Extracting Haplotypes from Diploid Organisms

Jianping Xu

Department of Biology, McMaster University, 1280 Main St. West, Hamilton, Ontario, L8S 4K1, Canada

Abstract

Each diploid organism has two alleles at every gene locus. In sexual organisms such as most plants, animals and fungi, the two alleles in an individual may be genetically very different from each other. DNA sequence data from individual alleles (called a haplotype) can provide powerful information to address a variety of biological questions and guide many practical applications. The advancement in molecular technology and computational tools in the last decade has made obtaining large-scale haplotypes feasible. This review summarizes the two basic approaches for obtaining haplotypes and discusses the associated techniques and methods. The first approach is to experimentally obtain diploid sequence information and then use computer algorithms to infer haplotypes. The second approach is to obtain haplotype sequences directly through experimentation. The advantages and disadvantages of each approach are discussed. I then discussed a specific example on how the direct approach was used to obtain haplotype information to address several fundamental biological questions of a pathogenic yeast. With increasing sophistication in both bioinformatics tools and high-throughput molecular techniques, haplotype analysis is becoming an integrated component in biomedical research.

Introduction

With increasing availability and accessibility of DNA-based molecular typing methods, the size and scale of genetic variation data sets have increased dramatically in the last several years. The increase has occurred for almost all groups of organisms, from viruses to prokaryotic bacteria and archaea, from simple microscopic eukaryotes to complex large plants and animals. At present, by far the biggest increase has come from human population studies. Over 10 million common single nucleotide polymorphisms (SNPs) have been identified in the human genome (the International HapMap Consortium 2005). Indeed, it was the Human Genome Project that drove the initial technological developments that have subsequently made obtaining such large databases possible for populations of both human and other organisms. Some technologies allow assaying over 500,000 SNPs in a single experiment for human populations.

In sexual diploids such as most plants and animals, including humans, each nuclear gene has two alleles with each allele coming from a different parent. As a result, the two alleles may have different gene sequences and different evolutionary histories. Such genetic differences have traditionally been analyzed through techniques such as protein electrophoresis and restriction enzyme digest profiling. Protein electrophoresis was commonly used from the 1960s up to the early 1990s. The discovery of extensive isozyme polymorphisms in populations of many groups of organisms in the 1960s has often been credited as setting the foundation for the current revolution in molecular evolution and population genetics. Isozyme electrophoresis assays amino acid changes that cause proteins to migrate differently in a supporting matrix such as polyacrylamide gels or starch gels. Because only a subset of nucleotide substitutions will lead to protein polymorphisms that can be detected on a gel, following the discovery of restriction enzymes and the invention of DNA sequencing techniques, direct analyses of DNA polymorphisms were soon developed and actively pursued. These and other associated molecular techniques such as the polymerase chain reaction (PCR) and Southern hybridization allowed the development of many DNA-based markers, including restriction fragment length polymorphisms (RFLP), random amplified polymorphic DNA (RAPD), PCR-RFLP, amplified fragment length polymorphisms (AFLP), microsatellites, PCR-fingerprinting, and SNPs. Among the complex patterns seen in RAPD, AFLP and PCR-fingerprinting, the nucleotide variation for individual loci are often difficult to infer from the observed data (Xu 2005). In contrast, for molecular markers based on RFLP, PCR-RFLP, microsatellite and SNP, the allelic assignment for individual polymorphic nucleotides can be usually assigned without ambiguity in an individual (Xu 2006).

Among the many categories of population genetic data in diploids, that containing long stretches of DNA sequences (in the range of several hundred bases) has been gaining importance in the last several years, due largely to increasing accessibility and affordability of DNA sequencing to typical researchers. Therefore, this review will focus on DNA sequence-based information. For each locus in a diploid, direct sequencing typically generates a mixture of two allelic sequences, with variable nucleotide sites shown in the form of SNPs. However, the haplotype phases of these SNPs are typically unknown. Having the haplotype information can be highly valuable for many areas of theoretical and applied research, including the inferences of parentage and mating system, the analyses of clonality and recombination, gene flow and migration, cryptic speciation, and hybridization (e.g. Akey *et al*. 2001; the International HapMap Consortium 2005; Marchini *et al*. 2006; Xu 2005). In human population studies, haplotypes have been used to infer selection, parentage, forensic identification, recombination, and both historical and recent migration events (e.g. Sabeti *et al* 2002; Fearnhead and Donnelly, 2001; De Iorio and Griffiths, 2004). In the last several years, one application that has attracted the most attention is to use haplotypes

For correspondence: jpxu@mcmaster.ca

to locate markers and mutations associated with both monogenic and multigenic disorders in humans. Indeed, SNPs and haplotypes associated with many monogenic and multigenic disordered have already been identified (Lazzeroni 2001, the International HapMap consortium 2005; Chapman *et al*. 2003).

In this review, I group current haplotyping methods into two broad categories: the indirect approach and the direct approach. The indirect approach has two components: (i) experimentally obtaining diploid sequence information directly from individuals in a population; and (ii) computationally inferring haplotype sequences based on the distribution of polymorphic nucleotides in diploid sequences in the population. In contrast, the direct approach is completely experimentally based and involves the cloning of individual haplotypes, followed by sequencing of these cloned haplotypes. The advantages and disadvantages of the two approaches will be compared and discussed. This is then followed by a specific example on how the direct approach was used to obtain haplotype information to address several fundamental biological questions of a human pathogenic yeast *Cryptococcus neoformans*. I will conclude with a brief summary and future perspectives.

The indirect approach

As was briefly mentioned above, the indirect approach for obtaining haplotype sequence information from diploid organisms includes two components. The first is to obtain DNA sequence information from the target loci. At present, this process is most commonly achieved by first using gene-specific PCR primers to amplify both alleles in diploid individuals. The PCR products are then sequenced directly using automated DNA sequencers, yielding a composite chromatogram for each locus in each individual organism. The composite DNA sequence data for each locus from many individuals are then imported into computer program(s) to estimate the likely haplotypes in each individual and/or the likely haplotype frequencies in a population. Because PCR and DNA sequencing methods are straightforward, familiar to most molecular biology laboratories, I will skip the discussions of these two techniques but focus on the computer programs that have been developed to infer haplotypes based on diploid sequence data.

Most of the computer programs currently in use for haplotype inferences have been developed for analyzing human SNP data sets and, in principle, these programs should be applicable for haplotyping in other organisms. However, there may be subtle differences in data acquisition between human samples and samples from other species. This is mainly because modern humans are of relatively recent origin and the density of SNPs is relatively low, about one common SNP for every 2–5 kb in a typical individual (e.g. Lander *et al*. 2001; Venter *et al*. 2001; the International HapMap Consortium, 2005). With such a low density of polymorphic nucleotide sites, a typical stretch of human DNA sequence (from several hundred bp to about 1kb) that current DNA sequencers can directly obtain in one sequencing reaction may contain none or very few nucleotide substitutions. In these DNA fragments, the SNP itself in each sequenced fragment can be directly interpreted for its allelic status and there might be no need to use any computer programs for haplotype inferences for such short DNA fragments from humans. Instead, in human populations, the challenge is to infer the associations (linkage disequilibrium) among SNPs located far from each other (tens to hundreds of kb sequences) on the same chromosome.

