-2316714	N/A	N/A	N/A
NTRANS_0025	NC_013967.1: 2334880-2335076	N/A	N/A	N/A
NTRANS_0026	NC_013967.1: 2368676-2368877	N/A	N/A	N/A
NTRANS_0027	NC_013967.1: 2388678-2388899	N/A	N/A	N/A
NTRANS_0028	NC_013967.1: 2499852-2500053	N/A	N/A	Periplasmic_Binding_Protein_Type_1 super family[cl10011], Type 1 periplasmic binding fold superfamily
NTRANS_0029	NC_013967.1: 2649993-2650476	transposase (ISH5) [Haloferax volcanii DS2] (YP_003537069)	3.00E-40	N/A
NTRANS_0030	NC_013967.1: 2705350-2705601	N/A	N/A	N/A
NTRANS_0031	NC_013967.1: 2760270-2760455	N/A	N/A	TIM_phosphate_binding super family[cl09108]
NTRANS_0032	NC_013967.1: 2846906-2847206	N/A	N/A	N/A

Table 3-1. Novel transcripts identified in the *Hfx. volcanii* transcriptome. Of these transcripts, NTRANS_0004 was the most abundant transcript in the transcriptome, after the 6 rRNA genes. Homology data was obtained using BLASTX with a BLOSUM45 matrix against the non-redundant protein sequence database (Altschul et al. 1990). Conserved domains were identified using the Conserved Domain Database (Marchler-Bauer et al. 2011).

3.2.3 Conserved chromatin architecture

In eukaryotes, the TSS of the majority of expressed genes is characterized by a nucleosome-depleted region (NDR) (Jiang and Pugh 2009). This NDR is flanked by the well-positioned −1 and the +1 nucleosomes. These regions direct RNA polymerase II to initiate transcription and influence the binding of promoter regulatory elements (Jiang and Pugh 2009). This stereotypical pattern of nucleosome depletion at promoters and well-ordered nucleosomes in gene bodies is found in all eukaryotes, including yeast, *Drosophila*, *Arabidopsis* and humans. Using the RNA-seq-derived transcripts for *Hfx. volcanii*, we computed the degree of aggregate nucleosome occupancy for the 2343 transcripts on the main chromosome, and found that the NDR and −1 and +1 nucleosomes are conserved in *Hfx. volcanii* (Fig. 3-3) suggesting that the interplay between chromatin and transcription is conserved in archaeal promoters. We generated nucleosome occupancy profiles for each transcript and clustered them hierarchically. Differential nucleosome density was observed with profiles encompassing between four to six nucleosomes in a 400bp DNA segment spanning 200bp on each side of the TSS (Fig. 3-3c). NDRs at transcription termination sites (TTSs) are also observed, and similar to those found in eukaryotes (Lee et al. 2007) they are less prominent than promoter NDRs in *Hfx. volcanii* (Fig. 3-4).

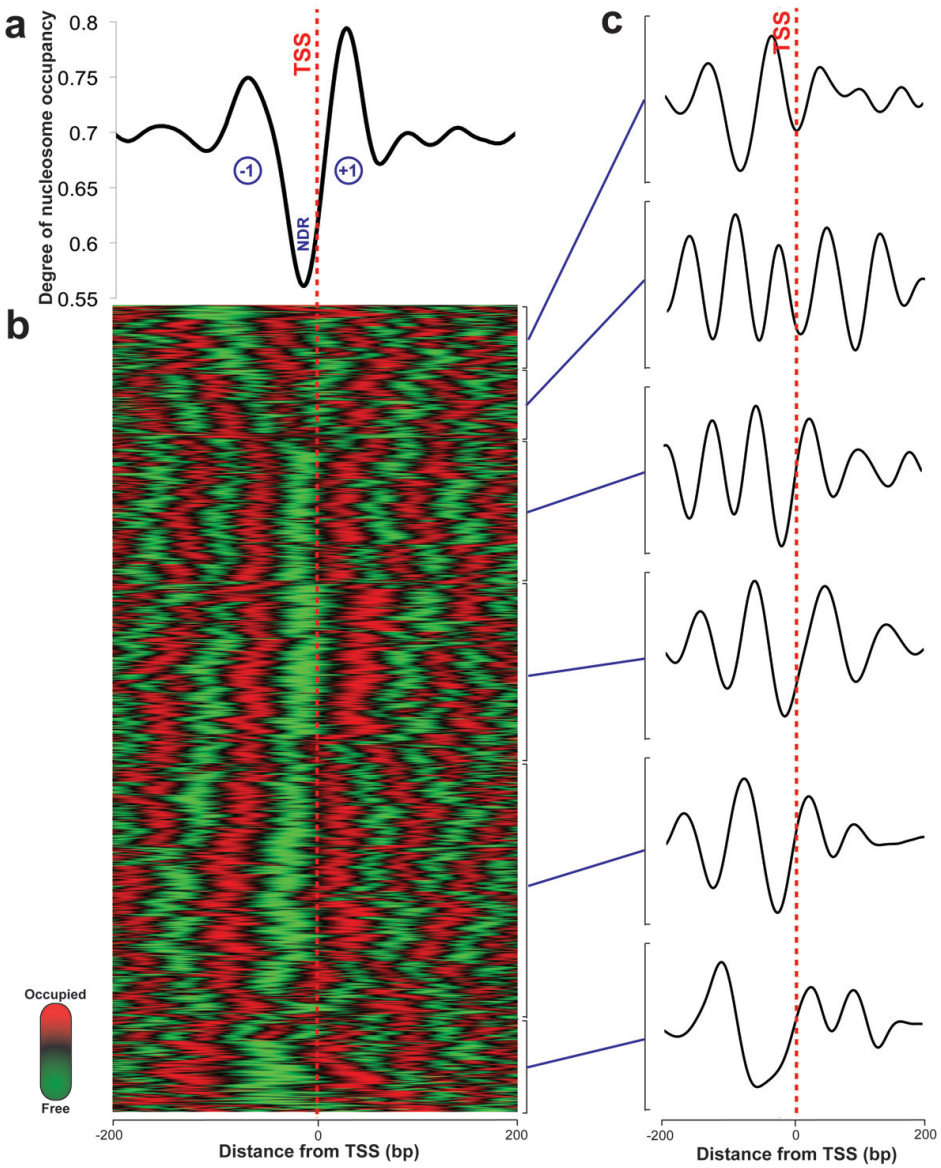


Figure 3-3. Nucleosome occupancy in *Haloferax volcanii*. (**A**) Degree of normalized nucleosome occupancy in aggregate for the main chromosome. As observed in eukaryotes, there is a prominent nucleosome-depleted region (NDR) at the transcriptional start site (TSS) preceded by a −1 nucleosome and followed by a +1 nucleosome, demonstrating that promoter genome architecture is conserved between archaea and eukaryotes. (**B**) Hierarchical clustergram for the 2343 expressed transcripts on the main *Haloferax* chromosome. Green represents nucleosome-depleted regions and red represents occupied regions. (**C**) The clustered heatmap was subdivided into the largest 6 subclades, and differential density of nucleosomes can be observed with occupancy profile clusters containing between 4 to 6 nucleosomes.

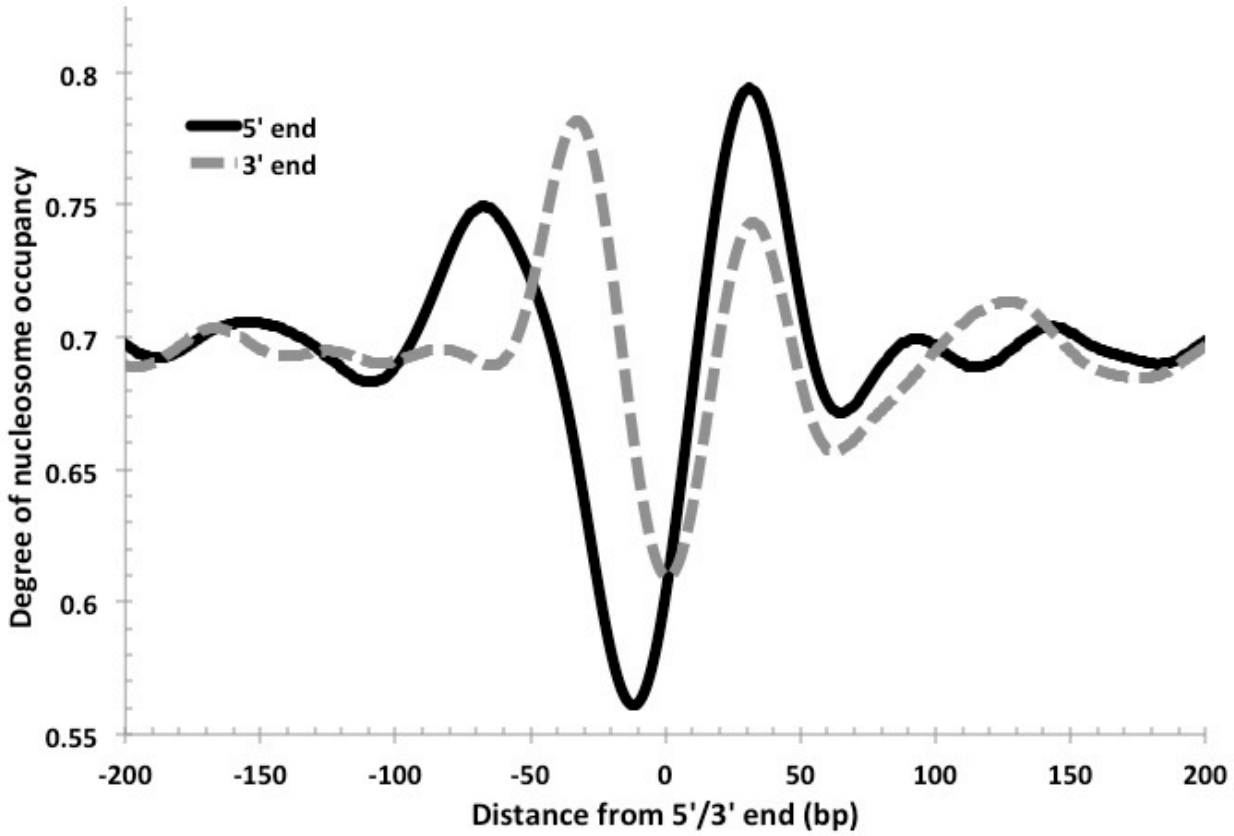


Figure 3-4. Nucleosome-depleted regions at the 5’ and 3’ ends of transcripts. As observed in eukaryotes, NDRs are also found at the transcriptional termination sites in *Hfx. volcanii*. Both 5’ and 3’ end profiles are overlaid in this figure for comparison. The 5’ NDR is, on average, more depleted and longer.

3.2.4 A universal sequence-based nucleosome occupancy predictor

The positioning of nucleosomes is related to the flexibility of DNA as it wraps around histone octamers or tetramers, but nucleosomes can be relocated by chromatin-remodelling complexes (Drew and Travers 1985; Cairns 2009). *In vivo* nucleosome occupancy is partially governed by ATP-dependent remodeling enzymes, and abolition of these enzymatic activities has been shown to lead to partial loss of global nucleosome positioning (Gkikopoulos et al. 2011; Zhang et al. 2011). Occupancy has also been shown to be greatly dictated by DNA sequence features such as GC-rich promoter sequences in humans and GC-depleted promoter sequences in yeast (Kaplan et al. 2009; Zhang et al. 2009; Valouev et al. 2011). Nucleosomal fragments can also display a ~10bp periodicity of AA/AT/TA/TT dinucleotides at the minor groove and CC/CG/GC/GG dinucleotides at the major groove, although this has not been observed in all occupancy maps (Satchwell et al. 1986; Valouev et al. 2011; Brogaard et al. 2012). The reconstitution of genomic DNA onto histone octamers *in vitro* has demonstrated correlation ($r = 0.74$) with *in vivo* occupancy data, suggesting that intrinsic DNA sequence preferences have a significant role in determining nucleosome occupancy (Kaplan et al. 2009). This observation has provided impetus for building computational models that can predict nucleosome occupancy, most notably for the yeast *Saccharomyces cerevisiae* (Segal et al. 2006; Lee et al. 2007; Kaplan et al. 2009; Tillo and Hughes 2009). To date, no such systematic efforts have been undertaken for archaea, and it is pertinent to establish such a model because the chromatin architecture around transcripts is conserved across eukaryotes and histone fold-containing archaea (Sandman and Reeve 2006; Ammar et al. 2012). Nucleosome maps in both eukaryotes and *Hfx. volcanii* exhibit increased GC content at the nucleosome dyad with the greatest increase in GC probability observed at the precise midpoint of nucleosomal protected fragments.

Currently, there are multiple methods for predicting nucleosome occupancy, however a few methods have been shown to be more accurate when tested on diverse eukaryotic nucleosome occupancy maps. The method of Field et al. (2008) was a probabilistic approach trained on features from an *in vivo* yeast nucleosome map. These features were dinucleotide frequency within nucleosomes and enrichment or depletion of specific 5-mers from nucleosomal DNA (Field et al. 2008). A subsequent model by Kaplan et al. (2009) applied an identical approach, but

was trained on *in vitro* nucleosome occupancy data when yeast genomic DNA was reconstituted onto chicken histones by salt gradient dialysis. This model used *in vitro* data to predict occupancy solely based on DNA sequence features, independent of chromatin remodeling proteins and additional factors (Kaplan et al. 2009). Tillo and Hughes (2009) also made use of the Kaplan et al. (2009) *in vitro* data, but trained a simpler linear regression Lasso algorithm. This model was based on GC content in 150bp moving average windows and 4-mers within nucleosome fragments (Tillo and Hughes 2009). To date, these methods are the most accurate predictors of nucleosome occupancy, and are based on genomic sequence alone. A key limitation of these methods, however, is that they are learned and may exhibit overfitting. In addition, they have been designed specifically for eukaryotes, and, with the release of the first nucleosome map for archaea, we have an opportunity to discover a universal model.

I present a novel model for the prediction and elucidation of nucleosome occupancy based on a Gaussian convolution sum across GC bases of genomic sequence. This approach is not probabilistic, and, as such, requires no prior training and is simpler than previously described methods. Since the sole parameter is the Gaussian interval length, based on the experimentally-derived nucleosome fragment length, the method is easily generalized to archaea and eukaryotes.

In eukaryotes it has been shown that there is an increase of G/C nucleotides and a decrease in A/T nucleotides at the dyad (Satchwell et al. 1986; Bailey et al. 2000; Albert et al. 2007; Valouev et al. 2011; Brogaard et al. 2012). This trend also exists in the halophilic archaeon *Haloferax volcanii*, which exhibited an enrichment with 61.4% GC at the edge of the protected fragment to 74.6% GC at the midpoint (Ammar et al. 2012). Since this GC content follows a Gaussian distribution with the mean point at the dyad, I created a novel method that computes the convolution sum of the GC content within a predefined interval of nucleotides using a Gaussian kernel. This allows one to effectively reverse engineer the trend of highest GC at midpoints. The sole parameter for this method is an integer based on the length of the nucleosome fragment (used to create a Gaussian interval; see Materials and Methods).

To identify the optimal Gaussian interval, I assessed the correlation between known nucleosome occupancy and predicted occupancy in a parameter search space. In *S. cerevisiae*, the optimal

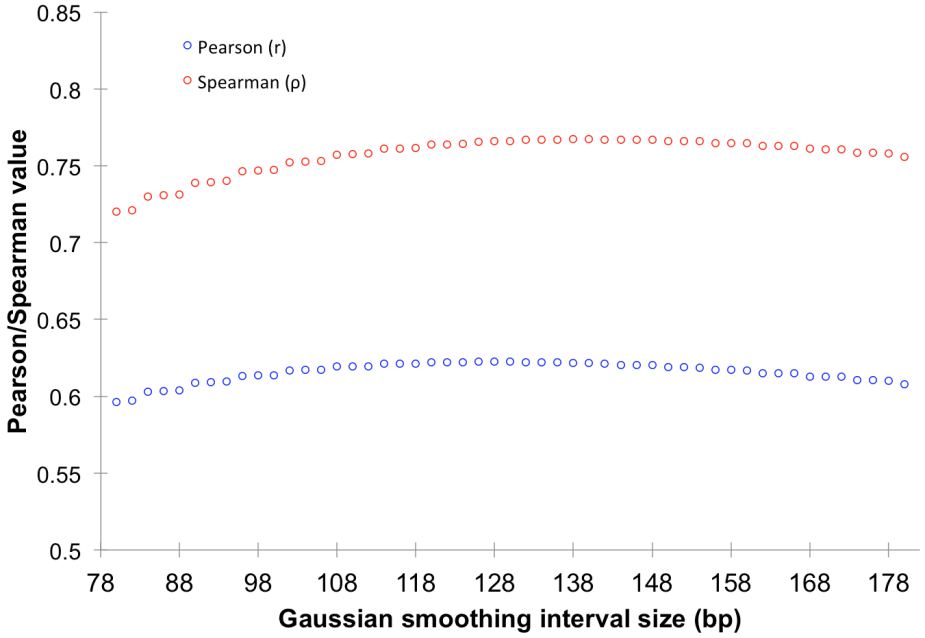


Figure 3-5. Predicting nucleosome occupancy of *S. cerevisiae* using our GC-based predictor. The GC-based predictor performance was evaluated by correlating occupancy prediction with the MNase-seq-derived nucleosome map of Van Bakel et al. (2013). Predictions are not sensitive to variation in the Gaussian interval, its sole input parameter. Based on my experience, testing with multiple data sets, I recommend choosing an interval 12bp shorter than the known protected fragment length for an organism of interest, in either eukaryotic or archaeal species.

interval was 138bp (Fig. 3-5), and in *Hfx. volcanii* it was 48bp. Since the eukaryotic nucleosome is ~150bp in length and the archaeal nucleosome is ~60bp, I suggest that optimal prediction is computed at 12bp less than the experimentally determined protected fragment size. Based on this finding, I fixed the Gaussian interval at 12bp less than the known protected fragment size for all subsequent predictions. The finding that maximal performance is observed at interval sizes less than protected fragment sizes is reminiscent of the suggestion of Widom and colleagues, stating that when processing sequence data "algorithms for aligning the selected DNA sequences should seek to optimize the alignment over much less than the full 147bp of nucleosomal DNA" based on their observation of high-affinity nucleosome-forming DNA sequences (Thastrom et al. 2004).

Using the fixed interval lengths of 138bp for eukaryotes and 48bp for archaea, I predicted genome-wide occupancy of several organisms with known occupancy maps across a spectrum of genomic GC content. These included the yeast *S. cerevisiae*, the AT-rich malaria parasite *Plasmodium falciparum* and the GC-rich halophilic archaeon *Hfx. volcanii* (Kaplan et al. 2009; Bartfai et al. 2010; Ammar et al. 2012). I observed improved performance in the predicted occupancy when compared to other published occupancy models, including those trained on GC content with a simple moving average (SMA), dinucleotide frequency or tetramer frequency data (Table 3-2) (Field et al. 2008; Kaplan et al. 2009; Tillo and Hughes 2009). In particular, the model performed better on the newest sequence data for yeast, but slightly poorer on the Kaplan et al. (2009) *in vivo* and *in vitro* data when compared to the Kaplan et al. and Tillo et al. models. This is likely because these two models were initially trained on the Kaplan et al. (2009) *in vitro* data and evaluated against the *in vivo* data. Our method outperforms the other predictors with the *P. falciparum* genome and the *Hfx. volcanii* genome, even with the vast genomic GC content differences between these organisms ranging from 19 - 65% (Table 3-2). The Tillo et al (2009) model performed slightly better that the Kaplan et al. and Field et al models with the *Hfx. volcanii* data because it considers GC content and weights it most heavily when predicting occupancy. Overall, our method performs consistently well when compared to all 5 experimentally derived nucleosome maps and can function well as a universal predictor that is not overfit to any particular data set.

		S. cerevisiae (Van Bakel et al. *In review*)	*P. falciparum* (Bartfai et al. 2010)	*Hfx. volcanii* (Ammar et al. 2012)	*S. cerevisiae* (Kaplan et al. 2009)	Chicken histones reconstituted onto *S. cerevisiae* genomic DNA *in vitro* (Kaplan et al. 2009)
Genomic GC content (%)		38.2	19.4	65.5	38.2	38.2
Model	**Method**	**Performance (Spearman's correlation coefficient, *ρ*)**				
Guassian GC smoothing	Convolution of Gaussian kernel across GC content. No probabilistic framework.	**0.738**	**0.740**	**0.725**	0.564	0.710
Lasso model (Tillo and Hughes 2009)	Linear regression model trained on *in vitro* yeast nucleosome occupancy data with primary weighting on 150bp simple moving average-smoothed GC content.	0.670	0.640	0.511	**0.603**	**0.778**
Kaplan et al., 2009	Probabilistic framework using periodic dinucleotide frequency and occupancy of all nucleosome 5-mers, trained on *in vitro* yeast data.	0.706	0.648	0.339	0.594	0.767
Field et al., 2008	Probabilistic framework using periodic dinucleotide frequency and occupancy of all nucleosome 5-mers, trained on *in vivo* yeast data.	0.650	0.645	0.296	0.549	0.636

Table 3-2. Contrasting nucleosome occupancy prediction models across different organisms. Coefficients represent average performance within the specified data set. Bold coefficients indicate the model that performed best.

3.3 Discussion

This study establishes that nucleosome occupancy is conserved between archaea and eukaryotes (Fig. 3-6). I further show that the nucleosomal protected fragments and NDRs are shorter in archaea than in eukaryotes. These findings are particularly noteworthy because *Hfx. volcanii* likely resembles a deeply rooted ancestor that possessed eukaryotic genome architecture hallmarks such as histones, as well as bacterial hallmarks such as the Shine-Dalgarno sequence (Sartorius-Neef and Pfeifer 2004). Archaeal histone tetramers likely resemble an ancestral state of chromatin, as it has been observed that functional $(H3\text{-}H4)_2$ tetramers can be formed *in vitro* from eukaryotic histones, and these tetramers are functional; they facilitate more rapid transcription *in vitro* compared to native histone octamers(Puerta et al. 1993). The observation that archaea contain $(H3\text{-}H4)_2$ tetramers is consistent with the proposal that formation of the canonical eukaryotic nucleosome octamer begins with $(H3\text{-}H4)_2$ tetramer assembly (Talbert and Henikoff 2010).

This study demonstrates that both histones and chromatin architecture arose before the divergence of Archaea and Eukarya, suggesting that the fundamental role of chromatin in the regulation of gene expression is ancient. As well, owing to the small bacterial-sized archaeal genome, I suggest the primary function of archaeal chromatin is for gene regulation and not for genome compaction. This leads one to postulate that higher-order chromatin (Sajan and Hawkins 2012) is a eukaryotic invention and that archaeal chromatin is necessary but not sufficient for genome compaction. Additionally, these observations provide a rich dataset that addresses the evolution of chromatin and its fundamental role in the regulation of gene expression.

.

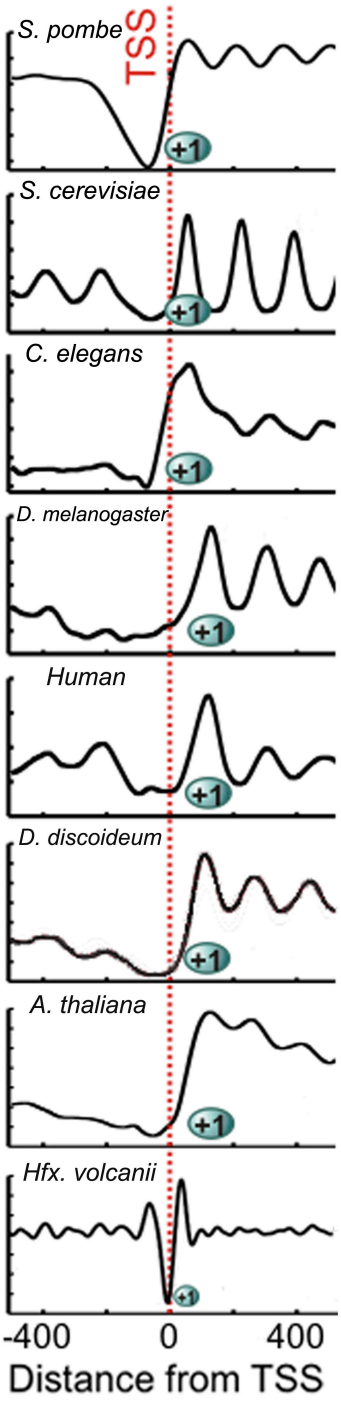


Figure 3-6. Chromatin architecture is conserved at the 5' end of transcripts across eukaryotes and archaea. Due to the smaller size of archaeal nucleosome DNA, the occupancy has a shorter periodicity. Figure adapted with permission from Chang *et al*. (2012).

A heated debate in the field of primary chromatin structure exists as to whether nucleosome positioning is directed by intrinsic DNA sequence preference or by chromatin-remodelling complexes and other proteins (Hughes et al. 2012). A survey of these studies and their respective data sets would instead suggest that both of these factors are vital in the genome-wide positioning of nucleosomes with clear roles for GC depletion/enrichment at promoter sequences, GC preference at nucleosome dyads and essential roles for chromatin remodelling enzymes. I demonstrate that nucleosomes are positioned along genomic DNA in a periodic manner most resembling a Gaussian distribution of G/C nucleotides with the nucleosome dyad at the mean and a standard deviation proportionate to the protected fragment length (see Materials and Methods). I observe that this model is valid for both eukaryotes and archaea (Table 3-2), suggesting that nucleosome occupancy evolved based on the DNA-directed positioning of the core histone $(H3\text{-}H4)_2$ tetramers, which exist in both eukaryotes and archaea (Thastrom et al. 2004; Talbert and Henikoff 2010). While previous models for nucleosome occupancy prediction perform reasonably well, they appear to be overfit to their training data and perform poorly when GC content varies significantly ranging from 19% for *P. falciparum* to 65% for *H. volcanii*. Since my model is not trained on any specific data set, this algorithm cannot overfit any particular data. In addition, other methods are based on frequencies of specific sequences in yeast, which can vary across organisms (Kaplan et al. 2009; Tillo and Hughes 2009). Interestingly, all models performed relatively poorly on the Kaplan et al. (2009) *in vivo* occupancy data set, suggesting that these data are less accurate than the newer eukaryotic data from Van Bakel et al. (*in review*) and Bartfai et al. (2010). I also highlight the observation that nucleosome occupancy is universally governed by GC content even when global genomic GC content is high (65%) or low (19%). I anticipate that this predictor may aid in the discovery of TSSs in existing and newly-sequenced genomes, and expand our understanding of chromatin-based gene regulation. The identification of TSSs can be useful in improving gene-finding algorithms.

3.4 Materials and Methods

Sample preparation. Haloferax volcanii DS2 cells (obtained from the ATCC) were grown to mid-log phase at 42°C in ATCC 974 Halobacterium medium containing with 2M NaCl. Cells were fixed with 2% formaldehyde for 30 min then quenched with 125mM of glycine for 5 min. An

unfixed control sample was also prepared to serve as a deproteinized, "naked" DNA control, as described previously (Chung et al. 2010). Cells were pelleted and snap frozen prior to MNase digestion and DNA extraction. Frozen cells were processed according to a modified protocol from Rizzo *et al.* (Rizzo et al. 2011; Tsui et al. 2012). Samples were digested with increasing concentrations of MNase and a no MNase control. After digestion, fragments 50-60bp in length were size-selected using an Agilent Bioanalyzer High Sensitivity chip (Agilent, part# 5067-4626) and further processed for Illumina deep sequencing. This size-selection was critical, as the formaldehyde crosslinking causes both histones as well as other DNA-binding proteins to crosslink with bound DNA. Nucleosomal and genomic libraries were pooled equally according to qPCR quantitation, and sequenced using v3 chemistry on one single-read HiSeq2000 lane (50x8). Samples were demultiplexed using an 8bp index read at the end of read 1.

Sequence read filtering and alignment. Illumina sequencers require the ligation of an adapter oligonucleotide to facilitate cluster formation on the flow cell. Because the library inserts were short (~60bp), many sequence reads extended into the Illumina adapter sequences. The adapter subsequences were computationally trimmed to ensure maximal read mapping. Then, using a sequence quality cutoff of Phred20, reads were trimmed from both 5' and 3' ends to ensure accurate mapping. These trimmed reads from control and MNase-treated genomic DNA were aligned to the *Hfx. volcanii* DS2 genome using the Bowtie 2 gapped short read aligner (Langmead and Salzberg 2012). Sequence coverage was computed using the Genome Analysis Toolkit (GATK) depth of coverage walker, which revealed the periodicity in the occupancy data (DePristo et al. 2011).

Nucleosome identification. To detect nucleosome midpoint positions, sequence data were Gaussian-smoothed as described previously by Shivaswamy *et al.* (2008) and Kaplan *et al.* (2009) (Shivaswamy et al. 2008; Kaplan et al. 2009). This is appropriate because signals generated by processes that are random, such as sequence coverage noise, usually have a probability density function defined by a Gaussian distribution (Smith 1997).

The Gaussian filter was defined as:

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)}$$

where μ is the mean of the distribution and σ is the standard deviation.

A symmetrical convolution sum was applied with the following format:

$$y[i] = \sum_{j=-\frac{M}{2}}^{\frac{M}{2}} h[j] \cdot x[i-j]$$

where M is an integer bandwidth, $y[j]$ is the output, $x[j]$ is the input and $h[j]$ is an M-point function.

So, to smooth the coverage data, we applied the following convolution sum:

$$y[i] = \sum_{j=-\frac{M}{2}}^{\frac{M}{2}} G[j] \cdot x[i-j]$$

where $\sigma = \frac{M}{6}$. The interval length M is constrained to 6σ because this encompasses 99.75% of the Gaussian (Smith 1997).

We also optimized nucleosome midpoint detection by convoluting a 2-pass simple moving average (SMA) filter, but the Gaussian filter detected midpoints with greater resolution. Optimal interval size for the Gaussian convolution sum, as determined by Pearson's correlation coefficient with the raw data, was 27bp. For the 2-pass SMA it was 40bp for first-pass and 15bp for second-pass.

