Universidade de Lisboa

Faculdade de Ciências

Departamento de Informática

Analysis of RNA-seq data from the interaction of *Coffea* spp. - *Colletotrichum kahawae*

Joana Rita Vieira Fino

Dissertação

Mestrado em Bioinformática e Biologia Computacional

Especialização em Bioinformática

2014

Universidade de Lisboa

Faculdade de Ciências

Departamento de Informática

Analysis of RNA-seq data from the interaction of *Coffea* spp. - *Colletotrichum kahawae*

Joana Rita Vieira Fino

Dissertação

Mestrado em Bioinformática e Biologia Computacional

Especialização em Bioinformática

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

2014

Agradecimentos

Em primeiro lugar, gostaria de agradecer ao Professor Doutor Octávio Paulo e à Doutora Dora Batista por me terem orientado durante este longo percurso. Ao Professor Octávio pelo constante otimismo, entusiasmo, e confiança que me transmitiu mesmo nas alturas mais difíceis. À Doutora Dora pelo apoio e conhecimento que me transmitiu, que me ajudou a tornar uma melhor cientista.

Agradeço também a toda a equipa do CIFC, que apesar do pouco tempo que passei com eles, me fizeram sentir sempre em casa. À Doutora Maria do Céu pela sua simpatia e conhecimento cientifico que ajudou a tornar esta tese muito mais rica do que eu pensei que poderia ser.

À Doutora Andreia Figueiredo agradeço a amizade, o espírito critico e o conhecimento sobre expressão génica tão importante desde o inicio de todo este trabalho.

A todo o CoBiG² que durante os últimos dois anos foram praticamente uma segunda família para mim. A constante boa disposição, amizade e entreajuda torna-vos mesmo especiais. Em especial ao Francisco Pina-Martins por todo o conhecimento informático que me transmitiu e a constante disponibilidade para quando os scripts teimosamente não funcionavam. Prometo continuar a ser uma boa padawan.

À Telma, que passou de apenas uma colega, a uma verdadeira amiga. A forma critica com que olhou para todo o trabalho, o seu dom na construção de figuras explicativas, e a sua constante má disposição sarcástica tornou tudo mais fácil.

Ao João Pedro pelo seu constante apoio e encorajamento até quando tudo parecia impossível. Tudo isto tinha sido mais difícil sem a tua paciência e compreensão. Tu bem dizias que eu conseguia.

Por fim, mas não menos importante, à minha família. Aos meus pais, pela educação e apoio que sempre me deram, independentemente do caminho que escolhi. Vocês deram-me asas para voar. Ao Ricardo e à Rute, por serem simplesmente os meus irmãos mais velhos.

Em suma, obrigada a todos, sem vocês não seria possível ter chegado onde cheguei.

Esta tese é dedicada a ti avô.

Nota prévia

A escrita desta tese de mestrado encontra-se em 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. 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 baseados nos resultados aqui obtidos. Esta foi a razão pela qual esta tese foi escrita em formato de publicação científica. Desta forma, visa-se acelerar o processo de elaboração dos manuscritos e suas subsequentes publicações. Como os resultados aqui obtidos têm de ser complementados com as subsequentes validações biológicas dos dados de expressão genética, o manuscrito encontra-se escrito de acordo com as instruções para autores de uma das revistas de referência da área: *"Molecular Ecology"*. No entanto, para facilitar a leitura, as figuras e tabelas foram incluídas ao longo do texto.

As referências bibliográficas da Introdução Geral foram também elaboradas segundo os parâmetros da revista científica internacional, *"Molecular Ecology"*. Trata-se de uma revista relevante com um sistema de citações cómodo para a leitura de textos de revisão científica. Adicionando o seu elevado fator de impacto na sociedade científica, pareceu apropriada a escolha desta revista como referência para a apresentação da bibliografia.

Resumo

O café é um dos produtos mais comercializados no mundo, com extrema importância económica e social, influenciando milhões de pessoas que dependem direta ou indiretamente desta indústria. No entanto, a cultura do café é extremamente afetada por agentes patogénicos, nomeadamente fungos. *Colletotrichum kahawae* Waller and Bridge é um desses agentes, sendo responsável pela antracnose dos frutos verdes do cafeeiro, conhecida como "Coffee Berry Disease". Esta doença afeta a espécie *Coffea arabica* L., a espécie de maior importância no mercado, apresentando os maiores volumes de produção. Atualmente, a antracnose dos frutos verdes do cafeeiro incide sobretudo em zonas de alta altitude, encontrando-se confinada ao continente africano. Contudo tal não significa que não se possa dispersar para outras zonas de cultivo onde as condições de desenvolvimento, tanto para a planta como para o fungo, sejam favoráveis. Foram desenvolvidas várias estratégias de melhoramento para o combate à doença, levando ao desenvolvimento de algumas variedades resistentes no Quénia. Apesar de já serem atualmente conhecidos vários genótipos com um carácter de resistência a esta doença, as bases genéticas e moleculares da mesma são ainda desconhecidas. Com o intuito de compreender as bases subjacentes ao processo de resistência, recorreu-se à sequenciação comparativa do transcriptoma de dois genótipos de cafeeiro, um susceptível (Caturra) e outro resistente (Catimor 88) durante as primeiras horas de interação de *C. kahawae,* através da plataforma Illumina. A análise destes dados visou a identificação de genes diferencialmente expressos, envolvidos na resistência da planta à doença. Os dados desta sequenciação foram previamente analisados pela empresa ARK genomics (UK), embora utilizando softwares e parâmetros padronizados, normalmente aplicados para todo o tipo de análises deste género, desde bactérias a plantas. Com o objetivo de melhorar e aprofundar a análise, foi desenvolvida uma nova análise customizada, que aqui se apresenta, em comparação com a análise anterior. Várias ferramentas e abordagens foram aplicadas nesta nova análise, tendo em conta a inexistência de um geno ma de referência. Neste trabalho foi possível identificar vários problemas e cuidados a ter desde o tratamento das "reads", até ao cálculo de diferenças de expressão, bem como simples diferenças entre softwares. Neste novo estudo de expressão teve-se ainda em conta análises comparativas a diferentes níveis que não tinham sido efetuadas na análise anterior. A anotação de "unigenes" diferencialmente expressos indica uma tendência para categorias funcionais

diretamente relacionadas com a produção de energia, envolvida no crescimento e desenvolvimento da planta, e com processos já identificados como envolvidos na resposta de defesa a agentes patogénicos tais como o metabolismo de açucares ou a biosíntese de fenilalanina e fenilpropanoides.

De um modo geral, os objetivos deste trabalho foram cumpridos, tendo-se desenvolvido uma linha de análise que permitiu uma melhor e mais adequada exploração dos dados gerados por sequenciação de transcriptoma. Espera-se assim que os resultados obtidos venha a contribuir para o aumento do conhecimento científico sobre a resposta de defesa por parte da planta, gerando informações úteis para o estabelecimento de programas de melhoramento que apoiem a produção sustentável de uma cultura tão relevante a nível económico e social.

Por outro lado, espera-se que este trabalho mostre a necessidade de uma análise cuidada de dados de "next generation sequencing", em especial dados resultantes da sequenciação de RNA, tecnologia ainda bastante recente e sem um processo universalmente aceite para a análise correta dos dados gerados.

Palavras-Chave: *Cafeeiro*.; Antracnose dos frutos verdes ; Mecanismos de defesa; *Assemblagem* do Transcriptoma; Expressão diferencial; Análise comparativa

Abstract

Coffee is one of the most traded products in the world, with extremely social and economic importance, and millions of people who depend directly or indirectly on it. Coffee berry disease (CBD), caused by the fungus *Colletotrichum kahawae* Waller & Bridge, is considered the biggest threat to Arabica coffee production in Africa at high altitude. In *Coffea arabica L.* plantations, CBD can cause up to 20-50% of crop losses, reaching 80% in years of severe epidemics if chemical control is not applied. In order to control this disease, several coffee improvement strategies were developed which leaded to the selection of few hybrid commercial resistant varieties in Kenya. Therefore, breeding for coffee resistance remains a powerful strategy to fight CBD, in an economic and sustainable manner. With the purpose of gaining some insights on coffee resistance process, a RNA Illumina sequencing approach was used to characterize the transcriptional profile of two coffee genotypes, respectively susceptible (Caturra) and resistant (Catimor 88) to *C. kahawae*, during the early stages of the infection process. The differential expression analysis of this data aimed to identify genes putatively involved in the resistance process. Although a previous analysis was made by the sequencing company ARK genomics (UK), this was only based on non-specific methods generally applied to a wide range of organisms. To improve the analysis and consequently the results obtained, a new approach was taken aiming to produce a more customized workflow. Comparatively with the previous analysis, the present approach showed some improvement regarding the transcriptome assembly quality and size, or the level of confidence of the differential expression results, despite the CPU and RAM limitations. It was possible to account for additional comparative analyses for the differential expression assessment and to identify the enriched functional categories representing the differential expressed unigenes. Regarding the biological results, the resistant genotype showed a high effective response to the infection while the susceptible genotype showed an early stress-leaded response by the infection. The KOG and KEGG annotation of the differential expressed unigenes, was able to identify two main domains: plant development and defense response. It is expected that the results obtained here will contribute to increase the scientific knowledge on the plant defense response , generating useful information able to guide the establishment of breeding programs that support sustainable production.

Moreover, it is expected that this study show the necessity of careful analysis of next generation sequencing data, especially when dealing with recent methods like RNA-seq, for which there is no clear consensus about the best analysis practices.

Keywords: Coffee plant; Anthracnose; Plant Defense mechanisms; Transcriptome assembly; Differential expression; Comparative analyses

Table of Contents

Chapter I

I. General introduction

Coffee is one of the most valuable agricultural products in the world, and one of the greatest economic income generators for several developing countries where a considerable percentage of the population depends on coffee-related activities such as production, processing, transport and commercialization. The worlds' consumption of coffee is constantly growing, which makes the coffee industry prosperous. Nevertheless, recurrent rock bottom prices cause immense hardship both to countries where coffee is a key economic activity, and to the farmers involved in coffee production. The origin of this situation lies on the oscillation of prices due to the current imbalance between supply and demand. Meanwhile, the costs of production, transport, machinery and disease control continue to grow. The subsequent effects force the coffee farmers to economize and this has often led to a reduction in the use of agricultural inputs necessary for optimal coffee production. On the other hand, the occurrence of major severe diseases is one of the main limiting factors of coffee production. Coffee berry disease (CBD) caused by the fungus *Colletotrichum kahawae* Waller and Bridge, is the most devastating threat to *Coffea arabica* L. production in Africa at high altitude, and its dispersal to Latin America and Asia represents a serious concern. This pathogen is a highly destructive specialist that infects expanding green berries, leading to their premature dropping and mummification. Despite the existence of effective methods for CBD control such as chemical control, their prices and the application procedures can be too high and complex especially for small producers. Thus, the utilization of methods such as the cultivation of disease resistant varieties seems to be the most reliable way to manage disease control. In order to accomplish long-lasting resistance using breeding strategies, a better knowledge of the molecular bases of coffee resistance is essential, so that a sustainable system of coffee production can be created.

Deep transcriptome sequencing studies are becoming more and more common, presenting inumerous advantages towards the unprecedent amount of knowledge that can generate, but the bioinformatics analysis of the data is still a major limitation. RNA-Seq analysis is mostly used for expression studies, and is suitable for the understanding of transcriptomic dynamics between conditions.

In the present work, a RNA-seq comparative analysis was made between a susceptible and a resistant genotype of coffee when infected with *C. kahawae*, with the aim of identifying genes potentially involved in the resistance response. Before presenting this work, a brief introduction is made on the host, *Coffea* spp., the pathogen *C. kahawae* and the plant-pathogen interaction in order to highlight the most relevant aspects of the pathosystem studied. It is also presented a little introduction to the NGS technology, methods of analysis and software used.

1. The Host – Coffee Plants

1.1. General Characteristics

Coffee plants belong to the genus *Coffea* from the Rubiaceae family. This classification encounters 103 described species, with the most economically relevant species belonging to the subgenus *Coffea,* including the three species that are commercially explored: *Coffea arabica* L. (Arabica coffee), *Coffea canephora* Pierre ex A.Froehner (Robusta coffee) and *Coffea liberica* Hiern with a marginal expression in total coffee production, grown only at a regional scale (Bridson 1994; Davis 2003; Davis *et al.* 2006).

