Fundamentals of Fungal Molecular Population Genetic Analyses

Jianping Xu

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

Abstract
The last two decades have seen tremendous growth in the development and application of molecular methods in the analyses of fungal species and populations. In this paper, I provide an overview of the molecular techniques and the basic analytical tools used to address various fundamental population and evolutionary genetic questions in fungi. With increasing availability and decreasing cost, DNA sequencing is becoming a mainstream data acquisition method in fungal evolutionary genetic studies. However, other methods, especially those based on the polymerase chain reaction, remain powerful in addressing specific questions for certain groups of taxa. These developments are bringing fungal population and evolutionary genetics into mainstream ecology and evolutionary biology.

Introduction
Fungi play critical roles in human and animal health, agriculture and food industry, biotechnology, and as model organisms for basic scientific inquiries. However, until recently, their importance in these areas had not been fully appreciated. For example, fungal infections have always been part of our humanity but their significance was not recognized by the public until the dramatic increases of fungal infections in the last two decades due to the rising incidences of immunocompromised patients. As a result of this and other developments, there is a pressing need to improve the accuracy and speed of the diagnosis of fungal infections, to identify the sources of individual cases and outbreaks of these infections, and to understand the patterns of genetic variation and evolutionary potentials in populations of pathogenic fungi. These and other fundamental issues apply not only to human pathogens but also to plant and non-human animal pathogens as well as non-pathogenic species. For edible fungi, strain identification and population genetic analysis could help uncover novel, economically important genetic variations for breeding purposes.

Until about two decades ago, most studies of natural fungal strains and populations focused on phenotypic differences in morphology and physiology, and when possible, mating. However, these phenotypic features are often difficult to observe, quantify, standardize, and/or analyze. In recent years, there has been substantial progress in the development of innovative methods to analyze fungi (and other organisms) at the molecular level. This paper will review the common molecular methods currently used for typing species and strains of fungi and discuss the selection of methods to address specific questions.

This paper is divided into two major sections. The first section deals with specific molecular methods. For each method, I briefly introduce the concept and its underlying principles, including its advantages and limitations. The second section deals with the common analytical methods, including their underlying assumptions and the types of data and questions that they are most appropriate to help interpreting and addressing. I will conclude with a brief discussion on future directions. The literature on the development and application of molecular techniques and their associated analytical tools is rapidly expanding. Indeed, we are entering a golden age of fungal evolutionary genetics.

Molecular methods for genotyping fungi
The current revolution in molecular biology has provided techniques to identify numerous fungi. These include the identifications of strains and clonal lineage within a species, and the discriminations and classifications of populations, species, genera, families, orders, classes, and even kingdoms and domains. Molecular methods exploit the tremendous naturally occurring variations in the DNA. This section focuses on protein polymorphisms in isozymes, electrophoretic karyotypes, hybridization of probes to DNA, PCR-based fingerprints, restriction fragment length polymorphisms, amplified fragment length polymorphisms, and DNA sequencing. All these methods could generate molecular markers to compare fungi strains. It should be stressed from the outset that there is no molecular marker ideal for every organism and for addressing every question. Some markers may be better at discriminating individual strains, separate species or higher taxonomic groups. For some purposes, it is important to use markers in specific genes. In other situations, markers in non-coding, usually anonymous portions of the genome are preferable because they are assumed to be neutral or unaffected by selective pressure imposed by the environments. Even with the same organism, different markers may be used to address different questions.

A molecular marker refers to any detectable property that identifies a specific region of the genome.

Isozyme electrophoresis
The electrophoretic migration of proteins such as enzymes is among the most cost effective methods to investigate genetic variation at the molecular level. There are five common methods of protein electrophoresis and they differ in the nature of the supporting medium (or gel) and whether they are run horizontally or vertically: starch horizontal; starch vertical; polyacrylamide vertical; agarose horizontal; and cellulose acetate. These methods have been compared and reviewed in detail by Murphy

For correspondence: jpxu@mcmaster.ca

et al. (1996). Regardless of the supporting medium and orientation of running direction, the basic principle of protein electrophoresis is the same and can be described as follows.

The migration (M) of a protein is influenced by many factors, including its net charge (Q), molecular size as measured by its radius (r), the strength of the electric field (E), and the viscosity of the supporting gel (V). The relationship between M and the other four factors can be described as:

$$M = QE/4\pi r2V$$

Under appropriate conditions, the rate of M increases with the net charge of the protein, which is influenced by the pH of the buffering system and the strength of the electric field, and M decreases as the molecular size of the protein and the viscosity of the suspension medium are increased. The above formula assumes that the protein is globular. Differences in protein shapes can also affect migration.

Most useful isozymes are functional enzymes that differ in amino acid sequences. These differences in amino acid sequences can contribute to both the charge of the molecule and its 3-dimensional structure. After electrophoresis, the variant bands of an enzyme are detected and recognized by adding the appropriate substrate and a detection system. The substrate is often coupled with a dye that is released upon enzymatic activity and can be visualized using naked eyes.

The accurate acquisitions of isozyme data require that the observed banding patterns on gels are correctly interpreted. Conventional interpretations have two basic assumptions. The first is that changes in the mobility of an enzyme in an electric field reflect a change in its amino acid sequence and thus by inference, the encoding DNA sequence. Therefore, if the enzyme banding patterns of two individual organisms differ, such differences are assumed to be DNA-based and heritable. The second assumption is that enzyme expression is co-dominant, that is, every allele at a locus is expressed.

In population genetic terms, enzyme electrophoretic data we obtained from a gel can be divided into two types: isozymes and allozymes. Isozymes are functionally similar forms of an enzymatic protein, including all its subunits, which may be produced by different gene loci in the genome or by different alleles at the same locus. In contrast, allozymes are a subset of isozymes, in which polypeptide variants of the enzyme are formed by different allelic alternatives at the same gene locus in the genome. Strictly speaking, only allozyme data can be used to assign alleles and calculate allele frequencies. Complex isozyme data are useful for certain population genetic analyses only when the genomic locations of migrating bands are correctly interpreted. Often, obtaining allozyme data from isozyme data requires genetic crosses and the analyses of meiotic progeny. In asexual diploid fungi, such as the human pathogenic yeast *Candida albicans*, it is impossible to infer allelic status correctly from isozyme patterns that involve either multiple loci and/or enzymes with polymeric structures.

Electrophoretic karyotype (EK)

Fungi are highly variable in the number and size of chromosomes (Zolan 1995). The variation in both the number and the size of fungal chromosomes can be detected by electrophoresis under conditions that provide alternating fields of electric current, often referred to as the pulsed field gel electrophoresis (PFGE) (Zolan 1995). Several instruments and procedures have been developed for PFGE, and perhaps the most common is the contour-clamped homogeneous electric field (CHEF). In these procedures, intact chromosomes migrate through an agarose gel matrix under the influence of the pulsed fields. Following electrophoresis and optimal separation of chromosomes, the gels can be stained with ethidium bromide and viewed under ultraviolet light to analyze the patterns of chromosomal banding, or electrophoretic karyotype (EK).

This technique can potentially detect large deletions, insertions, duplications and translocations among chromosomes. However, identifying genes and chromosomes that cause these polymorphisms requires additional analyses, such as digestion with certain endonucleases, analysis of restriction fragments, blotting and probing the chromosomal gels with specific probes. A common modification of EK is to first digest the chromosomes with rare cutting restriction endonucleases (e.g., the 8-base cutting enzymes *Not1* and *Sfi*), and then use PFGE to separate these large restriction fragments for further analysis.

In some species, the EK method compares favorably with other strain typing methods. Indeed, in some cases, fungal EKs may be too variable and unstable. Diverse EK types have been observed following asexual propagation and sub-culturing of a single genotype (Fries *et al*., 1996). Additional drawbacks of EKs include the difficulty in (i) determining homologous chromosomes; and (ii) quantifying the differences among different EKs. Because chromosomes must pair during meiosis, it has been assumed that the EKs of sexual species should be less variable than those of asexual ones. However, the EK types of sexual fungi have been found to vary greatly in size and gene arrangement, similar to many presumed “asexual” strains and species (Suzuki *et al*., 1988; Zolan 1995).

DNA–DNA hybridization

Several methods of molecular typing rely on hybridization between complementary strands of DNA. DNA–DNA hybridization techniques can offer a quantitative measure of overall genomic similarity among strains. The most common DNA–DNA hybridization is the DNA re-association kinetics. However, hybridization using oligonucleotide probes and DNA microarrays are becoming powerful tools for assaying fine-scale genetic differences. Whole-genome microarrays are capable of detecting such differences on a whole genome scale.