Nucleosome occupancy was normalized genome-wide by transforming sequence coverage data into binary-like data that existed in states of "occupied", "depleted" or transitioning between those two states. This final occupancy map was used to define nucleosome positions. Nucleosome occupancy profiles were clustered hierarchically by average linkage using Pearson's correlation coefficient as the similarity metric in the Cluster 3.0 software package. Clusters were visualized with Java Treeview (Fig. 3-3b).

Transcript identification and genome annotation. RNA was extracted with Trizol reagent (Invitrogen, 15596-026), and DNase treated (Invitrogen, AM1907) according to manufacturer specifications. A cDNA library was generated using 100ng of total RNA according to Illumina TruSeq RNA Sample Prep protocol (Illumina, RS-122-2001) prior to duplex-specific nuclease (DSN) treatment. 100ng of cDNA library was incubated in hybridization buffer (50mM HEPES,

500mM NaCl) for 2 minutes at 98°C, followed by 1 hour at 68°C. Ribosomal RNA (rRNA) was not specifically depleted (He et al. 2010). Instead, we used duplex-specific nuclease (DSN) normalization to remove recurrent RNA (rRNA, tRNA) from the total RNA sample, thereby enriching mRNA (Zhulidov et al. 2004). Samples were immediately treated with 4 units of DSN enzyme (Evrogen, EA001) in 1X DSN buffer and incubated for an additional 25 minutes at 68°C, prior to addition of stop solution, and purification with Ampure XP beads (Beckman Coulter, A63881). RNA libraries were pooled equally according to qPCR quantitation, and sequenced using v3 chemistry on a paired-end single HiSeq2000 lane (100x8x100). Samples were demultiplexed using an 8bp index read at the end of read 1. Total RNA was sequenced at extremely high coverage (2587× mean coverage) so that rRNA sequences (~77% of all sequence reads) could be computationally excluded, as described by Wurtzel *et al.* (Wurtzel et al. 2010).

After quality score trimming (described earlier), sequence reads were aligned using TopHat (Trapnell et al. 2009). The RNA-seq data displayed a great deal of overlap with the predicted annotations(Hartman et al. 2010), with 92.1% of the existing annotations being confirmed. Of the 4073 predicted annotations, 3751 were confirmed, and, of these, 744 were merged with other transcripts to form longer transcripts. A heuristic approach was applied to adjust the transcript 5' and 3' positions of the Hartman *et al.* predicted annotations based on the boundaries of high RNA-seq coverage regions. This was vital as TSS accuracy is of great importance for NDR identification (Fig. 3-7).

Because 85% of the *Haloferax* genome is predicted to be coding (Hartman et al. 2010), transcript detection is complicated by transcript overlap. To overcome this, computationally identified transcripts were manually curated yielding a total of 3059 expressed transcripts in *Hfx. volcanii*. Of these, 32 transcripts are novel (Table 3-1). Of these transcripts, NTRANS_0004 was the most abundant transcript in the transcriptome, excluding the 6 rRNA genes. Homology data was obtained using BLASTX with a BLOSUM45 matrix against the non-redundant protein sequence database (Altschul et al. 1990). Conserved domains were identified using the Conserved Domain Database (Marchler-Bauer et al. 2011).

Data availability. Sequence data, nucleosome, transcriptome maps and supplemental tables are available at http://chemogenomics.med.utoronto.ca/supplemental/chromatin/. The sequence data has been deposited to the NCBI Short Read Archive with the accession SRA057981.

Nucleosome occupancy source data. Data for the performance analyses were retrieved from the NCBI's Gene Expression Omnibus (GEO) and Short Read Archive (SRA) databases. The Kaplan et al (2009) pre-normalized *in vivo* and *in vitro* occupancy datasets were obtained from GEO series accession GSE13622. These occupancy data corresponded to DNA sequence from release 53 of the *S. cerevisiae* S288c genome (Kaplan et al. 2009; Cherry et al. 2012). Occupancy data from Van Bakel et al (*In review*) was also obtained as pre-normalized data (data obtained from the authors). Since these yeast nucleosome data were obtained using reads shorter than 75bp, the nucleosomal centre positions were inferred, and, for this reason, we used the pre-normalized data. *P. falciparum* 3D7 nucleosome occupancy data was available at NCBI GEO accession GSE23787 as single-end 75bp read data, which required no inference of centre position (Bartfai et al. 2010). These data were aligned using Bowtie2 to the current release of the *P. falciparum* 3D7 genome (PlasmoDB 9.0) and further processed with Picard (http://picard.sourceforge.net) and the Genome Analysis Toolkit (GATK) software packages to determine sequence coverage across the genome (Aurrecoechea et al. 2009; McKenna et al. 2010; Langmead and Salzberg 2012). Periodic sequence coverage corresponded to the nucleosome map. This approach was also applied to the nucleosomal fragment reads from *Hfx. volcanii* DS2 from NCBI SRA accession SRA057981 (Ammar et al. 2012), which were aligned to the current *Hfx. volcanii* genome (Hartman et al. 2010).

The reference genomes that were paired with the occupancy data were the same genomes used to predict occupancy with each predictor (even if a newer genome release was available), so that the occupancy curves were congruent.

Nucleosome occupancy prediction. The GC-based predictor determines the presence of a cytosine or guanine nucleotide at each position in the genome and converts the genome into binary sequence data (G/C = 1, A/T = 0). Given our findings that GC frequency is enriched at nucleosome midpoints with a Gaussian distribution, we applied a Gaussian convolution sum to the binary GC data to obtain a smoothed GC curve. The Gaussian kernel was defined as

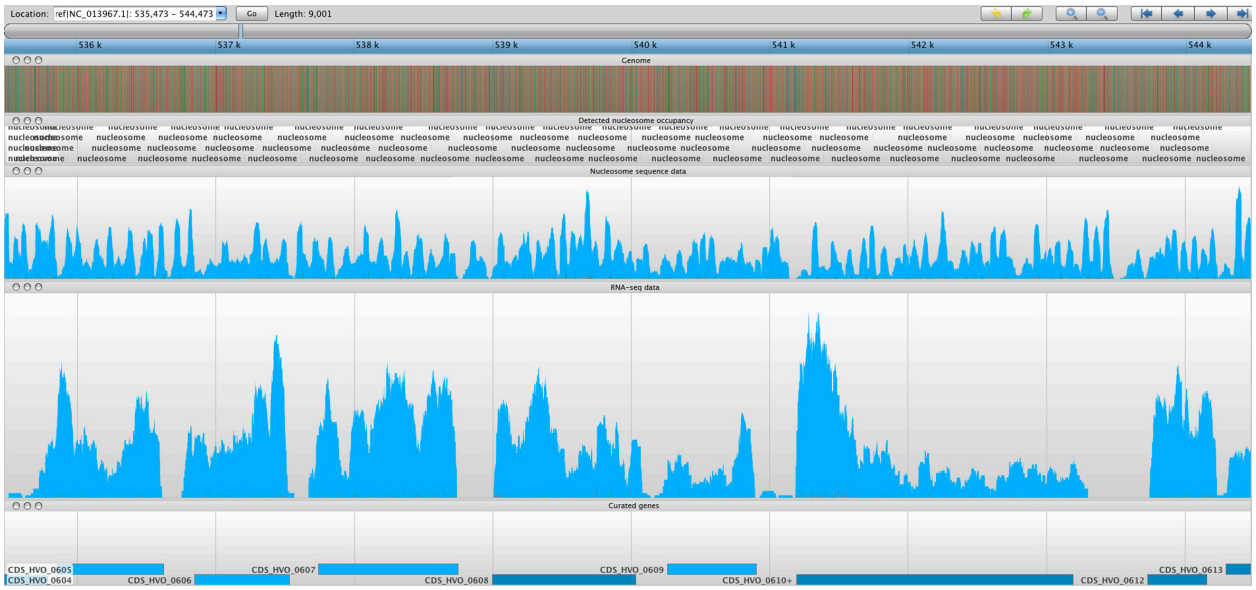


Figure 3-7. Sample screenshot of all data tracks loaded into the Savant genome browser (Fiume et al. 2010). The nucleosome sequence data is displayed, and the periodicity reflects protected and unprotected fragments after MNase digestion (magnitude of peak is not considered). Peaks represent nucleosome midpoints, which were detected and marked. Below are the corresponding RNA-seq and curated gene tracks. In this screenshot, one can observe seven entire ORFs in line with their NDRs and –1 and +1 nucleosomes.

described above in the methods subsection "nucleosome identification". The convoluted sum output was used as a direct measure of nucleosome occupancy.

Predictor performance. Predictive performance was benchmarked using sequence-based mono-nucleosome occupancy data instead of array-based data because, like the smoothed GC curve, sequence reads offer continuous occupancy coverage, whereas arrays do not (continuous periodicity in array data is interpolated from tiled probes).

The predictor script *gc_nucleosome_predictor.py* is available for download at http://baderlab.org/Software/nucleosome-prediction, and operates on Unix platforms. Based on our optimizations, best performing occupancy predictions can be achieved in eukaryotes or archaea when choosing a Gaussian interval size 12 bp shorter than the estimated protected fragment length. In our optimizations, the best performance was obtained at 48bp for *Hfx. volcanii* (~60bp protected fragment) and 138bp for *S. cerevisiae* (~150bp protected fragment). After scanning the interval space near the known protected fragment size, we found that the predictor is not sensitive to variation in the interval (Fig. 3-5).

4 Tools for identifying human drug targets

Modern drug discovery methods have been developed for the challenging endeavor of identifying candidate drug targets and target pathways. Here I report a novel assay to identify human drug targets, called human Multi-copy Suppression Profiling (hMSP). This assay is optimized for use with a custom DNA microarray platform that was designed for high-throughput ORFeome studies and can be adapted for a massively parallel sequencing readout. In hMSP, a collection of *Saccharomyces cerevisiae* strains, each expressing a different human protein, is exposed *en masse* to a drug at growth inhibitory concentrations. Strains that are resistant to drug inhibition are prioritized as candidate drug targets. I confirmed the utility of hMSP by identifying human dihydrofolate reductase (DHFR) as the target of methotrexate. hMSP was then applied to study the β adrenergic receptor (βAR) antagonist propranolol, a drug used in the treatment of cardiovascular diseases. Propranolol has been shown to be an effective and safe therapeutic via a mechanism independent of its βAR activity. I identified yeast strains overexpressing the dual specificity phosphatases (DUSPs) 10 and 16 as resistant to propranolol treatment. These observations were confirmed in human embryonic kidney cells, where the overexpression of DUSP16 was able to rescue the cells from propranolol toxicity. Propranolol was subsequently shown to inhibit DUSP10 activity *in vitro*, suggesting that these dual-specificity phosphatases are bona fide targets of propranolol. This study emphasizes the value of *in vivo* chemical genomic assays in yeast for the purpose of identifying novel human drug targets in a rapid and unbiased manner.

Portions of this chapter have been adapted from the following publications:

1. Ammar, R.*, Smith, A. M.*, Heisler, L. E., Giaever, G. & Nislow, C. 2009. A comparative analysis of DNA barcode microarray feature size. *BMC Genomics*, 10, 471.

   This work has been adapted under the BioMed Central Open Access license agreement.

   Author contributions:
   RA, AMS, GG, CN conceived of the project and designed experiments. RA and AMS performed the experiments. RA analyzed the data. RA, AMS, GG, CN wrote the paper.

2. Smith, A. M.*, Ammar, R.*, Nislow, C. & Giaever, G. 2010. A survey of yeast genomic assays for drug and target discovery. *Pharmacol Ther,* 127**,** 156-64.

   This work has been adapted with permission from Elsevier Ltd.

3. Ketela, T.*, Heisler, L. E.*, Brown, K. R., Ammar, R., Kasimer, D., Surendra, A., Ericson, E., Blakely, K., Karamboulas, D., Smith, A. M., Durbic, T., Arnoldo, A., Cheung-Ong, K., Koh, J. L., Gopal, S., Cowley, G. S., Yang, X., Grenier, J. K., Giaever, G., Root, D. E., Moffat, J. & Nislow, C. 2011. A comprehensive platform for highly multiplexed mammalian functional genetic screens. *BMC Genomics,* 12**,** 213.

   This work has been adapted under the BioMed Central Open Access license agreement.

   * denotes equal contribution

4.1 Introduction to yeast chemical genomics

4.1.1 A survey of yeast genomic assays for drug and target discovery

Current approaches to drug discovery are typically target-oriented, making use of validated targets as the starting point for discovery and development efforts. Typically, promising targets are selected based on several criteria including: 1) prior knowledge of a target's biological role(s) and potential for therapeutic intervention 2) proven value based on approved drugs (so-called "me too" targets) 3) a target's essentiality for cell growth and 4) druggability (Hopkins and Groom 2002). As a consequence of these constraining criteria, the selection of targets is biased toward well-characterized proteins or pathways. Once a target has been selected in this manner, biochemical assays are developed such that the target can be screened in a high-throughput assay. Because these assays are performed *in vitro* using purified components, once a lead compound is introduced into the context of the cell, the contributions of other potential protein-compound interactions are unpredictable.

During the past two decades, target-based approaches to drug discovery have produced novel lead compounds and therapeutic candidates, however, the overall approval rate for new chemical entities has remained relatively flat despite skyrocketing research development costs (Higgins and Graham 2009). Due in part to this lack of increased productivity, cell-based phenotypic screens have gained renewed interest. Advantages of cell-based screens include 1) identified compounds are cell-permeable and 2) sophisticated tools are available to screen a wide range of desired phenotypes. However, a major challenge remains; once a compound producing the desired phenotype is identified, the cellular target of the compound is yet to be determined (Chan et al. 2009). New technologies and experimental approaches for identifying drug targets have been recently developed including *in silico* docking approaches (Teotico et al. 2009), computational predictions (Keiser et al. 2009; Song et al. 2009), novel compound derivation strategies (Schreiber 2000; Stockwell 2004), chemical proteomics (Rix and Superti-Furga 2009) and many others which have been the subject of several recent reviews (Butcher et al. 2004; Li and Vederas 2009; Mandal et al. 2009; Quon and Kassner 2009; Wagner and Clemons 2009). However, these approaches are not yet amenable to genome-wide approaches to identify targets

in vivo. Here, I focus on the *in vivo* chemical genomic assays developed in the yeast *Saccharomyces cerevisiae* as they are currently the only tools available that allow the relative sensitivities of all potential drug targets to be simultaneously measured, leading to identification of the most likely compound/drug target candidates.

The model organism *Saccharomyces cerevisiae* has proved an effective test bed for the development of virtually all "omics" techniques (Bader et al. 2003; Sidhu et al. 2003; Provart and McCourt 2004; Rual et al. 2004a; Costanzo et al. 2006; Dixon et al. 2009; Snyder and Gallagher 2009). The *S. cerevisae* genome is one of most well-characterized (Pena-Castillo and Hughes 2007) in part due to its rapid generation time, inexpensive cultivation and facile genetics. Molecular genetic efforts have led to the generation of a complete molecular-barcoded gene deletion collection (Winzeler et al. 1999; Giaever et al. 2002). Because of these experimental attributes, *S. cerevisiae* will continue to be a major player in biological studies aimed at understanding specific proteins and pathways that can be modulated to ameliorate disease (Dixon and Stockwell 2009). Yeast can also be used to model processes in metazoans, e.g. approximately 45% of the genes in yeast are homologous to mammalian genes (BLAST e-value $<10^{-10}$) (Hughes 2002), encouraging efforts aimed at translating assays and results from yeast to metazoans (Chervitz et al. 1998).

Despite its numerous advantages, yeast assays are not without limitations for the purposes of drug discovery. Chief among these is the high concentration of compound often required to produce a biological response, likely due to the barrier presented by the cell wall, and the presence of numerous active efflux pumps and detoxification mechanisms (Leppert et al. 1990; Wehner et al. 1993; Miyahara et al. 1996; Molin et al. 2003; Cowen and Steinbach 2008). In addition, although many core processes are conserved between yeast and human, several "metazoan-specific" processes are not. Nonetheless, a number of labs have designed clever screens to study processes such as neurodegeneration (Miyano 2005), diabetes (Kohlwein 2010), and angiogenesis (McGary et al. 2010) in yeast models.

4.1.1.1 Drug-induced HaploInsufficiency Profiling (HIP)

The Yeast KnockOut (YKO) collection consists of a complete set of deletion strains, including haploid strains of both yeast mating types and heterozygous and homozygous diploid deletions. Each strain carries a precise start to stop deletion of a single gene (Winzeler et al. 1999; Giaever et al. 2002). A key feature of these collections is that each deletion strain is tagged or "barcoded" with two unique 20 base pair sequences that serve as strain identifiers. These collections can be pooled and grown competitively in any condition of choice which allows the identification of genes most important for growth in a given condition (e.g. compound/drug treatment) because strains carrying deletions of these genes will become depleted from the pool over time. The relative abundance of each strain is measured by the abundance of the barcodes. Specifically, following pooled cell growth, genomic DNA is extracted from cells, barcode PCR amplified using the primers common to every strain, and relative strain abundance quantified based on hybridization signal from a DNA barcode microarray (TAG4 microarray; Affymetrix part no. 511331) containing the barcode complements (Winzeler et al. 1999; Giaever et al. 2004; Pierce et al. 2006) (Fig. 4-1a). Alternatively, barcodes can be detected by next-generation sequencing (Smith et al. 2009). Barcodes that decrease in abundance over the time course of the experiment versus the control identify strains deleted for genes required for survival in the tested condition.

Drug-induced HaploInsuffiency Profiling or HIP was one of the first assays to take advantage of this parallelized growth strategy. HIP is based on the observation that a heterozygous deletion strain is specifically sensitized to a drug that targets the product of the heterozygous locus (as measured by a decrease in growth rate or fitness) (Giaever et al. 1999). By screening all possible heterozygous deletion strains in parallel, the heterozygous deletion strain most sensitive to a particular drug often identifies the drug target(s) (Giaever et al. 1999; Giaever et al. 2004; Lum et al. 2004). A key advantage of this assay is that it simultaneously identifies both the inhibitory compound and candidate targets without prior knowledge of either. These candidate targets represent those genes most important for growth and are therefore relevant for identification of antiproliferative targets that may have potential as either antifungal or oncology targets. The feasibility and robustness of this assay has been demonstrated by screening well-characterized and novel compounds (Giaever et al. 1999; Giaever et al. 2004; Lum et al. 2004; Pierce et al. 2007; St Onge et al. 2007; Hillenmeyer et al. 2008; Yan et al. 2008; Smith et al. 2009). In addition, such

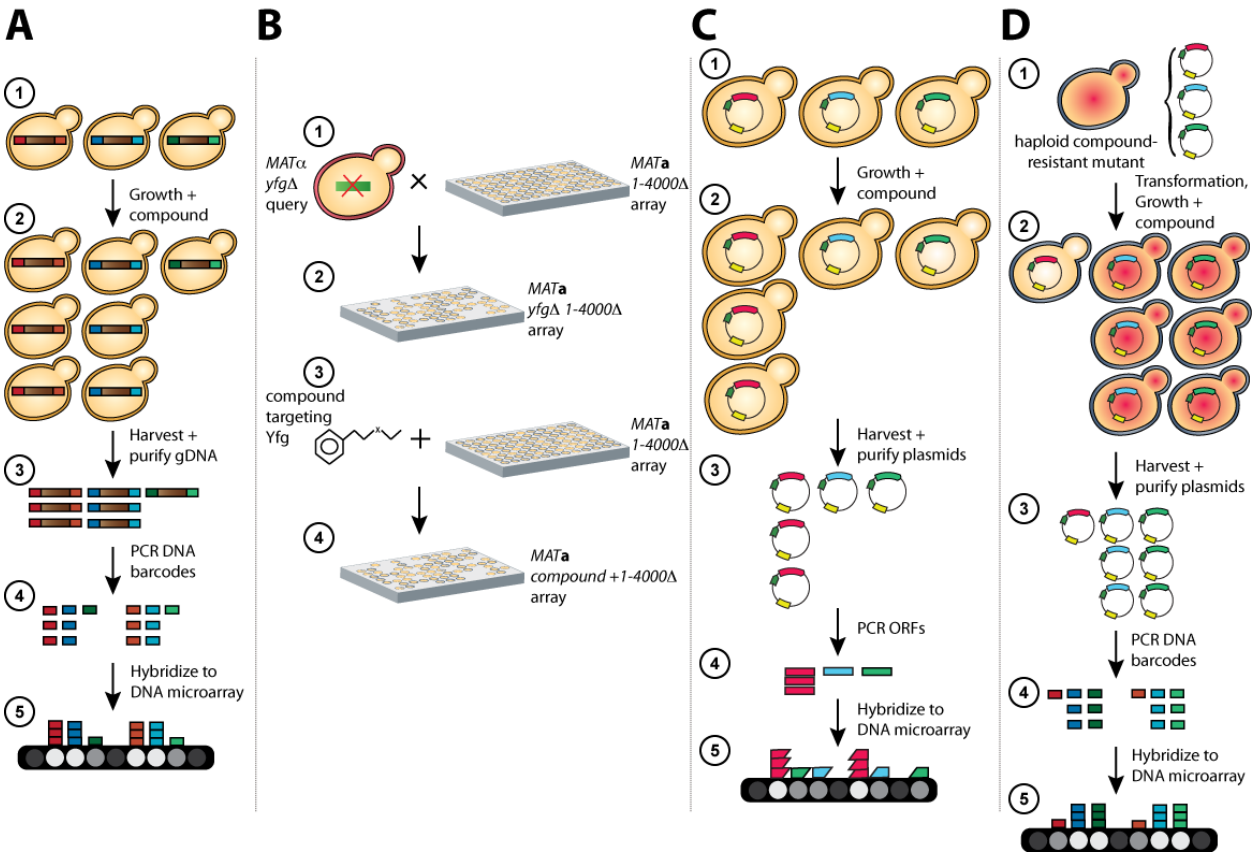


Figure 4-1. (**A**) HIPHOP assay: Yeast deletion collection is pooled with each strain at approximately equal abundance (1). Pool is grown competitively in a compound of choice (2), Genomic DNA is isolated from the treated sample (3) and barcodes are PCR amplified (4). The PCR product is hybridized to a barcode microarray to assess relative abundance of each strain by hybridization intensity. (**B**) Comparison of genetic interactions and compound-gene interactions: A query strain consisting of a mutation in Your Favourite Gene (*yfgΔ*) is crossed into an array of ~4000 non-essential deletion strains using the Synthetic Genetic Array (SGA) protocol (1). Resulting double mutant haploid progeny are selected on plates, and colonies reduced in sized represent genetic interactions (2). The array of ~4000 strains is pinned onto plates containing drug targeting Yfg (3) and colony size is used to identify deletion mutants sensitive to compound (4). (**C**) Multi-copy suppression profiling: An ORFeome library is transformed *en masse*, into a wildtype yeast strain (1). The resulting pool is grown in a compound of choice (2), plasmid DNA is isolated (3) and inserts are amplified (4). Amplicons are then labelled and hybridized to a microarray carrying ORF-specific probes (5). Intensities that are increased on the microarray identify ORFs that confer drug resistance. (**D**) Complementation of compound-resistant mutants: A haploid drug-resistant strain is isolated and the resistant phenotype is confirmed to be recessive by crossing with a wildtype strain (1). Resistant strain is transformed with the MoBY-ORF library, and the resulting pool is grown in a high concentration of drug where strains that are sensitive, due to plasmid complementation, are depleted (2). Plasmid DNA is isolated (3), barcodes are amplified (4) and hybridized to a barcode microarray. Intensities that are significantly reduced contain the ORF responsible for drug resistance (5).

screens, when the direct target does not exist in yeast, provide insight into the off-target mechanism of action (Ericson et al. 2008).

An alternative to the competitive pooling approach is to pin the heterozygous (or other) yeast strain collections onto agar plates that contain a compound of interest and to monitor fitness based on colony size (Baetz et al. 2004; Carroll et al. 2009). A drawback of plate-based assays are that they typically require significantly more compound (1 to 2 orders of magnitude) versus pooled liquid assays. Nonetheless, recent results show that genetic interactions monitored in liquid media correlate well with interactions identified on solid media when using robotic technologies combined with data analysis (Costanzo et al. 2010).

Despite the successes of HIP, reducing gene copy by half (in heterozygotes) may be insufficient to identify well-characterized drug targets may be required. To address this issue, Yan *et al.* (2008) barcoded a yeast allele collection of haploid essential gene mutants to create a set of DAmP (Decreased Abundance by mRNA Perturbation) where a drug resistance marker is placed upstream of the 3' UTR (UnTranslated Region) of each gene. These DAmP truncations or strains have been shown to express, on average, about 10% of the wildtype protein levels (Schuldiner et al. 2005). This collection of hypomorphic alleles was able to detect drug-induced haploinsufficiency not observed in the heterozygote case (Yan et al. 2008), thereby widening the scope of the assay to identify compound target candidates. Because the collections of heterozygote and DAmP essential alleles carry non-overlapping barcode sequences, both assays can be performed in parallel (and hybridized to the same microarray), resulting in an increase in the dynamic range and sensitivity of the pooled assays. However, if the known target requires a dosage decrease beyond that of the heterozygous strains and if other heterozygous strains are identified as conferring sensitivity this indicates that the target that is well-characterized does not represent the primary mechanisms of drug action.

Haploinsufficiency Profiling is a timely and powerful approach, particularly in light of recent studies that have made it apparent that few drugs target single gene products (e.g. imatinib (Gleevec) (Buchdunger et al. 1996; Druker et al. 1996)), therefore, an *in vivo* view of the relative sensitivity of all targets in the cell is invaluable to understand the complete mechanism of drug action. Yeast cells are obviously not equivalent to human cells, and, therefore, any human targets

that lack a yeast homolog will not be identified. Moreover, the HIP assay relies on a growth phenotype resulting from drug/target binding; these targets will also not be identified. Despite these caveats, based on our screening of >2000 compounds, we have not yet failed to identify a target in yeast when that target is 1) well-characterized and 2) target inhibition impairs cell growth. We do observe off-target effects that likely reflect actual *in vivo* interactions. Another important caveat is that, decreasing gene dosage by a single copy may not be sufficient to reveal drug-induced haploinsufficiency for a particular target. In principle, further lowering gene dose may reveal the true target when simply raising the compound concentration would not, due to general cellular toxicity which could obscure the results. However, the failure to detect a target as a heterozygote (and only in a more severe DAmP or temperature-sensitive allele) may imply that the suspected/known target is actually not the major mechanism of action of a particular compound. For example, 5-FU is thought to act by inhibiting Cdc21. However, yeast lacks thymidine kinase and therefore the inhibition of Cdc21 can only occur indirectly through a series of metabolic interactions (Goodman et al. 2001). Indeed, genome-wide yeast assays reveal the primary mechanism of action is via misincorporation of fluorinated nucleotides into RNA(Scherf et al. 2000; Goodman et al. 2001; Giaever et al. 2004; Lum et al. 2004). Finally, targets that are either not essential and/or are highly redundant are unlikely to be detected in the loss of function assays because inhibition of all homologs would be required.

4.1.1.2 Homozyous Profiling (HOP)/Haploid deletion chemical-genetic profiling

Homozygous profiling (HOP) is analogous to the HIP assay, except that the strains are completely deleted for non-essential genes in either haploid or diploid strains. Relative growth rate, in the condition of choice (e.g. drug treatment), is measured by microarray signal intensity as described above.

In the HOP or haploid assays (Parsons et al. 2004; Lee et al. 2005; Parsons et al. 2006; Hillenmeyer et al. 2008), strains most sensitive to a drug become depleted from a pool over time, as in the HIP assay. However, because these strains carry complete deletions of non-essential genes they do not identify the target directly because the target is absent. Rather, these assays

identify genes that act to buffer the drug target pathway and are therefore required for growth in the presence of compound. This assay can be particularly informative for compounds that lack a direct protein target. For example, genes involved in the DNA damage response, while non-essential under standard growth conditions, are required for survival when challenged with DNA damaging agents (Birrell et al. 2001; Chang et al. 2002; Lee et al. 2005; Workman et al. 2006; Yu et al. 2008). One study (Lee et al., 2005) defined the relative importance of different DNA-repair modules for resistance to 12 DNA damaging agents and revealed functional interactions that comprise the DNA-damage response. While many of these compounds share similar mechanisms of actions (e.g. a subset were alkylating agents), each compound produced a unique genome-wide profile, or "signature". By screening a collection of compounds across non-essential genes, these genome-wide profiles can be clustered which allow one to infer the mechanism of action (Parsons et al. 2004; Parsons et al. 2006) when compared to those profiles obtained from drugs with well-characterized mechanisms (Fig. 4-1b). Like HIP, this assay can be performed either competitively in pools using barcode-based assays (Giaever et al. 1999; Giaever et al. 2004; Parsons et al. 2006; Pierce et al. 2007; St Onge et al. 2007; Hillenmeyer et al. 2008; Yan et al. 2008; Smith et al. 2009; Xu et al. 2009a) or on agar plates where drug sensitivity is measured by colony size (Parsons et al. 2004). A drawback of this approach in identifying the drug target is that markers that include gold-standard well-characterized drug-target relationships must be included to best infer the drug target.

A related approach to identifying drug-target interactions compares genome signatures by correlating HOP profiles with Synthetic Genetic Analysis (SGA) profiles of genetic interactions (Tong et al. 2001; Tong et al. 2004) where a conditionally essential gene is used as a query gene (Costanzo et al. 2010). When the query is essential, genetic interactions identified often correlate with non-essential deletion strains detected by HOP in the presence of drug, and the essential gene used as a query can be inferred to be the drug target. A recent example of the power of this approach identified Ero1 as the target of a novel small molecule (Costanzo et al., 2010). In a variation of this approach, Carroll *et al.* (2009) screened a yeast mutant collection to probe the mechanism of action of the yeast K28 toxin. In this screen, the inhibition of growth by secreted K28 toxin was monitored using a traditional halo assay to identify novel genes involved in cellular pathways essential for the response to this toxin (Carroll et al. 2009).

Because HIP and HOP assays are complementary, combining the results of both heterozygous and homozygous/haploid loss-of-function chemical genomic screens can be particularly powerful for understanding the mode-of-action (MOA) of compounds. A caveat of all HIP and HOP-based screens is that while these assays screen compounds against all potential targets simultaneously, definitive demonstration of a drug-target interaction requires independent confirmatory approaches such as *in vitro* binding or activity assays (Chan et al. 2009).