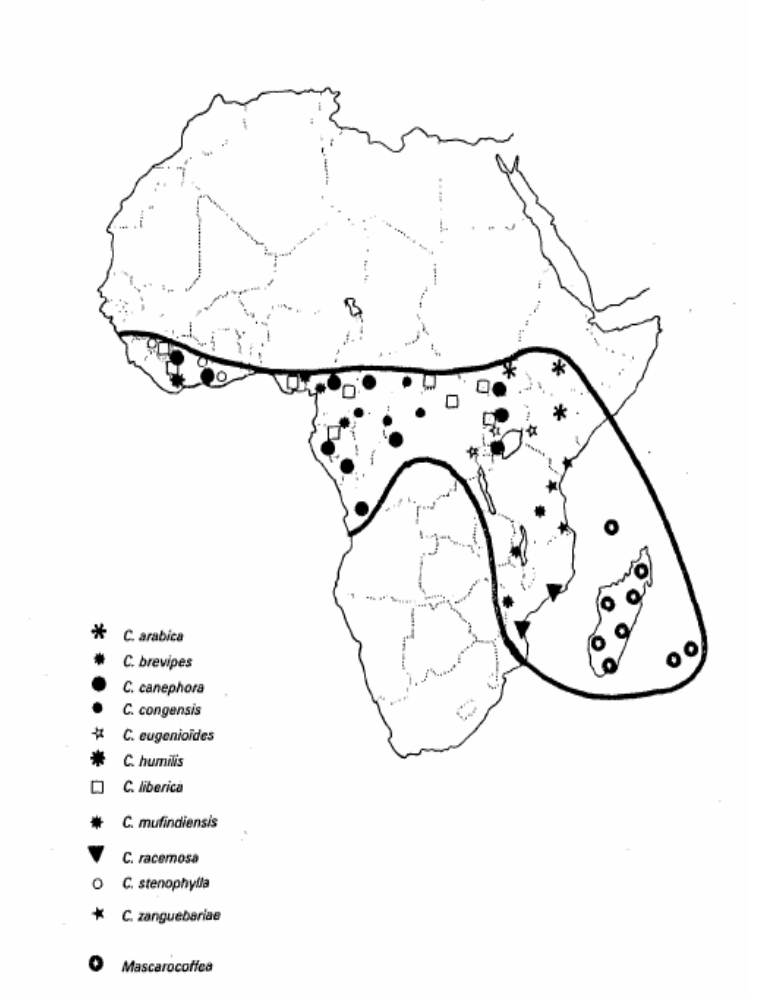


Figure 1 - The distribution of native species of *Coffea spp.* adapted from Charrier & Berthaud 1985

Coffea spp. are evergreen, glossy-leaved shrubs or trees 5–10 m high from tropical and sub-tropical forest habitats. Native of the African continent, *Coffea spp*. occur mostly in humid, evergreen forests, but their habitat also includes other forest types (Waller *et al.* 2007) The three commercially relevant species are better adapted to different kinds of forests: *C. arabica* need cool and humid environmental conditions at high altitudes, while *C.canephora* and *C. liberica* are usually found in humid and relatively warmer environments, typical of the lowlands (Wrigley 1988; Lashermes & Anthony 2007). The natural distribution of *Coffea spp.* is represented in Figure 1. Coffee plants features include elliptical leaves with pointed tips, which occur 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. Flower clusters are produced in leaf axils. The fruit is a two-seeded drupe with a fleshy epicarp. The stems exhibit dimorphic branching due to the different development of two buds that occur, one above the other in each leaf axil of the main stem (Waller *et al.* 2007).

As common in the family Rubiaceae, most of the species of the genus *Coffea* are diploid with 2n=22 chromosomes, except *C. arabica* which is allotetraploid (2n=44 chromosomes), resulting from a natural hybridization between *C. eugenioides* and *C. canephora* genomes (Lashermes *et al.* 1999). *C. arabica* is further considered a relatively new species, due to the lack of differentiation from its parental species (Raina *et al.* 1998; Lashermes & Anthony 2007).*C. arabica* also differs from the other species due to being self-fertile, which is a trait that is not present in other species (self-incompatible) (Charrier & Berthaud 1985).

C. arabica, is one of the most important species in coffee industry, since the best quality coffee, with low caffeine content is produced from its fruits, however is highly susceptible to various diseases.

1.2. History

The history of the coffee plant is not accurate, since it dates back to ancient times, and covers so many episodes that the version presented here is most likely a mix of facts and fiction that cannot be easily dissociated from each other.

According to Ferrão, 1993, coffee has its origin in the mountainous area of Abyssinia (actual Ethiopia) from where it spread to South-East Arabia possibly carried by pilgrims to Mecca, which used coffee berries for its stimulating effect. These pilgrims later introduced the plant in India around the 16th or 17th century (Bigger 2006; Ukers 1935), but their cultivation was known to be first started as early as 575 AD in Yemen (Anthony *et al.* 2002; Topik 2004; Bigger 2006; Lécolier *et al.* 2009)

In the 16th century, the Europeans become aware of coffee cultivation and use as beverage, which led to their dissemination around the colonies (Anthony *et al.* 2002; Topik 2004; Bigger 2006), turning coffee into one of the major sources of income, as it remained until today. The Dutch were the first to recognize the potential of coffee, and manage to ship a coffee plant from Yemen to Java (Ferrão 1993; Topik 2004; Bigger 2006; Ukers 1935). In

1706 the first coffee plants were received at the Amsterdam botanical gardens, from Java, and soon they were being shipped to other gardens all around Europe (Ferrão 1993; Topik 2004; Bigger 2006; Ukers 1935). The French soon started the dispersion along the West Indies, between 1715 and 1730, introducing coffee into places like the Dominican Republic, Haiti, Martinique, Jamaica and Reunion island (Bigger 2006; Ukers 1935). The dispersion continued to Central America, including Costa Rica, Cuba, Mexico and Venezuela, due to Spanish intervention (Topik 2004; Bigger 2006; Ukers 1935). The Portuguese seems to be responsible for the introduction of coffee in Brazil, and later on, in other colonies, such as the African colonies of São Tomé, Mozambique and Cape Verde, on the 17-18th century (Ferrão 1993). Just like Portugal, other European Countries introduced coffee on their African colonies: in the 19th century, the Dutch, established plantations on Gana and the French in the Ivory Cost (Bigger 2006).

The dissemination and domestication of coffee was thus mainly conducted from the 16th to the 19th century and was subjected to an intensive selection of phenotypes, optimized for better economic performance (Stukenbrock & McDonald 2008). This new and rapidly created agro-ecosytem provided genetically uniform populations, ideal as a host for the emergence and dispersal of plant pathogens (Anthony *et al.* 2002). This apparent lack of genetic variation in *C. arabica* crops makes them highly vulnerable to disease outbreaks since 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(Friesen *et al.* 2006; Butler *et al.* 2009).

1.3. Production and Commercialization

Nowadays, coffee is one of the world's most valuable export commodities, ranking second on the world market after petroleum products and a primary export of many developing countries that rely, to a greater or lesser extent, on the revenues generated. This means that any decline on coffee trading earnings can have major economic repercussions in those countries (Davis 2003).

According to the International Coffee Organization, coffee is the world's most widely traded tropical agricultural commodity, accounting for exports estimated in US$ 15.4 billion for 2009/10. Coffee also plays an important role at the social level of the producing countries, due to the high number of jobs provided by this industry. For example, in 2010 the total

coffee sector employment was estimated at about 26 million people in 52 producing countries (van Hilten *et al.* 2011).

Coffee production relies mainly on two species: *Coffea arabica* (70%) and *Coffea canephora* (30%) (Davis 2003; Ukers 1935)**.** This distribution of production is related to the superior cup quality of *C. arabica. C. arabica* is predominantly produced in Central and South America and *C. canephora* in West Africa and Asia (http://www.ico.org, accessed on October 16th 2013).

Brazil encounters itself on the top of the list of the world's biggest producer of coffee both in Arabica and Robusta coffee, followed by Colombia for *C. arabica*, and Vietnam for *C. canephora* (http://www.ico.org, accessed on October 16th 2013).

The coffee industry is prosperous and stable due to the exports of most of the production to European countries (Vega *et al.* 2003; Waller & Masaba 2006) (for example, in 2010-11, Brazil consumed 19130000 bags against 29603000 exported) (http://www.ico.org, accessed on October 16th 2013). Despite that, the coffee crisis is a fact: the oscillation of prices due to the current imbalance between supply and demand has severe consequences at several levels. On top of that, coffee diseases can potentially aggravate this crisis, especially major ones, such as coffee leaf rust and coffee berry disease (Osorio 2002; Vega *et al.* 2003).

2. *Colletotrichum kahawae*, the agent of coffee Berry Disease

Coffee Berry disease (CBD) is an extremely severe disease of *C. arabica* caused by the fungus *Colletotrichum kahawae* Waller & Bridge resulting in anthracnose of the green fruits. It is the largest threat to *Coffea arabica* production in Africa, to where it is presently still confined.

The most recent speciation hypothesis showed that *Colletotrichum kahawae* emerged from the *C. gloeosporioides* complex as a specialist on Arabica coffee (Silva *et al.* 2012), producing anthracnose symptoms on the green berries, expressed by dark sunken leasons leading to their premature dropping or mummification – Coffee Berry Disease.

2.1. Origin and Distribution

Only in 1993 the CBD agent was well characterized as a distinct species, based on morphological, cultural and biochemical characters, as *Colletotrichum kahawae* Waller &

Bridge belonging to the Family Glomerellaceae (Waller *et al.* 1993).

A few years ago, the *C. gloeosporioides* species complex was reclassified, and the CBD agent was then classified as a subspecies of *C. kahawae*, *C. kahawae subsp. kahawae* (Weir *et al.* 2012).

The specific origin of this pathogen and the disease emergence is still a subject of debate. Nonetheless the first known report goes back to 1922 in two small districts of Kenya, located at high altitudes, where most of the crops were ruined by a new unknown disease (McDonald 1926; Vermeulen 1970). After that, the disease was described in Angola (1930), Congo RD (1938), Mount Kenya district (1939), and afterwards, it rapidly spread to almost all the Arabica coffee cultivation areas of Africa (Nutman & Roberts 1960; Manga 1997). More recently Silva et al. (2012) in their study hypothesized that *C. kahawae* emergence may have taken place in Angola, as opposed to Kenya.

Currently the disease is confined to the African Continent and is rarely reported below 1600 meters (Manuel et al. 2010). This preference is due to the cooler and wetter conditions of high altitudes that favor both pathogen and disease development (Vermeulen 1970; Mulinge 1971; Waller & Masaba 2006). However, the spread of the disease is a big concern for non African coffee production countries bearing similar environmental conditions, due to the terrible consequences that it could bring for production.

2.2. Infection process and disease symptoms:

In the infection process, *C. kahawae* uses a hemibiotrophic strategy, which includes a post-penetrative asymptomatic biotrophy phase, followed by a destructive necrotrophy phase that culminates in the appearance of disease symptoms and the reproduction of the fungus (Loureiro *et al.* 2012). *C. kahawae´s* infection starts with the germination of the conidia (asexual spore) and differentiation of melanized apressoria on the plant´s surface, a structure used by the fungus to penetrate the cuticule (Fig 2) by mechanical pressure, secretion of cutin degrading enzymes, or a combination of both processes (Chen *et al.* 2004; Silva *et al.* 2006). Following the penetration, the fungus starts to colonize the host tissues: an infection vesicle is formed from which several other hyphae emerge and grow. This phase involves the transition of hyphae growth in living cells (biotrophy - which may last 24 to 48 hours after inoculation) to dead cells (necrotrophy). Finally, a new conidia is formed and emerges from the cuticule, setting free a new generation of *C. kahawae* spores (Figure 2) (Silva *et al.* 2006; Loureiro *et*

al. 2012).

Depending on the resistant or susceptible response of the coffee genotype, two types of symptomatology can occur. **Scab lesions** are typical of a resistance plant response, which restricting fungal development, only allows the formation of superficial little black spots . In this case, the deeper layers of the fruit are not invaded, and the lesion appears stationary, not affecting the normal development of the green berry (Anthony *et al.* 2002; Topik 2004). On the other hand, in susceptible plants the development of **Active lesions** is observed, starting as little black spots, which in the presence of good conditions can form dark, sunken, active lesions that rapidly expand and destroy the entire fruit (Nutman 1970; Ntahimpera *et al.* 1999; Schroth *et al.* 2000).

2.3. Dissemination:

To a properly *C. kahawae* spore dispersion, environmental conditions such as temperature, precipitation and humidity, are crucial because these conditions directly interfere with the infection process, affecting also the distribution and severity of the disease (Mulinge 1971). Temperatures within the range of 17 – 28ºC are favorable for the development of the infection, while temperatures outside this range slow down this process. In addition, the maturation stage of the host plant's fruit is also a parameter known to have influence in the infection process (Nutman & Roberts 1960).

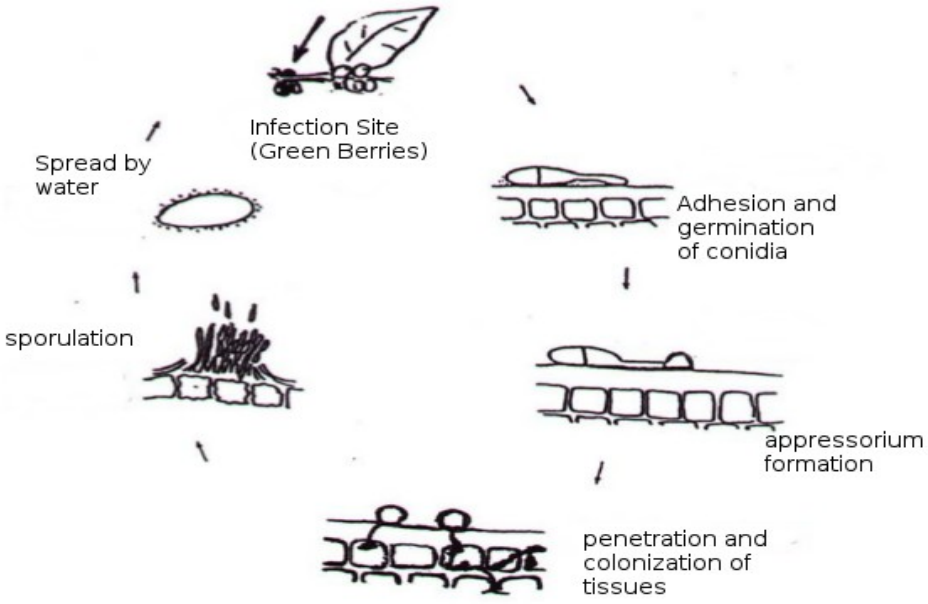


Figure 2 - Schematic representation of the infection process of *Colletotrichum kahawae* adapted from Jeffries & Koomen 1992.

Sporulation occurs mainly in already infected berries with high humidity conditions (Gibbs 1969) and the spread, responsible for new infections, is exclusively dependent on rain, since the mucilage surrounding the spores also prevents dispersal by wind (Fitt *et al.* 1989). Thereby, the spread occurs in a vertical way in the same plant, where the highest twigs infect the lowest twigs, and, in different plants by rain splash, , but only in short distances (never more than 1 m) (Ntahimpera *et al.* 1999). Taking into account the high distribution of *C. kahawae*, there is, however, a very good possibility that spread by rain is not the only mechanism for the *C. kahawae* dispersal (Waller 1972). Human activity and animals could play also a responsible part in fungus dispersal (Nutman & Roberts 1960; Schroth *et al.* 2000).

2.4. Economical impact

CBD is considered one of the main problems to coffee production, with great repercussions in economics.

