

Article

"Applying comparative genomics to plant disease epidemiology"

Linda M. Kohn

Phytoprotection, vol. 85, n° 1, 2004, p. 45-48.

Pour citer cet article, utiliser l'adresse suivante :

<http://id.erudit.org/iderudit/008906ar>

Ce document est protégé par la loi sur le droit d'auteur. L'utilisation des services d'Érudit (y compris la reproduction) est assujettie à sa politique d'utilisation que vous pouvez consulter à l'URI <http://www.erudit.org/documentation/eruditPolitiqueUtilisation.pdf>

Érudit est un consortium interuniversitaire sans but lucratif composé de l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal. Il a pour mission la promotion et la valorisation de la recherche. Érudit offre des services d'édition numérique de documents scientifiques depuis 1998.

Pour communiquer avec les responsables d'Érudit : erudit@umontreal.ca

Applying comparative genomics to plant disease epidemiology

Linda M. Kohn¹

PHYTOPROTECTION 85 : 45-48

Phylogenetic or genealogical interpretation of DNA sequence data from multiple genomic regions has become the gold standard for species delimitation and population genetics. Precise species concepts can inform quarantine decisions but are likely to reflect evolutionary events too far in the past to impact disease management. On the other hand, multilocus approaches at the population level can identify patterns of endemism or migration directly associated with episodes of disease, including host shifts and associated changes in determinants of pathogenicity and avirulence. We used the genome database of *Magnaporthe grisea* to frame a comparative, multilocus genomics approach from which we demonstrate a single origin for rice infecting genotypes with concomitant loss of sex in pandemic clonal lineages, and patterns of gain and loss of avirulence genes. In the *Sclerotinia sclerotiorum* pathosystem, we identified significant associations of multilocus haplotypes with specific pathogen populations in North America. Following the introduction of a new crop, endemic pathogen genotypes and newly evolved migrant genotypes caused novel, early-season symptoms.

[Application de la génomique comparative à l'épidémiologie des maladies des plantes]

L'interprétation phylogénétique ou généalogique des données sur les séquences d'ADN provenant de plusieurs régions des génomes est devenue la méthode de choix pour l'établissement des limites des espèces et en génétique des populations. Des concepts d'espèces bien définies peuvent aider à prendre des décisions en matière de quarantaine, mais ils sont probablement le reflet d'événements évolutifs qui se sont produits voilà trop longtemps pour avoir un impact sur la lutte contre les maladies. Par contre, une approche multilocus effectuée au niveau des populations peut permettre l'identification de schémas d'endémisme ou de migration en lien direct avec des épisodes de maladie, y compris des changements d'hôte et des modifications associées aux facteurs déterminants de la pathogénicité et de l'avirulence. Nous avons utilisé la base de données sur le génome du *Magnaporthe grisea* comme base d'une approche génomique comparative et multilocus qui nous a permis de démontrer une origine unique pour les génotypes qui infectent le riz avec la perte concomitante de la sexualité chez les lignées clonales pandémiques ainsi que des schémas de gain et de perte de gènes avirulents. Dans le pathosystème du *Sclerotinia sclerotiorum*, nous avons identifié d'importantes associations entre des haplotypes multilocus et des populations pathogènes en Amérique du Nord. À la suite de l'introduction d'une nouvelle culture, les génotypes pathogènes endémiques et les génotypes migrants d'évolution récente ont été la cause de symptômes nouveaux qui se sont produits tôt en saison.

Taken broadly, the term "comparative genomics" usually conveys parallel study of genome structure, dynamics, and function among different individuals or groups of individuals (**taxa**) at the different ranks in the taxonomic hierarchy, from species to kingdom. Genome comparisons at different **phylogenetic distances** answer different sorts of questions (Hardison 2003). The reader is directed to Figure 1 in Hardison at <http://www.plosbiology.org/plosonline/?request=index-html>. Short distances are defined by a small number of steps, such as nucleotide substitutions or changes in microsatellite repeats, often associated with recent evolution on a contemporary timescale. In a branching phylogenetic tree, these patterns are usually represented as the **tip branches** (branches are also termed **clades**). In contrast, long distances are defined by a large number of steps, associated either

with more ancient ancestry or with accelerated evolutionary rates. In a phylogeny, these are often represented on the **internal branches**.

