UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Analysis of the neutral and adaptive genetic variation of *Colletotrichum kahawae* and its relationship with the *C. gloeosporioides* complex

Diogo Nuno Proença Rico Silva

Mestrado em Biologia Humana e Ambiente

2010

UNIVERSIDADE DE LISBOA

FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Analysis of the neutral and adaptive genetic variation of *Colletotrichum kahawae* and its relationship with the *C. gloeosporioides* complex

Diogo Nuno Proença Rico Silva

Dissertação orientada por:

Prof. Doutor Octávio Fernando de Sousa Salgueiro Godinho Paulo
Doutora Dora Cristina Vicente Batista Lyon de Castro

Mestrado em Biologia Humana e Ambiente

2010

Nota prévia

A escrita desta tese de mestrado encontra-se na língua Inglesa uma vez que esta é a língua científica universal. Por esta razão, o conhecimento e treino da sua escrita e gramática revestem-se de uma importância acrescida para quem tenciona seguir uma carreira em investigação científica em Biologia. A escrita da presente tese nesta língua representa assim um exercício apropriado que poder-se-á revelar proveitoso no futuro.

No decorrer deste mestrado foram reunidas as condições para a escrita de artigos científicos em revistas internacionais, razão pela qual a presente tese foi escrita sob a forma de duas publicações científicas de acordo com o Artº 4 das Regras e Recomendações Para a Elaboração de Dissertações de 2º Ciclo (Mestrado) do Departamento de Biologia Animal. Desta forma visa-se acelerar o processo de elaboração dos manuscritos e suas subsequentes publicações. Cada um dos manuscritos foi escrito de acordo com as instruções para autores das respectivas revistas científicas a que se pretende submeter. Especificamente, o 1º artigo segue as directrizes da revista "*Fungal Diversity*" e o 2º artigo segue as directrizes da revista "*Molecular Ecology*". No entanto, para facilitar a leitura, e seguindo as normas do Artº 4 previamente referido, as figuras e tabelas foram incluídas ao longo do texto.

As referências bibliográficas da Introdução Geral foram elaboradas segundo os parâmetros da revista científica internacional, *Trends in Ecology and Evolution*. Esta é uma das revistas mais relevantes na área em que esta tese foi desenvolvida e possui um sistema de citações cómodo para a leitura de textos de revisão científica. Adicionando o seu elevado factor de impacto na sociedade científica, pareceu apropriada a escolha desta revista como referência para a apresentação da bibliografia.

Acknowledgements

To my advisors, Dr Dora Batista and Prof. Octávio Paulo, I give a special thanks for all the confidence deposited in me and for their guidance in so many aspects of my training to become a better scientist. In particularly, I thank Drª Dora for passing on her enthusiastic and communicative way of doing science, and Profº Octávio for further stimulating my inquisitive way of treating biology and the scientific though.

I also deeply appreciate Dr Pedro Talhinhas for his friendship and guidance throughout the mycological world and constant availability to contribute and improve this work. To Dr Andreia Loureiro I give my thanks for her friendship, cheering, and for all the help and guidance during my early training in several laboratory procedures. To Engº Vitor Várzea, I sincerely thank his friendship and guidance and appreciate for passing on his vast knowledge of CBD and experience in the field and with the people that truly suffer from this disease. To Sandra Sousa Emídio, I am thankful for all the help and training in several technical mycological procedures and constant friendship. I also thank the entire CIFC group for accepting and providing me an excellent working environment.

To Ana Vieira, I express my deepest gratitude for all the help, support, and encouragement, even in the bleakest days. Her almost constant liveliness and optimism were invaluable to endure the accomplishment of this work.

To Tiago Jesus, I offer my special thanks for his friendship, brainstorming sessions, Dr. Phil sessions, vegetable suicide sessions, and invaluable contributions (constructive and destructive) to the elaboration of this work. I also sincerely hope we get those chats published in Science.

To André Barros I am thankful for showing me the immense (and underestimated) power of a series of bad jokes in a row. So many days would have been a burden, if not for this unspeakable contribution. Of course, the friendship and support were also appreciated.

To Joana Costa I give my thanks for taking me in the first steps of the technical molecular biology procedures. To Francisco Pina-Martins and Bruno Vieira, I sincerely appreciate all the help and tips in bioinformatics and data analysis and, most of all, for dazzling me with the Linux OS. In return, I will make sure to evolve my pokemons. To all the CoBiG^2 members in general, I extend my thanks for their friendship, support and help, whenever needed.

Last but not the least, I deeply thank my parents and grandparents for their immense patient, infinite support, and huge help, without which I would miss so many opportunities to enrich this work.

Resumo: O café é uma dos produtos mais comercializados em todo o mundo, contribuindo significativamente para à economia de mais de 60 países tropicais. A sua importância a nível económico e social é patente quando se considera que centenas de milhões de pessoas dependem, directa ou indirectamente, do rendimento que esta industria lhes fornece para garantir o seu bem-estar e qualidade de vida. No entanto, a sua produção pode ser extremamente sensível a distúrbios provenientes de factores extrínsecos, tal como as doenças fúngicas, e as consequências que daí advêm acarretam efeitos socioeconómicos devastadores para muitos dos países produtores. Por outro lado, a maioria das variedades de cafeeiro actualmente cultivadas em todo o mundo, resultam de um recente período de domesticação que, apesar de ter permitido a criação de variedades extremamente produtivas e de alta qualidade, diminuiu grandemente a variabilidade genética das suas populações. Em termos práticos, tem-se reconhecido cada vez mais que este evento homogeneizador fomenta a emergência e disseminação de agentes patogénicos, facilitando a sua adaptação e transmissão entre hospedeiros. Deste modo, e dada a crescente ameaça que as doenças causadas por estes agentes patogénicos estão a gerar, torna-se essencial que as estratégias para o seu controlo incluam a maior quantidade possível de informação no que toca aos aspectos evolutivos e demográficos desses agentes, analisando as assinaturas deixadas na estrutura das suas populações e que se reflectem na variação neutral e adaptativa na

A antracnose dos frutos verdes do cafeeiro, conhecida como Coffee Berry Disease, é uma devastadora doença que *Coffea arabica*, a espécie de cafeeiro mais importante e valiosa no mercado. Esta doença, causada pelo fungo *Colletotrichum kahawae*, é relativamente recente, tendo sido documentada pela primeira vez no Quénia em 1922. Desde então, tem-se disseminado por todos as regiões onde o cafeeiro Arábica é cultivado no continente Africano, onde ainda hoje se encontra confinada. No entanto, uma vez que esta doença é dada como o principal factor limitativo na produção de café Arábica neste continente, o receio da sua dispersão para outros países produtores, particularmente para América Latina, é motivo de grande preocupação.

Vários estudos têm sido realizados na perspectiva de entender como é que *C. kahawae* surgiu e como é que as suas populações estão estruturadas, de maneira a retirar informação acerca do seu potencial patogénico, assim de como essas populações evoluíram e dispersaram ao longo do tempo e do espaço. No entanto, apesar do esforço empreendido, este agente patogénico tem revelado uma variabilidade genética extremamente baixa, o que até agora tem limitado possíveis inferências acerca da sua história evolutiva. Deste modo, a maioria da informação que se tem acerca da sua origem e

disseminação baseia-se em dados históricos.

Duas razões principais têm sido invocadas para explicar a uniformidade genética de *C. kahawae*. Em primeiro lugar, supõe-se que tenha tido uma origem recente do grupo de espécies próximo, *C. gloeosporioides*, e três hipóteses têm sido propostas nas quais a sua emergência terá ocorrido através de: 1) mutação a partir de uma linhagem ligeiramente patogénica de *C. gloeosporioides* existente em cafeeiros Arábica; 2) hibridação entre duas linhagens de *C. gloeosporioides* presentes nas espécies parentais de *C. arabica*, como *C. canephora* ou *C. eugenioides*; 3) transferência de uma das espécies parentais de *C. arabica*, onde estaria presente como saprófita, para se tornar um agente patogénico agressivo no cafeeiro Arábica quando este foi reintroduzido no continente africano para fins comerciais. No entanto, *C. gloeosporioides* é um complexo de espécies extremamente abrangente, que ocorre em mais de 1800 espécies vegetais, e tem permanecido taxonomicamente mal descrito, pelo que a simples afirmação da sua proximidade com *C. kahawae* traz muito pouca informação. Deste modo, a questão mais relevante que se coloca é de que linhagem dentro deste complexo terá surgido *C. kahawae*.

A segunda razão diz respeito à assexualidade de *C. kahawae*. A ausência de fase sexuada, quer em meio de cultura, quer no campo, associada à sua estrutura populacional aparentemente clonal, têm sido mencionadas como evidências sugestivas da assexualidade desta espécie, embora inconclusivas. Neste aspecto, uma análise complementar útil e que tem levado a uma informação mais completa sobre o modo sexual de várias espécies de fungos, tem sido a investigação da evolução molecular dos genes que regulam a reprodução sexual, designados genes MAT, entre espécies próximas, tal como *C. kahawae* e *C. gloeosporioides*. Por outro lado, o género *Colletotrichum* possui um sistema genético de reconhecimento sexual único em todo o reino dos fungos, e cujo conhecimento acerca da evolução destes genes tem permanecido desconhecido.

Neste trabalho, recorrendo a várias ferramentas laboratoriais e analíticas da biologia molecular, pretendeu-se cumprir dois amplos objectivos. Em primeiro lugar, desenvolver novos marcadores moleculares mais informativos do ponto de vista filogenético e filogeográfico, que pudessem ser postos em prática para concretizar o segundo objectivo, de desvendar não só as relações filogenéticas entre isolados de *C. kahawae* e de *C. gloeosporioides*, mas também a diversidade genética e estrutura populacional de *C. kahawae*, de maneira a se poderem realizar inferências acerca da sua origem e dispersão.

Neste trabalho foi conseguido com sucesso o desenvolvimento de seis novos marcadores

moleculares na região genómica Apn1/MAT, que se revelaram extremamente variáveis e mais informativos que os marcadores tradicionalmente usados. Usando uma amostragem representativa de 22 isolados de várias espécies dentro do complexo *C. gloeosporioides* e incluindo *C. kahawae*, aplicaram-se estes novos marcadores, verificando-se que o aumento do nível de resolução das relações evolutivas entre espécies, e até entre populações, era extremamente significativo. Um dos marcadores desenvolvidos, que compreende uma região inter-génica, revelou um potencial singular em termos informativos, permitindo por si só uma inferência filogenética equivalente à concatenação de sete marcadores. Por outro lado, uma vez que o tempo de divergência entre várias espécies próximas dentro de complexos tende a ser muito reduzido, detectou-se ainda uma presença acentuada do fenómeno de *lineage sorting* incompleto, que tem como uma consequência a discordância entre as topologias das árvores filogenéticas reconstruídas a partir dos vários marcadores utilizados. Salienta-se deste modo, ser essencial o uso de marcadores de vários *loci* no genoma para a resolução deste problema.

Com base na selecção dos *loci* mais variáveis, analisou-se um total de 85 isolados, compreendendo uma extensa amostragem de isolados provenientes de uma vasta área onde *C. kahawae* é encontrado, assim como uma selecção representativa de isolados de *C. gloeosporioides* de *Coffea* spp. e de outros hospedeiro. Utilizando um conjunto de seis marcadores moleculares (ITS, β-tub2, ApMAT, Apn15L, MAT1-2-1 e MAT5L) e recorrendo a abordagens analíticas filogenéticas e filogeográficas, os resultados obtidos revelaram uma divergência muito superior ao esperada entre *C. kahawae* e isolados de *C. gloeosporioides* de cafeeiros, o que não suportava a hipótese de uma origem recente. Por outro lado, uma linhagem de *C. gloeosporioides* proveniente de *Mangifera indica* apresentou uma relação filogenética extremamente próxima com *C. kahawae*, ainda que biologicamente fosse uma espécie distinta devido à sua incapacidade de causar sintomas de antracnose nos frutos verdes. Estes dados sugerem que o agente patogénico, *C. kahawae*, possa ter emergido saltando de um hospedeiro diferente para os cafeeiros, ao contrário do pensamento corrente em que *C. kahawae* evoluiu e emergiu do género *Coffea*. Por outro lado, foi também possível desvendar alguma variação genética dentro de *C. kahawae*, distinguindo-se três grupos de haplótipos que estavam intimamente relacionados e estruturados de acordo com a sua distribuição geográfica. Reconstruindo o estado ancestral desses três grupos e com base na amostragem utilizada, verificou-se que o grupo da região Angolana era o mais ancestral, enquanto os isolados do Quénia e de outras regiões do Este Africano eram derivados desse grupo. Estes resultados vêm contrariar a ideia actual da origem de *C. kahawae* centrada no Quénia e sugerem uma hipótese alternativa, que aparentemente não é suportada e

dificilmente seria prevista à luz dos dados históricos disponíveis. No entanto, a origem assumida para *C. kahawae* baseia-se na premissa de que este emergiu a partir de espécies de cafeeiros, o que pode não corresponder à realidade e que de alguma forma fornece uma possível explicação para o padrão evolutivo encontrado no nosso estudo. Finalmente, com a mesma amostragem, foi realizada uma análise da evolução molecular do gene MAT, *MAT1-2-1*, baseada em modelos estatísticos de máxima verosimilhança (*Maximum likelihood*), que se revelou que este gene parece estar sob uma intensa selecção purificadora, possivelmente devido a uma função biológica importante. Apesar disso, dois haplótipos separados por uma mutação não-sinónima foram encontrados nos grupos de *C. kahawae*, revelando a inesperada presença de duas proteinas diferentes ao nível da sequência de aminoácidos numa espécie tão geneticamente uniforme. Apesar de não se conseguir excluir a hipótese uma variação fenotípica neutral, este evento pode sugerir que esta proteína tenha um papel adaptativo a nível populacional dentro de *C. kahawae*. No geral, este trabalho cumpriu os objectivos a que foi proposto, revelando resultados inesperados e que irão certamente requerer uma investigação futura para apurar as hipóteses aqui levantadas. Espera-se assim que ele venha a ter impacto na comunidade científica interessada nesta doença e agente patogénico e que venha a estimular futuros estudos e trabalhos na biologia evolutiva deste agente patogénico, tão relevante ao nível económico e social. No geral, este trabalho cumpriu os objectivos propostos, revelando resultados inovadores e inesperados, particularmente na origem do agente patogénico estudado, e que irão certamente requerer uma investigação futura para apurar as hipóteses aqui levantadas. Espera-se assim que ele venha a contribuir para o aumento do conhecimento e que seja útil para a comunidade científica interessada nesta doença e agente patogénico, tão relevante a nível económico e social.

Palavras-chave: *Coffea* spp., Doença vegetal, Antracnose, *Host-shift*, Filogenética, Filogeografia

Abstract: In this work, a range of phylogenetic, phylogeographic and statistical methods of molecular evolution were used to investigate the evolutionary biology of an emergent pathogen in coffee crops (*Coffea* spp.), *Colletotrichum kahawae*. *C. kahawae* was first reported in Kenya, 1922, and causes severe disease (Coffee Berry Disease) in Arabica coffee throughout the African continent, where it is still restricted. However, the origin and spread of this pathogen has hitherto relied on historical data, and molecular studies have not been able to unearth enough genetic variation to infer about these processes, possibly due to its putative asexual nature and assumed recent evolution from *Coffea* spp. inhabiting strains of the *C. gloeosporioides* complex. To address this issue, a set of molecular markers was successfully developed, including a mating-type gene (*MAT1-2-1*), showing a greater informative potential than traditional markers. The most valuable markers were selected and used to analyze a representative sampling of *C. kahawae*, throughout most of its range, and *C. gloeosporioides*, mainly from *Coffea* spp. worldwide. Unexpectedly, the obtained results from the evolutionary relationships between *C. kahawae* and *C. gloeosporioides* suggest that the former may have emerged through a host-shift from hosts other than *Coffea* spp.. Moreover, the phylogeographic analysis of *C. kahawae* revealed a geographically structured population and the Angolan group as the most ancestral state inferred. Although our results come in opposition to the current Kenyan hypothesis and are not supported by the historical reports, the prevailing view follows the premise that *C. kahawae* evolved from *Coffea* spp., which our data suggests otherwise. Regarding the molecular evolution of *MAT1-2-1*, our results suggest a possible adaptive role within *C. kahawae* populations, due to the presence of two haplotypes in a highly conserved gene. These results have significantly contributed to fill a void in the current knowledge of the evolutionary biology of this pathogen and will hopefully stimulate further research.

Keywords: *Coffea* spp., Plant disease, Anthracnose, Host-shift, Phylongenetic, Phylogeography

Table of Contents

Part III

List of Abbreviations

AFLP – Amplified Fragment Length Polymorphisms

bp – Base pair

CBD – Coffee Berry Disease

cca – *Colletotrichum caffeanum* acervuli

ccm – *Colletotrichum caffeanum* mycelial

ccp – *Colletotrichum caffeanum* pink

CIFC – Centro de Investigação das Ferrugens do Cafeeiro

CLR – Coffee Leaf Rust

EtBr – Ethidium bromide

HDT – Híbrido de Timor

HMG – High Mobility Group

ICO – International Coffee Organization

ITS – Internal Transcribed Spacer

rDNA – Ribosomal DNA

MAT – Mating-type

mtDNA – Mitochondrial DNA

RFLP – Restriction Fragment Length Polymorphism

RAPD – Random Amplified Polymorphic DNA

VCG – Vegetative Compatibility Group

Part I

General Introduction

Coffee is a primary export of many developing countries that rely to a greater or lesser extent on its foreign exchange earnings for financing essential imports and services. Its industry in developed countries is generally perceived as prosperous and uncontroversial. But, although the coffee business is booming in consuming developed countries, current rock bottom prices are causing immense distress to countries where coffee is a key economic activity, as well as to the farmers who produce it. Coffee has been subjected to the rigorous discipline of market forces, with depressed prices resulting from excess of production over demand, interspersed with short periods of high prices stimulated by temporary setbacks in production.

Meanwhile the costs of inputs, such as transport, machinery, labor and materials, have continued to increase. The decreasing profit margins resulting from these opposing trends have forced coffee farmers to economize and this has often led to a reduction in the use of agricultural inputs necessary for optimal coffee production. The effects have been felt most by the millions of smallholder farmers who rely on coffee as their only or main cash crops and who lack financial resources and the possibilities of economy of scale. Resources allocated to crop protection have often been the first to be cut, and this has added further distress in pest management.

Over the last decades there has been a marked shift away from the reliance on pesticides for management of pest and diseases, towards a more integrated approach using a variety of methods. As a tropical perennial crop, coffee has a wider environmental effect as it is grown in some of the world's most ecologically sensitive regions. Moreover, the way in which coffee crops are currently grown using agricultural practices, has led to a highly susceptible agro-ecosystem not only for the emergence of new pathogens but also for their transmission. An understanding of the biology and ecology of such pathogens is therefore essential, as a sustainable management system of coffee crops is unlikely to be achieved without it. Diseases on coffee crops are emerging and many examples can illustrated how sensitive are both crops and the populations that rely on them, to the nefarious disturbances from these extrinsic factors. The Coffee Leaf Rust (CLR), caused by *Hemileia vastatrix* Berkley and Broome and Coffee Berry Disease, caused by *Colletotrichum kahawae* Waller and Bridge, are the most impactful diseases ravaging coffee crops and have limited their production to a fairly high extent.

Objectives

In this research the focus has been given to the evolutionary biology of the emergent pathogen *C. kahawae* and its closely related taxa, using a wide range of phylogenetic, phylogeographic and statistical methods of molecular evolution. The problematic of its origin and spread has been elusive so far, much due to the lack of appropriate molecular markers to unravel those issues. This has left a void in the current understanding of the evolutionary processes that led to its emergence and dissemination and of the evolutionary potential that this pathogen may present. The present work intents to, at least, contribute to fill these gaps by studying the patterns of neutral and adaptive genetic variation of *C. kahawae* and its closest taxa, providing information on the events that shaped their past evolutionary histories in order to more effectively prepare for what the future may reserve. Specifically, this research aimed at the:

1- Development of a molecular marker suite that may provide a significant improvement in phylogenetic and phylogeographic resolution for the *C. kahawae – C. gloeosporioides* complex

2- Analysis of the genetic variability and differentiation of *C. kahawae* and *C. gloeosporioides* populations through a multi-locus sequencing approach, to infer evolutionary relationships and dissemination patterns.

3- Analysis of the molecular evolution of a mating-type gene (*MAT1-2-1*) to assess selective constraints and haplotype distribution in inter- and intra-specific datasets.

Before presenting this work, a brief general introduction was made on the most relevant aspects of the host, *Coffea arabica* L., the disease, Coffee Berry Disease, and the pathogen, *C. kahawae*, on the scope of the issues that will be addressed latter. A first article will follow, in which the development of a new set of molecular markers is described and presented as more suitable to address the problematic above mentioned. The application of those and other markers to unravel the longstanding issues of *C. kahawae* emergence, dissemination and evolution is presented in the second article. Finally, some final remarks will be made concerning how this research has contributed to enhance the current knowledge of this severe disease and possible future directions on its research are suggested.

1. The Host – Arabica coffee plants

1.1 Taxonomy

According to Cronquist (1988) [1], the taxonomic classification of the *Coffea* genus is as follows:

Kingdom: Plantae

Division: Magnoliophyta

Class: Magnoliopsida

Order: Rubiales

Family: Rubiaceae

Genus: *Coffea*

At the infrageneric level, Chevalier's (1947) classification [2], based on geographical distribution and fruit characteristics, remains one of the most used and comprises four sectors: *Paracoffea*, *Argocoffea*, *Mascarocoffea* and *Eucoffea*. Nonetheless, taxonomic progress has been made, particularly since the 80's [3,4], in order to clarify this classification given the increasing number of described species and the new molecular, biochemical, cytogenetic and geographical methods available [5-7]. The current subgeneric classification of the genus encompasses 103 described species and only two subgenera [4,8,9]: *Coffea* subgen. *Coffea* and *Coffea* subgen. *Baracoffea*., with most of the species belonging to the *Coffea* subgen. *Coffea*, including those used for producing the beverage coffee. It is noteworthy that 70% of these species are threatened with extinction as a combination of a decline in quantity and quality of the habitat [9].

Only three species are used in commercial coffee production: *C. arabica*, *C. canephora* L. and *C. liberica* Bull ex Hiern [9]. Among these three, *C. arabica*, also known as Arabica coffee, is by far the most important commercial species, accounting for over 75% of world production. *C. canephora*, also known as Robusta coffee accounts for most of the rest of the trade, with *C. liberica* contributing with less than 1% [10].

1.2 General Characteristics

Coffea spp. are evergreen, glossy-leaved shrubs or trees with 5-10m high and most are adapted to a

tropical forest habitat [10]. Their general botanical features include elliptical leaves, with pointed tips and occurrence in pairs. They have short petioles with small stipules, and domatia (small pits) are present on the undersides of leaves at the junction of the main veins. Following the rainy season, flower clusters are produced in leaf axils and usually, a period of nine months spans between flowering and fruit ripening. The fruit starts as a two seeded green drupe, becoming red or yellow as it ripens. The stems exhibit a dimorphic branching due to the different development of two buds that occur, one above the other in each leaf axil of the main stem. The upper bud develops to produce a lateral or primary branch. The primaries develop in succession from the base upwards and grow horizontally (plagiotropic) on opposite sides of each node, and they bear the flowers and fruits. The lower bud can only develop into a vertical (orthotropic) branch, and remains dormant until the main stem has been damaged or pruned, when it grows around the primary to produce a new vertical vegetative shoot.

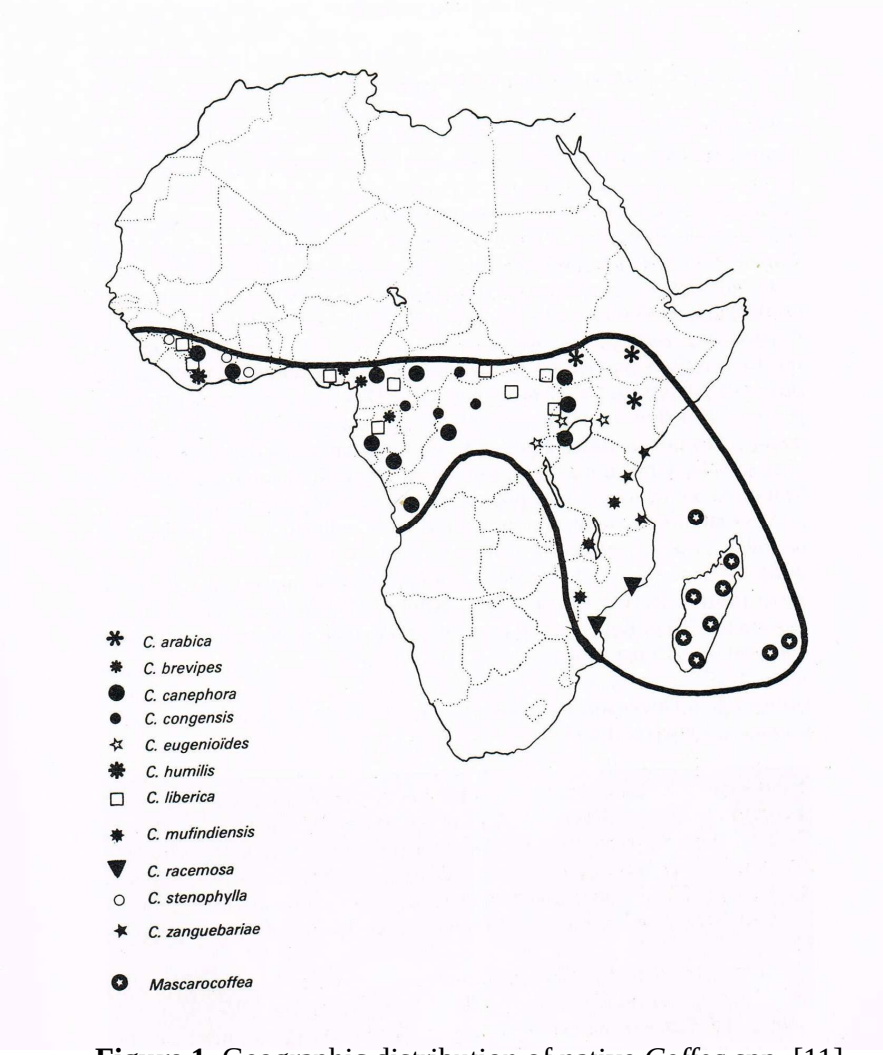

Figure 1. Geographic distribution of native *Coffea* spp. [11].

One particular and interesting aspect of this genus is not only its morphological variation but also the adaptation of its species to a wide range of environmental conditions. The habitats of the various coffee species often correspond closely with specific biotopes [11]. *C. arabica* is better adapted to cool and humid environmental conditions at higher altitudes (1300-1800m)[12] and has a center of genetic diversity on the Ethiopian plateau, which lies outside the distribution area of the other species[13]. However, the most likely center of origin is in Uganda where the distribution of the two parental species overlaps [13]. *C.*

canephora and *C. liberica* are native to the tropical regions of west Africa and are usually found in humid and warmer environments of the lowlands [11]. The current geographic occurrence of natural *Coffea* spp. populations is depicted in Figure 1.

All *Coffea* species are generally self-sterile diploids (2n = 22) except for *C. arabica*, which is a self-fertile allotetraploid (2n = 44) with a diploid-like meiotic behaviour [13]. This event has been shown to be the result of a natural hybridization between a maternal *C. eugenioides*-like and a *C. canephora*-like genomes, followed by an unreduced gamete formation and enhanced autogamous reproduction [14,15]. Moreover, the lack of genetic divergence from the parental genomes and within *C. arabica* can argue for a unique and relatively recent origin [13]. Absolute crossing barriers appear to be absent in this genus, even though there is a varying degree of successful hybridizations [11].

Despite being rather susceptible to diseases, *C. arabica* is also responsible for the production of the best quality coffee, with lower caffeine content. In combination with the ease that its traits are genetically stabilized due to an autogamous reproduction, it has long been a desirable target for domestication and crop exploitation.

1.3 Domestication and Dissemination

Hosts and their pathogens are often entwined in a close interdependence, which led to the realization that the nature and evolution of the host itself may play a critical role in the emergence and spread of pathogens. Arguably, one of the most dramatic events in the evolutionary history of a plant host is the transition from naturally occurring wild populations to dense and uniform agricultural ecosystems (agro-ecosystems), with the advent of agriculture, which has enormous consequences in the host-pathogen system [16]. Moreover, the subsequent human mediated movements of plant material associated with the agricultural practice were instrumental in shaping the present distribution of pathogens [17].

The history of the domestication and dissemination of coffee throughout the world is relatively recent and incomplete, with much of the information scattered and often a mixture of fact and fantasy (Figure 2). Despite its African origin, Arabic coffee cultivation may have begun as early as AD 575 in Yemen [18,19] but only in the 16th century, prompted by the discoveries, reports of its cultivation and beverage quality reached Europe [20]. The early diffusion of coffee crops was slow, mainly due to the desire of the Arabians to maintain the monopoly of the profitable business that coffee was already becoming

[11]. It was not long until the Europeans appreciate the potential of coffee as a crop in the overseas territories and in the 17th century, the Netherlands East India Company managed to ship coffee plants to Java, which took over Yemen as the main source of coffee, before other European nations entered the trade [10,21]. In the early 18th century, a single plant from Java was taken to the Amsterdam Botanic Gardens, where it flowered and produced berries – this was latter known as "typica" variety of coffee (*C. arabica* "Typica" Cramer), which is currently the most common by far. The French East Indian Company soon followed and, during the 17th century, several introductions were made from Yemen to Reunion, an island on the East African coast, where the crop flourished - the "Bourbon" variety (*C. arabica* "Bourbon" Choussy) [18]. These early introductions created the two narrow genetic basis that fostered most of the commercial Arabic coffee crops worldwide and represent the onset of the coffee cultivation in a large and global scale [22].

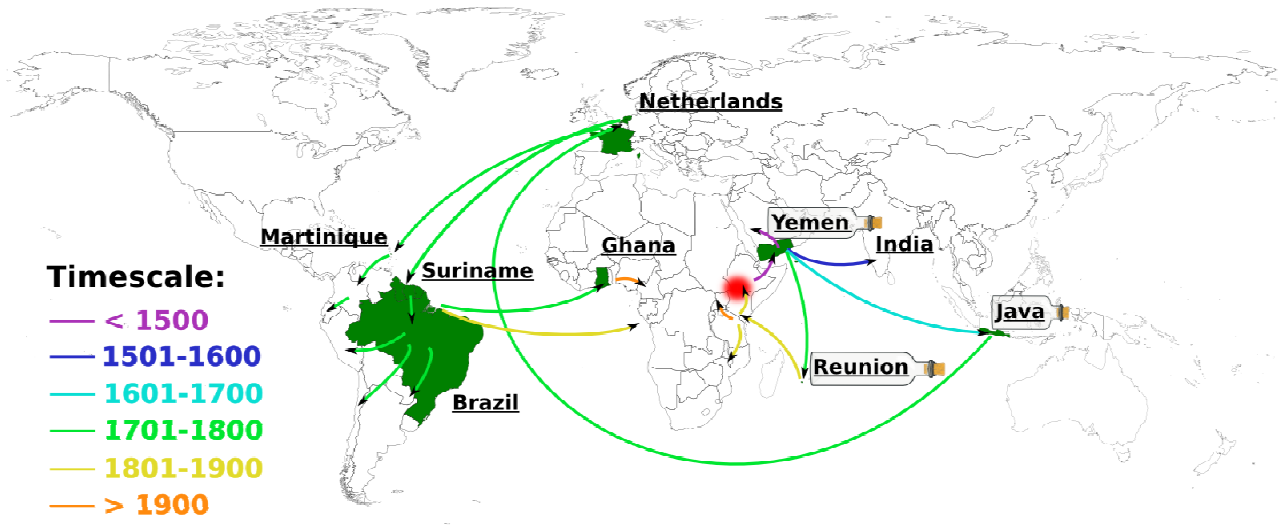

Figure 2. Schematic overview of the main steps in Arabica coffee dissemination with a timescale. The red circular blur indicates the geographic origin of *C. arábica*. Bottles represent countries in which severe historcial bottlenecks have occurred.