DNA re-association kinetics

The DNA re-association kinetics method takes advantage of the double-stranded nature of genomic DNA in which nucleotides on opposing strands are held together

by hydrogen bonds. Two hydrogen bonds are formed between adenine and thymine while three hydrogen bonds link guanine and cytosine. When double-stranded DNA is heated to around 100°C, the hydrogen bonds between complementary base pairs are broken and the two single strands separate. During subsequent cooling of the solution, the complementary DNA strands can re-anneal. If DNA from two different species are combined, denatured, and then allowed to re-anneal, the double stranded molecules that form between complementary strands from the two species will contain base pair mismatches (Kurtzman 1993). The extent of mismatching determines the temperature at which these hybrid molecules melt when they are placed in a thermal gradient. The more mismatches, the lower the temperature at which the hybrid strands will separate. The decrease in the melting temperature of a heteroduplex hybrid relative to a homoduplex control provides an index of divergence and similarity between the DNA samples (Kurtzman 1993).

It is critical that the conditions of re-association be standardized because the amount of base pair mismatches that form in the hybrid molecules can be affected by salt concentration, temperature, buffering conditions, and DNA fragment size. Under highly stringent conditions of re-association, base pairing between DNA strands will only occur between well-matched sequences. High stringency conditions include increased temperature and/or decreased salt concentration. In contrast, under conditions of progressively lower stringencies, more mismatches will be tolerated during re-association.

Oligonucleotide hybridization

This technique utilizes known single nucleotide polymorphisms (SNPs) in a species. Oligonucleotides of more than 20 bases can be designed and synthesized with known polymorphic sites placed near the middle of the oligonucleotide. This oligonucleotide can be end-labeled with a radioactive tag or a fluorescent dye. The labeled probes can be then hybridized by conventional Southern hybridization to either total genomic DNA or specific gene fragments amplified by the PCR. This technique has been applied to detect polymorphisms in several fungal species, including *C. albicans* (Cowan *et al*., 1999). The presence or absence of a hybridization signal for each probe can be scored as alternative alleles at a specific site. The drawbacks of this technique are that (i) it requires two hybridizing procedures for every locus in diploid individuals, and (ii) unknown mutations at any of the 20 or so nucleotides may cause the loss of a hybridizing signal. Therefore, the loss of a hybridizing signal may not be attributable to the specific nucleotide site.

DNA chip

The new technology of DNA microarrays on chips represents a miniature but mass version of individual oligonucleotide hybridization. This method has been used widely to detect whole genome expression profiles. However, it has also been used to detect genetic variations among strains. Briefly, DNA chips are glass surfaces to which arrays of specific DNA fragments have been attached at discrete locations. These fragments serve as probes for hybridization. Under conditions suitable for hybridization, the DNA spots on the chip are exposed to a solution containing a complex sample of fluorescent-labeled DNA. Though this technique has been established only recently, it has been used for many species, including fungal species. The first successful application for genomic variation identification involved the model yeast *Saccharomyces cerevisiae* (Winzeler *et al*., 1998). In this example, the array consisted of 20 complementary pairs of oligonucleotides (each with 25-mers) for each of the 6400 genes in the *S. cerevisiae* genome. In addition, there were three permutations of each consensus 25-mer with each permutation having a single base change in the central nucleotide position to account for all four possible nucleotides A, T, G, and C. Thus, many base pair substitutions of the gene are represented on the chip. A DNA sample of each isolate is then labeled with a fluorescent dye and hybridized to the array. The arrays are scanned to retrieve signals for each spot and the hybridization patterns are then compared among strains. Specific mutations across the whole genome can be inferred (Winzeler *et al*., 1998, 2003).

Polymerase chain reaction (PCR)-based methods

The PCR technology has spawned many procedures for typing strains and species, some of which have become standard methods for species and strain identifications. PCR methods are easy to set up and have the advantage of requiring only minute amounts of starting material or template DNA. Although simple in concept, PCR methods have unrivaled, often overlooked complexity. The source of this complexity includes multi-ionic interactions, kinetic constants, and enzymatic activities etc. These factors can repeatedly affect the reactants in a typically small PCR reaction volume over an extended time period. Despite these potential problems, many methods have been developed and are widely used. Below I describe some of the common PCR-based strain typing techniques.

Random amplified polymorphic DNA (RAPD)

In RAPD analysis, genomic or template DNA is primed at a low annealing temperature (30–38°C) with a single short oligonucleotide (ca. 10 bases) in the PCR. Multiple PCR products of different electrophoretic mobility are typically generated (Williams *et al*., 1990). RAPD analysis detects two types of genetic variations: (i) in the length of DNA between the two primer binding sites, and (ii) in sequence variation at the priming regions. Nucleotide substitutions in the region of PCR primer binding, particularly at the 3′ ends, can prevent binding of the primer to the DNA template. As a result, this band will be missing in a PCR reaction. Similarities in banding profiles among strains (i.e., the number and mobility, but not the density of the bands) can be calculated and used to infer strain relationships. When multiple primers are screened, RAPD analysis can be very sensitive to detect variation among isolates that cannot be observed using other methods.

Although technically fast and simple, there are some disadvantages to RAPD. The major drawback is reproducibility. RAPD analysis can detect minute variation among strains because, as noted above, even a single nucleotide mismatch in the priming region may prevent annealing and the absence of a characteristic band on gels. Small differences in any aspect of PCR conditions that affect binding of the primer may have similar effects;

consequently, RAPDs are sensitive to the vagaries of the testing procedure. This problem can be minimized if strains under study are treated identically. When multiple strains are compared, the same PCR buffer, the master mix (includes all four nucleotides, primers, appropriate ions, and DNA polymerase) and the same thermal cycler and PCR running program should be used at the same time (Xu *et al*., 1999a; 2000a).

The second concern for RAPDs is that bands with the same electrophoretic mobility may not share the same sequence. This problem may be common for interspecific studies and can be affected by the conditions of electrophoresis. Additionally, with the usual concentrations of agarose gels (1–1.5% of weight/volume), it is often difficult to distinguish RAPD bands that differ in sizes of less than 20 base pairs.

The third concern with RAPDs is the problem of dominant and null alleles. In haploid organisms, both the dominant (presence) and recessive/null (absence) alleles can be scored. However, in diploid organisms, it is often not possible to distinguish genotypes that are homozygous for the dominant allele (1/1) from those that are heterozygous (1/0). Therefore, RAPD data are generally not ideal for inferences of population genetic history in diploids.

Nevertheless, for distinguishing strains and developing fingerprints for molecular epidemiology, RAPDs can be highly effective (e.g. Xu *et al*., 1999a; 2000a).

PCR fingerprinting

PCR fingerprinting is similar to RAPD, except that primers are longer (>15 bases) and annealing temperatures are higher and PCR conditions more stringent. Most PCR fingerprinting primers are designed from repetitive DNA sequences (Xu *et al*., 1999a; 2000a). Commonly used PCR fingerprinting primers in fungi include M13, which is derived from the core sequence of phage M13; T3B, which originates from the internal sequences of tRNA genes; and TELO1, which is based on fungal telomere repeat sequences. Because of more stringent reaction conditions, PCR fingerprinting is generally more reproducible than RAPDs. Nonetheless, it suffers the same problems of interpretation as RAPDs. However, under standardized conditions, PCR fingerprinting has proven quite reliable for discrimination and the identification of species and strains.

Microsatellite loci

One emerging technique exploits the hypervariability of DNA regions composing multiple tandemly repeated units of di-, tri- or multiple nucleotides. This hypervariability can be caused by either strand slippage during DNA replication or unequal crossing-over during meiosis, both can occur much more frequently than nucleotide substitutions. Useful microsatellites can be located by probing a genomic library with simple repeated sequences or by searching databases of gene sequences. PCR primers flanking these repeat regions can be developed and PCR products can be run on polyacrylamide gels to detect differences in repeat numbers (Field *et al*., 1996). One potential drawback of this technique is that because multiple alleles are often found at a single locus, the relationships among alleles can be difficult to decipher, and alleles may be identical by convergence, not by descent.

PCR-RFLP of known genes

With increasing knowledge of genes and genomes from fungi, the supply of single-copy genes for genotyping is now feasible for many fungal species. These gene sequences can be used to investigate the variability among strains and the history of populations and species. One fast application is to design PCR primers to amplify a particular gene from representative strains followed by digestion of the amplified products with an array of restriction enzymes to screen for variability. Variable restriction sites can then be used to screen a larger sample of isolates (Xu, 2002). The gene specific PCR in combination with restriction digestions can generate excellent co-dominant markers that are highly stable and reproducible, ideal for both haploid and diploid organisms (Xu *et al*., 1999b).