DNA sequence-based studies of single or multiple genomic regions can delimit species or the patterns of gene flow that define populations of a species. While such studies seek genomic differences, comparative genomics often targets similarities among genomes to identify the genes that are relatively invariant, i.e. conserved, because of their significant function. Mutational variants of genes that do not improve the fitness of the individual tend to be lost through natural selection. Mutations that change the function or protein products of genes that have important functions tend to be weeded out through a kind of natural selection termed "purifying" selec-

1. Department of Botany, University of Toronto at Mississauga, Mississauga, Ontario L5L 1C6; e-mail: kohn@utm.utoronto.ca

tion, with the result that the DNA sequences of these genes tend to be conserved rather than highly variable. On a contemporary evolutionary scale, mating type genes are conserved within species and in some fungi, such as the *Fusarium graminearum* complex, among closely related species (O'Donnell *et al.* 2004). On a more ancient scale, comparative genomics seeks the genes for core proteins common to all organisms, such as MAP kinases in protein-protein signaling or ABC transporters as efflux pumps.

All of these aspects of comparative genomics can contribute to plant disease epidemiology. Accurately defined species are key for quarantine strategies, as well as for patenting of biocontrol agents and biotechnological products or processes. Knowing the closest relatives of a species predicts biological characteristics, most notably of fungi known only in their asexual state that lack the necessary morphological features for taxonomic diagnosis among fungi with known sexual states. A good example is *Sclerotium cepivorum*, cause of white rot of onion and garlic, which has no known sexual state. Species of the form genus *Sclerotium* produce only mycelium and sclerotia. Some are asexual Basidiomycetes, e.g. *S. rolfsii*, and some are Ascomycetes, e.g. *S. cepivorum*. Based on phylogenies of seven genes or genomic regions, there is substantial evidence that *S. cepivorum* is closely related to the genus *Sclerotinia* (Carbone and Kohn 1993; Carbone 2000; Holst-Jensen *et al.* 2004). Control strategies proven to be effective for sexual relatives have potential for such asexual orphan species. As more fungal plant pathogens are fully sequenced methods in comparative genomics extended to closely related species will fully exploit the resource information on gene regulation of pathogen development, pathogenesis, and mechanisms of virulence (Yarden *et al.* 2003).

Recently, we have traced the distribution of avirulence (AVR) genes in genotypes of the rice blast agent, *Magnaporthe oryzae* (of the *M. grisea* species complex), against an evolutionary framework that provides information on geographical and host distribution (Couch 2004). With powerful statistical approaches, such as coalescence, Bayesian, and statistical parsimony, comparative genomic data can be explored to address basic questions on the origins, movement and amplification of epidemiologically significant genotypes (Carbone and Kohn 2001, 2004; Milgroom and Peever 2003; Phillips *et al.* 2002).

Phylogenetic or genealogical interpretation of DNA sequence data from multiple genomic regions has become the gold standard for species delimitation and population genetics. The basic unit of analysis is the **haplotype**. All members of a haplotype share the same variable nucleotide sites in a sequence, as well as all of the invariant sites. A species defined by the multilocus standard is a clade based on concordant DNA sequences from multiple genes or genomic regions, termed **loci**, corroborated if possible by proof of interfertility among individuals within the clade and lack of fertility with individuals from other clades. The genomic part of this approach has been called the genealogical concordance phylogenetic species recognition (GCPSR) concept (Taylor *et al.* 2000; O'Donnell *et al.* 2004). In the absence of the interfertility crite-