In 1706, plants were brought from the Botanical Gardens in Amsterdam to Surinam, onwards to French Guiana and from there to Brazil, in South America [20,21]. Further introductions of "Typica" coffee were made by the French in Martinique [10]. Subsequently, during the 18th century, the crop rapidly dispersed throughout most of the suitable locations on Latin America, not only from sources within the continent but also from Reunion [21].

In the African continent, coffee crops were not so much interesting for the established Europeans and

thus, the history of its cultivation only began in the 18th and 19th century [20]. Despite, being the cradle of the Arabic coffee, most of the actual commercial plantations have resulted from re-introductions mainly from South America and Reunion [10,21]. Throughout this period, Arabic coffee crops began to emerge in East Africa mostly from Reunion and also in the West coast from several locations in South America [20,21].

The other currently significant commercial species, *C. canephora*, is indigenous to the forests of West Africa but it was not used or cultivated until the beginning of the 20th century (L. Manuel, PhD thesis, Instítuto Superior de Agronomia, 2007). By this time, it became widely produced in the low lands of many West African countries, where it is native, and also in regions seriously affect by coffee diseases such as Coffee Leaf Rust due to its higher resistance [21].

In summary, the domestication of coffee crops is recent compared with other agricultural crops, such as wheat and barley [23,24], emerging between the 15th and 16th centuries and its dissemination on a global scale only spanned three to four hundred years. Intensive selection, severe bottleneck events and cultivation of selected phenotypes altered the populations of the wild progenitor species into the domesticated varieties of the crops we know today, along with an astonishing decrease of their genetic variability [25]. This new and rapidly created agro-ecossytem provided denser and genetically uniform host populations, highly conducive for the emergence and dispersal of plant pathogens [16]. The lack of genetic variation in Arabica crops makes them highly vulnerable to disease outbreaks because virulent pathogen genotypes adapted to a particular host genotype can increase very rapidly in frequency, quickly generating a degree of host specificity or race specificity rarely seen in natural ecosystems [16,26,27].

1.4 Production and Economic Relevance

The importance of coffee for man cannot be overstated as it is the world's most heavily traded commodity after oil [28]. According to the International Coffee Organization (ICO), coffee production in 2008 was over 120 million 60Kg bags (7 million tons) with more than $70 billion total retail value (see: http://www.ico.org). Africa corresponded with 11,5% of world production, Asia and Oceania with 25,6% and Latin America with 62,9% (see: http://www.fao.org). The main coffee growing countries are Brazil, Vietnam and Indonesia, which account for almost 60% of the total coffee production worldwide.

Its importance is even greater for the over 50 producing countries where coffee can represent the kernel

of total exports revenues and influences directly or indirectly the lives of roughly 120 million people (Figure 3). Coffee is currently grown in most tropical and sub-tropical regions with a prevalence of *C. arabica* in Latin America and East Africa and *C. canephora* in West Africa and Asia. Arabica coffee is preferred over all other species because of its superior cup quality and if it had not been so vulnerable to diseases, it would certainly be the exclusive producer of all commercial coffee, as it was until the

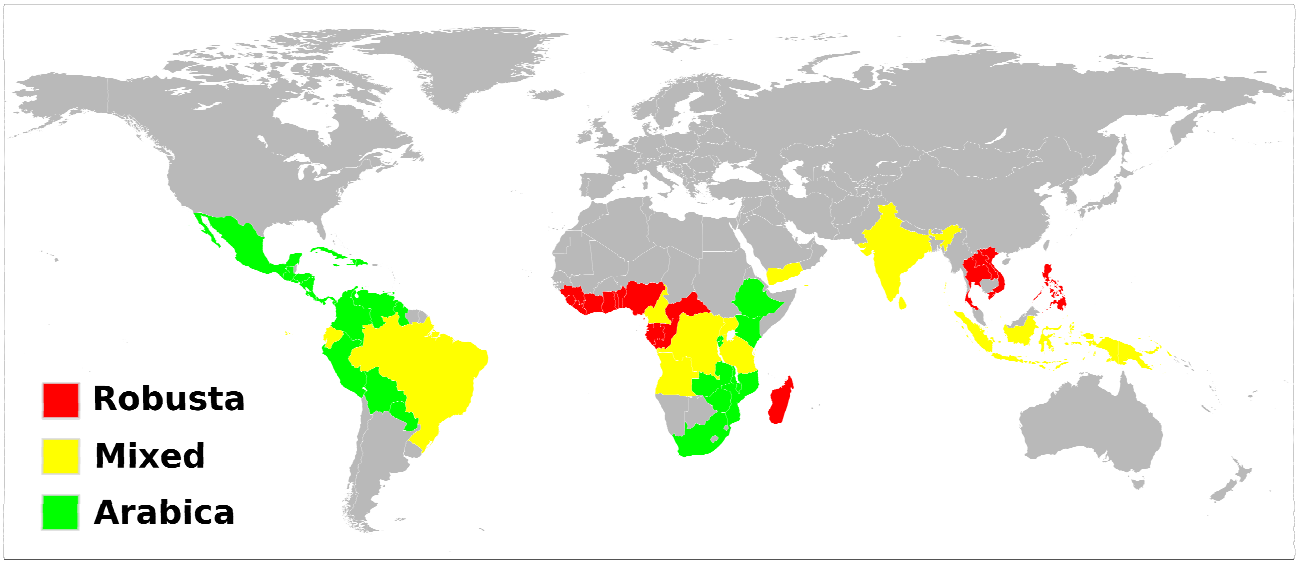

Figure 3. Coffee producing countries worldwide. China produces mixed species of coffee, although it is not highlighted (source: http://www.ico.org).

end of the 19th century [11]. Currently, it still contributes with 75% of the world coffee exports (see: http://www.ico.org).

The coffee industry is generally perceived as prosperous, largely due to the boom of coffee consumption in developed countries [29]. However, the disequilibrium between supply and demand has caused great oscillation in prices and consequently, the coffee market has endured multiple crises with severe social and economic consequences [29]. These can be further aggravated by limiting factors of coffee production, such as the major diseases: Coffee Leaf Rust (*Hemileia vastatrix*) [30] and Coffee Berry Disease (*Colletotrichum kahawae*) [31].

Since coffee is mainly grown in developing countries, it represents one of the few solutions to fight poverty, especially in some countries of the American and African continent [28].

2. The disease – Coffee Berry Disease

Coffee Berry Disease is an emergent and severe disease of Arabica coffee crops, originating anthracnose of green fruits, caused by the fungus *Colletotrichum kahawae* Waller & Bridge. It is considered the main limiting factor of Arabica coffee production in the African continent, where it is endemic, leading to severe yield losses in some regions. Thus far, no absolute effective control measure has been developed and therefore, there is an increasing concern of its dispersal to coffee growing areas in other continents, such as South America, where it could produce catastrophic consequences for coffee trade and production [32]. Some aspects of the disease will be reviewed, since they will closely reflect the life mode of the pathogen and consequently, will influence its genetic structure and evolutionary potential [33].

2.2 Origin and Distribution

The origin of CBD still remains poorly understood and largely based on historical data. It was first reported in 1922 in western Kenya [34] when it led to the destruction and abandon of *C. arabica* plantations in some regions [12]. Despite the little attention received during the early stages of its documented emergence, African coffee growers soon witnessed a swift dissemination of CBD throughout most of the continent [35]. The disease was reported in Angola (1930) Congo RD (1938), Cameroon (1955), Tanzania (1964), Ethiopia (1971), Malawi and Zimbabwe (1985) and eventually, most of the Arabica coffee areas of Africa were affected [28,36]. Nonetheless, the disease is currently limited to this continent and is rarely reported below 1400 m [37,38]. Climatic conditions commonly found in areas at higher altitudes, such as high humidity and cool temperatures of around 18 – 20 ºC, highly favor its occurrence and are also the most suitable habitat for the host [39]. *C. arabica* remains the only species affected by CBD in nature, despite its occasional sympatry with other *Coffea* spp. such as *C. canephora*, which seems to be resistant to the disease [31].

2.3 Economic Impact

Since its emergence, CBD has been associated with severe yield losses and could be a factor preventing cultivation of Arabica coffee in highland regions [38]. In African countries growing *C. arabica*, crop losses of 20-30% are common but can exceed 80% in extremely wet years, if no control measures are applied [12,35,39]. Although the majority of the Arabica coffee production is concentrated in the American continent, yield losses in African countries still reach several millions of US dollars annually,

which is particularly burdensome in developing countries [40]. Ethiopia, for example, is the largest coffee producing country in Africa and grows exclusively *C. arabica* [41]. Coffee production represents about 63% of the total commodities export revenues and the overall national loss due to CBD reaches 30% [41]. If these percentages are translated into monetary value, Ethiopia suffers losses of 84 million US annually only due to average CBD outbreaks. These losses sound even more dramatic when we consider that over 15 million Ethiopians depend directly or indirectly on the coffee industry and the socio-economic negative consequences that they might suffer [28]. Taking the remaining African coffee producing countries into account, overall losses are estimated to be around 300 – 500 million US dollars per year [45].

2.4 Symptoms

Although CBD symptoms can be found in other organs of the coffee plant, their effects are more devastating in the fruits, causing their premature drop or blackened mummification [12,42,43]. Generally, it affects any part of the green berries in the expansion phase and may take two forms. *Active lesions* (anthracnoses) may start as a dark brown, slightly sunken spot (necrosis) [37]. Under appropriate conditions the spot eventually enlarges and covers the whole berry, which is reduced to a black and deformed fruit with no commercial value, preventing further processing of the beans. Before the fruit decays completely, if the environment is sufficiently wet, small pinkish masses of spores develop on the surface of the decayed pulp [35]. In *scab lesions*, however, the lesion develops slowly due to a plant resistance response and is characterized by a buff color and the presence of small dark spots. The growth of the lesion is limited, as the fungus dies out, not affecting the normal bean development [44].

Anthracnose symptoms in other parts of the plant such as leaves and stems are seldom and economically unimportant [12]. Infection of the flowers may be economically important, as they are very susceptible at all stages, but it is difficult to measure since they drop shortly after the infection [38].

2.5 Epidemiology

Precipitation, humidity and temperature are the most important factors determining CBD outbreaks, playing a key role on the germination, production and dispersal of spores by the fungus [38]. In regions with two rainy seasons, two flowering events are created, which dramatically increases the emergence

of epidemics and prolongs the disease active period [28].

Germination of spores must occur in the presence of water and thus, humid environments are highly conducive for the development of the first stages of infection. The temperature will set the time that these early stages will take to unfold, which is minimum at cool temperatures of 22ºC and increases as temperature reaches its biological maximum (30-35ºC) or minimum (10-15ºC). Moreover, the germination rate depends on the maturation stage of the fruit, rather than the variety of the infected plant [45].

The sporulation of the pathogen is higher at the onset of rain events and it mainly occurs on the surface of infected berries, which represent the most relevant source of inoculum [46]. They are found in mucilaginous masses which prevent them from being dispersed by wind [47]. However, they are suited and readily dispersed by rain through “splash” dispersal, which is particularly effective for short range dissemination [48]. The impact of rain drops spreads the spores to a variable range of distances, which may be further enhanced by the wind [49,50]. The mucilaginous layer protects the spores from dissection and loss of viability in dry weather until subsequent rains transport the spores further away [48].

Despite the incremental nature of this mode of dispersal through time, its overall capacity is still relatively low as very few spores travel more than 1 m per rain event [48,49]. However, if one considers the increasing reports of anthropogenic dispersal of fungal diseases through agricultural activities, travelers, war events, it might be expected that the dispersal potential of *C. kahawae* is greater than the observed through natural phenomena alone [17,47]. Other animals such as birds and insects may also come into contact with infected fruits and become disease vectors [51]. Therefore, it might be prudent to take these factors into account when analyzing dispersal patterns of *C. kahawae* from historical and molecular data.

2.6 Control

Chemical control was the first attempt to manage disease outbreaks in the 60's and remains the most common until nowadays, namely using cupric fungicides [37]. However, it has revealed somewhat inefficient whether because chemicals are washed away in the rainy seasons, when the fungus strikes most, or because the fungus rapidly generates tolerance in the field [52,53]. Moreover, smallholders, who produce the majority of the Arabica coffee in most countries of East Africa, are usually unable to

carry out the recommended complete spray program, due to the costs associated with the use of fungicides [11]. Thus, a common major drawback is the counterproductive effect that irregular applications have on reducing the sporulating capacity of the saprophytic *Colletotrichum* spp., while favoring the proliferation of the CBD pathogen in the longer term, since the spores of *C. kahawae* represent only a small proportion of the overall *Colletotrichum* spp. spores present in coffee plants [12,37,54,55].

Recently, great expectations and results come from the breeding of resistant varieties in research centers, which are later introduced in the field. Since 1989 the Coffee Leaf Rust Research Center (CIFC) in Portugal, started a research line aimed at obtaining *C. kahawae* collections from different geographic locations where the disease exists, using them for screening coffee genotypes with different levels of resistance [12]. Hundreds of coffee genotypes arrive annually from those locations and from countries were the disease is still absent but presents a threat. Currently, none of the *C. arabica* genotypes show 100% resistance to all the *C. kahawae* isolates from CIFC's collection although some varieties like Rume Sudan, Ruiru 11 and derivatives from Timor Hybrid (HDT) shown high levels of resistance [56]. These resistant varieties are the result of decades of breeding and selection by coffee research stations, such as those in Kenya [22].

Nonetheless, the effects of resistance genes can be ephemeral in host – pathogen interactions. It is increasingly accepted that the durability of the resistance genes is mostly due to the nature of the pathogen population rather than to the nature of the resistance gene [57]. Population genetic parameters such as gene/genotypic flow, mutation rate and population size, as well as mating systems, are critical in the assessment of the potential risk of pathogen species in overthrowing hosts resistance and/or enhancing virulence [58,59]. For example, pathogens in homogeneous agro-ecosystems tend to exhibit large populations, with greater potential to generate genetic diversity through accumulation of mutations and less inbreeding. As a result, the pathogen's populations are able to respond more rapidly to the deployment of control measures, such as the introduction of new fungicides or resistant varieties [57].

For these reasons an expansion of our knowledge beyond the discovery of new resistant genes or new fungicides is required. It is vital to understand the dynamics and evolutionary history of the pathogen's populations as to effectively hamper their spread and improve disease control in economically and socially vital crops such as coffee.

3. The pathogen – *Colletotrichum kahawae*

3.1 Taxonomy

Until recently, the taxonomic description and position of the CBD causal agent was a subject of great confusion. From the range of *Colletotrichum* spp. that are isolated from coffee plants, four groups were initially described based on their morphological traits: ccm (*C. coffeanum* mycelial), cca (*C. coffeanum* acervuli), ccp (*C. coffeanum* pink) and the CBD strain [46]. The three former groups were latter recognized as *C. gloeosporioides* Penz (ccm and cca) and *C. acutatum* Simmonds (ccp), and proved to be non-pathogenic in green coffee berries [60]. Only the fourth group was able to infect both wounded and unwounded green berries and was formerly referred to as *C. coffeanum* [60]. However, *C. coffeanum* was described in 1901 based on *Colletotrichum* isolated from coffee in Brazil [61], where CBD does not exist, and was probably synonymous with *C. gloeosporioides*, which occurs as a saprophyte or weak pathogen of ripe berries and damaged coffee tissue worldwide [62]. Several authors attempted to emend this anomaly but it was not until 1993 that Waller and Bridge described *C. kahawae* as the causal agent of CBD and as a distinct species based on morphological, cultural and biochemical characters [63] and more recently on multi-locus datasets [64]. The current taxonomy of this fungus is as follows:

Kindgom: Fungi
Phylum: Ascomycota
Class: Sordariomycetes
Sub-class: Sordariomycetidae
Order: Incertae sedis
Family: Glomerellaceae
Genus: *Colletotrichum*
Species: *C. kahawae* Waller and Bridge

3.2 Phylogenetic Relationships

3.2.1. Onset of the molecular systematic shift

Morphological and cultural criteria for species recognition dominate fungal taxonomy, since the ca. 80.000 described fungi species have been diagnosed either by morphological or other phenotypic characters [65]. The greatest strength of these criteria is that they are so widely applied that comparisons can easily be made among described taxa and between new and existing taxa [66]. They

are also relatively easy to describe and apply, and so their use for species identification and delimitation has been widespread [67]. However, they are often unreliable, unpredictable, lack the ability to resolve cryptic species complexes and are limited by the genetic plasticity of many traits in culture [66,68]. Fresh isolates of *C. kahawae*, for example, may exhibit morphological and cultural distinct hallmarks but after subculturing several times they become morphologically indistinguishable from *C. gloeosporioides*, although growth rate in culture media is slower in *C. kahawae* [69]. Identification within genera such as *Colletotrichum* is further complicated as species have only a few distinguishing morphological characters [70]. Nevertheless, species within this genus have been primarily defined on the basis of morphology and host preference, leading to long-standing, unresolved taxonomic issues [68,70]. Therefore, as widespread and straightforward as morphological and cultural criteria may be, they should not be counted on to diagnose evolutionary meaningful species [66]. A pragmatic exception lies on the usefulness of pathogenicity tests to distinguish isolates of *C. kahawae* from other *Colletotrichum* spp., since this is one of the few unequivocal methods of recognizing this pathogen.

Molecular phylogenetic approaches have proven to be a more effective way to dissect evolutionary relationships and to recognize distinct species [66,71,72,73]. The analysis of variable nucleotide characters in sequence data has the potential to perform best because, once evolutionary species are formed from an ancestor, changes in gene sequences occur and can be recognized before changes in morphology or other phenotypic traits are evident. Moreover, they should not be as directly influenced by environmental factors as morpho-cultural traits [66]. Thus, the comparative analysis of homologous DNA sequences has been regarded as a more reliable method, with the ability to provide a considerable enhancement on the resolution and accuracy of phylogenies and, consequently, of species identification [74]. However, a major drawback in the reliance on a small proportion of the genome to understand phylogenetic relationships amongst species has been the risk of recreating gene trees rather than species trees, i.e. the phylogeny may not accurately recover the true species relationship, particularly when a single or few genes are used [74,75]. Thus, multi-gene phylogenies are beginning to be routinely employed to overcome those difficulties and to systematically characterize and diagnose species [74].

In the last two decades, molecular systematics have revolutionized fungal taxonomy and challenged many morpho-cultural based classifications, unearthing not only a much higher species richness among the already described groups but also providing insights on the evolutionary relationships between species [76,77]. Groups of cryptic species formerly lumped together on the basis of morphological criteria are now being revealed, sometimes with astounding results. *Fusarium graminearum*, causal agent of Fusarium Head Blight, was thought to be a single panmitic species spanning six continents

until multi-locus phylogentic studies identified eight geographically structured and phylogenetically distinct species [78]. In the *Colletotrichum* genus, the broadly defined *C. graminicola* species found on several cereal and grass hosts was recently subjected to the same approach, resulting in the identification of ten new species [79].

Thus, species concepts in the economically important genus *Colletotrichum* have been undergoing revolutionary change in recent years. The old morphology-based species are rapidly being superseded by taxa that are largely defined by molecular sequence. Cryptic species are increasingly being recognized that cannot be reliably separated using morphological or cultural methods. In the two next sub-chapters, focus will be on how these approaches have been applied to the understanding of the *C. kahawae – C. gloeosporioides* complex relationship and emergence of *C. kahawae* and which gaps remain unfilled.

3.2.2. The C. kahawae – C. gloeosporioides complex

From the early single gene phylogenetic studies [80,81] to the more recent multi-locus approaches [64,74], *C. kahawae* has revealed to position itself among the *C. gloeosporioides* species complex. An accurate phylogenetic arrangement of pathogens causing emergent fungal disease, particularly among the closest related taxa, is extremely relevant for practical and academic reasons [82]. However, in the *C. kahawae – C. gloeosporioides* relationship this is far from straightforward since *C. gloeosporioides* is though to be species complex, reported in at least 1800 plant species worldwide, with such a wide range of morphological and pathogenic variation, that the species name is of limited use for evolutionary inferences [68,81]. The 'gloeosporioides' complex arose because of the artificially enlarged spore range placed as a criterion for *C. gloeosporioides* identification [83] to overcome the instability of spore morphology under different conditions or from different hosts. As this criterion was followed by researchers, many *Colletotrichum* strains with similar cylindrical conidia were identified as *C. gloeosporioides,* lumping potentially distinct taxa into a species complex [84]. Molecular phylogenetic studies have partially addressed this complex, but the task has been quite challenging as the introduction of molecular data has been hitherto unfruitful. The Internal Transcribed Spacer (ITS) sequences of rDNA, which has dominated *Colletotrichum* systematics for the past 15 years [80,85-89] has left unanswered questions about overly broad species groups, due to its low variability (and thus low ability to discriminate and resolve the true species tree) escorted by a high rate of species misidentification in public nucleotide databases [90]. As much as 86% of the *C. gloeosporioides* ITS sequences deposited in GenBank show considerable divergence from the published epitype and most

likely represent other species [74]. Fortunately, epitypification, i.e. the selection of a fresh living specimen to serve as a representative type of a taxon species [91], has been recently applied to *C. gloeosporioides*, based on a collection from *Citrus* sp. in Italy, providing a solid starting basis for accurate species identification and relationships [92].

Nonetheless, this species complex remains taxonomically ill-defined and thus, simply stating a close relationship between *C. kahawae* and *C. gloeosporioides* is nearly meaningless. Therefore, in order to cope with this current limitation, one can only gather useful information about their relationship when isolates of both species are used and analyzed simultaneously. On the other hand, to improve the 'gloeosporioides' complex systematics it has revealed necessary to move beyond the use of nearly uninformative markers such as ITS. Only very recently are multi-locus datasets beginning to be employed and the demand for more informative markers is increasing considerably.

3.2.3. Emergence of C. kahawae

It is a longstanding view that *C. kahawae* has emerged fairly recently from the group species *C. gloeosporioides* based on the above mentioned similarities at the morphological, cultural and molecular level. However, it has already been shown that *C. gloeosporioides* is a large group species with considerable variation within its ranks, and thus a more relevant question is raised: From which sub-group did *C. kahawae* evolved?

Currently, there are three main hypotheses for the emergence of *C. kahawae*. Nutman and Roberts (1960) suggested that the CBD strain may have arisen by mutation from a mildly parasitic form, such as that causing brown blight in *Coffea* sp., some time prior to its first report in Kenya [35]. On the other hand, Robinson (1974) proposes that the CBD strain was originally a harmless pathogen of one of the diploid progenitors of *C. arabica*, such as *C. canephora* or *C. eugenioides*, but was able to cause a virulent disease when the cultivated varieties of *C. arabica* were re-introduced in the African continent [93]. The same author also suggests that given the high intesity of the disease, the CBD form may be a polyphyletic pathotype derived by hybridization between the two forms of the progenitor hosts, and possessed the natural maximum of horizontal pathogenicity [19]. What is remarkable, however, is the little solid data, other than observational and historical, that exists to support any of the mentioned hypotheses. At this point, there is not enough information to know which scenario fits best, but the authors seem to agree that *C. gloeosporioides* populations from *Coffea* spp. hosts are the most likely source of *C. kahawae*.

Several phylogenetic studies have included *C. kahawae* together with a collection of *C. gloeosporioides* from different hosts and geographic locations, but they generally lacked a representative sample of *C. gloeosporioides* from coffee hosts to properly address those hypotheses [42,81,94,95]. Although some of them reported a close relationship of *C. gloeosporioides* isolates from hosts other than coffee to *C. kahawae*, they relied on the ITS sequence region alone and thus, caution must precede inferences to be taken from these results [94,95]. An exception is the recent characterization of *Colletotrichum* spp. associated with coffee berries in northern Thailand, in which *C. kahawae* is also included. Several *Colletotrichum* spp. morphologically resembling *C. gloeosporioides*, were isolated from coffee hosts and a multi-locus phylogenetic approach revealed two interesting results [64]. Firstly, none of the analyzed isolates grouped together with the *C. gloeosporioides* epitype, suggesting that this strain does not affect coffee hosts in Thailand. Secondly, they uncovered and characterized three new species from the 'gloeosporioides' complex, *C. asianum*, *C. siamense* and *C. fruticola*, which were more closely related to *C. kahawae* than was the *C. gloeosporioides* epitype. Despite insightful, this study was limited to the geographic location of Thailand, where *C. kahawae* does not even occur, and may not have a representative sampling of *C. gloeosporioides sensu lato* from coffee hosts worldwide, particularly in the African continent where *C. kahawae* has emerged.

In this way, although the proposed hypotheses for *C. kahawae* emergence remain largely untested, the recent advances on *Colletotrichum* systematics, with the epitypification of *C. gloeosporioides* and the use of multi-locus datasets, as well as a more comprehensive understanding of the limitation and challenges of the current taxonomy, are creating a solid starting point to unravel this issue. Knowing the specific origin of *C. kahawae* would have deep implications in our knowledge of CBD since it would open the possibility to discover what events and changes led to the emergence of such a virulent pathogen amongst the *C. gloeosporioides* complex.

3.3 Population Studies on C. kahawae

Population genetics and molecular phylogeography are fields that can hold answers to important questions concerning the evolution of pathogens and the diseases caused by them, through the analysis of the amount and distribution of genetic variation within and among populations. The population genetic structure is determined by their evolutionary history and can give insights not only into the processes that shaped populations in the past, but also into the future evolutionary potential of the pathogen [96]. This knowledge can then be applied to optimize the management of resistance genes

and/or fungicides and maximize their useful life expectancy, minimizing the losses that result from the loss of efficiency of such control measures [17].

Vegetative compatibility and pathogenicity tests are often employed in population studies of *C. kahawae* and despite useful to identify and characterize distinct groups of isolates, they constitute indirect measures of the underlying genetic variability. Analysis of the genetic variation provides a more direct measure but has not been as used and explored as the two former methods.

3.3.1. Vegetative compatibility groups

Vegetative compatibility is a fungal biological process by which two different fungal haploid strains fuse their hyphae in their contact zones and form vegetative heterokaryons, i.e. hyphae with the presence of two haploid nuclei [97]. Strains that are compatible with one another are described as members of the same vegetative compatibility group (VCG). Otherwise, vegetative incompatible strains are unable to form heterokaryons with each other and are described as belonging to different VCG [97]. This compatibility acts to restrict transfer of nuclear and cytoplasmatic elements during vegetative growth and is controlled genetically, albeit independently from sexual reproduction [98].

In several fungal species, VCGs have been applied with varying degrees of success in characterizing isolates into pathotypes or geographic groups [99-101]. However, the most recent comprehensive study in *C. kahawae* using a wider sampling and mutants, found only one main VCG, supporting the existence of only one population but with indications of some geographical specialization [102].

3.3.2. Pathogenicity variation

Pathogenic variation of *C. kahawae* is usually assessed through pathogenicity tests on hypocotyls or detached berries, and is commonly used as it can have practical consequences for disease control and management. In such tests, small but significant differences in the aggressiveness of isolates from different geographic region are often revealed [41,56,103,104], and some authors further suggested the existence of physiological races [111]. Unfortunately, these studies were fragmented and sampled *C. kahawae* isolates from different geographic combinations and thus, tended to reach to different conclusions on the most aggressive strains. A more comprehensive study with a broad sampling of fungal isolates is still required.

3.3.3. Genetic variation

Until now, research on the genetic variation of *C. kahawae* populations has been a daunting and challenging task. In 1993, Sreenivasaprasad *et al.* carried the first study using a wide range of molecular techniques, such as rDNA and mtDNA RFLP, RAPDs and ITS locus sequencing, on a relatively small sampling of *C. kahawae* and *C. gloeosporioides* isolates, revealing an absence of genetic variation among *C. kahawae* isolates [106]. Strikingly, since 1993 the same results have been obtained by several subsequent authors, using the same and additional methods, and sampling *C. kahawae* isolates from several geographic locations [41,104,107]. For that reason the species began to be regarded as a genetically homogeneous population with a common origin and subsequent clonal dispersal throughout its geographic range. Only recently, additional and more sensitive markers for detecting polymorphisms within species, such as AFLP, were employed and unraveled the existence of a small amount of variation [32]. Despite the low sampling and subtlety of the differences, the authors suggested the existence of geographically differentiated populations based on the divergent AFLP banding pattern of three isolates from Cameroon and Malawi, from the remaining sample. Using the ITS sequence region, Manuel *et al.* also found slight differences among isolates from Angola and between Angola and other east African countries [42].

Nonetheless, population genetic studies on *C. kahawae* have revealed the species as a nearly clonal population with insufficient genetic variation to draw any inferences on its origin or dispersal patterns.

Overall, VCG, pathogenicity and genetic variation suggests the existence of some genetic variation, but no global and consistent structuring has been clearly revealed yet. Sequencing represents a good alternative to find polymorphic loci [96] and despite the employment of several molecular techniques, the analysis of nucleotide sequence variation from a multi-locus dataset in *C. kahawae* remains largely unexplored. Moreover, this approach has showed promising results in resolving phylogenetic relationships and phylogeographical patterns below species level, when other techniques are not so well succeeded [108-113].

3.4 Mating and Reproduction

In the *Colletotrichum* genus, as in other fungi in general, there is a nomenclature distinction between the sexual (teleomorph) and asexual (anamorph) stages of a species: *Glomerella* represents the teleomorphic stage, while *Colletotrichum* represents the anamorphic stage. Classification as

Colletotrichum or *Glomerella* is often based on the ability of single-spore cultures to produce perithecia [114], a sexual structure where meiosis and ascospores formation occur [98]. Following this description, strains designated by their anamorphic name are either presumably incapable of sexual reproduction or simply represent the imperfect stage of the fungus [114]. However, this morphological distinction is far from conclusive of the sexual reproduction capability of a fungus as it might be either rare or more likely to occur under unknown special conditions [115-117].

C. kahawae has no known sexual stage and is described as an asexual fungus, based on morphological criteria and field observations [35]. In fact, after the first studies demonstrated lack of variability in the limited number of loci sampled, it was assumed that this was a case of a true asexual pathogen [32]. However, it has been showed that the emergence of clonal populations is uncoupled with the capacity of a fungus to undergo, or not, sexual reproduction [116,118]. Nearly all of the fungi studied show recombining population structures in addition to clonality and the source of recombination can be other than sexual reproduction, such as parasexuality or mitotic crossing-over [116,119]. The same survey also showed that whenever genuinely asexual lineages occur, they do not persist for evolutionary meaningful lengths of time and therefore, are extremely rare [116].