4.1.1.3 Multi-copy Suppression Profiling (MSP)

One approach to identify or confirm a drug-target interaction is to demonstrate that overexpression of the target *in vivo* confers resistance to drug (Rine et al. 1983; Li et al. 2004; Butcher et al. 2006; Hoon et al. 2008). In a feasibility study demonstrating that drug targets can be identified *de novo*, Rine *et al.* (1983) used a high copy plasmid carrying randomly generated yeast genomic inserts of approximately 5kb to identify genes that, when overexpressed, conferred resistance to tunicamycin when plated on solid media containing this compound. Plasmids were then isolated from resistant colonies and sequenced to identify *ALG7*, which encodes the known target of tunicamycin. This random genomic library approach has been miniaturized to use pools of strains in liquid culture screened in parallel in a manner analogous to the HIP assay (Hoon et al. 2008). Specifically, a high copy plasmid collection containing yeast genomic DNA fragments (including native promoters) is screened in yeast at high inhibitory concentrations of compounds (e.g. doses that inhibit wildtype yeast by ~90%). Strains are grown competitively in compound, such that only one or a few strains that confer resistance are selected from the population. Plasmids are then isolated from surviving cells, and inserts are amplified by PCR and hybridized to a DNA TAG4 microarray carrying probes complementary to each yeast open reading frame (ORF). Microarray signal intensities are mean normalized (Hoon et al. 2008) (Fig. 4-1c) and resistance scored by comparing strain abundance between drug treatment pools and untreated reference pool. This approach correctly identified Dfr1, Erg11 and Tor1 as the targets of methotrexate, fluconazole and rapamycin, respectively. One caveat when using this approach as currently described is that drug pumps or other "indirect" targets may dominate the set of strains resistant to compound. A overexpression library transformed into diverse drug pump resistant mutants can alleviate this challenge (Paulsen et al. 1998).

Several recently constructed libraries offer advantages over the traditional genomic DNA library used by Hoon *et al.* (2008). The Yeast Genome Tiling collection (Jones et al. 2008) contains overlapping fragments of the yeast genome (~10Kb in size) cloned into high-copy 2μ vectors. The ends of each insert of this library have been sequenced, and the plasmids organized in a tiling fashion across the yeast genome, ensuring near-saturation of 97.2% coverage of the yeast genome. However, because each insert contains multiple genes, once a resistant fragment is identified, the exact gene target must be subcloned and confirmed. Another library consists of 3,900 yeast strains, each carrying a plasmid containing a single yeast ORF under expression of the GAL1 promoter (available from http://www.hip.harvard.edu/). Butcher *et al.* (2006) performed a proof-of-principle experiment with this library using the immunosuppressant rapamycin and found that plasmids carrying the Target Of Rapamycin (*TOR*) genes were correctly identified as conferring the greatest level of resistance (Chiu et al. 1994; Butcher et al. 2006). Sopko *et al.* (2006) created a yeast library consisting of over 80% of all yeast ORFs, with gene expression controlled by a galactose-inducible promoter (Zhu et al. 2001), such that high-levels of expression can be obtained (Johnston 1987). A benefit of using an inducible system is that gene expression can be induced at specific times during the course of an experiment. On the other hand, galactose induction often does not accurately reflect endogenous gene expression levels, and overexpression can cause toxicity to the host cell (Sopko et al. 2006). Because PCR was used to create this library, the ORFs may contain PCR-induced mutations, a concern addressed with a new, fully sequenced library (Hu et al. 2007). Arguably, the ideal ORF library for MSP is the Molecular-Barcoded Yeast Open Reading Frame (MoBY-ORF) collection where each plasmid was constructed by PCR to include individual yeast ORFs flanked by their endogenous 5' and 3' UTRs, representing 90% of all yeast ORFs (Ho et al. 2009). Because these plasmids are CEN-based, copy number is low (1-3 copies/cell) and predictable (Apostol and Greer 1988), and thus does not generally suffer from overexpression toxicity. Each ORF in the MoBY collection is linked to the same two DNA barcodes associated with the corresponding deletion strain from the YKO permitting abundance measurements by microarray hybridization or sequencing (Ho et al. 2009).

A caveat (and occasional advantage) of these yeast clone banks is the identification of those genes that, when overexpressed, are toxic to yeast. At least 15% of yeast genes are toxic to the cell when

overexpressed, and follow-up analyses of these genes can be informative regarding how regulation of gene products alter cell physiology (Sopko et al. 2006). This cohort of toxic genes can be used as the starting point for chemical suppressor screens to identify compounds that suppress the toxicity and which, by extension, may interact with that toxic gene product. Chemical suppression has been successfully employed to find inhibitors of the bacteria *Pseudomonas aeruginosa* by overexpressing Pseudomonas genes in yeast (Arnoldo et al. 2008). Tugenreich *et al.* (2001), similarly identified human genes that, when expressed in yeast, result in a growth defect (Tugendreich et al. 2001). The authors selected a p38 overexpressing strain from their collection of "toxic" human genes to screen commercial libraries to identify chemical suppressors of the fitness defect which represent potential p38 inhibitors.

An additional application of the MoBY-ORF library is the identification of genes responsible for drug resistance in a recessive manner (Ho et al. 2009). After a drug resistant mutant strain is identified, it is transformed with the MoBY-ORF collection (Ho et al. 2009). Complementation by one or more wildtype alleles from the collection that restores drug-sensitivity identifies the gene(s) conferring drug resistance (Fig. 4-1d). The feasibility of the assay was demonstrated by identifying *fpr1* as a mutant resistant to the drug rapamycin (Heitman et al. 1991a; Sabatini et al. 1994). This assay complements the MSP assay in that it identifies drug targets in cases where the compound must interact with another protein to become toxic. In this case, rapamycin binds to Fpr1 to form a toxic complex, which, in turn, inhibits the Tor1 protein (Heitman et al. 1991a; Sabatini et al. 1994). Complementation of drug resistant alleles can be used to systematically uncover general and specific resistance mechanisms. For example, Ho *et al.* (2009) used MoBY-ORF complementation to identify an essential enzyme in the ergosterol biosynthesis pathway as resistant to the natural product theopalauamide. Subsequent confirmations indicated that theopalaumide binds to ergosterol, defining a novel class of sterol-binding compounds.

MSP is flexible in that it can be used with diverse genomes. For example, ORF libraries exist for other organisms, all of which can be cloned into yeast expression vectors and expressed in yeast. As a test case, a genomic DNA library from *Candida albicans* expressed in *Saccharomyces cerevisiae* was used to identify a *Candida* ortholog of *Saccharomyces cerevisiae* Glc7 (a type 1 protein phosphatase) as resistant to the inhibitor calyculin A (Hoon et al. 2008). Indeed, yeast

mutants have been rescued using human genes by several groups to identify human gene function by complementation (Mushegian et al. 1997; Tugendreich et al. 2001; Zhang et al. 2003; Osborn and Miller 2007).

4.1.2 The human ORFeome collection and human MSP

While yeast-mediated MSP designs have been successfully applied to identify known and novel drug targets, until now they have been restricted to yeast genes or open reading frames (ORFs). Pertinent to this, ORFeome libraries have been generated from various model organisms, including humans (Rual et al. 2004b; Lamesch et al. 2007; Yang et al. 2011). The human ORFeome is based on the Mammalian Gene Collection (MGC), a cDNA library containing clones of each human protein coding gene (currently spanning 92% of human cDNAs) (Gerhard et al. 2004; Temple et al. 2009). MGC cDNAs were used as templates, PCR-amplified from start to stop codons, flanked by Gateway® recombination sites and sequence-verified (Brasch et al. 2004; Yang et al. 2011). This collection is easily cloned into Gateway-compatible destination vectors for transformation into an organism of interest, such as yeast (Hartley et al. 2000; Alberti et al. 2007).

I describe a novel assay where a high copy plasmid collection containing human ORFs from the human ORFeome was transformed into *S. cerevisiae* for the purpose of human drug target identification. This yeast human ORFeome collection was used in a variation of MSP termed human MSP (hMSP). *S. cerevisiae* was a suitable system in which to study expression of human genes due to a relatively neutral cytosolic pH, facilitating formation of proper protein tertiary structure as well as the presence of some protein modification systems (Buckholz and Gleeson 1991; Young 1998). In addition to being amenable to study in yeast assays, such as two-hybrid assays, human proteins have been shown to rescue yeast mutants from growth defects (Mushegian et al. 1997; Osborn and Miller 2007). By creating a yeast chemical genomic assay that expresses human proteins, one can study the mechanisms of compounds whose primary targets do not exist in yeast, while benefitting from established robust chemical genomic strategies.

4.2 Results

4.2.1 A comparative analysis of DNA barcode microarray feature size

Having conceived of the yeast-based human MSP assay, we sought to construct a novel microarray framework for these and other human ORF and shRNA experiments. The microarray would have to contain multiple probes for each ORF in the human ORFeome, among other probes, in a small form factor in order to be cost effective when mass produced. As a result, I explored using smaller microarray "features". Microarrays are made up of thousands to millions of these microscopic features, clusters of identical oligonucleotide probes, which are used to detect hybridized gene products. The microarrays used for HIPHOP assays have gone through several iterations of development, beginning with a feature size of 103μm on the TAG1 microarray, which consisted of 20bp probes (Shoemaker et al. 1996; Giaever et al. 1999). The *S. cerevisiae* cassette was originally designed for detection using the TAG1 microarray, which used 20bp-long oligonucleotide probes. Current Affymetrix microarrays use up to 25bp probes to detect complementary DNA sequences, and this length is more appropriate for newer barcoded collections as it improves hybridization specificity and increases the number of resolvable potential barcodes (Xu et al. 2009b).

The features on these chips were subsequently miniaturized to 30μm and provided full deletion pool coverage on the TAG3 microarray (P/N 510318) (Giaever et al. 2002). The current TAG4 chips (P/N 511331) with 8μm feature sizes were designed for improved performance and affordability. This scheme omitted uninformative probes present on previous tag microarrays and added five replicates to report non-uniform hybridization and allow adjustment of intensities accordingly (Pierce et al. 2006). No smaller yeast deletion pool barcode microarray exists due to manufacturing size constraints, however, these barcode probes are also present on the 5μm yeast whole genome tiling array (S288c genome tiling microarray; P/N 520055) representing 0.25% of the total 6.5 million probes on this microarray (Juneau et al. 2007). The area of the features scale quadratically, such that the tiling array features at 5μm on a side correspond to 25μm^2, and TAG3 features at 30μm on a side correspond to 900μm^2, or 36 times the area of the tiling features. It is important to note that all microarrays have the same oligonucleotide probe density of approximately 4000 probes/μm^2 (personal communication with Affymetrix technical support).

All three microarray generations, the TAG3, TAG4 and *S. cerevisiae* whole genome tiling arrays, identified *ALG7* as the primary target of tunicamycin, as expected (Fig. 4-2). The tiling array also identified several other genes as additional potential targets. This list of targets includes *ADO1*, *FYV8*, *GET2*, *HAC1* and *IRE1*, all of which have been shown to be sensitive to tunicamycin when knocked out, as well as *BCK1*, a gene which has previously been shown to be resistant to tunicamycin when overexpressed (Cherry et al. 1997; Chen et al. 2005; Krause et al. 2008; Schuldiner et al. 2008; Tan et al. 2009). In particular, *ADO1* is a prime example of a gene deletion strain exhibiting increased sensitivity on the tiling array, since it is detected at a $\log_2$ ratio of 2.59 in the tiling array data, but at 0.50 and 0.66 in the TAG3 and TAG4 data, respectively. In addition to known sensitive strains, our screen identified *COP1* and *RER2*, which are involved in ER to Golgi vesicle-mediated transport (see Table 4-1 for summary of sensitive strains) (Sutterlin et al. 1997; Belgareh-Touze et al. 2003). As with most sensitive strains, these genes were detected at slightly higher levels on the tiling array than on the other microarray generations. The tiling array appears to have slightly higher variance in its $\log_2$ ratios than the other microarrays (standard deviation of 0.58 in tiling, compared to 0.37 and 0.43 in TAG4 and TAG3 microarrays, respectively). I determined this to be due to its increased sensitivity to hybridized barcode abundance since sometimes strains that appear sensitive on the tiling array, fall into the background signal of the other microarrays, as with *ADO1*. It is reassuring to observe both the primary target of tunicamycin and genes annotated as sensitive to tunicamycin in our results. Additionally, I also identified genes associated with the endoplasmic reticulum and involved in the unfolded protein response because tunicamycin promotes protein misfolding.

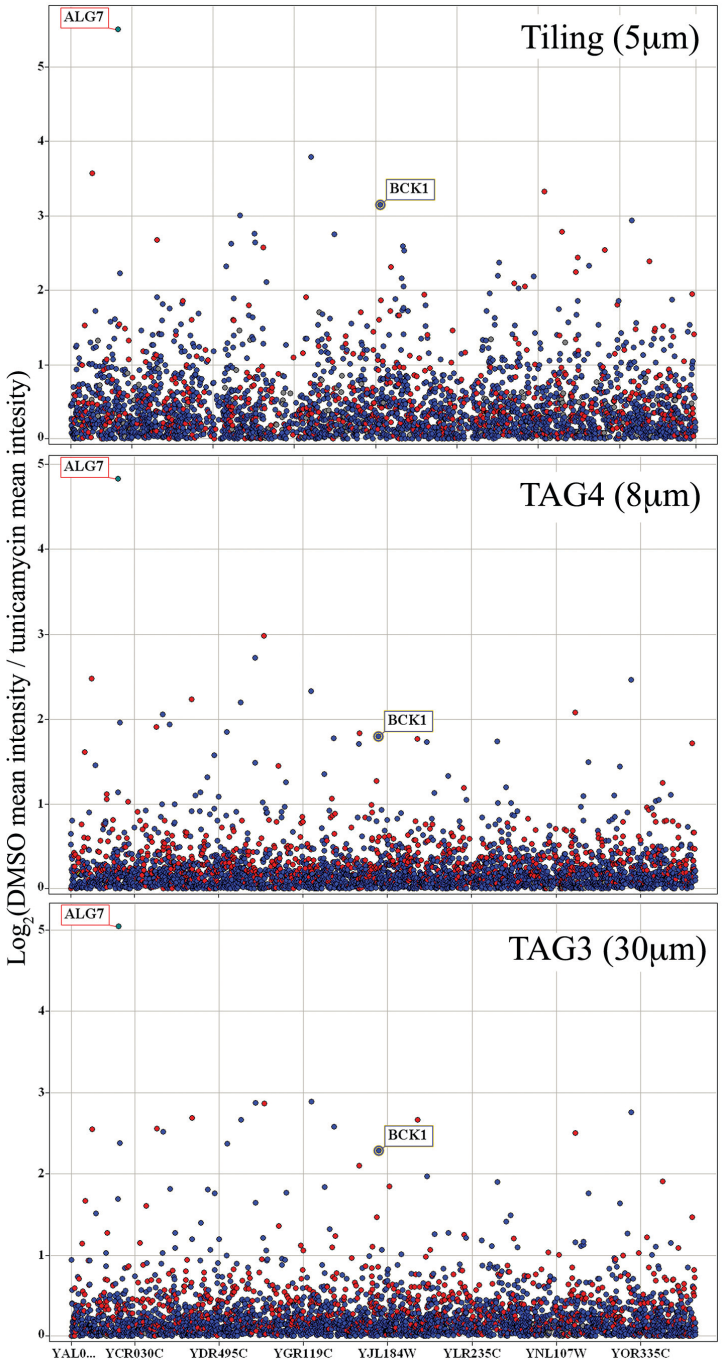


Figure 4-2. Identifying strains sensitive to tunicamycin on three microarray generations. Barcode intensity data are normalized according to a DMSO reference treatment. Blue dots represent non-essential genes, red dots represent essential genes and grey dots are genes that are not annotated. Log_2 ratios are calculated as a measure of change in barcode intensity (vertical axis) across all genes (horizontal axis). Ratios below 0 have been removed for clarity. Log_2 scales differ based on optimal dynamic range between baseline and *ALG7*. Higher ratios correspond to greater abundance of barcode from reference to treatment. In all three analyses, *ALG7* was correctly identified as the primary target of tunicamycin. Several additional genes previously determined to be resistant to tunicamycin, were most discernibly identified in the tiling data, but less so using TAG4 (the current microarray standard) and TAG3. These include *ADO1*, *BCK1*, *FYV8*, *GET2*, *HAC1* and *IRE1*. Furthermore, the genes *COP1* and *RER2*, known to be involved in ER to Golgi vesicle-mediated transport, showed up as sensitive to tunicamycin in our screen.

ORF	*Gene Name*	*GO Biological Process*	*tunicamycin treatment relevance*
YJR105W	*ADO1*	purine base metabolic process	knockout sensitive to tunicamycin (Tan et al. 2009)
YBR243C	*ALG7*	protein amino acid N-linked glycosylation *and others*	known target of tunicamycin (Barnes et al. 1984; Kukuruzinska and Robbins 1987; Kukuruzinska and Lennon 1995)
YJL095W	*BCK1*	endoplasmic reticulum unfolded protein response *and others*	knockout sensitive (Chen et al. 2005; Krause et al. 2008; Tan et al. 2009), overexpressor resistant (Chen et al. 2005) to tunicamycin
YDL145C	*COP1*	ER to Golgi vesicle-mediated transport *and others*	involved in ER to Golgi vesicle-mediated transport (Sutterlin et al. 1997)
YGR196C	*FYV8*	unknown	knockout sensitive to tunicamycin (Chen et al. 2005)
YER083C	*GET2*	protein insertion into ER membrane *and others*	knockout sensitive to tunicamycin (Schuldiner et al. 2008)
YFL031W	*HAC1*	specific RNA polymerase II transcription factor activity *and others*	knockout sensitive to tunicamycin (Chen et al. 2005; Tan et al. 2009)
YHR079C	*IRE1*	endoplasmic reticulum unfolded protein response *and others*	knockout sensitive to tunicamycin (Chen et al. 2005; Tan et al. 2009)
YOR246C	*N/A*	unknown	unknown
YFL032W	*N/A*	unknown	likely deletes *HAC1* promoter (Cherry et al. 1997)
YER010C	*N/A*	unknown	interacts with kinases Ptk2, Tpk1 (Ptacek et al. 2005)
YMR308C	*PSE1*	protein import into nucleus *and others*	interacts with Ulp1, regulating ubiquitination (Collins et al. 2007a)
YBR002C	*RER2*	ER to Golgi vesicle-mediated transport *and others*	involved in ER to Golgi vesicle-mediated transport (Belgareh-Touze et al. 2003)
YFR051C	*RET2*	ER to Golgi vesicle-mediated transport *and others*	interacts with Bre5, Hsc82, Hsp92 (Schuldiner et al. 2005; Collins et al. 2007a; McClellan et al. 2007) which are involved in protein processing
YNL151C	*RPC31*	transcription from RNA polymerase III promoter	interacts with Mms1, Shp1, Ubi4, regulating ubiquitination (Briand et al. 2001; Collins et al. 2007b)
YJR102C	*VPS25*	ubiquitin-dependent protein catabolic process via the multivesicular body sorting pathway *and others*	involved in ubiquitin-dependent protein catabolism (Bowers et al. 2004)

Table 4-1. Deletion strains sensitive to tunicamycin identified in the tiling array experiment.

Because the tiling array has millions of probes, only a few thousand of which are barcode probes, I hypothesized that non-specific hybridization of barcode DNA to the genome tiling probes could potentially contribute to noise in target identification. This may have been problematic because the tiling probes were not designed for explicit use with the barcode probes, which could lead to unanticipated cross-hybridization of barcode samples to tiling probe features. To determine if non-specific binding was a factor in our experiments, I co-hybridized barcode DNA with unlabelled digested genomic DNA (gDNA). The digested gDNA (20-150bp) competitively hybridized to tiling probes of the microarray to which barcodes may have had a non-specific affinity. I asked if the addition of gDNA could result in an increase of specific binding of barcodes to barcode probes, yielding a HIPHOP profile with greater dynamic range and more distinct targets (making the millions of tiling probes unavailable for barcode hybridization) analogous to the addition of salmon or herring sperm to a Southern blot to prevent non-specific hybridization (Wahl et al. 1979; Sambrook and Russell 2001). However, in practice, I found that the addition of gDNA did not improve resolution of the target *ALG7* when compared to a microarray without competitive gDNA co-hybridization.

Our initial experiments used protocols for each microarray that were optimized for that particular technology. For example, each microarray type has particular hybridization, washing and staining protocols. To minimize the effect of these subtle variations and to accurately compare intensity data across microarray generations, I hybridized a reference sample (treated with 2% DMSO) to TAG3, TAG4 and tiling microarrays and applied TAG4 wash protocols to each microarray type. The hybridization conditions were fixed so that I could be certain that any changes observed were attributed solely to feature size and not protocol variation. The microarrays were scanned following this protocol, and I subsequently applied the tiling array antibody stain wash step to all three chips and, once again, scanned them. In this manner, each microarray was treated identically. In general, the observed median downtag intensity was higher than median uptag intensity, an observation that was also reported by Pierce *et al* (Pierce et al. 2006; Pierce et al. 2007). In addition, the median intensities differed across generations, with TAG3 intensity lower than TAG4 intensity, which was lower than tiling intensity.

I found that TAG4 and tiling array intensities were very highly correlated ($r = 0.927$). This correlation increased slightly once the microarrays had been antibody stained during the tiling wash protocol ($r = 0.952$; Fig. 4-3). In contrast, TAG3 intensities did not correlate as well with either TAG4 or tiling ($r = 0.751$ and $r = 0.733$, respectively), and this decreased significantly after antibody staining ($r = 0.642$ and $r = 0.605$, respectively). However, this low correlation is unlikely to affect identification of drug targets on TAG3 microarrays, as these strains are often the most distinguishable from the background, as shown previously (Fig. 4-2).

The relatively recent design of the TAG4 microarray includes five replicates of each barcode probe (Pierce et al. 2006). However, I noticed that intensity values do not vary greatly between these replicates, and, therefore, a minimum of three replicates should be included to allow for appropriate trim mean calculations and masking of unusable barcode probes (Pierce et al. 2007). This finding confirms an earlier assertion by Pierce *et al*. that suggests that the minimum number of replicates required to achieve high correlation is three replicates, and that the increase in correlation from the fourth and fifth replicates is marginal (Pierce et al. 2006). Although the TAG3 and tiling results contain only single data points for each barcode and are able to determine *ALG7* as the primary target of tunicamycin (Fig. 4-2), replicate data points are advised to accommodate hybridization, washing and staining inconsistencies.

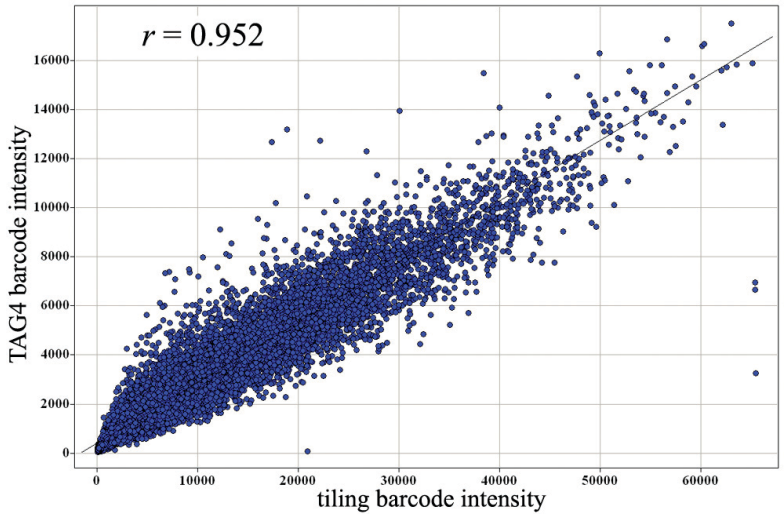


Figure 4-3. TAG4 and tiling array data correlation after antibody staining. This example shows that the signal intensity for common barcodes between TAG4 and tiling arrays are highly correlated (r = 0.952), demonstrating that tiling arrays are as accurate as TAG4 microarrays when determining relative signal intensity (compared to a DMSO reference on the same chip generation).

4.2.2 The Gene Modulation Array Platform

Having determined that the 5μM feature size was optimal for hybridization intensity and dynamic range, in collaboration with Jason Moffatt, we designed a custom Affymetrix microarray called the Gene Modulation Array Platform (GMAP). The GMAP facilitates cost-effective data collection from genome-scale pooled gene-dosage modulation screens performed in human, mouse, and yeast cells using commercially available libraries on a standard platform. Specifically, the GMAP enables readout of clone/strain enrichments and depletions from pooled screens using the RNAi consortium (TRC) human and mouse libraries (Moffat et al. 2006; Root et al. 2006; Luo et al. 2008), human ORF expression pools (Rual et al. 2004b; Lamesch et al. 2007), and pooled screens using gene deletion-associated barcodes or ORFs from budding yeast (Giaever et al. 2002; Hoon et al. 2008) (Table 4-2).

My specific contribution to the GMAP pertained to the design of human ORF probes for pooled gene-dosage modulation experiments using the human ORFeome library. To accommodate the functional genomics approach of ORF overexpression screening on the GMAP, I designed features against human ORFs in the Mammalian Genome Collection (MGC) (Strausberg et al. 1999; Strausberg et al. 2002; Lamesch et al. 2007; Temple et al. 2009). At the time of design, the library (version 5.1) was incomplete, consisting of 15483 of a predicted ~25000 maximum human ORFs. Anticipating that updates to the ORFeome library would be generated from the remaining MGC transcripts, I designed probes against the members of the MGC cDNA collection not currently found in the hORFeome library. In total, probes were designed against 24363 distinct ORF sequences. Microarray probes were designed using the software OligoArray 2.1 (Rouillard et al. 2003). For each of the 24363 distinct ORF sequences, up to 8 probes were generated. All probes were 25 nucleotides long and had GC content between 30-75%, modeled after the existing Affymetrix Human Gene 1.0 ST array. Probes were designed such that all regions of each ORF were interrogated, and the presence of incomplete ORFs would be apparent. In addition to these novel features, and for array comparison purposes, for each gene I included up to three human gene (huGene) microarray features present on the human expression profiling Affymetrix Human Gene 1.0 ST array (Table 4-2).

Probe ID	Description	Unique Probes	Replicates	Total number of probes	Probe length
hORF	Human ORFeome	134901	1	134901	25
huORF	HuGene ORFs	58087	1	58087	25
HP	shRNA sequences (mouse and human)	248049	3	744147	22
HPC	shRNA negative controls	138	33	4554	22
HSPI	Hybridization spike-in probes	200	25	5000	22
HPTMM	Hairpin mismatch control probes	8097	3	24291	22
TAG	Yeast bar-code probes	26801	3	80403	20
yORF	Yeast open reading frames	11421	1	11421	25
Total		**487694**	**NA**	**1062804**	**NA**

Table 4-2. Description of the features on the UT GMAP 1.0 microarray.

To assess the GMAP performance with human ORF hybridization, we developed 41 plasmid pools of entry clones (15,332 cDNAs representing >12,000 genes) from the human ORFeome v5.1 collection. Subsequently, 15,332 ORFs were amplified in pooled format with common flanking primers, labelled and hybridized to both the Human Gene 1.0 ST array and GMAP. Signal for features shared between the two microarrays was highly correlated ($r = 0.953$), with similar distribution of signal across the features for each microarray and similar signal-to-noise ratios. These results demonstrate that the GMAP has robust reporting of ORF data, and suggests that the GMAP can be used for a number of human gene assays including hORF overexpression screens. To compare the dynamic range of signal for huORF and huGene features on the GMAP, varying amounts of probe generated from ORFeome plasmid pools were hybridized to microarrays. The resulting data indicated that two-fold changes in probe input produce highly correlated signals across a 16-fold dynamic range, thus we combined huORF and huGene features into discrete sets and poor features were removed using a custom filtering algorithm.

4.2.3 Human multi-copy suppression profiling in yeast

The human ORFeome was transformed into *S. cerevisiae en masse* (Fig. 4-4). To establish the number of human ORFs that were integrated in plasmids in the pool, I performed a detection above background analysis using microarrays to determine that 96.1% of the ORFs from the human ORFeome plasmids were detected above background levels in the yeast pool ($p < 0.005$). The number of human ORFs that were expressed in the pool was determined by inducing expression with galactose, extracting total RNA, enriching for mRNA, generating double-stranded cDNA and PCR amplifying the cDNA via common primers to up/downstream untranslated regions. Performing the detection above background analysis with these data, 84.3% of the hORFeome was found to be expressed in yeast ($p < 0.05$). I also characterized the pool by identifying human genes that were toxic to yeast when expressed to high levels with galactose. The yeast hORFeome pool was grown in expression-inducing (galactose) and repressing (glucose) conditions and 4 biological replicates (representing 2 time points) of each condition

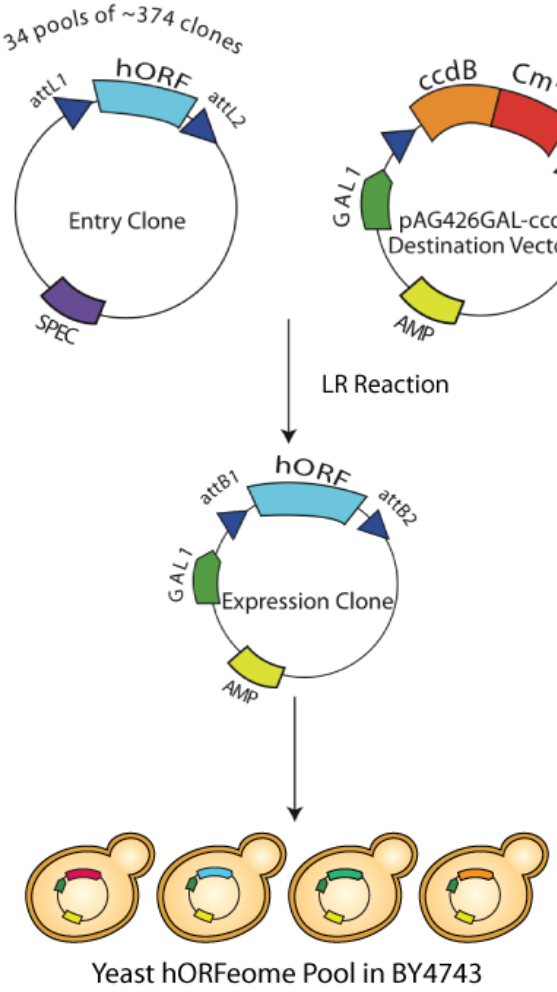


Figure 4-4. Pool Construction. Yeast human ORFeome pool of strains was created using the human ORFeome (v3.1) containing 12212 distinct ORFs. Clones were transferred to the destination yeast expression vector, transformed into *E. coli* and plasmids were isolated and then transformed into *S. cerevisiae* (BY4743) *en masse*. In yeast, human ORF expression is induced by the addition of galactose. Detection above background using microarrays found 96.1% of the ORFs integrated in the pool ($p < 0.005$) and 84.3% of the hORFeome was expressed in yeast ($p < 0.05$).

were collected. After ORFs were amplified and hybridized to microarrays, the hybridization intensity of the 4 induced experiments were compared with the paired 4 repressed experiments to yield a list of human genes toxic to yeast (Benjamini-Hochberg-corrected Student's T-test cutoff < 0.1).