riion, the GCPSR concept may not distinguish population divergence from speciation, e.g. individuals in the same species but in two evolutionarily diverging populations would be expected to be able to mate but could easily be represented in a phylogeny as two distinct tip clades (Carbone and Kohn 2001). Species and populations within species can be distinguished by sampling of more than one population within each putative species, then screening sequences of multiple loci. The screen should include some loci that are highly polymorphic, variable within population samples of a single species, i.e. at the contemporary time scale, as well as loci that are invariant within populations, but variable among species, i.e. at the more ancient time scale. With this type of sample and dataset, species and divergent populations can be distinguished, even in the absence of interfertility tests. This approach also has the advantage of removing sampling bias when only a few isolates are studied for each species pre-determined to be of interest (Hare 2001). For rapid throughput of large samples, we initially sequence a small set of isolates (usually 12) that are likely to vary. Once polymorphic sites are identified, samples are screened for nucleotide variation by means of single-strand conformation polymorphisms (SSCPs; Carbone *et al.* 1999).

Comprehensive multi-locus sequencing of large samples reveals several important aspects of molecular evolution in fungi. In recent studies in our laboratory, we screened 383 isolates of *Sclerotinia sclerotiorum* and related species (Carbone and Kohn 2001) and 497 isolates of *Magnaporthe oryzae* (Couch 2004) plus isolates of additional species (Couch and Kohn 2002). First, in both studies, it is evident that one gene or genomic region may be invariant at the population level but highly polymorphic among isolates representing different species, e.g., the MPG1 gene in the *M. grisea* complex with 12 polymorphic sites among population samples of *M. oryzae*, which is associated with rice and other grasses excluding *Digitaria* (crabgrass), and 49 polymorphic sites in the species complex that includes *M. oryzae*, and what is now recognized as *M. grisea*, which is associated only with crabgrass (Couch 2004; Couch and Kohn 2002). This feature is also evident in *Sclerotinia*, in which some loci yield many haplotypes and branches within species but short branches separating species, while other loci yield only one or two haplotypes within species but long branches separating species. For example, the intergenic spacer of the nuclear rDNA repeat (IGS), with 55 informative sites yields 47 haplotypes within *S. sclerotiorum*, offers solid evidence of population divergence within the species. In species other than *S. sclerotiorum*, the IGS was present in more than one form as highly divergent paralogues. Consequently for this level of comparison the promoter region of the IGS was used, yielding 18 haplotypes and distinguishing all species of *Sclerotinia* on relatively long branches. Louisiana and Ontario populations of a *Trillium* parasite could not be distinguished but this taxon and *Dumontinia* and *S. cepivorum* formed a clade. In contrast, the Actin gene with 14 haplotypes based on 44 informative sites, yielded only one haplotype, based on one informative site, in *S. sclerotiorum*. However at species level it clearly distinguishes on long branches all four species of *Sclerotinia* (*S.*

minor, *S. sclerotiorum*, *S. trifoliorum*, and one new species from Norwegian potatoes, wild *Caltha palustris* (marsh marigold) and *Taraxacum* (dandelion). In the Actin trees, the *Dumontinia* clade, including *S. cepivorum*, was marked by shorter branches (i.e. fewer differences as evidenced by nucleotide substitutions) than in the IGS tree.

A second aspect of molecular evolution is that evolutionary rates may vary among loci, as shown by different branch lengths in phylogenies for one sample of isolates in the Sclerotiniaceae. This point is well illustrated in comparing branch lengths among species within *Sclerotinia* and among species representing other genera of the Sclerotiniaceae for the IGS and Actin loci, as well as the calmodulin, RAS protein, chitin synthase, and the translation elongation factor-1 alpha (TEF or EF1-alpha) (Carbone 2000; Carbone and Kohn 2001; Holst-Jensen *et al.* 2004). Basically, each locus may tell a different evolutionary story, as in the Sclerotiniaceae, which is why a multi-locus approach is important to approximate the true evolutionary tree. In contrast, in the *M. grisea* complex, there is remarkable concordance among three loci, both with respect to rates and topology (Couch and Kohn 2002).