In a wide range of fungi, complex developmental traits such as cell identity, morphogenesis and sexual development are controlled by the mating type (MAT) loci [120]. While the vast majority of sexually reproducing organisms occur as just two sexes or mating types, transitions in sexuality from two to multiple mating types, and vice versa, have occurred in the fungal kingdom [121]. Since the great majority of Ascomycetes have two mating types and one locus [122], focus has henceforth been given on this system. Based on general Ascomycete case studies, the single MAT locus (MAT1) is bipolar: exists as two alternate forms, defining two mating types (MAT1-1 and MAT1-2) [122,123]. In compatible heterothallic matings, two different strains, each bearing one of the two idiomorphs, are required. In homothallic interactions, a single fungal strain carries copies of both idiomorphs, either linked at the MAT1 locus, or, less commonly, in another area of the genome at a second locus MAT2. The alternate forms at the single MAT locus are called “idiomorphs” rather than alleles since they lack similarity to each other and do not normally recombine. Each idiomorph consists of a unique DNA sequence flanked by almost identical sequences between chromosomes [122,124]. Further molecular analysis revealed the presence of DNA-binding motifs in the protein, suggesting that they encode master transcription regulators [122]. These binding motifs allow the two forms to be distinguished: one idiomorph, MAT1-1, encodes a protein with an alpha-domain, while the other, MAT1-2, encodes a protein with a high-mobility group (HMG) [98]. Through gene disruption it was revealed that these

genes are strictly needed for sexual reproduction and meiosis, though in some asexual fungi loss of sexual reproduction seems to be related with other important genes [122]. Currently, there is an ongoing debate regarding MAT functions besides those reproduction-related. However, their mutually exclusive character in heterothallic species with completely different translated proteins harboring different binding motifs is a strong argument against the retention of vital functions for fungi survival other than those related to reproduction [122].

However, the *Colletotrichum* genus has been documented as an apparent deviation from this system [125,126]. Studies in some species of the genus with known sexual stage such as *Glomerella lindemuthiana* (anamorph: *Colletotrichum lindemuthianum*), *Glomerella graminicola* (anamorph: *Colletotrichum graminicola*) and *Glomerella cingulata* (anamorph: *Colletotrichum gloeosporioides*) support the idea that the standard ascomycete configuration of two idiomorphs at the MAT locus does not hold true [126]. Individuals from these species can be both homothallic and heterothallic and segregation studies suggest that a single locus with multiple forms controls mating in *G. cingulata* and two unlinked loci control mating in *G. graminicola*. [114,125-127]. Most of these conclusions are based on crosses and segregation patterns of genetic markers but few studies have directly analyzed MAT genes. One of these few studies, in *G. lindemuthiana*, reported that in both parental strains as well as in the analyzed progeny, sequences homologous to the HMG box of the MAT1-2 idiomorph were always found and named *Mat1-2-1* [128]. Recently, in an extensive sampling of 11 graminicolous *Colletotrichum* species, only a single idiomorph, *Mat1-2-1*, was identified in the ~30 kb of the MAT locus gene cluster analyzed, regardless of the reproductive lifestyle [129]. On the other hand, the MAT1-1 idiomorph was never revealed. Apparently, the *Colletotrichum* genus has a unique mating strategy in need of empirical data to unravel its mechanism.

Several population studies have already been carried out using MAT genes sequencing in a wide array of fungi with the standard homothallic or heterothallic mating systems [130-133]. These studies revealed that MAT genes are under intense purifying selection in individuals from the same species due to their importance in mating and nuclei recognition [130]. Even putative asexual species showed the presence and high conservation of MAT genes [134,135]. Further research revealed that this selection was not homogeneous through the entire gene and that some regions seemed to be under a more neutral selection, launching an ongoing debate [136].

However, until now there are no reports of population studies using the complete MAT1-2 gene on *Colletotrichum* species. Yet, the acquisition of such knowledge would expand our understanding not

only of the selective constrains operating on this gene but also of the asexuality of *C. kahawae*. Many reports on other putative asexual fungi revealed that analysis of MAT genes provided evidence that led to a shift in their sexual status [116,126,137]. The conservation of the nucleotide sequence as well as protein conformation in asexual species when compared with closely related sexual species gave a strong support to question their true sexual nature. The same approach could be used in the *C. kahawae*/*C. gloeosporioides* complex, though with the additional caution that the mating system is apparently not the same and that the *Mat1-2-1* gene described may not have the exact same function as its homologous in other fungi species.

Part II

Application of the Apn1/MAT locus to improve the systematics of the *Colletotrichum gloeosporioides* complex: An example from coffee (*Coffea* spp.) hosts

Diogo Nuno Silva [1,2*], Pedro Talhinhas [1], Vítor Várzea [1], Lei Cai [3], Octávio Salgueiro Paulo [2], Dora Batista [1]

[1] *Centro de Investigação das Ferrugens do Cafeeiro (CIFC) /Instituto de Investigação Científica Tropical (IICT), Quinta do Marquês, 2784-505 Oeiras, Portugal.*

[2] *Computational Biology and Population Genomics group, Environmental Biology Center, Animal Biology Department, Faculty of Science, University of Lisbon, Campo Grande, P-1749-016 Lisbon, Portugal*

[3] *Key Laboratory of Systematic Mycology & Lichenology, Institute of Microbiology, Chinese Academy of Sciences, No.10, North 4th Ring Rd West, Beijing 100190, People's Republic of China*

*Corresponding author: Diogo Nuno Silva. e-mail: diogo_nuno_silva@hotmail.com; Address: Centro de Investigação das Ferrugens do Cafeeiro (CIFC)/ Instituto de Investigação Científica Tropical (IICT), Quinta do Marquês, 2784-505 Oeiras, Portugal; Telephone: +351214544680

One of the most longstanding issues in *Colletotrichum* taxonomy has resided in how to properly address species complexes. *C. gloeosporioides* is one such complex, and hitherto the most challenging one due to its wide morphological, biological and host range. Systematics of this complex, and also of *Colletotrichum* in general, are now witnessing a shift from the morphological and cultural based species identification, towards molecular phylogenetic approaches with multi-locus datasets. These approaches show promise in recognizing stable and well resolved species within complexes, but their usefulness is limited to the information that molecular markers are able to convey. To improve and expand our limits in this regard, we have successfully developed and applied a new set of molecular markers from the Apn1/MAT locus, to a case study of 22 isolates, mostly from coffee hosts, representing five well characterized species formerly within the *C. gloeospoiroides* complex. Markers from the new locus revealed an outstanding informative potential when compared to other commonly used molecular markers such as ITS, β-tub2 and GS. Among those, the ApMAT marker alone was almost as informative in terms of phylogenetic resolution as the seven gene concatenated dataset. However, results further reveal that gene tree discordance may come to be a common issue in the process of species delimitation in the *C. gloeosporioides* complex, highlighting the importance of multi-locus approaches. The use of state-of-the-art data analysis techniques and a highly informative datasets as employed here may abate this issue and hopefully assist in disentangling the *C. gloeosporioides* complex.

Keywords: (Species phylogeny – Incomplete lineage sorting – Bayesian Estimate of Species Tree (BEST) – Mating type (MAT) locus)

Introduction

Disentangling species complexes in the *Colletotrichum* genus remains a considerable challenge for taxonomists in the years to come. Given the worldwide importance of this genus constituted mostly of plant pathogens, it became essential to accurately identify species and/or pathotypes to improve biosecurity and disease control (Cai *et al*, 2009; Johnston and Jones, 1997; Freeman *et al*, 1998). Morphological, cultural and host preference criteria have been the primary basis for species identification and delimitation but due to their unreliability and limited number of diagnosable characteristics they have led to longstanding and unresolved taxonomic issues (Cannon *et al*, 2000; Du *et al*, 2005; Hyde *et al*, 2009a; Afanador-Kafuri *et al*, 2002). As Sutton (1992) noted, morphology alone is unlikely to provide enough information to improve systematic of the *C. gloeosporioides* species complex.

C. gloeosporioides is regarded as the most challenging species complex to resolve, comprising the broadest host range of all *Colletotrichum* species (Du *et al*, 2005). Fungal strains from this complex were already reported in at least 1800 plant species and present such a wide range of morphological and pathogenic variation that the species name is of limited practical use (Phoulivong *et al*, 2010; Du *et al*, 2005). A major contribution for the emergence of this complex was the revision of the genus by von Arx (1957) in which an artificially enlarged spore range was placed as a criterion for *C. gloeosporioides* identification. As this criterion was followed by researchers, many *Colletotrichum* strains with similar cylindrical conidia were identified as *C. gloeosporioides*, lumping together potentially distinct taxa (Chakraborty *et al*, 2002; Freeman *et al*, 2000; Hindorf *et al*, 1970; Munaut *et al*, 1998; Xiao *et al*, 2004, Hyde *et al*, 2009b). Since the last major revision of the *Colletotrichum* genus (Sutton, 1992), most studies on *C. gloeosporioides* systematics have addressed only specific hosts (e.g., Abang *et al*, 2002; Lubbe *et al*, 2004; Xiao *et al*, 2004; Suzuki *et al*, 2010), frequently moving pathogens to other species, such as *C. acutatum* (e.g., Brown *et al*, 1996; Martín and Garcia-Figueres; 1999, Sreenivasaprasad *et al*, 1994) and *C. boninense* (e.g., Ramos *et al*, 2006), among others. The complexity of *C. gloeosporioides* genetic structure and the breath of hosts it inhabits hampers studies addressing the entire complex, which remains mainly delimited by Sutton (1992) broad criteria, in practical terms supported by simple-to-use tools such as the ITS-based specific PCR primers developed by Mills *et al* (1992).

On the long road towards a comprehensive understating of this complex, the recent epitypification of *C. gloeosporioides* (Cannon *et al*, 2008) was an important first step. Ever since, the availability of living type strains and sequence data has provided a solid reference basis to which isolates can be compared in future research. The criterion to consider a genetic entity as a species has been subject of much debate throughout the years and the genus *Colletotrichum* has been an example of such (Cannon *et al*, 2008). Currently, the diversity threshold considered sufficient to define a species has been lowering in *Colletotrichum*, supported by molecular data. An example of this is the division of the monophyletic species *C. acutatum* into five species (Shivas and Tan, 2009), based on previous work characterizing the genetic diversity of the "acutatum" cluster (Sreenivasaprasad and Talhinhas, 2005). Polyphyletic taxa such as *C. gloeosporioides* are following this trend. For instance, Prihastuti *et al* (2009) were able to identify three new distinct species on coffee hosts (*Coffea* spp.) in Thailand from the *C. gloeosporioides* complex, using a multi-locus phylogenetic approach. Likewise, Phoulivong *et al* (2010) revealed that the typified *C. gloeosporioides* does not seem to occur in tropical fruits of Thailand and Laos, as previously thought.

The emergence of molecular systematics using multi-locus phylogenies in combination with morphological and cultural characters, is revealing to be a more effective way to resolve species complexes and also to understand species evolutionary relationships (Cai *et al*, 2009; Crouch *et al*, 2006; Crouch *et al*, 2009a; Prihastuti *et al*, 2009; Taylor *et al*, 2000; Damm *et al*, 2009). Even though single gene phylogenies, mainly using the Internal Transcribed Spacer (ITS) region, have been frequently applied over the last decades to the 'gloeosporioides' complex (Sreenivasaprasad *et al*, 1996; Sreenivasaprasad *et al*, 1993; Johnston and Jones,

1997), their resolving ability has been fairly limited and sometimes casted even more confusion (Cai *et al*, 2009; Crouch *et al*, 2009b). The limitations of ITS are already being well recognized and thus, the development and use of more informative loci has become increasingly necessary if one aims at unraveling the existent distinct species and their relationships among the *C. gloeosporioides* complex.
For this purpose, we developed and used several markers from the Apn1/MAT locus to improve systematic resolution and knowledge. As previously shown by Crouch *et al* (2009a), markers from this locus had an outstanding performance in resolving the graminicolous *Colletotrichum* group, although they have been impossible to apply to other groups, due to their highly variable nature. Since the 'gloeosporioides' complex is huge and unlikely to be settled in one sitting, we have further exploited this region for marker development and assessed their usefulness when compared to other commonly used nuclear genes or regions, focusing on a well characterized group of species from coffee hosts. These included the known pathogen, *C. kahawae*, and other opportunistic pathogenic species recently characterized by Prihastuti *et al* (2009), as well as type representatives of *C. gloeosporioides*. In addition, considering that molecular markers from different loci are prone to reveal incongruent species relationships due to incomplete lineage sorting or deep coalescence resulting from their recent divergence (Degnan & Rosenberg, 2009; Edwards *et al*, 2008), we also applied a recent Bayesian data analysis technique for species tree reconstruction and compared it to the routinely used method of concatenation.

Materials and Methods

Fungal material and DNA extraction

The list of *Colletotrichum* isolates used in this study is provided in Table 1. Nineteen isolates were obtained from the collection maintained at CIFC/IICT, Portugal, and three from the Mae Fah Luang University, Thailand, representing 6 previously characterized species from the *C. gloeosporioides* complex: *C. fragariae*, *C. gloeosporioides*, *C. kahawae*, *C. asianum*, *C. siamense* and *C. fructicola*. For the four later species, type specimens were available and employed. On the other hand, the *C. gloeosporioides* sample was constituted by isolates which are similar to the described epitype at the ITS and *β-tubulin 2* loci (GenBank accession numbers: EU371022 and FJ907445, respectively). Likewise, Blast analysis of the MAT1-2 HMG locus enabled the identification of five studied isolates as *C. fragariae*, by comparison with the species epitype (GenBank accession number: DQ002827). These *C. fragariae* isolates were used as outgroup taxa, since recent studies show this species as the most basal lineage of the *C. gloeosporioides* complex (Cai *et al*, 2009). Except for *C. gloeosporioides* isolates, which were isolated from *Olea europaea* and *Citrus lemon*, all others were obtained from coffee hosts (*Coffea* spp.). Isolates were revived on Malt Extract Agar 1% (MEA) with a bacterial inhibitor (KCNS, 50mM) and grown for 5-7 days. Isolates were then grown in liquid media containing Malt Extract (3%) and Peptone (0.5%) for 12-14 days at 25ºC in the dark and DNA was extracted from freeze-dried mycelium with the DNeasy plant Minikit (Qiagen, Hilden, Germany) according to manufacturer's instruction.

Apn1/MAT locus marker development strategy

A novel set of specific primers was developed to amplify a portion of the Apn1/MAT locus (Fig. 1). This locus comprises two genes: *Apn1* (~2,244bp) and *MAT1-2-1* (842bp), separated by an intergenic region of ~713bp. A representative sample of 10 isolates, comprising the six species, was used to amplify and sequence the full portion of the locus. Primers were designed using PerlPrimer v1.1.18 (Marshall, 2004). The first two sets of specific primers (AM-F/AM-R and M5L-F/M5L-R)(Fig. 1) were designed within and flanking the conserved HMG domain of *MAT1-2-1*, using a single *C. gloeosporioides* accession in GenBank

Table 1 *Colletotrichum* isolates used in this study

Isolate	***Species or undetermined group***	***Host***	***Origin***	***GeneBank Accession Numbers***						
				ITS	***b-tub2***	***GS***	***Apn15L***	***ApMAT***	***MAT1-2-1***	***MAT5L***
Que2*	***C. kahawae***	*C. arabica*	Kenya	---	---	---	---	---	---	---
Tan12	*C. kahawae*	*C. arabica*	Tanzania	---	---	---	---	---	---	---
Mal2	*C. kahawae*	*C. arabica*	Malawi	---	---	---	---	---	---	---
Bur2	*C. kahawae*	*C. arabica*	Burundi	---	---	---	---	---	---	---
PT111	***C. gloeosporioides***	*Olea sp.*	Portugal	---	---	---	---	---	---	---
PT21*	*C. gloeosporioides*	*Citrus sp.*	Portugal	---	---	---	---	---	---	---
PT220	*C. gloeosporioides*	*Olea sp.*	Portugal	---	---	---	---	---	---	---
BDP-I 2*	***C. siamense***	*C. arabica*	Thailand	---	---	---	---	---	---	---
Mal5*	*Group B*$^{+}$	*Coffea sp.*	Malawi	---	---	---	---	---	---	---
CCM5	*Group B*$^{+}$	*Coffea sp.*	Kenya	---	---	---	---	---	---	---
Bra9*	*Group B*$^{+}$	*Coffea sp.*	Brazil	---	---	---	---	---	---	---
Col1	*Group B*$^{+}$	*Coffea sp*	Colombia	---	---	---	---	---	---	---
Chi4*	*Group B*$^{+}$	*Coffea sp.*	China	---	---	---	---	---	---	---
BDP-I 4*	***C. asianum***	*C. arabica*	Thailand	---	---	---	---	---	---	---
BDP-I 16*	***C. fruticola***	*C. arabica*	Thailand	---	---	---	---	---	---	---
Ang40	*Group C*$^{+}$	*Coffea sp.*	Angola	---	---	---	---	---	---	---
Ang97	*Group C*$^{+}$	*Coffea sp.*	Angola	---	---	---	---	---	---	---
Bra8	***C. fragariae***	*Coffea sp.*	Brazil	---	---	---	---	---	---	---
Bra5*	*C. fragariae*	*Coffea sp.*	Brazil	---	---	---	---	---	---	---
Ang52*	*C. fragariae*	*Coffea sp.*	Angola	---	---	---	---	---	---	---
Ang91	*C. fragariae*	*Coffea sp.*	Angola	---	---	---	---	---	---	---
Ang84	*C. fragariae*	*Coffea sp.*	Angola	---	---	---	---	---	---	---

* Isolates included in the representative sample for sequencing of the full Apn1/MAT locus; $^{+}$ Isolates received as *C. gloeosporioides*.

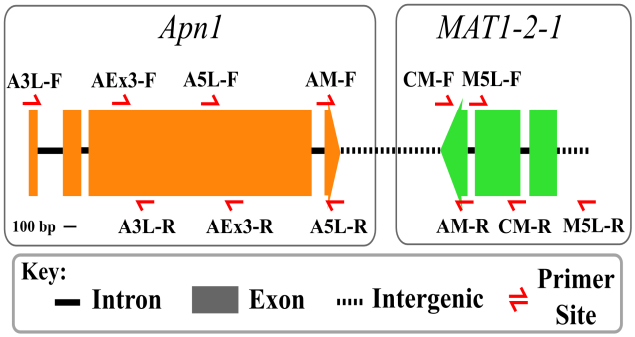

Fig. 1 Apn1/MAT locus showing the position of the primers used in this study. The arrow shaped end of the genes indicates the direction of the transcription

(AY357890) as a template, which contains 11,592bp of sequence including the Apn1/MAT locus and flanking regions. The following conditions were used in PCR amplification: 3min at 94ºC followed by 30 cycles of 45s at 94ºC, 45s at 62ºC and 1min at 72ºC, with a final extension of 7min at 72ºC. The amplification products were sequenced and used as templates, along with the *C. gloeosporioides* accession, for further primer design. Based on these, primer pairs CM-F/CM-R and A5L-F/A5L_R were designed to amplify the remainder *MAT1-2-1* gene and the 5' end of the *Apn1* gene, respectively (Fig. 1). PCR conditions were identical except for the use of a 66ºC annealing temperature for the CM-F/CM-R pair. With the obtained sequences from the 5' end of *Apn1* used as templates, primers AEx3-F/AEx3-R, and subsequently primers A3L-F/A3L-R, were designed to amplify the remaining portion of the *Apn1* gene (Fig. 1). Amplification reactions were performed as follows : 3min at 94ºC followed by 30 cycles with 45s at 94ºC, touchdown annealing step for 45s starting at 62ºC and decreasing 0,5ºC per cycle until stabilizing at 57ºC during the remaining 20 cycles, 1min at 72ºC and a final extension of 7min at 72ºC. As described in Table 2, the nomenclature of the molecular markers developed is as follows (from the 3' of *Apn1* and onwards): Apn13L, with A3L-F/A3L-R; Apn1Ex3, with AEx3-F/AEx3-R; Apn15L, with A5L-F/A5L-R, ApMAT, with AM-F/AM-R; *MAT1-2-1* and MAT5L, with CM3-F/M5L-R.

Molecular data

Using the complete sampling of this study, five nuclear genes and two intergenic regions were sequenced. From the Apn1/MAT locus described above, the two intergenic regions (ApMAT and MAT5L) as well as the *MAT1-2-1* gene and the Apn15L gene fragment were selected. The complete rDNA Internal Transcribed Spacer (ITS) (Brown *et al*, 1996) nuclear region and two gene fragments from *β-tubulin2* (β-tub2) (O'Donnel and Cigelnik, 1997) and *Gluthamine syntethase* (GS) (Prihastuti *et al*, 2009), were also analyzed, using primers and PCR conditions previously described. PCR products were verified by staining with ethidium bromide (EtBr) on 1.2% agarose electrophoresis and purified using SureClean (Bioline). Sequencing reactions were carried out using the BigDye version 3.1 chemistry (Applied Biossystems) on an ABI prism 310 automated sequencer. Amplicons were sequenced in both directions and chromatographs were manually checked for errors in SEQUENCHER v4.0.5 (Gene Codes Corporation).

Data analysis

Datasets for each marker were aligned in MAFFT v6.717b (Katoh *et al*, 2009) using the L-INS-i method, followed by manual refinement in BioEdit v7.0.5.1. Using the complete sampling, the nucleotide diversity (π) and its standard deviation were estimated for each selected marker in DNasp v5 (Rozas *et al*, 2003). Additionally, in order to compare the polymorphism between and among all used markers in the representative sample, a sliding window analysis of nucleotide diversity was performed, with a window length of 50bp and step size of 25bp. Sites with alignment gaps were not considered in the length of the windows and all estimations were performed using the standard parameters of the program. Output graphics were produced on the R package v2.10.1 using the ggplot2 library.

To create concatenated matrixes from the individual datasets, the Concatenator software was used (Pina-Martins and Paulo, 2008). Three combinations were assembled for subsequent comparative purposes: the Apn-MAT dataset with the combination of only Apn1/MAT locus makers, except *MAT1-2-1*;

Table 2 Primer pairs used to amplify the Apn1/MAT locus, according to Fig. 1

Locus Region	Primers	Sequence 5'-3'	Amplicon size (bp)
Apn13L	A3L-F	TGACATGGAACGGTGAGTG	740
	A3L-R	TTCCAGTCCTCGACCGTCA	
Apn1Ex3	AEx3-F	CTCATCGGCACCTACAGC	840
	AEx3-R	CGGAGACATCTTGCTCGTG	
Apn15L	A5L-F	CAAGCGACGAAGTATACGAG	900
	A5L-R	GCATCACGGGAATAACTAGG	
ApMAT	AM-F	CCAGAAATACACCGAACTTGC	910
	AM-R	TCATTCTACGTATGTGCCCG	
MAT1-2-1/ MAT5L	CM-F	TCTACCTCATCGACGCTGCT	550
	CM-R	CATGTGGGCAAAGGATGGC	
	M5L-F	ATCTTTGCGGTAGAGAATGAAGG	830
	M5L-R	GACCCTTCTATGAACGAGCC	

the Traditional dataset, with the combination of ITS, β-tub2 and GS; and the total combined dataset. *MAT1-2-1* was not included in the Apn-MAT dataset in order to balance the sequence length of both Apn-MAT and Traditional datasets, providing a more even comparison. Phylogenetic trees were constructed from the individual and combined analyses of the seven nuclear regions using Maximum Likelihood (ML) and a Bayesian framework with Markov Chain Monte Carlo (BMCMC). Gaps were treated as missing data. The ML analysis was run in PAUP* v4.0d99 (Swofford and Begle, 1993) with heuristic searches of 100 replicates with random sequence addition and a Tree-Bisection-Reconnection (TBR) branch swapping algorithm. Nonparametric bootstrapping was conducted using 1000 pseudoreplicates with 10 random additions and TBR branch swapping. ModelTest v3.7 (Posada and Crandall, 1998) was used to select the best fit model of nucleotide evolution, under the Akaike Information Criterion (AIC), for each dataset. The BMCMC analysis was run in MrBayes v3.1.2 (Ronquist and Huelsenbeck, 2003) with the optimal model selected under the AIC, as implemented in MrModelTest v2.3 (Nylander, 2004), specified as prior for each partition. For the individual and combined datasets, Bayesian posterior probabilities were generated from $3x10^7$ and $1x10^8$ generations respectively, sampling at every 1000^{th} generation. The analysis was run three times with one cold and three incrementally heated Metropolis-coupled Monte Carlo Markov chains, starting from random trees. $1x10^6$ generations were discarded as a burn-in. Trees were then combined and summarized on a 50% majority-rule consensus tree.

The total combined dataset was also analyzed using the species tree approach implemented in BEST (Liu, 2008) to incorporate the signals of each marker. BEST has been shown to deal with the common issue of deep coalescence in recently diverged species. It uses a hierarchical Bayesian approach to estimate the joint posterior distribution of multiple gene trees and is known to provide more accurate estimates than the concatenation method in some multi-locus datasets (Liu and Pearl, 2007; Edwards *et al*, 2007). The BEST analysis was run in MrBayes v2.3 and model priors for each partition were estimated as in the BMCMC analysis. Priors included an inverse gamma distribution (3, 0.003) for theta and a uniform distribution (0.2, 2) for gene mutation, and were estimated as part of the analysis. A total of $2.5x10^9$ generations were sampled at every 2000 generations. Convergence and mixing were assessed for all parameters using Tracer v1.4 (Drummond and Rambaut, 2007). The species tree was constructed from the combined runs.

Topology tests

The current taxonomic relationship of the studied species was also assessed through topology tests on each relevant dataset. Alternative hypothesis were tested against the most recent hypothesis of relationships between our species of interest (Phoulivong *et al*, 2010) using the topological test of Shimodaira & Hasegawa (Shimodaira and Hasegawa, 1999) as implemented in PAUP* v4.0d99 (Swofford and Begle, 1993). Although, this current taxonomy was not formally addressed, it provided an objective starting point for our analysis. First, we tested the unconstrained phylogenies of each dataset that revealed discordant, to assess if their branching order was significantly incongruent with the current taxonomy. In addition, since *C. siamense* received only moderate support in the current taxonomy, we also tested the alternative hypothesis of its monophyly to the second closest taxa, *C. asianum* (i.e. [[*C. siamense*, *C. asianum*], *C. fructicola*], instead of the current taxonomy [[*C. siamense*, *C. fructicola*], *C. asianum*]) for congruent datasets. One thousand replicates were performed by resampling the partial likelihoods for each site (RELL model).

Results

Molecular markers analysis

We have successfully developed and tested a new set of specific primers that allowed the amplification of ~4061bp from the Apn1/MAT locus. Each primer pair produced partially overlapping fragments that, in combination, spanned all but the first 21 codons of *Apn1*, the intergenic region between *Apn1* and *MAT1-2-1*,

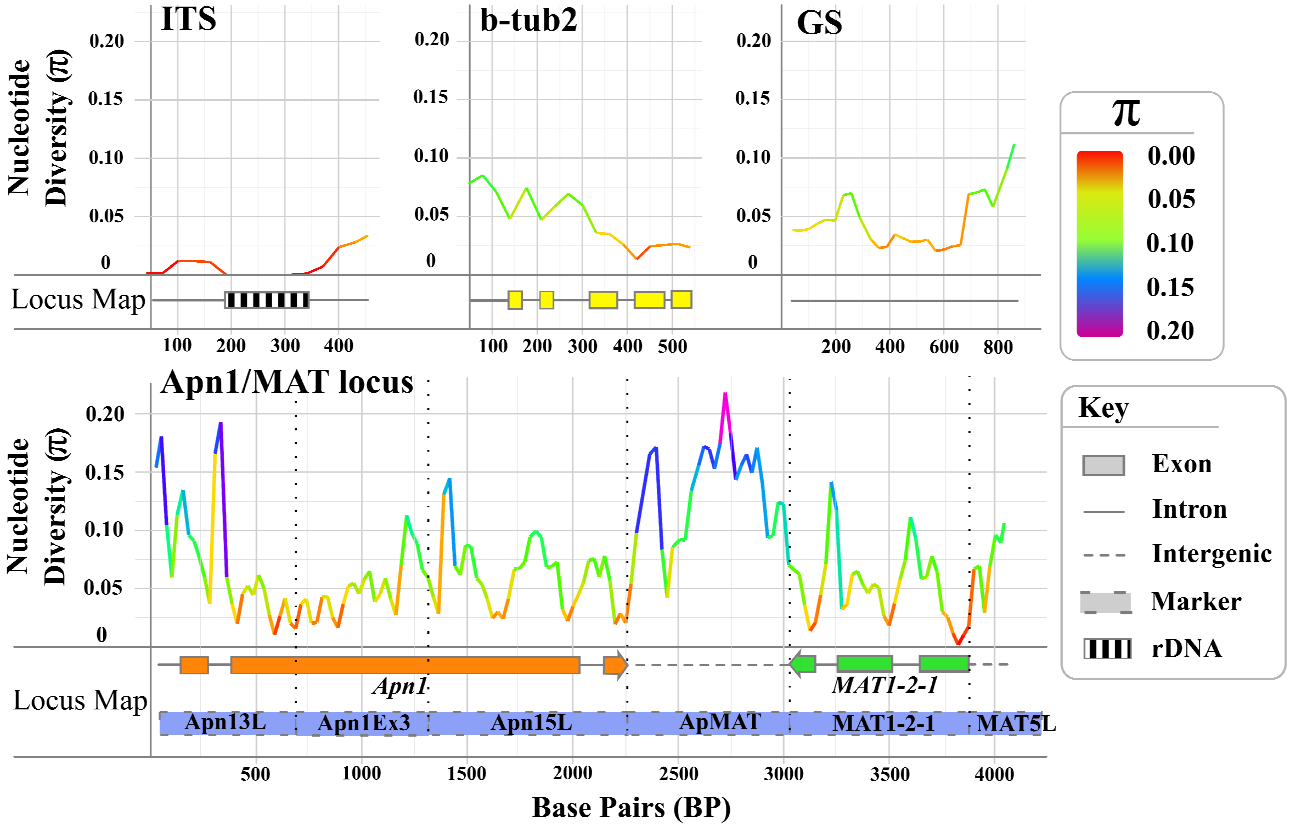

Fig. 2 Sliding window analysis of the nucleotide diversity (π) estimated from the molecular markers used, based on a 10 isolates representative sample. The window length is 50 and the step size is 25 bp. For each locus, below the sliding window graphic, a schematic representation of gene, intergenic or rDNA segments along their length is presented. In the Apn1/MAT locus graphic, the vertical dot lines represent the molecular markers boundaries. Graphics were produced and adapted from DNasp v5 estimates using the R package.