As a first step to determine if the yeast hORFeome pool could be used in an MSP assay, I treated the pool with the anti-neoplastic antimetabolite methotrexate. Methotrexate is a synthetic folate analog that inhibits the catalysis of folate to tetrahydrofolate by competitively inhibiting the activity of dihydrofolate reductase (DHFR) (Villafranca et al. 1983). Using hMSP, methotrexate consistently selected for the yeast strain overexpressing the human ORF encoding DHFR (Fig. 4-5a). Given that hMSP makes use of a pooled hORFeome collection, individual strains containing DHFR were constructed and confirmed that the overexpression of DHFR conferred resistance to methotrexate toxicity. Hemoglobin epsilon 1 (HBE1), a protein that was not identified as resistant to methotrexate, did not confer resistance to methotrexate toxicity when overexpressed (Fig. 4-5b,c). A gene that was also consistently identified by the microarray as resistant to methotrexate was DHFR-like 1 (DHFRL1), a protein with unannotated function that has been shown to be both expressed and have a protein sequence homologous to DHFR (BLAST E-value = 2e-102). The overexpression of DHFRL1 in yeast was not sufficient to rescue the cells from methotrexate toxicity (DHFRL1 was DNA sequence-verified and DHFRL1 protein was present at high levels) and was likely identified due to non-specific cross-hybridization.

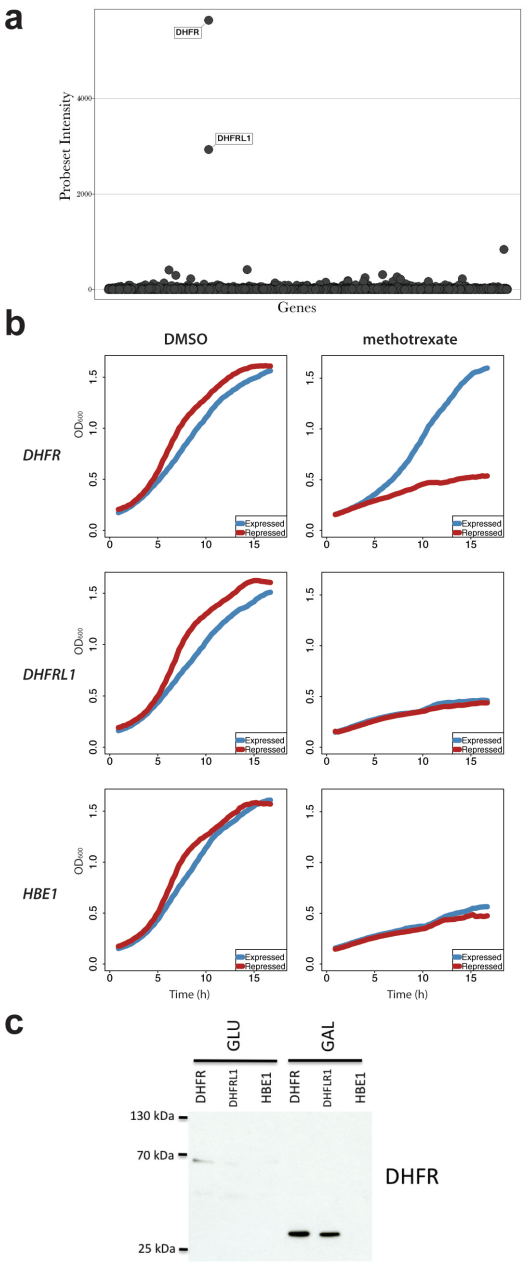


Figure 4-5. Identifying the human drug target of methotrexate with hMSP. (**A**) I treated the pool with the methotrexate, a synthetic folate analog that inhibits the catalysis of folate to tetrahydrofolate by competitively inhibiting dihydrofolate reductase (DHFR). Using hMSP, methotrexate consistently selected for the strain overexpressing human DHFR. (**B**) Individual unpooled strains confirmed this observation. Hemoglobin epsilon 1 (HBE1), a protein that was not identified as resistant to methotrexate, did not confer resistance to methotrexate toxicity when overexpressed. DHFR-like 1 (DHFRL1), which is homologous to DHFR (BLAST E-value = 2e-102), was not found to be resistant and was identified due to non-specific cross-hybridization. (**C**) DHFR and DHFRL1 protein expression was observed in inducing conditions only, as confirmed by western blot.

4.2.4 A novel candidate target for the β-blocker propranolol

Since I was able to demonstrate that the human drug target DHFR could be identified in an hMSP run, I applied the hMSP assay to test for drug-gene interactions with drugs that do not have known targets in yeast. One such drug is propranolol, a competitive non-selective β-adrenergic receptor (βAR) antagonist (Black et al. 1964; Black 1989). Despite over 40 years of clinical use in infants, propranolol has only recently emerged as a therapeutic in the treatment of infantile hemangiomas (IHs), a common and occasionally life-threatening tumor of infancy (Leaute-Labreze et al. 2008; Siegfried et al. 2008; Sans et al. 2009). Propranolol treatment results in a rapid change in hemangioma appearance followed by a shortened period of involution, improving upon traditional systemic corticosteroid therapy while exhibiting considerably fewer side-effects (Sans et al. 2009). This therapy is now a first-line treatment for IHs, and, while its mechanism of action in the treatment of this disease is not known, it appears to be independent of its known βAR antagonist activity (Drolet et al. 2013; Marqueling et al. 2013).

Propranolol was used in hMSP to select for strains that were resistant to its toxicity when overexpressing a particular human ORF. In replicate hMSP runs, dual specificity phosphatase 10 (DUSP10) and dual specificity phosphatase 16 (DUSP16) were observed to provide the greatest rescue from drug toxicity (Fig. 4-6a). Individual strains overexpressing either DUSP10 or DUSP16 were shown to rescue yeast from propranolol toxicity, confirming the observations from the pooled hMSP selection experiment (Fig. 4-6b). It is worth noting that of the known targets of propranolol, only the β_2AR (ADRB2) was present in the hORFeome pool. However, to establish functional integrity of the human β_2AR in yeast, it has been shown that the ORF must be modified to include a yeast localization signal (for example, from the α factor receptor gene *STE2*) (King et al. 1990). As a result, I did not expect the yeast strain overexpressing ADRB2 to be resistant to propranolol treatment.

Given that I found DHFR overexpression to confer resistance to methotrexate, and since it is known that the two form an interacting complex (Matthews et al. 1977), I hypothesized that propranolol toxicity may be alleviated due to an interaction with DUSP10/16. Both of these DUSPs contain a highly conserved DUSP catalytic domain (DSPc)

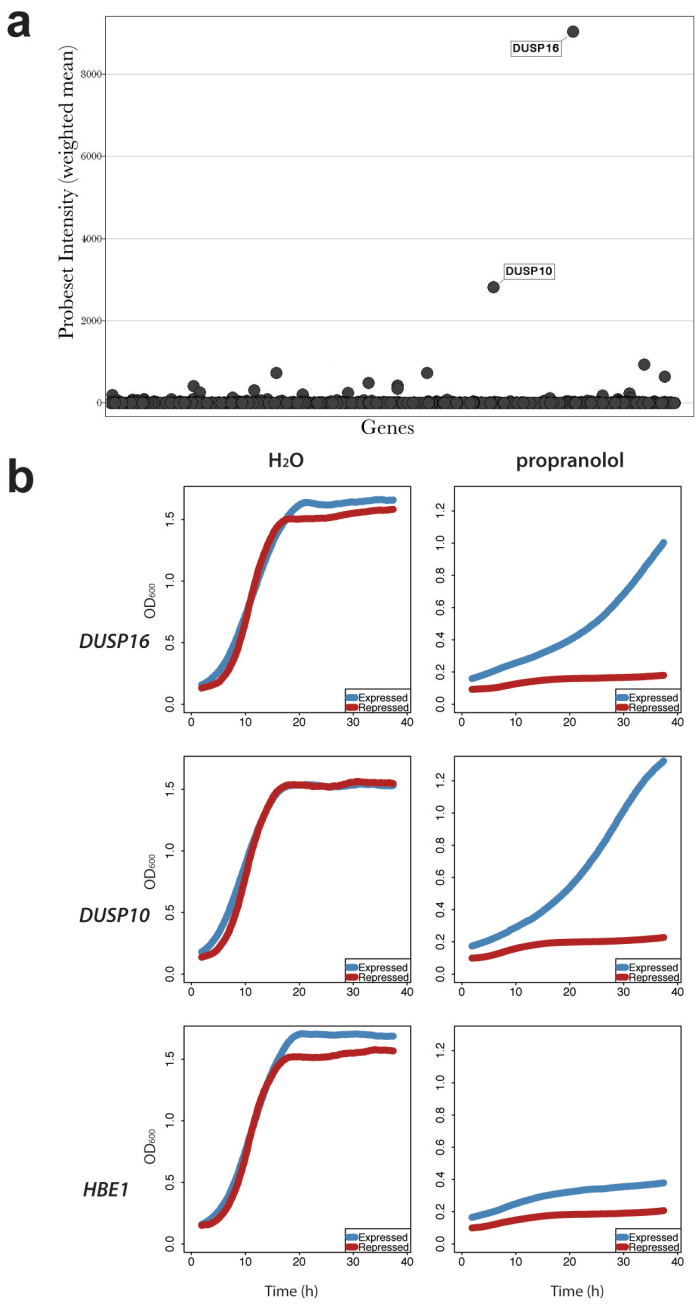


Figure 4-6. Propranolol hMSP profile. (**A**) In 3 replicate hMSP runs, dual specificity phosphatases 10 (DUSP10) and 16 (DUSP16) were observed to provide rescue from drug toxicity. (**B**) This was confirmed with individual strains. HBE1 was used as a control because it was not identified in the competitive assay. While the known target of propranolol is the β_2AR (ADRB2), it has been shown that the ORF must be modified to include a yeast localization signal for functionality in yeast. As a result, I did not expect the yeast strain overexpressing ADRB2 to be appearing in this screen.

(Theodosiou and Ashworth 2002; Marchler-Bauer et al. 2009). The aspartate, cysteine and arginine residues are essential for catalysis (Theodosiou and Ashworth 2002). A protein encoding the STYX gene, a close relative of the MKP family, contains a naturally occurring C to G substitution in the DSPc, rendering it inactive (Wishart and Dixon 1998). Based on these data, I used site-directed mutagenesis to introduce a C408G mutation in DUSP10 to render it phosphatase-dead. Yeast overexpressing the *dusp10* C408G mutant protein were found to be as resistant to propranolol relative to the wildtype DUSP10. This suggests DUSP10 rescue is mediated by the presence of the protein and not by its phosphatase activity, suggesting that the drug may bind to the protein.

Since the DUSP10/16 ORFs were heterologously expressed in *S. cerevisiae*, we tested whether DUSP10/16 overexpression in human cells would also confer resistance to propranolol toxicity. DUSP10 or DUSP16 proteins were overexpressed via doxycyclin-induction in HEK293 cells, which were subsequently treated with propranolol (collaboration with A. Arnoldo). Cell counts after drug treatment were determined with the sulforhodamine B (SRB) assay, measuring cellular protein content as a proxy for viability (Vichai and Kirtikara 2006). Induced overexpression of DUSP16 was sufficient to rescue HEK293 cells from propranolol toxicity at concentrations of 75μM confirming our observations in yeast ($p < 1 \times 10^{-8}$, n=18), while rescue was not observed in non-inducing conditions ($p = 0.29$, n=6; Fig. 4-7). We were unable to confirm whether DUSP10 overexpression had a similar rescue because, of the multiple cell lines overexpressing DUSP10 that were generated, the protein was not detected by western blot.

Next, I tested whether propranolol could affect DUSP10/16 activity *in vitro*. I measured the hydrolysis of a synthetic phosphatase substrate 3-O-methylfluorescein (OMFP), which fluoresces when dephosphorylated (Lazo et al. 2006; Molina et al. 2009). I observed inhibition of basal activity of DUSP10 *in vitro* by propranolol at drug concentrations similar to those of other published methods (Fig. 4-8) (Lazo et al. 2006; Molina et al. 2009). At 100μM of propranolol, DUSP10 basal activity was inhibited by 60%. DUSP10 had greater activity than DUSP16 *in vitro,* and due to narrow dynamic range of DUSP16 fluorescence measurements, I was unable to confidently determine whether propranolol could significantly affect the activity of DUSP16 *in*

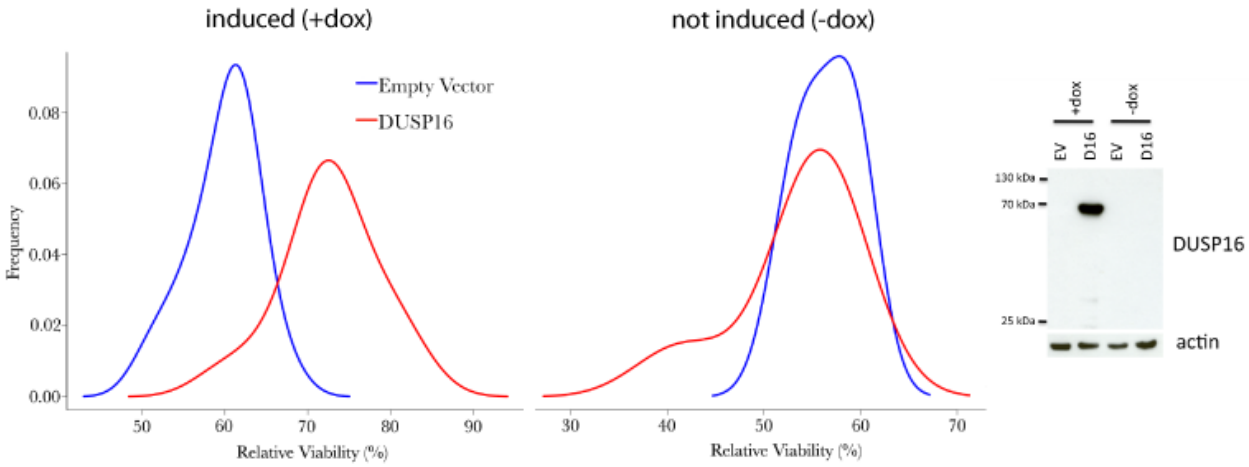


Figure 4-7. Confirming DUSP16 resistance in human cells. DUSP16 was expressed via doxycyclin-induction in HEK293 cells, which were subsequently treated with propranolol. Cell counts were obtained with the sulforhodamine B (SRB) assay, measuring cellular protein content as a proxy for viability. Induced expression of DUSP16 rescued HEK293 cells from propranolol toxicity at concentrations of 75μM confirming the observations in yeast ($p < 10^{-8}$, n=18). Rescue was not observed in non-inducing conditions ($p = 0.29$, n=6). EV, empty vector; D16, DUSP16.

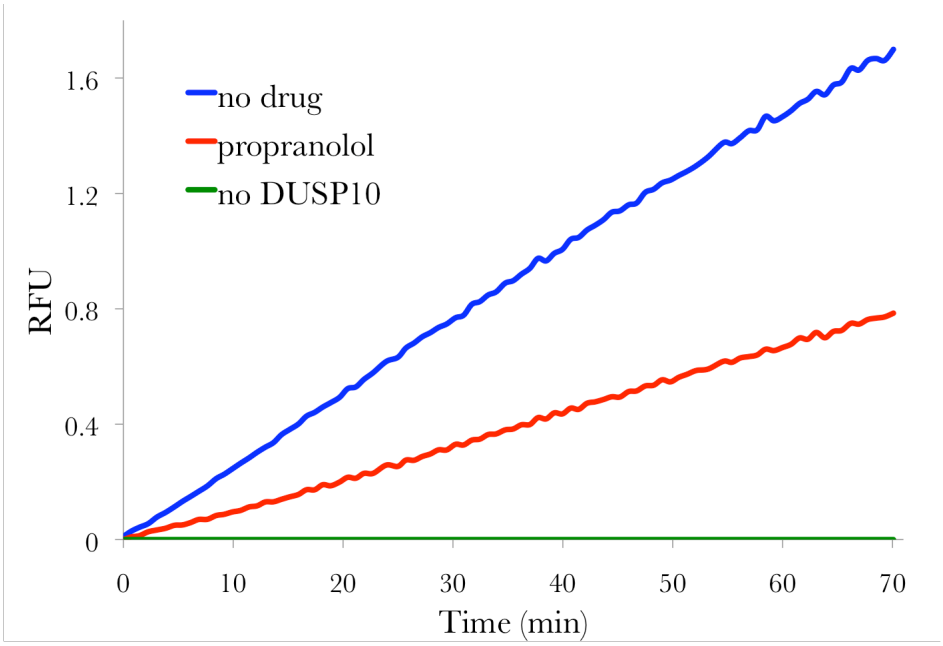


Figure 4-8. Propranolol inhibits DUSP10 *in vitro*. I measured the hydrolysis of a synthetic phosphatase substrate 3-O-methylfluorescein (OMFP), which fluoresces when dephosphorylated by DUSPs. Basal activity of DUSP10 was inhibited *in vitro* by propranolol at concentrations similar to those of other published methods (Lazo et al. 2006; Molina et al. 2009). At 100μM propranolol, DUSP10 activity was inhibited by 60%.

vitro. To determine the nature of DUSP10 inhibition, I applied classical Michaelis-Menten kinetics to the measurements of DUSP10 activity using nonlinear regression (Montgomery and Swenson 1976; Ritz and Streibig 2008). However, the regression analysis yielded inconclusive results, and these data were not discriminately informative with regard to a putative mechanism of phosphatase inhibition.

As an antihypertensive therapeutic, (S)-propranolol is 130-fold more potent than the (R)-entantiomer (Ng 2009). However, I found both the (S)-(–) enantiomer and the racemate to have equal potency when applied to both the hMSP and phosphate release assays. A similar observation was made by Sozzani *et al.* when demonstrating that both propranolol enantiomers inhibit protein kinase C equally, citing the amphipathic nature of the drug (Sozzani et al. 1992).

Based on these *in vitro* DUSP activity data, I tested whether propranolol was capable of binding directly to DUSP10/16 proteins. To assay this, I performed an affinity-based capture using biotin-conjugated propranolol and streptavidin-coated paramagnetic beads. After drug was bound to the beads, I incubated the beads with purified DUSP10 or DUSP16. No protein was detected in the pull-down assays, suggesting that biotinylated propranolol is unable to bind to DUSP10/16. However, it is possible that potential interactions may be hindered due to interference caused by the short linker region between the biotin tag and the propranolol moiety (Sato et al. 2010). Isothermal titration calorimetry (ITC) is one approach that may allow one to determine the binding affinity for propranolol to the DUSP proteins.

4.3 Discussion

Propranolol was developed for the treatment of angina pectoris, a disease characterized by severe chest pain brought about by a lack of oxygen delivery to the myocardium (Black et al. 1964). Pioneer pharmacologist Sir James W. Black sought to treat this disease by decreasing heart rate in patients suffering from angina, thereby reducing myocardial oxygen demand (Stapleton 1997). In a reverse engineering approach, Black "emasculated" the hormone epinephrine to synthesize a compound with reduced agonist efficacy, yielding propranolol, which was used to reduce mortality and morbidity in angina sufferers (Black 1989; Stapleton 1997). Propranolol is known to act by reducing the level of activity of βARs (agonist-induced and agonist-independent),

decreasing cAMP-dependent phosphorylation of factors that regulate the intracellular calcium concentration (Chidiac et al. 1994; Dorn 2010). In turn, this reduces heart contraction, which is controlled by calcium-dependent proteins.

Propranolol is also widely used therapeutically in the treatment of heart failure (Goodman et al. 2001). However, the rationale for the use of a βAR blocker in treating this disease is actually counterintuitive since its pathogenesis involves chronically enhanced sympathetic tone, which leads to a reduction in βARs (Karoor et al. 2004). As mentioned, propranolol has also been demonstrated to have great therapeutic effect in the treatment of IHs (Leaute-Labreze et al. 2008), and a recent study has shown propranolol is safer and has greater efficacy than the standard therapy of corticosteroid treatment (Annual Congress Of The European Academy Of Dermatology And Venereology, 2010). Since the βAR activity of propranolol appears to be independent of its therapeutic mechanisms in these diseases, it has been suggested that it acts downstream in the signaling pathway via an auxiliary mode of action that modulates Mitogen-Activated Protein Kinase (MAPK) activation (Karoor et al. 2004; Leaute-Labreze et al. 2008; Boye and Olsen 2009).

It has been shown that IH tumors exhibit high levels of MAPK activity (Arbiser et al. 2001). Also, the Extracellular-Signal-Regulated (ERK) subfamily of MAPKs has been implicated in IHs, and it has been suggested that ERK activity, mediated by Vascular Endothelial Growth Factor (VEGF) and basic Fibroblast Growth Factor (bFGF), is responsible for the development of IHs (Arbiser et al. 2001; Boye and Olsen 2009; Pramanik et al. 2009). MAPK activation is regulated by the DUSPs known as MAPK Phosphatases (MKPs), which downregulate activity of MAPKs by dephosphorylating tyrosine and threonine residues (Jeffrey et al. 2007). Both DUSP10 and DUSP16 are MKPs and have been shown to interact with and dephosphorylate residues in the activation loops of the three major subfamilies of the MAPKs (Theodosiou et al. 1999; Masuda et al. 2001; Theodosiou and Ashworth 2002; Bandyopadhyay et al. 2010).

Recent evidence suggests that MAPK activation can be modulated by propranolol. Meier *et al.* found that propranolol treatment of rat vascular smooth muscle cells resulted in increased levels of phospho-ERK (Meier et al. 1998). Similarly, propranolol treatment of HEK293 cells expressing the β_2AR increased cellular levels of phospho-ERK (Azzi et al. 2003). Conversely, Karoor *et al.*

observed that in mice overexpressing the G-protein $G_{s\alpha}$, propranolol treatment was shown to decrease the amount of phosphorylated MAPKs in protein extract from heart tissue (Karoor et al. 2004). It has been noted that the phosphorylation of ERK in response to βAR agonists is dependent on cell type, therefore, one may anticipate that the activity of MAPKs varies in response to propranolol from one cell type to another (Pullar et al. 2006).

I propose that the mode of action by which propranolol modulates MAPK activity is via interaction with the DUSP10/16 proteins. Modulation of DUSP protein activity may be the reason for the therapeutic benefit of propranolol in IH and heart failure, since the DUSPs have been implicated in various vascular developmental processes. The MKP DUSP5 was recently shown to regulate angioblast development, and a *dusp5* S147P mutation was observed to be associated with different vascular malformations and tumors, including lymphatic, arteriovenous, and venous malformations, as well in a single IH sample (Pramanik et al. 2009). If the somatic S147P mutation were to increase phosphatase activity of DUSP5 in a disease, one may envision a model wherein propranolol acts therapeutically by regulating the phosphatase activity of DUSP10/16 to return MAPK phosphorylation to wild-type levels. When DUSP5 was overexpressed in yeast, I found that it did not rescue the cells from propranolol toxicity. Research has noted that mice lacking the DUSP10 gene exhibit an accelerated vascular response to lipopolysaccharide injection in a mouse model of the Shwartzman reaction, wherein thrombosis is observed in the tissue where a toxin has been applied (Qian et al. 2009). As well, DUSP6 null mice exhibit greater rates of myocyte proliferation during embryonic development, resulting in enlarged hearts, which protected the mice from cardiomyopathy (Maillet et al. 2008).

The finding that the overexpression of DUSP10/16 can confer resistance to *S. cerevisiae* in the presence of propranolol was facilitated by the unbiased nature of the hMSP assay. This assay is largely a tool that allows one to select for yeast strains with the greatest fitness in the presence of a toxic challenge. In this manner, there is no predilection toward a candidate drug target, such that all strains have an equal opportunity to grow in the presence of a compound. This is the first high-throughput chemical genomic assay using human proteins in yeast for the purpose of drug target identification. While I have demonstrated its utility in identifying novel and potential drug targets, the assay has limitations. In particular, in order to apply a strong selection to identify the

fittest strains, the compound that is used to treat the cells must be applied at both a toxic and water-soluble concentration. For this reason, I was unable to find strains that were most resistant to propranolol analogs, including epinephrine, metoprolol, timolol and atenolol (atenolol has been shown to have similar therapeutic effects in IH). However, preliminary results have exhibited a similar pattern of strain selection using the propranolol analog alprenolol, a βAR antagonist which possesses no inverse agonist activity (Chidiac et al. 1994), warranting further investigation.

Given the chemical genomic utility of the current assay, it would be beneficial to include the current version of the human ORFeome (version 8.1) as it would increase the number of genes that can be queried at once (Yang et al. 2011). To overcome the limitation that a drug must be toxic at a soluble concentration, the ORFeome can be transformed into a broadly drug-sensitive yeast strain, allowing one to significantly dial down the dose applied in hMSP (Suzuki et al. 2011). Further iteration of this study will increase the effectiveness of our novel chemical genomic assay with the goal of expediting human drug target identification.

4.4 Materials and Methods

Feature size comparative analysis: Yeast deletion pools were thawed from frozen stocks and heterozygote essential gene deletion mutants were grown for 20 generations, while homozygous deletion mutants were grown for 5 generations as described (Hoon et al. 2008). After growth, heterozygous essential deletion mutants were mixed with correspondingly treated homozygous non-essential deletion mutants. Genomic DNA was isolated and molecular barcodes amplified by PCR. Amplicons were then hybridized to microarrays overnight, washed, stained and scanned the following day. For further details regarding sample preparation and data analysis, consult Pierce *et al* (Pierce et al. 2007) and Hoon *et al* (Hoon et al. 2008). I performed a HIPHOP screen (pooled heterozygous essential strains and homozygous deletion non-essential strains) with tunicamycin treatment ($IC_{10\text{-}20}$ = 0.35μM). Tunicamycin is a known glycosylation inhibitor, targeting the yeast essential gene *ALG7* (Barnes et al. 1984; Kukuruzinska and Robbins 1987; Kukuruzinska and Lennon 1995), which encodes UDP-N-acetyl-glucosamine-1-P transferase, a vital protein in the dolichol pathway of protein asparagine-linked glycosylation (Rine et al. 1983;

Cherry et al. 1997). Upon treatment with tunicamycin, unfolded proteins remain in the ER (endoplasmic reticulum) (Parodi 2000). A sample treated with 2% DMSO was used as a control. Yeast pools were grown in liquid culture in 48 well plates in a shaking spectrophotometer interfaced to liquid handling robots. After the cells had grown for the desired number of generations, corresponding to a specific optical density (OD), they were robotically harvested (Pierce et al. 2007). Genomic DNA was isolated from each pool, and the DNA barcodes were amplified by PCR using common primers. These barcodes were subsequently hybridized to three generations of barcode microarrays: the aforementioned TAG3, TAG4 and *S. cerevisiae* whole genome tiling arrays. Each chip was prepared using the optimal hybridization and wash/stain protocols recommended for that microarray type. Deletion strain abundance was resolved by averaging scanned downtag and uptag intensities for each strain and comparing intensities between the tunicamycin-treated pool and the DMSO-treated pool (Pierce et al. 2007).

Pool construction. The yeast human ORFeome pool of strains was generated using the human ORFeome version 3.1 encompassing 12212 distinct ORFs (obtained from Invitrogen) (Lamesch et al. 2007). Clones were transferred in pools of ~374 to the destination yeast expression vector pAG426GAL-ccdB (Alberti et al. 2007) by Gateway reaction (Gateway LR Clonase II enzyme mix, Invitrogen catalog no. 11791-100) to yield 34 pools. Individual pools from each Gateway reaction were transformed into E. coli (TransforMaxTM EC100TM cat. no. EC10010)by electroporation and plated onto LB media with carbenicillin selection (50µg/ml). Colonies were pooled and plasmids were isolated (QIAprep Spin Miniprep Kit cat. no. 27104). Each pool was then transformed into S. cerevisiae (BY4743) by lithium acetate/single-stranded carrier DNA/PEG procedure (Gietz et al. 1995; Gietz and Schiestl 2007) and selected on SD-URA (synthetic dropout media without uracil) selection media. Finally, yeast colonies were pooled to create a final pool containing all clones and stored at -80°C until use.

Pool growth. Yeast human ORFeome pools were thawed from -80°C and diluted to OD_{600} 0.0625 in SD-URA media containing 2% raffinose. Pools were grown for 6 hours at 30°C while shaking at 200rpm. During this time period, the pool recovers from the thaw from -80°C and undergoes a generation doubling, such that yeast growth is in the lag phase entering the logarithmic phase. To induce ORF expression, galactose was added to the culture to a final concentration of 2% and

cultures were grown for 1 hour at 30°C while shaking at 200rpm. 686µl of the pool were pipetted into wells of a 48-well microplate and, compounds were added in a 1/25 or 1/50 dilution to the culture at ~IC_{90} to select for cells containing ORFs that conferred resistance to yeast against specific compounds. Cells were grown in a GENios microplate reader (Tecan Group Ltd.) reading the OD_{600} every 15 minutes while shaking at 30°C. After drug-treated cultures reached an OD_{600} of 0.8, they were harvested by a MultiPROBE II PLUS EX robotic liquid handling system (PerkinElmer Inc.) and temporarily stored on a save plate at 0°C. The microplate reader and liquid handling robot were controlled by Yeast Grower software (Michael Proctor, Stanford Genome Technology Center) generated using LABVIEW (National Instruments Corp.) (Proctor et al. 2011).