In *C. arabica* plantations, CBD can cause up to 20-50% of crop losses, reaching 80% in years of severe epidemics if chemical control is not applied (Van der Vossen *et al.* 1976; Griffiths et al 1971). This can signify the loss of millions of dollars, especially in countries where coffee production is almost exclusively restricted to *C. arabica* such as the case of Ethiopia (Derso & Waller 2003). The scenario can be even more concerning, if we take in consideration that in Ethiopia, more than 700,000 families are involved in coffee production and more than 15 million people depend directly or indirectly of coffee (Vega *et al.* 2010).

The use of chemical control measures can decrease the losses by CBD, but that measures may account for 30-40% of total production costs. Annual economic damage to *C. arabica* production in Africa, due to crop loss by CBD and cost of chemical control, is estimated at US$ 300–500 million (Van Der Vossen 2009).

2.5. Control

The first attempts to control the disease consisted on the application of copper and or systemic organic fungicides (Nutman 1970). The application of this treatment, however, not only caused toxicity problems in the plants and soils, but in other cases was not effective as the fungicide itself could be mostly washed away (Nutman 1970; Chung *et al.* 2006; van den Bosch & Gilligan 2008).

Currently the main focus for disease control is the development and cultivation of resistant coffee varieties to CBD, through plant breeding programs (Silva *et al.* 2006). A major breakthrough for the improvement of coffee breeding programs was the discovery in the late 1950s, of Hibrido de Timor (HDT), a coffee hybrid of *C. arabica* and *C. canephora*. HDT was discovered in an Arabica coffee plantation in Timor. Initially was known for being resistant to coffee leaf rust, but later was recognized as having some degrees of CBD resistance too. In that way, some lines of HDT and Rume Sudan, has been used in breeding programs as resistance sources (Wrigley 1988; Varzea 1993; Silva *et al.* 2006). One such examples, is the commercial variety Ruiru 11 and Catimor 88 in Kenya, which were bred for resistance to CBD and coffee leaf rust.

3. Coffee – *C. kahawae* interaction

Plants and pathogens have evolved together in a dynamic system of interaction. While plants have the ability to recognize potential invading pathogens, and have developed several defense mechanisms; pathogens, at the same time, have developed new infection strategies, compromising the defense mechanisms of the host, effectively playing an evolutionary "ping-pong" game.

There are essentially three reasons for a pathogen not to be able to infect a host, leading to an incompatible interaction:

i) The plant is unable to support the niche requirements of the potential pathogen, constituting a non host (Hammond-Kosack & Jones 1996);

ii) The plant possesses means to confine successful infections, which are constitutively expressed, like structural characteristics that prevent the entrance of micro-organisms or the presence of some antimicrobial compounds, forming physical and chemical barriers (Hammond-Kosack & Jones 1996);

iii) Upon recognition of the attack, the plant initiates mechanisms that can keep the invasion localized, such as structural alterations of the cell wall, production and accumulation of antimicrobial compounds, deposition of compounds between the plasma membrane and the cell wall or even cell death at the pathogen's site of penetration, which involves a network of signal transduction and rapid activation of gene expression (Hammond-Kosack & Jones 1996). In this process, a non-specific first line of plant defense is activated by the recognition of common pathogen elicitors (pathogen-associated molecular patterns, PAMPs), which

trigger the subsequent responses, such as the production of pathogenesis-related (PR) proteins (Hammond-Kosack & Jones 1996; Gururani *et al.* 2012)

However, pathogens can "bend" these rules, by suppressing host defenses and subsequently colonizing host tissues, which corresponds to a compatible interaction (susceptibility).

3.1. Coffee resistance mechanisms to *C. kahawae*

In Arabica coffees resistance mechanisms to *C. kahawae* are both preformed and induced, and operate at different stages of pathogenesis (Gichuru 1997). The coffee berry cuticle could act as a physical barrier to the penetrating pathogen. Moreover, several investigations on the occurrence and possible role in CBD resistance of preformed antifungal compounds in the cuticle have been carried out, although the chemical nature of these compounds was not identified (Silva *et al.* 2006 and references therein).

Resistant coffee genotypes can rapidly initiate a specific defense response to the infection of *C. kahawae*, leaving only a scab lesion on the infection site (Anthony *et al.* 2002; Topik 2004). According to Masaba & van der Vossen (1992), this type of lesion is related with the formation of a suberin barrier under the local of infection – a mechanic barrier to the development of the fungus - and the capacity to form layers of suberised cells under the local of infection. Apparently these mechanisms are dependent on metabolic activity, because when the fruit is detached from the plant, this capacity of response is completely lost.

Cytological analysis showed that for certain coffee genotypes resistance to *C. kahawae* is characterized by the restriction of fungal growth associated with the hypersensitive host cell death (hypersensitive response), accumulation of phenolic compounds, encasement of intracellular hyphae with callose and modifications in cell walls (lignification and thickening) (Silva *et al.* 2006; Loureiro *et al.* 2012).

In susceptible plants, some of these responses, such as deposition of callose and phenols are delayed, not being able to prevent the fungus development and reproduction (Silva *et al.* 2006; Loureiro *et al.* 2012).

Although coffee responses have been well described in a citological context, the genetic molecular bases of coffee resistance against *C. kahawae* remain unknown. Previous studies have identified some genes as being involved in resistance mechanism, however, their annotation and characterization has not been possible (Silva *et al.* 2006; van der Vossen & Walyaro 2009). Thus, it is extremely important to increase the knowledge on the structure of

the transcriptome, through the comparison of infected resistant and susceptible coffee genotypes, to get some insights on the distinctive processes underlying plant resistance response.

4. RNA-sequencing and data analysis

Studying the transcriptome profile is essential for fully understand the biological pathways that are active in various physiological conditions or developmental stages (Wang *et al.* 2009; Ozsolak & Milos 2010). Knowledge about functional elements of the genome and molecular specificities of cells and tissues can be retrieved from this type of analysis (Wang *et al.* 2009; Martin & Wang 2011). For a long time, the utilization of the Sanger technology led to a limited knowledge of the transcriptome, since this technology can only allow sequencing of limited sets of samples with a high time and resource consumption (Martin & Wang 2011). Recently, the development of novel deep-sequencing technologies (Next generation sequencing, NGS) opened exciting new approaches to transcriptome profiling (Bohnert *et al.* 2009).

4.1. NGS technologies

Currently, there are three NGS technologies in major use: Roche/454 (entering into disuse, but still viable for a number of goals as proved by many recently published studies, such as the study of Oak root response to ectomycorrhizal symbiosis establishment (Sebastiana *et al.* 2014)), Ion torrent, and Illumina (Mardis 2011; Loman *et al.* 2012).Table 1 resumes some technical specifications of these platforms of next generation sequencing methods.

Roche/454 was the first to achieve commercial success, and uses an alternative sequencing technology known as pyrosequencing. Although this technology offers long reads (~ 600bp), which facilitates the assembly step in comparison with other technologies, it cannot interpret long stretches of the same nucleotides (homopolymers), introducing errors on base calling, resulting in a low throughput (Mardis 2008, 2011; Metzker 2010; Liu *et al.* 2012; Loman *et al.* 2012).

Like 454, Ion Torrent technology exploits emulsion PCR. This platform is based on the detection of hydrogen ions that are released during the polymerization of DNA (Rothberg *et al.* 2011; Loman *et al.* 2012). Also, Ion Torrent technology suffers from errors in homopolymer regions , although to a lesser extent, and produces shorter reads than 454

technology (up to 400bp). This technology presents lower accuracy but a lower price per Gigabase, comparatively with IIllumina sequencing (http://allseq.com, accessed on January 11th 2014; Loman *et al.* 2012). Despite their reasonable throughput, the main advantage of ion torrent, relatively to the other sequencing technologies, is the price of the equipment which is much more cheaper than the others (Quail *et al.* 2012).

Table 1 - Technical specifications of Next Generation Sequencing platforms. (http://allseq.com, accessed on January 11th 2014;Gilles *et al.* 2011; Liu *et al.* 2012; Quail *et al.* 2012; Loman *et al.* 2012)

	Roche/454 (Titanium)	**Ion torrent** (Proton 318 chip v2)	**Illumina** (Illumina HiSeq 2000)
Equipment price	$500k	$50k	$654k
Sequencing yield per run	700Mb	Up to 2Gb	600Gb
Sequencing cost per Gb	~$10k	~$16	~$41
Observed raw error rate	1.07%	>1%	0.26%
Read length	~600bp	Up to 400bp	~150bp
Paired-end reads	no	no	yes

Lastly, the Illumina system utilizes a sequencing-by-synthesis approach in which all four nucleotides are added simultaneously to the flow cell channels, along with DNA polymerase, for incorporation into the oligo-primed cluster fragments. Illumina, produces the shortest reads (~150bp, but it is already commercialized equipment that can produce reads up to 300bp, and so, fragments of 600bp), but yields the best throughput/cost relation. Plus, it presents the highest accuracy among the mentioned technologies and is suitable for a large range of applications, such as mRNA sequencing (RNA-Seq) and whole genome sequencing (Mardis 2008, 2011; Metzker 2010).

4.2. RNA-Seq

RNA-Seq is a recent method for both mapping transcriptomes and quantifying transcripts, measuring gene expression, based on the latest developed deep-sequencing technologies.
In general, RNA is converted to a library of cDNA fragments with adapters in both ends. Each molecule, with or without amplification, is then sequenced in a high-throughput manner to obtain short sequences from one end (single-end sequencing) or both ends (pair-end sequencing) (Wang *et al.* 2009). In principle, deep-sequencing technology can be used for RNA-Seq, such as Illumina or Roche 454 systems, which are commonly applied for this

purpose (Wang *et al.* 2009).

Although RNA-Seq is still a technology under active development, it offers several key advantages over existing technologies. Comparatively with Microarrays (Table 2), RNA-seq is not limited to identifying transcripts corresponding to existing genomic sequences. For example, Illumina based RNA-seq can be used when no reference genome is available as reported for the gene expression analysis of Paulownia infected by *Phytoplasma* (Paulownia witches'-broom) (Mou *et al.* 2013). This feature makes it very attractive for non-model organisms with or without reference genome (Wang *et al.* 2009). Furthermore, RNA-seq is particularly useful for transcriptome assembly and hence to provide information on how exons are connected, and can be used for base variation calling in the transcribed regions. Other advantages of RNA-seq relative to DNA microarrays include: the absence of background noise caused by unambiguity when mapped against a reference genome, bigger sensitivity for low and extremely high expression regions, and a higher accuracy (Nagalakshmi *et al.* 2010; Xu *et al.* 2013). Consequently, the volume of expressed genes detected are much higher just as the sensitivity of the different degrees of expression. At last, RNA-seq has shown high levels of accuracy, when confirmed through quantitative real-time PCR.

Table 2 - Differences between Microarrays and RNA-Seq (adapted from Wang *et al.*, 2009)

Technology	**Microarray**	**RNA-Seq**
Specifications		
Resolution	Up to 100bp	Single base
Throughput	High	High
Reliance on genomic sequence	Yes	In some cases
Background noise	High	Low
Application		
Simultaneously map transcribed regions and gene expression	Yes	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes
Ability to distinguish allelic expression	Limited	Yes
Practical issues		
Amount of RNA needed	High	Low
Cost of mapping transcriptome	High	Relatively low

RNA-Seq is the first method that allows the survey of the entire transcriptome in a very high-throughput and quantitative manner, with countless advantages over other methods. This method offers both single-base resolution for annotation and 'digital' gene expression levels at the genome scale, often at a much lower cost than microarrays.

4.3. Data Analysis

Like other deep-sequencing technologies, RNA-seq implies several bioinformatic challenges including methods and infrastructures to store and process large amounts of data in a fast, error-free and "less memory consuming" way (Wang *et al.* 2009; Oshlack *et al.* 2010).

The first step of RNA-seq analysis, after "cleaning" the reads, is to map the reads against a reference genome, or assemble the reads all together (*de novo* assembly), to unravel the structure of the transcriptome. When a reference genome exists, the assembly process is relatively simple: The reads are mapped against the genome (normaly called "backbone") originating the transcriptome (Wang *et al.* 2009; Oshlack *et al.* 2010). There is a wide choice of software available for this task, which does not require great computing power or time, such as Cufflinks (Garber *et al.* 2011). These software make the best use of the reference genome, reporting isoforms and identifying novel transcripts. On the other hand, when a reference genome is missing, the task becomes much more complicated: the software for transcriptome assembly without a "backbone" (usually called *de novo* assembly, as Velvet/Oases (Schulz *et al.* 2012) and TransAbyss (Garber *et al.* 2011)) is time and resource consuming, and the final result usually entails a high level of redundancy. This redundancy can be the result of assembly bias, already identified in several *de novo* assembly programs, or simply the result of a mixed assembly of different isoforms (due to alternative splicing), which without a reference, cannot be distinguished (Wang *et al.* 2009; Oshlack *et al.* 2010; Martin & Wang 2011). For reducing this redundancy bias, some software already exists, such as CD-HIT (Li & Godzik 2006; Surget-Groba & Montoya-Burgos 2010; Miller *et al.* 2012), but sometimes it is not sufficient to remove all the redundancy generated. In a study of *Pinus sp.* Transcriptome (Parchman *et al.* 2010), blastx hits were used for redundancy evaluation, which unraveled redundant genes when their hits were the same. This strategy can be imperfect, if several genes do not match with any of the databases sequences, hampering their redundancy analysis (Parchman *et al.* 2010).

The next step includes read mapping against the reference genome or *de novo* assembled

transcriptome, for expression quantification. This task can be very difficult, especially in large transcriptomes with short reads (like Illumina sequencing) because reads can match several locations in the transcriptome/genome (Oshlack *et al.* 2010). Several solutions have been proposed, including the assignment of the multi-matched reads based on the number of reads mapped, to their neighboring unique sequences (for low-copy repetitive sequences) (Mortazavi *et al.* 2008), or the assignment of multi-matched reads based on the probability of a fragment being derived from a certain transcript, computed by maximum likelihood (Li & Dewey 2011; Garber *et al.* 2011). This last method is used in the software RSEM, which uses Bowtie (Langmead *et al.* 2009) for read mapping, and relies on this same method to quantify the expression of different isoforms without a reference genome (Li & Dewey 2011). On the other hand, the use of longer reads, such as those obtained with 454 technology, and paired-end sequencing can help on this multi-matching problem. Also, the advance of the sequencing technologies may proportionate a bigger read length in the near future (Wang *et al.* 2009).