Single-strand conformation polymorphism (SSCP)

SSCP is a promising technique that allows efficient detection of nucleotide substitutions in short fragments (<500 bp) of DNA. SSCP analysis typically involves the amplification by PCR of a unique segment of genomic DNA, melting the PCR products, and running the single strands on a non-denaturing polyacrylamide gel (Hauser *et al*., 1997). The detection system can be accomplished by either radioactive labeling of DNA during the PCR amplification step or by silver staining of DNA after gel electrophoresis. Polymorphic differences in strand mobility result from the effects of primary sequence changes on the folded structure of a single DNA strand. The primary sequence differences alter the intra-molecular interactions that generate a three-dimensional folded structure. The molecules may thus move at different rates through a non-denaturing polyacrylamide gel. Because these conformational variations are subtle, the success of any particular SSCP experiment depends heavily on the following two factors: (i) the particular DNA fragments being investigated, including the primary DNA sequence organization and the size of the DNA fragments, and (ii) the optimization of experimental conditions to maximize differential migration among fragments. Investigators have used a variety of methods to improve the resolving power of SSCP, including adding glycerol to polyacrylamide gels, reducing temperatures, and increasing the length of the gels or the duration of gel electrophoresis. Nonetheless, differentiation among polymorphic molecules on a polyacrylamide matrix is not entirely predictable, and the method can result in false negatives, ambiguous results and experimental artifacts.

Heteroduplex

The analysis of heteroduplex is dependent on conformational differences in double stranded DNA. In this technique, PCR products from two different strains in equal quantities (e.g., from wild and mutant DNA samples) are combined in a non-denaturing buffer (Olicio *et al*.,

1999). The DNA is melted at high temperature (e.g. 95°C) and is then slowly cooled to room temperature. During the cooling process, the complementary single strands from the same origin strain anneal to form homoduplex DNA, and the complementary single strands from different origins also re-anneal but form heteroduplex DNA. The mismatch in the heteroduplex DNA causes the re-annealed double strands to have a different flexibility and three-dimensional shape than homoduplex DNAs. As a result, the mobility of heteroduplex DNAs will be slower than that of homoduplex DNA. The running and detection conditions for heteroduplex analysis are similar to that for SSCP. Heteroduplex analysis works well for fragments with 200–600bp in length.

Amplified fragment length polymorphism

The development of amplified fragment length polymorphism (AFLP) method has had a significant impact in its relatively short history. AFLP is a powerful method for fingerprinting strains and for generating a large number of dominant markers for the analysis of genetic crosses (Vos *et al*., 1995). The procedure is briefly described as follows. Genomic DNA samples are first digested with two endonucleases (usually a frequent cutter and a rare cutter). Double stranded DNA adapters are then ligated to the ends of the DNA fragments to create template DNA for PCR. The adapters consist of a core sequence and an enzyme-specific sequence that allow the ligations to occur. The ligated products are then amplified with AFLP amplification primers. Amplification primers consist of the core sequence (same as the adaptor core sequence), the enzyme-specific sequence and a selective extension of one to several nucleotides, depending on the complexity of the study genome. These selective bases will allow amplification of a subset of the restriction fragments. AFLP usually involves two PCR steps. The first step is the pre-amplification step that uses unlabelled primers with a single selective nucleotide in the primer. After the first step, the reaction mixtures are diluted for second PCR amplifications. In the second amplification, additional selective nucleotides are often added to enhance specificity. The selective second step often uses fluorescently or radioactively-labeled primers.

AFLP has several powerful advantages over the other methods. Many more fragments can be generated and analyzed in a simple reaction. It can detect restriction site variations as well as insertions and deletions within a genomic region. Different enzymes and/or selective extension nucleotides can be used to create new sets of markers. Therefore, AFLP can provide an almost limitless set of genetic markers. In addition, the fragments are stable and highly reproducible since they are amplified with two specific primers under stringent conditions.

Restriction fragment length polymorphisms

Restriction polymorphisms have been used to discriminate species and strains of fungi as well as other biological taxa. One approach is to digest genomic DNA with a restriction enzyme and directly examine the resulting bands in agarose or polyacrylamide gels after electrophoresis. Depending upon the size of the genome and the frequency of restriction recognition sites in the genome, it may be possible to directly compare digests of whole genomic DNAs from different species/strains. In complex genomes such as those in fungi, this direct comparison can only detect differences in high copy number DNA molecules, e.g. the ribosomal DNA genes and mitochondrial DNA. For low copy number genetic elements, it is almost impossible to observe restriction site polymorphisms through simple digestion and electrophoresis on agarose or polyacrylamide gels.

Alternatively, PCR product of a gene can be digested and analyzed as described above. For between species comparisons, it is often possible to obtain restriction site differences in the ribosomal DNA motif by PCR-RFLPs. Because it is frequently difficult to accurately determine the migration of bands (i.e., DNA fragment sizes), comparisons should be made on samples in adjacent lanes of the same agarose gel with size gradients on both sides of the gel.

The most widely used RFLP method is a DNA–DNA hybridization-based technique that involves cutting genomic DNA with restriction endonuclease(s), separating the DNA fragments in agarose gels with electrophoresis, transferring DNA onto membranes, and hybridizing the membranes with labeled specific probes (e.g. Xu *et al*., 1997, 1998).

RFLPs generated only through total genomic digests or with additional Southern hybridization using repetitive elements are challenging for interpretations. This is because these kinds of banding patterns are fingerprints and are difficult to relate to specific alleles of individual loci. In addition, when the number of bands is high, the accuracy in determining the number and size of DNA bands will be limited and affected by the electrophoretic conditions.

For RFLPs detected by Southern hybridization with probes targeted to single copy DNA markers, the interpretations are straightforward and data can be used in a variety of ways. Single copy RFLP markers are excellent for addressing population and evolutionary genetic questions in diploid or dikaryotic fungi (e.g. Xu *et al*., 1997, 1998).

DNA sequencing

The most accurate but laborious method to catalog differences at the molecular level is directly sequencing cloned genes or PCR products. This approach can provide data for both high level phylogenetic analyses among species and for the analyses of genetic variations among strains within populations and species. For phylogenetic analysis among species or higher taxonomic levels, the most common genes to be sequenced and compared reside in the ribosomal RNA (rRNA) gene cluster, including the internal transcribed spacer (ITS) regions ITS1 and 2, the inter genic spacer IGS, 5.8S rRNA, 18S rRNA, and 26S rRNA genes. This is because these multi-copy genes are high conserved within a species but can be quite variable among species. Other commonly used genes include the mitochondrial ATPase subunits, beta-tubulin, and elongation factor.

For comparisons among strains within a species, protein coding and non-functional DNA fragments are usually more informative than the conserved rRNA genes.

This is because rRNA genes are under strong concerted evolution pressure whereas non-coding sequences and third-base substitutions are less constrained. With increasing accessibility and decreasing cost, multilocus DNA sequencing is becoming a mainstream genetic data-gathering tool in many studies (e.g. Xu *et al*., 2000b, 2002, 2003b).

Is there an ideal, cost-effective molecular method for all organisms that is capable of addressing every evolutionary genetic question?

The short answer to this question is no. Practically speaking, there is no best or worst method among the above described molecular typing techniques. Different typing methods are appropriate for addressing different population and evolutionary questions in different species. As I will briefly show below, appropriate sampling can be more important than the typing techniques used in addressing many questions. Below are a few example questions in medical mycology and how different molecular techniques could help addressing them.

Which species does the infectious agent belong to?

While traditional/classical identification schemes (e.g. API-20C and API-32 for pathogenic yeasts, and morphological features for filamentous fungi) are still the mainstream clinical microbiological methods and are usually adequate to address this question, there is a growing need for more efficient and rapid diagnostic method. Furthermore, most phenotypic characters used by traditional methods can vary among strains within a species. At present, portions of the 26S rRNA sequence are available in the Genbank for almost all human pathogenic fungi. Therefore, DNA sequence-based identification is becoming feasible and will probably play an increasingly significant role in the future for species identification. All other molecular methods described in the last section also have the potential for species identification. However, to do so, a standard large database for these methods needs to be established and confirmed by a large number of investigators. At present, only the GenBank ribosomal RNA gene database is widely applicable.

What is the source of the infecting strain?

Once the species is identified for the pathogen, the next question is where the causal agent comes from. This is a more difficult question to address and a lot more information is needed. Most human fungal pathogens are either commensal organisms (i.e. part of the normal human microflora, e.g. *Candida* spp.) or ubiquitous in certain environments (e.g. the worldwide distribution of *Cryptococcus neoformans* associated with bird droppings, soil and certain tree species). For commensal organisms, there are at least three possibilities. The first is that the causal agent is the original colonizing strain, i.e. part of the host's commensal microflora. The second is that the causal agent is a mutated form of the original colonizing strain. The third possibility is that the causal agent comes from a specific source outside the host (e.g. other hosts or the physical environments). For pathogens with significant environmental niches, the causal agent could come from either other hosts or a specific environment.

To test these hypotheses, a variety of time-sensitive materials should be taken. These include the commensal microflora of the host, local environmental samples (e.g. various hospital and residential settings), and people sharing common environments (family members, doctors, and nurses). When these samples are available, we can use a variety of techniques to define the similarities and differences among strains. In this case, DNA sequencing of specific genes might be of little use, but AFLP, RAPD, and PCR-fingerprinting could be potentially very useful. Other techniques can also be used, especially those generating large number of polymorphisms that can be unambiguously scored and quantified. Molecular markers detecting no variation among strains within a species is of little use in strain diagnostics, even though these intra-specific invariable markers can be useful for species identifications. When data are collected, appropriate statistical tests can be used to reject or confirm the hypothetic origins.