Plasmid isolation, ORF amplification and DNA labeling. The ORF-containing plasmids (with the yeast expression vector pAG426GAL-ccdB backbone) were isolated from cell pellets using a Zymoprep Yeast Plasmid Miniprep Kit II (Zymo Research catalog no. D2004). Plasmid was eluted in 20µl of ddH_2O and amplified using the FailSafe PCR System (EPICENTRE Biotechnologies catalog no. FSE51100) with primers 5'-GCGAAGCGATGATTTTTGAT-3' and 5'-CTTTTCGGTTAGAGCGGATG-3'. Amplicons were purified using a QIAquick PCR purification kit (Qiagen catalog no. 28104) and eluted in 20µl of ddH_2O. Purified amplicons were biotinylated using the Biotin DecaLabel DNA labeling kit (Fermentas Life Sciences catalog no. K0652). Unincorporated biotin-11-dUTP was removed using columns containing illustra Sephadex G-50 Fine DNA Grade gel (GE Healthcare catalog no. 17-0573-01).

RNA isolation, ds-cDNA synthesis, ds-CDNA amplification. Total yeast RNA was isolated using the hot acid phenol extraction method (Sambrook and Russell 2001) and treated with DNAse I. Enrichment of ploy(A) mRNA was accomplished using oligo(dT) bead-based NucleoTrap mRNA kit (Macherey-Nagel GmbH & Co.). First strand cDNA synthesis was performed as described previously as was second strand synthesis (Sambrook and Russell 2001). Double-stranded (ds) cDNA was subsequently subjected to PCR amplification to enrich for human ORFs expressed of the total mRNA expressed in yeast. Primers were designed downstream/upstream of the GAL1 promoter/CYC1 terminator: 5'-AATATACCTCTATACTTTAACGTC-3' and 5'-GCGTGAATGTAAGCGTGAC-3'.

Microarray hybridization. ORF abundance from the drug-treated pools was determined indirectly by hybridization to a DNA microarray. Purified biotinylated PCR products were hybridized to the GeneChip Human Gene 1.0 ST Array (Affymetrix Inc. catalog no. 901086). The hybridization mixture (75µl 2× Hyb buffer, 2.5µl 5mM b213 control oligonucleotide, 3µl 50× Denhardt's, 30µl ddH2O and 25µl biotinylated PCR product) was denatured at 100°C and subsequently chilled on ice for 2 minutes before hybridizing to the microarray at 45°C for 17±1 hours while rotating at 60rpm. Microarrays were stained using SAPE (Streptavidin-Phycoerythrin) without an antibody stain (staining mix: 300µl 2× MES staining buffer, 60µl BSA 20mg/ml, 2µl SAPE at 1mg/ml, 238µl ddH_2O; blocking mix: 300µl 2× MES staining buffer, 60µl BSA 20mg/ml, 240µl ddH_2O) using the Fluidics Station 450 (Affymetrix Inc.). Microarrays were scanned using the GeneChip Scanner 3000 7G (Affymetrix Inc.).

Microarray analysis. Microarray probe sequences from the GeneChip Human Gene 1.0 ST Array are mapped, by Affymetrix, to transcripts including untranslated regions (UTRs) upstream and downstream of ORFs. As well, not all ORF sequences from the hORFeome (v3.1) (Lamesch et al. 2007) were assigned designated probes on the microarray. Probe sequences were mapped to ORF sequences in the hORFeome (v3.1) by modifying Affymetrix probeset grouping and annotation files. Microarray data analysis was performed by using these files with both the Expression Console software and Affymetrix Power Tools (Affymetrix Inc.). Probeset hybridization values were summarized using the integrated PLIER algorithm with GC background subtraction and quantile normalization. ORF sequence annotations were obtained from NCBI.

Confirmation strain construction. Individual strains of E. coli containing specific ORFs were obtained from the hORFeome (v3.1) stock of frozen 96-well plates (Lamesch et al. 2007). After growth in SOC media with spectinomycin selection (50µg/ml), a diagnostic PCR (using primers 5'-CACGACGTTGTAAAACGACGGCCAGTC-3' and 5'-GAGCTGCCAGGAAACAGCTATGACCATG-3') was performed to determine if the amplicon sizes matched their corresponding ORF sizes. If the sizes matched, plasmid-containing ORFs were isolated using a GeneJet Plasmid Miniprep Kit (Fermentas Inc.), and transferred to the destination yeast expression vector pAG426GAL-ccdB (Alberti et al. 2007) by Gateway reaction. The cloning products were transformed into E. coli (DH5α) and plated on SOC media with

ampicillin selection (100μg/ml). Plasmids were purified from the resultant colonies, as described previously, and transformed into S. cerevisiae (BY4743) using the lithium acetate/single-stranded carrier DNA/PEG method (Gietz et al. 1995; Gietz and Schiestl 2007). Transformed yeast was selected on SD-URA selection media, and verified using the diagnostic ORF amplification PCR mentioned previously.

The yeast strains were grown overnight, and diluted to OD_{600} 0.0625 in SD-URA media containing 2% raffinose. Pools were grown for 3 hours at 30°C while shaking at 200rpm. To induce ORF expression, galactose was added to the half of the culture to a final concentration of 2%. Repression of expression was accomplished by adding glucose to the remaining half of the culture to a final concentration of 2%. Cultures were grown for 1 hour at 30°C while shaking at 200rpm. 686μl of a strain were pipetted into wells of a 48-well microplate, and compound concentrations were the same ~IC_{90} that was used in the hMSP experiment. Cells were grown in a microplate reader, as described previously, to obtain growth curves.

Drug treatment in hMSP and confirmation strains. Methotrexate (Sigma-Aldrich Co. catalog no. M9929) was used at an IC_{90} of 1mM on yeast cells, to achieve optimal selection for resistant strains in hMSP. As well, yeast cultures were treated with (S)-(–)-propranolol (Sigma-Aldrich Co. catalog no. P8688) at IC_{90}. Since yeast cell growth was generally robust in the presence of lower concentrations of propranolol, I achieved an IC_{90} at a final concentration of 6.76mM (an order of magnitude greater than inhibitory concentrations in mammalian cells). Other βAR antagonists, such as atenolol (Sigma-Aldrich Co. catalog no. A7655), metoprolol (Sigma-Aldrich Co. catalog no. M5391) and timolol (Sigma-Aldrich Co. catalog no. T6394) were unable to inhibit yeast growth at soluble concentrations. The βAR antagonists exhibited greater solubility in water, compared with methotrexate, which was more soluble in DMSO.

Site-directed mutagenesis. Identification of the active site consensus motif was accomplished with a Python script using a regular expression to identify the catalytic cysteine in the sequence DX_{26}(V/L)X(V/I)HCXAG(I/V)SRSXT(I/V)XXAY(L/I)M (Theodosiou and Ashworth 2002; Marchler-Bauer et al. 2009). Site-directed mutagenesis was accomplished with a QuikChange Lightning site-directed mutagenesis kit (Agilent Technologies catalog no. 210518). An online QuikChange application was used to design the primers 5'-

GCTTCTCATCCACGGCCAGGCTGGGGT-3' and 5'-ACCCCAGCCTGGCCGTGGATGAGAAGC-3' changing a TGC codon to a GGC codon. This results in an amino acid substitution at position 408 from a C to a G. The mutation was DNA sequence verified in three separate clones.

HEK293 confirmation cell lines and toxicity rescue. DUSP10wt, DUSP16wt and DUSP16-C408G were cloned into PB-TGcMV-Neo transposon vector using the Gateway reaction. HEK293_M2 (rtTA expressing) cells were seeded in DMEM+10%iFBS and were cotransfected with pCyL43 (transposase expression plasmid) and PB-TGcMV-Neo transposon vectors using FuGene according to manufacturer instructions. Selection started the next day using 1mg/ml Geneticin (Invitrogen) for 5 days. Note, initial expression of DUSP10/16 in HEK293_M2 cells was unsuccessful due to the cytotoxicity of induced DUSP10/16 expression, therefore we decided to start protein induction and immediately treat with propranolol. 10000 transfected HEK293_M2 cells were seeded in each well of a 48 well plate and grown in 300µL of DMEM+10%iFBS+Pen/Stept. 15 hours after seeding, DUSP10wt, DUSP16wt and DUSP16-C408G expression was induced by addition of doxycyclin (2µg/ml) and 1-2 hours later, propranolol was added at final concentrations of 25µM, 50µM, 75µM and 100µM, similar to previous studies (Meier et al. 1998). We subsequently focused on treating cells only at 75µM. The drug toxicity was assessed 4 days after the induction of protein expression using a sulphorhodamine B assay and compared to reference cells containing the empty vector (Vichai and Kirtikara 2006).

Western blots. Yeast cell lysates were prepared by vortexing cell pellets in equal volumes of RIPA buffer (50mM Tris-HCl [pH 8.0], 150mM NaCl, 1.0% Triton X-100, 0.5% sodium deoxycholate, 0.1% SDS) and glass beads (~0.5mm). HEK293 cell lysates were prepared by incubating cell pellets (1.0×10^6 cells) with RIPA buffer for 5 minutes at room temperature. Protein concentrations were determined by Bradford assay (Sigma-Aldrich Co. catalog no. B6916). After boiling, proteins were separated by SDS-PAGE (10%) at 140V for 90 minutes and transferred to nitrocellulose at 300mA for 90 minutes. Blots were blocked with 5% non-fat milk (in TBST) and all antibodies were incubated in 0.5% non-fat milk at room temperature for 90 minutes. Blots were imaged using enhanced chemilumescent (ECL) substrate for horseradish peroxidase (HRP)

(Thermo Fisher Scientific Inc. catalog no. 34087). Antibodies: DHFR (Sigma-Aldrich Co. catalog no. D1067), DUSP10 (Abcam plc. catalog no. ab87842, ab71532), DUSP16 (Abcam plc. catalog no. ab65151), actin (Abcam plc. catalog no. ACTN05), tubulin (AbD Serotec catalog no. MCA78G).

Phosphate release assay. The protocol to measure in vitro activity of DUSP10/16 was adapted from Biomol/Enzo Life Sciences and Lazo et al. (2006). 0.74μg of DUSP10 (Enzo Life Sciences Inc. catalog no. BML-SE467-0100) was used in a total volume of 70μL containing buffer (100mM Tris-HCl [pH 8.2], 40mM NaCl, 1mM DTT, 20% glycerol, 0.33% BSA), propranolol (1/35 dilution) and 3-O-methylfluorescein (OMFP; Sigma-Aldrich Co. catalog no. M2629). OMFP was used at multiple final concentrations ranging from 12.5μM to 500μM, based on activity measurements in previous studies (Lazo et al. 2006; Molina et al. 2009). Activity was measured using 96-well half-volume flat bottom black microplates at 30°C in a Safire II-Basic microplate reader (Tecan Group Ltd.) with an excitation wavelength of 485nm and an emission wavelength of 530nm (additional parameters: Gain 60, z-position 8500μm). Fluorescence reads were obtained every 30-40 seconds for 4 hours. Propranolol concentrations ranged from 25μM to 300μM, based on previous in vitro and in vivo studies (Morlock et al. 1991; Sozzani et al. 1992). To determine whether DUSP10 was functioning as a classical enzyme, I applied nonlinear regression techniques in the statistical language R (Ritz and Streibig 2008) to observe Michaelis-Menten kinetics (Montgomery and Swenson 1976). The *in vitro* data were presented as raw Relative Fluorescence Units (RFU) similarly to the previous studies (Molina et al. 2009). Due to narrow dynamic range of DUSP16 (Enzo Life Sciences Inc. catalog no. BML-SE495-0100) fluorescence measurements (phosphatase activity is ~240× lower per μg than DUSP10, as reported by Biomol/Enzo Life Sciences), significant changes in DUSP16 *in vitro* activity were not recorded.

5 Summary and perspectives

Massively parallel genome-scale microarray or sequencing analyses have diverse applications in the field of molecular biology, and I have reported my studies that were entirely facilitated by these approaches. In the absence of these technologies, the identification of resistance-causing mutations in yeast would be a more laborious task likely involving time-consuming fine genetic mapping. Studies of global genomic trends such as nucleosome occupancy would be impossible without sequencing. Not to mention that these studies are dependent on established reference genomes that are a product of large sequencing projects. Now that these methods are widely available and economical, we can ask biological questions on a greater scale than ever before as well as questions which were impractical to ask with the previous technology. This affords us great opportunity in understanding genome structure and response to perturbation.

5.1 Yeast chemical genomics and human genes

I reported a novel assay to identify human drug targets, called human Multi-copy Suppression Profiling (hMSP) and confirmed its utility by identifying human DHFR as the target of methotrexate. hMSP was applied to study the βAR antagonist propranolol, identifying DUSPs 10 and 16 as potential novel targets. This study emphasizes the value of *in vivo* chemical genomic assays in yeast for the purpose of identifying novel human drug targets in a rapid and unbiased manner.

In human MSP, a heterologous ORFeome is overexpressed in a model organism to identify human drug targets. This is a novel approach, as previous studies have demonstrated that MSP methods are effective at identifying proteins that interact with small molecules using endogenous

ORFeomes, as shown for *S. cerevisiae* and *Schizosaccharomyces pombe* (Hoon et al. 2008; Nishimura et al. 2010).

The next step for hMSP is to use it in chemical genomic screens. To acheive this with the most accurate results, there are several updates that need to be made to the assay. The original hORFeome pool was constructed by *en masse* transformation with the human ORFeome version 3.1, and this should be updated to the human ORFeome version 8.1 which is clonally prepared and fully sequence verified (Yang et al. 2011). In addition, microarrays are no longer required to process hMSP results as it is amenable to a high degree of multiplexing. Since each experiment yields only a few strains that are resistant to the compound of interest, only a couple of strains will dominate the composition of the competitively grown strain pool. Depending on the sequencing technology, at least 1000 hMSP experiments can be combined at once, greater than for HIPHOP experiments (Smith et al. 2010). With this degree of throughput, each sequencing run can cover hundreds of drugs with multiple biological replicates for each, yielding high confidence results.

A limiting factor of MSP experiments is that in order to achieve a strong selection for resistant strains, the pool must be treated with a very high concentration of drug (IC_{90}). Often, I found this was unachievable because many drugs would simply not reach this level of toxicity at soluble concentrations. A solution to this is to create the hORFeome pool in a drug-sensitive mutant, such as the "green monster" *S. cerevisiae* strain which bears deletions of all 16 ATP-binding cassette (ABC) drug transporters (Suzuki et al. 2011).

Ultimately, this assay does not definitively identify human drug targets for every drug. It is limited primarily to drugs that can directly bind to non-toxic overexpressed proteins when they are cytosolically localized. Even then, a large amount of follow-up experimentation is necessary to ensure that the target candidates are true drug targets. However, hMSP offers an unbiased first step to aid drug researchers in the identification of potential targets from a pool of all human ORFs. While similar methods now exist using RNAi-mediated knockdown or overexpression of human genes in mammalian cell lines (Kittanakom, *in review*; Arnoldo, *in preparation*), the yeast-based hMSP method offers a comparatively simple, fast and consistent profile of drug activity.

5.2 Understanding modes of antifungal resistance

We provided the first analysis of the genomic changes that underpin the evolution of resistance to antifungal drug combinations in *C. albicans* and *S. cerevisiae*. Using whole genome sequencing, diverse resistance mutations were identified among the lineages that evolved resistance to the drug combination. These included mutations in genes encoding the drug targets, a transcriptional regulator of multidrug transporters, a transcriptional repressor of ergosterol biosynthesis enzymes and a regulator of sphingolipid biosynthesis. Aneuploidies in several *C. albicans* lineages were also found. This study reveals multiple mechanisms by which resistance to drug combinations can evolve, suggesting novel strategies to combat drug resistance.

The mechanisms of yeast resistance to antifungal compounds is a well-studied topic, and many of the underlying targets are known (Shapiro et al. 2011). However, with the limited number of antifungal drugs available, the use of drug combinations will become more critical to treat fungal infections, and the mechanisms of resistance to these combinations are not as well understood (Hill et al. 2013). This same problem is being investigated in targeted combination therapy of melanoma, and a recent study has described the evolutionary dynamics of cancer, concluding that combination therapy is more effective then sequential therapy (Bozic et al. 2013).

We identified several key mutations that conferred resistance of yeast strains to azole and stress-response inhibitors, but the frequency of these mutations is unknown. With the cost of massively parallel sequencing decreasing significantly, future work will benefit from a larger sample size of resistant strains (including more replicates). Mechanisms that have the highest frequency will likely be good candidates for drug targeting.

The genome resequencing workflow developed in this study can easily be applied to other drug-resistant fungal strains. As a recent example, I identified several SNVs in drug-hypersensitive *S. cerevisiae* strains in collaboration with Tanvi Shekhar-Guturja and Leah Cowen. In this study, it was observed that the combination of an azole and the natural product beauvericin exhibits potently synergistic antifungal activity in *S. cerevisiae*. Wildtype yeast are resistant to beauvericin in the absence of an azole, therefore, beauvericin is not amenable to target-identifying assays

such as HIPHOP or yeast MSP (Giaever et al. 1999; Hoon et al. 2008). However, the "green monster" *S. cerevisiae* mutant in which all 16 ABC drug transporters have been deleted is sensitive to beauvericin (Suzuki et al. 2011). To identify the mode of action of this natural product, T. Shekhar-Guturja evolved resistance to beauvericin in the ABC-16 strain background, and identified several clones that are highly resistant to beauvericin. Three of these strains have been sequenced, and a missense mutation was observed in either beta regulatory subunit of casein kinase 1 or 2 (*CKB1* or *CKB2*) for each strain. T. Shekhar-Guturja is currently following up on these results to identify if the *CKB* genes play a role in the mode of action of beauvericin.

In addition to identifying resistance-conferring mutations in yeast, our study also highlights an observation that has been made previously: *C. albicans* responds to many stresses by increasing its chromosomal copy number, while under identical stress conditions *S. cerevisiae* resorts to aneuploidy less frequently. In all my analysis of evolved *S. cerevisiae* strains, I have not observed aneuploidies or copy number variation, however, in *C. albicans*, aneuploidy is a known mechanism to increase fitness in the presence of an antifungal compound (Selmecki et al. 2006; Selmecki et al. 2009a). A collection of haploid *S. cerevisiae* strains exists where each strain contains an additional copy of one or more of the 16 chromosomes. These strains exhibit decreased fitness including defects in cell cycle progression, increased glucose uptake and increased sensitivity to protein folding and synthesis inhibitors (Torres et al. 2007). As a result, it has been proposed that protein folding inhibitors such as 17-AAG may be effective therapeutic options for the treatment of aneuploid malignant tumors (Tang et al. 2011). However, it is worth noting that these studies involve research with *S. cerevisiae* strains and mouse embryonic fibroblast (MEF) cell lines, both of which exhibit proliferation defects due to aneuploidy (Tang et al. 2011). Aneuploidy in *C. albicans*, on the other hand, has a relatively minor fitness cost, as is observed in tumor cells (Selmecki et al. 2009a; Hanahan and Weinberg 2011).

These observations have two implications: first, the inhibition of protein synthesis and folding may not be as toxic to malignant cancer cells as it is to cells that are already exhibiting growth defects due to aneuploidy. In this case, *C. albicans* is a more appropriate model system to study the sensitivity of cells that bear additional chromosomal copies. Secondly, while *S. cerevisiae* and *C. albicans* are very divergent, the two species have similar morphologies, many orthologs and a

similar genome size with ~6000 genes (Dujon 2010). Yet these two fungal species have very different fitness costs associated with aneuploidy. Future studies can examine factors that contribute to aneuploidy tolerance in *C. albicans*, similar to aneuploidy-tolerating mutations that have been identified in *S. cerevisiae* (Torres et al. 2010). For example, using highly-multiplexed massively parallel sequencing, CNV-seq can be performed for the collection of *C. albicans* deletion strains after exposure to heat stress, as a stimulus known to increase copy number (Bouchonville et al. 2009; Noble et al. 2010). Since low sequence coverage would be sufficient to compare chromosome copy number between a deletion and a reference strain, this experiment can likely be performed in a single sequencing run. Also, it is worth exploring whether the aneuploidy-tolerating mutations in *S. cerevisiae* ubiquitin-specific protease *UBP6* are conserved in *C. albicans* (Torres et al. 2010). A study of this nature may identify novel factors that enable tumor cells to tolerate aneuploidies while other mammalian cells cannot.

5.3 Archaeal nucleosomes

Nucleosomes and chromatin have historically been considered to be unique to eukaryotes, and I reported the first archaeal genome-wide nucleosome occupancy map in the halophile *Haloferax volcanii*. I found that archaeal transcripts possess hallmarks of eukaryotic chromatin structure: nucleosome-depleted regions at transcriptional start sites and conserved −1 and +1 promoter nucleosomes. This discovery demonstrates that histones and chromatin architecture evolved before the divergence of Archaea and Eukarya, suggesting that the fundamental role of chromatin in the regulation of gene expression is ancient.

The *Hfx. volcanii* nucleosome occupancy map presented in this thesis is an *in vivo* map, and it is reproducible as I observed when comparing two biological replicates prepared several months apart while compiling data for this study. While this correlation was not quantified, based on preliminary observation, the two samples exhibited identical nucleosome occupancy. Since I successfully designed a sequence-based occupancy predictor based on this reproducible conserved positioning trend, I hypothesize that there is an intrinsic sequence preference for nucleosomes, in concordance with the proposal by Segal and Colleagues (Kaplan et al. 2009). In their study, Kaplan et al (2009) trained a model of occupancy on an *in vitro* nucleosome

occupancy map generated using chicken histones bound to *S. cerevisiae* genomic DNA. Due to the high degree of conservation among eukaryotic histones and a hypothesized sequence-based intrinsic positioning signal, they observed great correlation between the *in vitro* and *in vivo* yeast maps (Kaplan et al. 2009). *Hfx. volcanii* histones likely also have an intrinsic sequence preference in the formation of nucleosomes, and I propose a method to test in future experiments.

The most straightforward way to determine if there is an intrinsic sequence preference for archaeal histones is to compare *in vivo* and *in vitro* maps. This requires the binding of Haloferax histones to Haloferax genomic DNA *in vitro*. To accomplish this, I have cloned C-terminal hexahistidine-tagged *Hfx. volcanii* histone hstA into a T7 expression vector to be expressed in *Escherichia coli* cells based on previously defined methods for recombinant human histones (Tanaka et al. 2004). The hstA histone is homologous to the histone protein of the archaeon *Methanopyrus kandleri*, HMk, the monomer of which is homologous to eukaryotic H3-H4 dimers. HMk homodimerizes, forming a $(H3\text{-}H4)_2$ tetramer-like structure (Fahrner et al. 2001; Talbert and Henikoff 2010).

His_6-tagged hstA was ligated into the pET15b-based vector p11 (Novagen) containing a Tobacco Etch Virus (TEV) protease cleavage site in place of a thrombin site. Since we anticipated that expressing highly positively charged histones in bacteria would lead to toxicity due to binding of negatively charged genomic DNA, p11 has a copy of the lacI repressor on the plasmid for tighter hstA regulation in the absence of IPTG induction (Fig. 5-1). Initially, I expressed recombinant hstA in both T7 Express and T7 Express lysY *E. coli* strains (New England Biolabs), since the lysY strains produce a mutant T7 lysozyme which binds to T7 RNA polymerase to reduce the basal level of the expression of hstA. I observed expressed hstA in the insoluble fraction of both strains, suggesting that the strains are circumventing toxicity by forming inclusion body aggregates (Goeddel 1991).

Future experiments require the optimization of hstA induction to produce greater yield. Once purified, tag-cleaved and refolded, nucleosomes will be reconstituted by the salt dialysis method (Thastrom et al. 2004). These samples can be subjected to MNase-seq to determine the *in vitro* nucleosome occupancy of *Hfx. volcanii*, and subsequently compared to the *in vivo* occupancy map.

In vivo, nucleosome occupancy is believed to be relatively static, however, under certain stress conditions, such as oxidative stress, nucleosomes can be gained and lost to facilitate transcription

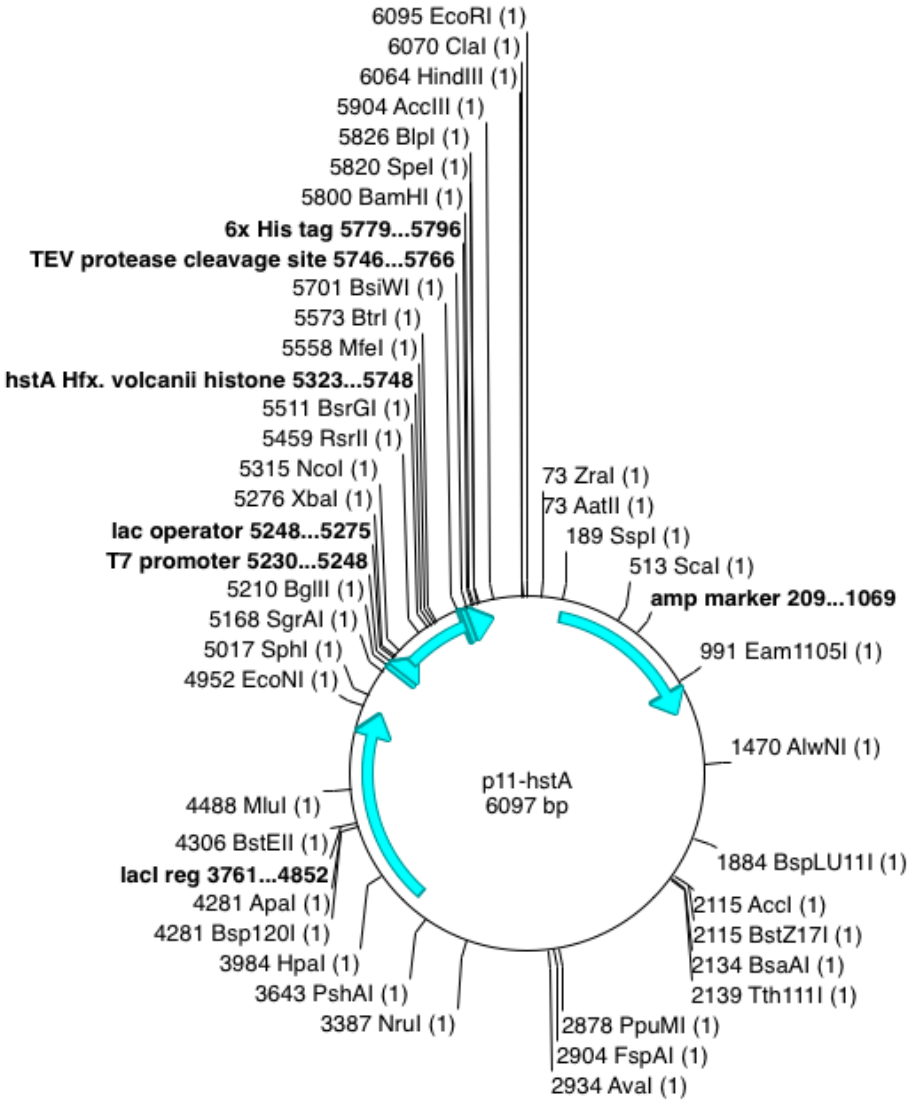


Figure 5-1. Vector expressing recombinant hstA. Plasmid map of His_6-tagged hstA is expressed under the regulation of a T7 promoter coupled to a lac operator. A TEV protease cleavage site is included to remove the His_6 tag post-purification. A local copy of the lacI repressor is present to regulate basal expression (-IPTG).

factor binding and differential gene expression (Huebert et al. 2012). Studies in *S. cerevisiae* have shown that a significant number of nucleosomes can be repositioned after hydrogen peroxide (H_2O_2) treatment to allow stress-activated transcription factors to regulate downstream pathways (Huebert et al. 2012). I am currently working on MNase-seq and RNA-seq data from a set of *Hfx. volcanii* samples that have been subjected to various stresses in attempt to identify occupancy changes in archaea. Included in our samples are treatments with H_2O_2, heat shock (60°C, 45min) and low salt (0.5M, 3h; note that because *Hfx. volcanii* can grow in a saturated salt solution, it is not possible to perform a high salt stress experiment). These samples can be studied using the methods of Huebert et al (2012) to identify nucleosome gains or losses in order to determine how stresses can affect chromatin architecture and gene expression regulation.

In addition, as mentioned earlier, *Hfx. volcanii* is a polyploid organism with ~15 copies of its genome present during the exponential growth phase and ~10 copies at early stationary phase (Breuert et al. 2006). The increase in genomic material as well as the resultant copy number variation of the histone-encoding hstA gene likely affects chromatin formation and structure. To test this, I have additional *Hfx. volcanii* samples that have been harvested at early stationary, mid-exponential, late exponential and saturation phases of growth. Applying the same analysis as described above for stress conditions, one can determine if nucleosome occupancy changes at different phases of archaeal growth or with different ploidy.

Furthermore, while it is known that the deletion of components of the histone octamer from *S. cerevisiae* is lethal, this has not been explored in archaea (Dollard et al. 1994). Using existing gene knockout methods in *Hfx. volcanii* may establish the essentiality of archaeal histones (Allers and Mevarech 2005). If these histones are found to be nonessential for Haloferax viability, one can perform a differential expression profile analysis to determine how the absence of histones affects global gene regulation.

Finally, while histone proteins are absent in bacteria, there are several histone-like DNA-binding proteins that exist in bacteria. One example is histone-like nucleoid structuring protein (H-NS) which uses a GC-based sequence preference analogous to histones to selectively silence horizontally-acquired genes (Navarre et al. 2006). Another histone-like protein is HU, an essential protein that localizes throughout the *Deinococcus radiodurans* nucleoid body. It is

proposed that HU enables the extreme radioresistance of *D. radiodurans* by preventing dispersion of DNA fragments following irradiation (Nguyen et al. 2009). A global MNase-seq or DNase-seq analysis of these genomes may reveal that these histone-like proteins may in fact exhibit regulation of transcription in bacteria.

6 References

Adachi T, Mizuuchi M, Robinson EA, Appella E, O'Dea MH, Gellert M, Mizuuchi K. 1987. DNA sequence of the E. coli gyrB gene: application of a new sequencing strategy. *Nucleic acids research* **15**(2): 771-784.

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF et al. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**(5013): 1651-1656.