Errors in sequencing or polymorphisms can present other types of mapping problems, besides ambiguous locations on the genome/transcriptome. Small differences can be overcome by the software, which can accommodate one or two base differences. However, resolving large differences is much harder, and will usually require great genome annotation for polymorphisms and deeper coverage (Wang *et al.* 2009).

For a proper RNA-seq expression quantification, considerable sequencing depth is needed. Insufficient depth would result in lower coverage, which lead to a less accurate quantification, in a method that depends directly on read quantity for accurate results (Wang *et al.* 2009). In general, the larger the genome, the more complex the transcriptome and consequently, more sequencing depth is needed for a decent coverage (Wang *et al.* 2009). For simple transcriptomes such as yeast, with no evidence of alternative splicing, 30 million of 35 nucleotide reads are sufficient to observe transcription for most genes on a single condition (Wilhelm *et al.* 2008).

Nevertheless, there is no way to better compute the coverage needed for transcriptome sequencing, as the true number and level of different transcript isoforms is not usually known and transcription activity varies greatly across the genome. However, analyzing different conditions can further increase coverage (Wang *et al.* 2009).

Lastly, it is possible to use gene quantification across conditions to obtain their differences

and gain insights about gene regulation, allowing differential gene expression analysis (Garber *et al.* 2011). RNA-Seq is capable of capturing transcriptome dynamics across different conditions, times and tissues offering a robust and accurate way to compute differentially expressed genes. For calculating the fold change of genes between conditions, several packages are available, with different features adapted to different data. For instance, for diferentially expressed genes analysis of *Citrus reticulata* infected vs not infected by *Xylella fastidiosa,* Cufflinks-Cuffdiff was used for mapping, quantifying and comparing expression levels, based on a reference genome (Rodrigues *et al.* 2013); on the other hand, the RNA-seq analysis of catfish (susceptible and resistant) when infected with *Flavobacterium columnare*, in different time points was made using CLC Genomics Work-bench, with a reference trancriptome (Peatman *et al.* 2013). There are some other packages for expression analysis, mostly R packages, such as EdgeR or Ebseq (Garber *et al.* 2011; Leng *et al.* 2013). As it happens with the remaining software for the bioinformatics analysis, there is no perfect software for any type of data. Depending on the software, differential expressed genes calling (DE calling) can be more restrictive or liberal, be indifferent or work better with higher number of replicates or even perform better or worse with the heterogeneity of the samples (Soneson & Delorenzi 2013; Seyednasrollah *et al.* 2013). It is up to the technician to choose the most adequate software for his analysis.

Although RNA-seq is still a recent technology, its advantages over other transcriptomic methods are quite clear. It can be valuable for understanding transcriptomic dynamics across different conditions, where it allows a robust comparison between them. The biggest challenge of this recent technology is to be able to target more complex transcriptomes in order to identify and track the expression changes of rare RNA isoforms from all genes, even without a reference genome (Wang *et al.* 2009).

Objectives

The research presented in this Thesis is integrated in project PTDC/AGR-GPL/112217/2009, "Unravelling defense mechanisms underlying coffee resistance to *Colletotrichum kahawae*" developed at Centro de Investigação das Ferrugens do Cafeeiro/Instituto de Investigação Científica Tropical (CIFC/IICT) and funded by Fundação para a Ciência e Tecnologia (FCT). This work was focused on the bioinformatic analysis of Illumina RNA-seq data obtained from 24 cDNA libraries representing three key points of two *Coffea* spp.- *Colletotrichum kahawae* interaction (compatible vs incompatible), in order to identify coffee genes putatively involved in the plant resistance mechanism and quantify differences in gene expression during the defense response of coffee to *C. kahawae*.

The present work intends to contribute to a better understanding of the molecular genetic bases of coffee resistance to *C. kahawae* as well to increase the available genomic resources of both the fungus and the plant, which can be used in future studies.

Specifically, this research aimed at:
1 – Assembling a coffee transcriptome to use as basis for gene discovery and expression analysis, including a plant-fungus separation pipeline.
2 – Analyzing differential gene expression to characterize the defense response of two coffee genotypes, respectively resistant and susceptible to *C. kahawae*, during the early stages of the infection process.
3 – Assessing the differences between a custom and a standard RNA-sequencing data analysis and subsequently improving and optimizing data analysis towards the achievement of higher quality results regarding coffee transcriptome assembly and differential gene expression.

Chapter II

Differential expression profiling of coffee resistance vs susceptible response in the early stages of *Colletotrichum kahawae* infection

Joana Fino[1,2*], Andreia Figueiredo[3] , Andreia Loureiro[2], Elijah K. Gichuru[4] , Vitor Várzea[2], Maria C. Silva[2], Dora Batista[2]; Octávio S. Paulo[1]

[1]Computational Biology and Population Genomics Group, Centro de Biologia Ambiental, DBA/FCUL, P-1749-016 Lisboa, Portugal.
[2]CIFC-Biotrop/IICT-Centro de Investigação das Ferrugens do Cafeeiro-Biotrop/Instituto de Investigação Científica Tropical, Quinta do Marquês, 2784-505 Oeiras, Portugal
[3]Centro de Biodiversidade e Genómica Integrativa e Funcional, BioFIG, Faculdade de Ciências da Universidade de Lisboa (FCUL), Lisboa, Portugal.
[4]Coffee Research Foundation (CRF), Ruiru, Kenya

Abstract

Coffee berry disease (CBD), caused by the fungus *Colletotrichum kahawae*, is considered one of the biggest threats to Arabica coffee production, at high altitude, in Africa. Some coffee genotypes are known to be resistant to CBD, but the molecular genetic basis of coffee resistance is still unknown. With the purpose of gaining some insights on coffee resistance process, a RNA Illumina sequencing approach was used to characterize the transcriptional profile of two coffee genotypes, respectively resistant (Catimor 88) and susceptible (Caturra) to *C. kahawae*, during the early stages of the infection process. Twenty four cDNA libraries were sequenced and data was analysed by ARK-Genomics (UK) in order to assess differential gene expression when comparing inoculated with control samples. Here, a *de novo* transcriptome assembly was carried out with special care in the inoculated libraries for Coffee-fungus sequence separation. A differential expression pipeline was performed using the *de novo* assembled coffee transcriptome as reference. Our results were compared with ARK genomics analysis, revealing some variation on the transcriptome and differentially expressed unigenes, influenced by different aproaches. Finally, our analysis allowed the identification of genes putatively involved in coffee resistance, their expression profiles and the pathways in which they are involved.

Keywords: Coffee transcriptome; Defense mechanisms, plant-fungus sequence separation; gene expression

***Corresponding author:** Joana Rita Vieira Fino. E-mail: joana.fino@gmail.com; Address: Centro de Investigação das Ferrugens do Cafeeiro (CIFC)/ Instituto de Investigação Científica Tropical (IICT), Oeiras, Portugal;

Telephone: +351927854702

Runing title: Differential gene expression analysis of *Coffea* spp. - *C. kahawae* interaction

1. Introduction

Coffee is one of the most important commodities in the world economy, accounting for a trade worth of approximately 16.5 billion dollars in 2010 (van Hilten *et al.* 2011). Coffee growing countries are mainly located in Africa, Central and South America, and Asia where coffee production represents a major income, but particularly in Africa, people can depend entirely on this resource for their livelihoods (Lashermes & Anthony 2007). The commercial production relies mostly on two species: *Coffea arabica* L. and *Coffea canephora* Pierre ex A. Froehner, which represent about 70% and 30% of the market supply, respectively (Charrier & Berthaud 1985; Vieira & Andrade 2006). Despite of an increase in coffee production over the years, current production is still insufficient to satisfy the commercial demand (Muñoz et al 2010)

Coffee berry disease (CBD), caused by the pathogenic fungus *Colletotrichum kahawae* J.M. Waller & P.D. Bridge, is one of the limiting factors of *C. arabica* production. *C. kahawae* affects several organs of the crop, but major production losses occur when green berries are infected, leading to the formation of dark sunken lesions with sporulation, which results in fruit premature dropping and mummification (Silva *et al.* 2006; Hindorf & Omondi 2011).

The first report of this disease goes back to 1922, in Kenya, rapidly disseminating afterwards throughout Eastern Africa (McDonald 1926; Silva *et al.* 2006).The disease has stronger impact at high altitudes (>1700m) and is still, reportedly, confined to the African continent. However, at such similar altitudes and under appropriate climatic conditions, the disease may be able to colonize other continents (van der Vossen & Walyaro 2009).

Currently, chemical control has been successfully applied but its high cost, makes it unreachable for small scale producers. Crop damages due to CBD, along with chemical control costs, accounts annually for a loss of US$ 300–500 millions in Arabica coffee production (van der Vossen & Walyaro 2009). This severe problem stimulated the development of breeding programmes in several countries (such as Kenya, Ethiopia and Tanzania) giving rise to several resistant coffee varieties for coffee growers (Vossen & Walyaro 1980; Silva *et al.* 2006). In Kenya, the most relevant example is the hybrid commercial variety Ruiru 11, which was bred for

resistance to CBD and coffee leaf rust (*Hemileia vastatrix*) using lines of the coffee cultivar Catimor as resistance sources. In resistant coffee plants, , several mechanisms of defense can be observed, both constitutive and induced, working at different stages of the infection (Gichuru 1997): formation of cork barriers, early callose deposition around intracellular hyphae, hypersensitive-like cell death and early accumulation of phenolic compounds in the cytoplasm and the cell walls (Masaba & van der Vossen 1992; Silva *et al.* 2006; Loureiro *et al.* 2012b).

Despite the insights gathered so far about the cellular mechanisms of pathogen infection and host resistance, there is still no information about the molecular and genetic basis of coffee resistance to CBD. Gaining new insights into the defense response of *C. Arabica* to *C. kahawae* is of the utmost importance.

RNA-Seq has been successfully used to accurately quantify transcript levels, with potential advantages over microarray-based methods (Griffith *et al.* 2010; Nagalakshmi *et al.* 2010). Global gene expression analysis has emerged as an important tool for studying how organisms, such as plants, respond to stresses, such as abiotic stress, or biotic stress caused by pathogen infections (Liu *et al.* 2012; Peatman *et al.* 2013). Several studies in other host-pathogen interactions recurring to RNA-seq approaches, reported the use of the technique to perform *de novo* transcriptome assembly and annotation, estimate expression of specific isoforms and compare gene expression between a pair of contrasting conditions (Griffith *et al.* 2010). Successful results were achieved, being an example the case of *Citrus reticulata* infected by *X. fastidiosa*, in which expression analysis identified several defense response-related genes (Rodrigues *et al.* 2013). Congruent results were found through the sequencing of Sorghum infected by *Bipolaris sorghicola,* for which both plant and pathogen transcriptomes were analysed,, identifying genes involved in the host defense response (Yazawa *et al.* 2013).

In our study, Illumina RNA-seq data was produced for two interactions of *Coffea sp – C. kahawae* (compatible and incompatible, corresponding to susceptible and resistant coffee genotypes), during the early stages of infection, aiming to characterize transcriptional differences. A first analysis by the sequencing company ARK genomics (UK) was made with a pipeline used for a generality of types of data, including the softwares SOAPdenovo-Trans

(http://soap.genomics.org.cn/SOAPdenovo-Trans.html, last access April 11th 2014) and EdgeR (Robinson *et al.* 2009). However, the use of a standard analysis with standard software and parameters may not be perfectly suitable for this data. Yang & Smith (2013) have shown the possible qualities of the unpublished software SOAPdenovo-Trans but the use of a well documented software, with detailed information, and the ability to test different parameters, adjusting the analysis to our data, may be preferable (Wilson *et al.* 2014). On the other hand, EdgeR, developed for analysis with few replicates, can be too liberal for differential expression assigning (Soneson & Delorenzi 2013; Seyednasrollah *et al.* 2013). In addition, the potential of deeply exploring and getting more revenue from the data, showed the demand for a different and more focused approach. Therefore, here we report the deployment of a new expression analysis, with a custom workflow, and the subsequent advantages provided on result quality. Also, functional categories and metabolic pathways were identified as putatively involved in coffee resistance to *C. kahawae.*

2. Material and Methods

2.1. Inoculation of coffee hypocotyls and sampling

Hypocotyls were used as a model material to study CBD because previous studies have shown a correlation between the pre-selection test on hypocotyls and mature plant resistance in the field (r=0.73–0.80) (Van der Vossen *et al.* 1976). Hypocotyls of the cultivars Catimor 88 (resistant genotype) and Caturra (susceptible genotype) were inoculated with the *C. kahawae* isolate Que2 (from Kenya), as described by Figueiredo *et al.* 2013. After inoculation, hypocotyls were vertically placed on plastic trays containing a wet nylon sponge and sprayed with a conidia suspension ($2x10^6$/ml) (inoculated samples) or with water (mock-inoculated hypocotyls – control samples). Afterwards, trays were covered with plastic bags and kept in a Phytotron 750 E at 22ºC in the dark for 24h, and then under a photoperiod of 12 hours during the inoculation time-course.

Hypocotyls were harvested at 24, 48 and 72 hours post inoculation (hpi), corresponding to different stages of the infection process, as described in Loureiro *et al.* 2012a: i) differentiation of melanised appressoria (in both coffee genotypes) at 24 hpi; ii) fungal penetration and establishment of biotrophic phase (susceptible

genotype) or beginning of hypersensitive cell death (HR) and accumulation of phenols (resistant genotype) at 48hpi; iii) switch to the necrotrophic phase (susceptible genotype) or display of HR and phenols deposition in more that 50% of infection sites (resistant genotype) at 72 hpi. Two independent experiments were conducted and 40 hypocotyls were collected for each coffee genotype (Catimor 88 and Caturra) and time points, both at control and inoculated conditions. Plant material was immediately frozen in liquid nitrogen and stored at -80ºC.