For computational inferences of haplotypes, the fundamental principles for inferring human haplotypes using long tracts of DNA sequence data are the same as those using short stretches of DNA in more ancient and polymorphic species. For species with higher densities of SNPs in their genomes than the human genome, multiple polymorphic sites may be found in short stretches of DNA.

The wealth of human SNP datasets and the great potential these datasets can offer in biomedical research and clinical applications have generated tremendous amounts of interests among computational biologists and statistical geneticists. Indeed, over 40 computer programs have been developed over the last decade and hundreds of papers have been published to describe, test, and compare these programs (see two recent reviews by Salem *et al*. 2005 and Marchini *et al*. 2006). These programs can be broadly classified into following three groups based on their fundamental inference principles: (i) parsimony methods, (ii) maximum-likelihood methods, and (iii) Bayesian methods. Below is a brief description of these methods.

Parsimony methods

Using parsimony method to construct haplotypes was first proposed by Clark in 1990. His method relies extensively the data from highly homozygous individuals to first obtain a set of known haplotypes. The frequencies of these known haplotypes are then used to draw inferences about the most likely haplotypes for individuals whose haplotypes are ambiguous. Because this procedure depends on the existence of homozygous individuals, in a population of unrelated individuals with few homozygous loci in each individual, the haplotypes of many individuals may not be assigned. This problem has since been partially solved through integrating the assumption of perfect phylogeny. The idea of a perfect phylogeny was based on the fact that recombination among closely linked SNPs may be uncommon and that there is significant linkage disequilibrium among such SNPs. With no recombination, a phylogeny of haplotypes should have a consistency index of 1, without any homoplasmy (Xu 2006). The haplotyping algorithm was then reduced to the search for haplotypes that yield a perfect phylogeny. Several computer programs such as HAPAR (Wang and Xu 2003), HAPINFERX (Clark 1990), and GPPH (Chung and Gusfield 2003) are based on the parsimony approach.

Maximum-likelihood methods

Maximum-likelihood methods make up the majority of haplotyping computational programs developed so far (Salem *et al*. 2005). These methods are based on the principle that the set of haplotypes giving the highest likelihood score for the observed diploid genotype data is the most likely set of haplotypes. The likelihood of a set of

haplotypes is the probability of observing the data (diploid genotypes) with a given set of haplotypes under specific assumptions about the population differentiation (e.g. differentiated vs. non-differentiated sub-populations), mating structure (e.g. random mating vs. non-random mating), and recombination rate among the nucleotide sites. Estimates for the sets of haplotypes and their frequencies are iteratively updated to maximize the likelihood function for the observed diploid genotypes. The most commonly used algorithm for maximum-likelihood estimation is the two-step process known as Expectation-Maximization (EM). Because the EM algorithm may converge to a non-global maximum and a large sample size is required to allow accurate estimates, a variety of implementations have been developed to increase the accuracy of haplotype estimation. At present, over 30 haplotype inference programs have been developed based on the maximum-likelihood approach. These programs differ in their assumptions (e.g. loci in Hardy-Weinberg equilibrium), the types of data the program can handle (bi-allelic, multi-allelic), the size of data (the number of individuals and the number of loci), and computational platforms etc (PC, MAC, UNIX, WEB-based etc) (Salem *et al*. 2005). Programs in this category include SNPHAP (Clayton 2001), THESIAS (Tregouet *et al*. 2004), HPLUS (Zhao *et al*. 2003), and Arlequin (Schneidler *et al*. 2002).

Bayesian methods

Similar to maximum-likelihood methods, Bayesian methods are also based on the likelihood theory. However, Bayesian methods use different prior assumptions to model haplotype frequencies. Bayesian methods can be further divided into two subgroups: the simple Bayesian methods (e.g. Lin *et al*. 2002) and the coalescent-based Bayesian methods (Stephens *et al*. 2001). The difference between these two groups of methods is that the simple methods make no assumption about the history of the analyzed populations while the coalescent-based methods take into account the overall similarities among putative haplotypes. The simple Bayesian programs include HAPLOTYPER (Niu *et al*. 2002) that has a statistical procedure similar to the EM mentioned above, and HAPLOREC (Eronen *et al*. 2004) that has a Variable Length Markov Chain function incorporated into the Bayesian algorithm. The coalescent-based Bayesian programs include the widely used PHASE program (Stephens and Donnelly 2003). The PHASE algorithm incorporates the idea that, over short genomic regions, because recombination is likely to be uncommon, sampled chromosomes tend to cluster together into groups of similar haplotypes due to shared common ancestries. One significant improvement in the recent version PHASE v2.0 is the explicit incorporation of recombination that allows haplotypes to change and be updated as the analysis moves along a chromosome (Stephens and Donnelly 2003).

Despite the large number of programs, a critical and exhaustive comparison among all methods using the same natural diploid genotype data with known haplotypes has not been conducted. Based on limited computer simulation studies and the analyses of some real data with known haplotype information, most compared programs developed so far were found to provide very similar estimates of population haplotype frequencies and/or haplotype sets for individuals. However, as was noted in the recent comparisons of haplotyping software, most current programs and methods have limitations and many were designed with specific tasks in mind by the software developer (Salem 2005; Marchini *et al*. 2006). For example, about a third of the existing programs takes only bi-allelic data and cannot analyze multi-allelic data, some cannot handle missing data, and still some can only estimate haplotype frequencies but do not assign haplotypes to individuals. Most methods are highly sensitive to genotyping errors and missing data – common problems with direct DNA sequence typing in diploids.

The direct approach

In the direct approach, individual haplotypes are first separated, followed by the identification and determination of individual DNA sequences. The technical issues for extracting haplotypes from diploid organisms may differ depending on the size of the genomic region from where haplotypes are to be inferred. Below is a brief review of the technical issues involved in haplotype extractions. For convenience, these issues are discussed in separate sections based on the size of DNA fragments from where SNPs are inferred.