Determining the sources of environmental pathogens can be more tedious than those of commensal organisms. This is because we often have very little knowledge of the extent of genetic variation among environmental populations of pathogenic fungi. However, samples similar to the above mentioned should be collected from the environment for critical evaluation. Additional strains from suspected areas where infection likely occurred will be particularly useful.

What is the origin of antibiotic resistant strain?

The origins of antibiotic resistant strains can be addressed in a similar way as to those regarding the sources of infection. However, susceptibility to the antibiotic under investigation should be evaluated for strains from different samples. Briefly, when strains are genotyped and compared using clustering analyses, independent placements (i.e. lack of clustering in an evolutionary tree) of different antibiotic resistant strains suggest independent origins of resistance (Xu *et al*., 2000a). Alternatively, the clustering of resistant strains in a genotype similarity tree suggests a clonal origin of the resistant genotype and the horizontal spread of this genotype among hosts (Xu *et al*., 2000a).

What are the structures of the fungal populations and their evolution potential?

Addressing this question requires genetically interpretable markers as well as meaningful population samples of appropriate sizes (see next section). The selection of markers will depend on the ploidy of the species and the amount of standing genetic variation in populations. For haploid species, dominant-recessive markers (e.g. AFLP and RAPD) can be as useful as co-dominant markers. However, co-dominant markers will be more informative than dominant-recessive markers even in haploid species because multiple alleles (> 2) can be detected at the same time. For diploid species, only co-dominant genetic markers can give you enough information to infer mode of reproduction in nature as well as genetic differences among populations. The analytical methods for understanding the genetic structure of fungal population will be discussed below.

Future considerations of molecular typing techniques

With the completion of a number of fungal genome sequencing projects, abundant genomic information is now available to design primers for species-specific identification systems. The present database can also be used to generate gene-specific products for further comparisons among strains within individual species. Furthermore, developing a set of genetic markers from genes with known functions could facilitate the analysis of inter-species population genetic studies between closely related species.

Population genetic analyses

In this section, I will review the common analytical approaches in understanding the patterns of natural genetic variation as well as the potential mechanisms responsible for the observed patterns of variation. The intention here is to introduce the basic concepts, issues and rationales for population genetic-based studies of fungi. Their limitations and pitfalls will also be introduced and discussed.

What is a population?

There are several different but overlapping definitions of "population" (Xu and Mitchell 2003a; Mish, 1996). A population can be (i) a group of organisms inhabiting a particular locality; (ii) a group of interbreeding organisms that represents the level of organization at which speciation begins; and (iii) a group of objects from which samples are taken for statistical measurement. The first definition is based on geographic locations, although the size and boundaries of individual populations can vary widely and are often arbitrary. This is the definition where most population genetic studies of fungi and other organisms are conventionally used. The second definition is genetically based and more restrictive than the other two. By this definition, a population refers to groups of individuals that are genetically isolated but still capable of interbreeding with individuals in other such groups within the same species. For many species, it is usually difficult to establish the precise breeding boundaries for groups of individuals. In species of plants and animals, both geographical and ecological factors have been found to play important roles in determining the population breeding boundaries. Another problem associated with this definition is that over 20% of the 80,000 or so fungal species identified so far are not known to have observable sexual mating and meiosis under laboratory conditions (Hawkesworth *et al*., 1995)

The third definition is the most versatile of the three. It allows multidimensional analyses on the distribution of genetic variation within species. This definition of a population is an operational one. For example, a population of *Candida albicans* can be a collection of individual strains from a continent, a country, a state/province within a country, a county, a city, a town or village, a specific ecological niche (e.g. certain associated disease conditions, specific body sites of hosts), or different body sites of a single host. Furthermore, a population could be defined based on the sex, age, and/or ethnic backgrounds of the host, regardless of other characteristics. The levels of sample organizations can vary and often depend on sample sizes and analytical objectives. The size of a sample appropriate for detecting differences between populations depends on the patterns of genetic variation of population. For example, the smaller the actual difference in allele frequencies, the larger the sample sizes needed to reliably detect them at a statistically meaningful level. When different populations are compared, the sample size (N) per population needed to detect a given level of differentiation at a diploid locus among the populations for at least 50% of the time (i.e. a power of 0.5) with type I error of 0.05 can be approximated as:

$$2N = 1/F_{ST}$$

(F_{ST} represents the proportion of the total genetic variation explained by the difference between samples. For detailed explanations and calculations, see below). Specifically, to detect a statistically significant F_{ST} value with a P of 0.05 in a diploid species, a sample of 10 individual organisms per population is needed (or 20 strains per population for haploids).

The third definition is commonly used in research on epidemiology and prevention of infectious diseases. It has been used for identifying risk factors among humans and potential virulence factors in pathogens.

Molecular markers, biological systems, and analytical methods

The types of markers amenable for population genetic-based analysis depend on the ploidy level and mating system of the species under investigation. An important issue is that the marker information should be interpretable to alleles of specific locus and that the alleles at each locus can be obtained for all strains. In haploid species, most markers can be easily analyzed by population genetic approaches. This is because there is only one set of genetic material (1N) for each strain and each specific marker could be scored as a locus with two alternative alleles: one allele representing the presence of the marker and the other the absence of the marker. In diploid species (2N), the interpretation on the number of loci and the number of alleles per locus for many dominant-recessive molecular markers (e.g. PCR fingerprinting) can be tedious and often not feasible. In species with a sexual reproductive system, crosses could be constructed and meiotic progenies be analyzed to determine the number of locus and the number of alleles per locus for each marker system. Such an analysis would reduce ambiguity in locus and allele assignments for complex fingerprinting patterns. For diploid species with no known sexual cycle, the interpretation of complex fingerprinting patterns is highly problematic. Generally speaking, co-dominant, single copy genetic markers are best suited for allelic assignments for strains of diploid species. Examples of these markers include allozymes (a subset of isozymes, see above), single locus RFLP, and DNA sequence-based single nucleotide polymorphisms (SNPs).

Some markers are better for addressing certain questions than others. For example, selectively neutral, co-dominant markers are more suitable for examining the

roles of recombination and gene flow between populations. Loci under selection will have higher probability of convergent and parallel evolution than neutral loci. For example, genes involved in drug resistance and in response to host defenses in human pathogenic fungi are likely under severe selection pressure in clinical settings, therefore these genes may not be best suited for examining recombination and genetic differentiation in natural populations.

Analysis of genetic variation within a population

Issues and rationales

Many simple measures can be used to describe the genetic variation within a population. These include, but not restricted to: (i) the number of alleles per locus; (ii) the frequencies of individual alleles; (iii) the observed heterozygosity; (iv) the gene diversity; (v) the mean genetic distances between strains; and (vi) the observed genotypic diversity.

The observed heterozygosity, Ho, represents the percentage of observed heterozygotes at each locus. Gene diversity, He, is defined as the expected heterozygosity, $He=1-\sum p_i^2$, where p_i is the frequency of the *i*th allele at a locus. Both the observed heterozygosity and the gene diversity are individual locus-based measures of genetic variation for population samples. The mean observed heterozygosity and mean gene diversity of a sample are typically estimated as the arithmetic mean of all loci tested.

A common measure of multilocus population genetic variation is the observed genotypic diversity. This diversity measure is calculated as:

$$(1-\sum p_i^2)N/(N-1)$$

Where p_i is the frequency of the *i*th multilocus genotype and N is the sample size. This diversity represents the probability that random pairs of isolates in the sample will have different multilocus genotypes.

One of the most frequently discussed issues dealing with within-population genetic variation in fungi (and other microbes) has been the role recombination plays in the patterns of genetic variation in natural populations.

Since all microbes are known to reproduce asexually via mitosis, it is therefore expected that most microbial species would show some evidence of clonal structure in nature (Xu, 2004). Therefore, one commonly asked question in fungal population genetic studies has been whether recombination plays any role in generating genetic variation in natural populations.

Why is the understanding of recombination in natural populations of microbes important? First, this understanding is of intrinsic interest from an evolutionary point of view (Maynard Smith 1978). Historically, sexual reproduction and recombination pose an evolutionary paradox. An organism that reproduces asexually passes on all its genes to each of its individual progeny, whereas one that reproduces sexually passes on only half to each. Other things being equal, natural selection favors asexual reproduction, because given the same number of progeny, the asexual individual has double the fitness (gene copy) of the sexually reproducing one.