2.2. Extraction and sequencing

Total RNA was isolated from hypocotyls of all samples with Spectrum™ Plant Total RNA Kit (Sigma-Aldrich, USA), according to the manufacturer's instructions. Total RNA purity and concentration was measured at 260/280 nm and 260/230 nm using a spectrophotometer (NanoDrop- 1000, Thermo Scientific), while RNA integrity was verified by gel electrophoresis. mRNA-seq library construction for each independent sample and replicate (Table 1), in a total of 24, was performed at ARKs Genomics (UK) for subsequent 100bp paired-end sequencing on a flow cell composed of 4 lanes on a Illumina HiSeq2000.

Table 1 - List of the cDNA libraries produced with information about the genotype, condition (inoculated and control), time-points (hpi-hours post inoculation), experimental replicates and respective identification.

Control												
Genotype	Resistant – Catimor						Susceptible - Caturra					
Time (hpi)	24		48		72		24		48		72	
Exp Replicate	I	II	I	II	I	II	I	II	I	II	I	II
Identification	R1C24	R2C24	R1C48	R2C48	R1C72	R2C72	S1C24	S2C24	S1C48	S2C48	S1C72	S2C72
Inoculated												
Genotype	Resistant - Catimor						Susceptible - Caturra					
Time (hpi)	24		48		72		24		48		72	
Exp Replicate	I	II	I	II	I	II	I	II	I	II	I	II
Identification	R1Q24	R2Q24	R1Q48	R2Q48	R1Q72	R2Q72	S1Q24	S2Q24	S1Q48	S2Q48	S1Q72	S2Q72

2.3. Read processing

Previously to the assembly steps, two approaches of sequence cleaning were taken: One applying contaminant cleaning, using SeqTrimNext version 2.0.59 (Falgueras *et al.* 2010) for the control libraries-derived reads (from this point onwards designated as

control reads for simplicity), and other excluding contaminant cleaning, using TrimGalore! Version0.3.3 (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/, accessed in March 9th 2013), for the inoculated libraries-derived reads (from now on designated as inoculated reads), in order to also retrieve *C. kahawae´*s sequences. For the subsequent step of transcriptome mapping, the inoculated reads were then further processed by SeqTrimNext, for contaminant cleaning.

2.4. Transcriptome assembly and Scaffolding

Two transcriptome assemblies (one with the control reads and another with the inoculated reads) were performed using Velvet version 1.2.08 (Zerbino & Birney 2008) and Oases version 0.2.08 (Schulz *et al.* 2012), with a k-mer value of 31, a coverage cutoff of 0.377 and a minimum contig length of 200 bp. As a first step, the transcriptome assembled from the inoculated reads was surveyed for the presence of fungus sequences.

Afterwards, in order to complete the reference transcriptome, the contigs from the control reads assembly and the contigs classified as "plant" and "possibly plant" in the plant-fungus contig identification step (from the transcriptome assembly with the inoculated reads) were clustered together using the software CD-HIT-EST version 4.6.0 (Li & Godzik 2006) with a contig identity > 90%.

The clusters were then scaffolded with SSPACE version 2.0 (Boetzer *et al.* 2011) without extension and a minimal number of read pairs to compute a scaffold of 4.

Due to a highly repetitive transcriptome assembly, a redundancy pipeline using blastn's version 2.2.25+ (Camacho *et al.* 2009) was applied using the two best hits and the whole transcriptome as both query and subject (discarding the 1st hit since it is always a self match), with a minimum e-value of 10^{-5}, and an alignment length with at least half of the length of the query sequence (based on Calduch-Giner *et al.* 2013). The scaffolds with the same hits were grouped as being sufficiently similar to be considered the same. Only the longest sequence of each group was considered as part of the final transcriptome. The entire previous process was run two times, until no hits between different sequences were found.

2.5. Plant-Fungus contig identification

Using the transcriptome assembled with the inoculated reads and skipping the contaminant cleaning step, two methods were used to identify the contigs of plant sequences and the contigs of fungus sequences for subsequent transcript separation:

a) Mips-EST3 (Emmersen *et al.* 2007) that uses triplet nucleotide frequencies to classify contigs as plant or fungus;

b) A pipeline based on Fernandez *et al.* 2012 that uses blastn searches against i) NCBI coffee and fungus available sequences (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/, downloaded April 27th 2013); ii) the control assembled transcriptome; and iii) 3 *Colletotrichum* genome databases:*C. gloeosporioides*, *C. graminicola* and *C. higginsianum* (Sequencing Project, Broad Institute of Harvard and MIT `http://www.broadinstitute.org/`, accessed April 20th 2013), with a minimum e-value of 10^{-5}, to evaluate the probability that each contig has to be considered plant or fungus.

MIPS-EST3 uses groups of sequences of *C arabica* and *C. kahawae* properly identified for the classificator training. The "training" step was performed using nucleotide sequences downloaded from NCBI databases. The trained classificator has a dinucleotide bias distance of genomes of 97.26 which, according to the authors, is sufficient for a confident separation of the sequences (Emmersen *et al.* 2007). Finally the classificator is applied to the transcriptomic contigs and classifies them as either "plant" or "fungus".

The blast pipeline is based on the X value which is calculated by subtracting the mean score of the best hits against the fungus databases to the mean scores of the best hits against the coffee databases. This value is then used as a measure of similarity with coffee and fungus sequences. Thus, the X is used to classify the contigs in "Plant", "Fungus" or "Unclassified" categories. Figure 1 shows a scheme of the process.

Finally, the results of the two methods were crossed and the contigs were separated into 5 categories: *Plant* or *Fungus* when the results of the two methods were concordant, *Potentially plant* or *Potentially fungus* when the blast pipeline lacked classification and *uncategorized* when the results of both methods were contradictory.

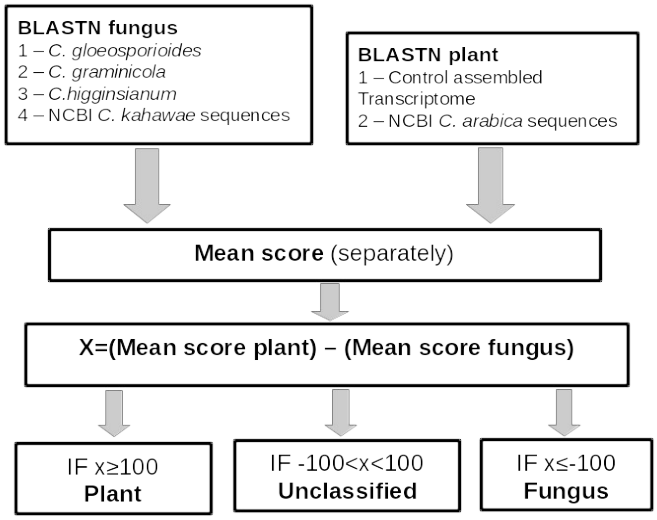


Figure 1 – Schematic diagram of the blastn's pipeline for the Plant-Fungus contig identification. Nucleotide blast was performed against each of the databases, mean score and X value calculated, and finally the contigs were classified as Plant, fungus or unclassified.

2.6. Read mapping, expression quantification and differential expression analysis

For the read mapping and expression quantification, both the transcriptome previously assembled and the coffee ESTs from a 454 assembly (Santos 2011) were used, separately, as reference. The program used for this task was Rsem version 1.2.10 (Li & Dewey 2011). This software runs Bowtie version 0.12.7 (Langmead *et al.* 2009) for the different libraries separately, to find all the possible alignments, with a maximum of 3 mismatches per read.

For differential expression estimations, the R package from Bioconductor, EBSeq version 1.1.5 (Leng *et al.* 2013) was used. Only unigenes with a posterior probability of being differentially expressed (PPDE) > 0.95 and a -1.0 ≥ log2 fold change ≥1.0 were considered as such.

2.7. Sequence Annotation

De novo functional annotation of the coffee transcriptome was obtained by similarity using Rapsearch2 (Zhao *et al.* 2012), Blast2GO (Conesa *et al.* 2005) and custom made scripts. Rapsearch2 was used to search against functional proteins from the KOG (euKaryotic Orthologous Groups) database which is a component of the Clusters of Orthologous Groups (COG) database (Tatusov *et al.* 2003), restricted by

A*rabidopsis thaliana* sequences and an E-value < 10^{-5}. Additionally, for the Gene Ontology (GO) annotation, Rapsearch2 similarity searches were locally conducted against non-redundant ("nr") peptide database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/ downloaded at November 26, 2013, including all "nr" GenBank CDS translations + PDB + SwissProt + PIR+PRF). Rapsearch2 search was carried out using default parameters and an E-value < 10^{-5}. The outputs were then converted to XML format which is similar to the Blastx output, with an in-house developed script Rapsearch2XML (https://github.com/Nymeria8/Rapsearch2Xml, last access May 6th 2014). The output was then used in the Blast2GO software for functional annotation using GO terms. Further, KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa *et al*. 2008) pathways were assigned to the assembled sequences using also the Blast2Go software.

2.8. Ark Genomics Workflow

The ARK Genomics analysis used the inoculated and control quality trimmed reads for a transcriptome assembly, using the software SOAPdenovo-Trans version 1.01 with a k-mer of 21. The reads were them mapped back to the assembled transcriptome and separately to the group of coffee ESTs from the 454 assembly (Santos 2011). This task was performed by BWA version 0.6.2. (Li & Durbin 2009) Potentially differentially expressed genes were identified using R package EdgeR version 2.13.0, and filtered by $-1.0 \geq \log2$ fold change ≥ 1.0 and a p-value ≤ 0.05.

3. Results

In this study, an analysis of RNA-seq data was performed to assess the process of coffee defense to *C. kahawae*. With the lack of a reference genome, a reference transcriptome was *de novo* assembled to use as base for the expression quantification. To get some insights on the plant reaction to the infection, three different data comparisons were made: Control vs infected, time-point vs time-point and infected resistant vs infected susceptible. Differentially expressed transcripts were then annotated using GO, KOG and KEGG annotations and the profiles of expression categorized. Lastly, a EST based differential expression analysis was used to compare the current and ARK genomics analysis.

3.1. Transcriptome assembly

Sequencing of the 24 libraries generated a total of 1,552,057,070 paired-end reads of 100bp. After the cleaning steps, it was possible to recover 1,540,810,726 paired-end reads of variable size: 65,734,996 inoculated reads and 888,075,730 control reads.

Due to the lack of a coffee reference genome for subsequent read mapping, and aiming to separate all the available fungus information, two independent assemblies were performed: one for a reference transcriptome construction using only control reads, and the other for plant-fungus sequence separation using the inoculated reads(Table 2).

This step produced 614041 contigs with an average length of 1056.56 bp for the coffee transcriptome, and 656839 contigs with an average length of 1092.40bp for the plant-fungus transcriptome.

The coffee transcriptome clustering step using the already assembled contigs, and also the sequences considered as "plant" and "potentially plant" in the plant-fungus sequence identification step (section 3.2 of results), with a minimum size restriction of 200bp. This resulted on a total of 284482 contigs, with an average length of 1470.43 bp, and a N50 of 2285. Finally, all the contigs were scaffolded, resulting in a total of 283928 scaffolds.

Due to RAM restrictions at our bioinformatics facility (178Gb), the only k-mer value that was possible to use was 31, which led to a higher level of redundancy than expected. This finding led to the inclusion of an extra step for redundancy cleaning, using a Blast pipeline. This step was repeated twice, which drastically decreased the redundancy, and consequently the length of the resulting transcriptome. At the end , a final set of 65759 unigenes was obtained, with an average length of 1398.64 bp.

3.2 Plant-fungus sequence identification

Combining the results from the two methods used for plant-fungus identification, the contigs were assigned to 5 different categories as follows: plant/fungus when both methods classified the contig as such; Potentially plant/fungus, when the blast pipeline result was inconclusive but the MIPS – EST3 method was conclusive; and Unclassified when the results of the two methods were discordant.

Table 2 - Summary of *de novo* assembly results of Illumina sequence data from *Coffea* sp. (Coffee transcriptome assembly and Plant-Fungus Transcriptome assembly)

Assembly Steps	Contigs	N50	% > 1kbp	Max. length(bp)	Average length(bp)
Coffee Transcriptome					
Oases Contigs	614041	1897	37.93	14662	1056.56
Clustering	284482	2285	52.51	14662	1470.43
Scaffolding	283928	2288	52.83	14662	1459.75
Redundancy cleaning step 1	83940	2598	50.45	14662	1509.95
Redundancy cleaning step 2	65759	2623	44.30	14662	1398.64
Plant-Fungus Transcriptome					
Oases Contigs	656839	1980	39.19	13181	1092.40
Clustering	209580	2220	51.24	13181	1410.75

From the positive identification provided by both methods, 198036 contigs were considered as “plant” and 653 were considered as “fungus”. From the inconclusive classification, 8564 and 119 were respectively considered as “Potentially plant” or “Potentially fungus”, while 2208 contigs were not classified due to contradictory results presented by both methods (Table 3). The unclassified category may include not only contigs that failed to be properly classified, but also contaminant sequences, which are neither from *Coffea spp*. nor from *C. kahawae.*

Table 3 - Plant-fungus contig identification summary.

	Classified	Potentially classified	Unclassified
Plant	198036	8564	2208
Fungus	653	119	

3.4 Transcriptome sequence annotation

The KOG annotation against *Arabidopsis thaliana* database revealed 23252 annotations (35.36% of the transcriptome) divided by 6 categories: “metabolism” (6.91%), “celular processes and signaling”(9.49%), “information storage and processing” (6.07%), “other function” (3.83%), “unknown function” (1.85%) and

"general prediction only" (7.22%) (Figure S1). From the 65759 unigenes of the transcriptome, 20335 were annotated using GO-terms, which represents 30.92% of the transcriptome. These annotations are distributed in 3 main GO domains, with a total of 83398 GO terms. Of these assigned GO terms, "Biological Process" was the predominant domain with 43.80%, followed by "Molecular Function" with 30.75% and "Cellular Component" with 25.44% (Figure 2). The KEGG annotation was also performed and only 5.85% of the transcriptome (3850 unigenes) was successfully identified in a total of 136 metabolic pathways (Figure S2). Predominantly represented pathways are "purine metabolism" with 553 unigenes, "starch and sucrose metabolism" with 368 and "phenilalanine metabolism" with 205.