Haplotyping within fragments < 2kb in length

To identify haplotypes among variable nucleotides within a short stretch of DNA fragment (< 2kb in length), the entire nucleotide sequence of the whole fragment may be directly obtained by using both the forward and reverse PCR primers for sequencing. Similar to the indirect approach, this step of direct sequencing from a diploid organism will result in a composite sequence profile, with a mixture of homozygous and heterozygous nucleotide sites along the entire sequence. To obtain the haplotype sequences from the diploid genotype, several methods can be used. A classical method is to construct a random cloning library from the PCR products of the target gene. Each clone in the library will contain the sequence of only one allele (haplotype). Sequencing the DNA fragment from any random clone in this library will give one of the two original haplotypes from each analyzed individual. From here on, one of two methods can be further used to obtain the other haplotype. In the first, the second haplotypes may be directly deduced by subtracting the first haplotype sequence information from the composite diploid sequences. In the second method, sequencing additional fragments contained in other random clones of the library may reveal the sequence identity of the second haplotype. To make sure there is a >95% probability that both haplotypes will be present among a collection of random clones in the library and assuming that the two haplotypes are equally represented in the library, 7 random clones (based on a binomial distribution function) will need to be sequenced for each locus for each individual. However, as will be shown below, the application of additional screening of the random clones from the library could rapidly reduce the number of clones need to be sequenced, thus reducing the cost of sequencing.

Other techniques have been recently developed to allow more efficient determination of haplotypes. For example, a high-throughput, bead-based capture-based haplotyping (CBH) assay was recently reported (Hurley *et al*. 2005). In this method, data collection was performed via flow cytometry and the assay yields plus/minus results for individual nucleotides, allowing for automated base calling by a simple computer application. The CBH assay required minimal setup, no centrifugation and can be performed in <1 h. It was shown to be effective in molecular haplotyping of 11 SNPs within the exon 2 (about 1.1 kb) of the *N*-acetyltransferase-2 gene (Hurley *et al*. 2005). In another development, Ding and Cantor (2003) described a technique called M1-PCR (M for "multiplex" and 1 for "single-copy DNA molecules") that enabled direct molecular haplotyping of several polymorphic markers separated by as many as 24 kb (See also below). To achieve this, genomic DNA samples were first diluted to approximately single-copy (i.e. one haploid genome content per sample). The haplotypes were then directly determined by simultaneously genotyping several polymorphic markers in the same reaction with a multiplex PCR and base extension reaction (Ding and Cantor 2003).

Haplotyping for genomic regions between 2kb and 1.5mb long

Most haplotype phasing in humans is concerned with genomic fragments in this size category. Several approaches are available to convert diploid genomes and genotypes into haploid genomes or haplotypes. One frequently used is to apply somatic cell hybrid technology and convert a diploid cell into haploid cell lines. Another classical approach involves cloning to first separate individual haplotypes, followed by sequencing or other SNP determination techniques. Current cloning vectors can accommodate DNA fragments between tens of nucleotides to over 1.5 million base pairs (Ausubel *et al*. 2000). Therefore, direct haplotyping for SNPs located within this size range is possible.

To directly obtain haplotype information among nucleotide sites located between 2kb and 1.5mb, random genomic libraries containing individual haplotypes can be first constructed. Each clone in the random library will contain only one of the two haplotypes of the diploid individual. Depending on the distance between the most distant SNPs for the analysis, different cloning vector can be used. Specifically, plasmid vectors can carry inserts up to about 10kb but the typical range is 0.5–2kb. Lambda phage-based cloning vectors have slightly larger capacity with a typical insert size between 7–10kb but can reach up to 20kb. Cosmid vectors have a capacity for inserts between 35–45kb. The bacterial artificial chromosome (BAC) vectors can clone fragments about 80–200kb and the yeast artificial chromosome (YAC) vector can clone fragments between 200kb and 1.5 million base pairs (Ausubel *et al*. 2000).

Once the random library is constructed (there are commercially available services for library construction), the desired clone(s) can be identified using a variety of common laboratory techniques such as PCR screening and Southern hybridization using specific probes. PCR primers targeting the specific genes in the clone can be then used to amplify the DNA fragments and all the nucleotide sequences for the target sites within the clone can be obtained. The composition of these nucleotides would constitute one haplotype. The second haplotype can be deduced by comparing the composite diploid sequence with the known haplotype sequence already identified. Alternatively, the second haplotype can be directly obtained through a similar process described above by first identifying the clone containing the second haplotype and then sequencing the target genes within the clone.

Similar to those described for fragments smaller than ~2kb, novel techniques have also been developed for this fragment size category. For example, for fragments smaller than 20 kb, the M1-PCR protocol described in the previous section can be directly applied to eliminate the cloning step (Ding and Cantor 2003). In addition, if haplotype-specific signature sequences are known, PCR primers targeting such haplotypes can be developed and long range PCR can be used to directly pull out the desired haplotype for further sequencing (Pont-Kingdon *et al*. 2004). Other techniques have also been developed to enhance the efficiency of haplotyping for fragments in this size category. One such technique is based on microfluid dynamics and is called the direct linear analysis (DLA, Chan *et al*. 2004). In this method, high molecular weight individual DNA molecules are first labeled with sequence-specific fluorescent tags. These DNA molecules are then passed through a microfluidic device to stretch and linearize DNA molecules in elongational flow. The microfluidic device is coupled to a multicolor detection system capable of single-fluorophore sensitivity. In their test of this device, double-stranded DNA molecules were tagged at sequence-specific motif sites with fluorescent bisPNA (Peptide Nucleic Acid) tags (Chan *et al*. 2004). The DNA molecules were then stretched in the microfluidic device and driven in a flow stream past confocal fluorescence detectors. Their analysis showed that DLA could provide the spatial locations of multiple specific sequence motifs along individual DNA molecules, and thousands of individual molecules could be analyzed per minute. They validated this technology using the 48.5 kb lambda phage genome with different 8-base and 7-base sequence motif tags. The distance between the sequence motifs was determined with an accuracy of ±0.8 kb, and these tags could be localized on the DNA with an accuracy of ±2kb.

Other techniques include (i) haplotyping of kilobase-sized DNA using carbon nanotube probes and multiplex detection of labeled probes by atomic force microscopy (Woolley *et al*. 2000); (ii) Direct haplotyping by the polony technique that relies on embedding intact chromosomal DNA in polyacrylamide gels on a glass slide and followed by PCR and sequencing (Mitra *et al*. 2003); and (iii) an optical mapping technology that can physically map very long strands of DNA by imaging individually stretched DNA molecules attached to a surface and digested with a restriction enzyme (Aston *et al*. 1999).

Haplotyping for genomic regions greater than 1.5mb long

Direct haplotyping for genomic regions greater than 1.5mb is still possible through cloning by first identifying

clones with perfectly overlapping sequences in the library, in a process called chromosome walking. Chromosome walking is a common technique for cloning specific genes in large genomes, including candidate disease genes in humans. It was one of the key steps for the public Human Genome Project (Lander *et al*. 2001). Chromosome walking can be used to construct haplotypes as long as an entire chromosome. Once the overlapping clones for a haplotype chromosome are identified, the polymorphic nucleotides along the chromosome can be easily obtained through direct PCR and DNA sequencing of the target sites, similar to those described in previous sections.