Why are there sexual reproductions then? Two possible advantages of sexual reproduction have been proposed: mixis of genes and DNA repair (Maynard Smith 1978; Bell 1982; Michod and Levin 1988). The mixis argument goes as follows. Without the mixis of genes generated by sexual recombination, adaptive evolution is limited to the accumulation of favorable mutations that happen successively in each independently evolving lineage. With sexual recombination, favorable mutations arisen in separate lineages can become combined in the same individual, thus providing an advantage in the adaptation to changing environments. The repair argument points out that the two gene copies from different parents provide an error-correction mechanism for repairing genetic damages. Genetic damages can be generated spontaneously and continuously in the DNA replication and possibly during transcription processes. With two copies, the intact DNA of one copy (or haplotype) can serve as a template for correcting the damaged DNA in the other haplotype. Moreover, deleterious mutations in one haplotype can be masked by complementary wild-type alleles in diploids. Whether either one or both of the purported advantages can account for the origin and maintenance of sexual reproduction is a subject of much debate and investigation (e.g. Michod and Levin 1988, Xu 2004).

The second reason for assessing the role of recombination in natural populations is its practical importance. This is because whether recombination occurs in natural populations has significant implications for the evolution and spread of genes related to antibiotic resistance, pathogenicity, host and vector specificity. A recombining population structure implies that selected entities are non-recombining genetic elements (usually distinct genes are assumed to be non-recombining units, even though intragenic recombination has been found in all groups of organisms critically examined so far). On the other hand, a clonal population structure implies that the selected units are clones or clonal lineages. In a clonal population, the study of medically important traits must select representatives of every clonal lineage. Contrary, in a recombining population, studies would be more fruitful if focused primarily on individual genes.

How are clonal and recombining population structures determined then? Unlike investigators of plants and animals who can study reproductive mode with a pair of naked eyes or binoculars, microbiologists must use molecular markers, microscopes and population genetic methods. The next section introduces the common genetic tests for determining whether recombination occurs in populations. Because of intrinsic differences in genetic systems, tests for haploid and diploid species are somewhat different, as shown below.

Analysis of clonality and recombination in haploids

Since there is only one set of chromosome in a haploid genome, there is only one allele for each locus for an individual strain. Tests of recombination in natural population therefore involve examining the associations among alleles from different loci.

Different scientists may define a genetic locus differently. In the current population genetic literature, a locus could mean one of several things: (i) a single

polymorphic nucleotide site, (ii) a polymorphic restriction endonuclease recognition site (several base pairs), (iii) a single insertion/deletion, (iv) a whole continuous stretch of DNA of arbitrary lengths with any difference among them treated as distinct alleles, and (v) an enzymatic staining profile. Because there is a big variation in the size of DNA/gene being recognized as a locus, analytical methods can be different. The basic assumption in all analytical methods include that each distinct allele is the result of a unique mutational event and occurred only once in the population history. In practice, this is to assume that indistinguishable alleles are identical-by-descent.

It should be emphasized here that there are two distinct questions in tests for the roles of recombination in natural populations. The first question is whether a population is panmictic. A panmictic structure implies that alleles at all loci are randomly associated with each other at the population level. However, if the panmictic hypothesis is statistically rejected and a predominantly clonal population structure is assumed, then the second question is whether recombination plays any role in generating genetic variation in the examined natural population. Since all medical fungi are capable of reproducing asexually through mitosis, a clonal component in populations is therefore expected. One widely used approach is to analyze representatives of each different multilocus genotype to distinguish between the null hypothesis of recombination and the alternative hypothesis of clonality. The rejection of panmixia for the total sample but the acceptance of panmixia for the clone-corrected sample is usually interpreted as evidence of a random mating genetic structure with episodes of clonal expansion.

However, there are several problems with censoring sample before testing. First, the decrease in sample sizes can sometimes greatly decrease the statistical power in rejecting the null hypothesis, thus increasing Type II error (Sokal and Rohlf 1981). Second, the justification for clonal censoring of sample sizes in the analysis should be biologically and ecological based. Thirdly, though often assumed, it is usually not confirmed that strains with identical multilocus genotypes derived from a certain genetic marker system are actually identical and clonal in origin.

Tests for whether haploid population is panmictic typically involve comparing observed allelic associations with those derived under the null hypothesis of random mating. There are three common tests (Table 1). The first is to determine the extent of allelic association (linkage equilibrium) between pairs of loci. The second is to exam the overall index of association (I_A) involving all examined loci. These two tests use panmixia as the null hypothesis. The third test uses complete clonality as the null hypothesis. It compares phylogenies from different genes. The existence of phylogenetic incompatibility indicates evidence of recombination. In this third test, any incongruence among gene genealogies from different genes would suggest recombination. These tests are briefly described below.

Linkage disequilibrium

Linkage disequilibrium (also called gametic phase disequilibrium or gametic disequilibrium) is a measure of association between alleles from pairs of loci. Random association between alleles at different loci is an indicator of recombination between these pairs of loci in the population. This test is briefly summarized as follows.

If alleles at two loci, A and B, segregate independently, then the expected frequency of the genotype A_iB_i is simply the product of the frequencies of the two alleles A_i and B_i. A Chi-square test or the Fisher's exact test can be performed to determine whether the observed genotypic counts are significantly different from the expected counts (Hartl and Clark 1989). If the observed and expected counts are not significantly different, then the two loci in the population under study are assumed to be recombining. On the other hand, if the observed and expected counts are significantly different, then this population is assumed to have a non-recombining structure as inferred by the loci. In a completely panmictic population, less than five

Table 1. Common criteria for distinguishing clonality and recombination for microbial populations with large population size and examined with neutral, genetically unlinked markers.

Criteria	Ploidy	Clonality	Recombination
Allelic association	≥Diploidy		
Within a locus			
Hardy–Weinberg equilibrium		No	Yes
Excess homozygosity		Yes	No
Excess heterozygosity		Yes	No
Allelic association	All ploidy		
Between loci			
Random		No	Yes
Non-random		Yes	No
Gene genealogy	All ploidy (but mostly for haploids)		
Congruence		Yes	No
Over-representation of certain multilocus genotype(s)	All ploidy	Yes	No

percent of locus-pairs are expected to have genotypic counts significant different from those expected.

One caveat of this test is the independence of the loci. If loci are linked, they are not totally independent. The strength of linkage between loci could affect the degrees of association among the alleles. It has generally been assumed and mathematically proven, however, that any positive and/or negative allelic association between loci will be broken down if sexual mating and meiosis are frequent enough and if the loci under study are selectively neutral (Hartl and Clark 1989).

When an organism reproduces clonally (through either mitotic division or homothallism), the entire genome is effectively linked since there is no segregation and re-assortment of alleles. Both linkage and clonal reproduction can cause deviations from the expected genotypic frequencies for any pairs of loci. The degree of deviations or non-random association between two loci, each with two alleles, A_1 and A_2, and B_1 and B_2, is measured by D, with:

$$D = pA_1B_1\ pA_2B_2 - pA_1B_2pA_2B_1$$

Where pA_1B_1 represents the observed frequency of genotype A_1B_1, etc.

Another weakness of this test is that with increasing number of loci, many comparisons are performed for the same loci. If there are n polymorphic loci in a population, there would be n(n–1)/2 unique pairwise loci combinations and that many comparisons. Therefore, some pairs may exhibit significant deviation from random mating simply due to chance even if a population is panmictic. On the other hand, even if a population is strictly clonal, some tests might still show random association. This probability increases when sample size decreases and when allele frequencies are highly skewed. To avoid some of these problems in linkage disequilibrium test, another test was introduced to measure the overall allelic association in a sample.

The index of association (I_A)

To provide an overall allelic association in haploid organisms, an index of association (I_A) has been widely used in microbial population genetic analyses. This index was first used by Brown *et al*. (1980) to measure population structure of the plant *Hordeum spontaneum* and was used later by Whittam *et al*. (1983) and Maynard Smith *et al*. (1993) for *Escherichia coli* and other bacteria respectively. I_A is a generalized measure of linkage disequilibrium and has an expected value of zero if there is no association between loci. I_A is calculated as follows.

Suppose M loci have been analyzed for N individuals. Let p_{ij} represent the frequency of the *i*th alleles at the *j*th locus. Then the gene diversity at the *j*th locus, h_j (=1–$\sum p_{ij}^2$), is the probability that two individuals have different alleles at the *j*th locus. Let K represent the number of loci having different alleles between two individuals. The observed variance of K, Vo, can be then calculated from the distribution of K. Since there are N(N–1)/2 possible pairs of individuals, the mean difference between any two individuals, K′, is $\sum h_j$. The expected variance of K is Ve = $\sum h_j(1–h_j)$. The index of association, I_A, is:

$$I_A = Vo/Ve - 1$$

There are two ways to test whether I_A is significantly different from zero, the null hypothesis assuming random association of alleles at different loci.

The first test assumes that the sampling distribution of the error variance of I_A approximates normality. Var(Ve) is calculated as:

$$Var(Ve) = [\sum h_i - 7\sum h_i^2 + 12\sum h_i^3 - 6\sum h_i^4 + 2(\sum h_i - h_i^2)^2]/N$$

With the upper 95% confidence limit for Var(Ve) calculated as:

$$L \approx \sum h_j - \sum h_j^2 + 2[Var(Ve)]^{1/2}$$

If Vo does not exceed L, then the null hypothesis of random association of alleles at all loci is not rejected. The population under study is therefore interpreted to have a structure not significantly different from panmixia. If Vo is greater than L, the hypothesis of a panmictic population structure is rejected and a significant clonal reproduction component is inferred.