Table 3 The length (bp), number and percentage of parsimony informative sites (PI;%PI), estimated nucleotide diversity (π) and Model of sequence evolution for each dataset using the complete sample

Data	Length (bp)	PI	%PI	π	Model
Individual datasets					
ITS	489	14	**2.86%**	0.0077	TrNef+I
b-tub2	578	77	**13.32%**	0.0436	TIM+G
GS	889	141	**15.86%**	0.0433	K81uf+I
Apn15L	883	168	**19.03%**	0.0559	TrN+G
ApMAT	772	268	**34.72%**	0.1282	HKY+G
MAT1-2-1	843	144	**17.08%**	0.0642	TIM+G
MAT5L	213	42	**19.72%**	0.0512	TIM+I
Combined datasets					
ApnMAT	1886	478	**25.34%**		
Traditional	1965	232	**11.81%**		
Total	4676	801	**17.13%**		

the complete *MAT1-2-1* gene and a small intergenic region adjacent to the 3' end of *MAT1-2-1* (Fig. 1). As in previous *Colletotrichum* studies, only one mating-type (MAT) locus idiomorph, *MAT1-2-1*, was obtained (García-Serrano *et al*, 2008). Clean bands were reproducibly amplified for all the isolates, with the occasional exception of Apn13L, whose amplification success was intermittent. PCR amplification generated amplicons of ~740bp for Apn13L, ~760bp for Apn1Ex3, ~840bp for Apn15L, ~910bp for ApMAT and ~1040bp for *MAT1-2-1* and MAT5L. An overview of the Apn1/MAT locus map is presented in Fig. 2, along with the position and length of each developed marker, excluding overlapping segments. The ITS, β-tub2 and GS markers generated products of the expected size, and their locus map is also presented (Fig. 2). Sequences were collected and submitted to GenBank under the accession numbers described on Table 1.

The complete set of markers was sequenced for a representative sample of 10 isolates comprising the five studied species (*C. kahawae*, *C. gloeosporioides*, *C. asianum*, *C. fructicola* and *C. siamense*), plus outgroup (*C. fragariae*). This preliminary dataset was subject to a sliding window analysis to quantify nucleotide polymorphism along the sequence of each marker and to provide a comparative exploratory analysis between markers (Fig. 2). The commonly used ITS region had the lowest overall nucleotide diversity (π = 0.0077), since it was mostly invariable, and should provide little usefulness for systematic analysis. The intron rich β-tub2 (π = 0.0436) and GS (π = 0.0433) gene fragments, presented a significant improvement but are still constrained by regions of low polymorphism. The greatest improvement, was provided by the Apn1/MAT locus as even the most conserved marker of this locus, Apn1Ex3, revealed higher polymorphism (π = 0.0467) than β-tub2 and GS. However, the most striking result was found for the intergenic region, ApMAT, which showed a very high nucleotide diversity (π = 0.1282), revealing a tremendous informative potential for the 'gloeosporioides' complex and related taxa. Its nucleotide diversity exceeded in more than two fold the diversity of b-tub2 or GS fragments, both in average and along the sequence. Moreover, this region is flanked by the relatively conserved regions on the final exons of *Apn1* and *MAT1-2-1*, which provides a suitable ground for primer development. The *MAT1-2-1* gene (π = 0.0512), although much more conserved in exon sequences, showed a considerable diversity on its introns, particulary the second in the HMG region. In the *Apn1* gene, the nucleotide diversity seems to be relatively high even in exon sequences, though it was highest in the Apn13L marker region (π = 0.0739) due to the presence of a pair of introns. The Apn15L region (π = 0.0559) also showed a considerable diversity, particularly along the exon sequence.

Taking into account the distribution of the nucleotide polymorphism along the locus as well as the reproducibility of the PCR amplification of the different markers, two gene fragments (Apn15L and *MAT1-2-1*) and two intergenic regions (ApMAT and MAT5l) were selected for further analysis. A total of 22 sequences were generated for seven nuclear markers (ITS, β-tub2, GS, ApMAT, Apn15L, MAT1-2-1 and MAT5L), totalizing 4676bp of sequence data for each isolate studied. Using parsimony and nucleotide diversity statistics as comparative benchmarks, all of the newly developed markers revealed to be more informative than the traditional markers (Table 2). The ApMAT, Apn15L, MAT5L and MAT1-2-1 were 35%, 19%, 20% and 17% informative, comparing with the 3%, 13% and 16% of the ITS, β-tub2 and GS, respectively (Table 2). Regarding the combined datasets, even excluding MAT1-2-1 to balance the sequence length of the two partially concatenated datasets, the Apn-MAT dataset exhibited 478 polymorphic sites (25%), while the tradi-

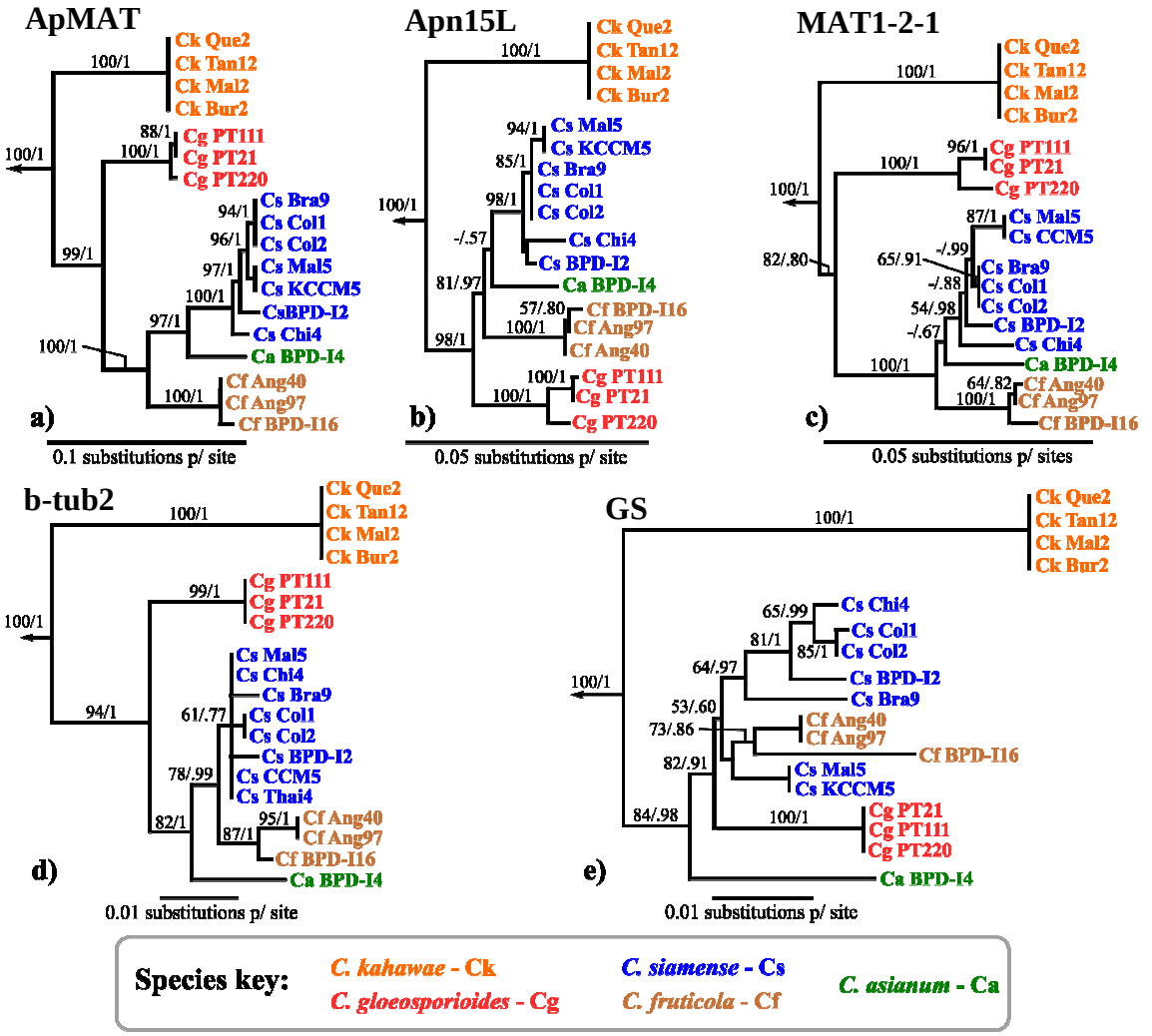

Fig. 3 ML estimated of gene trees (rooted with *C. fragariae*) for five individual datasets. The remaining ITS and MAT5L datasets provided a poor resolution and were not included. Bootstrap/Posterior values are shown above branches. The key to species/groups codes and colors is provided in the bottom.

tional markers contributed with 232 polymorphic sites (12%), less than half of the former combination (Table 2). Altogether, 801 (17%) nucleotide characters were informative using parsimony criteria.

Phylogenetic analysis

We first constructed phylogenies from each of the individual markers (Fig. 3). The ML and BMCMC analysis always resulted in identical topologies for each dataset, although not all gene trees had the same resolution. For example, ITS and MAT5L provided such poorly resolved trees that no significant topology could be recovered and thus, were not analyzed individually (data not shown). In a striking contrast, ApMAT provided an exceptional resolving ability with species/group and even intragroup nodes highly supported by both analyses (Fig 1a).

The *Colletotrichum* strains whose group was initially undefined were shown to cluster with the included species representatives, namely *C. siamense* and *C. fructicola* (Fig. 3). Generally, species/groups were reciprocally monophyletic, although with varying degrees of support. An exception was the GS gene tree, in which 2 isolates from the *C. siamense* group were monophyletic with *C. fructicola*. However, further incongruities can be found when comparing the relative taxonomic position of the ingroup species. Excluding *C. kahawae*, which was consistently revealed as the most anciently diverged lineage, all other species showed different taxonomic relationships in each of the three independent loci analyzed. *C. siamense* was found to

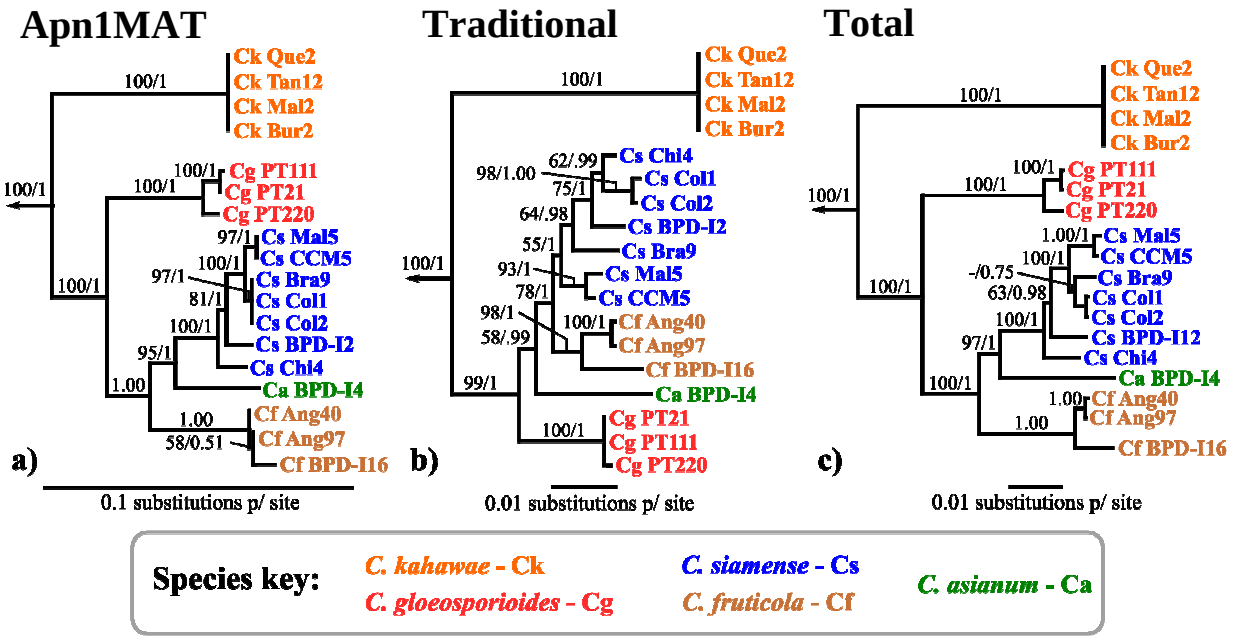

Fig. 4 BMCMC estimated of gene trees (rooted with *C. fragariae*) for the tree combined datasets. Bootstrap/Posterior values are shown above branches. Species/groups are highlighted with different colors: *C. kahawae*: Orange; *C. gloeosporioides*: Red; *C. siamense*: Blue; *C. asianum*: Green; *C. fruticola*: Brown

be monophyletic with either *C. asianum* (Fig. 3a-c) or *C. fructicola* (Fig. 3d,e) and the second most basal lineage obtained was either *C. gloeosporioides* (Fig. 3a-d) or *C. asianum* (Fig. 3e). Moreover, most of the discordant trees had their branching order fairly supported by bootstrap and posterior probability values. Consequently, the phylogenetic relationship of most species is not obvious based on the topologies of gene trees alone.

Phylogenetic reconstructions were then undertaken from the three combined datasets to further assess how these incongruities would affect the species relationship and their node support in true multi-locus datasets, and to compare their degrees of resolution. Overall, both the Traditional and Apn-MAT datasets provided almost completely resolved and highly supported trees. However, this means that they highly supported two different topologies, since *C. siamense* was again either reciprocally monophyletic with *C. asianum* (Fig. 4a) or *C. fructicola* (Fig. 4b). When all markers were combined, the prevailing phylogenetic relationships were congruent with the Apn-MAT dataset, supporting the reciprocal monophyly between *C. siamense* and *C. fructicola*. Interestingly, both the topology and node support of the single ApMAT marker were similar to the seven gene concatenated dataset, revealing the singularity of this region for systematic purposes in the 'gloeosporioides' complex.

Species tree approach and topology tests

The current taxonomic and evolutionary relationships of the included species from the 'gloeosporioides' complex, were further assessed by two statistical methods. First, the species tree approach of the BEST software was carried out, since it has the useful property of accounting for discrepancies between gene trees and uses the coalescent theory to combine those gene trees and infer a species tree. The obtained tree shared the same topology as the Apn-MAT and total concatenated datasets, albeit with a much smaller support for the reciprocal monophyly between *C. siamense* and *C. asianum* (Fig. 6). Nonetheless, the remaining relationships were highly supported.

For the topological tests, we assumed the null hypothesis of the most recently published species relationship, which is identical to that exhibited by the Traditional dataset (SH test, Table 3). Except for β-tub2, all individual datasets were incongruent, but only the branching order of ApMAT and MAT1-2-1 was significantly

deviated from the current taxonomic understanding (SH test, P = 0.035 and P = 0.000, respectively). Likewise, the combined Apn-MAT topology was significantly different (SH test, P = 0.000) but the deviation of the total combined dataset was only marginally significant (SH test, P = 0.076). We also tested whether the Traditional dataset presents a significant deviation from the alternative scenario, in which the true species relationship is as that recovered in the total dataset. Indeed, when the topology for this alternative scenario was constrained, it was not significantly different than the unconstrained phylogeny (SH test, P = 0.143).

Discussion

Resolving species complexes in the *Colletotrichum* genus, particularly the "*gloeosporioides*" complex, is a demanding quest in which molecular systematic tools are becoming of the utmost importance for taxonomists. However, genomic regions are not equally informative and thus, it is important to selected the most useful regions to address this issue. The Apn1/MAT locus has been successfully employed in phylogenetic studies in genera such as *Cochliobulos* (Turgeon, 1998), *Fusarium* (O'Donnell *et al*, 2004) and *Leptosphaeria* (Voigt *et al*, 2005). In our work, a newly developed set of molecular markers enabled the analysis of an enlarged region of the Apn1/MAT locus, revealing the potential to provide an outstanding improvement to face the incoming taxonomic challenges of the 'gloeosporioides' complex. The analysis conducted here focused in a small species group from coffee hosts belonging to a much wider complex of species but envisages demonstrating the potential of the developed markers and their application to resolve the phylogenetic relationships in the whole complex.

Regarding their informative potential, all of the new markers were superior to ITS and the two commonly used gene fragments, β-tub2 and GS, even though these have been useful and regarded as fairly informative in previous studies (Prihastuti *et al*, 2009; Talhinhas *et al*, 2005). The ITS sequence in particular, which is one of the most promising regions for fungi barcoding (Seifert, 2009), performed rather poorly in our benchmark comparisons being unable to distinguish species and recover their relationships, as other studies have noted (Cai *et al*, 2009; Crouch *et al*, 2009b; Yang *et al*, 2009). As evidenced by the sliding window analysis of polymorphism and parsimony statistics, the intergenic region between the *Apn1* and *MAT1-2-1* genes revealed to be the most variable segment and thus, a promising marker for both taxonomic and population genetic studies in the complex. Intergenic regions are expected to be much more variable since they are generally not under functional constraints or any direct selective pressures (O'Donnell *et al*, 2004; Thomson *et al*, 2010) but, for the same reason, they are unsuitable regions for primer design over a broad range of species. However, the ApMAT intergenic segment is flanked by fairly conserved regions in which primers can be designed, making it an ideal marker for the *C. gloeosporioides* complex. To a lesser extent, both the introns from *MAT1-2-1* gene and the two *Apn1* gene extremities, already partially applied in other *Colletotrichum* studies (Du *et al*, 2005; Crouch *et al*, 2009a; Crouch *et al*, 2006), were rather variable and should also prove useful for systematic purposes. Nevertheless, these markers are not barcode candidates, and attempts to use them fairly outside the range of the *C. gloeosporioides* complex will probably fail, due to their highly variable nature even in priming sites. For example, the markers developed by Crouch *et al* (2009a) to amplify partial fragments of the *Apn1* gene and intergenic region in the same locus for the graminicolous *Colletotrichum* species could not be used in the isolates analyzed in this study (data not

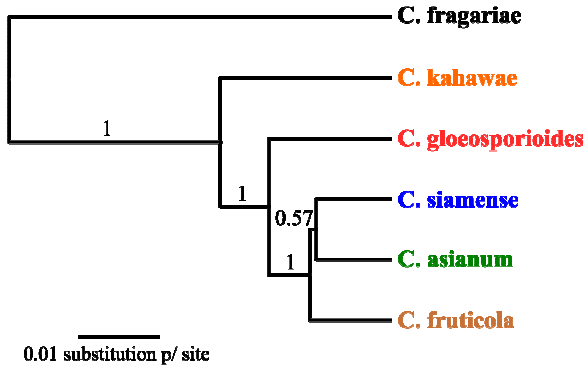

Fig. 5 The species tree (rooted with *C. fragariae*), for the total combined dataset using BEST software. Values above branches indicate the posterior probability support.

Table 4 Likelihood topology tests (SH) for each relevant dataset and hypothesis to be tested, based on the most recent taxonomic understanding of the studied species

Dataset	Hypothesis	Diff -ln L	P
Individual datasets			
Ap15L	Current	2.73	0.259
ApMAT	Current	12.93	0.035*
MAT1-2-1	Current	80.91	0.000*
GS	Current	5.02	0.132
Combined datasets			
ApnMAT	Current	202.42	0.000*
Traditional	Current	4.97	0.143
Total	Alternative	25.71	0.076^{+}

Note – Current (*C. asianum*, (*C. fruticola*, *C. siamense*)) and Alternative (*C. frutícola*, (*C. asianum*, *C.siamense*)) hypothesis tested; * Significance with $p < 0.05$; $^{+}$ Significance with $p < 0.10$

shown). In fact, after obtaining the homologous sequences from isolates of the 'gloeosporioides' complex, alignments revealed a high level of divergence between the two groups of species. However, these markers are much more suitable than barcode genes to accurately dissect and unravel the existent species diversity and their evolutionary relationships in a complex of closely related species. As Cai *et al* (2009) stated, the selection and decision of the most suitable barcode gene(s) can only be made in the process of species delimitation, since only then we know which taxa need to be distinguished. In other words, we first need to fully resolve the *C. gloeosporioides* complex with a set of highly informative genes in order to adequately search and select a barcode candidate.

Our study revealed the ubiquitous presence of only one MAT gene, *MAT1-2-1*, which comes in agreement with other studies in the *Colletotrichum* genus (García-Serrano *et al*, 2008; Chen *et al*, 2002; Vaillancourt *et al*, 2000). In the fungal kingdom, MAT genes are known transcription factors that regulate sexual development and recognition (Kronstad and Staben, 1997). However, their configuration in *Colletotrichum* species seems to deviate from the great majority of the ascomycetes, where a single MAT locus (MAT1) exists as two alternate forms (MAT1-1 and MAT1-2), called idiomorphs because they lack the similarity to be termed alleles (Turgeon, 1998; Rydholm *et al*, 2007). Thus, the exclusive presence of a MAT1-2 homologous gene in all of the species included in this work supports the idea that the genetic control underlying sexual reproduction in this genus does not conform to any known system in the entire fungal kingdom (Crouch and Beirn, 2009).

In our sampling, all isolates clustered into five groups, each corresponding to the previously characterized *Colletotrichum* species: *C. kahawae*, *C. gloeosporioides*, *C. asianum*, *C. fructicola* and *C. siamense*. These species have been recently described in coffee and other hosts and their phylogenetic relationships inferred from multi-locus datasets (Prihastuti *et al*, 2009; Phoulivong *et al*, 2010). From the individual gene trees, the ApMAT marker was singular in its ability to resolve species and even intraspecific relationships. Alone, it provided essentially the same information and support as the concatenated tree of seven markers, a feat that was not achieved by any other individual marker. The other markers had variable levels of resolution and support, ranging from low to moderate, as it is normally expected for nuclear gene markers when addressing recently diverged taxa (Thomson *et al*, 2010). However, gene trees are only single realizations of the true species tree (Degnan and Rosenberg, 2009), and the usefulness of individually reconstruct gene trees lies in enabling an overview of how each marker depicts species relationships. In this way, several incongruities were revealed by comparing the topology of each gene tree. For example, two isolates that were consistently shown to cluster with the *C. siamense* group (Mal5 and CCM5), were monophyletic with isolates from the *C. fructicola* group in the GS gene tree. Moreover, each analyzed locus provided a discordant species branching order with each other, with moderate to strong support values, and the correct relationship of the studied *Colletotrichum* species was not clear. In fact, only one gene tree (β-tub2) was congruent with the current taxonomic understanding of the sampled species, with ApMAT and *MAT1-2-1* gene trees revealing a significantly deviated phylogenetic signal. Even when different combinations of genes are concatenated, these dis-

cordances remain, as evidenced by the ambiguous relationship of *C. siamense*, *C. fructicola* and *C. asianum* between the Traditional and Apn-MAT datasets.

Even though the total concatenated dataset provided a single resolved topology, congruent with that of the combined Apn-MAT and individual ApMAT datasets, the concatenation method assumption of no heterogeneity between the embedded gene trees was violated. In addition, the longest and more informative markers, such as those in the Apn1/MAT locus, may bias tree topology towards their own simply because their phylogenetic signal overwhelms the information contained in less informative markers (Knowles and Carstens, 2007). These incongruities can be due to several factors, including Horizontal Gene Transfer (HGT) (Andersson *et al*, 2005) or Hibridization (Meng and Kubatko, 2009). However, in the initial stages of divergence, where many of the closely related species in complexes most likely are, incomplete lineage sorting of ancestral polymorphisms is a ubiquitous source of discordance (Carstens and Knowles, 2007; Carstens and Dewey, 2010). In our case study, incomplete lineage sorting seems to fit best the pattern of incongruence and given its magnitude, it was deemed necessary to formally account for it. The evolutionary history of species is widely known to be a stochastic process and thus, we have applied a recent statistical framework that accounts for discrepancies of the gene trees from multiple unlinked loci and is specifically conceived to estimate the species tree. Using a Bayesian hierarchical model, the BEST program is able not only to accommodate gene trees heterogeneity but also to retrieve useful information from it and provides a more accurate estimate of the true species tree (Edwards *et al*, 2007). The resultant species tree topology fairly supported that of the total concatenated dataset, except regarding the *C. siamense* and *C. asianum* relationships. Although all other nodes were highly supported, there was little support for the reciprocal monophyly between *C. siamense* and *C. asianum*, and the relationship of these two species plus *C. fructicola* was almost polytomic, i.e., it is not clear how these species relate to each other. This result becomes apparent even with the use of a large and highly informative sequence data from our analysis. Nonetheless, it is expected that the problems and difficulties derived from incomplete lineage sorting increase as the time since species divergence decreases (Degnan and Rosenberg, 2009). Thus, it seems likely that these species diverged from each other recently and consequently, their relationships are still not easily resolved.

Altogether, this study demonstrates the dramatic improvement that the Apn1/MAT locus may provide to the molecular systematic of the *C. gloeosporioides*. Using this new dataset, a new alternative view of the taxonomic position of the studied species is brought to light. Since we also showed that gene tree discordance may be a problem in recently diverged species of this complex, highly variable markers, such as ApMAT, are much more likely to surpass those problems and successfully distinguish taxa and understand how they are related. Indeed, according to Degnan and Rosenberg (2009) the probable gene tree discordance source of incomplete lineage sorting is less intense when addressed with more informative markers

This work is the first report of the development and application of this novel set of markers as valuable promising tools to dissect species and even population's relationships within the broader universe of the *C. gloeosporioides* complex. It is hoped that this contribution may assist and stimulate further research on this challenging complex, and eventually lead to a more comprehensive understanding of its structure, evolution and diversity.

Acknowledgements

At FCUL we thank our colleagues, Ana Vieira and Tiago Jesus, for invaluable contributions and criticisms during the elaborations of this manuscript. We are also grateful to Francisco Pina-Martins for helping in the first steps of the data analysis. At CIFC/IICT we appreciate the technical support provided by Sandra Sousa Emídio.

References

Abang MM, Winter S, Green KR, Hoffmann P, Mignouna HD, Wolf GA (2002) Molecular identification of *Colletotrichum gloeosporioides* causing yam anthracnose in Nigeria. Plant Pathol 51:63-71

Afanador-Kafuri L, Minz D, Maymon M, Freeman S (2002) Characterization of *Colletotrichum* isolates from Tamarillo, Passiflora, and Mango in Colombia and identification of a unique species from the genus. Phytopathology 93:579-587

Andersson JO (2005) Lateral gene transfer in eukaryotes. Cellular and Molecular Life Sciences 62:1182–1197

Brown A, Sreenivasaprasad S, Timmer, L (1996) Molecular characterization of slow-growing orange and key lime anthracnose strains of *Colletotrichum* from *Citrus* as *C. acutatum*. Phytopathology 86:523-527

Cai L, Hyde KD, Taylor PWJ, Weir BS, Waller J, Abang MM, Zhang JZ, Yang YL, Phoulivong S, Liu ZY, Prihastuti H, Shivas RG, McKenzie EHC, Johnston PR (2009) A polyphasic approach for studying Colletotrichum. Fungal Divers 39:183–204

Cannon PF, Bridge PD, Monte E (2000) Linking the past, present, and future of Colletotrichum systematics. In: Prusky D, Freeman S, Dickman M (eds) Colletotrichum: Host specificity, pathology, and host-pathogen interaction. APS, St. Paul, pp 1–20

Cannon PF, Buddie AG, Bridge PD (2008) The typification of *Colletotrichum gloeosporioides*. Mycotaxon 104:189-204

Carstens BC, Dewey TA (2010) Species delimitation using a combined coalescent and information-theoretic approach: An example from North American *Myotis* Bats. Syst Biol 59:400-414

Carstens BC, Knowles LL (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. Syst Biol 56:400-411

Chakraborty S, Fernandes C, Charchar MJD, Thomas M (2002) Pathogenic variation in *Colletotrichum gloeosporioides* infecting *Stylosanthes* spp. in a center of diversity in Brazil. Phytopathology 92:553-562

Chen F, Goodwin P, Khan A, Hsiang T (2002) Population structure and mating-type genes of *Colletotrichum graminicola* from *Agrostis palustris*. Can J Microbiol 48:427-436

Crouch J, Clarke B, Hillman B (2006) Unraveling evolutionary relationships among the divergent lineages of *Colletotrichum* causing anthracnose disease in turfgrass and corn. Phytopathology 96:46-60

Crouch J, Clarke B, Hillman B (2009) What is the value of ITS sequence data in *Colletotrichum* systematics and species diagnosis? A case study using the falcate-spored graminicolous *Colletotrichum* group. Mycologia 101:648-656

Crouch J, Tredway L, Clarke B, Hillman B (2009) Phylogenetic and population genetic divergence correspond with habitat for the pathogen *Colletotrichum cereale* and allied taxa across diverse grass communities. Mol Ecol 18:123-35

Crouch J, Beirn LA (2009) Anthracnose of cereals and grasses. Fungal Divers 39:19-44

Damm U, Woudenberg JHC, Cannon PF, Crous PW (2009) *Colletotrichum* species with curved conidia from herbaceous hosts. Fungal Divers 39:45-87

Degnan JH, Rosenberg NA (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol Evol 24:332-40

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutio nary analysis by sampling trees. BMC Evol Biol 7:214

Du M, Schardl C, Nuckles E, Vaillancourt L (2005) Using mating-type gene sequences for improved phylogenetic resolution of *Collectotrichum* species complexes. Mycologia 97:641-658

Edwards SV (2008) Is a new and general theory of molecular systematics emerging?. Evolution 63:1-19

Edwards SV, Liu L, Pearl DK (2007) High-resolution species trees without concatenation. Proc Natl Acad Sci USA 104:5936-41

Freeman S, Katan T, Shabi E (1998) Characterization of *Colletotrichum* species responsible for anthracnose diseases of various fruits. Plant Dis 82:596-605

Freeman S, Minz D, Jurkevitch E, Maymon M, Shabi E (2000) Molecular analyses of *Colletotrichum* species from almond and other fruits. Phytopathology 90:608-614

García-Serrano M, Laguna EA, Rodríguez-Guerra R, Simpson J (2008) Analysis of the *MAT1-2-1* gene of *Colletotrichum lindemuthianum*. Mycoscience 49:312-317

Hindorf H (1970) *Colletotrichum* spp. isolated from *Coffea arabica* L. in Kenya. Zeitschrift fur Pflanzenkrankheiten und Pflanzenschutz 77:328-331

Hyde KD, Cai L, McKenzie EHC, Yang Y, Zhang J, Prihastuti H (2009a) *Colletotrichum*: a catalogue of confusion. Fungal Divers 39:1-17

Hyde KD, Cai L, Cannon PF, Crouch JA, Crous PW, Damm U, Goodwin PH, Chen H, Johnston PR, Jones EBG, Liu ZY, McKenzie EHC, Moriwaki J, Noireung P, Pennycook SR, Pfenning LH, Prihastuti H, Sato T, Shivas RG, Tan YP, Taylor PWJ, Weir BS, Yang YL, Zhang JZ (2009b) Colletotrichum –names in current use. Fungal Divers 39: 147-182

Johnston P, Jones D (1997) Relationships among *Colletotrichum* isolates from fruit-rots assessed using rDNA sequences. Mycologia 89:420-430

Katoh K, Asimenos G, Toh H (2009) Bioinformatics for DNA Sequence Analysis. In: Posada D (ed) Methods in Molecular Biology, Humana Press, New Jersey. pp 537

Knowles LL, Carstens BC (2007) Delimiting species without monophyletic gene trees. Syst Biol 56: 887-895

Kronstad JW, Staben C (1997) Mating type in filamentous fungi. Annu Rev Genet 31:245-276

Liu, L (2008) BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24:2542-2543

Liu L, Pearl DK (2007) Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst Biol 56:504-514