Adams MD, Kerlavage AR, Fields C, Venter JC. 1993. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nature genetics* **4**(3): 256-267.

Adessi C, Matton G, Ayala G, Turcatti G, Mermod JJ, Mayer P, Kawashima E. 2000. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic acids research* **28**(20): E87.

Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, Schuster SC, Pugh BF. 2007. Translational and rotational settings of H2A.Z nucleosomes across the Saccharomyces cerevisiae genome. *Nature* **446**(7135): 572-576.

Alberti S, Gitler AD, Lindquist S. 2007. A suite of Gateway cloning vectors for high-throughput genetic analysis in Saccharomyces cerevisiae. *Yeast* **24**(10): 913-919.

Alberti S, Halfmann R, King O, Kapila A, Lindquist S. 2009. A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. *Cell* **137**(1): 146-158.

Allan J, Fraser RM, Owen-Hughes T, Keszenman-Pereyra D. 2012. Micrococcal nuclease does not substantially bias nucleosome mapping. *Journal of molecular biology* **417**(3): 152-164.

Allers T, Mevarech M. 2005. Archaeal genetics - the third way. *Nature reviews Genetics* **6**(1): 58-73.

Altman-Price N, Mevarech M. 2009. Genetic evidence for the importance of protein acetylation and protein deacetylation in the halophilic archaeon Haloferax volcanii. *Journal of bacteriology* **191**(5): 1610-1617.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* **215**(3): 403-410.

Ammar R, Torti D, Tsui K, Gebbia M, Durbic T, Bader GD, Giaever G, Nislow C. 2012. Chromatin is an ancient innovation conserved between Archaea and Eukarya. *elife* **1**: e00078.

Anderson JB. 2005. Evolution of antifungal-drug resistance: mechanisms and pathogen fitness. *Nat Rev Microbiol* **3**(7): 547-556.

Anderson JB, Funt J, Thompson DA, Prabhu S, Socha A, Sirjusingh C, Dettman JR, Parreiras L, Guttman DS, Regev A et al. 2010. Determinants of divergent adaptation and Dobzhansky-Muller interaction in experimental yeast populations. *Curr Biol* **20**(15): 1383-1388.

Anderson JB, Ricker N, Sirjusingh C. 2006. Antagonism between two mechanisms of antifungal drug resistance. *Eukaryot Cell* **5**(8): 1243-1251.

Anderson JB, Sirjusingh C, Parsons AB, Boone C, Wickens C, Cowen LE, Kohn LM. 2003. Mode of selection and experimental evolution of antifungal drug resistance in Saccharomyces cerevisiae. *Genetics* **163**(4): 1287-1298.

Annunziato A. 2008. DNA packaging: Nucleosomes and chromatin. *Nature Education* **1**(1).

Ansorge W, Sproat B, Stegemann J, Schwager C, Zenke M. 1987. Automated DNA sequencing: ultrasensitive detection of fluorescent bands during electrophoresis. *Nucleic acids research* **15**(11): 4593-4602.

Apostol B, Greer CL. 1988. Copy number and stability of yeast 2 mu-based plasmids carrying a transcription-conditional centromere. *Gene* **67**(1): 59-68.

Arbiser JL, Weiss SW, Arbiser ZK, Bravo F, Govindajaran B, Caceres-Rios H, Cotsonis G, Recavarren S, Swerlick RA, Cohen C. 2001. Differential expression of active mitogen-activated protein kinase in cutaneous endothelial neoplasms: implications for biologic behavior and response to therapy. *J Am Acad Dermatol* **44**(2): 193-197.

Arents G, Moudrianakis EN. 1995. The histone fold: a ubiquitous architectural motif utilized in DNA compaction and protein dimerization. *Proceedings of the National Academy of Sciences of the United States of America* **92**(24): 11170-11174.

Arnoldo A, Curak J, Kittanakom S, Chevelev I, Lee VT, Sahebol-Amri M, Koscik B, Ljuma L, Roy PJ, Bedalov A et al. 2008. Identification of small molecule inhibitors of Pseudomonas aeruginosa exoenzyme S using a yeast phenotypic screen. *PLoS Genet* **4**(2): e1000005.

Augenlicht LH, Kobrin D. 1982. Cloning and screening of sequences expressed in a mouse colon tumor. *Cancer research* **42**(3): 1088-1093.

Augenlicht LH, Wahrman MZ, Halsey H, Anderson L, Taylor J, Lipkin M. 1987. Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro. *Cancer research* **47**(22): 6017-6021.

Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, Gao X, Gingle A, Grant G, Harb OS et al. 2009. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic acids research* **37**(Database issue): D539-543.

Avery OT, Macleod CM, McCarty M. 1944. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types : Induction of Transformation by a

Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. *J Exp Med* **79**(2): 137-158.

Azzi M, Charest PG, Angers S, Rousseau G, Kohout T, Bouvier M, Pineyro G. 2003. Beta-arrestin-mediated activation of MAPK by inverse agonists reveals distinct active conformations for G protein-coupled receptors. *Proc Natl Acad Sci U S A* **100**(20): 11406-11411.

Bader GD, Heilbut A, Andrews B, Tyers M, Hughes T, Boone C. 2003. Functional genomics and proteomics: charting a multidimensional map of the yeast cell. *Trends in Cell Biology* **13**(7): 344-356.

Baetz K, McHardy L, Gable K, Tarling T, Reberioux D, Bryan J, Andersen RJ, Dunn T, Hieter P, Roberge M. 2004. Yeast genome-wide drug-induced haploinsufficiency screen to determine drug mode of action. *Proc Natl Acad Sci U S A* **101**(13): 4525-4530.

Bailey KA, Pereira SL, Widom J, Reeve JN. 2000. Archaeal histone selection of nucleosome positioning sequences and the procaryotic origin of histone-dependent genome evolution. *Journal of molecular biology* **303**(1): 25-34.

Balzi E, Wang M, Leterme S, Van Dyck L, Goffeau A. 1994. PDR5, a novel yeast multidrug resistance conferring transporter controlled by the transcription regulator PDR1. *J Biol Chem* **269**(3): 2206-2214.

Bandyopadhyay S, Chiang CY, Srivastava J, Gersten M, White S, Bell R, Kurschner C, Martin CH, Smoot M, Sahasrabudhe S et al. 2010. A human MAP kinase interactome. *Nat Methods* **7**(10): 801-805.

Barnes G, Hansen WJ, Holcomb CL, Rine J. 1984. Asparagine-linked glycosylation in Saccharomyces cerevisiae: genetic analysis of an early step. *Mol Cell Biol* **4**(11): 2381-2388.

Barnhart BJ. 1989. DOE Human Genome Program. In *Human Genome Quarterly*, Vol 2013.

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell* **129**(4): 823-837.

Bartfai R, Hoeijmakers WA, Salcedo-Amaya AM, Smits AH, Janssen-Megens E, Kaan A, Treeck M, Gilberger TW, Francoijs KJ, Stunnenberg HG. 2010. H2A.Z demarcates intergenic regions of the plasmodium falciparum epigenome that are dynamically marked by H3K9ac and H3K4me3. *PLoS pathogens* **6**(12): e1001223.

Belgareh-Touze N, Corral-Debrinski M, Launhardt H, Galan JM, Munder T, Le Panse S, Haguenauer-Tsapis R. 2003. Yeast functional analysis: identification of two essential genes involved in ER to Golgi trafficking. *Traffic* **4**(9): 607-617.

Bentley DR Balasubramanian S Swerdlow HP Smith GP Milton J Brown CG Hall KP Evers DJ Barnes CL Bignell HR et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218): 53-59.

Benton WD, Davis RW. 1977. Screening lambdagt recombinant clones by hybridization to single plaques in situ. *Science* **196**(4286): 180-182.

Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW, 3rd, Bulyk ML. 2006. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology* **24**(11): 1429-1435.

Berman J, Sudbery PE. 2002. Candida Albicans: a molecular revolution built on lessons from budding yeast. *Nat Rev Genet* **3**(12): 918-930.

Bernstein BE, Liu CL, Humphrey EL, Perlstein EO, Schreiber SL. 2004. Global nucleosome occupancy in yeast. *Genome biology* **5**(9): R62.

Berrar DP, Dubitzky W, Granzow M. 2003. *A Practical Approach To Microarray Data Analysis*. Kluwer Academic Publishers.

Bilguvar K, Ozturk AK, Louvi A, Kwan KY, Choi M, Tatli B, Yalnizoglu D, Tuysuz B, Caglayan AO, Gokben S et al. 2010. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature* **467**(7312): 207-210.

Birrell GW, Giaever G, Chu AM, Davis RW, Brown JM. 2001. A genome-wide screen in Saccharomyces cerevisiae for genes affecting UV radiation sensitivity. *Proc Natl Acad Sci U S A* **98**(22): 12608-12613.

Birren BW, Simon MI, Lai E. 1990. The basis of high resolution separation of small DNAs by asymmetric-voltage field inversion electrophoresis and its application to DNA sequencing gels. *Nucleic acids research* **18**(6): 1481-1487.

Black J. 1989. Drugs from emasculated hormones: the principle of syntopic antagonism. *Science* **245**(4917): 486-493.

Black JW, Crowther AF, Shanks RG, Smith LH, Dornhorst AC. 1964. A New Adrenergic Betareceptor Antagonist. *Lancet* **1**(7342): 1080-1081.

Blanchard AP, Kaiser R. J., Hood L. E. 1996. High-density oligonucleotide arrays. *Biosensors and Bioelectronics* **11**(6-7): 687-690.

Blankenship JR, Singh N, Alexander BD, Heitman J. 2005. Cryptococcus neoformans isolates from transplant recipients are not selected for resistance to calcineurin inhibitors by current immunosuppressive regimens. *J Clin Microbiol* **43**(1): 464-467.

Blat Y, Kleckner N. 1999. Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. *Cell* **98**(2): 249-259.

Blazquez J. 2003. Hypermutation as a factor contributing to the acquisition of antimicrobial resistance. *Clin Infect Dis* **37**(9): 1201-1209.

Botstein D, White RL, Skolnick M, Davis RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *American journal of human genetics* **32**(3): 314-331.

Bouchonville K, Forche A, Tang KE, Selmecki A, Berman J. 2009. Aneuploid chromosomes are highly unstable during DNA transformation of Candida albicans. *Eukaryotic cell* **8**(10): 1554-1566.

Bowers K, Lottridge J, Helliwell SB, Goldthwaite LM, Luzio JP, Stevens TH. 2004. Protein-protein interactions of ESCRT complexes in the yeast Saccharomyces cerevisiae. *Traffic* **5**(3): 194-210.

Boye E, Olsen BR. 2009. Signaling mechanisms in infantile hemangioma. *Curr Opin Hematol* **16**(3): 202-208.

Bozic I, Reiter JG, Allen B, Antal T, Chatterjee K, Shah P, Moon YS, Yaqubie A, Kelly N, Le DT et al. 2013. Evolutionary dynamics of cancer in response to targeted combination therapy. *elife* **2**: e00747.

Brasch MA, Hartley JL, Vidal M. 2004. ORFeome cloning and systems biology: standardized mass production of the parts from the parts-list. *Genome research* **14**(10B): 2001-2009.

Braslavsky I, Hebert B, Kartalov E, Quake SR. 2003. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America* **100**(7): 3960-3964.

Brazma A. 2009. Minimum Information About a Microarray Experiment (MIAME)--successes, failures, challenges. *ScientificWorldJournal* **9**: 420-423.

Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC et al. 2001. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature genetics* **29**(4): 365-371.

Brenner S. 2013. Reading the Human Genome. In *Human Biology Program Distinguished Lecture*, University of Toronto.

Breuert S, Allers T, Spohn G, Soppa J. 2006. Regulated polyploidy in halophilic archaea. *PloS one* **1**: e92.

Briand JF, Navarro F, Rematier P, Boschiero C, Labarre S, Werner M, Shpakovski GV, Thuriaux P. 2001. Partners of Rpb8p, a small subunit shared by yeast RNA polymerases I, II and III. *Mol Cell Biol* **21**(17): 6056-6065.

Brogaard K, Xi L, Wang JP, Widom J. 2012. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**(7404): 496-501.

Brown GD, Denning DW, Gow NA, Levitz SM, Netea MG, White TC. 2012a. Hidden killers: human fungal infections. *Sci Transl Med* **4**(165): 165rv113.

Brown GD, Denning DW, Levitz SM. 2012b. Tackling human fungal infections. *Science* **336**(6082): 647.

Buchdunger E, Zimmermann J, Mett H, Meyer T, Muller M, Druker BJ, Lydon NB. 1996. Inhibition of the Abl protein-tyrosine kinase in vitro and in vivo by a 2-phenylaminopyrimidine derivative. *Cancer Res* **56**(1): 100-104.

Buckholz RG, Gleeson MA. 1991. Yeast systems for the commercial production of heterologous proteins. *Biotechnology (N Y)* **9**(11): 1067-1072.

Buede R, Rinker-Schaffer C, Pinto WJ, Lester RL, Dickson RC. 1991. Cloning and characterization of LCB1, a Saccharomyces gene required for biosynthesis of the long-chain base component of sphingolipids. *J Bacteriol* **173**(14): 4325-4332.

Burks C, Fickett JW, Goad WB, Kanehisa M, Lewitter FI, Rindone WP, Swindell CD, Tung CS, Bilofsky HS. 1985. The GenBank nucleic acid sequence database. *Comput Appl Biosci* **1**(4): 225-233.

Bush K, Courvalin P, Dantas G, Davies J, Eisenstein B, Huovinen P, Jacoby GA, Kishony R, Kreiswirth BN, Kutter E et al. 2011. Tackling antibiotic resistance. *Nat Rev Microbiol* **9**(12): 894-896.

Bussey H, Kaback DB, Zhong W, Vo DT, Clark MW, Fortin N, Hall J, Ouellette BF, Keng T, Barton AB et al. 1995. The nucleotide sequence of chromosome I from Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* **92**(9): 3809-3813.

Butcher EC, Berg EL, Kunkel EJ. 2004. Systems biology in drug discovery. *Nat Biotechnol* **22**(10): 1253-1259.

Butcher RA, Bhullar BS, Perlstein EO, Marsischky G, LaBaer J, Schreiber SL. 2006. Microarray-based method for monitoring yeast overexpression strains reveals small-molecular targets in the TOR pathway. *Nature Chemical Biology* **2**: 103-109.

Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB. 2008. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome research* **18**(5): 810-820.

Cairns BR. 2009. The logic of chromatin architecture and remodelling at promoters. *Nature* **461**(7261): 193-198.

Cardenas ME, Muir RS, Breuder T, Heitman J. 1995. Targets of immunophilin-immunosuppressant complexes are distinct highly conserved regions of calcineurin A. *EMBO J* **14**(12): 2772-2783.

Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. 2012. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. *BMC genomics* **13**: 375.

Carroll SY, Stirling PC, Stimpson HE, Giesselmann E, Schmitt MJ, Drubin DG. 2009. A yeast killer toxin screen provides insights into a/b toxin entry, trafficking, and killing mechanisms. *Dev Cell* **17**(4): 552-560.

Chait R, Craney A, Kishony R. 2007. Antibiotic interactions that select against resistance. *Nature* **446**(7136): 668-671.

Chan JN, Nislow C, Emili A. 2009. Recent advances and method development for drug target identification. *Trends Pharmacol Sci*.

Chang GS, Noegel AA, Mavrich TN, Muller R, Tomsho LP, Ward E, Felder M, Jiang C, Eichinger L, Glockner G et al. 2012. Unusual combinatorial involvement of poly-A/T tracts in organizing genes and chromatin in Dictyostelium. *Genome research*.

Chang M, Bellaoui M, Boone C, Brown GW. 2002. A genome-wide screen for methyl methanesulfonate-sensitive mutants reveals genes required for S phase progression in the presence of DNA damage. *Proc Natl Acad Sci U S A* **99**(26): 16934-16939.

Chen B, Zhong D, Monteiro A. 2006. Comparative genomics and evolution of the HSP90 family of genes across all kingdoms of organisms. *BMC Genomics* **7**: 156.

Chen JK, Lane WS, Schreiber SL. 1999. The identification of myriocin-binding proteins. *Chem Biol* **6**(4): 221-235.

Chen Y, Feldman DE, Deng C, Brown JA, De Giacomo AF, Gaw AF, Shi G, Le QT, Brown JM, Koong AC. 2005. Identification of mitogen-activated protein kinase signaling pathways that confer resistance to endoplasmic reticulum stress in Saccharomyces cerevisiae. *Mol Cancer Res* **3**(12): 669-677.

Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M. 2012. Automated forward and reverse ratcheting of DNA in a nanopore at 5-A precision. *Nature biotechnology* **30**(4): 344-348.

Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK et al. 1997. Genetic and physical maps of Saccharomyces cerevisiae. *Nature* **387**(6632 Suppl): 67-73.

Cherry JM, Hong EL, Amundsen C, Balakrishnan R, Binkley G, Chan ET, Christie KR, Costanzo MC, Dwight SS, Engel SR et al. 2012. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic acids research* **40**(Database issue): D700-705.

Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T et al. 1998. Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* **282**(5396): 2022-2028.

Chidiac P, Hebert TE, Valiquette M, Dennis M, Bouvier M. 1994. Inverse agonist activity of beta-adrenergic antagonists. *Mol Pharmacol* **45**(3): 490-499.

Chiu MI, Katz H, Berlin V. 1994. RAPT1, a mammalian homolog of yeast Tor, interacts with the FKBP12/rapamycin complex. *Proc Natl Acad Sci U S A* **91**(26): 12574-12578.

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloglu A, Ozen S, Sanjad S et al. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* **106**(45): 19096-19101.

Chopra I. 2012. The 2012 Garrod Lecture: Discovery of antibacterial drugs in the 21st century. *The Journal of antimicrobial chemotherapy*.

Chung HR, Dunkel I, Heise F, Linke C, Krobitsch S, Ehrenhofer-Murray AE, Sperling SR, Vingron M. 2010. The effect of micrococcal nuclease digestion on nucleosome positioning data. *PloS one* **5**(12): e15754.

Church GM, Gao Y, Kosuri S. 2012. Next-generation digital information storage in DNA. *Science* **337**(6102): 1628.

Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G et al. 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature methods* **5**(7): 613-619.

Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. 2010. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic acids research* **38**(6): 1767-1771.

Cohen SN, Chang AC, Hsu L. 1972. Nonchromosomal antibiotic resistance in bacteria: genetic transformation of Escherichia coli by R-factor DNA. *Proceedings of the National Academy of Sciences of the United States of America* **69**(8): 2110-2114.

Collins SR, Kemmeren P, Zhao XC, Greenblatt JF, Spencer F, Holstege FC, Weissman JS, Krogan NJ. 2007a. Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol Cell Proteomics* **6**(3): 439-450.

Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M et al. 2007b. Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* **446**(7137): 806-810.

Connell C, Fung S, Heiner C, Bridgham J, Chakerian V, Heron E, Jones B, Menchen S, Mordan W, Raff M et al. 1987. Automated DNA sequence analysis. *Biotechniques* **5**: 342-348.

Conner BJ, Reyes AA, Morin C, Itakura K, Teplitz RL, Wallace RB. 1983. Detection of sickle cell beta S-globin allele by hybridization with synthetic oligonucleotides. *Proceedings of the National Academy of Sciences of the United States of America* **80**(1): 278-282.

Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, Ding H, Koh JL, Toufighi K, Mostafavi S et al. 2010. The genetic landscape of a cell. *Science* **327**(5964): 425-431.

Costanzo M, Giaever G, Nislow C, Andrews B. 2006. Experimental approaches to identify genetic networks. *Curr Opin Biotechnol* **17**(5): 472-480.

Cowen LE. 2008. The evolution of fungal drug resistance: modulating the trajectory from genotype to phenotype. *Nat Rev Microbiol* **6**(3): 187-198.

Cowen LE, Carpenter AE, Matangkasombut O, Fink GR, Lindquist S. 2006. Genetic architecture of Hsp90-dependent drug resistance. *Eukaryot Cell* **5**(12): 2184-2188.

Cowen LE, Kohn LM, Anderson JB. 2001. Divergence in fitness and evolution of drug resistance in experimental populations of Candida albicans. *J Bacteriol* **183**(10): 2971-2978.

Cowen LE, Lindquist S. 2005. Hsp90 potentiates the rapid evolution of new traits: drug resistance in diverse fungi. *Science* **309**(5744): 2185-2189.

Cowen LE, Sanglard D, Calabrese D, Sirjusingh C, Anderson JB, Kohn LM. 2000. Evolution of drug resistance in experimental populations of Candida albicans. *J Bacteriol* **182**(6): 1515-1522.

Cowen LE, Singh SD, Kohler JR, Collins C, Zaas AK, Schell WA, Aziz H, Mylonakis E, Perfect JR, Whitesell L et al. 2009. Harnessing Hsp90 function as a powerful, broadly effective therapeutic strategy for fungal infectious disease. *Proceedings of the National Academy of Sciences of the United States of America* **106**(8): 2818-2823.

Cowen LE, Steinbach WJ. 2008. Stress, drugs, and evolution: the role of cellular signaling in fungal drug resistance. *Eukaryot Cell* **7**(5): 747-764.

Crick FH. 1958. On protein synthesis. *Symp Soc Exp Biol* **12**: 138-163.

Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. 1961. General nature of the genetic code for proteins. *Nature* **192**: 1227-1232.

Cruz MC, Goldstein AL, Blankenship JR, Del Poeta M, Davis D, Cardenas ME, Perfect JR, McCusker JH, Heitman J. 2002. Calcineurin is essential for survival during membrane stress in Candida albicans. *The EMBO journal* **21**(4): 546-559.

Cyert MS, Kunisawa R, Kaim D, Thorner J. 1991. Yeast has homologs (CNA1 and CNA2 gene products) of mammalian calcineurin, a calmodulin-regulated phosphoprotein phosphatase. *Proceedings of the National Academy of Sciences of the United States of America* **88**(16): 7376-7380.

Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG. 2006. The affymetrix GeneChip platform: an overview. *Methods in enzymology* **410**: 3-28.

David L, Huber W, Granovskaia M, Toedling J, Palm CJ, Bofkin L, Jones T, Davis RW, Steinmetz LM. 2006. A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* **103**(14): 5320-5325.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* **43**(5): 491-498.

Derrington IM, Butler TZ, Collins MD, Manrao E, Pavlenok M, Niederweis M, Gundlach JH. 2010. Nanopore DNA sequencing with MspA. *Proceedings of the National Academy of Sciences of the United States of America* **107**(37): 16060-16065.

Dettman JR, Rodrigue N, Melnyk AH, Wong A, Bailey SF, Kassen R. 2012. Evolutionary insight from whole-genome sequencing of experimentally evolved microbes. *Mol Ecol* **21**(9): 2058-2077.

Dickson RC, Lester RL. 2002. Sphingolipid functions in Saccharomyces cerevisiae. *Biochim Biophys Acta* **1583**(1): 13-25.

Dixon SJ, Costanzo M, Baryshnikova A, Andrews B, Boone C. 2009. Systematic mapping of genetic interaction networks. *Annu Rev Genet* **43**: 601-625.

Dixon SJ, Stockwell BR. 2009. Identifying druggable disease-modifying gene products. *Curr Opin Chem Biol* **13**(5-6): 549-555.

Dolgin E, Motluk A. 2011. Heat shock and awe. *Nat Med* **17**(6): 646-649.

Dollard C, Ricupero-Hovasse SL, Natsoulis G, Boeke JD, Winston F. 1994. SPT10 and SPT21 are required for transcription of particular histone genes in Saccharomyces cerevisiae. *Molecular and cellular biology* **14**(8): 5223-5228.

Donis-Keller H, Green P, Helms C, Cartinhour S, Weiffenbach B, Stephens K, Keith TP, Bowden DW, Smith DR, Lander ES et al. 1987. A genetic linkage map of the human genome. *Cell* **51**(2): 319-337.

Dorn GW, 2nd. 2010. Refugee receptors switch sides. *Science* **327**(5973): 1586-1587.

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. 2003. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America* **100**(15): 8817-8822.

Drew HR, Travers AA. 1985. DNA bending and its relation to nucleosome positioning. *Journal of molecular biology* **186**(4): 773-790.

Drolet BA, Frommelt PC, Chamlin SL, Haggstrom A, Bauman NM, Chiu YE, Chun RH, Garzon MC, Holland KE, Liberman L et al. 2013. Initiation and use of propranolol for infantile hemangioma: report of a consensus conference. *Pediatrics* **131**(1): 128-140.

Druker BJ, Tamura S, Buchdunger E, Ohno S, Segal GM, Fanning S, Zimmermann J, Lydon NB. 1996. Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. *Nat Med* **2**(5): 561-566.

Du X, Takagi H. 2007. N-Acetyltransferase Mpr1 confers ethanol tolerance on Saccharomyces cerevisiae by reducing reactive oxygen species. *Applied microbiology and biotechnology* **75**(6): 1343-1351.

Dujon B. 2010. Yeast evolutionary genomics. *Nature reviews Genetics* **11**(7): 512-524.

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research* **30**(1): 207-210.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B et al. 2009. Real-time DNA sequencing from single polymerase molecules. *Science* **323**(5910): 133-138.

Engel SR, Balakrishnan R, Binkley G, Christie KR, Costanzo MC, Dwight SS, Fisk DG, Hirschman JE, Hitz BC, Hong EL et al. 2010. Saccharomyces Genome Database provides mutant phenotype data. *Nucleic acids research* **38**(Database issue): D433-436.

Ericson E, Gebbia M, Heisler LE, Wildenhain J, Tyers M, Giaever G, Nislow C. 2008. Off-target effects of psychoactive drugs revealed by genome-wide assays in yeast. *PLoS Genet* **4**(8): e1000151.

Ericson E, Hoon S, St Onge RP, Giaever G, Nislow C. 2010. Exploring gene function and drug action using chemogenomic dosage assays. *Methods Enzymol* **470**: 233-255.

Ewing B, Green P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome research* **8**(3): 186-194.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome research* **8**(3): 175-185.

Fahrner RL, Cascio D, Lake JA, Slesarev A. 2001. An ancestral nuclear protein assembly: crystal structure of the Methanopyrus kandleri histone. *Protein Sci* **10**(10): 2002-2007.

Farrar M. 2007. Striped Smith-Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics* **23**(2): 156-161.

Favre B, Didmon M, Ryder NS. 1999. Multiple amino acid substitutions in lanosterol 14alpha-demethylase contribute to azole resistance in Candida albicans. *Microbiology* **145 (Pt 10)**: 2715-2725.

Fedurco M, Romieu A, Williams S, Lawrence I, Turcatti G. 2006. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic acids research* **34**(3): e22.

Feldmann H, Aigle M, Aljinovic G, Andre B, Baclet MC, Barthe C, Baur A, Becam AM, Biteau N, Boles E et al. 1994. Complete DNA sequence of yeast chromosome II. *The EMBO journal* **13**(24): 5795-5809.

Ferragina P, Manzini G. 2000. Opportunistic data structures with applications. *Ann Ieee Symp Found*: 390-398.

Field Y, Kaplan N, Fondufe-Mittendorf Y, Moore IK, Sharon E, Lubling Y, Widom J, Segal E. 2008. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput Biol* **4**(11): e1000216.

Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Min Jou W, Molemans F, Raeymaekers A, Van den Berghe A et al. 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**(5551): 500-507.

Fiume M, Williams V, Brook A, Brudno M. 2010. Savant: genome browser for high-throughput sequencing data. *Bioinformatics* **26**(16): 1938-1944.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**(5223): 496-512.

Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. 1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* **251**(4995): 767-773.

Forbes AJ, Patrie SM, Taylor GK, Kim YB, Jiang L, Kelleher NL. 2004. Targeted analysis and discovery of posttranslational modifications in proteins from methanogenic archaea by top-down MS. *Proceedings of the National Academy of Sciences of the United States of America* **101**(9): 2678-2683.

Friedrich-Jahn U, Aigner J, Langst G, Reeve JN, Huber H. 2009. Nanoarchaeal origin of histone H3? *Journal of bacteriology* **191**(3): 1092-1096.

Gaali S, Gopalakrishnan R, Wang Y, Kozany C, Hausch F. 2011. The chemical biology of immunophilin ligands. *Curr Med Chem* **18**(35): 5355-5379.

Geer LY, Domrachev M, Lipman DJ, Bryant SH. 2002. CDART: protein homology by domain architecture. *Genome research* **12**(10): 1619-1623.

Gerhard DS Wagner L Feingold EA Shenmen CM Grouse LH Schuler G Klein SL Old S Rasooly R Good P et al. 2004. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res* **14**(10B): 2121-2127.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B et al. 2002. Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418**(6896): 387-391.

Giaever G, Flaherty P, Kumm J, Proctor M, Nislow C, Jaramillo DF, Chu AM, Jordan MI, Arkin AP, Davis RW. 2004. Chemogenomic profiling: identifying the functional interactions of small molecules in yeast. *Proc Natl Acad Sci U S A* **101**(3): 793-798.

Giaever G, Shoemaker DD, Jones TW, Liang H, Winzeler EA, Astromoff A, Davis RW. 1999. Genomic profiling of drug sensitivities via induced haploinsufficiency. *Nat Genet* **21**(3): 278-283.

Gietz RD, Schiestl RH. 2007. High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat Protoc* **2**(1): 31-34.

Gietz RD, Schiestl RH, Willems AR, Woods RA. 1995. Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure. *Yeast* **11**(4): 355-360.

Gkikopoulos T, Schofield P, Singh V, Pinskaya M, Mellor J, Smolle M, Workman JL, Barton GJ, Owen-Hughes T. 2011. A role for Snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization. *Science* **333**(6050): 1758-1760.

Glenn TC. 2011. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* **11**(5): 759-769.

Goeddel DV. 1991. *Gene Expression Technology*. Academic Press, Inc.

Goffeau A. 1997. The yeast genome directory. *Nature* **387**(6632 Suppl): 5.

Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M et al. 1996. Life with 6000 genes. *Science* **274**(5287): 546, 563-547.

Goodman LS, Hardman JG, Limbird LE, Gilman AG. 2001. *Goodman and Gilman's the pharmacological basis of therapeutics*. McGraw-Hill, New York.

Griffith F. 1928. The Significance of Pneumococcal Types. *J Hyg (Lond)* **27**(2): 113-159.