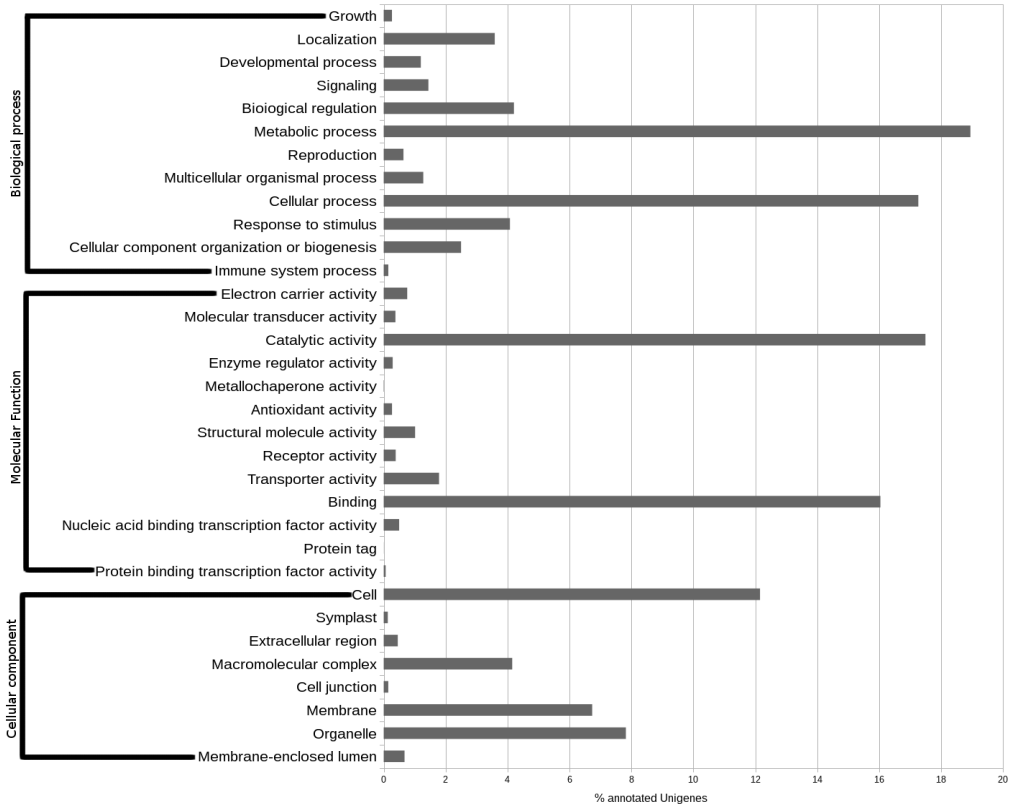


Figure 2 - GeneOntology classification of the transcriptome sequences. The categories are represented by percentage relatively to the total of the transcriptome unigenes. The classification are displayed in three main categories: cellular component, molecular function and biological process.

3.5 Differential expression analysis

3.5.1 Within genotypes

Differential expression analysis by comparison of inoculated with control samples (including the 3 time points) was carried out for both susceptible and resistant coffee genotypes (Table 4).

From the total differential expression analysis , 1713 unigenes were identified as differential expressed. A predominance of unigenes being differentially expressed in the resistant genotype (1617 vs 567 unigenes) was observed, as well as a predominance of upregulated unigenes. Consequently, a higher number of shared upregulated unigenes among both genotypes was found. It was possible to see the number of differentially expressed unigenes rising with time. For example, at 24 hpi a total of 238 and 25 unigenes were differentially expressed in resistant and susceptible samples, respectively, while at the 72 hpi the number of unigenes raised to 1423 and 622. For the resistant genotype, 27 unigenes were found to be expressed only at 24 hpi, and 845 at 72 hpi. At 48 hpi, both susceptible and resistant genotypes did not show unigenes expressed uniquely on that time-points.

Table 4 - Number of differentially expressed unigenes at the 3 sampled time-points of the inoculated resistant and susceptible genotypes relative to the control. Shared category indicates the number of differentially expressed genes in both genotypes. Values inside brackets correspond to unigenes only expressed in a respective time-point. Values indicate unigenes passing cutoff values of -1.0 ≥ log2 fold change ≥1.0 and PPDE > 0.95

	24h	48h	72h	Total
Up regulated				
Resistant	228(18)	671 (0)	1169 (610)	1320
Susceptible	22(8)	517 (0)	520 (235)	526
Shared	3	371	14	
Down regulated				
Resistant	10 (9)	52(0)	254 (235)	297
Susceptible	3(2)	38(0)	102 (94)	41
Shared	0	3	1	

3.5.2 Between time-points

Furthermore, to compare different infection time points per genotype, resistant and susceptible inoculated libraries were used: 24hpi vs 48hpi, 24hpi vs 72hpi and 48hpi vs 72hpi for each of the genotypes. A total of 521 unigenes were considered differentially expressed, with a predominance of upregulated unigenes. The results showed a higher number of differentially expressed unigenes between time points 24h and 72h comparatively with the other two pairs of conditions (Table 5). As in the comparison between control and inoculated, the resistant genotype presented a higher number of differentially expressed unigenes, with 397 against 176 unigenes for the susceptible genotype.

Table 5 - Statistics of differentially expressed unigenes between time points, for the two inoculated genotypes. Values indicate unigenes passing cutoff values of -1.0 ≥ log2 fold change ≥1.0 and PPDE > 0.95

	Susceptible	Resistant
48h over 24h		
Upregulated	67	82
Downregulated	10	37
Shared with 72h over 48h	14	4
Shared with 72h over 24h	2	13
Total	77	119
72h over 48h		
Upregulated	20	75
Downregulated	14	5
Shared with 72h over 24h	8	12
Total	34	80
72h over 24h		
Upregulated	68	198
Downregulated	21	50
Shared by the 3 comparisons	0	1
Total	89	248

3.5.3 Between genotypes

To recover a larger level of information from the data, an additional analysis was conducted from directly comparing differences in expression profiles between

resistant and susceptible inoculated genotypes, at 24 hpi, 48 hpi and 72 hpi. Designating the susceptible genotype as the "control" group, comparisons between the two genotypes in each time point were performed (Figure 3). A total of 699 unigenes were classified as differentially expressed for at least one time point. At 24 hpi the number of unigenes differentially expressed in the two genotypes is the most similar comparatively with the other times points: 124 unigenes for the susceptible and 145 for the resistant. In contrast, the susceptible genotype showed a pike of expression at 48 hpi (214 unigenes), and decreasing at 72 hpi (189 unigenes), while the expression of the resistant genotype, relatively to the susceptible genotype, increased with time – at 48 hpi, 156 unigenes and at 72 hpi 229 unigenes.

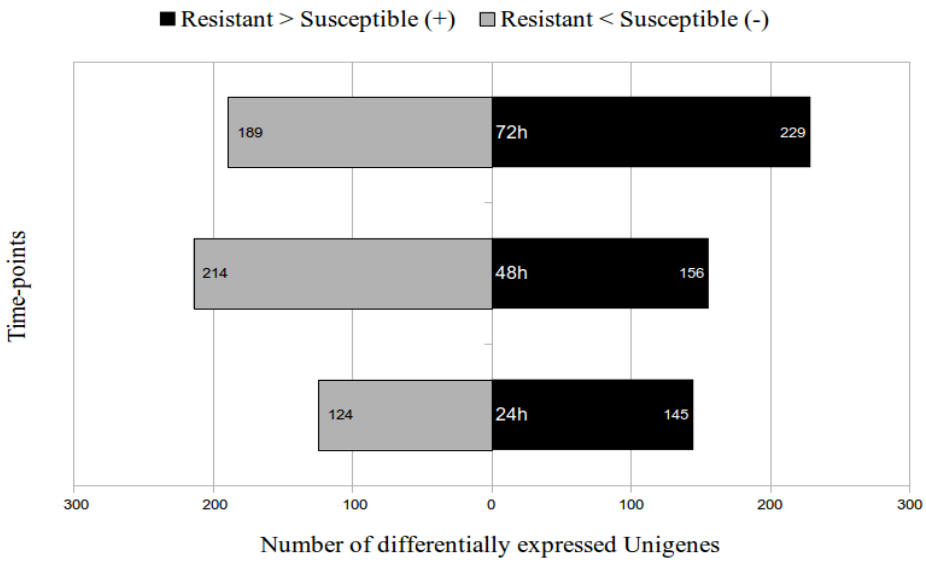


Figure 3 - Statistics of differentially expressed unigenes between resistant and susceptible genotypes at the three time points. Resistant > Susceptible indicates number of unigenes with significantly higher expression in resistant samples relative to susceptible samples, and vice-versa for Resistant < Susceptible. Values indicate unigenes passing cutoff values of -1.0 ≥ log2 fold change ≥1.0 and PPDE > 0.95

3.6 Annotation and expression profiles of the differential expressed unigenes

To obtain a general perspective of the biological processes influenced by the infection of *Colletotrichum kahawae*, we selected the KOG and KEGG annotations for the unigenes identified as differentially expressed against their control. From the KOG

annotation, 24 KOG categories were assign to 1224 differentially expressed unigenes (49.18% of the total differentially expressed unigenes). A high percentage of unigenes was assign to non descriptive categories: 37.34% unigenes in “other functions”, “Function unknown” and “General function only”. The category represented with a larger number of unigenes was “Cellular processes and Signalling” with 28.02%, followed by “Metabolism” with 26.39% and “Information storing and processing” with 8.25%. Dividing the annotations by pair of comparisons (Figure 4, S3, S4), it was possible to identify the category “signal transduction mechanisms” well represented in all the comparisons. At 24 hpi, the resistant genotype showed a higher amount of identified categories, including “defense mechanisms”. At 48 hpi and 72 hpi, there were few differences between the susceptible and the resistant genotype annotations, where “signal transduction mechanisms” and “posttranslational modification, protein turnover, chaperones” were the most represented categories.

The KEGG annotation included 33 of the total differentially expressed unigenes which were assigned to 100 different pathways. Observing the annotations of the pairs of comparisons, it was possible to identify different categories as mainly represented. In the resistant genotype comparisons, the pathways involved with phenylalanine and phenylpropaniod biosynthesis and metabolism, are the most representative, while in the susceptible genotype, “Starch and sucrose metabolism” is the most representative pathway, independently of the time-point (Figure 5, S5, S6).

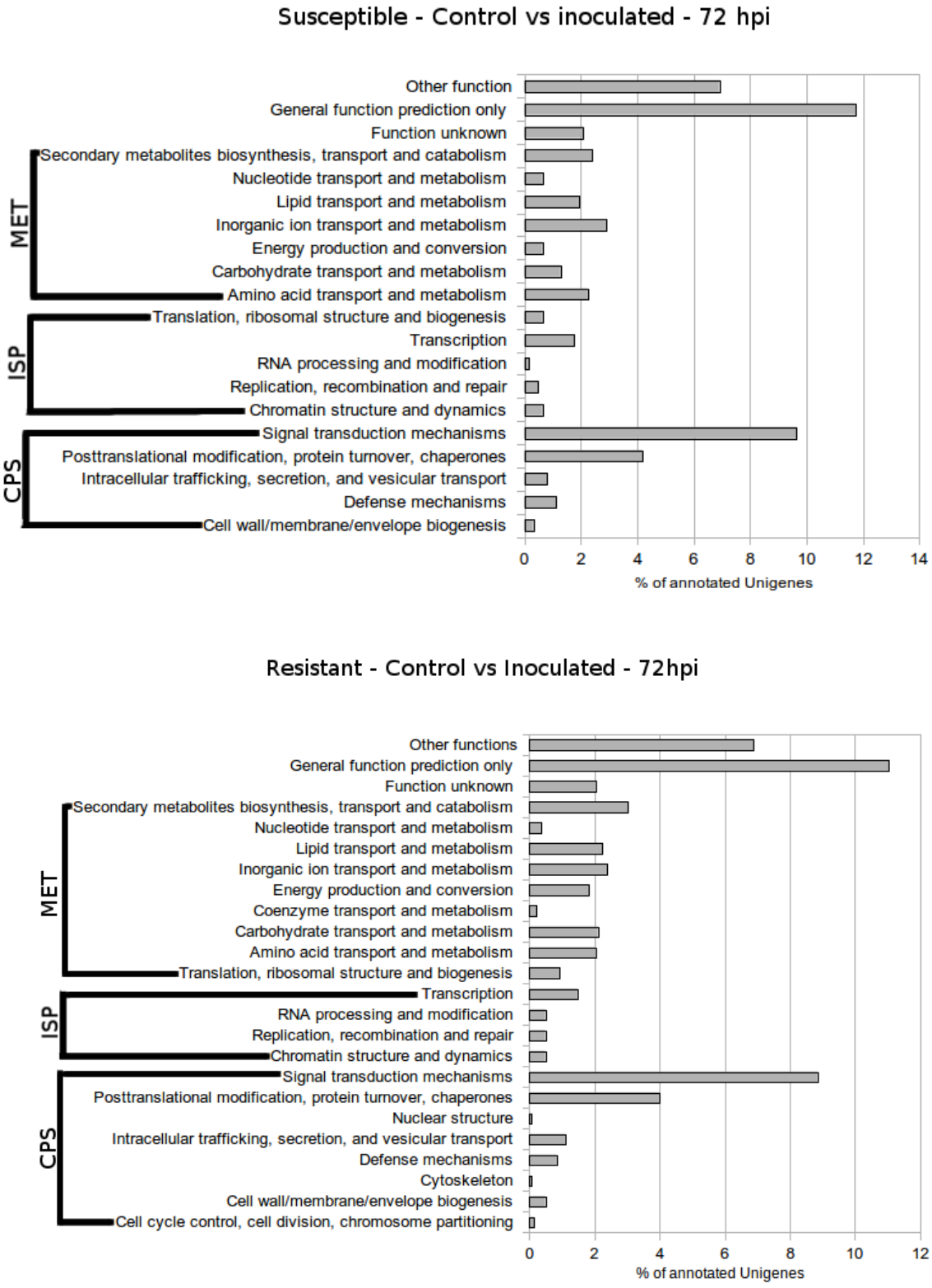


Figure 4 - KOG annotation of the unigenes identified as DE in the susceptible and resistant genotypes comparison between control and inoculated at 72hpi. The annotations are divided by 3 main categories: CPS - "Cellular processes and signalling"; ISP - "Information storage and processing"; MET – "Metabolism".The percentage of unigenes is relatively to the total of differentially expressed unigenes of each comparison.