Haplotyping a group of related individuals

Both the direct and indirect approaches discussed above can be used to obtain haplotypes from both unrelated individuals and related individuals. To obtain haplotypes from individuals with known relationships, the inferences may be more straightforward and techniques less demanding. With known relationships such as those among individuals from genetic crosses or pedigrees, recombinant haplotype blocks may be directly obtained through high-density genotyping of parents and offspring. In humans, such pedigrees may include grandparents, parents, offspring, siblings and cousins etc. The simplest family component is a trio that consists of two parents and one offspring. In this situation, given enough polymorphisms within and between the two parents, the haplotypes of the progeny as well as both parents may be obtained directly from the genotype data. The current International HapMap Project utilizes the inherent advantages of information from trios to increase accuracy and speed up haplotype inferences (the International HapMap Consortium 2005; Marchini *et al*. 2006).

Comparisons between the indirect and direct approaches

As is obvious from the above descriptions, both the direct and indirect approaches have advantages and disadvantages. While the direct methods can give 100% accuracy as to haplotype assignment for each individual in a population, they are slower, more labor-intensive and costlier than the indirect methods. In contrast, while fast and low cost, indirect methods are more susceptible to genotyping errors and missing data can have a significant effect on haplotype assignments (Salem *et al*. 2005; Marchini *et al*. 2006). In addition, most algorithms have a variety of assumptions (e.g. large random mating population and lack of recombination) that can be easily violated in specific populations. In fact, the assumptions such as Hardy-Weinberg equilibrium and linkage equilibrium are often themselves the objectives of population genetic research. Therefore, inferences based on such assumption can potentially confound the conclusions. The most significant disadvantage of the indirect methods is that estimated haplotypes are never 100% certain. Computer simulation studies identified that incorrect haplotype inferences typically occur for over 5% of the cases in a population of unrelated individuals (Marchini *et al*. 2006). This error rate can have a significant effect on further inferences between haplotypes and complex phenotypic traits where individual genetic contributions to such trait values can be very small.

A direct approach to obtain and analyze haplotype data in a human pathogenic fungus *Cryptococcus neoformans*

Much research has been done and many reviews have been written about human haplotyping, including the identification of human SNPs (e.g. International HapMap Consortium, 2005), inferring the haplotypes using bioinformatic methods (e.g. Salem *et al*. 2005), and analyzing the relationships between genetic information (SNPs or haplotypes) and phenotypic traits (Lazzeroni 2001; Akey *et al*. 2001). Readers interested in human haplotyping research are encouraged to look into these and other recent papers on the subject. In the following sections, I will describe a non-human example to illustrate how haplotyping can be used to address several fundamental evolutionary questions of a fungal pathogen *Cryptococcus neoformans*.

C. neoformans is a major pathogen of humans and other mammals throughout the world. Using commercial antibodies, strains of *C. neoformans* can be classified into five serotypes A, B, C, D and AD. Strains of serotypes A, B, C and D are haploid (1N) but those of serotype AD are diploid (2N) or aneuploid (between 1N and 2N). Recent studies showed that serotypes A, B, C, and D exhibited significant divergence at the molecular level (Xu *et al*. 2000). However, the origins and evolution of serotype AD remained unknown, until very recently. To investigate the origins and evolution of serotype AD strains, fourteen strains of serotype AD were analyzed (Table 1). These strains were from three geographic areas in the US obtained during a population-based active surveillance conducted by the Centers for Disease Control and Prevention (CDC). Twelve strains were from San Francisco, California, and one strain each from Georgia and Texas. Among the 14 strains, strains MAS94–0241 and MAS94–0244 were isolated from different body sites of the same patient while strains MAS93–0315 and MAS93–0610 were isolated at different times from the same body site (cerebrospinal fluid) of one patient. Each of the other 10 strains was from a different patient (Table 1).

Sequencing PCR products from serotype AD strains

To understand the origins of serotype AD strains, a phylogenetic approach was employed to analyze DNA sequence information. Since a large DNA sequence database was already established for two highly polymorphic genes diphenol oxidase (i.e. Laccase or LAC) and the orotidine monophosphate pyrophosphorylase (i.e. URA5) (Xu *et al*. 2000), fragments of these two genes were chosen for investigating strains of serotype AD.

However, direct sequencing of PCR products from the genomic yielded un-interpretable DNA sequences for both LAC and URA5 in all 14 strains of serotype AD (Xu *et al*. 2002; 2003). In contrast, the same protocol generated clean DNA sequence for strains of serotypes A, B, C, and D (Xu *et al*. 2000). The extensive sequence ambiguity suggested that each of the two loci must contain distinctly

Table 1. Strains of serotype AD in *Cryptococcus neoformans* and their summary haplotype information

Strain	Geographic origin	Source[1]	Allele	LAC haplotype	URA5 haplotype
MAS92-0022	San Francisco	CSF	1	1	1
			2	6	11
MAS92-0086	San Francisco	Blood	1	2	2
			2	7	12
MAS92-0153	San Francisco	Blood	1	1	3
			2	8	13
MAS92-0189	Georgia	CSF	1	1	4
			2	7	12
MAS92-0224	San Francisco	CSF	1	1	2
			2	7	12
MAS92-0668	San Francisco	CSF	1	3	5
			2	7	12
MAS92-0793	San Francisco	CSF	1	1	6
			2	6	11
MAS92-0855	San Francisco	CSF	1	4	7
			2	6	11
MAS93-0315	San Francisco	CSF	1	1	8
			2	10	11
MAS93-0610	San Francisco	CSF	1	1	8
			2	7	11
MAS94-0018	San Francisco	PF	1	1	9
			2	9	12
MAS94-0241	San Francisco	BW	1	5	2
			2	7	12
MAS94-0244	San Francisco	Prostate	1	1	2
			2	7	12
MAS94-0351	Texas	CSF	1	4	10
			2	6	14

1, CSF: cerebral spinal fluid; BW: bronchial wash; PF: peritoneal fluid.

different alleles within the 14 strains. Therefore, a direct approach was taken to obtain haplotype (or allelic) sequences from each locus for each of the 14 strains.

Direct approach to obtain haplotype sequences in C. neoformans

To obtain individual haplotype sequences, the PCR products from each of the two loci for each of the 14 strains (28 PCR products total) were cloned using a pGEM-T cloning kit (Promega) and transformed into competent *Escherichia coli* cells. From each of the 28 cloning experiments, ten random *E. coli* transformed colonies were picked. These colonies were then used to screen for distinct alleles, as follows.

Based on serotype and other molecular and biochemical information, it was hypothesized that strains of serotype AD might be hybrids between strains of serotypes A and D (Xu *et al*. 2002). If so, strains of serotype AD should contain haplotypes from both serotypes A and D. To test this hypothesis, known sequences of strains of serotypes A and D are used to design efficient screening methods to identify potential serotype AD haplotypes similar to those of serotypes A and D. Because sequences for both LAC and URA5 were available from the GenBank for a large number of strains from all four serotypes A, B, C, and D (GenBank accession numbers AF140150-AF140183 for LAC and AF140185-AF140217 for URA5), polymorphic sites from these existing sequences were used as baselines for screening distinct clones in the library.