Gronenborn B, Messing J. 1978. Methylation of single-stranded DNA in vitro introduces new restriction endonuclease cleavage sites. *Nature* **272**(5651): 375-377.

Gunderson KL, Kruglyak S, Graige MS, Garcia F, Kermani BG, Zhao C, Che D, Dickinson T, Wickham E, Bierle J et al. 2004. Decoding randomly ordered DNA arrays. *Genome research* **14**(5): 870-877.

Hanahan D, Weinberg RA. 2011. Hallmarks of cancer: the next generation. *Cell* **144**(5): 646-674.

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW et al. 2008. Single-molecule DNA sequencing of a viral genome. *Science* **320**(5872): 106-109.

Hartley JL, Temple GF, Brasch MA. 2000. DNA cloning using in vitro site-specific recombination. *Genome Res* **10**(11): 1788-1795.

Hartman AL, Norais C, Badger JH, Delmas S, Haldenby S, Madupu R, Robinson J, Khouri H, Ren Q, Lowe TM et al. 2010. The complete genome sequence of Haloferax volcanii DS2, a model archaeon. *PloS one* **5**(3): e9605.

Hayden EC. 2012. Nanopore genome sequencer makes its debut. In *Nature News*, Vol 2013.

He S, Wurtzel O, Singh K, Froula JL, Yilmaz S, Tringe SG, Wang Z, Chen F, Lindquist EA, Sorek R et al. 2010. Validation of two ribosomal RNA removal methods for microbial metatranscriptomics. *Nature methods* **7**(10): 807-812.

Hedgpeth J, Goodman HM, Boyer HW. 1972. DNA nucleotide sequence restricted by the RI endonuclease. *Proceedings of the National Academy of Sciences of the United States of America* **69**(11): 3448-3452.

Hegreness M, Shoresh N, Damian D, Hartl D, Kishony R. 2008. Accelerated evolution of resistance in multidrug environments. *Proceedings of the National Academy of Sciences of the United States of America* **105**(37): 13977-13981.

Heitman J, Movva NR, Hall MN. 1991a. Targets for cell cycle arrest by the immunosuppressant rapamycin in yeast. *Science* **253**(5022): 905-909.

Heitman J, Movva NR, Hiestand PC, Hall MN. 1991b. FK 506-binding protein proline rotamase is a target for the immunosuppressive agent FK 506 in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America* **88**(5): 1948-1952.

Hemenway CS, Heitman J. 1999. Calcineurin. Structure, function, and inhibition. *Cell Biochem Biophys* **30**(1): 115-151.

Hendrych T, Kodedova M, Sigler K, Gaskova D. 2009. Characterization of the kinetics and mechanisms of inhibition of drugs interacting with the S. cerevisiae multidrug resistance pumps Pdr5p and Snq2p. *Biochim Biophys Acta* **1788**(3): 717-723.

Hershey AD, Chase M. 1952. Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J Gen Physiol* **36**(1): 39-56.

Higgins DG, Sharp PM. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**(1): 237-244.

Higgins MJ, Graham SJ. 2009. Intellectual property. Balancing innovation and access: patent challenges tip the scales. *Science* **326**(5951): 370-371.

Hill JA, Ammar R, Torti D, Nislow C, Cowen LE. 2013. Genetic and Genomic Architecture of the Evolution of Resistance to Antifungal Drug Combinations. *PLoS Genet* **9**(4): e1003390.

Hillenmeyer ME, Fung E, Wildenhain J, Pierce SE, Hoon S, Lee W, Proctor M, St.Onge RP, Tyers M, Koller D et al. 2008. The chemical genomic potrait of yeast: uncovering a phenotype for all genes. *Science* **320**(5874): 362-365.

Ho CH, Magtanong L, Barker SL, Gresham D, Nishimura S, Natarajan P, Koh JL, Porter J, Gray CA, Andersen RJ et al. 2009. A molecular barcoded yeast ORF library enables mode-of-action analysis of bioactive compounds. *Nat Biotechnol* **27**(4): 369-377.

Hogg RS, Heath KV, Yip B, Craib KJ, O'Shaughnessy MV, Schechter MT, Montaner JS. 1998. Improved survival among HIV-infected individuals following initiation of antiretroviral therapy. *JAMA* **279**(6): 450-454.

Hoheisel JD, Pohl FM. 1987. Searching for potential Z-DNA in genomic Escherichia coli DNA. *Journal of molecular biology* **193**(3): 447-464.

Hongay C, Jia N, Bard M, Winston F. 2002. Mot3 is a transcriptional repressor of ergosterol biosynthetic genes and is required for normal vacuolar function in Saccharomyces cerevisiae. *EMBO J* **21**(15): 4114-4124.

Hoon S, Smith AM, Wallace IM, Suresh S, Miranda M, Fung E, Proctor M, Shokat KM, Zhang C, Davis RW et al. 2008. An integrated platform of genomic assays reveals small-molecule bioactivities. *Nat Chem Biol* **4**(8): 498-506.

Hopkins AL, Groom CR. 2002. The druggable genome. *Nat Rev Drug Discov* **1**(9): 727-730.

Horn DL, Neofytos D, Anaissie EJ, Fishman JA, Steinbach WJ, Olyaei AJ, Marr KA, Pfaller MA, Chang CH, Webster KM. 2009. Epidemiology and outcomes of candidemia in 2019 patients: data from the prospective antifungal therapy alliance registry. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **48**(12): 1695-1703.

Hossain MS, Azimi N, Skiena S. 2009. Crystallizing short-read assemblies around seeds. *BMC Bioinformatics* **10 Suppl 1**: S16.

Hu Y, Rolfs A, Bhullar B, Murthy TV, Zhu C, Berger MF, Camargo AA, Kelley F, McCarron S, Jepson D et al. 2007. Approaching a complete repository of sequence-verified protein-encoding clones for Saccharomyces cerevisiae. *Genome Res* **17**(4): 536-543.

Huebert DJ, Kuan PF, Keles S, Gasch AP. 2012. Dynamic changes in nucleosome occupancy are not predictive of gene expression dynamics but are linked to transcription and chromatin regulators. *Molecular and cellular biology* **32**(9): 1645-1653.

Hughes AL, Jin Y, Rando OJ, Struhl K. 2012. A Functional Evolutionary Approach to Identify Determinants of Nucleosome Positioning: A Unifying Model for Establishing the Genome-wide Pattern. *Molecular cell*.

Hughes TR. 2002. Yeast and drug discovery. *Funct Integr Genomics* **2**(4-5): 199-211.

Hunkapiller T, Kaiser RJ, Koop BF, Hood L. 1991. Large-scale and automated DNA sequence determination. *Science* **254**(5028): 59-67.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. *Nature genetics* **36**(9): 949-951.

Imai J, Yahara I. 2000. Role of HSP90 in salt stress tolerance via stabilization and regulation of calcineurin. *Mol Cell Biol* **20**(24): 9262-9270.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research* **31**(4): e15.

Jacq C, Alt-Morbe J, Andre B, Arnold W, Bahr A, Ballesta JP, Bargues M, Baron L, Becker A, Biteau N et al. 1997. The nucleotide sequence of Saccharomyces cerevisiae chromosome IV. *Nature* **387**(6632 Suppl): 75-78.

Jeffrey KL, Camps M, Rommel C, Mackay CR. 2007. Targeting dual-specificity phosphatases: manipulating MAP kinase signalling and immune responses. *Nat Rev Drug Discov* **6**(5): 391-403.

Jiang C, Pugh BF. 2009. Nucleosome positioning and gene regulation: advances through genomics. *Nature reviews Genetics* **10**(3): 161-172.

Johnson DS, Mortazavi A, Myers RM, Wold B. 2007. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**(5830): 1497-1502.

Johnston M. 1987. A model fungal gene regulatory mechanism: the GAL genes of Saccharomyces cerevisiae. [Review]. *Microbiological Reviews* **51**(4): 458-476.

Jones GM, Stalker J, Humphray S, West A, Cox T, Rogers J, Dunham I, Prelich G. 2008. A systematic library for comprehensive overexpression screens in Saccharomyces cerevisiae. *Nat Methods* **5**(3): 239-241.

Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT et al. 2004. The diploid genome sequence of Candida albicans. *Proceedings of the National Academy of Sciences of the United States of America* **101**(19): 7329-7334.

Juneau K, Palm C, Miranda M, Davis RW. 2007. High-density yeast-tiling array reveals previously undiscovered introns and extensive regulation of meiotic splicing. *Proc Natl Acad Sci U S A* **104**(5): 1522-1527.

Kafatos FC, Jones CW, Efstratiadis A. 1979. Determination of nucleic acid sequence homologies and relative concentrations by a dot hybridization procedure. *Nucleic acids research* **7**(6): 1541-1552.

Kaminska KH, Bujnicki JM. 2008. Bacteriophage Mu Mom protein responsible for DNA modification is a new member of the acyltransferase superfamily. *Cell Cycle* **7**(1): 120-121.

Kanehisa M, Fickett JW, Goad WB. 1984. A relational database system for the maintenance and verification of the Los Alamos sequence library. *Nucleic acids research* **12**(1 Pt 1): 149-158.

Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, LeProust EM, Hughes TR, Lieb JD, Widom J et al. 2009. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature* **458**(7236): 362-366.

Karoor V, Vatner SF, Takagi G, Yang G, Thaisz J, Sadoshima J, Vatner DE. 2004. Propranolol prevents enhanced stress signaling in Gs alpha cardiomyopathy: potential mechanism for beta-blockade in heart failure. *J Mol Cell Cardiol* **36**(2): 305-312.

Kasianowicz JJ, Brandin E, Branton D, Deamer DW. 1996. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America* **93**(24): 13770-13773.

Keiser MJ, Setola V, Irwin JJ, Laggner C, Abbas AI, Hufeisen SJ, Jensen NH, Kuijer MB, Matos RC, Tran TB et al. 2009. Predicting new molecular targets for known drugs. *Nature* **462**(7270): 175-181.

Kellogg DA, Doctor BP, Loebel JE, Nirenberg MW. 1966. RNA codons and protein synthesis. IX. Synonym codon recognition by multiple species of valine-, alanine-, and methionine-sRNA. *Proceedings of the National Academy of Sciences of the United States of America* **55**(4): 912-919.

Kelly TJ, Jr., Smith HO. 1970. A restriction enzyme from Hemophilus influenzae. II. *Journal of molecular biology* **51**(2): 393-409.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome research* **12**(4): 656-664.

Khorana HG. 1968. Nucleic acid synthesis in the study of the genetic code. *Nobel Lecture*.

King K, Dohlman HG, Thorner J, Caron MG, Lefkowitz RJ. 1990. Control of yeast mating signal transduction by a mammalian beta 2-adrenergic receptor and Gs alpha subunit. *Science* **250**(4977): 121-123.

Kissinger CR, Parge HE, Knighton DR, Lewis CT, Pelletier LA, Tempczyk A, Kalish VJ, Tucker KD, Showalter RE, Moomaw EW et al. 1995. Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex. *Nature* **378**(6557): 641-644.

Kohlwein SD. 2010. Obese and anorexic yeasts: experimental models to understand the metabolic syndrome and lipotoxicity. *Biochim Biophys Acta* **1801**(3): 222-229.

Kolaczkowska A, Goffeau A. 1999. Regulation of pleiotropic drug resistance in yeast. *Drug Resist Updat* **2**(6): 403-414.

Krause SA, Xu H, Gray JV. 2008. The synthetic genetic network around PKC1 identifies novel modulators and components of protein kinase C signaling in Saccharomyces cerevisiae. *Eukaryot Cell* **7**(11): 1880-1887.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. *Genome research* **19**(9): 1639-1645.

Kukuruzinska MA, Lennon K. 1995. Diminished activity of the first N-glycosylation enzyme, dolichol-P-dependent N-acetylglucosamine-1-P transferase (GPT), gives rise to mutant phenotypes in yeast. *Biochim Biophys Acta* **1247**(1): 51-59.

Kukuruzinska MA, Robbins PW. 1987. Protein glycosylation in yeast: transcript heterogeneity of the ALG7 gene. *Proc Natl Acad Sci U S A* **84**(8): 2145-2149.

LaFayette SL, Collins C, Zaas AK, Schell WA, Betancourt-Quiroz M, Gunatilaka AA, Perfect JR, Cowen LE. 2010. PKC signaling regulates drug resistance of the fungal pathogen Candida albicans via circuitry comprised of Mkc1, calcineurin, and Hsp90. *PLoS Pathog* **6**(8): e1001069.

LaFramboise T. 2009. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research* **37**(13): 4181-4193.

Lamesch P, Li N, Milstein S, Fan C, Hao T, Szabo G, Hu Z, Venkatesan K, Bethel G, Martin P et al. 2007. hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics* **89**(3): 307-315.

Lander E. 2012. Secrets of the Human Genome. In *GSA Model Organisms to Human Biology: Cancer Genetics*, Washington, DC.

Lander ES, Green P. 1987. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* **84**(8): 2363-2367.

Lander ES Linton LM Birren B Nusbaum C Zody MC Baldwin J Devon K Dewar K Doyle M FitzHugh W et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**(6822): 860-921.

Lander ES, Waterman MS. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**(3): 231-239.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods* **9**(4): 357-359.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* **10**(3): R25.

Lazo JS, Nunes R, Skoko JJ, Queiroz de Oliveira PE, Vogt A, Wipf P. 2006. Novel benzofuran inhibitors of human mitogen-activated protein kinase phosphatase-1. *Bioorg Med Chem* **14**(16): 5643-5650.

Leaute-Labreze C, Dumas de la Roque E, Hubiche T, Boralevi F, Thambo JB, Taieb A. 2008. Propranolol for severe hemangiomas of infancy. *N Engl J Med* **358**(24): 2649-2651.

Lee W, St Onge RP, Proctor M, Flaherty P, Jordan MI, Arkin AP, Davis RW, Nislow C, Giaever G. 2005. Genome-Wide Requirements for Resistance to Functionally Distinct DNA-Damaging Agents. *PLoS Genet* **1**(2): e24.

Lee W, Tillo D, Bray N, Morse RH, Davis RW, Hughes TR, Nislow C. 2007. A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* **39**(10): 1235-1244.

Leppert G, McDevitt R, Falco SC, Van Dyk TK, Ficke MB, Golin J. 1990. Cloning by gene amplification of two loci conferring multiple drug resistance in Saccharomyces. *Genetics* **125**(1): 13-20.

Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. 2003. Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**(5607): 682-686.

Levene PA. 1919. The Structure of Yeast Nucleic Acid. IV. Ammonia Hydrolysis. *J Biol Chem* **40**: 415-424.

Levy J. 1994. Sequencing the yeast genome: an international achievement. *Yeast* **10**(13): 1689-1706.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

-. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**(5): 589-595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Li H, Homer N. 2010. A survey of sequence alignment algorithms for next-generation sequencing. *Brief Bioinform* **11**(5): 473-483.

Li H, Ruan J, Durbin R. 2008a. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome research* **18**(11): 1851-1858.

Li JW, Vederas JC. 2009. Drug discovery and natural products: end of an era or an endless frontier? *Science* **325**(5937): 161-165.

Li R Fan W Tian G Zhu H He L Cai J Huang Q Cai Q Li B Bai Y et al. 2010a. The sequence and de novo assembly of the giant panda genome. *Nature* **463**(7279): 311-317.

Li R, Li Y, Kristiansen K, Wang J. 2008b. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**(5): 713-714.

Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009b. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* **25**(15): 1966-1967.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K et al. 2010b. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* **20**(2): 265-272.

Li X, Zolli-Juran M, Cechetto JD, Daigle DM, Wright GD, Brown ED. 2004. Multicopy suppressors for novel antibacterial compounds reveal targets and drug efflux susceptibility. *Chem Biol* **11**(10): 1423-1430.

Link AJ, Olson MV. 1991. Physical map of the Saccharomyces cerevisiae genome at 110-kilobase resolution. *Genetics* **127**(4): 681-698.

Lipman DJ, Pearson WR. 1985. Rapid and sensitive protein similarity searches. *Science* **227**(4693): 1435-1441.

Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR. 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. *Cell* **133**(3): 523-536.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V et al. 2009. Population genomics of domestic and wild yeasts. *Nature* **458**(7236): 337-341.

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology* **14**(13): 1675-1680.

Lum PY, Armour CD, Stepaniants SB, Cavet G, Wolf MK, Butler JS, Hinshaw JC, Garnier P, Prestwich GD, Leonardson A et al. 2004. Discovering modes of action for therapeutic compounds using a genome-wide screen of yeast heterozygotes. *Cell* **116**(1): 121-137.

Lundquist PM, Zhong CF, Zhao P, Tomaney AB, Peluso PS, Dixon J, Bettman B, Lacroix Y, Kwo DP, McCullough E et al. 2008. Parallel confocal detection of single molecules in real time. *Opt Lett* **33**(9): 1026-1028.

Luo B, Cheung HW, Subramanian A, Sharifnia T, Okamoto M, Yang X, Hinkle G, Boehm JS, Beroukhim R, Weir BA et al. 2008. Highly parallel identification of essential genes in cancer cells. *Proc Natl Acad Sci U S A* **105**(51): 20380-20385.

Luo W, Sun W, Taldone T, Rodina A, Chiosis G. 2010. Heat shock protein 90 in neurodegenerative diseases. *Mol Neurodegener* **5**: 24.

Maillet M, Purcell NH, Sargent MA, York AJ, Bueno OF, Molkentin JD. 2008. DUSP6 (MKP3) null mice show enhanced ERK1/2 phosphorylation at baseline and increased myocyte proliferation in the heart affecting disease susceptibility. *J Biol Chem* **283**(45): 31246-31255.

Mandal S, Moudgil M, Mandal SK. 2009. Rational drug design. *Eur J Pharmacol* **625**(1-3): 90-100.

Manrao EA, Derrington IM, Laszlo AH, Langford KW, Hopper MK, Gillgren N, Pavlenok M, Niederweis M, Gundlach JH. 2012. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature biotechnology* **30**(4): 349-353.

Marchetti O, Moreillon P, Glauser MP, Bille J, Sanglard D. 2000. Potent synergism of the combination of fluconazole and cyclosporine in Candida albicans. *Antimicrob Agents Chemother* **44**(9): 2373-2381.

Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M et al. 2009. CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res* **37**(Database issue): D205-210.

Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR et al. 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic acids research* **39**(Database issue): D225-229.

Mardis ER. 2011. A decade's perspective on DNA sequencing technology. *Nature* **470**(7333): 198-203.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**(7057): 376-380.

Marqueling AL, Oza V, Frieden IJ, Puttgen KB. 2013. Propranolol and infantile hemangiomas four years later: a systematic review. *Pediatr Dermatol* **30**(2): 182-191.

Marr KA. 2010. Fungal infections in oncology patients: update on epidemiology, prevention, and treatment. *Curr Opin Oncol* **22**(2): 138-142.

Maskos U, Southern EM. 1992. Oligonucleotide hybridizations on glass supports: a novel linker for oligonucleotide synthesis and hybridization properties of oligonucleotides synthesised in situ. *Nucleic acids research* **20**(7): 1679-1684.

-. 1993. A novel method for the parallel analysis of multiple mutations in multiple samples. *Nucleic acids research* **21**(9): 2269-2270.

Masuda K, Shima H, Watanabe M, Kikuchi K. 2001. MKP-7, a novel mitogen-activated protein kinase phosphatase, functions as a shuttle protein. *J Biol Chem* **276**(42): 39002-39011.

Matthaei JH, Jones OW, Martin RG, Nirenberg MW. 1962. Characteristics and composition of RNA coding units. *Proceedings of the National Academy of Sciences of the United States of America* **48**: 666-677.

Matthews DA, Alden RA, Bolin JT, Freer ST, Hamlin R, Xuong N, Kraut J, Poe M, Williams M, Hoogsteen K. 1977. Dihydrofolate reductase: x-ray structure of the binary complex with methotrexate. *Science* **197**(4302): 452-455.

Maxam AM, Gilbert W. 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**(2): 560-564.

McClellan AJ, Xia Y, Deutschbauer AM, Davis RW, Gerstein M, Frydman J. 2007. Diverse cellular functions of the Hsp90 molecular chaperone uncovered using systems approaches. *Cell* **131**(1): 121-135.

McGary KL, Park TJ, Woods JO, Cha HJ, Wallingford JB, Marcotte EM. 2010. Systematic discovery of nonobvious human disease models through orthologous phenotypes. *Proc Natl Acad Sci U S A* **107**(14): 6544-6549.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**(9): 1297-1303.

McKernan K, Blanchard A, Kotler L, Costa G. 2006. Reagents, methods, and libraries for bead-based sequencing. Vol 20080003571 (ed. U patent).

McKernan KJ, Peckham HE, Costa GL, McLaughlin SF, Fu Y, Tsung EF, Clouser CR, Duncan C, Ichikawa JK, Lee CC et al. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome research* **19**(9): 1527-1541.

Meier KE, Gause KC, Wisehart-Johnson AE, Gore AC, Finley EL, Jones LG, Bradshaw CD, McNair AF, Ella KM. 1998. Effects of propranolol on phosphatidate phosphohydrolase and mitogen-activated protein kinase activities in A7r5 vascular smooth muscle cells. *Cell Signal* **10**(6): 415-426.

Mewes HW, Albermann K, Bahr M, Frishman D, Gleissner A, Hani J, Heumann K, Kleine K, Maierl A, Oliver SG et al. 1997. Overview of the yeast genome. *Nature* **387**(6632 Suppl): 7-65.

Michel JB, Yeh PJ, Chait R, Moellering RC, Jr., Kishony R. 2008. Drug interactions modulate the potential for evolution of resistance. *Proceedings of the National Academy of Sciences of the United States of America* **105**(39): 14918-14923.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP et al. 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**(7153): 553-560.

Miller JR, Koren S, Sutton G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics* **95**(6): 315-327.

Millson SH, Prodromou C, Piper PW. 2010. A simple yeast-based system for analyzing inhibitor resistance in the human cancer drug targets Hsp90alpha/beta. *Biochem Pharmacol* **79**(11): 1581-1588.

Min Jou W, Haegeman G, Ysebaert M, Fiers W. 1972. Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. *Nature* **237**(5350): 82-88.

Mitra RD, Church GM. 1999. In situ localized amplification and contact replication of many individual DNA molecules. *Nucleic acids research* **27**(24): e34.

Miyahara K, Mizunuma M, Hirata D, Tsuchiya E, Miyakawa T. 1996. The involvement of the Saccharomyces cerevisiae multidrug resistance transporters Pdr5p and Snq2p in cation resistance. *FEBS Lett* **399**(3): 317-320.

Miyano S. 2005. *Research in computational molecular biology : 9th annual International Conference, RECOMB 2005, Cambridge, MA, USA, May 14-18, 2005 : proceedings*. Springer, Berlin.

Moffat J, Grueneberg DA, Yang X, Kim SY, Kloepfer AM, Hinkle G, Piqani B, Eisenhaure TM, Luo B, Grenier JK et al. 2006. A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**(6): 1283-1298.

Molin M, Norbeck J, Blomberg A. 2003. Dihydroxyacetone kinases in Saccharomyces cerevisiae are involved in detoxification of dihydroxyacetone. *J Biol Chem* **278**(3): 1415-1423.

Molina G, Vogt A, Bakan A, Dai W, Queiroz de Oliveira P, Znosko W, Smithgall TE, Bahar I, Lazo JS, Day BW et al. 2009. Zebrafish chemical screening reveals an inhibitor of Dusp6 that expands cardiac cell lineages. *Nat Chem Biol* **5**(9): 680-687.

Montgomery R, Swenson CA. 1976. *Quantitative problems in the biochemical sciences*. W. H. Freeman, New York.

Morlock KR, McLaughlin JJ, Lin YP, Carman GM. 1991. Phosphatidate phosphatase from Saccharomyces cerevisiae. Isolation of 45- and 104-kDa forms of the enzyme that are differentially regulated by inositol. *J Biol Chem* **266**(6): 3586-3593.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods* **5**(7): 621-628.

Mullakhanbhai MF, Larsen H. 1975. Halobacterium volcanii spec. nov., a Dead Sea halobacterium with a moderate salt requirement. *Arch Microbiol* **104**(3): 207-214.

Mushegian AR, Bassett DE, Jr., Boguski MS, Bork P, Koonin EV. 1997. Positionally cloned human disease genes: patterns of evolutionary conservation and functional motifs. *Proc Natl Acad Sci U S A* **94**(11): 5831-5836.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**(5881): 1344-1349.

Nakagawa F, Lodwick RK, Smith CJ, Smith R, Cambiano V, Lundgren JD, Delpech V, Phillips AN. 2012. Projected life expectancy of people with HIV according to timing of diagnosis. *AIDS* **26**(3): 335-343.

Navarre WW, Porwollik S, Wang Y, McClelland M, Rosen H, Libby SJ, Fang FC. 2006. Selective silencing of foreign DNA with low GC content by the H-NS protein in Salmonella. *Science* **313**(5784): 236-238.

Neckers L, Workman P. 2012. Hsp90 molecular chaperone inhibitors: are we there yet? *Clin Cancer Res* **18**(1): 64-76.

Nekrutenko A, Taylor J. 2012. Next-generation sequencing data interpretation: enhancing reproducibility and accessibility. *Nature reviews Genetics* **13**(9): 667-672.

Ng R. 2009. *Drugs : from discovery to approval*. John Wiley & Sons, Hoboken, N.J.

Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, Gildersleeve HI, Beck AE, Tabor HK, Cooper GM, Mefford HC et al. 2010. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature genetics* **42**(9): 790-793.

Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA et al. 2009. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics* **42**(1): 30-35.

Nguyen HH, de la Tour CB, Toueille M, Vannier F, Sommer S, Servant P. 2009. The essential histone-like protein HU plays a major role in Deinococcus radiodurans nucleoid compaction. *Mol Microbiol* **73**(2): 240-252.

Ning Z, Cox AJ, Mullikin JC. 2001. SSAHA: a fast search method for large DNA databases. *Genome research* **11**(10): 1725-1729.

Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, Rottman F, O'Neal C. 1965. RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proceedings of the National Academy of Sciences of the United States of America* **53**(5): 1161-1168.

Nirenberg MW, Matthaei JH. 1961. The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. *Proceedings of the National Academy of Sciences of the United States of America* **47**: 1588-1602.

Nishimura S, Arita Y, Honda M, Iwamoto K, Matsuyama A, Shirai A, Kawasaki H, Kakeya H, Kobayashi T, Matsunaga S et al. 2010. Marine antifungal theonellamides target 3beta-hydroxysterol to activate Rho1 signaling. *Nat Chem Biol* **6**(7): 519-526.

Noble SM, French S, Kohn LA, Chen V, Johnson AD. 2010. Systematic screens of a Candida albicans homozygous deletion library decouple morphogenetic switching and pathogenicity. *Nature genetics* **42**(7): 590-598.

Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E, Halliday K, Kamerewerd J, Kempken F, Knab B, Kuo HC et al. 2010. De novo assembly of a 40 Mb eukaryotic genome from short sequence reads: Sordaria macrospora, a model organism for fungal morphogenesis. *PLoS genetics* **6**(4): e1000891.

Okoniewski MJ, Miller CJ. 2008. Comprehensive analysis of affymetrix exon arrays using BioConductor. *PLoS Comput Biol* **4**(2): e6.

Oliphant A, Barker DL, Stuelpnagel JR, Chee MS. 2002. BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *BioTechniques* **Suppl**: 56-58, 60-51.

Oliver SG, van der Aart QJ, Agostoni-Carbone ML, Aigle M, Alberghina L, Alexandraki D, Antoine G, Anwar R, Ballesta JP, Benit P et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature* **357**(6373): 38-46.

Olson MV, Dutchik JE, Graham MY, Brodeur GM, Helms C, Frank M, MacCollin M, Scheinman R, Frank T. 1986. Random-clone strategy for genomic restriction mapping in yeast. *Proceedings of the National Academy of Sciences of the United States of America* **83**(20): 7826-7830.

Onyewu C, Blankenship JR, Del Poeta M, Heitman J. 2003. Ergosterol biosynthesis inhibitors become fungicidal when combined with calcineurin inhibitors against Candida albicans, Candida glabrata, and Candida krusei. *Antimicrob Agents Chemother* **47**(3): 956-964.

Osborn MJ, Miller JR. 2007. Rescuing yeast mutants with human genes. *Brief Funct Genomic Proteomic* **6**(2): 104-111.

Palella FJ, Jr., Delaney KM, Moorman AC, Loveless MO, Fuhrer J, Satten GA, Aschman DJ, Holmberg SD. 1998. Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. *N Engl J Med* **338**(13): 853-860.

Palmer JR, Daniels CJ. 1995. In vivo definition of an archaeal promoter. *Journal of bacteriology* **177**(7): 1844-1849.

Parodi AJ. 2000. Protein glucosylation and its role in protein folding. *Annu Rev Biochem* **69**: 69-93.

Parsons AB, Brost RL, Ding H, Li Z, Zhang C, Sheikh B, Brown GW, Kane PM, Hughes TR, Boone C. 2004. Integration of chemical-genetic and genetic interaction data links bioactive compounds to cellular target pathways. *Nat Biotechnol* **22**(1): 62-69.

Parsons AB, Lopez A, Givoni IE, Williams DE, Gray CA, Porter J, Chua G, Sopko R, Brost RL, Ho CH et al. 2006. Exploring the mode-of-action of bioactive compounds by chemical-genetic profiling in yeast. *Cell* **126**(3): 611-625.

Paszkiewicz K, Studholme DJ. 2010. De novo assembly of short sequence reads. *Brief Bioinform* **11**(5): 457-472.

Pauling L, Corey RB. 1953. A Proposed Structure For The Nucleic Acids. *Proceedings of the National Academy of Sciences of the United States of America* **39**(2): 84-97.

Paulsen IT, Sliwinski MK, Nelissen B, Goffeau A, Saier MH, Jr. 1998. Unified inventory of established and putative transporters encoded within the complete genome of Saccharomyces cerevisiae. *FEBS Lett* **430**(1-2): 116-125.

Paya CV. 1993. Fungal infections in solid-organ transplantation. *Clin Infect Dis* **16**(5): 677-688.

Pena-Castillo L, Hughes TR. 2007. Why are there still over 1000 uncharacterized yeast genes? *Genetics* **176**(1): 7-14.

Pennisi E. 2012. Genome sequencing. Search for pore-fection. *Science* **336**(6081): 534-537.