Lubbe CM, Denman S, Cannon PF, Groenewald JZ, Lamprecht SC, Crous PW (2004) Characterization of *Colletotrichum gloeosporioides* and similar species associated with anthracnose and dieback of Proteaceae. Mycologia 96:1268-1279

Marshall OJ (2004) PerlPrimer: cross-platform, graphical primer design for standard, bisulphite and real-time PCR. Bioinformatics 20:2471-2472

Martin MP, Garcia Figueres F. (1999) *Colletotrichum acutatum* and *C. gloeosporioides* cause anthracnose on olives. Eur J Plant Pathol 105:733-741

Meng C, Kubatko LS (2009) Detecting hybrid speciation in the presence of incomplete lineage sorting using gene tree incongruence: a model. Theor Popul Biol 75:35-45

Mills PR, Sreenivasaprasad S, Brown AE (1992) Detection and differentiation of *Colletotrichum gloeosporioides* isolates using PCR. FEMS Microbiol Lett 98:137-144

Munaut F, Hamaide N, Stappen JV, Maraite H (1998) Genetic relationships among isolates of *Colletotrichum gloeosporioides* from *Stylosanthes* spp. in Africa and Australia using RAPD and ribosomal DNA markers. Plant Pathol 47:641-648

Nylander JAA (2004) MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University

O'Donnell K, Ward T, Geiser D, Kistler HC, Aoki T (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

Phoulivong S, Cai L, Chen H, McKenzie EHC, Abdelsalam K, Chukeatirote E, *et al* (2010) *Colletotrichum gloeosporioides* is not a common pathogen on tropical fruits. Fungal Divers. doi: 10.1007/s13225-010-0046-0

Pina-Martins F, Paulo OS (2008) Concatenator: sequence data matrices handling made easy. Mol Ecol Resour 8:1254–1255

Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. Bioinformatics 14:817-818

Prihastuti H, Cai L, Chen H, McKenzie, EHC, Hyde KD (2009) Characterization of *Colletotrichum* species associated with coffee berries in northern Thailand. Fungal Divers 39:89-109

Ramos AP, Merali Z, Talhinhas P, Sreenivasaprasad S, Oliveira H (2006) Molecular and morphological characterisation of *Colletotrichum* species involved in citrus anthracnose in Portugal. Bul OILB/SROP 29:317-326

Rydholm C, Dyer PS, Lutzoni F (2007) DNA sequence characterization and molecular evolution of MAT1 and MAT2 Mating-type loci of the self-compatible ascomycete mold *Neosartorya fischer*. Eukaryotic Cell 6:868-874

Rodríguez-Guerra R, Ramírez-Rueda MT, Cabral-Enciso M, García-Serrano M, Lira-Maldonado Z, Guevara-González RG, González-Chavira M, Simpson J (2005) Heterothallic mating observed between Mexican isolates of *Glomerella lindemuthiana*. Mycologia 97:793-803

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Shimodaira H, Hasegawa M (1999) Multiple comparisons of loglikelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114–1116

Shivas RG, Tan YP (2009) A taxonomic re-assessment of *Colletotrichum acutatum*, introducing *C. fioriniae* comb. et stat. nov. and *C. simmondsii* sp. nov. Fungal Diver 39:111-122

Sreenivasaprasad S, Brown A, Mills P (1993) Coffee Berry Disease pathogen in Africa: genetic structure and relationship to the group species *Colletotrichum gloeosporioides*. Mycol Res 87:995-1000

Sreenivasaprasad S, Mills PR, Brown AE (1994) Nucleotide sequence of the rDNA spacer 1 enables identification of isolates of *Colletotrichum* as *C. acutatum*. Mycol Res 98:186-188

Sreenivasaprasad S, Mills P, Meehan BM, Brown A (1996) Phylogeny and systematics of 18 *Colletotrichum* species based on ribosomal DNA spacer sequences. Genome 39:499-512

Sreenivasaprasad S, Talhinhas P (2005) Genotypic and phenotypic diversity in *Colletotrichum acutatum*, a cosmopolitan pathogen causing anthracnose on a wide range of hosts Mol Plant Pathol 6:361-378

Sutton BC (1992) The genus *Glomerella* and its anamorph *Colletotrichum*. In Bailey JA, Jeger MJ (eds.), *Colletotrichum: Biology, Pathology and Control* CAB International, Wallingford, pp 1-26

Suzuki T, Tanaka-Miwa C, Ebihara Y, Ito Y, Uematsu S (2010) Genetic polymorphism and virulence of *Colletotrichum gloeosporioides* isolated from strawberry (*Fragaria* x *ananassa* Duchesne). J Gen Plant Pathol 76:247-253

Swofford DL (2003) PAUP* Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts

Talhinhas P, Neves-Martins J, Oliveira H, Sreenivasaprasad S (2005) Molecular and phenotypic analyses reveal association of diverse *Colletotrichum acutatum* groups and a low level of *C. gloeosporioides* with olive anthracnose. Appl Environ Microbiol 71:2987-2998

Taylor J, Jacobson D, Geiser D, Kroken S, Hibbett D, Fisher M, *et al* (2000) Phylogenetic species recognition and species concepts in fungi. Fungal Genet Biol 31:21-32

Thomson RC, Wang IJ, Johnson JR (2010) Genome-enabled development of DNA markers for ecology, evolution and conservation. Mol Ecol 19:2184-2195

Turgeon B (1998) Application of mating type gene technology to problems in fungal biology. Annu Rev of Phytopathol 36:115-137

Vaillancourt LJ, Wang J, Hanau RM (2000) Genetic regulation of sexual compatibility. In: Host Specificity, Pathology and Host Pathogen Interaction of Colletotrichum (eds. D. Prusky, S. Freeman and M.B. Dickman). APS Press, St. Paul, MN: 24-44

Voigt K, Cozijnsen A, Kroymann J, Pöggeler S, Howlett B (2005) Phylogenetic relationships between members of the crucifer pathogenic *Leptosphaeria maculans* species complex as shown by mating type (MAT1-2), actin, and beta-tubulin sequences Mol Phylogenetic Evol 37:541-557

von Arx JA (1957) Die Arten der Gattung Colletotrichum. Phytopathologische Zeitshrift 29:413-468

Xiao C, MacKenzie S, Legard D (2004) Genetic and pathogenic analyses of *Colletotrichum gloeosporioides* isolates from strawberry and noncultivated hosts. Phytopathology 94:446-453

Yang YL, Liu ZY, Cai L, Hyde KD, Yu ZN, McKenzie EHC (2009) *Colletotrichum* anthracnose of *Amaryllidaceae*. Fungal Divers 39:123-149

Unraveling the phylogenetic origin and spread of *Colletotrichum kahawae* epidemics on *Coffea arabica* and the evolutionary relationships with the *C. gloeosporioides* species complex

Diogo Nuno Silva[1,2]*, Pedro Talhinhas[1], Andreia Loureiro[1], Luzolo Manuel[3], Elijah K. Guichuru[4], Vítor Várzea[1], Octávio Salgueiro Paulo[2], Dora Batista[1]

[1] *Centro de Investigação das Ferrugens do Cafeeiro (CIFC) / Instituto de Investigação Científica Tropical (IICT), Oeiras, Portugal.*

[2] *Computational Biology and Population Genomics group, Environmental Biology Center, Animal Biology Department, Science College, Lisbon University, Campo Grande, P-1749-016 Lisbon, Portugal*

[3] *Instituto Nacional do Café de Angola, Luanda, Angola*

[4] *Coffee Research Foundation, P.O.Box 4, Ruiru, Kenya*

Fungal plant diseases are emerging at an increasing rate and there is an urgent need to understand disease epidemics in order to respond with efficient disease control measures and strategies. *Colletotrichum kahawae* is an emergent pathogen causing severe epidemics of Coffee Berry Disease on Arabica coffee crops. The disease was first reported in 1922 in Kenya but has currently spread to most of the Arabica coffee growing regions in Africa and its potential imminent dissemination to other continents is cause for great concern. However, information regarding the origin of *C. kahawae* is mostly limited to historical data and the low genetic variation of its populations has hindered inferences about their structure and dispersal patterns. Here we describe the use of phylogenetic and phylogeographic approaches to address the origin of this emergent pathogen within the closely related group species *C. gloeosporioides* and to assess the population genetic variability to infer potential dissemination routes. Using a six gene multi-locus dataset we found a high divergence between *C. kahawae* and *C. gloeosporioides sensu lato* from coffee hosts, in contrast to a close relationship to isolates from mango and strawberry. In addition, we were able to identify three groups within *C. kahawae*, in which the Angolan group seemed the most ancestral, based on our sampling. Additionally, we analyzed the molecular evolution of a mating type gene, *MAT1-2-1*, which appeared to have been unimportant for *C. kahawae* speciation but may have a role at the populational level, as revealed by the presence of two haplotypes of a gene under strong purifying selection. These results suggest an alternative view on the origin and spread of *C. kahawe* and highlight the importance of molecular data when addressing these issues.

Keywords: Emerging diseases; Host-shift; Coffee crops; anthracnose

*Corresponding author: Diogo Nuno Silva. e-mail: diogo_nuno_silva@hotmail.com; Address: Centro de Investigação das Ferrugens do Cafeeiro (CIFC)/ Instituto de Investigação Científica Tropical (IICT), Oeiras, Portugal; Telephone: +351962656755

Running title: Phylogenetic origin and spread of *C. kahawae*

Introduction

Arabica coffee (*Coffea arabica* L.) is the most economically relevant coffee species, accounting for 75% of coffee production worldwide (Waller *et al.* 2007). Over the last century, *C. arabica* plantations in Africa have experienced devastating anthracnose epidemics due to the emergent and highly virulent pathogen, *Colletotrichum kahawae* Waller & Bridge (Hindorf 1970; Gordon 1988; Waller *et al.* 1993; Silva *et al.* 2006). Unlike other *Colletotrichum* spp. present on coffee plants, *C. kahawae* is able to cause a severe anthracnose disease on green berries, known as Coffee Berry Disease (CBD), leading to their premature drop or blackened mummification and preventing the subsequent processing of the beans (Nutman & Roberts 1960; Firman & Waller 1977; Silva *et al.* 2006). During average outbreaks, yield losses of 20-30% are common but can exceed 80% in extremely wet years if no control measures are applied, which can be particularly burdensome in developing countries where these epidemics translate into the loss of several million US Dollars per year (Várzea *et al.* 2002; Derso & Waller 2003). Despite several efforts, no absolute resistant coffee plant variety has been developed and fungicides have been somewhat ineffective (Waller & Masaba 2006; van der Vossen 2009). Thus, CBD is regarded as a major factor limiting the production of Arabica coffee in the African continent (Várzea *et al.* 1993; Omondi *et al.* 2000; van der Vossen 2009).

This disease seems to be very recent, as it was first reported in 1922 in western Kenya, when it led to the destruction and abandon of coffee plantations in some districts (McDonald 1926). Despite the little attention received during the initial stages of its documented emergence, African coffee growers soon witnessed a swift spread of CBD throughout most of the continent (Gordon 1988; Silva *et al.* 2006). In roughly 50 years, the presence of the disease was documented in nearly every region where Arabica coffee is grown, from Ethiopia to

Zimbabwe (Gordon 1988). Nonetheless, the causal agent of CBD is still restricted to the African continent, more frequently above 1400m of altitude and in cool and humid environments, which are also the most suitable for *C. arabica* (Schultes 1986; Waller & Masaba 2006). Still, an eventual dispersal to coffee plantations of Latin America and Asia is cause for great concern and has motivated a growing interest on the evolutionary potential of *C. kahawae* (Bridge *et al.* 2008).

Several studies have attempted to unveil the genetic diversity and population structure of this pathogen, mostly driven by the information that these factors may provide on the pathogenic potential of the fungus and to take a glimpse on how its populations had evolved and dispersed through temporal and geographical scales. However, they have revealed an astounding lack of genetic variability suggesting that the species may be comprised by a single clonal population (Sreenivasaprasad *et al.* 1993; Beynon & Várzea 1995; Omondi *et al.* 1997; Gichuru *et al.* 2000; Derso & Waller 2003). Only recently, more sensitive molecular techniques were able to unravel the existence of a subtle variation but their limited sampling hindered more general conclusions (Bridge *et al.* 2008; Manuel *et al.* 2009). Until now, no global and consistent structuring has been clearly revealed and, therefore, little information is known regarding the origin, structure and spread of this pathogen's populations besides that provided by historical data.

Two main reasons have been advocated to explain the overall genetic uniformity of *C. kahawae*. First, a recent evolution from the closely related and cosmopolitan group-species *C. gloeosporioides* has been suggested. Three competing hypotheses were proposed, in which *C. kahawae* would have emerged either by: 1) mutation from a mildly pathogenic strain of the *C. gloeosporioides* gene pool from coffee hosts, which cause the economically unimportant brown blight on ripe berries (Nutman & Roberts 1960); 2) hybridization between two *C.*

gloeosporioides strains from other *Coffea* spp. hosts, such as *C. canephora* and *C. eugenioides* (Robinson 1976); 3) or emerging from one of the previously mentioned *Coffea* spp., where it would have been present as an harmless fungal strain, to become a severe pathogen in *C. arabica*, when Arabica coffee was re-introduced in Africa for commercial purposes (Robinson 1974). At this point, there is not enough data available to know which scenario fits best but it seems to be a general agreement that *C. gloeosporioides sensu lato* populations from *Coffea* spp. are the most likely source for *C. kahawae*'s origin. However, *C. gloeosporioides* is a large species complex, reported in at least 1800 plant species, that has remained taxonomically ill-defined and simply stating its close relationship to *C. kahawae* is nearly meaningless (Du *et al.* 2005; Cannon *et al.* 2008). For example, a recent study has addressed isolates from this complex from coffee hosts in Thailand and revealed the existence of three new species, clearly distinct from the recently published *C. gloeosporioides* epitype (Prihastuti *et al.* 2009; Cannon *et al.* 2008). Thus, even when only considering populations from coffee hosts, a more relevant question is raised: From which lineage within *C. gloeosporioides sensu lato* has *C. kahawae* emerged?

The second reason concerns the putative asexual nature of *C. kahawae*, which contrasts with the sexual status of the *C. gloeosporioides* epitype. Indeed, the absence of a sexual stage, both in culture and in the field, as well as the clonal structure of its populations described so far, have served as evidence for the asexuality of *C. kahawae*, although they are not conclusive (Firman & Waller 1977). In this regard, some studies have been addressing the evolution of mating-type (MAT) genes in closely related fungal species with different sexual status (O'Donnell *et al.* 2004; Voigt *et al.* 2005; Stergiopoulos *et al.* 2007), and, in fact, they were able to change our current perspective on the most likely reproductive mode of some species (Kerényi *et al.* 2004; Stergiopoulos *et al.* 2007). MAT genes are known transcription factors

that govern sexual reproduction and compatibility in fungi (Turgeon 1998). In the great majority of the ascomycetes, there is a single bipolar MAT locus (MAT1) that exists as two alternate forms (MAT1-1 and MAT1-2), called idiomorphs because they lack the similarity to be termed alleles (Rydholm *et al.* 2007). Fungal species can be either heterothallic, if two different strains bearing one of the two idiomorphs are required for mating, or homothallic, if a single strain carries both idiomorphs (Kronstad & Staben 1997). However, the *Colletotrichum* genus has been documented as an apparent deviation to any known system in the entire fungal kingdom (García-Serrano *et al.* 2008; Crouch & Beirn 2009). Mating studies have shown that species from this genus can be both homothallic and heterothallic, such as *C. graminicola* and *C. gloeosporioides* (Cisar & TeBeest 1999; Vaillancourt *et al.* 2000). Moreover, only the MAT1-2 gene, termed *MAT1-2-1*, has been ubiquitously found in all studied *Colletotrichum* strains and there is no evidence for the presence of the MAT1-1 gene (Chen *et al.* 2002; Crouch *et al.* 2007; García-Serrano *et al.* 2008).

In this study, we made use of a multi-locus sequencing approach of five nuclear genes, including *MAT1-2-1*, to explore the longstanding problematic of *C. kahawae*'s origin within *C. gloeosporioides sensu lato* and its subsequent spread throughout the African continent. Specifically, using an extensive sampling of *C. kahawae* across most of its range, we intended to reassess the genetic variability and structure of its populations in order to infer possible ancestral states and patterns of dispersal. Moreover, with a worldwide sampling of *C. gloeosporioides sensu lato* from coffee (including other hosts), we intended to, at some extent, test the three current hypotheses for the emergence of *C. kahawae*. We have included the epitypes of three recently described species from coffee hosts (*C. asianum*, *C. siamense* and *C. fructicola*), as well as type representatives of *C. gloeosporioides*, to provide a backbone reference for the phylogenetic analysis. This inclusion of a broad range of species

will also allow a much more rigorous and informative analysis of the evolution of the single mating-type gene found in *Colletotrichum*, *MAT1-2-1*, with potential implications to our current understanding of *C. kahawae*'s sexuality and evolution.

Materials and Methods

Fungal material

A total of 85 isolates were obtained from collections maintained at CIFC/IICT, Portugal, Mae Fah Luang University, Thailand and Instituto Superior de Agronomia, Portugal, and are described in Table S1. This sampling comprises at least five species belonging to the *C. gloeosporioides* complex. The *C. kahawae* sampling was obtained from nine African countries, covering most of its current range (Angola, Cameroon, Ethiopia, Kenya, Tanzania, Rwanda, Burundi, Malawi and Zimbabwe). Isolates from *Citrus lemon* and *Olea europaea* hosts, similar to the *C. gloeosporioides* epitype (based on Blast searches of the ITS and β-tub2 sequences; data not shown) were also used. One *C. gloeosporioides sensu lato* isolate from *Mangifera indica* showing a distinct ITS haplotype and was included (Ramos, personal communication). Five samples of *C. fragariae* were selected as outgroup taxa to root the phylogenetic trees (Cai *et al.* 2009; Silva *et al.* Unpublished). Culturing and DNA extraction of fungal isolates were as described previously (Silva *et al.* Unpublished).

Molecular data

Six nuclear gene markers were chosen for a detailed sequencing analysis, based on a preliminary screening of the following markers: The complete rDNA Internal Transcribed Spacer (ITS), a partial fragment of *β-tubulin 2* (β-tub2), two intergenic regions from the Apn1/MAT locus (ApMAT and MAT5L), the 5' end fragment of the *Apn1* gene (Apn15L) and the complete *MAT1-2-1* gene (Silva *et al.* Unpublished). Regarding the *C. kahawae*

samples, all isolates were sequenced for β-tub2, *MAT1-2-1* and MAT5L, but only a subset (1-6) was selected from each sampled country for sequencing using the other nuclear markers. The remaining isolates from other species were sequenced for the total six marker set. Primers and PCR conditions for ApMAT, MAT5L, Apn15L and *MAT1-2-1* were as described by Silva *et al.* (Unpublished). PCR amplification of ITS with primers ITS1Ext/4Ext (Brown *et al.* 1996) and β-tub2 using primers T1/T2 (O'Donnell & Cigelnik 1997). Amplification products were resolved on 1.2% agarose gels stained with ethidium bromide (EtBr) to verify the amplification success. PCR products yielding one clear band were purified using SureClean (BioLine). When products presented a multi-band profile, the band of the expected size was excised and purified using the Silica Bead DNA Gel Extraction kit (Fermentas). Sequencing reactions were carried out using the BigDye version 3.1 chemistry (Applied Biosystems) on an ABI prism 310 automated sequencer. Amplicons were sequenced in both directions and chromatographs were manually checked for errors in SEQUENCHER v4.0.5 (Gene Codes Corporation).

Phylogenetic analysis

Multiple sequence alignments were constructed for each dataset in MAFFT v6.717b (Katoh *et al.* 2009), using the L-INS-i method, followed by manual refinement in BIOEDIT v7.0.5.1. Individual datasets were concatenated into a combined matrix using the CONCATENATOR software (Pina-Martins & Paulo 2008). We used maximum likelihood (ML) and Bayesian methods to reconstruct phylogenies from the separate and combined datasets. ML searches were performed in PAUP* v4.0d99 (Swofford 2000) with 100 replicates, random sequence addition and a Tree-Bissection Reconnection (TBR) branch swapping algorithm. Nonparametric bootstrap was conducted using 1000 pseudoreplicates with 10 random

Table 1 Characteristics of the individual and combined sequence datasets used in this study

Parameter	Sequenced region						
	ITS	**β-tub2**	**Apn15L**	**ApMAT**	**MAT5L**	**MAT1-2-1**	**Combined**
Nucleotide characters (bp)	489	578	883	772	213	843	3778
Indels	2	31	13	79	19	4	130
Total characters	487	547	870	693	194	839	3648
Parsimony informative	14	77	184	289	42	148	764
% Parsimony informative	**3 %**	**13%**	**21%**	**37%**	**20%**	**18%**	**20%**
Variable, uninformative	0	4	17	38	5	7	72
Variable in *C. kahawae*	0	2	0	0	0	1	3
Model	TrNef+I	TIM+G	TrN+G	HKY+G	TIM+G	TIM+G	

additions and TBR branch swapping. MODELTEST v3.7 (Posada & Crandall 1998) was used to select the best fit model of DNA sequence evolution, under the Akaike Information Criterion (AIC), for each dataset. The Bayesian analysis was run in MRBAYES v3.1.2 (Ronquist & Huelsenbeck 2003) with the optimal model of sequence evolution selected under the AIC, as implemented in MRMODELTEST v2.3 (Nylander 2004), specified as prior for each partition. Posterior probabilities were generated from $1x10^8$ generations, sampled at every 1000^{th} generation. The analysis was run three times with one cold and three incrementally heated Metropolis-coupled Monte Carlo Markov chains, starting from random trees. Log files were analyzed and the convergence and mixing of the independent runs were assessed for all parameters using TRACER v1.4 (Rambaut & Drummond 2007). A suitable burin phase was established and $1x10^6$ generations were discarded. Trees from the different runs were then combined and summarized in a 50% consensus tree.

ITS phylogenetic context

Aiming at a rough phylogenetic contextualization of *C. kahawae*, a complementary analysis was also performed focusing on ITS sequences, since this is the most common nuclear region sequenced for the 'gloeosporioides' complex. A sample of 547 sequences named "*C.* gloeosporioides" or "*Glomerella cingulata*" curated in GenBank, accessed at 13/07/10, was

analyzed and a subset with highest homology to *C. kahawae* was selected. Sequences were aligned in MAFFT v6.717b (Katoh *et al.* 2009), using the FFT-NS-2 method, and collapsed into haplotypes using SNAP Combine and MAP (Aylor *et al.* 2006), to reduce the computational load of the subsequent analysis. An uncorrected pairwise distance matrix, excluding gaps, was calculated in PAUP* v4.0d99 (Swofford 2000) and, based on the distance between the isolates Rua1 (*C. kahawae*) and Mal5 (*C. gloeosporioides sensu lato lato* – closest relative to *C. kahawae* from coffee crops in our sample), an objective threshold was established above which sequences would be discarded. After this selection, published ITS sequences of *C. gloeosporioides sensu lato* isolated from coffee hosts were added. The final dataset was used to construct a Median Joining (MJ) network, using NETWORK v.4.516 (Bandelt *et al.* 1999).

Phylogeographic analysis

The relationship of *C. kahawae*'s populations was estimated by ML and Bayesian methods as described above. For each of the polymorphic sites within *C. kahawae*, the ancestral states were estimated using SNAP Workbench (Price & Carbone 2005). Relevant alternative topologies were tested to assess the robustness of the unconstrained inference of population's relationship, using the topological test of Shimodaira & Hasegawa (Shimodaira & Hasegawa 1999). One thousand replicates were performed by resampling the partial likelihoods for each site (RELL model). Using the IDRISI 15 Andes (Eastman 2006), a topographical map of a partial region of the African continent was modeled, in order to highlight regions above 1400m of altitude.

Analysis of MAT1-2-1 molecular evolution

The molecular evolution analysis of the *MAT1-2-1* gene was initiated by creating a MJ network from a dataset containing only nonsynonymous sites, using NETWORK v.4.516

(Bandelt *et al.* 1999). Patterns of selection within the *MAT1-2-1* gene were investigated using the codeml program of the PAML4 package (Yang 2007), using a fully bifurcating tree, as inferred above from the phylogenetic analysis, with single representatives for each species/haplotype. Selective pressure was measured by using the nonsynonimous/synonymous (dS/dN) rate ratio, also referred to as ω. An $\omega < 1$ suggests that purifying selection is acting on a gene, $\omega = 1$ is consistent with a gene evolving in a neutral fashion and $\omega > 1$ generally reflects that positive selection may be occurring. Using a maximum likelihood analysis, nested codon and branch models can be estimated and compared by means of a likelihood ratio test (LRT). Using codon models, the existence of positive selection on particular sites was tested, using two recommended nested models: M1a vs M2a and M7 vs M8. The null model M1a assumes two site classes with $1 > \omega > 0$ and $\omega = 1$, which implicitly supposes that no site is under positive selection, and can be compared to the alternative model M2a, which adds an extra class of sites with $\omega > 1$, allowing the presence of positively selected sites. Alternatively, the null model M7, which assumes a beta distribution within the interval (0, 1) for ω values across all sites, was compared with model M8, which adds an extra class where ω can take values >1. In this way, positive selection can be detected if a model allowing for positively selected sites is significantly better (as estimated through an LRT) than the null model. Since very little information is available for the *MAT1-2-1* sequence evolution in *Colletotrichum* and information regarding the reproductive mode of the studied species may not be completely accurate, we have based one of our hypotheses for the branch models on a preliminary analysis of the *MAT1-2-1* haplotype distribution. Using the null hypothesis of one ω ratio for every branch of the phylogenetic tree, we tested both two and three ratio models to assess if: 1) the ancestral branch to *C.*

kahawae had experienced positive selection; 2) the terminal branches within the 'gloeosporioides' complex from coffee hosts were under significant less selective constraints

Results

Phylogenetic relationships of Colletotrichum spp. from coffee and other hosts

The evolutionary relationships of *C. kahawae* and *C. gloeosporioides sensu lato* species from *Coffea* spp. and other hosts, were established by constructing a multi-locus phylogeny from 85 samples. Since species assignment prior to this study was uncertain for most of the *C. gloeosporioides sensu lato* isolates, the inclusion of species epitypes (or type representatives) from *C. gloeosporioides*, *C. asianum*, *C. fructicola* and *C. siamense* provided a useful backbone reference with which phylogenetic relationships could be used to enlighten their taxonomy. The employed six gene sequence dataset from three loci (ITS, β-tub2, Apn15L, ApMAT, MAT1-2-1 and MAT5L), provided a substantial resolution of species and populations relationships, especially within *C. gloeosporioides sensu lato* (Fig. 1). Using parsimony statistics as a comparative benchmark of the six nuclear regions, the ApMAT, Apn15L, MAT5L and *MAT1-2-1* were 37%, 21%, 20% and 18% parsimony informative, while β-tub2 and ITS were 13% and 3% informative, respectively (Table 1). Altogether, 764 characters (20% of the total dataset) were informative when analyzed using parsimony criteria.

Phylogenetic reconstructions using the combined 3778bp dataset yielded the same topology between ML and Bayesian methods, despite discrepancies in the support of some nodes (Fig. 1). The phylogeny was mostly resolved with high node support (i.e., branches supported by > 75% bootstrap and > 90% posterior probabilities) and the topology revealed a pattern of divergent evolution between three main lineages (Fig. 1). The first was comprised by all *C. kahawae* isolates plus a single *C. gloeosporioides sensu lato* isolate from *Mangifera indica*.

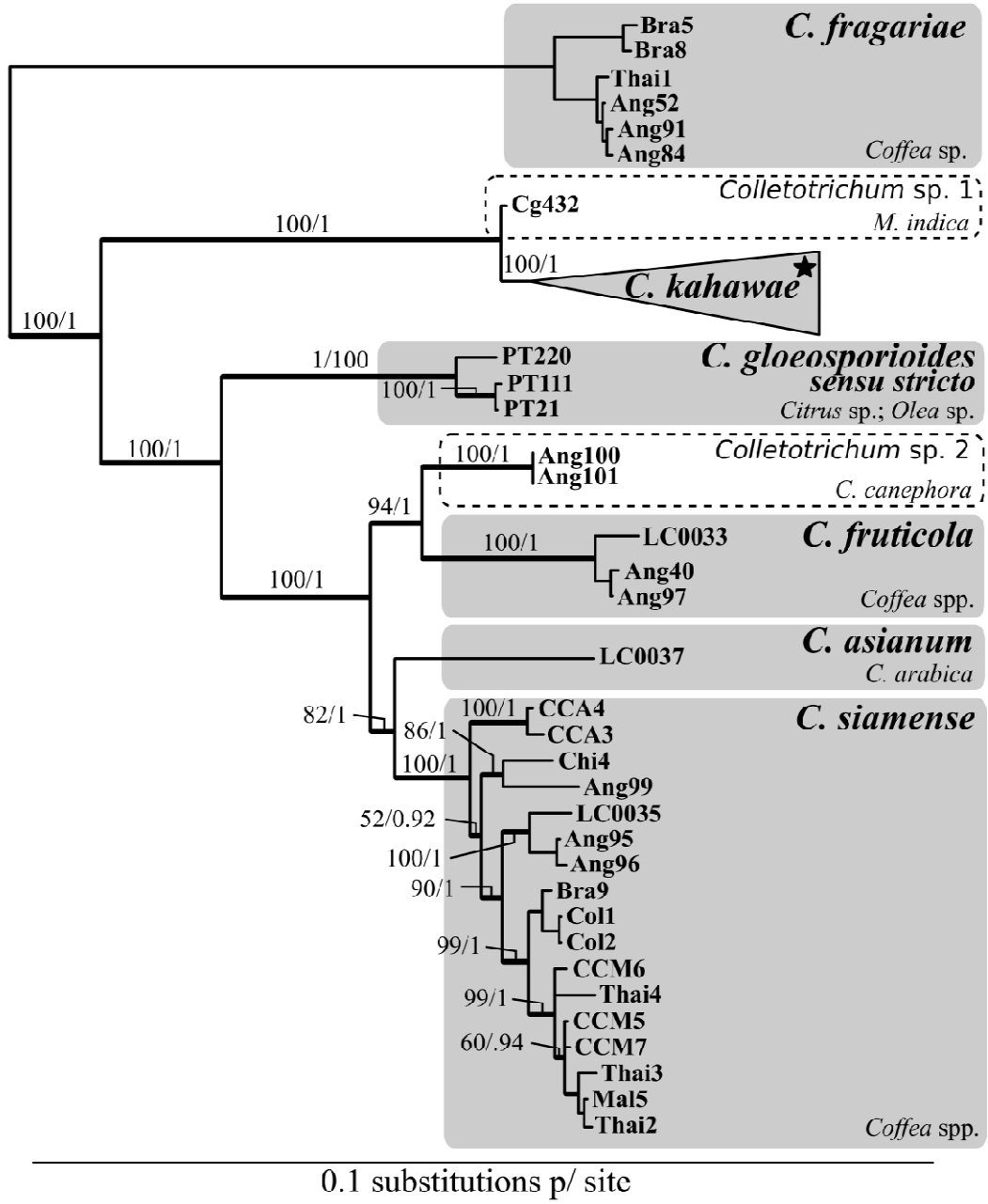

Fig 1 50% majority rule Bayesian ML tree with the concatenated six-gene (3778 bp) dataset illustrating the evolutionary relationships between *C. kahawae* and *C. gloeosporioides sensu lato* from coffee and other hosts. The tree was root with *C. fragariae* as outgroup taxa. *Colletotrichum kahawe* is represented by a single clade (marked with a black start) for clarity. The detailed phylogenetic reconstruction of the *C. kahawae* clade is provided in Fig 3a. The numbers above branches represent bootstrap and posterior probability values, respectively.