First, 3 representative sequences for each gene were retrieved from strains of both serotypes A and D from the GenBank and restriction maps were generated for each sequence using the program Web-Cutter (http://www.firstmarket.com/cutter/cut2.html). The restriction maps were then compared between strains of serotypes A and D for each gene and unique restriction sites were identified that could distinguish alleles from the two serotypes. Because of the extensive divergence between serotypes A and D, a large number of polymorphic restriction sites were identified for each of the two genes. To increase the efficiency of allelic identification, enzymes that generate easily distinguishable allelic patterns were chosen (in our case, the restriction enzymes *DpnII* and *DdeI* for LAC and URA5 genes respectively) (Xu *et al*. 2002; 2003).

For each of the 280 random clones (ten clones for each of the 28 cloning experiments), the primer pair T7/SP6 flanking the cloning site on the vector was used to amplify the cloned inserts in a PCR reaction directly from fresh bacterial cells. The PCR products were then digested using enzymes *DpnII* (for LAC) or *DdeI* (for URA5). Upon agarose gel electrophoresis, two restriction-digest patterns were identified for each gene in each of the 14 serotype AD strains. Each pattern represented a distinct haplotype. A representative gel is presented in Fig. 1.

Based on restriction site polymorphism data, a total of 56 *E. coli* clones were selected for sequencing from the 280 screened with one representative for each restriction digestion pattern for each gene in each strain (14 strains x 2 genes per strain x two restriction digest patterns per gene = 56 unique clones). Plasmid DNA was isolated and purified from each of the 56 *E. coli* clones. Each of the 56 inserts was then amplified using the T7/SP6 primer pair and PCR products were cleaned and sequenced. Unambiguous DNA sequences were obtained for all PCR products. To ensure reproducibility of the results, additional clones were sequenced from several strains for each of the two genes to identify potential genotyping errors.

Haplotype analyses

For each gene, all sequences were aligned using CLUSTALX and unique haplotypes were identified and presented in Table 1. For each of the two genes, two distinct alleles were identified within each of 13 strains. One strain (MAS94-0241) showed three alleles at the URA5 locus: haplotypes 2, 12 (Table 1), and a third that differed by a single nucleotide from haplotype 12. To investigate further, several additional colonies established from the URA5 gene PCR product were sequenced from this strain. These additional clones had the same restriction fragment pattern as haplotype 12. Sequencing results confirmed that these additional clones contained sequences identical to haplotype 12. Therefore, the most likely hypothesis was that the single nucleotide difference was the result of an error introduced during PCR or plasmid replication within bacterial cells.

This hypothesis was supported by the observation that another strain MAS94-0244 isolated from the same patient as strain MAS94-0241 had the same haplotype (haplotype 12) at URA5. In another pair of isolates MAS93-0315 and MAS93-0610 from a different patient, their allele 2s at LAC gene also differed by a single nucleotide (Table 1, Fig. 2). However, this single nucleotide difference was confirmed by additional sequencing (Xu *et al*. 2002). Point mutations could be introduced every time DNA replicates, both *in vivo* and *in vitro* during PCR. Therefore, it's essential that suspicious variations be confirmed by empirical investigations.

Evolutionary inferences on the origins of serotype AD strains and evidence for clonality and recombination of serotypes A and D populations

The obtained haplotype sequences were analyzed using phylogenetic methods implemented in the computer program PAUP (6). To test for the specific hypothesis about the origins of serotype AD strains and the reproductive modes of serotypes A and D populations, representative sequences from strains of serotypes A and D obtained earlier were also included for comparison and analyses (Xu *et al*. 2000). One representative maximum parsimonious tree for each of the two genes was presented in Fig. 2. Our analyses suggested the following conclusions.

First, each strain had two evolutionary distinct haplotypes at both LAC and URA5 loci, with one haplotype highly similar or identical to sequences from serotype A strains and the other highly similar or identical to sequences from serotype D strains. Because of the unambiguous clustering and the strong statistical supports, haplotypes at both loci from serotype AD strains were assigned into serotypes A and D groups. The result suggested recent hybrid origins of serotype AD strains from between strains of serotypes A and D (Xu *et al*. 2002; 2003).

Second, based on the patterns of (i) sequence diversity among serotype AD strains; (ii) similarity to existing serotypes A and D sequences; and (iii) bootstrap values, the gene genealogical analyses suggested that multiple hybridization events were responsible for generating the 14 serotype AD strains. The LAC gene genealogy identified at least three hybridization events while the URA5 gene genealogy identified at least four hybridization events. When these two genes were considered together, at least five hybridization events could be inferred from these data (Xu *et al*. 2002; 2003).

Third, the comparisons between the LAC gene genealogy and the URA5 gene genealogy revealed statistically significant genealogical incongruence

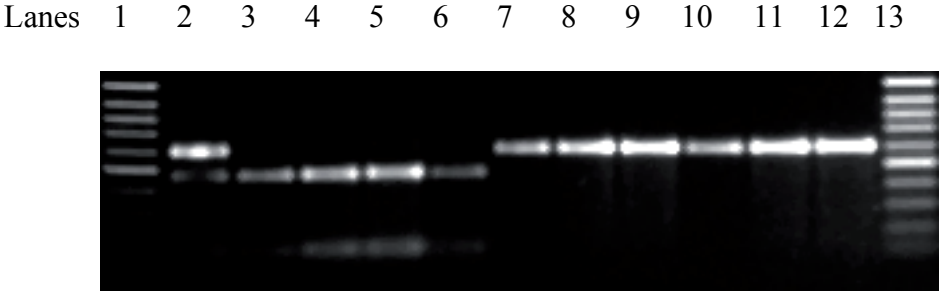


Fig. 1. A representative gel showing two distinct alleles at LAC locus from a serotype AD strain MAS92-0022. Lanes 1 and 13 contained the 100bp ladder size standard. Lane 2 contained digested PCR product from the original strain MAS92-0022 that showed heterozygosity. Lanes 3–12 contained digested PCR products from ten random transformed bacterial colonies. Two alleles are apparent: lanes 3–6 showing the typical serotype A allele (allele 2 in Table 1) and lanes 7–12 showing the typical serotype D allele (allele 1 in Table 1).