The second statistical test of I_A is to use a randomization approach in which the null distribution of Vo is generated by randomly permuting the alleles among all individuals within each locus and calculating Vo many times. The Vo from the observed sample is then compared to the null distribution from permuted samples to determine whether there is a significant difference. If the observed variance Vo is greater than (1 – a) of the Vo from randomized data sets, where a is the acceptable Type I error rate (0.05 or 0.01), then the sample deviates significantly from random mating. Since the sampling distribution of Vo is not known to be normal, randomization tests for significance are preferable, especially when sample size is small.

Gene genealogical comparisons

As shown above, both linkage disequilibrium and association index test against the null hypothesis of random mating. However, testing against a null hypothesis of random mating can create a significant type II error, the probability of accepting a false hypothesis. Type II error can be significant when sample size is small and/or when allele frequencies are highly skewed. Furthermore, the above two tests can't determine whether recombination plays any role in a population known to have a predominantly clonal component.

To overcome these problems associated with testing against the null hypothesis of panmixia, phylogenetic analysis offers a test against the alternative hypothesis of strict clonality. There are two types of phylogenetic tests for clonality and recombination. The first one is called the phylogenetic incompatibility test. The null hypothesis of this test is strict clonality, opposite that of linkage equilibrium. The basic underlying assumptions of the phylogenetic incompatibility test are: (i) that mutation to a specific allele occurs only once in the population history; (ii) that alleles identical in state are identical by descent; and (iii) the existence of phylogenetic incompatibility is evidence for recombination. In the simplest case in a haploid species, assuming two loci (A and B) with two alleles each (A1

and A2; B1 and B2), if all four possible genotypes (A1B1, A1B2, A2B1, and A2B2) are found in the population, these two loci are considered phylogenetically incompatible and must have been resulted from recombination at the population level. However, if only two or three genotypes were found, there would be no clear evidence of phylogenetic incompatibility and no robust evidence of recombination (Hudson and Kaplan 1985). This test can be extended to multiple alleles at each locus.

The second phylogenetic test uses gene sequences from several genes. Clonality can be distinguished from recombination by comparing phylogenetic trees built for different genes. If the trees are congruent, then there is strong evidence for clonality, but if the trees are incongruent, recombination is likely involved. The Partition Homogeneity Test (PHT) has been used to assess the statistical significance of gene genealogy consistencies (e.g. Geiser *et al*., 1998; Xu *et al*., 2000, 2002, 2003). For congruent gene trees, the sum of the lengths of the most parsimonious trees for each gene should not change significantly if the polymorphic nucleotides in each gene are swapped among genes. Contrary, for incongruent gene trees, the sum of the gene trees for the observed data should be shorter than the sum for gene trees made after polymorphic nucleotides have been swapped among genes. This is because recombination is assumed to be correlated with linkage relationships in the genome. Here, intra-genic recombination is assumed to be non-existent or if exists, occurs at a significantly lower frequencies than inter-genic recombination. Statistical significance of this test is established by making many re-sampled data sets and comparing the observed sum of length to the distribution of re-sampled data sets.

Aside from the three common tests used above, there is a fourth but rarely used test. This test compares the observed genotypic diversity with those expected under the null hypothesis of random mating (Stoddart and Taylor, 1988).

Analysis of clonality and recombination in diploid organisms

Hardy–Weinberg equilibrium test

Since there are two alleles at each locus in each diploid strain, tests for recombination in diploid organisms are somewhat different from that in haploid organisms. Specifically, in diploid species, the association of alleles within a locus is quite often used as a measure of recombination. A simple Chi-square goodness of fit test can be performed to compare the observed and expected genotypic counts at each locus and summed across loci (Weir 1996; Xu *et al*., 1997). This is conventionally called the Hardy–Weinberg equilibrium (HWE) test. This test is briefly described below.

Assuming there is a single locus, A, with two alleles, $A_1$and A_2. In a diploid organism, there would be three possible genotypes at this locus, A_1A_1, A_1A_2 and A_2A_2. The expected HWE frequencies of genotypes A_1A_1, A_1A_2 and A_2A_2 are p^2, 2pq and q^2 respectively, where p is the frequency of allele A_1 and q is the frequency of allele A_2, and $A_1 + A_2 = 1$. If the observed counts of genotypes are not significantly different from the expected counts, then the population is consistent with a recombining structure.

Hardy–Weinberg equilibrium is expected when populations meet a number of assumptions: large population size, no selection on the marker loci being analyzed, negligible migration and mutation, and random mating (Hartl and Clark 1989). Therefore, cautions must be made about the weaknesses in inferring population structure from HWE tests. First, the violation of any one of the above assumptions can cause significant deviations between observed and expected genotype frequencies even though the population may in fact be randomly mating. Secondly, failure to reject the null hypothesis of HWE does not guarantee that the population is in fact randomly mating (Type II error).

Composite genotypic equilibrium

Composite genotypic equilibrium test applies to diploids. It is similar to the index of association test applied to haploids described above. In the composite genotypic equilibrium test, the diploid genotypes for each individual locus are fixed with each unique allelic combination treated as a new “allele”. The associations among the new “alleles” at different loci are then calculated. An exact test for composite genotypic equilibrium was developed by Zaykin *et al*. (1995). In this test, the probability of the set of multilocus genotypes in a sample, conditioned on allelic counts, is calculated from the multinomial theory under the hypothesis of no association. Alleles are then permuted and the conditional probability calculated for the permuted genotypic arrays. The proportion of arrays no more probable than the original sample provides the significance level of the test. This test is versatile in the number of loci that can be examined. It also allows the calculation of the probability for individual multilocus genotypes conditioned on genotypic counts at individual locus. This test separates allelic association test within a locus (i.e. HWE test) from genotypic association among loci. Therefore, it is very useful for diploid species.

Analysis of genetic variation between populations

Comparisons of population genetic parameters

Depending on the research questions, there are many ways genetic variation among populations can be compared. One general approach is to compare descriptive parameters calculated for each population. These parameters include gene diversity, heterozygosity, genotypic diversity, relative percentage of polymorphic loci, the mean number of alleles per locus, and the relative importance of recombination and clonality between populations. These parameters could be used to infer a variety of population processes, for example, population histories, mutation rates, environmental conditions, and the selection pressure. The associations between specific environmental factors and genetic elements can be determined through general statistical analyses. The other approach is to directly examine the contributions of population subdivisions on the patterns of genetic variation.

Population subdivision

Population subdivision entails an inbreeding-like effect in terms of excess homozygosity (Wright, 1951; Weir, 1996). Therefore, it is possible to measure this effect in term of the decrease in the proportion of heterozygous genotypes. A subdivided population of a diploid organism has three distinct levels of complexity: individual organisms (I), subpopulations (S), and the total population (T). Let:

- N= number of subpopulations
- R_j= relative size of the jth subpopulation
- P_{ij}= frequency of the ith allele in the jth subpopulation
- P_i= frequency of the ith allele in the total population
- H_{oj}= observed heterozygosity in the jth subpopulation
- H_{ej}= expected heterozygosity in the jth subpopulation ($=1-\sum p_{ij}^2$)
- H_I = average observed heterozygosities over all subpopulation ($=\sum R_j H_{oj}$)
- H_S = average expected heterozygosities over all subpopulation ($=\sum R_j H_{ej}$)
- H_T = expected heterozygosity in the total population ($=1-\sum p_i^2$)

The effects of population subdivision are measured by a quantity called fixation index (symbolized F_{ST} or other comparable measures, see below). F_{ST} represents the reduction in heterozygosity of a subdivided population due to random genetic drift. It is calculated as:

$$F_{ST}= (H_T–H_S)/H_T$$

The greater the F_{ST} values, the greater divergence among the subpopulations. The F_{ST} is typically greater than (or equal to) zero. If all subpopulations are in Hardy–Weinberg equilibrium and with the same allele frequencies, F_{ST}=0.

There are several variations of F_{ST}. These include G_{ST}, R_{ST}, θ, N_{ST}, and Φ_{ST}. These are briefly described below.