(*Colletotrichum* sp. 1) and represents the most ancestral lineage to have diverged (11.6%) from the remaining *Colletotrichum* spp. in this study. Although the *C. kahawae* clade is clearly distinguishable from *Colletotrichum* sp. 1 with high bootstrap and posterior probabilities (100/1) using a multi-locus dataset and by the inability of *Colletotrichum* sp. 1 to cause CBD symptoms (data not shown), their divergence is rather small (0.2%). In fact, for

the ApMAT, MAT5L and β-tub2 gene trees, *Colletotrichum* sp. 1 is positioned within the *C. kahawae* clade (Fig. S1a, e, f). The remaining *C. gloeosporioides sensu lato* isolates clustered in a large monophyletic group, comprised by the other two main lineages. The representative isolates of *C. gloeosporioides sensu stricto* from *O. europaea* and *C. lemon* clustered well apart (8.4%) from the remaining isolates and were the first lineage to diverge within the *C. gloeosporioides* complex. The third main lineage is exclusively composed by isolates from coffee hosts, belonging at least to four distinct species, and revealed a fairly high divergence from *C. kahawae* (11.8%). In this lineage, species assignment was difficult due to a high genetic heterogeneity and haplotype diversity (H_D = 0.964). Most of the unassigned *C. gloeosporioides sensu lato* strains grouped with *C. siamense* but their distance to the epitype varied widely, which greatly decreased our confidence in assigning those strains to a single species. Thus, even though we treated the whole group as part of the *C. siamense* clade, it may be possible that it encompasses more than one species. On the other hand, isolates within the *C. fructicola* group were more genetically homogeneous and could be considered as the same species. We also found two clonal samples (Ang100 and Ang101) that revealed to be a distinct monophyletic lineage with *C. fructicola* but, given their divergence from any included epitype, they may represent a new undescribed species (*Colletotrichum* sp. 2). In our sample we did not recover closely related strains to the *C. asianum* epitype. Most of the individual gene supported this main topology, though with varying levels of resolution. An exception comes with the ITS gene tree, which was unable to recover clear phylogenetic relationships among the studied species.

We attempted to complement the phylogenetic analysis of *C. kahawae*, taking advantage of the universality and ubiquity of ITS sequences for strains of the *C. gloeosporioides* species complex, deposited in the GenBank public database. Since a tree-like phylogenetic analysis is

unlikely to be useful, as proved by our own ITS gene tree, we resorted to a network construction approach to depict relationships among a selection of sequences collected from GenBank, in addition to our own sampling. In a first step, 19 sequences were selected out of a sample of 547 sequences named "*C. gloeosporioides*" or "*G. cingulata*", for having the highest homology with *C. kahawae*. For the sake of reliability, we excluded five of these sequences for not being associated to any publication, while the remaining 14 sequences, obtained by Polashok *et al.* (2009) from *Fragaria* x *ananassa*, were retained. Then, 19 additional ITS sequences from *C. gloeosporioides sensu lato* isolates from *Coffea* spp. hosts were included (Nguyen *et al.* 2009; Prihastuti *et al.* 2009), in order to provide a broader perspective of the "gloeosporioides" complex presence in this host. Altogether, the network analysis comprised 89 ITS sequences (Fig. 2). Given that the overall variability of the ITS region was very low, species did not cluster away from each other completely, particularly *C. siamense*, *C. fructicola* and *C. asianum*, whose corresponding strains were scattered in the part B of the network. Focusing on part A, *C. kahawae* was found to be more closely related to isolates from *F.* x *ananassa* and *M. indica* by one and two mutations, respectively, than to isolates from *Coffea* spp. worldwide.

Phylogeography of C. kahawae

Regarding the *C. kahawae* clade, our results confirmed the extremely low genetic variability of this species throughout most of its range (Nucleotide diversity, $\pi = 0.00076$; Segregating sites, S = 3). However, both β-tub2 and *MAT1-2-1* datasets revealed informative enough to distinguish three divergent but clonal populations: Angola, Cameroon and East African populations, with the later including isolates from the remaining seven east African countries (Fig. 3a). For the three polymorphic sites within *C. kahawae*, ancestral states were inferred from the *C. gloeosporioides* sample (Fig. 3a), which were monomorphic for all sites (i.e., all

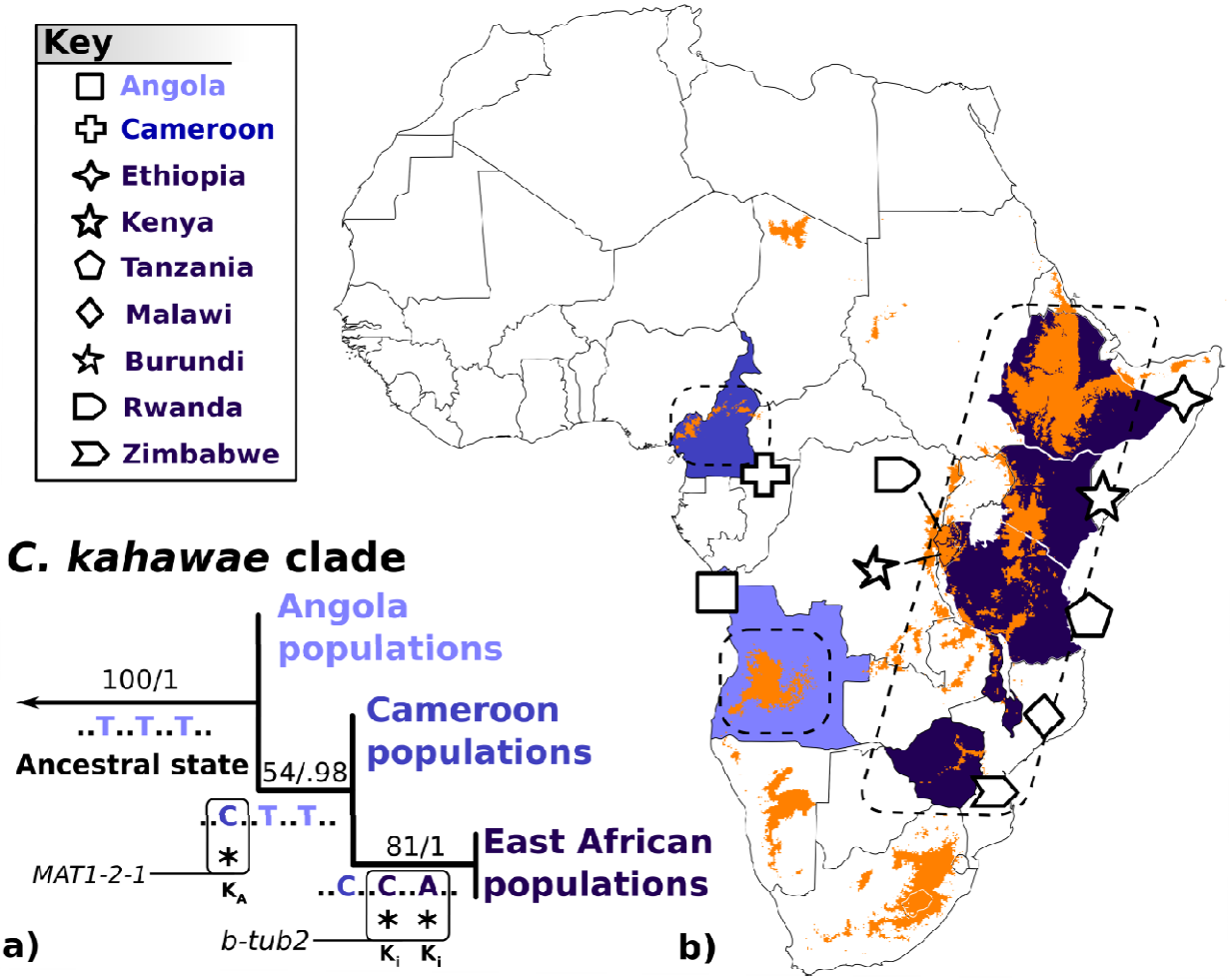

Fig 3 a) Detailed phylogenetic reconstruction of the population's relationship within the *C. kahawae* clade, using the combined six-gene dataset. Bootstrap and posterior probability values are provided above each branch. The three site combination provided below each branch represents the mutational events that occurred along the evolution of the three populations. The combination from the Angola population was inferred as the ancestral state. Mutations are highlighted with an asterisk and the source gene is provided. K_A, non-synonymous mutation; K_i, intronic mutations. **b)** Geographic distribution of the three divergent *C. kahawae* populations, with the respective location key provided on the superior left corner. Countries were highlighted with different colors corresponding to the existent haplotypes. Orange areas on the map represent regions above 1400m of altitude.

C. gloeosporioides isolates were T/T/T for the three segregating sites). The Angola population presented a nucleotide sequence identical to the inferred ancestral state, which makes it the most ancestral by inference. The Cameroon population diverged in one non-synonymous mutation at the *MAT1-2-1* gene, which replaces a serine residue for a proline residue. The East African population shared the same non-synonymous mutation and diverged from the other populations by two additional intronic mutations at the β-tub2 gene. To assess the robustness of this inference we tested an alternative topology, in which we constrained the

East African population as the most ancestral lineage and the Angola population as the most derived. The likelihood score of the resulting tree was worse than the unconstrained topology, with a marginally significant p-value (SH Test, P = 0.056), which means that we can assume the initial inference with fairly high confidence. Despite small, these differences were consistent across the individual datasets (Fig. S1) and when combined, revealed a population structure that seems correlated with geographical location (Fig. 3b). On the other hand, we could not detect the presence of migrant haplotypes in our sample, that is, there was no mixing of different haplotypes in the same geographic location. We further extended our phylogeographical analysis by modeling a topographical map of a partial region of the African continent, in order to highlight all regions above 1400m of altitude, and the result was embedded on Fig. 3a, as orange areas. As it can be observed, east African regions present a larger extension of highlands, which coincides with the Great Rift Valley area. Outside this area, highland regions are sparsely distributed, mainly through South Africa, Namibia, Angola and Cameroon, and isolated by long distances of lowland regions.

Molecular evolution of the MAT1-2-1 gene

A MJ network based on a dataset of non-synonymous mutations of the *MAT1-2-1* gene was constructed (Fig. 4), revealing eight different *MAT1-2-1* haplotypes that were not equally distributed throughout the sampled species. For example, *C. kahawae* possessed two different haplotypes (H2 and H3) separated by a single non-synonymous mutation and one of these haplotypes (H2) was shared with the closely related *Colletotrichum* sp. 1. On the other hand, the H4 haplotype was found in two different and well diverged (5.3%, based on the combined phylogenetic analysis) clades, *C. gloeosporioides* and *C. siamense*. However, the *C. siamense* epitype and two other isolates (Ang95 and Ang96) possessed a unique haplotype (H8) that diverged from the remaining *C. siamense* and *C. gloeosporioides* isolates by one non-

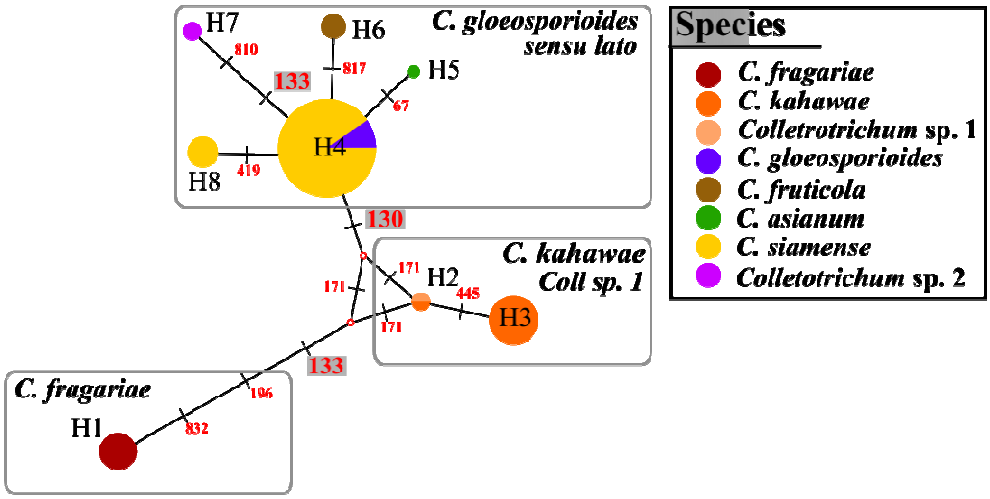

Fig 6 MJ network constructed from a non-synonymous sites dataset of the *MAT1-2-1* gene. Non-synonymous mutations are represented by dashes along the lines connecting haplotypes and their position in the gene sequence is provided by the numbers in red. Sites 130 and 133 are highlighted in grey due to the possibility of being under positive selection, according to the sites model M8 (Table 3).

synonymous mutation. All of the other haplotypes were exclusive for each of the remaining species.

For the analysis of the molecular evolution of *MAT1-2-1*, likelihood scores and estimates of ω under different models of codon and branch evolution are listed in Table 3 and the LRT results comparing nested models are listed in Table 4. Regarding the site models, model M0 was used as a first approach, which assumes a single ω ratio for all nucleotide sites and branches in the phylogeny, and estimated a $\omega_0 = 0.038$ for *MAT1-2-1*. Significant variation in the selective constraints among sites ($P < 0.05$) was identified by comparing model M0 with model M3 with two ω classes ($K = 2$). While parameter estimation under model M3 further indicated that the majority of sites were dominated by strong purifying selection ($p_0 = 0.986$; $\omega_0 = 0.023$), a sub-set of sites had $\omega > 1$ ($p_1 = 0.014$, $\omega_1 = 1.94$). However, LRTs specific for positive selection at individual sites gave discordant results. While the comparison between

Table 2 Parameter estimates and likelihood scores using Sites and Branch specific models for the *MAT1-2-1* gene.

	model		*p*	likelihood	κ	TL	Estimates of parameters
Sites Models							
A	One ratio (M0)	Nsites = 0	17	-1402,34	3,04	0,476	ω_0 = 0.03765
C	Nearly Neutral (M1a)	Nsites = 1	18	-1398,63	3,13	0,491	p_0 = 0.97433, ω_0 = 0.01934
D	Discrete (M3)	Nsites = 3, K=2	19	-1398,39	3,19	0,498	p_0 = 0.98556, p_1 = 0.01444; ω_0 = 0.02337, ω_1 = 1.93704
E	Selection (M2a)	Nsites =2	20	-1398,39	3,19	0,498	p_0 = 0.98556, p_1 = 0, p_2 = 0.01444; ω_0 = 0.02337, ω_2 = 1.93704
F	Beta (M7)	Nsites =7	18	-1401,05	3,04	0,478	p = 0.07918, q = 0.53983
G	Beta&w (M8)	Nsites = 8	20	-1398,40	3,19	0,498	p_0 = 0.98563, p_1 = 0.01437; p = 2.55244, q = 99.00000, ω_2 = 1.93605
Branch specific models							
A	One ratio	Model = 0	17	-1402,34	3,04	0,476	0,0377
B	Free ratio	Model = 1	31	-1389,78	3,10	0,485	Free
C	Two ratios	Model = 2	18	-1397,10	3,06	0,485	$\omega_0 = \omega_{CK}$ = 0.0197, $\omega_T = \omega_{SP2}$ = 0.1615
C1	Two ratios	Model = 2	18	-1400,68	3,05	0,478	$\omega_0 = \omega_{CK} = \omega_{SP2}$ = 0.0290, ω_T = 0.1102
C2	Two ratios	Model = 2	18	-1398,81	3,05	0,478	$\omega_0 = \omega_T = \omega_{CK}$ = 0.0294, ω_{SP2} = 0.4287
D	Three ratios	Model = 2	19	-1396,33	3,06	0,478	$\omega_0 = \omega_{CK}$ = 0.01966, ω_T = 0.11023,
E	Two ratios	Model =2	18	-1401,87	3,05	0,478	$\omega_0 = \omega_T = \omega_{SP2}$ = 0.03509, ω_{CK}= 0.20635

p, the number of free parameters; **κ**, the transition/transversion ratio; TL, the tree length.
Parameters ω_{CK}, ω_T and ω_{NS2} are the dN/dS ratios for the ancestral branch of *C. kahawae*, the terminal branches leading to *C. fructicola*, *C. siamense* and *C. asianum*, and the terminal branch leading to *Colletotrichum* sp. 2, respectively.

M1a and M2a nested models was unable to reject the null hypothesis of no positive selection, the comparison between M7 and M8 nested models detected the presence of positive selection at two specific sites (Sites 131 and 133), with a marginally significant p-value (P = 0.070). The posterior probabilities of those sites being under positive selection were 0.984 and 0.955, using the Naïve Empirical Bayes (NEB)(Fig. 5), and 0.808 and 0.730, using Bayes Empirical Bayes (BEB), for each site, respectively.

Using branch models of evolution, two specific hypothesis of *MAT1-2-1* evolution along the branches of the phylogenetic tree were tested. Our first hypothesis, concerning whether positive selection would have occurred along the branch that leads to *C. kahawae*, was not supported by the data (P > 0.05) when compared to the one-ratio model. Our second

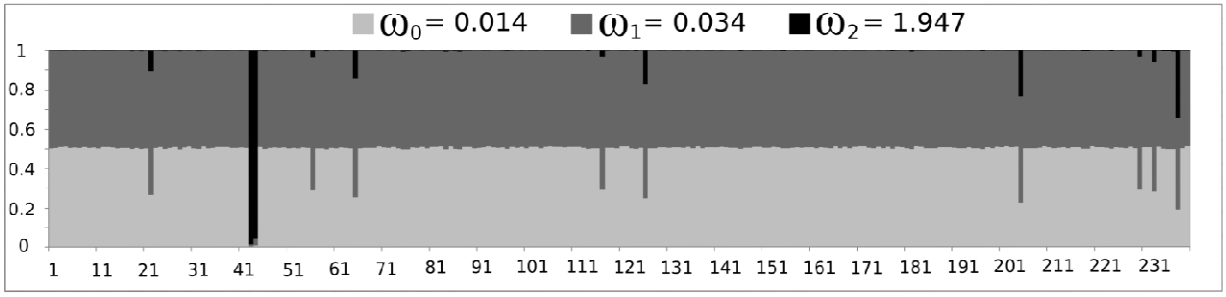

Fig 5 Posterior probabilities of site classes with different ω ratios along the *MAT1-2-1* sequence, according to the estimates of the M8 model using the Naïve Empirical Bayes (NEB) method. Two sites on the 42nd and 43rd aminoacids were shown to be under positive selection (ω_2 = 1.947) with posterior probabilities of 0.984 and 0.955, respectively. Using the Bayes Empirical Bayes (BEB) method, the posterior probabilities for both sites were somewhat lower, with 0.808 and 0.730 values, respectively.

hypothesis was based on the observation of the *MAT1-2-1* haplotype distribution and suggested that the branches leading to the closely related species *C. asianum*, *C. fructicola*, *C. siamense* and the *Colletotrichum* sp. 2, were under a significant relaxation of the general selective pressure operating on the gene. This hypothesis was supported by the data, with $\omega_T = \omega_{SP2}$=0.162 and ω_0=0.020 (P < 0.01) (Fig. 6). It should be noted that the branch leading to the *Colletotrichum* sp. 2 contains two non-synonymous site changes, one of which was suggested as being under positive selection by the M8 model and thus, it may be strongly influencing the result. However, when only the three other terminal branches were considered, the model was still supported by the data with a marginally significant p-value (P = 0.069; ω_T = 0.110) (Fig. 6), albeit with a much lesser ω estimate than when the *Colletotrichum* sp. 2 branch is considered alone (P = 0.008; ω_{SP2} = 0.429) (Fig. 6). Given this heterogeneity, a three-ratio model was tested, which adds an independent ω estimate for the *Colletotrichum* sp. 2 branch, providing a significant better fit to the data than the previous two-ratio models (P < 0.05; ω_0 = 0.0294, ω_T = 0.1102, ω_{SP2} = 0.472) (Fig. 6). Therefore, our data supports the hypothesis that these branches are under a significant and heterogeneous relaxation of the selective pressure.

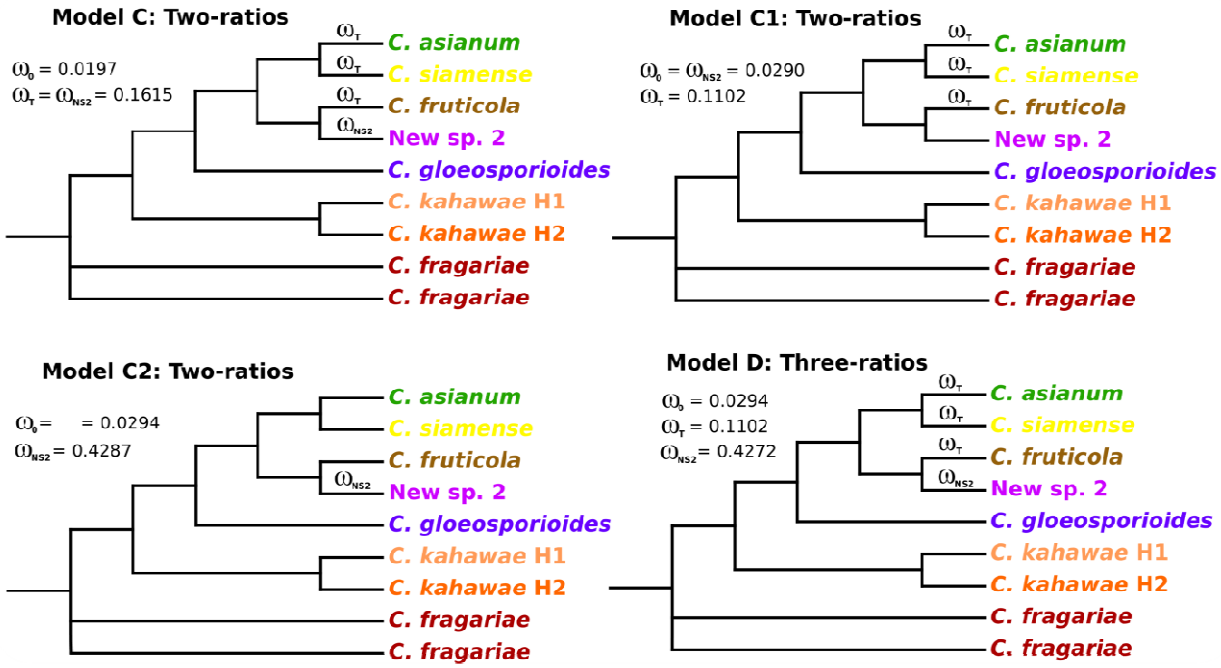

Fig 6 ML phylogenetic reconstruction for the Branch speficic models, using the six gene concatenated data set with only species/haplotype representatives. Only significantly better models than the one-ratio model are shown and detailed characteristics of each model are available on Table 2. Under the two-ratio and three-ratio models, branches whose ω was estimated as a free parameter were targeted by ω_T or ω_{NS2} above the respective branch.

Discusion

On the phylogenetic origin of C. kahawae

To our knowledge, this study documents the first comprehensive analysis of the evolutionary relationships of *C. kahawae* with *C. gloeosporioides sensu lato* species, mainly from coffee hosts worldwide, aiming at enlightening its phylogenetic origin based on previous longstanding hypotheses. One of the most striking finding, however, was that none of the initial hypotheses could be supported by our results in several aspects. First, all *C. gloeosporioides sensu lato* isolates sampled from *Coffea* spp. were highly divergent from the *C. kahawae* clade, which is inconsistent with a scenario of recent evolution either by mutation and adaptation from a sympatric *C. gloeosporioides* population causing brown blight (Nutman & Roberts 1960) or hybridization (Robinson 1976). Both these scenarios are liable to occur and have been proposed to underlie the recent emergence of fungal epidemics, such as *Batrachochytrium dendrobatidis* (Fisher *et al.* 2009), the cause of chytridiomycosis in

amphibians, and *Phytophthora alni* in Alder (Ioos *et al.* 2006), but the source populations or parental species are generally much more closely related, specially if they inhabit the same

Table 3 Likelihood Ration Tests (LRT) results

LRT	2Δ*l*	df	P-value
Branch-Models			
Model A vs B	25.104	14	0.0336
Model A vs C	10.463	1	0.0012
Model A vs C1	3.311	1	0.0688
Model A vs C2	7.053	1	0.0079
Model C2 vs E	4.956	1	0.0260
Model C1 vs E	8.699	1	0.0032
Model A vs E	0.924	1	0.3363
Sites Models			
Model A vs B	189.512	1	0.0000
Model A vs D	7.883	2	0.0194
Model C vs E	0.479	2	0.7871
Model F vs G	5.293	2	0.0709

2Δ*l*, log likelihood difference between models; df, degrees of freedom

host species and the event was recent. However, even though our *C. gloeosporioides* sampling is extensive (covering four continents), it is not intensive and thus, it would be possible that such populations or species would remain unsampled. Moreover, the hypothesis of emergence from one of the diploid progenitors of *C. arabica*, such as *C. canephora* or *C. eugenioides* (Robinson 1976), cannot be disregarded as we did not sample from *C. eugenioides*. The second line of evidence comes to fill these gaps by revealing the existence of an extremely similar phylogenetic lineage to the *C. kahawae* clade, sampled from *M. indica*. In fact, this isolate has diverged from the *C. kahawae* clade so recently that some of our markers could not distinguish them, although these are quite different biological entities as revealed by the inability of *Colletotrichum* sp. 1 to cause CBD symptoms. Our data suggests that this new lineage may be much closer to the phylogenetic source of *C. kahawae* epidemics, than those

sampled from coffee hosts, and even if it were present on coffee hosts but remained unsampled, our results show that it is not confined to the *C. gloeosporioides* gene pool of *Coffea* sp., as previously thought. In addition to this evidence, we also showed that *C. gloeosporioides sensu lato* strains from *Fragariae* sp. were very closely related to *C. kahawae*, based on published ITS sequences (Polashock *et al.* 2009). Although the authors noticed this close resemblance, the lack of a representative sampling of *Colletotrichum* spp. from coffee hosts as well as the sole employment of the ITS region in the phylogenetic analysis, hindered a clearer interpretation of the results.

Taken together and in the scope of our sampling, the closest lineages to *C. kahawae* were represented by isolates from hosts other than *Coffea* spp, which tantalizingly suggests the alternative hypothesis that a host shift/jump from other host may underlie the emergence of *C. kahawae*. Indeed, over the last years, a large body of evidence has accumulated with examples of emergent epidemics following a host shift (Slippers *et al.* 2005; Woolhouse *et al.* 2005; Stukenbrock & McDonald 2008; Giraud *et al.* 2010). It has been argued that host shift speciation is one of the main routes for emergence of new fungal diseases in plants and there are several factors in the life history of fungi that may facilitate this process, even accounting for the usual restrictions of ecological speciation (Giraud *et al.* 2010). In particular, host specialization may have been a major factor in the emergence of *C. kahawae* as it is the only known *Colletotrichum* species that is able to consistently colonize and infect green coffee berries (Waller *et al.* 1993; Chen *et al.* 2005; Manuel *et al.* 2009). Since most of the Ascomycetes, including the *Colletotrichum* genus (Cisar & TeBeest 1999), mate on their hosts or substrate after mycelia development, this means that mutations providing adaptation to a new host, or host part, will pleotropically affect local adaptation and matting patterns, hence providing an efficient barrier to gene flow and promoting speciation (Giraud *et al.*

2010). One example of such scenario is provided by *Venturia inaequalis*, causing scab disease on apples and pyracantha, whose natural populations on each host are highly differentiated despite being sympatric and interfertile in laboratory conditions (Cam *et al.* 2002; Gladieux *et al.* 2010). Another example in which host specialization prevents gene flow is presented by the Ascochyta fungi causing blights of chickpea, faba bean, lentil and pea (Peever 2007). However, Future research is certainly necessary to unravel such a complex event as speciation, and a new direction in pursuing the phylogenetic origin of *C. kahawae* will be focused on a broader host range and will incorporate additional analysis such as cross-pathogenicity tests.

Phylogeography of C. kahawae – *bound to the heights*

Regarding relationships within the *C. kahawae* clade, we have provided the first consistent and unambiguous population structure of *C. kahawae*, revealed by polymorphisms in nuclear gene markers. The overall low genetic variability supports the hypothesis of a recent emergence but, based on this data, three slightly divergent populations, yet highly correlated with their geographic distribution were found: Angola, Cameroon and East Africa. These results confirm previous indications of some geographic structuring, particularly between east and West African populations (Bella Manga *et al.* 1997; Bridge *et al.* 2008), and were able to resolve the western populations even further. Strikingly, the ancestral state inference suggests that the Angola population is the most ancestral lineage among the studied isolates, while those from Kenya and the remaining East African countries clustered together as the most derived lineage. This provides an alternative view for the geographic origin of *C. kahawae*, while, at the same time, shows that the current understanding of this event is considerably less parsimonious in light of our results. It is worth noting that this current understanding is based on historical data, such as disease reports and field observations and follows the premise that

C. kahawae evolved in coffee hosts, which offers very little support for an Angolan origin as the Arabica coffee culture on this country is relatively recent (McDonald 1926; Hindorf 1970; Firman & Waller 1977). However, as we suggested above, the ancestral population of *C. kahawae* may have emerged from hosts other than *Coffea* spp., which may by-pass the difficulties in reconciling our results with the current historical data. Speciation via host-shift/jump can be practically instantaneous and does not require the long timescales involved in the emergence of pathogens through co-evolution with its host (Giraud *et al.* 2010). Adding the fact that the transport of infected plant material has reached an unprecedented global scale (Stukenbrock & McDonald 2008), this greatly increases the potential for host-shift/jumps of pathogens, for example, of the ancestral population of *C. kahawae* into the newly arrived *C. arabica* plants in Angola. Moreover, information obtained from historical data can be flawed because it is biased towards regions where the disease densities are higher, and where substantial scientific effort has been focused on monitoring plant diseases (Fisher *et al.* 2009; Estoup & Guillemaud 2010).

Furthermore, this unexpectedly organized population structure reveals no evidence of present migration between the geographical locations of each haplotype, or the migration rate is too low to be detected. This may be explained by seldom sequential introductions with subsequent geographic isolation. Arabica coffee growing areas in Angola, Cameroon and East African are separated by extensive lowland areas, which are not suitable for the pathogen or the host (Firman & Waller 1977; Waller & Masaba 2006; van der Vossen 2009), thus representing a potential effective barrier to migration. In such scenario, bottleneck events during these seldom introductions can generate drift pulses in each of the introduced populations, which become genetically differentiated from each other whilst retaining their source-introduction relationship (Desprez-Loustau *et al.* 2007; Estoup & Guillemaud 2010;).