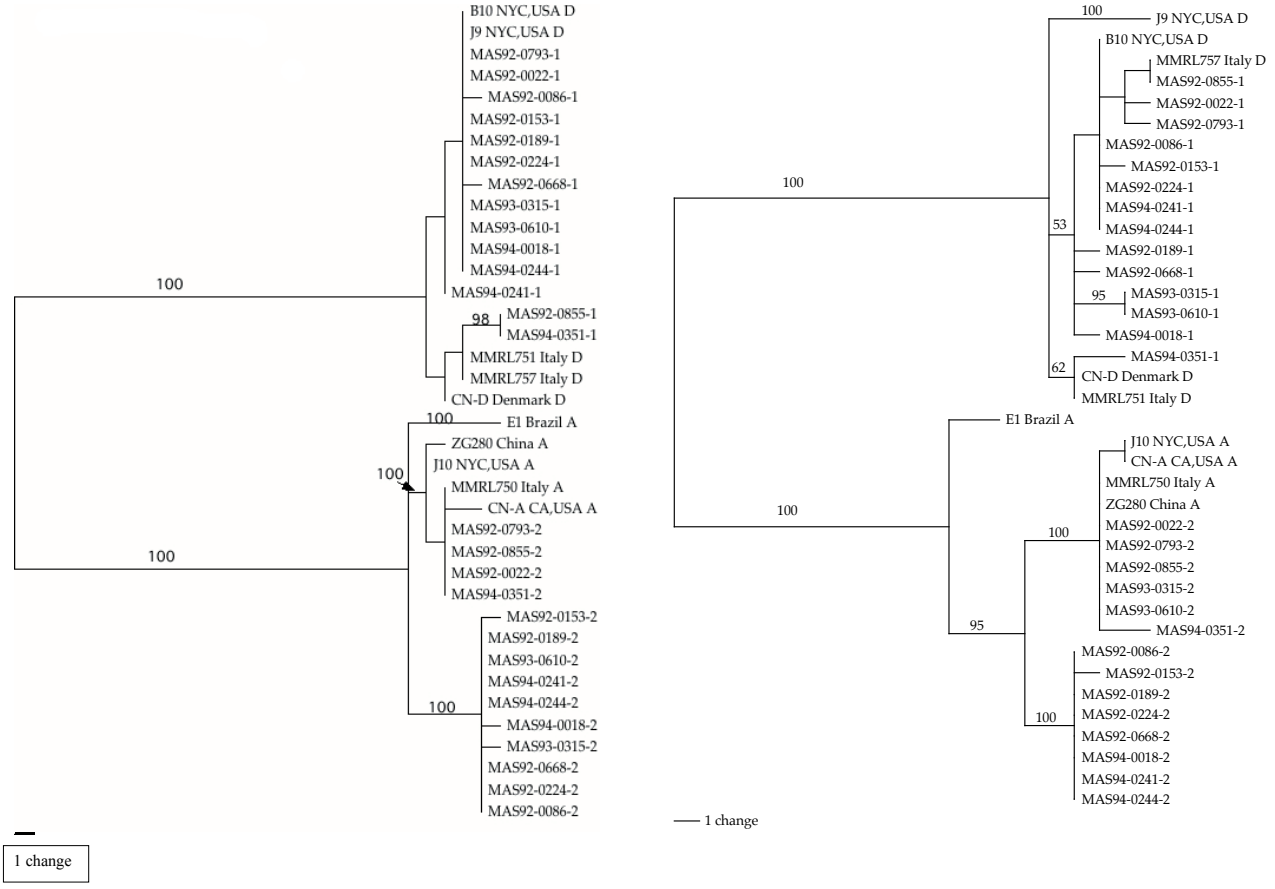


Fig. 2. Genealogies of the LAC (A) and URA5 (B) genes for 14 strains of serotype AD (from refs. 3–4). Branch-lengths are proportional to the amount of sequence divergence. Bootstrap values greater than 50% are presented above branches. Five representative strains of serotype A (E1, CN-A, MMRL750, J10 and ZG280) and five for serotype D (B10, CN-D, J9, MMRL751 and MMRL757) are included as references for comparisons. These 10 strain names were followed by their geographic origins (CA, California; NYC, New York City) and serotype identifications. The GenBank accession numbers for sequences from serotype AD strains are AY463634-AY463661 for LAC and AY463662-AY463689 for URA5.

within each of the two gene clusters (i.e. the serotype A gene cluster and the serotype D gene cluster). The genealogical incongruence was consistent with the hypothesis of recombination within both serotype A and serotype D populations. These conclusions were novel and not reached by analyzing only serotypes A and D populations. The detailed interpretations and discussions of hybridization and recombination from gene genealogical data can be found in Xu *et al*. (2002; 2003).

Due to significant sequence divergence (5–10% nucleotide differences) between strains of serotypes A and D and the existence of small insertions/deletions in intronic sequences of these two analyzed genes (Xu *et al*. 2000), the indirect approach could not obtain the haplotype sequences from either gene in the 14 serotype AD strains. Without the haplotype sequence information, none of the above conclusions could have been drawn about the evolution and population biology of this important human fungal pathogen.

Conclusions and perspectives

This paper provided an overview of the approaches for obtaining haplotype sequences from diploid organisms. I then presented a specific example on how the direct approach could be used to obtain haplotypes and how the haplotype sequence information can be used to address fundamental biological questions. This and other examples presented elsewhere (e.g. Johnson *et al*. 2001) highlights the utility of haplotype information in biomedical investigations. Both the direct and indirect approaches have advantages and disadvantages. While the indirect methods are cost-effective, there are many problems. The incorporation of new computational tools and more realistic assumptions will make these inference programs more useful and allow them to provide more accurate assessments of haplotypes. In contrast, the direct methods can provide accurate information on haplotype sets in diploid organisms and all the essential techniques are already developed. However, at present, direct methods for chromosome-wide and genome-wide haplotype inferences are still quite labor-intensive and costly.

The labor-intensive and costly protocols for obtaining large-scale haplotypes may be overcome in the near future with the continued development of high-throughput platforms that incorporate many of the techniques discussed above. Individual large-scale high throughput technologies are available for library construction (from plasmid vector-based to BAC based), selection for positive clones, tiling of clones or chromosome walking to obtain a minimum number of clones that cover the whole chromosome/genome, SNP screening (e.g.

denaturing gradient gel electrophoresis/temperature gradient gel electrophoresis or DGGE/TGGE; single strand conformational polymorphism analysis or SSCP; and heterduplex analysis), and DNA sequencing. For example, one process called the clone-based systematic haplotyping combined the high throughput platforms of cosmid library tiling, PCR and sequencing in a 96-well format that allowed rapid haplotyping of SNPs separated by about 50kb (Burgtorf *et al*. 2003). The integration of various existing platforms and the further development of additional techniques into these high throughput platforms should make direct haplotyping analysis increasingly affordable. It is likely that most future platforms will include both the direct and indirect approaches.

The discussions here have been focused on diploids, however, the approaches, specific techniques and analytical methods should be applicable for obtaining haplotypes from triploids, tetraploids, and organisms with higher ploidy levels. A significant proportion of plants have ploidy level greater than 2N. At present, haplotype analyses in these organisms are barely explored.

Acknowledgments

During the preparation of this manuscript, research in my lab is support by grants from Natural Sciences and Engineering Research Council (NSERC) of Canada, the Premier's Research Excellence Award, and Genome Canada.