For the differentially expressed unigenes identified between genotypes, a GO annotation was made to identify categories of interest (Figure 6). Defense related categories were selected with the aim of excluding the bias that could be introduced by comparing different genotypes. A reduced number of unigenes were annotated in defense related categories, with a total of 64 unigenes, 39 more expressed in the susceptible genotype and 25 more expressed in the resistant genotype. “Response to stimulus” is the most representative category for both genotypes, especially at 72 hpi, with 6 and 10 unigenes for susceptible and resistant, respectively. The categories “response to stress” and “response to reactive oxygen species” appeared represented at 48 hpi for the susceptible genotype, but only at 72 hpi for the resistant genotype. In both genotypes the category “response to other organism” is only represented at 72hpi.

To study the differentially expressed unigenes over the time-course, 14 profiles of expression were identifiedusing comparisons between control and inoculated samples (Figure 7).

These profiles involve different groups of time points, depending on the presence or absence of the unigenes as differentially expressed. In this way, the profiles can include the three time points (24 hpi, 48 hpi and 72 hpi) or just two of them (24 hpi and 48 hpi, 24 hpi and 72 hpi or 48 hpi and 72 hpi). Differences in expression were evaluated using fold change logarithm values: when the differences between fold change logarithm of two time-points was higher than one, it was assumed that the two time-points have differences of expression; otherwise, they are considered as stably expressed. So, the differences of expression showed in the different profiles do not correspond to absolute values but to comparisons between the fold changes along the three time-points.

The results showed a majority of upregulated unigenes, that stayed upregulated besides the differences of expression over time. The same was verified for the downregulated unigenes, with the exception of the differentially expressed unigenes of the susceptible genotype of the n) profile, which shifts from downregulated to upregulated in the time course.

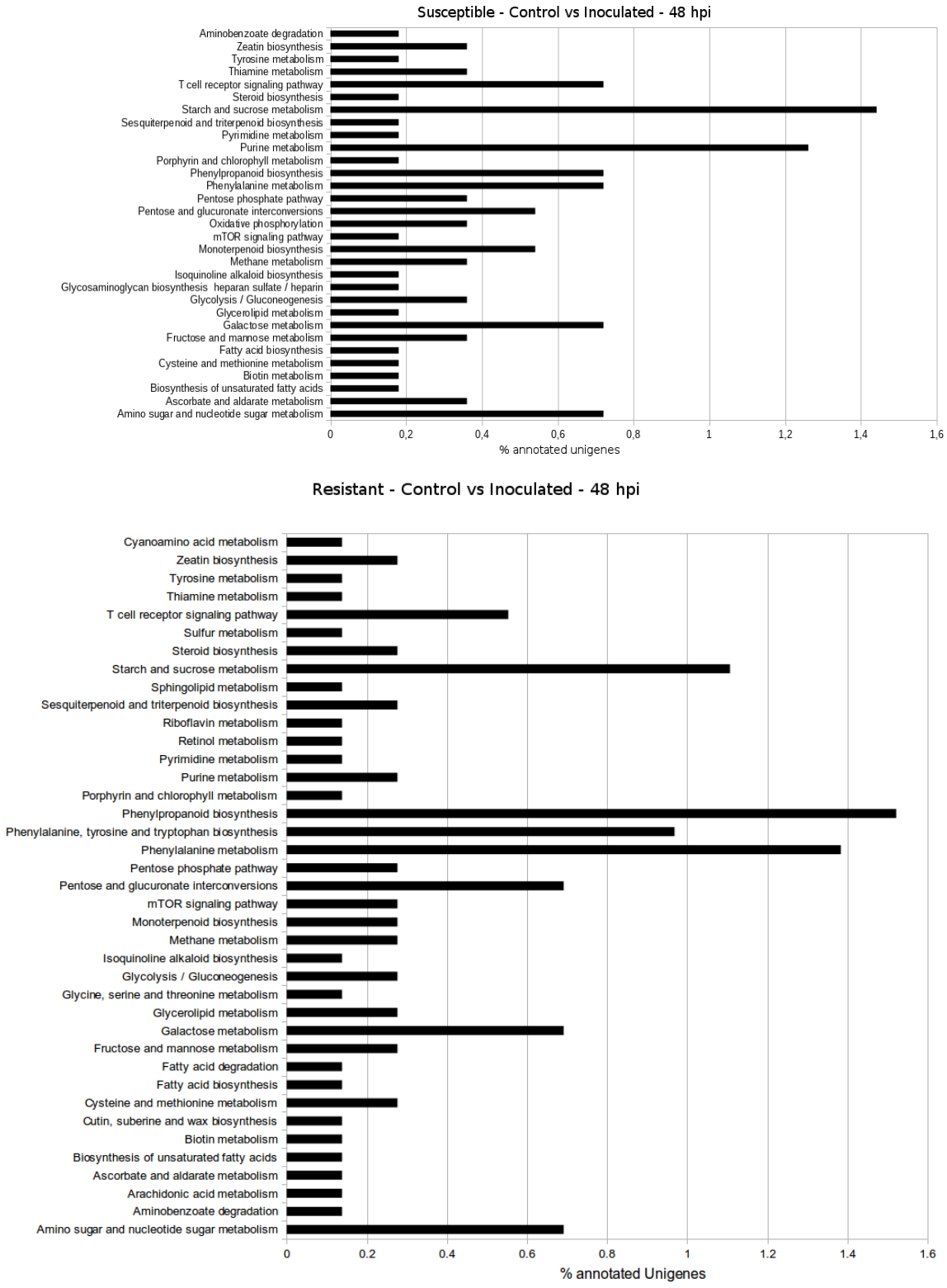


Figure 5 - KEGG annotation of the unigenes identified as DE in the susceptible and resistant genotypes comparison between control and inoculated at 48hpi.The percentage of unigenes is relatively to the total of differentially expressed unigenes of each comparison.

Unigenes identified with a three time points expression profile were few, especially for the susceptible genotype. Among theses, the profile a) (stable over the three times) was the most representative for the susceptible genotype with four unigenes. On the other hand, the resistant genotype was mainly represented by f) (increases the expression from 24 to 48 hpi, and then stabilizes at 72 hpi). The expression profiles with unigenes only at 24 and 48 hpi are reduced, where g) presents three and five unigenes for the susceptible and resistant genotypes, respectively, and h) with one for each genotype. The profiles accounting only two time-points were identified as the most common, with the profiles i) (stable at 48 and 72 hpi) and j) (increases from 48 hpi to 72 hpi) as the most representatives of all. The expression profile i) presented 265 for the susceptible genotype and 434 unigenes for the resistant genotype. The expression profiles with absent values at 48hpi (l), m) and n) were poorly represented with a total of two and 22 unigenes for the susceptible and resistant genotypes, respectively.

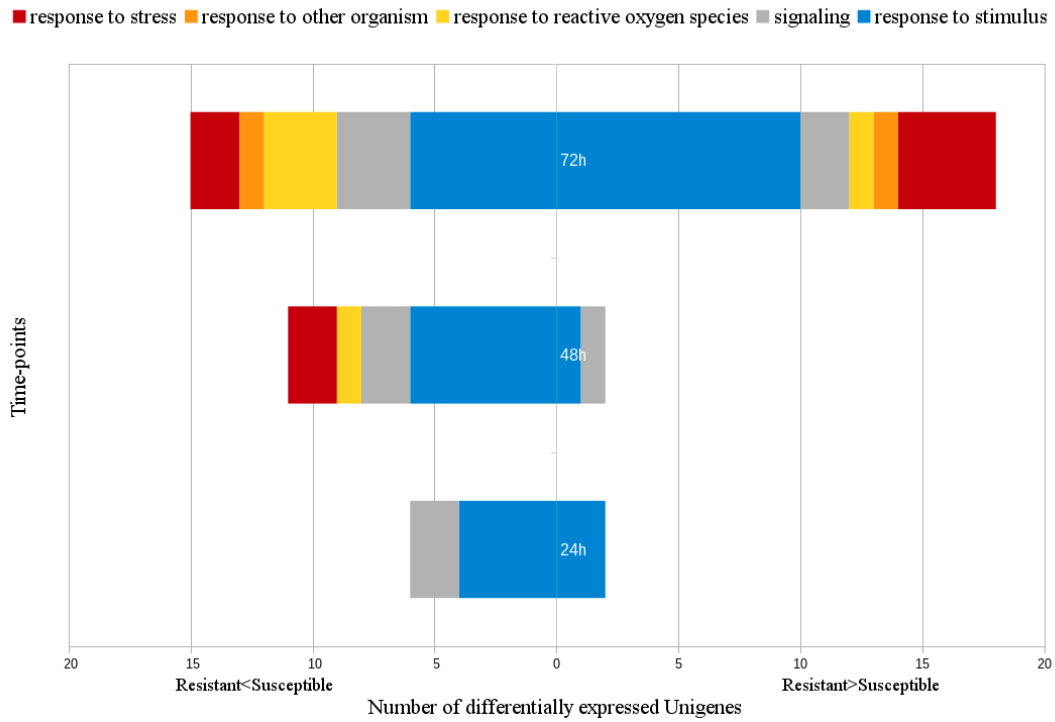


Figure 6 - GO annotation of the unigenes identified as differentially expressed between genotypes in the 3 time-points. Only categories of interest are represented. The unigenes at the left of the axis are more expressed in the susceptible genotype, and at the right of the axis the unigenes more expressed at the resistant genotype

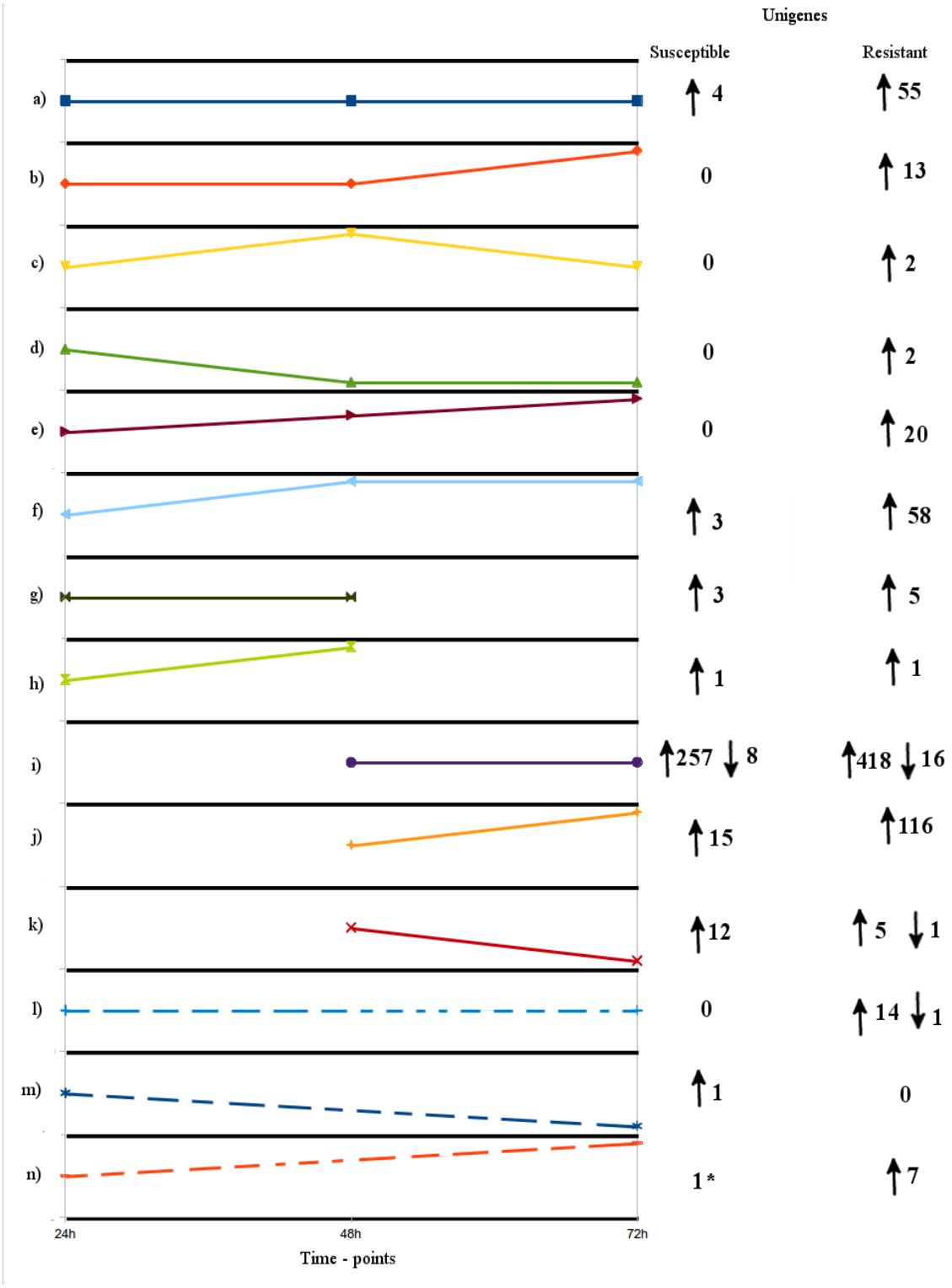


Figure 7 - Profiles of expression identified in the comparisons between control and inoculated samples in the two genotypes. Each coloured line represents a different profile of expression along the 3 times of inoculation. The dashed lines represents unigenes only differentially expressed at 24 and 72 hpi. At the right side we can see the number of unigenes upregulated (arrow up) and downregulated (arrow down) per profile and genotype. The asterisk represents the unigene that shifted from downregulated to upregulated in the time course The unigenes were considered up and down regulated in relation to the previous time point if the difference between fold change ≥1.0.