Accordingly, our results suggest that after a hypothetical origin of the Angola population, an introduction in the Cameroon followed and from there to the east Africa countries, while each of the established populations remained isolated to their respective highland areas. However, in invasion biology, evolutionary scenarios are often characterized by small divergence times, which may decrease the likelihood of identifying the true source-sink relationship between populations due to the stochasticity of the process (Estoup & Guillemaud 2010). Thus, even though our dataset suggest an Angolan origin for *C. kahawae*, sampling more isolates and polymorphic loci will certainly provide a much more reliable and robust insight on the evolution of *C. kahawae*'s populations.

Molecular evolution of MAT1-2-1

Despite the singularity of the genetic control underlying sexual development in the *Colletotrichum* genus, the results from our analysis of the mating-type gene *MAT1-2-1* are in agreement with those obtained by O'Donnell *et al.* (2004) for the evolution of mating-type genes in *Fusarium* spp., which follow the standard MAT configuration. In our study, the *MAT1-2-1* gene seemed to be under an intense purifying selection along most of its sequence and phylogenetic lineages of the studied species, indicating that its evolution is constrained by the need to maintain functionality. Unexpectedly, not only the genetically homogeneous *C. kahawae* possessed two haplotypes (H2 for the Angola population and H3 for the Cameroon and East African populations) but also shared one with the closely related *Colletotrichum* sp. 1. Hitherto we are not aware of significant phenotypic differences between populations carrying each *MAT1-2-1* haplotype, despite several studies have addressed isolates from these populations at the pathogenic (Rodrigues *et al.* 1992; Várzea *et al.* 1993) and vegetative compatibility grouping level (Várzea *et al.* 2002). However, taking into account the phylogeographical processes described above for this species, changes in phenotypic traits

during dispersal, establishment and range expansion of its populations may reflect neutral phenotypic changes rather than adaptive evolution (Keller & Taylor 2008). Random sampling of the genetic diversity of the source population, the sudden and dramatic decrease in population size during the introduction phase, and the low density of the introduced population during the establishment phase may lead to such intense drift pulses that phenothypic variation may become fixed purely by chance (Estoup & Guillemaud 2010). Furthermore, the haplotype sharing between the two different biological species, *C. kahawae* and *Colletotrichum* sp. 1, suggests that this gene was not important in the speciation process.

Regarding the *C. gloeosporioides* complex, purifying selection alone is unlikely to explain the observed pattern of the haplotype distribution. It seems that the H4 haplotype has remarkably endured since the split of *C. gloeosporioides sensu stricto* and the remaining *Colletotrichum* spp. from coffee hosts until the present time, possibly due a relevant functional role. Only more recently are the distinct species among the "gloeosporioides" complex beginning to diverge from this main haplotype, as revealed by the selective pressure relaxation at the terminal branches, but some are still retaining the ancient H4 haplotype. However, further studies on the functionality of these proteins and respective haplotypes will be essential to fully understand and integrate this knowledge in the species evolution as well as to comprehend the role of the specific sites that were suggested as being under positive selection. It is hoped that this study will contribute to subsequent research of the MAT gene evolution on the unique *Colletotrichum* genus.

Final remarks

World's globalization and agro-ecosystems, such as coffee crops, are providing a prone environment to the emergence of pathogens through mechanisms that would be rare in natural ecosystems. Several aspects in our results suggest that *C. kahawae* may have emerged

through a host-shift/jump in a fairly recent timescale, from an ancestral population from hosts other than *Coffea* spp., and several examples show that this type of event can be rather swift. However, given the existence of an extremely similar *C. gloeosporioides sensu lato* strain to isolates of *C. kahawae*, even when using a highly informative molecular dataset, we highlight the usefulness of pathogenicity tests on green coffee berries as a fairly reliable method for purposes of identifying genuine *C. kahawae* isolates. The overall genetic uniformity of *C. kahawae* supports a recent emergence but our results suggest that this event may have taken place in an alternative region to that previously hypothesized, such as Angola. Although based on a limited sampling, the population structure of *C. kahawae* was fragmented, possibly due to the topography of the African continent, and revealed a clonal pattern. However, even though pathogens with asexual and fragmented populations are regarded as having a low evolutionary potential, there are striking examples of similar pathogens uprising with patterns of sexual reproduction and higher evolutionary potential when faced with more efficient control measures (Stergiopoulos *et al.* 2007). Moreover, the analysis of the mating-type gene revealed two apparent functional haplotypes within *C. kahawae*, although it is not clear whether this divergence event is due to neutral phenotypic variation or adaptive evolution. Future research is essential and will be carried out in order to continue unraveling the complex events that led to the emergence of this aggressive pathogen.

Acknowledgements

At FCUL we thank our colleagues, Ana Vieira and Tiago Jesus, for invaluable contributions and criticisms during the elaborations of this manuscript. We are also grateful to Francisco Pina-Martins for helping in the first steps of the data analysis. At CIFC/IICT we appreciate the technical support provided by Sandra Sousa Emídio. At ISA, we appreciate Ana Paula Ramos for providing us the Cg432 isolates.

References

Aylor DL, Price EW, Carbone I (2006) SNAP: combine and map modules for multilocus population genetic analysis. *Bioinformatics*, **22**, 1399–1401.

Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16**, 37–48.

Beynon S, Várzea VMP, (1995) Genetic variation in the Coffee Berry Disease pathogen, *Colletotrichium kahawae*. *Physiological and Molecular Plant Pathology*, **46**, 457-470.

Bella Manga, Bieysse D, Mouen Bedimo JA *et al.* (1997) Observations sur la diversité de la population de *Colletotrichum kahawae* agent de l'anthracnose des baies du caféier Arabica. In *Proceedings of the 17th International Conference on Coffee Science (ASIC) Nairobi, Kenya*, pp 604-612.

Bridge P, Waller JM, Davies D, Buddie, AG (2008). Variability of *Colletotrichum kahawae* in relation to other *Colletotrichum* species from tropical perennial crops and the development of diagnostic techniques. *Journal of Phytopathology*, **156**, 274-280.

Brown A, Sreenivasaprasad S, Timmer L (1996) Molecular characterization of slow-growing orange and key lime anthracnose strains of *Colletotrichum* from *Citrus* as *C. acutatum*. *Phytopathology*, **86**, 523-527.

Cai L, Hyde KD, Taylor PWJ (2009) A polyphasic approach for studying *Colletotrichum*. *Fungal Diversity*, **39**, 183-204.

Cannon PF, Buddie AG, Bridge PD (2008) The typification of *Colletotrichum gloeosporioides*. *Mycotaxon*, **104**, 189-204.

Chen F, Goodwin PH, Khan A, Hsiang T (2002) Population structure and mating-type genes of *Colletotrichum graminicola* from *Agrostis palustris*. *Canadian Journal of Microbiology*, **48**, 427-436.

Chen Z, Liang J, Rodrigues C (2005) *Colletotrichum gloeosporioides* can overgrow *Colletotrichum kahawae* on green coffee berries first inoculated with *C. kahawae*. *Biotechnology Letters*, **27**, 679-682.

Cisar C & TeBeest D (1999) Mating system of the filamentous ascomycete, *Glomerella cingulata*. *Current Genetics*, **35**, 127-133.

Crouch J, Beirn LA (2009) Anthracnose of cereals and grasses. *Fungal Diversity*, **39**, 19-44.

Crouch J, Thon M, Clarke B (2007) Genomic architecture of the mating-type gene cluster in graminicolous species of the genus *Colletotrichum* and across the Ascomycota. In *2007 APS Annual Meeting*, p. S25.

Derso E, Waller J (2003) Variation among *Colletotrichum* isolates from diseased coffee berries in Ethiopia. *Crop Protection*, **22**, 561-565.

Desprez-Loustau M, Robin C, Buée M (2007) The fungal dimension of biological invasions. *Trends in Ecology & Evolution*, **22**, 472-480.

Du M ,Chardl CL, Nuckles EM, Vaillancourt, LJ (2005) Using mating-type gene sequences for improved phylogenetic resolution of *Collectotrichum* species complexes. *Mycologia*, **97**, 641-658.

Eastman JR (2006) IDRISI Andes (Worcester, MA: Clark University).

Estoup A, Guillemaud T (2010) Reconstructing routes of invasion using genetic data: why, how and so what? *Molecular Ecology*, **19**, 4113-4130.

Firman I, Waller J (1977) Coffee Berry Disease and other *Colletotrichum* diseases of Coffee. *Phytopathological Papers*, **20**, 1-53.

Fisher MC, Garner TWJ, Walker SF (2009) Global emergence of *Batrachochytrium dendrobatidis* and amphibian chytridiomycosis in space, time, and host. *Annual Review of Microbiology*, **63**, 291-310.

García-Serrano M, Laguna EA, Rodríguez-Guerra R, Simpson J (2008) Analysis of the *MAT1-2-1* gene of *Colletotrichum lindemuthianum*. *Mycoscience*, **49**, 312-317.

Gichuru EK, Várzea VMP, Rodrigues C, Masaba DM (2000). Vegetative Compatibility Grouping of *Colletotrichum kahawae* in Kenya. *Journal of Phytopathology*, **148**, 233-237.

Giraud T, Gladieux P, Gavrilets S (2010) Linking the emergence of fungal plant diseases with ecological speciation. *Trends in Ecology & Evolution*, **25**, 387-395.

Gladieux P, Caffier V, Devaux M, Cam B (2010) Host-specific differentiation among populations of *Venturia inaequalis* causing scab on apple, pyracantha and loquat. *Fungal Genetics and Biology*, **47**, 511-21.

Gordon W (1988) *Coffee. Tropical Agriculture series*, Wiley.

Hindorf H (1970) *Colletotrichum* spp. isolated from *Coffea arabica* L. in Kenya. *Zeitschrift fur Pflanzenkrankheiten und Pflanzenschutz*, **77**, 328-331.

Ioos R, Andrieux A, Marçais, B (2006) Genetic characterization of the natural hybrid species *Phytophthora alni* as inferred from nuclear and mitochondrial DNA analyses. *Fungal Genetics and Biologyi*, **43**, 511-29.

Katoh K, Asimenos G, Toh H (2009) Bioinformatics for DNA Sequence Analysis. *Methods in Molecular Biology*, **537**, 134-166.

Keller SR, Taylor DR (2008) History, chance and adaptation during biological invasion: separating stochastic phenotypic evolution from response to selection. *Ecology Letters*, **11**, 852-66.

Kerényi Z, Moretti A, Waalwijk C (2004) Mating type sequences in asexually reproducing *Fusarium* species. *Applied and Environmental Microbiology*, **70**, 4419-4423.

Kronstad J, Staben C (1997) Mating type in filamentous fungi. *Annual Review of Genetics*, **31**, 245-276.

Cam B, Parisi L, Arene L (2002) Evidence of two formae speciales in *Venturia inaequalis*, responsible for apple and pyracantha scab. *Phytopathology*, **92**, 314-320.

Manuel L, Talhinhas P, Várzea VMP, Neves-Martins J (2009) Characterization of *Colletotrichum kahawae* isolates causing Coffee Berry Disease in Angola. *Journal of Phytopathology*, **158**, 310-313.

McDonald J (1926) A preliminary account of a disease of green coffee berries in Kenya. *Transactions of the British Mycological Society*, **11**, 145-154.

Nguyen PTH, Petterson OV, Olsson P, Liljeroth E (2009) Identification of *Colletotrichum* species associated with anthracnose disease of coffee in Vietnam. *European Journal of Plant Pathology*, **127**, 73-87.

Nutman F, Roberts F (1960). Investigations on a disease of Coffea arabica caused by a form of *Colletotrichum coffeanum* Noack. I. Some factores affecting infection by the pathogen. *Transactions of the British Mycological Society*, **43**, 489-505.

Nylander JAA (2004) MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Omondi C, Ayiecho PO, Mwang'ombe AW, Hindorf H (2000) Reaction of some *Coffea arabica* genotypes to strains of *Colletotrichum kahawae*, the cause of Coffee Berry Disease. *Journal of Phytopathology*, **148**, 61-63.

Omondi C, Hindorf H, Welz H *et al.* (1997) Genetic diversity among isolates of *Colletotrichum kahawae* causing Coffee Berry Disease. In *Proceedings of the 17th International Conference on Coffee Science (ASIC) Nairobi, Kenya*. pp. 800-804.

O'Donnell K, Ward TJ Geiser DM *et al.* (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 Genetics and Biology*, **41**, 600-623.

O'Donnell K, Cigelnik E, (1997) Two divergent intragenomic rDNA ITS2 types within a monophyletic lineage of the fungus *Fusarium* are nonorthologous? *Molecular Phylogenetics and Evolution*, **7**, 103-116.

Peever T (2007) Role of host specificity in the speciation of Ascochyta pathogens of cool season food legums. *European Journal of Plant Pathology*, **119**, 119-126.

Pina-Martins F, Paulo OS (2008) Concatenator: sequence data matrices handling made easy. *Molecular Ecology Resources*, **8**, 1254–1255.

Polashock JJ, Caruso FL, Oudemans PV*et al.* (2009) The North American cranberry fruit rot fungal community: a systematic overview using morphological and phylogenetic affinities. *Plant Pathology*, **58**, 1116-1127.

Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics*, **14**, 817-818.

Price EW, Carbone I (2005) SNAP: workbench management tool for evolutionary population genetic analysis. *Bioinformatics*, **21**, 402–404.

Prihastuti H, Cai L, Chen H, *et al.* (2009) Characterization of *Colletotrichum* species associated with coffee berries in northern Thailand. *Fungal Diversity*, **39**, 89-109.

Rambaut A, Drummond AJ (2007) Tracer v1.4, Available from http://beast.bio.ed.ac.uk/Tracer

Robinson RA (1976) *Plant Pathosystems*, Springer-Verlag.

Robinson RA (1974). Terminal report of the FAO coffee pathologist to the government of Ethiopia. In *FAO, Rome, AGO/74/443*. pp. 16.

Rodrigues C, Várzea VMP, Medeiros E (1992) Evidence for the existence of physiological races of *Colletotrichum coffeanum* Noack *sensu* Hindorf. *Kenya Coffee*, **57**, 1417-1420.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572-1574.

Rydholm C, Dyer P, Lutzoni F (2007) DNA sequence characterization and molecular evolution of MAT1 and MAT2 mating-type loci of the self-compatible ascomycete mold *Neosartorya fischeri*. *Eukaryotic cell*, **6**, 868-874.

Schultes R (1986) *Coffee: Botany, Biochemistry and Production of Beans and Beverage* (Clifford MN, Willson KC (eds)). Springer.

Shimodaira H, Hasegawa M, (1999) Multiple comparisons of loglikelihoods with applications to phylogenetic inference. *Molecular Biology and Evolution*, **16**,1114–1116.

Silva M, Várzea VMP, Guerra-Guimarães L (2006) Coffee resistance to the main diseases: Leaf Rust and Coffee Berry Disease. *Brazilian Journal of Plant Physiology*, **18**, 119-147.

Slippers B, Stenlid J, Wingfield MJ (2005) Emerging pathogens: fungal host jumps following anthropogenic introduction. *Trends in Ecology & Evolution*, **20**, 420-421.

Sreenivasaprasad S, Brown A, Mills P (1993). Coffee Berry Disease pathogen in Africa: genetic structure and relationship to the group species *Colletotrichum gloeosporioides*. *Mycological Research*, **87**, 995-1000.

Stergiopoulos I, Groenewald M, Staats M, *et al.* (2007) Mating-type genes and the genetic structure of a world-wide collection of the tomato pathogen *Cladosporium fulvum*. *Fungal Genetics and Biology*, **44**, 415-429.

Stukenbrock E, McDonald B (2008) The origins of plant pathogens in agro-ecosystems. *Annual Review of Phytopathology*, **46**, 75-100.

Swofford DL (2000) PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods). Sinauer & Associates, Sunderland, Massachusetts.

Turgeon B (1998) Application of mating type gene technology to problems in fungal biology. *Annual Review of Phytopathology*, **36**, 115-137.

Vaillancourt L, Du M, Wang J, *et al.* (2000) Genetic analysis of cross fertility between two self-sterile strains of *Glomerella graminicola*. *Mycologia*, **92**, 430-435.

van der Vossen HAM (2009) The cup quality of disease-resistant cultivars of Arabica coffee (*Coffea arabica*). *Experimental Agriculture*, **45**, 323-332.

Várzea VMP, Rodrigues C, Lewis B (2002) Distinguishing characteristics and vegetative compatibility of *Colletotrichum kahawae* in comparison with other related species from coffee. *Plant Pathology*, **51**, 202- 207.

Várzea VMP, Rodrigues C, Medeiros E (1993) Different pathogenicity of CDB isolates on coffee genotypes. In *Proceedings of the 15th International Conference on Coffee Science (ASIC) Montpellier, France.* pp. 303-308.

Voigt K, Cozijnsen AJ, Kroymann J, *et al.* (2005) Phylogenetic relationships between members of the crucifer pathogenic *Leptosphaeria maculans* species complex as shown by mating type (MAT1-2), actin, and beta-tubulin sequences. *Molecular Phylogenetics and Evolution*, **37**, 541-557.

Waller JM, Bigger M, Hillocks RJ (2007) *Coffee Pests, Diseases and their Management*. CABI Publishing.

Waller JM, Bridge PD, Black R, Hakiza G (1993) Characterization of the Coffee Berry Disease pathogen, *Colletotrichum kahawae* sp. nov. *Mycological Research*, **97**, 989-994.

Waller J, Masaba D (2006) The microflora of coffee surfaces and relationships to Coffee Berry Disease. *International Journal of Pest Management*, **52**, 89-96.

Woolhouse MEJ, Haydon DT, Antia R (2005). Emerging pathogens: the epidemiology and evolution of species jumps. *Trends in Ecology & Evolution*, **20**, 238-244.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**, 1586-91.

Supplementary material

Table S1 List of the isolates with information regarding their species, host and origin and GenBank accession numbers

Isolate	Species/Group	Host	Country/Region	GenBank Acession Nº					
				ITS	b-tub2	ApMAT	MAT1-2-1	MAT5L	Apn15L
Ang21	*C. kahawae*	*C. arabica*	Angola, Kwanza Sul, Amboim	---	---	---	---	---	---
Ang28	*C. kahawae*	*C. arabica*	Angola, Benguela, Ganda	---	---	---	---	---	---
Ang 29	*C. kahawae*	*C. arabica*	Angola, Benguela, Ganda	---	---	---	---	---	---
Ang30	*C. kahawae*	*C. arabica*	Angola, Benguela, Ganda	---	---	---	---	---	---
Ang56	*C. kahawae*	*C. arabica*	Angola, Benguela, Ganda	---	---	---	---	---	---
Ang70	*C. kahawae*	*C. arabica*	Angola, Benguela, Ganda	---	---	---	---	---	---
Ang81	*C. kahawae*	*C. arabica*	Angola, Huambo	---	---	---	---	---	---
Ang65	*C. kahawae*	*C. arabica*	Angola, Benguela, Ganda	---	---	---	---	---	---
Ang69	*C. kahawae*	*C. arabica*	Angola, Benguela, Ganda	---	---	---	---	---	---
Ang92	*C. kahawae*	*C. arabica*	Angola, Benguela, Ganda	---	---	---	---	---	---
Ang34	*C. kahawae*	*C. arabica*	Angola, Benguela, Ganda	---	---	---	---	---	---
Ang37	*C. kahawae*	*C. arabica*	Angola, Benguela, Ganda	---	---	---	---	---	---
Ang27	*C. kahawae*	*C. arabica*	Angola, Benguela, Ganda	---	---	---	---	---	---
Ang6	*C. kahawae*	*C. arabica*	Angola	---	---	---	---	---	---
Ang30	*C. kahawae*	*C. arabica*	Angola, Benguela, Ganda	---	---	---	---	---	---
Ang52	*C. fragariae*[+]	*C. arabica*	Angola, Benguela, Ganda	---	---	---	---	---	---
Ang84	*C. fragariae*[+]	*C. arabica*	Angola, Kwanza Sul, Amboim	---	---	---	---	---	---
Ang91	*C. fragariae*[+]	*C. arabica*	Angola	---	---	---	---	---	---
Ang40	*C. fructicola*[+]	*Coffea* sp.	Angola, Kwanza Sul, Amboim	---	---	---	---	---	---
Ang95	*C. siamense*[+]	*C. canephora*	Angola, Uíge, Est.Exp.Uíge	---	---	---	---	---	---
Ang96	*C. siamense*[+]	*C. canephora*	Angola, Uíge, Est.Exp.Uíge	---	---	---	---	---	---
Ang97	*C. fructicola*[+]	*C. canephora*	Angola, Uíge, Est.Exp.Uíge	---	---	---	---	---	---
Ang99	*C. siamense*[+]	*C. canephora*	Angola, Uíge, Est.Exp.Uíge	---	---	---	---	---	---
Ang100	*Colletotrichum* sp. 2[+]	*C. canephora*	Angola, Uíge, Est.Exp.Uíge	---	---	---	---	---	---
Ang101	*Colletotrichum* sp. 2[+]	*C. canephora*	Angola, Uíge, Est.Exp.Uíge	---	---	---	---	---	---
Bur2	*C. kahawae*	*C. arabica*	Burundi	---	---	---	---	---	---
Cam1	*C. kahawae*	*C. arabica*	Cameroon	---	---	---	---	---	---
Cam2	*C. kahawae*	*C. arabica*	Cameroon	---	---	---	---	---	---
Cam5	*C. kahawae*	*C. arabica*	Cameroon	---	---	---	---	---	---
Cam8	*C. kahawae*	*C. arabica*	Cameroon	---	---	---	---	---	---
Cam3	*C. kahawae*	*C. arabica*	Cameroon	---	---	---	---	---	---
Cam12	*C. kahawae*	*C. arabica*	Cameroon	---	---	---	---	---	---

Isolate	Species/Group	Host	Country	Region	GenBank Acession Nº					
					ITS	b-tub2	ApMAT	MAT1-2-1	MAT5L	Apn15L
Cam13	*C. kahawae*	*C. arabica*	Cameroon		---	---	---	---	---	---
Eti9	*C. kahawae*	*C. arabica*	Ethiopia		---	---	---	---	---	---
Eti17	*C. kahawae*	*C. arabica*	Ethiopia		---	---	---	---	---	---
Eti10	*C. kahawae*	*C. arabica*	Ethiopia		---	---	---	---	---	---
Eti20	*C. kahawae*	*C. arabica*	Ethiopia		---	---	---	---	---	---
Mal2	*C. kahawae*	*C. arabica*	Malawi		---	---	---	---	---	---
Mal5	*C. siamense*[+]	*Coffea sp.*	Malawi		---	---	---	---	---	---
Mal7	*C. kahawae*	*C. arabica*	Malawi		---	---	---	---	---	---
Mal9	*C. kahawae*	*C. arabica*	Malawi		---	---	---	---	---	---
Que2	*C. kahawae*	*C. arabica*	Kenya		---	---	---	---	---	---
Que9	*C. kahawae*	*C. arabica*	Kenya		---	---	---	---	---	---
Que42	*C. kahawae*	*C. arabica*	Kenya		---	---	---	---	---	---
Que48	*C. kahawae*	*C. arabica*	Kenya		---	---	---	---	---	---
Que72	*C. kahawae*	*C. arabica*	Kenya		---	---	---	---	---	---
CBD10	*C. kahawae*	*C. arabica*	Kenya, CRS, Kitale		---	---	---	---	---	---
CBD11	*C. kahawae*	*C. arabica*	Kenya, CRS, Kitale		---	---	---	---	---	---
CBD12	*C. kahawae*	*C. arabica*	Kenya, Mgumguri		---	---	---	---	---	---
CBD13	*C. kahawae*	*C. arabica*	Kenya, Mgumguri		---	---	---	---	---	---
CCA3	*C. siamense*[+]	*Coffea* sp.	Kenya, Azamia		---	---	---	---	---	---
CCA4	*C. siamense*[+]	*Coffea* sp.	Kenya, Azamia		---	---	---	---	---	---
CCM5	*C. siamense*[+]	*Coffea* sp.	Kenya, Koru		---	---	---	---	---	---
CCM6	*C. siamense*[+]	*Coffea* sp.	Kenya, Koru		---	---	---	---	---	---
CCM7	*C. siamense*[+]	*Coffea* sp.	Kenya, Koru		---	---	---	---	---	---
Rua1	*C. kahawae*	*C. arabica*	Rwanda		---	---	---	---	---	---
Tan8	*C. kahawae*	*C. arabica*	Tanzania		---	---	---	---	---	---
Tan12	*C. kahawae*	*C. arabica*	Tanzania		---	---	---	---	---	---
Tan13	*C. kahawae*	*C. arabica*	Tanzania		---	---	---	---	---	---
Tan17	*C. kahawae*	*C. arabica*	Tanzania		---	---	---	---	---	---
Tan18	*C. kahawae*	*C. arabica*	Tanzania		---	---	---	---	---	---
Tan3	*C. kahawae*	*C. arabica*	Tanzania		---	---	---	---	---	---
Tan10	*C. kahawae*	*C. arabica*	Tanzania		---	---	---	---	---	---
Tan16	*C. kahawae*	*C. arabica*	Tanzania		---	---	---	---	---	---
Tan22	*C. kahawae*	*C. arabica*	Tanzania		---	---	---	---	---	---

Isolate	Species/Group	Host	Country	Region	GenBank Accession					
					ITS	b-tub2	ApMAT	MAT1-2-1	MAT5L	Apn15L
Tan23	*C. kahawae*	*C. arabica*	Tanzania		---	---	---	---	---	---
Tan25	*C. kahawae*	*C. arabica*	Tanzania		---	---	---	---	---	---
Zim12	*C. kahawae*	*C. arabica*	Zimbabwe		---	---	---	---	---	---
Zim1	*C. kahawae*	*C. arabica*	Zimbabwe		---	---	---	---	---	---
Zim14	*C. kahawae*	*C. arabica*	Zimbabwe		---	---	---	---	---	---
Bra5	*C. fragariae*[†]	*Coffea* sp.	Brazil		---	---	---	---	---	---
Bra8	*C. fragariae*[†]	*Coffea* sp.	Brazil		---	---	---	---	---	---
Bra9	*C. siamense*[†]	*Coffea* sp.	Brazil		---	---	---	---	---	---
Col1	*C. siamense*[†]	*Coffea* sp.	Colombia		---	---	---	---	---	---
Col2	*C. siamense*[†]	*Coffea* sp.	Colombia		---	---	---	---	---	---
Thai1	*C. fragariae*[†]	*C. canephora*	Thailand, Chumporn		---	---	---	---	---	---
Thai2	*C. siamense*[†]	*C. arabica*	Thailand, Chiang Mai		---	---	---	---	---	---
Thai3	*C. siamense*[†]	*C. canephora*	Thailand, Chumporn		---	---	---	---	---	---
Thai4	*C. siamense*[†]	*C. arabica*	Thailand, Chiang Mai		---	---	---	---	---	---
LC0033	*C. frutícola*	*Coffea* sp.	Thailand, Chiang Mai		---	---	---	---	---	---
LC0035	*C. siamense*	*Coffea* sp.	Thailand, Chiang Mai		---	---	---	---	---	---
LC0037	*C. asianum*	*Coffea* sp.	Thailand, Chiang Mai		---	---	---	---	---	---
Chi4	*C. siamense*[†]	*Coffea* sp.	China		---	---	---	---	---	---
Cg21	*C. gloeosporioides*	*C. lemon*	Portugal		---	---	---	---	---	---
Cg111	*C. gloeosporioides*	*O. europaea*	Portugal		---	---	---	---	---	---
Cg220	*C. gloeosporioides*	*O. europaea*	Portugal		---	---	---	---	---	---
Cg432	*Colletotrichum* sp. 1[†]	*M. indica*	Portugal		---	---	---	---	---	---

[†] Isolates received as *C. gloeosporioides* and assigned to a new species based on the phylogenetic inference

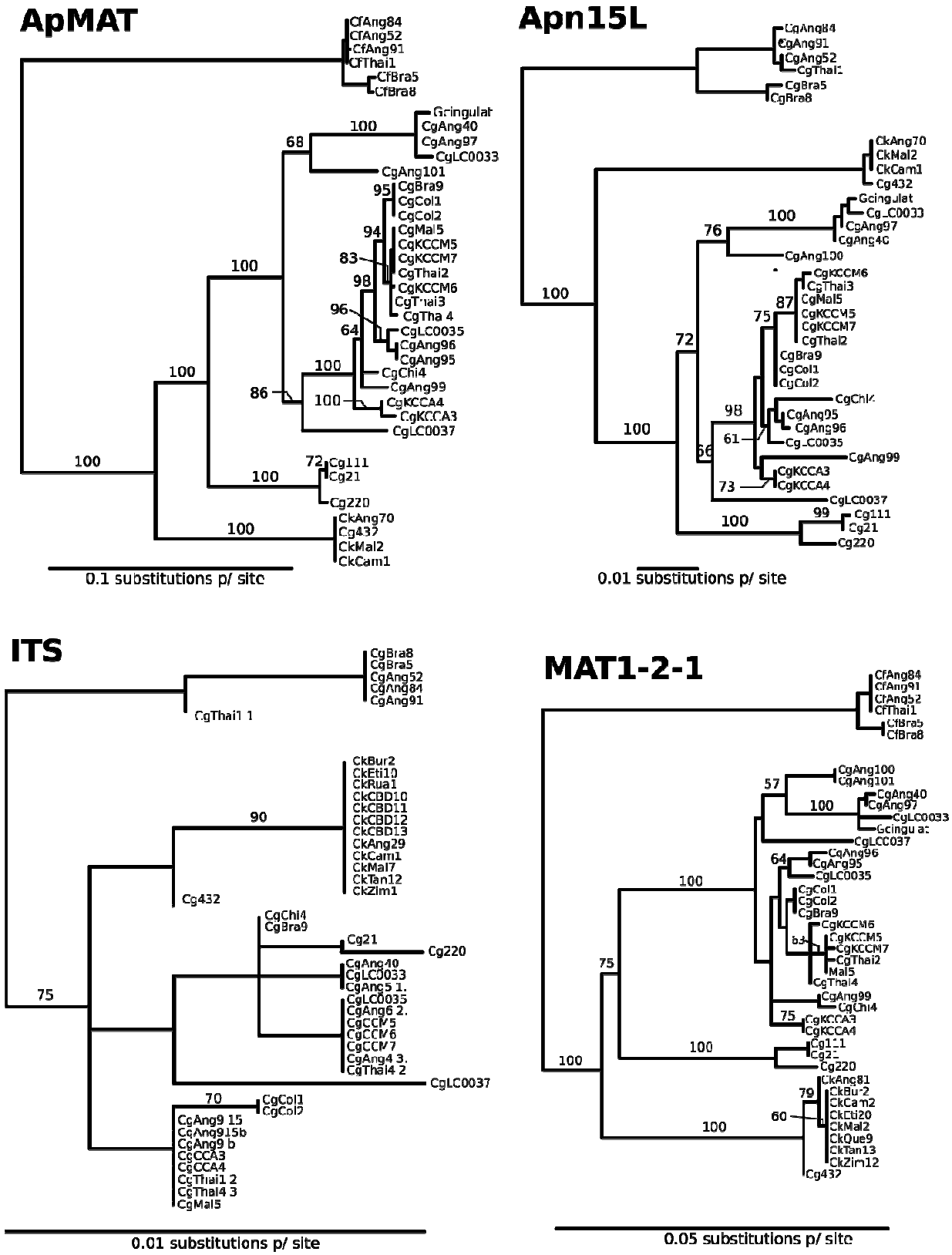

Fig S1 ML gene tree estimates for each individual dataset. Trees were rooted with *C. fragariae*. Bootstrap values are given above branches. **a)** ApMAT; **b)** Apn15L; **c)** ITS; **d)** *MAT1-2-1*; **e)** MAT5L; **f)** β-tub2

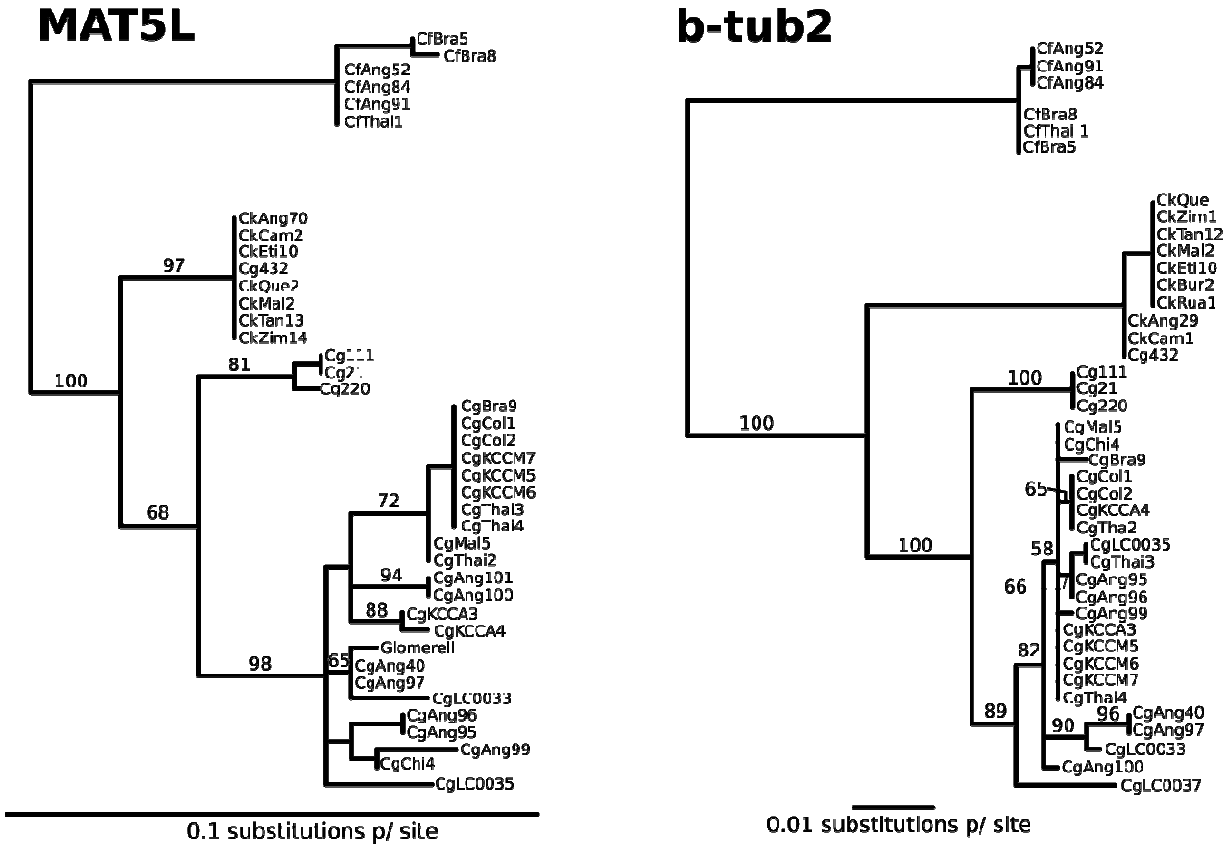

Fig S1 (Continuation)

Part III

Concluding Remarks

Plant pathogens are still emerging and will continue to do so in agro-ecosytems throughout the world [16,138-141]. Fungi (*sensu lato*) are responsible for ~30% of the emerging diseases in plants and they can impact negatively on human wellbeing through agricultural and economical losses [82,138]. However, advances in DNA analysis are allowing the reconstruction of the evolutionary history of pathogens and to understand the processes that caused their emergence [16,33].