Thirdly, the rate of molecular evolution may exceed that of morphological evolution. *M. oryzae* and *M. grisea*, two distinct biological species, are clearly demarcated phylogenetic species based on data from several loci, but present no detectable morphological difference (Couch and Kohn 2002). Despite the rather long branches separating species, most of the species of the Sclerotiniaceae produce sexual states with remarkably similar morphology.

A fourth characteristic of molecular evolution, albeit with certain caveats, is that based on multi-locus data in a contemporary sample, molecular evolution reflects pathogen evolution. Within our phylogeny of *M. oryzae* haplotypes, we detected evidence for one host jump from millet (*Setaria*) to rice with a corresponding increase in copies of a transposable element in rice haplotypes (Couch 2004), consistent with expectations for asexual populations in the absence of meiosis (when phenomena, such as repeat induced point mutation or "ripping", can eliminate copies). As no evidence of recombination was detected among rice haplotypes, the indication is that rice populations are indeed asexual. Interestingly, jumps from rice to weeds of rice were observed. These weed haplotypes had few copies of the transposable element. They were weakly or non-pathogenic on rice, consistent with high host specificity associated with clades, as evidenced in co-inoculation experiments in collaboration with B. Valent (Kansas State University) and Didier Tharreau (Montpellier, France).

Last in this list of intriguing molecular evolutionary phenomena is phylogenetic conflict, indicative of mutational hotspots or of recombination. Rather than viewing conflict with dread as an obstacle to inferring a well-supported phylogeny, conflict within or between loci should be investigated with enthusiasm, especially at the level of divergent populations within a species. At higher taxonomic rank, problems with inferring a phylogeny can also arise as a result of

hybridization or gene duplication events and can be difficult to sort out. In the context of epidemiology and disease management, recombination in pathogen populations can be a quick route to building a super pathogen by combining phenotypic traits, such as virulence, fungicide resistance, or enhanced tolerance or response to environment in one genotype. On the other hand, in the absence of conflict due to recombination, clonality through asexual reproduction or self-fertilization (in a haploid organism) is often associated with epidemic spread. Conflict can be detected by a lack of resolution in phylogenies (a comb- or star-like structure for the whole phylogeny, or a fan in one or more clades), and localized to specific loci or parts of a phylogeny by a variety of methods, including compatibility matrices, partitioned Bremer support, or inference of networks rather than branching phylogenies (Carbone *et al.* 1999; Carbone and Kohn 2001; Posada and Crandall 2001; Posada *et al.* 2002). Recombination "hot-spots", blocks of sequence that recombine with each other (with comparatively less or no recombination within a block) are an object of interest in current studies of human evolution (Pennisi 2004); recombination blocks which may or may not currently be hotspots are also found in fungi, as we have observed in *M. oryzae* and *S. sclerotiorum* (Carbone and Kohn 2001; Couch 2004). Such recombination blocks are evident in a compatibility matrix as large blocks of incompatibility within a locus and are most readily observed in large, contiguous sequences, such as the 4 Kb span of the intergenic spacer region of the nuclear rDNA repeat in *S. sclerotiorum*.

So what is next? As of July 1, 2004, *Sclerotinia sclerotiorum* will be sequenced at the Broad Institute, Massachusetts Institute of Technology, financed by the USDA as part of the NSF-USDA Microbial Genome Sequencing Program (M. Dickman, L.M. Kohn, J. Rollins). Like the *Magnaporthe grisea* Genome project (<http://www.broad.mit.edu/annotation/fungi/magnaporthe/>), an invaluable resource for our marker development, the *Sclerotinia sclerotiorum* genome will be publicly available.