- **G_{ST}:** The original F_{ST} assumed a diploid population with only two alleles at a locus (Wright, 1951). Therefore, F_{ST} is not applicable for haploid populations or for populations with more than two alleles at each locus. To accommodate these two issues, Nei (1973) introduced a parameter called the coefficient of gene differentiation and called it G_{ST}. $G_{ST}=(H_T–H_S)/H_T$. H_T represents gene diversity in the total population ($1–\sum p_i^2$) and H_S the average gene diversity among sub-populations.
- **R_{ST}**: R_{ST} is used specifically for microsatellite data (Slatkin, 1995). It differs from F_{ST} and G_{ST} in that it considers allele size and uses the stepwise mutation model to infer allelic relationships. R_{ST} is, in essence, a ratio of the variance of allele sizes (measured as the number of repeat units) among subpopulations to the variance of allele sizes in the total sample.
- **0**: θ is essentially the same as F_{ST} except they differ in their assumptions of the allelic sampling processes. Specifically, θ represents the ratio of the variance of allele frequencies among subpopulations to the overall variance in allele frequencies (Weir and Cockerham 1994).
- **N_{ST}**: N_{ST} is quite different from the above-mentioned parameters and there are several variations of N_{ST}. The most commonly used form defines N_{ST} based on nucleotide differences among haplotypes (Lynch and Crease 1990). Briefly, $N_{ST}= V_b/(V_w + V_b)$, where V_b is the average proportion of nucleotide substitutions between subpopulations and V_w is the average proportion of nucleotide substitutions within subpopulations.
- **Φ_{ST}**: Φ_{ST} is a variation of θ but it incorporates the generalized analysis of variance (ANOVA) approach (Excoffier et al., 1992). As a result, it has several advantages associated with ANOVA. It can be used to analyze many different types of molecular data such as RFLP and DNA sequences and incorporate the distances among the alleles. This approach is now commonly called AMOVA, short for the Analysis of MOlecular VAriance.

Gene flow

There are several demographic models of gene flow: continent-island model, island model, one-dimensional stepping stone model, and two-dimensional stepping stone model. Estimates of gene flow among subpopulations differ depending on the demographic models as well as other factors such as the nature of the genetic marker, their mutation rates and the breeding system of the organism under consideration.

Gene flow in animals and plants may be estimated through both direct (observational) and indirect (population genetic) means. For fungal and other microbial populations, direct estimate is usually not possible and the estimated parameters of population subdivision are often used to infer gene flow among subpopulations. The calculations of gene flow based on F_{ST} and other similar parameters are different between haploid and diploid organisms. This is because each migrant individual carries two alleles per locus in diploid organisms but only one allele per locus in haploid organisms. Typically, the number of migrants per generation, Nm, equals $(1–F_{ST})/4F_{ST}$ for diploids, while that for haploids equals $(1–F_{ST})/2F_{ST}$ for haploids (Cockerham and Weir, 1993). N is the average number of individuals in a subpopulation, and m the migration rate between pairs of subpopulations. F_{ST} can be substituted by other parameters described in the previous section.

It must be cautioned that the estimation of gene flow based on F_{ST} values entails many assumptions (Cockerham and Weir, 1993; Weir, 1996). These assumptions includes: low mutation rates, stable environments, no selection on the marker(s) under investigation, large population sizes, and genetic equilibrium within each subpopulation. If these assumptions are not met, Nm could be highly biased and may not reflect actual gene flow between subpopulations.

The statistical test for whether pairs of populations are significantly different can also be achieved through direct comparison of gene frequencies (Hudson *et al.*, 1992). This is done through either Chi-square contingency table tests or Fisher's exact test when sample sizes are small or when the lowest expected allele count is low (e.g.

less than five). Tests of this kind can be found in general biostatistic and epidemiological books.

Population genetic relationships and genetic isolation by geographical distance

The genetic relationships among subpopulations can be described in dendrograms. These dendrograms are obtained based on pairwise population genetic distances. There are several measures of population genetic distances, all based on population gene frequencies and inbreeding coefficients (Nei, 1975). The most widely used is Nei's genetic distance D (Nei, 1973). The basis of this measure is a normalized identity. This identity represents the probability that a randomly chosen allele from two different subpopulations will be identical, relative to the probability that the two randomly chosen alleles from the same subpopulation will be identical. The calculations are shown as follows.

Suppose there are two hypothetical subpopulations, A and B. At locus X, the frequency of allele i in subpopulation A is p_i and in subpopulation B is qi. Let J_{AA} represent the probability that two alleles chosen at random from subpopulation A are identical, then:

$$J_{AA} = \sum p_i^2$$

This is because with random mating, J_{AA} equals the homozygosity in subpopulation A. Likewise, J_{BB} is defined as the probability that the two alleles chosen at random from subpopulation B are identical. J_{BB} equals $\sum q_i^2$.

Next, we define J_{AB} as the probability that two alleles are identical when one allele is chosen from subpopulation A and the other is chosen from subpopulation B, we have:

$$J_{AB} = \sum p_i q_i$$

Nei (1973, 1975) defines the normalized identity, I, as:

$$I = J_{AB}/(J_{AA}\, J_{BB})^{1/2}$$

And the standard genetic distance, D, as D = –ln(I). D ranges from 0 to infinity. This measure of genetic distance is more reliable when data from many genes are available. With many genes, the quantities J_{AA}, J_{BB}, and J_{AB} are the arithmetic means of the values from individual loci. If two populations are identical, then J_{AA}= J_{BB}= J_{AB}, and the normalized identity is 1. Lack of genetic divergence between populations results in a distance of 0. Since the statistic is calculated from samples from populations, the exact value of D between populations may vary from one sample to another. The degree to which different samples produce different D values depends on the amount of genetic divergence, the number of genes, and the sample sizes. Reynolds *et al*. (1983) derived modified genetic distance measures using ANOVA.

Once pairwise population genetic distances are calculated, population similarities can be described on a dendrogram. Genetic distances are often used in comparison with geographic distances. Assuming roughly constant rate of migration, one might expect that pairs of populations that are situated farther apart geographically would show greater genetic divergence. This is known as isolation by distance (IBD; Wright, 1943).

Conclusions and perspectives

This paper described modern molecular techniques and the common analytical methods for studying fungal populations. The clear trend in the application and development of molecular techniques is DNA sequencing and large-scale analysis of single nucleotide polymorphisms using oligonucleotide microarrays. On the analytical methods front, there is an urgent need for method development to clearly identify underlying mechanisms responsible for the origin and maintenance of genetic variation both within a population as well as between populations. Another area of development is the integration of discrete population genetic data with quantitative trait data and various environmental and ecological parameters for fungal populations. The successful developments in these and other areas should bring fungal molecular population genetics to mainstream evolutionary genetics.

Acknowledgments

Research in my laboratory is provided by the Natural Science and Engineering Research Council (NSERC) of Canada, the Ontario Premier's Research Excellence Award (PREA), Genome Canada (GC), the Canadian Foundation for Innovation (CFI), and the Ontario Innovation Trust (OIT).

References

Archie, J.W. (1989). A randomization test for phylogenetic information in systematic data. Syst. Zool. *38*, 239–252.

Bell G. (1982). The Masterpiece of Nature. San Francisco: University of California Press. 1982.

Brown, A.H.D., Feldman, M.W., and Nevo, E. (1980). Multilocus structure of natural populations of *Hordeum spontaneum*. Genetics *96*, 523–536.

Cockerham, C.C., and Weir, B.S. (1993). Estimation of gene flow from F-statistics. Evolution *47*, 855–863.

Cowen, L.E., Sirjusingh, C., Summerbell, R.C., Walmsley, S., Richardson, S., Kohn, L.M., and Anderson, J.B. (1999). Multilocus genotypes and DNA fingerprints do not predict variation in azole resistance among clinical isolates of *Candida albicans*. Antimicrob. Agents Chemother. *43*, 2930–2938.

Excoffier, L., Smouse, P.E., and Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics *131*, 479–491.

Field, D., Eggert, L., Metzgar, D., Rose, R., and Wills, C. (1996). Use of polymorphic short and clustered coding-region microsatellites to distinguish strains of *Candida albicans*. FEMS Immunol. Med. Microbiol. *15*, 73–79.

Fries, B.C., Chen, F.Y., Currie, B.P., and Casadevall, A. (1996). Karyotype instability in *Cryptococcus neoformans* infection. J. Clin. Microbiol. *34*, 1531–1534.

Geiser, D.M., Pitt, J.I., and Taylor, J.W. (1998). Cryptic speciation and recombination in the aflatoxin producing fungus *Aspergillus flavus*. Proc. Natl. Acad. Sci. USA *95*, 388–393.

Hartl, D.L., and Clark, A.G. (1989). Principles of Population Genetics. Sinauer Associates, Sunderland, MA.

Hauser, P.M., Francioli, P., Bille, J., Telenti, A., and Blanc, D.S. (1997). Typing of *Pneumocystis carinii* f. sp. *hominis* by single-strand conformation polymorphism of four genomic regions. J. Clin. Microbiol. *35*, 3086–3091.

Hawkesworth, D.L., Kirk, P.M., Sutton, B.C., and Pegler, D.N. (1995). Ainsworth and Bisby's Dictionary of the Fungi. 8th ed. International Mycological Institute, Surrey, England.

Hudson, R.R., and Kaplan, N.L. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics *111*,147–164.

Hudson, R.R., Boos, D.D., and Kaplan, N.L. (1992). A statistical test for detecting geographic subdivision. Mol. Biol. Evol. *9*, 138–151.