In this work, the importance of such analyses is highlighted when addressing questions about the origin, structure and dissemination of the pathogen *C. kahawae*. First of all, the development of useful and informative markers is of vital importance, as they are responsible for revealing the patterns and structure that will allow inferences to be drawn about species and populations [108]. The development of molecular markers on the Apn1/MAT locus show promise in their informative potential and, hopefully, will open doors to address new questions that will have an impact on the whole *C. gloeosporioides* species complex. Several species within this complex have been considered distinct based on their pathogenicity, host range or certain morphological features, although phylogenetic studies with traditional markers have provided no basis to separate them from the *C. gloeosporioides* group, such as *C. musae* and *C. magna* [68,70]. Moreover, species within a complex are inherently closely related with relatively reduced divergence times, which may cause incongruities in the phylogenetic estimation of gene trees due to incomplete lineage sorting [96,142]. This is particularly true for less variable regions of the genome that take more time to coalesce and thus, have a higher probability of retaining ancestral polymorphisms between divergent taxa [143]. Acknowledging that the evolutionary history of these shallow species is greatly governed by stochastic processes, and employing methods that explicitly account for them in addition to highly informative markers, will be important in future research of the “gloeosporioides” complex.

The application of these and other markers to the problematic of *C. kahawae* emergence and evolution revealed fruitful, in that a new and more objective perspective was unearthed. This new perspective challenged the current vision of these events in a way that was not expected, suggesting not only a different place for the origin of *C. kahawae*, but also that the speciation event underlying its emergence from the group species *C. gloeosporioides* is much more complex than previously hypothesized. Even though the results presented here tantalizingly suggest an emergence of *C. kahawae* by host shift speciation, future research will be essential and the quest for the true phylogenetic origin of *C. kahawae*

is only beginning. On the other hand, the ability to resolve population relationships within this species allowed further inferences about how it has disseminated and how its populations are distributed throughout its range, revealing the importance of the altitude in their structure. This may also be useful for plant pathologists interested in breeding for resistance to CBD, who should integrate isolates from the three divergent populations whenever possible.

Finally, statistical methods of molecular evolution were used to investigate the only mating-type gene, *MAT1-2-1*, known in the singular *Colletotrichum* genus. Since this issue has remained largely unexplored, the presented results lacked a proper reference basis with which more general inferences could have been drawn. Nonetheless, the analysis itself may represent a starting point to generate hypothesis about the evolution and function of this apparent singular gene and mating system in the entire fungal kingdom.

References

1 Cronquist, A. (1988) *The evolution and classification of flowering plants,* New York Botanical Garden Press

2 Chevalier, A. (1947) *Les caféiers du globe. Encyclopédie Biologique. XXVIII: Fascicule III-Systématique des caféiers et faux-caféiers, maladies et insectes nuisibles,* Paul Lechevalier Editeur

3 Leroy J.F. (1982) L'origine Kenyane du genre *Coffea* L. et la radiation des espèces à Madagascar. In *Association Scientifique Internationale du Café (ASIC) 10th Colloque, Salvador-Bahia, Brazil*

4 Bridson D. (1994) Additional notes on *Coffea* (Rubiacea) from tropical East Africa. *Kew Bull.* 49, 331–342

5 Sonké, B. *et al.* (2006) A new dwarf *Coffea* (Rubiaceae) from southern Cameroon. *Bot. J. Linn. Soc.* 151, 425–430

6 Davis, A.P. and Mvungi, E.F. (2004) Two new and endangered species of *Coffea* (Rubiaceae) from the eastern arc mountains (Tanzania) and notes on associated conservation issues. *Bot. J. Linn. Soc.* 146, 237–245

7 Davis, A.P. and Rakotonasolo, F. (2000) Three new species of *Coffea* L. (Rubiaceae) from Madagascar. *Kew Bull.* 55, 405– 416

8 Davis A.P. (2003) A new combination in *Psilanthus* (Rubiaceae) for Australia, and nomenclatural notes on *Paracoffea*. *Novon* 13, 182–184

9 Davis A.P *et al.* (2006) An annotated taxonomic conspectus of the genus *Coffea* (Rubiaceae). *Bot. J. Linn. Soc.* 152, 465-512

10 Waller, J.M. *et al.* (2007) *Coffee Pests, Diseases and their Management,* CABI Publishing

11 Clifford, M.N. and Willson, K.C. (1985) *Botany, Biochemistry and Production of Beans and Beverage,* AVI Publishing Company, Inc

12 Gordon, W. (1988) *Coffee. Tropical Agriculture series,* Wiley

13 Lashermes, P. *et al.* (1999) Molecular characterization and origin of the *Coffea arabica* L. genome. *Mol. Gen. Genet.* 261, 259-266

14 Lashermes, P. *et al.* (1996) Phylogenetic relationships of coffee-tree species (*Coffea* L.) as inferred from ITS sequences of nuclear ribosomal DNA. *Theor. Appl. Genet.* 94, 947–955

15 Raina, S.N. *et al.* (1998) In situ hybridization identifies the diploid progenitor species of *Coffea arabica* (Rubiaceae). *Theor. Appl. Genet.* 97, 1204-1209

16 Stukenbrock E. and McDonald B. (2008) The origins of plant pathogens in agro-ecosystems. *Annu. Rev. Phytopathol.* 46, 75-100

17 Palumbi S. (2001) Humans as the world's greatest evolutionary force. *Science* 293, 1786-1790

18 Lécolier A. *et al.* (2009) Unraveling the origin of *Coffea arabica* 'Bourbon pointu' from La Réunion: a historical and scientific perspective. *Euphytica* 168, 1-10

19 Robinson, R.A. (1976) *Plant pathosystems,* Springer-Verlag

20 Ferrão, J. (1993) *A aventura das plantas e os descobrimentos portugueses*, Chaves Ferreira Publicações

21 Bigger M. (2006) The dissemination of coffee cultivation throughout the world. *Tropi. Agric. Assoc. Newsl* 26, 15-19

22 van der Vossen H.A.M. (2009) The cup quality of disease-resistant cultivars of Arabica coffee (*Coffea arabica*). *Exp. Agric.* 45, 323-323

23 Badr, A. *et al.* (2000) On the origin and domestication history of barley (*Hordeum vulgare*). *Mol. Biol. Evol.* 17, 499-510

24 Lev-Yadun, S. *et al.* (2006) How and when was wild wheat domesticated?. *Science*, 296 - 297

25 Anthony F. *et al.* (2002) The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. *Theor. Appl. Genet.* 104, 894-900

26 Friesen T. *et al.* (2006) Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat. Genet.* 38, 953-956

27 Butler G. *et al.* (2009) Evolution of pathogenicity and sexual reproduction in eight candida genomes. *Nature* 459, 657-662

28 Vega F. *et al.* (2003) Global project needed to tackle coffee crisis. *Nature* 425, 343-343

29 Osório N. (2002) The global coffee crisis: a threat to sustainable development. In *World summit on sustainable development*. International Coffee Organization

30 Zambolim L. *et al.* (2005) *Durable resistance to Coffee Leaf Rust*. Universidade Federal de Viçosa, Viçosa

31 Silva M. *et al.* (2006) Coffee resistance to the main diseases: Leaf Rust and Coffee Berry Disease. *Braz. J. Plant Physiol* 18, 119-147

32 Bridge P. *et al* (2008) Variability of *Colletotrichum kahawae* in relation to other *Colletotrichum* species from tropical perennial crops and the development of diagnostic techniques. *J of Phytopathol* 156, 274-280

33 Barrett L. *et al.* (2008) Life history determines genetic structure and evolutionary potential of host-parasite interactions. *Trends Ecol. Evol.* 23, 678-685

34 McDonald J. (1926) A preliminary account of a disease of green coffee berries in Kenya. *Trans. Br. Mycol. Soc.* 11, 145-154

35 Nutman F. and Roberts F. (1960) Investigations on a disease of *Coffea arabica* caused by a form of *Colletotrichum coffeanum* Noack. I. some factores affecting infection by the pathogen. *Trans. Br. Mycol. Soc.* 43, 489-505

36 Bella Manga (1997) Observations sur la diversité de la population de *Colletotrichum kahawae* agent de l'anthracnose des baies du cafíer arabica. *Association Scientifique Internationale du Café (ASIC) 17th Colloque, Nairobi, Kenya*

37 Firman I. and Waller J. (1977) Coffee Berry Disease and other *Colletotrichum* diseases of coffee. *Phytopathol. Pap.* 20, 1-53

38 Mulinge S.K. (1971) Effect of altitude on the distribution of the fungus causing Coffee Berry Disease in Kenya. *Ann. Appl. Biol.* 67, 93-98

39 Hindorf H. (1997) Correct identification of the pathogen *Colletotrichum kahawae* causing Coffee Berry Disease. *Association Scientifique Internationale du Café (ASIC) 17th Colloque, Nairobi, Kenya*

40 van der Vossen H.A.M. (2006) State-of-the-art of developing Arabica coffee cultivars with durable resistance to Coffee Berry Disease (*Colletotrichum kahawae*). *Association Scientifique Internationale du Café (ASIC) 17th Colloque, Montpellier, France*

41 Derso E. and Waller J. (2003) Variation among *Colletotrichum* isolates from diseased coffee berries in Ethiopia. *Crop Prot.* 22, 561-565

42 Manuel L. *et al.* (2010) Characterization of *Colletotrichum kahawae* isolates causing Coffee Berry Disease in Angola. *J. of Phytopathol.* 158, 310-313

43 Chen Z. *et al.* (2005) *Colletotrichum gloeosporioides* can overgrow *Colletotrichum kahawae* on green coffee berries first inoculated with *C. kahawae*. *Biotechnol. Lett.* 27, 679-682

44 Muller R.A. (1964) L'anthracnose des baies du caféier d'Arabie (*Coffea arabica*) due à *Colletotrichum coffeanum* Noack (sensu Hindorf), *Hemileia vastatrix* B. et Br., *Hemileia coffeicola* Maulanc et Roger. *Institut Français du Café et du Cacao, Montpellier, Bulletin* 6, 9-38

45 Nutman F. and Roberts F. (1960) Investigations on a disease of coffea arabica caused by a form of *Colletotrichum coffeanum* Noack. II. some factores affecting germination and infection by the pathogen. *Trans. Br. Mycol. Soc.* 43, 643-659

46 Gibbs J.N. (1969) Inoculum sources for Coffee Berry Disease. *Ann. Appl. Biol.* 64, 515-522

47 Waller J. (1972) Water-borne spore dispersal in Coffee Berry Disease and its relation to control. *Annu. Appl. Biol.* 71, 1-18

48 Fitt B. *et al.* (1989) The role of rain in dispersal of pathogen inoculum. *Ann. Rev. Phytopathol.* 27, 241-270

49 Pielaat A. and van den Bosch F. (1998) A model for dispersal of plant pathogens by rainsplash. *J. Math. Appl. Med. Biol.* 15, 117-134

50 Wilson L. *et al.* (1999) Comparison of rain effects on splash dispersal of three *Colletotrichum* species infecting strawberry. *Phytopathology* 89, 555-563

51 Schroth G. *et al.* (2000) Pests and diseases in agroforestry systems of the humid tropics. *Agrofor. Syst* 50, 199-241

52 Chung W. *et al.* (2006) Fungicide sensitivity and phylogenetic relationship of anthracnose fungi isolated from various fruit crops in Japan. *Plant Dis.* 90, 506-512

53 van den Bosch F. and Gilligan C. (2008) Models of fungicide resistance dynamics. *Annu. Rev. Phytopathol.* 46, 123-147

54 Mulinge S.K. (1972) Variation in the level of *Colletotrichum coffeanum* Noack in the bark of *Coffea arabica* cultivars. *Kenyan Coffee* 57, 9-47

55 Waller J.M. and Masaba D. (2006) The microflora of coffee surfaces and relationships to Coffee Berry Disease. *Int. J. Pest Manag.* 52, 89-96

56 Várzea, V.M.P. (1993) Different pathogenicity of CBD isolates on coffee genotypes. *Association Scientifique Internationale du Café (ASIC) 15th Colloque, Montpellier, France*

57 McDonald B. and Linde C. (2002) Pathogen population genetics, evolutionary potential, and durable resistance. *Annu. Rev. Phytopathol.* 40, 349-379

58 Hovmøller M. *et al.* (2008) Rapid global spread of two aggressive strains of a wheat rust fungus. *Mol. Ecol.* 17, 3818-3826

59 Wolfe M. and Mcdermott J. (1994) Population genetics of plant pathogen interactions: the example of the *Erysiphe graminis-Hordeum vulgare* pathosystem. *Annu. Rev Phytopathol* 32, 89-113

60 Hindorf H. (1970) *Colletotrichum* spp. isolated from *Coffea arabica* L. in Kenya. *Zeitschrift fur Pflanzenkrankheiten und Pflanzenschutz* 77, 328-331

61 Noack F. (1901) Die krankheiten das kaffeebaumes in brasilien. *Zeitschrift fur Pflanzenkrankheiten* 11, 202

62 Freeman S. *et al.* (1998) Characterization of *Colletotrichum* species responsible for anthracnose diseases of various fruits. *Plant Dis.* 82, 596-605

63 Waller J.M. *et al.* (1993) Characterization of the Coffee Berry Disease pathogen, *Colletotrichum kahawae* sp. nov. *Mycol. Res.* 97, 989-994

64 Prihastuti H. *et al.* (2009) Characterization of *Colletotrichum* species associated with coffee berries in northern Thailand. *Fungal Divers.* 39, 89-109

65 Kirk P.M. (1996) *Ainsworth's and bisby's dictionary of the fungi,* CABI Publishing

66 Taylor J. *et al.* (2000) Phylogenetic species recognition and species concepts in fungi. *Fungal Genet. Biol.* 31, 21-32

67 Seifert K. (2009) Progress towards DNA barcoding of fungi. *Mol. Ecol. Resour* 9, 83-89

68 Du M. *et al.* (2005) Using mating-type gene sequences for improved phylogenetic resolution of *Collectotrichum* species complexes. *Mycologia* 97, 641-658

69 Sutton, B.C. (1992) The genus *Glomerella* and its anamorph *Colletotrichum*. In *Colletotrichum - biology, pathology and control* (Bailey J.A. and Jeger M.J. (eds.)), pp. 1-26, CAB International Mycological Institute

70 Hyde K. *et al.* (2009) *Colletotrichum*: a catalogue of confusion. *Fungal Divers.*39, 1-17

71 Dettman J. *et al.* (2003) A multilocus genealogical approach to phylogenetic species recognition in the model eukaryote *Neurospora*. *Evolution* 57, 2703-2720

72 Fournier E. *et al.* (2005) Partition of the *Botrytis cinerea* complex in France using multiple gene genealogies. *Mycologia* 97, 1251-1267

73 Carstens B.C. and Dewey T.A. (2010) Species delimitation using a combined coalescent and information-theoretic approach: an example from North American *Myotis* bats. *Syst. Biol.* 59, 400-414

74 Cai L. *et al.* (2009) A polyphasic approach for studying. *Fungal Divers.* 39, 183-204

75 Edwards S.V. (2008) Is a new and general theory of molecular systematics emerging?. *Evolution* 63-1, 1-19

76 Hibbett D.S. *et al.* (2007) A higher-level phylogenetic classification of the fungi. *Mycol. Res.* 111, 509-547

77 Anderson J. and Kohn L. (1998) Genotyping, gene genealogies and genomics bring fungal population genetics above ground. *Trends Ecol. Evol.* 13, 444-449

78 O'Donnel K. *et al.* (2000) Gene genealogies reveal global phylogeographic structure and reproductive isolation among lineages of *Fusarium graminearum*, the fungus causing wheat scab. *Proc. Natl. Acad. Sci. U.S.A.* 97, 7905-7910

79 Crouch J. *et al.* (2009) Phylogenetic and population genetic divergence correspond with habitat for the pathogen *Colletotrichum cereale* and allied taxa across diverse grass communities. *Mol. Ecol.* 18, 123-135

80 Sreenivasaprasad S. *et al.* (1996) Phylogeny and systematics of 18 *Colletotrichum* species based on ribosomal DNA spacer sequences. *Genome* 39, 499-512

81 Lubbe C. *et al.* (2004) Characterization of *Colletotrichum* species associated with diseases of Proteaceae. *Mycologia* 96, 1268-1279

82 Giraud T. *et al.* (2010) Linking the emergence of fungal plant diseases with ecological speciation. *Trends Ecol. Evol.* 25, 387-395

83 von Arx J.A. (1957) Die arten der gattung colletotrichum cda. *Phytopathologische Zeitschrift* 29, 413–468

84 Phoulivong S. *et al.* (2010) *Colletotrichum gloeosporioides* is not a common pathogen on tropical fruits. *Fungal Divers.* 44, 33-43

85 Nguyen P.T.H. *et al.* (2009) Identification of *Colletotrichum* species associated with anthracnose disease of coffee in Vietnam. *Eur. J. Plant Pathol.* 127, 73-87

86 Afanador-Kafuri L. *et al.* (2003) Characterization of *Colletotrichum* isolates from tamarillo, passiflora, and mango in Colombia and identification of a unique species from the genus.. *Phytopathology* 93, 579-587

87 Denoyes-Rothan B. *et al.* (2003) Genetic diversity and pathogenic variability among isolates of *Colletotrichum* species from strawberry. *Phytopathology* 93, 219-228

88 Freeman S. *et al.* (2000) Molecular analyses of *Colletotrichum* species from almond and other fruits. *Phytopathology* 90, 608-614

89 Johnston P. and Jones D. (1997) Relationships among *Colletotrichum* isolates from fruit-rots assessed using rDNA sequences. *Mycologia* 89, 420-430

90 Crouch J. *et al.* (2009) What is the value of ITS sequence data in *Colletotrichum* systematics and species diagnosis? A case study using the falcate-spored graminicolous *Colletotrichum* group. *Mycologia* 101, 648-656

91 Hyde K.D. and Zhang Y. (2008) Epitypification: should we epitypify?. *J. Zhejiang Univ. Sci.* 9, 842-846

92 Cannon P.F. *et al.* (2008) The typification of *Colletotrichum gloeosporioides*. *Mycotaxon* 104, 189-204

93 Robinson R.A. (1974) Terminal report of the FAO coffee pathologist to the government of Ethiopia. In *TA3208, FAO, Rome, AGO/74/443*. pp. 16

94 Polashock J.J. *et al.* (2009) The North American cranberry fruit rot fungal community: a systematic

overview using morphological and phylogenetic affinities. *Plant Pathol.* 58, 1116–1127

95 Munaut F. *et al.* (1998) Genetic relationships among isolates of *Colletotrichum gloeosporioides* from *Stylosanthes* spp. in Africa and Australia using RAPD and ribosomal DNA markers. *Plant Pathol* 47, 641-648

96 Knowles L.L. (2009) Statistical phylogeography. *Annu. Rev. Ecol. Evol. Syst.* 40, 593-612

97 Leslie J. (1993) Fungal vegetative compatibility. *Annu. Rev. Phytopathol.* 31, 127-150

98 Coppin E. *et al.* (1997) Mating types and sexual development in filamentous ascomycetes. *Microbiol. Mol. Biol. Rev.* 61, 411-428

99 Grubisha L.C. and Cotty P.J. (2010) Genetic isolation among sympatric vegetative compatibility groups of the aflatoxin-producing fungus *Aspergillus flavus*. *Mol. Ecol.* 19, 269-280

100 Liu Y.C. and Milgroom M.G. (2007) High diversity of vegetative compatibility type in *Cryphonectria parasitica* in Japan and China. *Mycologia* 99, 279–284

101 Pál K. *et al.* (2007) Sexual and vegetative compatibility genes in the *Aspergilli*. *Stud. Mycol.* 59, 19-30

102 Varzea V.M.P. *et al.* (2002) Distinguishing characteristics and vegetative compatibility of *Colletotrichum kahawae* in comparison with other related species from coffee. *Plant Pathol* 51, 202-207

103 Omondi C. *et al.* (2000) Reaction of some coffea arabica genotypes to strains of *Colletotrichum kahawae*, the cause of Coffee Berry Disease. *J. Phytopathol.* 148, 61-63

104 Beynon S. and Várzea V.M.P (1995) Genetic variation in the Coffee Berry Disease pathogen, *Colletotrichium kahawae*. *Physiol. Mol. Plant Pathol.* 46, 457-470

105 Rodrigues C. *et al.* (1992) Evidence for the existence of physiological races of *Colletotrichum coffeanum* Noack sensu Hindorf. *Kenya Coffee* 57, 1417-1420

106 Sreenivasaprasad S. *et al.* (1993) Coffee Berry Disease pathogen in Africa: genetic structure and relationship to the group species *Colletotrichum gloeosporioides*. *Mycol. Res.* 87, 995-1000

107 Omondi C. (1997) Genetic diversity among isolates of *Colletotrichum kahawae* causing Coffee Berry Disease. *Association Scientifique Internationale du Café (ASIC) 17th Colloque, Montpellier, France*

108 Thomson R.C. *et al.* (2010) Genome-enabled development of DNA markers for ecology, evolution and conservation. *Mol. Ecol.* 19, 2184-2195

109 Rosendahl S. *et al.* (2009) Lack of global population genetic differentiation in the arbuscular mycorrhizal fungus *Glomus mosseae* suggests a recent range expansion which may have coincided with the spread of agriculture. *Mol. Ecol.* 18, 4316-4329

110 Goss E.M. *et al.* (2009) Ancient isolation and independent evolution of the three clonal lineages of the exotic sudden oak death pathogen *Phytophthora ramorum*. *Mol. Ecol.* 18, 1161-1174

111 Zaffarano P. *et al.* (2009) Phylogeographical analyses reveal global migration patterns of the barley scald pathogen *Rhynchosporium secalis*. *Mol. Ecol.* 18, 279-293

112 Zaffarano P. *et al.* (2006) Global hierarchical gene diversity analysis suggests the fertile crescent is not the center of origin of the barley scald pathogen *Rhynchosporium secalis*. *Phytopathology* 96,

941-950

113 Banke S. and McDonald B.A. (2005) Migration patterns among global populations of the pathogenic fungus *Mycosphaerella graminicola*. *Mol. Ecol.* 14, 1881-1896

114 Cisar C. and TeBeest D. (1999) Mating system of the filamentous ascomycete, *Glomerella cingulata*. *Curr. Genet.* 35, 127-133

115 O'Gorman C. *et al.* (2009) Discovery of a sexual cycle in the opportunistic fungal pathogen *Aspergillus fumigatus*. *Nature* 457, 471-474

116 Taylor J. *et al.* (1999) The evolution of asexual fungi: reproduction, speciation and classification. *Annu. Rev. Phytopathol.* 37, 197-246

117 Alby K. *et al.* (2009) Homothallic and heterothallic mating in the opportunistic pathogen *Candida albicans*. *Nature* 460, 890-893

118 Burt A. *et al.* (1996) Molecular markers reveal cryptic sex in the human pathogen *Coccidioides immitis*. *Proc. Natl. Acad. Sci. U.S.A.* 93, 770-773

119 Kronstad J. and Staben C. (1997) Mating type in filamentous fungi. *Annu. Rev. Genet.* 31, 245-276

120 Hsueh Y. and Heitman J. (2008) Orchestration of sexual reproduction and virulence by the fungal mating-type locus. *Curr. Opin. Microbiol.* 11, 517-524

121 Metin B. *et al.* (2010) The mating type locus (MAT) and sexual reproduction of *Cryptococcus heveanensis*: insights into the evolution of sex and sex-determining chromosomal regions in fungi. *PLoS Genet.* 6, e1000961

122 Turgeon B. (1998) Application of mating type gene technology to problems in fungal biology. *Annu. Rev. Phytopathol.* 36, 115-137

123 Casselton L. (2002) Mate recognition in fungi. *Heredity* 88, 142-147

124 Rydholm C. *et al.* (2007) DNA sequence characterization and molecular evolution of MAT1 and MAT2 mating-type loci of the self-compatible ascomycete mold *Neosartorya fischeri*. *Eukaryot. Cell* 6, 868-874

125 Crouch J.A. and Beirn L. (2009) Anthracnose of cereals and grasses. *Fungal Divers.* 39, 191-144

126 Rodríguez-Guerra R. *et al.* (2005) Heterothallic mating observed between Mexican isolates of *Glomerella lindemuthiana*. *Mycologia* 97, 793-803

127 Chen F. *et al.* (2002) Population structure and mating-type genes of *Colletotrichum graminicola* from *Agrostis palustris*. *Can. J. Microbiol.* 48, 427-436

128 García-Serrano M. *et al.* (2008) Analysis of the MAT1-2-1 gene of *Colletotrichum lindemuthianum*. *Mycoscience* 49, 312-317

129 Crouch J.A. *et al*. (2007) Genomic architecture of the mating-type gene cluster in graminicolous species of the genus *Colletotrichum* and across the ascomycota. In *APS Annual Meeting*

130 O'Donnell K. *et al.* (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

131 Kanematsu S. *et al.* (2007) Mating-type loci of heterothallic *Diaporthe* spp.: homologous genes are

present in opposite mating-types. *Curr. Genet.* 52, 11-22

132 Pöggeler S. (1999) Phylogenetic relationships between mating-type sequences from homothallic and heterothallic ascomycetes. *Curr. Genet.* 36, 222-231

133 Rau D. *et al.* (2007) Phylogeny and evolution of mating-type genes from *Pyrenophora teres*, the causal agent of barley net blotch disease. *Curr. Genet.* 51, 377-392

134 Kerényi Z. *et al.* (2004) Mating type sequences in asexually reproducing *Fusarium* species. *Appl. Environ. Microbiol.* 70, 4419-4423

135 Arie T. *et al.* (2000) Mating-type genes from asexual phytopathogenic ascomycetes *Fusarium oxysporum* and *Alternaria alternata*. *Mol. Plant-Microbe Interact.* 13, 1330-1339

136 Voigt K. *et al.* (2005) Phylogenetic relationships between members of the crucifer pathogenic *Leptosphaeria maculans* species complex as shown by mating type (MAT1-2), actin, and beta-tubulin sequences. *Mol. Phylogenet. Evol.* 37, 541-557

137 Stergiopoulos I. *et al.* (2007) Mating-type genes and the genetic structure of a world-wide collection of the tomato pathogen *Cladosporium fulvum*. *Fungal Genet. Biol.* 44, 415-429

138 Anderson P.K. *et al.* (2004) Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends Ecol. Evol.* 19, 535-544

139 Parker I. and Gilbert G. (2004) The evolutionary ecology of novel plant-pathogen interactions. *Annu. Rev. Ecol. Evol. Syst.* 35, 675-700

140 Woolhouse M.E.J. *et al.* (2005) Emerging pathogens: the epidemiology and evolution of species jumps. *Trends Ecol. Evol.* 20, 238-244

141 Desprez-Loustau M. *et al.* (2007) The fungal dimension of biological invasions. *Trends Ecol. Evol.* 22, 472-480

142 Carstens B.C. and Knowles L.L. (2007) Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. *Syst. Biol.* 56, 400-411

143 Degnan J.H. and Rosenberg N.A. (2009) Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* 24, 332-340