ACKNOWLEDGMENTS

This work is supported by the Natural Sciences and Engineering Research Council of Canada, the United States Department of Agriculture, and the North Central Research Program for research on white mold of soybean.

REFERENCES

- Carbone, I. 2000. Population history and process: nested clade and coalescent analysis of multiple gene genealogies in a parasite of agricultural and wild plants. Ph.D. Thesis, University of Toronto, Toronto. 272 pp.
- Carbone, I., and L.M. Kohn. 1993. Ribosomal DNA sequence divergence within ITS 1 of the Sclerotiniaceae. *Mycologia* 85 : 415-427.
- Carbone, I., and L.M. Kohn. 2001. A microbial population-species interface: nested cladistic and coalescent inference with multilocus data. *Mol. Ecol.* 10 : 947-964.

- Carbone, I., and L.M. Kohn. 2004.** Inferring process from pattern in fungal population genetics. Pages 29-58 *in* D.K. Arora and G.G. Khachatourians (eds.), *Fungal Genomics, Applied Mycology and Biotechnology Series*, Vol. 4. Elsevier Science.
- Carbone, I., J.B. Anderson, et al. 1999.** Patterns of descent in clonal lineages and their multilocus fingerprints are resolved with combined gene genealogies. *Evolution* 53 : 11-21.
- Couch, B.C. 2004.** The origin of rice-infecting populations of *Magnaporthe oryzae*. Ph.D. Thesis, University of Toronto, Toronto, 215 p.
- Couch, B. C., and L.M. Kohn. 2002.** A multilocus gene genealogy concordant with host preference indicates segregation of a new species, *Magnaporthe oryzae*, from *M-grisea*. *Mycologia* 94 : 683-693.
- Hardison, R. 2003.** Comparative genomics. *PLoS Biol.* 1 : 156-160.
- Hare M.P. 2001.** Prospects for nuclear gene phylogeography. *Trends Ecol. Evol.* 16 : 700-706.
- Holst-Jensen, A., T. Vralstad, and T. Schumacher. 2004.** *Kohninia linnaeicola*, a new genus and species of the Sclerotiniaceae pathogenic to *Linnaea borealis*. *Mycologia* 96 : 135-142.
- Milgroom, M.G., and T.L. Peever. 2003.** Population biology of plant pathogens -The synthesis of plant disease epidemiology and population genetics. *Plant Dis.* 87 : 608-617.
- O'Donnell K.O., T.J. Ward, D.M. Geiser, H.C. Kistler, and T. Aoki. 2004.** Genealogical concordance between the mating type locus and seven other nuclear genes supports formal recognition of nine phylogenetically distinct species within the *Fusarium graminearum* clade. *Fungal Genet. Biol.* 41 : 600-623.
- Pennisi, E. 2004.** The biology of genomes meeting: The case of the disappearing DNA hotspots. *Science* 304 : 1590.
- Phillips, D.V., I. Carbone, et al. 2002.** Phylogeography and genotype-symptom associations in early and late season infections of canola by *Sclerotinia sclerotiorum*. *Phytopathology* 92 : 785-793.
- Posada D., and K.A. Crandall. 2001.** Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol. Evol.* 16 : 37-45.
- Posada, D., K.A. Crandall, and E.C. Holmes. 2002.** Recombination in evolutionary genomics. *Annu. Rev. Genet.* 36 : 75-97
- Taylor, J.W., D.J. Jacobson, S. Kroken, T. Kasuga, D.M. Geiser, D.S. Hibbett, and M.C. Fisher. 2000.** Phylogenetic species recognition and species concepts in Fungi. *Fungal Genet. Biol.* 31 : 21-31.
- Yarden, O., D.J. Ebbole, S. Freeman, R.J. Rodriguez, and M.B. Dickman. 2003.** Fungal biology and agriculture: Revisiting the field. *Mol. Plant-Microbe Interact.* 16 : 859-866.