Pereira SL, Grayling RA, Lurz R, Reeve JN. 1997. Archaeal nucleosomes. *Proceedings of the National Academy of Sciences of the United States of America* **94**(23): 12633-12637.

Pereira SL, Reeve JN. 1998. Histones and nucleosomes in Archaea and Eukarya: a comparative analysis. *Extremophiles* **2**(3): 141-148.

Pfaller MA. 2012. Antifungal drug resistance: mechanisms, epidemiology, and consequences for treatment. *Am J Med* **125**(1 Suppl): S3-13.

Pfaller MA, Diekema DJ. 2007. Epidemiology of invasive candidiasis: a persistent public health problem. *Clin Microbiol Rev* **20**(1): 133-163.

-. 2010. Epidemiology of invasive mycoses in North America. *Critical reviews in microbiology* **36**(1): 1-53.

Pierce SE, Davis RW, Nislow C, Giaever G. 2007. Genome-wide analysis of barcoded Saccharomyces cerevisiae gene-deletion mutants in pooled cultures. *Nat Protoc* **2**(11): 2958-2974.

Pierce SE, Fung EL, Jaramillo DF, Chu AM, Davis RW, Nislow C, Giaever G. 2006. A unique and universal molecular barcode array. *Nat Methods* **3**(8): 601-603.

Poncz M, Solowiejczyk D, Ballantine M, Schwartz E, Surrey S. 1982. "Nonrandom" DNA sequence analysis in bacteriophage M13 by the dideoxy chain-termination method. *Proceedings of the National Academy of Sciences of the United States of America* **79**(14): 4298-4302.

Pramanik K, Chun CZ, Garnaas MK, Samant GV, Li K, Horswill MA, North PE, Ramchandran R. 2009. Dusp-5 and Snrk-1 coordinately function during vascular development and disease. *Blood* **113**(5): 1184-1191.

Prober JM, Trainor GL, Dam RJ, Hobbs FW, Robertson CW, Zagursky RJ, Cocuzza AJ, Jensen MA, Baumeister K. 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**(4825): 336-341.

Proctor M, Urbanus ML, Fung EL, Jaramillo DF, Davis RW, Nislow C, Giaever G. 2011. The automated cell: compound and environment screening system (ACCESS) for chemogenomic screening. *Methods in molecular biology* **759**: 239-269.

Prodromou C, Nuttall JM, Millson SH, Roe SM, Sim TS, Tan D, Workman P, Pearl LH, Piper PW. 2009. Structural basis of the radicicol resistance displayed by a fungal hsp90. *ACS Chem Biol* **4**(4): 289-297.

Provart NJ, McCourt P. 2004. Systems approaches to understanding cell signaling and gene regulation. *Curr Opin Plant Biol* **7**(5): 605-609.

Ptacek J, Devgan G, Michaud G, Zhu H, Zhu X, Fasolo J, Guo H, Jona G, Breitkreutz A, Sopko R et al. 2005. Global analysis of protein phosphorylation in yeast. *Nature* **438**(7068): 679-684.

Puerta C, Hernandez F, Gutierrez C, Pineiro M, Lopez-Alarcon L, Palacian E. 1993. Efficient transcription of a DNA template associated with histone (H3.H4)2 tetramers. *The Journal of biological chemistry* **268**(35): 26663-26667.

Pullar CE, Grahn JC, Liu W, Isseroff RR. 2006. Beta2-adrenergic receptor activation delays wound healing. *FASEB J* **20**(1): 76-86.

Pushkarev D, Neff NF, Quake SR. 2009. Single-molecule sequencing of an individual human genome. *Nature biotechnology* **27**(9): 847-850.

Putney SD, Herlihy WC, Schimmel P. 1983. A new troponin T and cDNA clones for 13 different muscle proteins, found by shotgun sequencing. *Nature* **302**(5910): 718-721.

Qian F, Deng J, Cheng N, Welch EJ, Zhang Y, Malik AB, Flavell RA, Dong C, Ye RD. 2009. A non-redundant role for MKP5 in limiting ROS production and preventing LPS-induced vascular injury. *EMBO J* **28**(19): 2896-2907.

Quon K, Kassner PD. 2009. RNA interference screening for the discovery of oncology targets. *Expert Opin Ther Targets* **13**(9): 1027-1035.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W et al. 2006. Global variation in copy number in the human genome. *Nature* **444**(7118): 444-454.

Reedy JL, Husain S, Ison M, Pruett TL, Singh N, Heitman J. 2006. Immunotherapy with tacrolimus (FK506) does not select for resistance to calcineurin inhibitors in Candida albicans isolates from liver transplant patients. *Antimicrob Agents Chemother* **50**(4): 1573-1577.

Rice PM, Elliston K, Lüthy R, Eisenberg D, States DJ, Boguski MS, Caballero L. 1991. *Sequence analysis primer*. Oxford University Press.

Richmond TJ, Davey CA. 2003. The structure of DNA in the nucleosome core. *Nature* **423**(6936): 145-150.

Richterich P. 1998. Estimation of errors in "raw" DNA sequences: a validation study. *Genome research* **8**(3): 251-259.

Rine J, Hansen W, Hardeman E, Davis RW. 1983. Targeted selection of recombinant clones through gene dosage effects. *Proc Natl Acad Sci U S A* **80**(22): 6750-6754.

Ritz C, Streibig JC. 2008. *Nonlinear regression with R*. Springer, New York.

Rix U, Superti-Furga G. 2009. Target profiling of small molecules by chemical proteomics. *Nat Chem Biol* **5**(9): 616-624.

Rizzo JM, Mieczkowski PA, Buck MJ. 2011. Tup1 stabilizes promoter nucleosome positioning and occupancy at transcriptionally plastic genes. *Nucleic acids research* **39**(20): 8803-8819.

Robbins N, Collins C, Morhayim J, Cowen LE. Metabolic control of antifungal drug resistance. *Fungal Genet Biol* **47**(2): 81-93.

Robbins N, Uppuluri P, Nett J, Rajendran R, Ramage G, Lopez-Ribot JL, Andes D, Cowen LE. 2011. Hsp90 governs dispersion and drug resistance of fungal biofilms. *PLoS Pathog* **7**(9): e1002257.

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A et al. 2007. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods* **4**(8): 651-657.

Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nature biotechnology* **29**(1): 24-26.

Ronaghi M. 2001. Pyrosequencing sheds light on DNA sequencing. *Genome research* **11**(1): 3-11.

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry* **242**(1): 84-89.

Root DE, Hacohen N, Hahn WC, Lander ES, Sabatini DM. 2006. Genome-scale loss-of-function screening with a lentiviral RNAi library. *Nat Methods* **3**(9): 715-719.

Roth A, Ding J, Morin R, Crisan A, Ha G, Giuliany R, Bashashati A, Hirst M, Turashvili G, Oloumi A et al. 2012. JointSNVMix: a probabilistic model for accurate detection of somatic mutations in normal/tumour paired next-generation sequencing data. *Bioinformatics* **28**(7): 907-913.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M et al. 2011. An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**(7356): 348-352.

Rouillard JM, Zuker M, Gulari E. 2003. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. *Nucleic acids research* **31**(12): 3057-3062.

Rual JF, Hill DE, Vidal M. 2004a. ORFeome projects: gateway between genomics and omics. *Curr Opin Chem Biol* **8**(1): 20-25.

Rual JF, Hirozane-Kishikawa T, Hao T, Bertin N, Li S, Dricot A, Li N, Rosenberg J, Lamesch P, Vidalain PO et al. 2004b. Human ORFeome version 1.1: a platform for reverse proteomics. *Genome Res* **14**(10B): 2128-2135.

Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. 2009. SHRiMP: accurate mapping of short color-space reads. *PLoS Comput Biol* **5**(5): e1000386.

Sabatini DM, Erdjument-Bromage H, Lui M, Tempst P, Snyder SH. 1994. RAFT1: a mammalian protein that binds to FKBP12 in a rapamycin-dependent fashion and is homologous to yeast TORs. *Cell* **78**(1): 35-43.

Sajan SA, Hawkins RD. 2012. Methods for Identifying Higher-Order Chromatin Structure. *Annu Rev Genomics Hum Genet*.

Sambrook J, Russell DW. 2001. *Molecular cloning : a laboratory manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.

Sandman K, Reeve JN. 2006. Archaeal histones and the origin of the histone fold. *Curr Opin Microbiol* **9**(5): 520-525.

Sanger F. 1981. Determination of nucleotide sequences in DNA. *Science* **214**(4526): 1205-1210.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M. 1977a. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**(5596): 687-695.

Sanger F, Brownlee GG, Barrell BG. 1965. A two-dimensional fractionation procedure for radioactive nucleotides. *Journal of molecular biology* **13**(2): 373-398.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology* **94**(3): 441-448.

Sanger F, Coulson AR, Barrell BG, Smith AJ, Roe BA. 1980. Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. *Journal of molecular biology* **143**(2): 161-178.

Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. 1982. Nucleotide sequence of bacteriophage lambda DNA. *Journal of molecular biology* **162**(4): 729-773.

Sanger F, Nicklen S, Coulson AR. 1977b. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**(12): 5463-5467.

Sanglard D. 2002. Resistance of human fungal pathogens to antifungal drugs. *Curr Opin Microbiol* **5**(4): 379-385.

Sanglard D, Ischer F, Marchetti O, Entenza J, Bille J. 2003. Calcineurin A of Candida albicans: involvement in antifungal tolerance, cell morphogenesis and virulence. *Mol Microbiol* **48**(4): 959-976.

Sans V, Dumas de la Roque E, Berge J, Grenier N, Boralevi F, Mazereeuw-Hautier J, Lipsker D, Dupuis E, Ezzedine K, Vergnes P et al. 2009. Propranolol for Severe Infantile Hemangiomas: Follow-Up Report. *Pediatrics*: e423-e431.

Sartorius-Neef S, Pfeifer F. 2004. In vivo studies on putative Shine-Dalgarno sequences of the halophilic archaeon Halobacterium salinarum. *Mol Microbiol* **51**(2): 579-588.

Sasse C, Dunkel N, Schafer T, Schneider S, Dierolf F, Ohlsen K, Morschhauser J. 2012. The stepwise acquisition of fluconazole resistance mutations causes a gradual loss of fitness in Candida albicans. *Mol Microbiol* **86**(3): 539-556.

Satchwell SC, Drew HR, Travers AA. 1986. Sequence periodicities in chicken nucleosome core DNA. *Journal of molecular biology* **191**(4): 659-675.

Sato S, Murata A, Shirakawa T, Uesugi M. 2010. Biochemical target isolation for novices: affinity-based strategies. *Chemistry & biology* **17**(6): 616-623.

Schena M, Shalon D, Davis RW, Brown PO. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**(5235): 467-470.

Scherf U, Ross DT, Waltham M, Smith LH, Lee JK, Tanabe L, Kohn KW, Reinhold WC, Myers TG, Andrews DT et al. 2000. A gene expression database for the molecular pharmacology of cancer. *Nat Genet* **24**(3): 236-244.

Schreiber SL. 2000. Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287**(5460): 1964-1969.

Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF et al. 2005. Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123**(3): 507-519.

Schuldiner M, Metz J, Schmid V, Denic V, Rakwalska M, Schmitt HD, Schwappach B, Weissman JS. 2008. The GET complex mediates insertion of tail-anchored proteins into the ER membrane. *Cell* **134**(4): 634-645.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305**(5683): 525-528.

Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J. 2006. A genomic code for nucleosome positioning. *Nature* **442**(7104): 772-778.

Selmecki A, Forche A, Berman J. 2006. Aneuploidy and isochromosome formation in drug-resistant Candida albicans. *Science* **313**(5785): 367-370.

-. 2010. Genomic plasticity of the human fungal pathogen Candida albicans. *Eukaryot Cell* **9**(7): 991-1008.

Selmecki A, Gerami-Nejad M, Paulson C, Forche A, Berman J. 2008. An isochromosome confers drug resistance in vivo by amplification of two genes, ERG11 and TAC1. *Mol Microbiol* **68**(3): 624-641.

Selmecki AM, Dulmage K, Cowen LE, Anderson JB, Berman J. 2009a. Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance. *PLoS genetics* **5**(10): e1000705.

-. 2009b. Acquisition of aneuploidy provides increased fitness during the evolution of antifungal drug resistance. *PLoS Genet* **5**(10): e1000705.

SGD SGD. Systematic Sequencing Table: Saccharomyces cerevisiae chromosome statistics and retrieval options. **2013**.

Shapiro RS, Robbins N, Cowen LE. 2011. Regulatory Circuitry Governing Fungal Development, Drug Resistance, and Disease. *Microbiol Mol Biol Rev* **75**(2).

Shchepinov MS, Case-Green SC, Southern EM. 1997. Steric factors influencing hybridisation of nucleic acids to oligonucleotide arrays. *Nucleic acids research* **25**(6): 1155-1161.

Shendure J, Ji H. 2008. Next-generation DNA sequencing. *Nature biotechnology* **26**(10): 1135-1145.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**(5741): 1728-1732.

Shendure JA, Porreca GJ, Church GM. 2008. Overview of DNA sequencing strategies. *Curr Protoc Mol Biol* **Chapter 7**: Unit 7 1.

Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, Iyer VR. 2008. Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS biology* **6**(3): e65.

Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW. 1996. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat Genet* **14**(4): 450-456.

Sidhu SS, Bader GD, Boone C. 2003. Functional genomics of intracellular peptide recognition domains with combinatorial biology methods. *Curr Opin Chem Biol* **7**(1): 97-102.

Siegfried EC, Keenan WJ, Al-Jureidini S. 2008. More on propranolol for hemangiomas of infancy. *N Engl J Med* **359**(26): 2846; author reply 2846-2847.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome research* **19**(6): 1117-1123.

Singh SD, Robbins N, Zaas AK, Schell WA, Perfect JR, Cowen LE. 2009. Hsp90 governs echinocandin resistance in the pathogenic yeast Candida albicans via calcineurin. *PLoS Pathog* **5**(7): e1000532.

Sionov E, Lee H, Chang YC, Kwon-Chung KJ. 2010. Cryptococcus neoformans overcomes stress of azole drugs by formation of disomy in specific multiple chromosomes. *PLoS pathogens* **6**(4): e1000848.

Skrzypek MS, Arnaud MB, Costanzo MC, Inglis DO, Shah P, Binkley G, Miyasato SR, Sherlock G. 2010. New tools at the Candida Genome Database: biochemical pathways and full-text literature search. *Nucleic acids research* **38**(Database issue): D428-432.

Smith AM, Heisler LE, Mellor J, Kaper F, Thompson MJ, Chee M, Roth FP, Giaever G, Nislow C. 2009. Quantitative phenotyping via deep barcode sequencing. *Genome research* **19**(10): 1836-1842.

Smith AM, Heisler LE, St Onge RP, Farias-Hesson E, Wallace IM, Bodeau J, Harris AN, Perry KM, Giaever G, Pourmand N et al. 2010. Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic acids research* **38**(13): e142.

Smith HO, Wilcox KW. 1970. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *Journal of molecular biology* **51**(2): 379-391.

Smith SW. 1997. *The scientist and engineer's guide to digital signal processing*. California Technical Pub., San Diego, Calif.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *Journal of molecular biology* **147**(1): 195-197.

Snyder M, Gallagher JE. 2009. Systems biology from a yeast omics perspective. *FEBS Lett* **583**(24): 3895-3899.

Sobel JD. 2007. Vulvovaginal candidosis. *Lancet* **369**(9577): 1961-1971.

Song CM, Lim SJ, Tong JC. 2009. Recent advances in computer-aided drug design. *Brief Bioinform* **10**(5): 579-591.

Sopko R, Huang D, Preston N, Chua G, Papp B, Kafadar K, Snyder M, Oliver SG, Cyert M, Hughes TR et al. 2006. Mapping pathways and phenotypes by systematic gene overexpression. *Mol Cell* **21**(3): 319-330.

Southern EM. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* **98**(3): 503-517.

-. 2001. DNA microarrays. History and overview. *Methods in molecular biology* **170**: 1-15.

Southern EM, Maskos U, Elder JK. 1992. Analyzing and comparing nucleic acid sequences by hybridization to arrays of oligonucleotides: evaluation using experimental models. *Genomics* **13**(4): 1008-1017.

Sozzani S, Agwu DE, McCall CE, O'Flaherty JT, Schmitt JD, Kent JD, McPhail LC. 1992. Propranolol, a phosphatidate phosphohydrolase inhibitor, also inhibits protein kinase C. *J Biol Chem* **267**(28): 20481-20488.

Spitzer M, Griffiths E, Blakely KM, Wildenhain J, Ejim L, Rossi L, De Pascale G, Curak J, Brown E, Tyers M et al. 2011. Cross-species discovery of syncretic drug combinations that potentiate the antifungal fluconazole. *Mol Syst Biol* **7**: 499.

St Onge RP, Mani R, Oh J, Proctor M, Fung E, Davis RW, Nislow C, Roth FP, Giaever G. 2007. Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* **39**(2): 199-206.

Staden R. 1980. A new computer method for the storage and manipulation of DNA gel reading data. *Nucleic acids research* **8**(16): 3673-3694.

Stapleton MP. 1997. Sir James Black and propranolol. The role of the basic sciences in the history of cardiovascular pharmacology. *Tex Heart Inst J* **24**(4): 336-342.

Steinbach WJ, Reedy JL, Cramer RA, Jr., Perfect JR, Heitman J. 2007. Harnessing calcineurin as a novel anti-infective agent against invasive fungal infections. *Nat Rev Microbiol* **5**(6): 418-430.

Stockwell BR. 2004. Exploring biology with small organic molecules. *Nature* **432**(7019): 846-854.

Strausberg RL, Feingold EA, Grouse LH, Derge JG, Klausner RD, Collins FS, Wagner L, Shenmen CM, Schuler GD, Altschul SF et al. 2002. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci U S A* **99**(26): 16899-16903.

Strausberg RL, Feingold EA, Klausner RD, Collins FS. 1999. The mammalian gene collection. *Science* **286**(5439): 455-457.

Studier FW. 1989. A strategy for high-volume sequencing of cosmid DNAs: random and directed priming with a library of oligonucleotides. *Proceedings of the National Academy of Sciences of the United States of America* **86**(18): 6917-6921.

Sutterlin C, Doering TL, Schimmoller F, Schroder S, Riezman H. 1997. Specific requirements for the ER to Golgi transport of GPI-anchored proteins in yeast. *J Cell Sci* **110 (Pt 21)**: 2703-2714.

Suzuki Y, Onge RP, Mani R, King OD, Heilbut A, Labunskyy VM, Chen W, Pham L, Zhang LV, Tong AH et al. 2011. Knocking out multigene redundancies via cycles of sexual assortment and fluorescence selection. *Nat Methods*.

Swerdlow H, Gesteland R. 1990. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic acids research* **18**(6): 1415-1419.

Taipale M, Jarosz DF, Lindquist S. 2010. HSP90 at the hub of protein homeostasis: emerging mechanistic insights. *Nat Rev Mol Cell Biol* **11**(7): 515-528.

Takahashi K, Yamanaka S. 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**(4): 663-676.

Talbert PB, Henikoff S. 2010. Histone variants--ancient wrap artists of the epigenome. *Nat Rev Mol Cell Biol* **11**(4): 264-275.

Tan SX, Teo M, Lam YT, Dawes IW, Perrone GG. 2009. Cu, Zn superoxide dismutase and NADP(H) homeostasis are required for tolerance of endoplasmic reticulum stress in Saccharomyces cerevisiae. *Mol Biol Cell* **20**(5): 1493-1508.

Tanaka Y, Tawaramoto-Sasanuma M, Kawaguchi S, Ohta T, Yoda K, Kurumizaka H, Yokoyama S. 2004. Expression and purification of recombinant human histones. *Methods* **33**(1): 3-11.

Tang YC, Williams BR, Siegel JJ, Amon A. 2011. Identification of aneuploidy-selective antiproliferation compounds. *Cell* **144**(4): 499-512.

Temple G Gerhard DS Rasooly R Feingold EA Good PJ Robinson C Mandich A Derge JG Lewis J Shoaf D et al. 2009. The completion of the Mammalian Gene Collection (MGC). *Genome research* **19**(12): 2324-2333.

Teotico DG, Babaoglu K, Rocklin GJ, Ferreira RS, Giannetti AM, Shoichet BK. 2009. Docking for fragment inhibitors of AmpC beta-lactamase. *Proc Natl Acad Sci U S A* **106**(18): 7455-7460.

Thastrom A, Bingham LM, Widom J. 2004. Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. *Journal of molecular biology* **338**(4): 695-709.

Theodosiou A, Ashworth A. 2002. MAP kinase phosphatases. *Genome Biol* **3**(7): REVIEWS3009.

Theodosiou A, Smith A, Gillieron C, Arkinstall S, Ashworth A. 1999. MKP5, a new member of the MAP kinase phosphatase family, which selectively dephosphorylates stress-activated kinases. *Oncogene* **18**(50): 6981-6988.

Thorvaldsdottir H, Robinson JT, Mesirov JP. 2012. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*.

Tillo D, Hughes TR. 2009. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics* **10**: 442.

Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H et al. 2001. Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**(5550): 2364-2368.

Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M et al. 2004. Global mapping of the yeast genetic interaction network. *Science* **303**(5659): 808-813.

Torella JP, Chait R, Kishony R. 2010. Optimal drug synergy in antimicrobial treatments. *PLoS computational biology* **6**(6): e1000796.

Torres EM, Dephoure N, Panneerselvam A, Tucker CM, Whittaker CA, Gygi SP, Dunham MJ, Amon A. 2010. Identification of aneuploidy-tolerating mutations. *Cell* **143**(1): 71-83.

Torres EM, Sokolsky T, Tucker CM, Chan LY, Boselli M, Dunham MJ, Amon A. 2007. Effects of aneuploidy on cellular physiology and cell division in haploid yeast. *Science* **317**(5840): 916-924.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**(9): 1105-1111.

Trepel J, Mollapour M, Giaccone G, Neckers L. 2010. Targeting the dynamic HSP90 complex in cancer. *Nat Rev Cancer* **10**(8): 537-549.

Tsui K, Durbic T, Gebbia M, Nislow C. 2012. Genomic approaches for determining nucleosome occupancy in yeast. *Methods in molecular biology* **833**: 389-411.

Tugendreich S, Perkins E, Couto J, Barthmaier P, Sun D, Tang S, Tulac S, Nguyen A, Yeh E, Mays A et al. 2001. A streamlined process to phenotypically profile heterologous cDNAs in parallel using yeast cell-based assays. *Genome Res* **11**(11): 1899-1912.

Turcatti G, Romieu A, Fedurco M, Tairi AP. 2008. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. *Nucleic acids research* **36**(4): e25.

Uppuluri P, Nett J, Heitman J, Andes D. 2008. Synergistic effect of calcineurin inhibitors and fluconazole against Candida albicans biofilms. *Antimicrob Agents Chemother* **52**(3): 1127-1132.

Valouev A, Johnson SM, Boyd SD, Smith CL, Fire AZ, Sidow A. 2011. Determinants of nucleosome organization in primary human cells. *Nature* **474**(7352): 516-520.

van het Hoog M, Rast TJ, Martchenko M, Grindle S, Dignard D, Hogues H, Cuomo C, Berriman M, Scherer S, Magee BB et al. 2007. Assembly of the Candida albicans genome into sixteen supercontigs aligned on the eight chromosomes. *Genome biology* **8**(4): R52.

Venter JC Adams MD Myers EW Li PW Mural RJ Sutton GG Smith HO Yandell M Evans CA Holt RA et al. 2001. The sequence of the human genome. *Science* **291**(5507): 1304-1351.

Vichai V, Kirtikara K. 2006. Sulforhodamine B colorimetric assay for cytotoxicity screening. *Nat Protoc* **1**(3): 1112-1116.

Villafranca JE, Howell EE, Voet DH, Strobel MS, Ogden RC, Abelson JN, Kraut J. 1983. Directed mutagenesis of dihydrofolate reductase. *Science* **222**(4625): 782-788.

Wagner BK, Clemons PA. 2009. Connecting synthetic chemistry decisions to cell and genome biology using small-molecule phenotypic profiling. *Curr Opin Chem Biol* **13**(5-6): 539-548.

Wahl GM, Stern M, Stark GR. 1979. Efficient transfer of large DNA fragments from agarose gels to diazobenzyloxymethyl-paper and rapid hybridization by using dextran sulfate. *Proc Natl Acad Sci U S A* **76**(8): 3683-3687.

Wahl LM, Gerrish PJ. 2001. The probability that beneficial mutations are lost in populations with periodic bottlenecks. *Evolution* **55**(12): 2606-2610.

Wallace RB, Shaffer J, Murphy RF, Bonner J, Hirose T, Itakura K. 1979. Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic acids research* **6**(11): 3543-3557.

Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**(5366): 1077-1082.

Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J, Guo Y et al. 2008. The diploid genome sequence of an Asian individual. *Nature* **456**(7218): 60-65.

Warren RL, Sutton GG, Jones SJ, Holt RA. 2007. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23**(4): 500-501.

Watson JD, Crick FH. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**(4356): 737-738.

Wehner EP, Rao E, Brendel M. 1993. Molecular structure and genetic regulation of SFA, a gene responsible for resistance to formaldehyde in Saccharomyces cerevisiae, and characterization of its protein product. *Mol Gen Genet* **237**(3): 351-358.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT et al. 2008a. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**(7189): 872-876.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S et al. 2008b. Database resources of the National Center for Biotechnology Information. *Nucleic acids research* **36**(Database issue): D13-21.

White TC. 1997. Increased mRNA levels of ERG16, CDR, and MDR1 correlate with increases in azole resistance in Candida albicans isolates from a patient infected with human immunodeficiency virus. *Antimicrob Agents Chemother* **41**(7): 1482-1487.

White TC, Marr KA, Bowden RA. 1998. Clinical, cellular, and molecular factors that contribute to antifungal drug resistance. *Clin Microbiol Rev* **11**(2): 382-402.

Whitesell L, Lindquist SL. 2005. HSP90 and the chaperoning of cancer. *Nat Rev Cancer* **5**(10): 761-772.

WHO. 2011. World Malaria report 2011. .

-. 2012. Global tuberculosis report 2012.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J. 2008. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**(7199): 1239-1243.

Winzeler E, A., Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H et al. 1999. Functional Characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**: 901-906.

Wishart MJ, Dixon JE. 1998. Gathering STYX: phosphatase-like form predicts functions for unique protein-interaction domains. *Trends Biochem Sci* **23**(8): 301-306.

Wodicka L, Dong H, Mittmann M, Ho MH, Lockhart DJ. 1997. Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nature biotechnology* **15**(13): 1359-1367.

Wolber PK, Collins PJ, Lucas AB, De Witte A, Shannon KW. 2006. The Agilent in situ-synthesized microarray platform. *Methods in enzymology* **410**: 28-57.

Workman CT, Mak HC, McCuine S, Tagne JB, Agarwal M, Ozier O, Begley TJ, Samson LD, Ideker T. 2006. A systems approach to mapping DNA damage response pathways. *Science* **312**(5776): 1054-1059.

Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, Serpe JM, Dasu T, Tschannen MR, Veith RL et al. 2011. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med* **13**(3): 255-262.

Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, Sorek R. 2010. A single-base resolution map of an archaeal transcriptome. *Genome research* **20**(1): 133-141.

Xie C, Tammi MT. 2009. CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics* **10**: 80.

Xu D, Sillaots S, Davison J, Hu W, Jiang B, Kauffman S, Martel N, Ocampo P, Oh C, Trosok S et al. 2009a. Chemical genetic profiling and characterization of small-molecule compounds that affect the biosynthesis of unsaturated fatty acids in Candida albicans. *J Biol Chem* **284**(29): 19754-19764.

Xu Q, Schlabach MR, Hannon GJ, Elledge SJ. 2009b. Design of 240,000 orthogonal 25mer DNA barcode probes. *Proc Natl Acad Sci U S A* **106**(7): 2289-2294.

Yan Z, Costanzo M, Heisler LE, Paw J, Kaper F, Andrews BJ, Boone C, Giaever G, Nislow C. 2008. Yeast Barcoders: a chemogenomic application of a universal donor-strain collection carrying bar-code identifiers. *Nat Methods* **5**(8): 719-725.

Yang X, Boehm JS, Salehi-Ashtiani K, Hao T, Shen Y, Lubonja R, Thomas SR, Alkan O, Bhimdi T, Green TM et al. 2011. A public genome-scale lentiviral expression library of human ORFs. *Nature methods* **8**(8): 659-661.

Yeh PJ, Hegreness MJ, Aiden AP, Kishony R. 2009. Drug interactions and the evolution of antibiotic resistance. *Nat Rev Microbiol* **7**(6): 460-466.

Young KH. 1998. Yeast two-hybrid: so many interactions, (in) so little time. *Biol Reprod* **58**(2): 302-311.

Yu L, Lopez A, Anaflous A, El Bali B, Hamal A, Ericson E, Heisler LE, McQuibban A, Giaever G, Nislow C et al. 2008. Chemical-genetic profiling of imidazo[1,2-a]pyridines and -

pyrimidines reveals target pathways conserved between yeast and human cells. *PLoS Genet* **4**(11): e1000284.

Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research* **18**(5): 821-829.

Zhang N, Osborn M, Gitsham P, Yen K, Miller JR, Oliver SG. 2003. Using yeast to place human genes in functional categories. *Gene* **303**: 121-129.

Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, Liu XS, Struhl K. 2009. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. *Nat Struct Mol Biol* **16**(8): 847-852.

Zhang Z, Wippo CJ, Wal M, Ward E, Korber P, Pugh BF. 2011. A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science* **332**(6032): 977-980.

Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, Bertone P, Lan N, Jansen R, Bidlingmaier S, Houfek T et al. 2001. Global analysis of protein activities using proteome chips. *Science* **293**: 2101-2105.

Zhulidov PA, Bogdanova EA, Shcheglov AS, Vagner LL, Khaspekov GL, Kozhemyako VB, Matz MV, Meleshkevitch E, Moroz LL, Lukyanov SA et al. 2004. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic acids research* **32**(3): e37.

zur Wiesch PA, Kouyos R, Engelstadter J, Regoes RR, Bonhoeffer S. 2011. Population biological principles of drug-resistance evolution in infectious diseases. *Lancet Infect Dis* **11**(3): 236-247.