References

Akey, J., Jin, L., and Xiong, M. (2001). Haplotypes vs single marker linkage disequilibrium tests: what do we gain? Eur. J. Hum. Genet. *9*:291–300

Aston, C., Hiort, C., and Schwartz, D.C. (1999). Optical mapping: An approach for fine mapping. Methods Enzymol. *303*: 55–73

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, K. (2000). Current Protocols in Molecular Biology. John Wiley and Sons, Inc.

Burgtorf, C., Kepper, P., Hoehe, M., Schmitt, C., Reinhardt, R., Lehrach, H., and Sauer, S. (2003). Clone-Based Systematic Haplotyping (CSH): A procedure for physical haplotyping of whole genomes. Genome Res. *13*: 2717–2724

Chan, E.Y., Goncalves, N.M., Haeusler, R.A., Hatch, A.J., Larson, J.W., Maletta, A.M., Yantz, G.R., Carstea, E.D., Fuchs, M., Wong, G.G., Gullans, S.R., and Gilmanshin, R. (2004). DNA mapping using microfluidic stretching and single-molecule detection of fluorescent site-specific tags. Genome Res. *14*: 1137–1146

Chapman, J.M., Cooper, J.D., Todd, J.A., and Clayton, D.G. (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum. Hered. *56*:18–31

Chung, R.H. and Gusfield, D. (2003), 'Empirical exploration of perfect phylogeny haplotyping and haplotypers', in: Warnow, T. and Zhu, B. (eds), Lecture Notes in Computer Science, Springer, Big Sky, MT, pp. 5 – 19.

Clark, A.G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. Mol. Biol. Evol. *7*:111–122

Clayton, D. (2001). SNPHAP a program for estimating frequencies of haplotypes of large numbers of diallelic markers from unphased genotype data from unrelated subjects. Ver 1.0

Ding, C., and Cantor, C.R. (2003). Direct molecular typing of long-range genomic DNA with M1-PCR. Proc. Natl. Acad. Sci. USA *100*:7449–7453

Eronen, L., Geerts, F. and Toivonen, H. (2004). A Markov chain approach to reconstruction of long haplotypes. Pac. Symp. Biocomput, Fearnhead, P., and Donnelly, P. (2001). Estimating recombination rates from population genetic data. Genetics *159*:1299–1318

Hoehe, M.R. (2003). Haplotypes and the systematic analysis of genetic variation in genes and genomes. Pharmacogenomics. *4*:547–570

Hurley, J.D., Engle, L.J., Davis, J.T., Welsh, A.M., and Landers, J.E. (2005). A simple, bead-based approach for multi-SNP molecular haplotyping. Nucleic Acids Res. 32(22): e186 – e186

International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*:1299–1320

De Iorio, M., and Griffiths, R. (2004). Importance sampling on coalescent histories. II. Subdivided population models. Adv. Appl. Probab. *36*:434–454

Johnson, G.C., Esposito, L., Barratt, B.J., *et al*. (2001) Haplotype tagging for the identification of common disease genes. Nat. Genet. *29*:233–237

Lander, E.S., Linton, L.M., Birren, B., *et al* (2001). Initial sequencing and analysis of the human genome. Nature *409*:860–921

Lazzeroni, L. (2001). A chronology of fine-scale gene mapping by linkage disequilibrium. Stat. Methods Med. Res. *10*:57–76

Lin, S., Cutler, D.J., Zwick, M.E., and Chakravarti, A. (2002). Haplotype inference in random population samples. Am. J. Hum. Genet. *71*:1129–1137

Marchini, J., Cutler, D., Patterson, N., Stephens, M., Eskin, E., Halperin, E. Lin, S., Qin, Z.S., Munro, H.M., Abecasis, G.R., and Donnelly, P. for the International HapMap Consortium. (2006). A comparison of phasing algorithms for trios and unrelated individuals. Am. J. Hum. Genet. *78*:437–450

Mitra, R.D., Butty, V.L., Shendure, J., Williams, B.R., Housman, D.E., and Church, G.M. (2003). Digital genotyping and haplotyping with polymerase colonies. Proc. Natl. Acad. Sci. USA *100*: 5926–5931

Niu, T., Qin, Z.S., Xu, X., Liu, J.S. (2002). Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. Am. J. Hum. Genet. *70*:157–169

Pont-Kingdon, G., Jama, M., Miller, C., Millson, A., and Lyon, E. (2004). Long-range (17.7 kb) allele-specific polymerase chain reaction method for direct haplotyping of R117H and IVS-8 mutations of the cystic fibrosis transmembrane regulator gene. Mol. Diagn. *6*(3): 264–270

Sabeti, P.C., Reich, D.E., Higgins, J.M. *et al*. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*:832–837

Salem, M., Wessel, J., and Schork, N.J. (2005). A comprehensive literature review of haplotyping software and methods for use with unrelated individuals. Hum. Genomics *2*:39–66

Schneidler, S., Roessli, D., and Excoffier, L. (2002). 'Arlequin version 2.*001*: A software for population genetics data analysis', Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Stephens, M., and Donnelly, P. (2003). A compariosn of Bayesian methods for haplotype reconstruction from population genotype data. Am. J. Hum. Genet. *73*:1162–1169

Stephens, M., Smith, N.J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data Am J. Hum. Genet. *68*:978–989

Swofford, D.L. (2003). PAUP*: Phylogenetic Analysis Using Parsimony. Mass: Sinauer Associates.

Tregouet, D.A., Escolano, S., Tiret, L. *et al*. (2004). A new algorithm for haplotype-based association analysis: The stochastic-EM algorithm. Ann. Hum. Genet. *68*:165–177

Venter, J.C., Adams, M.D., Myers, E.W. *et al*. (2001). The sequence of the human genome. Science *291*:1304–1351

Wang, L. and Xu, Y. (2003). Haplotype inference by maximum parsimony. Bioinformatics *19*: 1773–1780

Woolley, A.T., Guillemette, C., Li-Cheung, C., Housman, D.E., and Lieber, C.M. (2000) Direct haplotyping of kilobase-size DNA using carbon nanotube probes. Nat. Biotechnol. *18*:760–763

Xu, J. (2005). Evolutionary Genetics of Fungi. Horizon Biosciences Inc. UK.