Kurtzman, C.P. (1993). DNA–DNA hybridization approaches to species identification in small genome organisms. Meth. Enzymol. *224*, 335–348.

Lynch, M., and Crease, T.J. (1990) The analysis of population survey data on DNA sequence variation. Mol. Biol. Evol. *7*, 377–394.

Maynard Smith J. (1978). The Evolution of Sex. Cambridge University Press.

Maynard Smith, J., Smith, N.H., O'Rourke, M., and Spratt, B.G. (1993). How clonal are bacteria? Proc. Natl. Acad. Sci. USA *90*, 4384–4388.

Michod, R.E., and Levin, B.R. (1988). The Evolution of Sex: an examination of current ideas. MA: Sinauer Associates, Inc.

Mish, F. C. (Editor-in-Chief) (1996). Merriam Webster's Collegiate Dictionary (10th Edition). MA: Merriam-Webster, Inc.

Murphy, R.W., Sites, J.W., Jr., Buth, D.G., and Haufler, C.H. (1996). Proteins: isozyme electrophoresis. In: Hillis DM, Moritz C, Mable BK, editors. Molecular Systematics. Sunderland, Massachusetts: Sinauer Associates, pp.51–120.

Nei, M. (1973) Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA. *70*, 3321–3323.

Nei, M. (1975). Molecular Population Genetics and Evolution. New York: American Elsevier.

Olicio, R., Almeida, C.A., and Seuanez, H.N. (1999). A rapid method for detecting and distinguishing clinically important yeasts by heteroduplex mobility assays (HMAs). Mol. Cell Probes *13*, 251–255.

Reynolds, J., Weir, B.S., and Cockerham, C.C. (1983). Estimating the coancestry coefficient: basis for a short-term genetic distance. Genetics *105*, 767–779.

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. Genetics *139*, 457–462.

Sokal, R. R., and F. J. Rohlf. (1981). Biometry, 2nd ed., W. H. Freeman, New York.

Stoddart, J.A., and Taylor, J.F. (1988). Genotypic diversity: estimation and prediction in samples. Genetics *118*, 705–711.

Suzuki, T., Kobayashi, I., Mizuguchi, I., Banno, I., and Tanaka, K. (1988). Electrophoretic karyotypes in medically important *Candida* species. J. Gen. Microbiol. *34*, 409–416.

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M Frijters A., Pot, J., Peleman, J., Kulper, M., and Zabeau, M. (1995). AFLP: a new technique for DNA fingerprinting. Nucl. Acids Res. *23*, 4407–4414.

Weir, B.S. (1996). Genetic Data Analysis, 2nd ed. Sinauer Associates, Sunderland, MA.

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 1358–1370.

Whittam, T.S., Ochman, H., and Selander, R.K. (1983). Multilocus genetic structure in natural populations of *Escherichia coli*. Proc. Natl. Acad. Sci. USA *80*, 1751–1755.

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A., and Tingey, S.V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucl. Acids Res. *18*, 6531–6535.

Winzeler, E.A., Richards, D.R., Conway, A.R., Goldstein, A.L., Kalman, S., McCullough, M.J., McCusker, J.H., Stevens, D.A., Wodicka, L., Lockhart, D.J., and Davis, R.W. (1998) Direct allelic variation scanning of the yeast genome. Science *281*, 1194–1197.

Winzeler, E.A., Castillo-Davis, C.I., Oshiro, G., Liang, D., Richards, D.R., Zhou, Y., and Hartl, D.L.. (2003). Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. Genetics. *163*, 79–89.

Wright, S. (1943). Isolation by distance. Genetics *28*, 114–138.

Wright, S. (1951). The genetical structure of populations. Ann. Eugen. *15*, 323–354.

Xu, J. (2002). Mitochondrial DNA polymorphisms in the human pathogenic fungus *Cryptococcus neoformans*. Curr. Genet. *41*, 43–47.

Xu, J. (2004). The prevalence and evolution of sex in microorganisms. Genome *47*, 775–780.

Xu, J., Boyd, C.M., Livingstone, E., Meyer, W., Madden, J.F., and Mitchell, T.G. (1999a). Species and genotypic diversities and similarities of pathogenic yeasts colonizing women. J. Clin. Microbiol. *37*, 3835–3843.

Xu, J., Kerrigan, R. W., Callac, P., Horgen, P.A., and Anderson, J.B. (1997). The genetic structure of natural populations of *Agaricus bisporus*, the commercial mushroom. J. Heredity *88*, 482–494.

Xu, J., Kerrigan, R.W., Sonnenberg, A.S., Callac, P., Horgen, P.A., and Anderson, J.B. (1998). Mitochondrial DNA variation in natural populations of the mushroom *Agaricus bisporus*. Mol. Ecol. *7*, 19–33.

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.

Xu, J. and Mitchell, T.G. (2003a). Population genetic analysis of medical fungi. Chapter 13. Pp. 703–722. In Fungi Pathogenic to Human and Animals (2nd edition).

Editor Dexter H. Howard. Marcel Dekker, Inc. New York.

Xu, J., and Mitchell, T.G. (2003b). 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., Mitchell, T.G., and Vilgalys, R.J. (1999b). PCR-restriction fragment length polymorphism (RFLP) analyses reveal both extensive clonality and local genetic differences in *Candida albicans*. Mol. Ecol. *8*, 59–73.

Xu, J., Ramos, A., Vilgalys, R., and Mitchell, T.G. (2000a). Clonal and spontaneous origins of fluconazole resistance in *Candida albicans*. J. Clin. Microbiol. *38*, 1214–1220.

Xu, J., Vilgalys, R.J., and Mitchell, T.G. (2000b). Multiple gene genealogies reveal recent dispersion and hybridization in the human pathogenic fungus *Cryptococcus neoformans*. Mol. Ecol. *9*, 1471–1481.

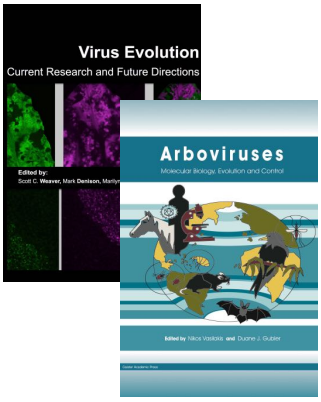
Zaykin, D., Zhivotovsky, L., and Weir, B.S. (1995). Exact tests for association between alleles at arbitrary numbers of loci. Genetica *96*, 169–178.

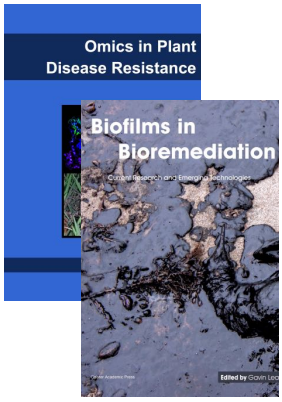
Zolan, M.E. (1995). Chromosome-length polymorphism in fungi. Microbiol. Rev. *59*, 686–698.

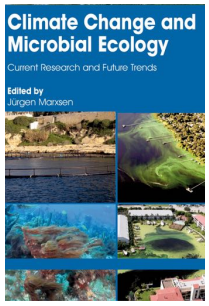
Further Reading

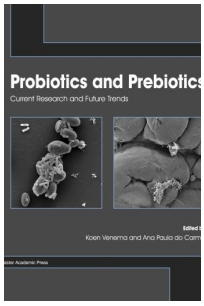
Caister Academic Press is a leading academic publisher of advanced texts in microbiology, molecular biology and medical research. Full details of all our publications at **caister.com**

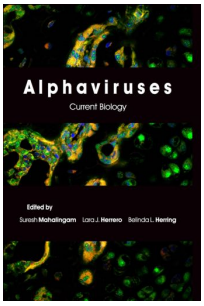
- **MALDI-TOF Mass Spectrometry in Microbiology**
  Edited by: M Kostrzewa, S Schubert (2016)
  www.caister.com/malditof
- ***Aspergillus* and *Penicillium* in the Post-genomic Era**
  Edited by: RP Vries, IB Gelber, MR Andersen (2016)
  www.caister.com/aspergillus2
- **The Bacteriocins: Current Knowledge 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

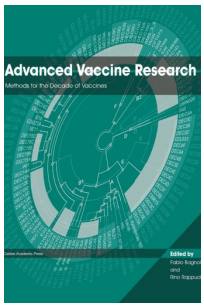


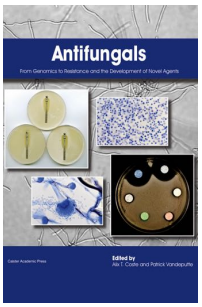












- **Flow Cytometry in Microbiology: Technology and Applications**
  Edited by: MG Wilkinson (2015)
  www.caister.com/flow
- **Probiotics and Prebiotics: Current Research and Future Trends**
  Edited by: K Venema, AP Carmo (2015)
  www.caister.com/probiotics
- **Epigenetics: Current Research and Emerging 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