Xu, J. (2006). Fundamentals of fungal molecular population genetic analysis. Curr. Issues Mol. Biol. *8*:75–90

Xu, J., and Mitchell, T.G. (2003). Comparative gene genealogical analyses of strains of serotype AD identify recombination in populations of serotypes A and D in the human pathogenic yeast *Cryptococcus neoformans.* Microbiology *149*:2147–2154

Xu, J., Vilgalys, R., and Mitchell, T.G. (2000). Multiple gene genealogies reveal recent dispersion and hybridization in the human pathogenic fungus *Cryptococcus neoformans*. Mol. Ecol. *9*:1471–1482

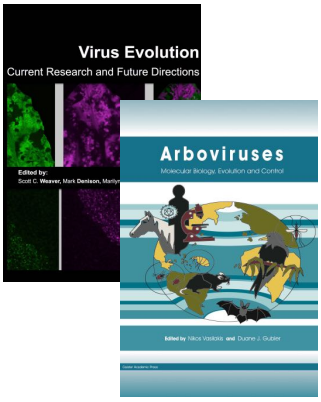
Xu, J., Luo, G., Vilgalys, R., Brandt, M.E., and Mitchell, T.G. (2002). Multiple origins of hybrid strains of *Cryptococcus neoformans* with serotype AD. Microbiology *148*:203–212

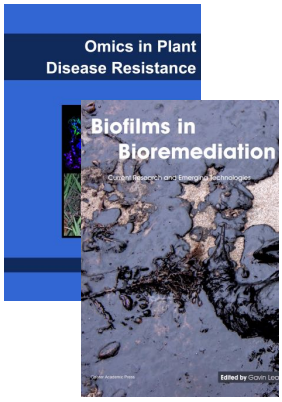
Zhao, L.P., Li, S.S. and Khalid, N. (2003). A method for the assessment of disease associations with single-nucleotide polymorphism haplotypes and environmental variables in case-control studies. Am. J. Hum. Genet. *72*:1231–1250

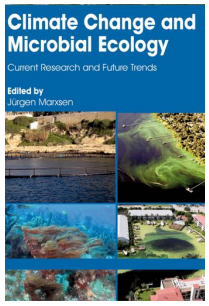
Further Reading

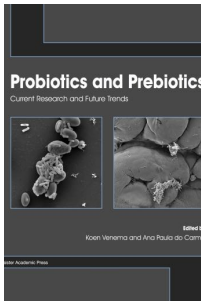
Caister Academic Press is a leading academic publisher of advanced texts in microbiology, molecular biology and medical research. Full details of all our publications at **caister.com**

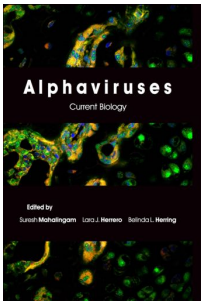
- **MALDI-TOF Mass Spectrometry in Microbiology**
  Edited by: M Kostrzewa, S Schubert (2016)
  www.caister.com/malditof
- ***Aspergillus* and *Penicillium* in the Post-genomic Era**
  Edited by: RP Vries, IB Gelber, MR Andersen (2016)
  www.caister.com/aspergillus2
- **The Bacteriocins: Current Knowledge and Future Prospects**
  Edited by: RL Dorit, SM Roy, MA Riley (2016)
  www.caister.com/bacteriocins
- **Omics in Plant Disease Resistance**
  Edited by: V Bhadauria (2016)
  www.caister.com/opdr
- **Acidophiles: Life in Extremely Acidic Environments**
  Edited by: R Quatrini, DB Johnson (2016)
  www.caister.com/acidophiles
- **Climate Change and Microbial Ecology: Current Research and Future Trends**
  Edited by: J Marxsen (2016)
  www.caister.com/climate
- **Biofilms in Bioremediation: Current Research and Emerging Technologies**
  Edited by: G Lear (2016)
  www.caister.com/biorem
- **Microalgae: Current Research and Applications**
  Edited by: MN Tsaloglou (2016)
  www.caister.com/microalgae
- **Gas Plasma Sterilization in Microbiology: Theory, Applications, Pitfalls and New Perspectives**
  Edited by: H Shintani, A Sakudo (2016)
  www.caister.com/gasplasma
- **Virus Evolution: Current Research and Future Directions**
  Edited by: SC Weaver, M Denison, M Roossinck, et al. (2016)
  www.caister.com/virusevol
- **Arboviruses: Molecular Biology, Evolution and Control**
  Edited by: N Vasilakis, DJ Gubler (2016)
  www.caister.com/arbo
- ***Shigella*: Molecular and Cellular Biology**
  Edited by: WD Picking, WL Picking (2016)
  www.caister.com/shigella
- **Aquatic Biofilms: Ecology, Water Quality and Wastewater Treatment**
  Edited by: AM Romaní, H Guasch, MD Balaguer (2016)
  www.caister.com/aquaticbiofilms
- **Alphaviruses: Current Biology**
  Edited by: S Mahalingam, L Herrero, B Herring (2016)
  www.caister.com/alpha
- **Thermophilic Microorganisms**
  Edited by: F Li (2015)
  www.caister.com/thermophile

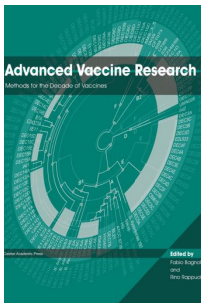


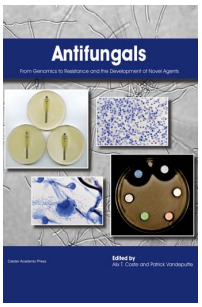












- **Flow Cytometry in Microbiology: Technology and Applications**
  Edited by: MG Wilkinson (2015)
  www.caister.com/flow
- **Probiotics and Prebiotics: Current Research and Future Trends**
  Edited by: K Venema, AP Carmo (2015)
  www.caister.com/probiotics
- **Epigenetics: Current Research and Emerging Trends**
  Edited by: BP Chadwick (2015)
  www.caister.com/epigenetics2015
- ***Corynebacterium glutamicum*: From Systems Biology to Biotechnological Applications**
  Edited by: A Burkovski (2015)
  www.caister.com/cory2
- **Advanced Vaccine Research Methods for the Decade of Vaccines**
  Edited by: F Bagnoli, R Rappuoli (2015)
  www.caister.com/vaccines
- **Antifungals: From Genomics to Resistance and the Development of Novel Agents**
  Edited by: AT Coste, P Vandeputte (2015)
  www.caister.com/antifungals
- **Bacteria-Plant Interactions: Advanced Research and Future Trends**
  Edited by: J Murillo, BA Vinatzer, RW Jackson, et al. (2015)
  www.caister.com/bacteria-plant
- ***Aeromonas***
  Edited by: J Graf (2015)
  www.caister.com/aeromonas
- **Antibiotics: Current Innovations and Future Trends**
  Edited by: S Sánchez, AL Demain (2015)
  www.caister.com/antibiotics
- ***Leishmania*: Current Biology and Control**
  Edited by: S Adak, R Datta (2015)
  www.caister.com/leish2
- ***Acanthamoeba*: Biology and Pathogenesis (2nd edition)**
  Author: NA Khan (2015)
  www.caister.com/acanthamoeba2
- **Microarrays: Current Technology, Innovations and Applications**
  Edited by: Z He (2014)
  www.caister.com/microarrays2
- **Metagenomics of the Microbial Nitrogen Cycle: Theory, Methods and Applications**
  Edited by: D Marco (2014)
  www.caister.com/n2

Order from caister.com/order