To associate these profiles to different biological processes, the most relevant ones were selected, and the corresponding unigenes organized by their KOG and KEGG annotations.

In the KOG annotation (Figure 8) for the stable profile a), the unigenes of the resistant genotype were annotated in the categories “Signal transduction mechanisms”

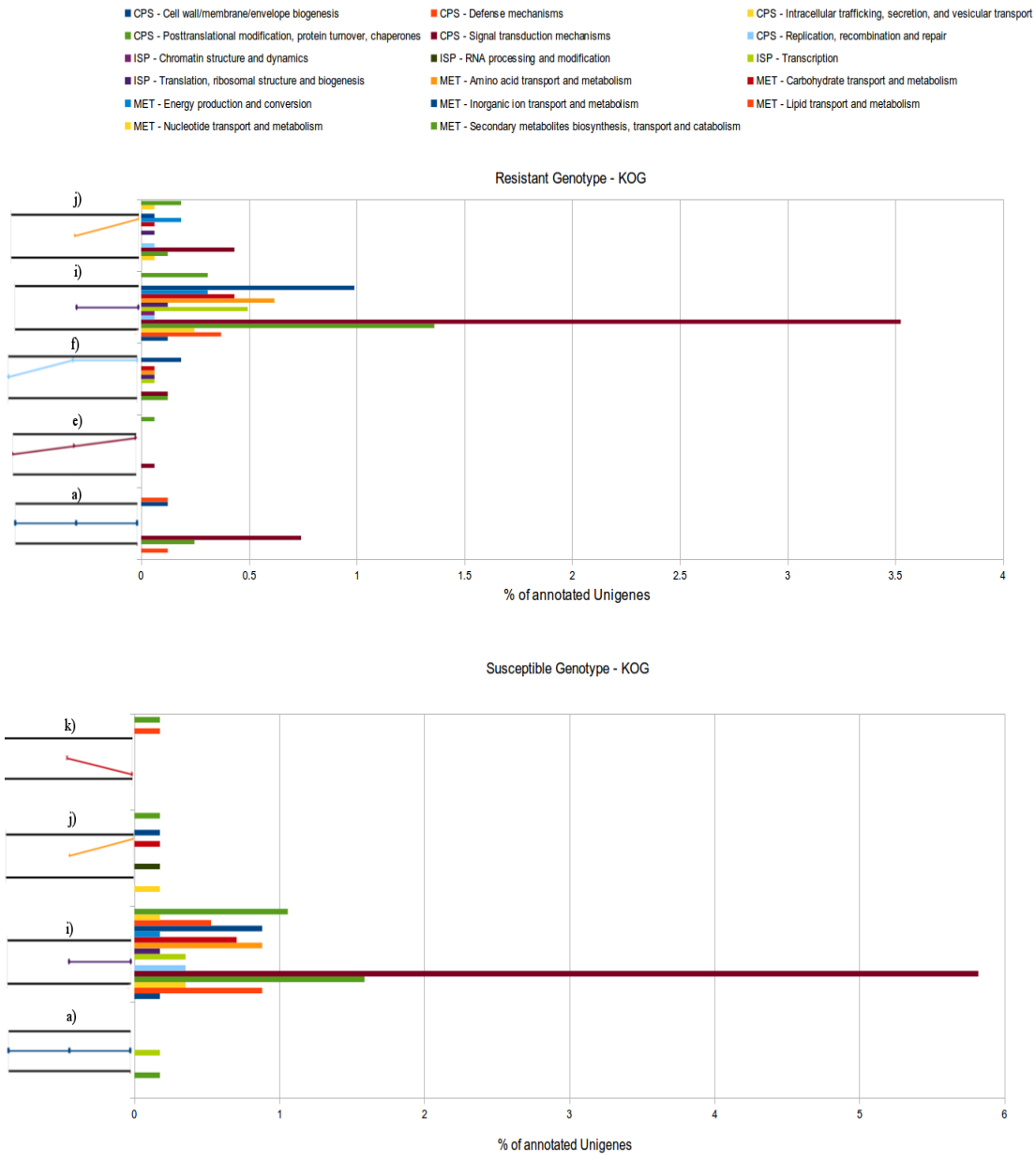


Figure 8 - KOG annotations by profile of expression for the resistant and susceptible genotypes. The percentage of unigenes is relatively to the total of differentially expressed unigenes of the susceptible comparisons – Control vs Inoculated. Three categories are represented: MET - Metabolism, ISP - Information Storing and Processing, CPS - Cellular Processes and Signaling.

"Posttranslational modification, protein turnover,chaperones" and "Defense mechanisms" while the susceptible genotype, only had annotations in the categories "Transcription" and "Posttranslational modification, protein turnover,chaperones".

The profile e), only represented in the resistant genotype, was associated to the categories "Signal transduction mechanisms" and "Secondary metabolites biosynthesis, transport and catabolism". Also, the profile i), which is similar between genotypes, presented two categories as most represented: "Signal transduction mechanisms" and "Posttranslational modification, protein turnover,chaperones". The j) and k) profiles, for the susceptible genotype, were associated to categories related with the production of energy, while for the resistant genotype the j) profile, showed also "signaling transduction mechanisms" associated.

The KEGG annotation revealed that in the susceptible genotype the differentially expressed unigenes mostly belong to categories related with energy production, such as "Amino sugar and nucleotide sugar metabolism" and "Starch and sucrose metabolism" (Figure S7a). In the resistant genotype, in addition to the above categories, pathways related with phenylpropanoid and phenylalanine are represented (Figure S7b).

3.7 Workflow comparisons

The ARK genomics assembly made with the software SOAPdenovo-trans resulted on 62579 contigs with an average length of 785.31bp (Table 6). Comparing this with our assembly (velvet/oases step of assembly), it was possible to see an increase on the amount of information retrieved, both in the total number of contigs and in the size of these contigs (614041 contigs, with N50 of 1897 and an average length of 1056.56 bp). Consequently, our transcriptome presents a higher percentage of contigs with more than 1000 bp.

Table 6 – Statistics of the ARK genomics assembly made with the software SOAPdenovo-trans.

Contigs	N50	% > 1000bp	Maximum length(bp)	Average length(bp)
62579	1838	28.33	11462	785.31

To evaluate the difference between the workflows of the current and ARK genomics approaches, the differential expression analysis using the coffee ESTs as base, was

used to take off the bias introduced by the different transcriptomes used. After submiting the coffee ESTs to the blast redundancy pipeline (used in our assembly to eliminate highly similar sequences), it was possible to identify a high level of redundancy. So, the use of the ESTs as reference in the two approaches allowed only comparisons of the mapping, quantification and differential expression steps and was not used for biological purposes.

The results of the current and ARK genomics differential expression analysis based on the ESTs are resumed in Table 7. A total of 2565 ESTs was identified as differentially expressed in the current analysis, while the ARK Genomics analysis identified 3634 ESTs.

Table 7- Statistics of differentially expressed ESTs of the current and ARK genomics analysis at the 3 sampled time-points of the control vs inoculated comparisons. Cut-off values of -1.0 ≥ log2 fold change ≥1.0 and PPDE > 0.95 for the current analysis and a p-value ≤ 0.05 for the previous analysis.

	Current analysis				**ARK genomics Analysis**			
	24h	**48h**	**72h**	**Total**	**24h**	**48h**	**72h**	**Total**
Up regulated								
Resistant	114	828	1376	1715	537	1317	2184	2388
Susceptible	25	575	670	927	492	1060	1353	1490
Down regulated								
Resistant	118	107	339	519	39	320	672	897
Susceptible	17	76	176	239	15	87	269	316

The differentially expressed ESTs of the two analysis were crossed. Figure 9 resumes the results of these comparisons using Venn's diagrams. A total of 1631 ESTs were shared by the two analyses at least at one of the comparisons. The number of shared differential expressed ESTs is consistent at 48 and 72 hpi in the susceptible and resistant comparisons. At 24 hpi, the number of shared ESTs is reduced or absent, in the susceptible and resistant comparisons, respectively.

Also, to evaluate the differences between the softwares used for differential expression analysis, all differentially expressed ESTs common to both analyses were considered. Then, the ESTs with more than 2 degrees of fold change of difference between analyses (the minimum difference between considering an up and down regulated EST) were selected. From this selection, only one EST had a contradictory

result (up or down regulated depending on the analysis), but several others had values of fold change highly dissimilar. To evaluate fold change calculation differences and its statistical support, the counts from the ARK genomics analysis were used with the package of differential expression used in the current analysis. From the 46 ESTs selected, only 17 were statistical validated by EBSeq, and the fold changes were very dissimilar (Table S1).

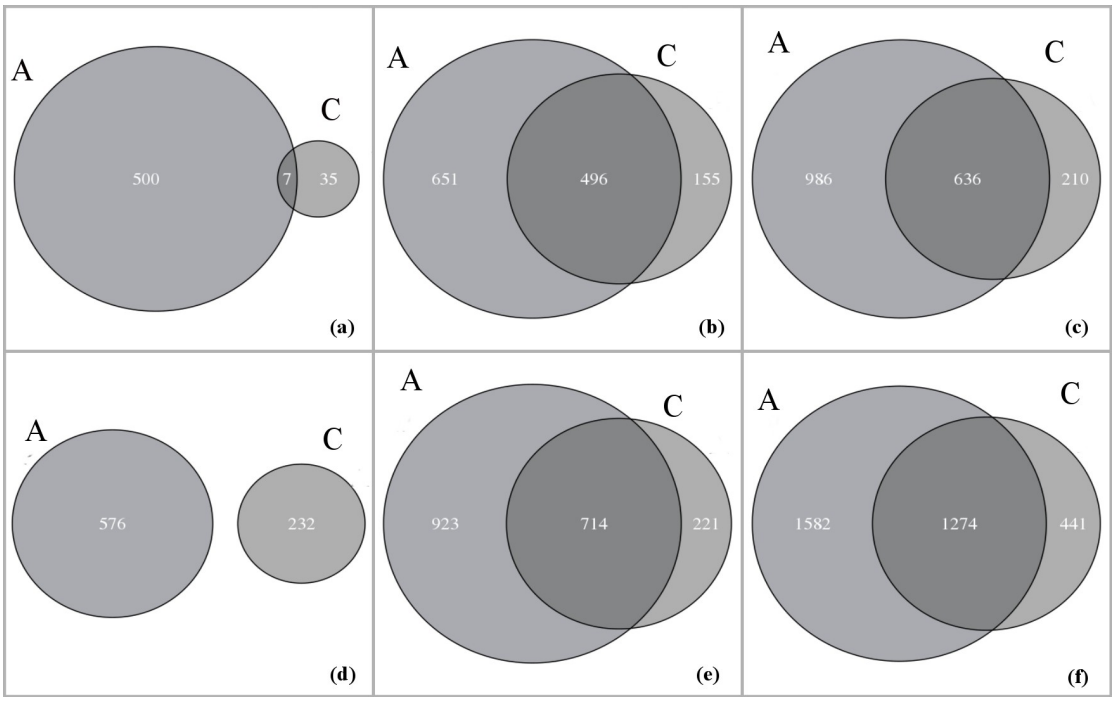


Figure 9 - Comparison of the results obtained from the ARK genomics analysis (A) and the current analysis (C), using as common base the EST sequences. The top line corresponds to Susceptible control vs inoculated at (a)24h (b)48h (c)72h. The bottom line corresponds to Resistant control vs inoculated at (d) 24h (e)48h (f)72h. This selection only considered ESTs with -1.0 ≥ log2 fold change ≥1.0. For the current analysis the cut-off was made by a PPDE > 0.95, and for the ARK genomics analysis a p-value ≤ 0.05.

4. Discussion

Despite all the previous studies regarding the defense mechanisms of Coffee to *C. kahawae,* there are very few insights on the molecular and genetic basis of coffee resistance. With the aim of unravelling these mechanisms, an RNA-seq approach was taken to study coffee resistance vs susceptible response in the early stages of *C. kahawae* infection. The data was analysed by the sequencing company ARK genomics, but to take the best revenue of the data, a new analysis, adjusted to data specifications, was made. In addition, the two approaches were compared taking into account not only the results but the entire workflow.. The results, including expression

profiles and unigene annotation, were analysed to unravel some of the biological bases possibly involved in the defense response.

4.1 – Workflow overview and comparison

High-throughput automated sequencing has enabled an exponential growth rate of sequencing data. This requires an increasing sequence quality and reliability in order to avoid database contamination with artefactual sequences (Falgueras *et al.* 2010). In this way, the preparation of NGS data before every analysis is crucial. Otherwise, the bad quality of the reads or the presence of contaminated reads, can compromise the downstream analysis, leading to inaccurate results (Seluja *et al.* 1999; Coker & Davies 2004). Therefore, in the analysis developed in our work, the presence of such artifacts were taken into account and properly cleaned.

However, in comparison, the ARK genomics analysis used reads only with quality based trimming. Contigs beginning with the same sequence of nucleotides – adapters – are an example of the lack of read proper pre-processing in the ARK Genomics transcriptome. The reads were also not subjected to a contaminant survey, which in a case like this, where we have libraries sequenced from plants purposely contaminated with fungus, is particularly important. Thus, one of the differences between transcriptomes relies on read cleaning. Besides that, differences between the parameters used, could also have influenced the differences between the transcriptomes obtained, since the two different softwares of assembly are based in the same algorithm – Bruijn Graphs – and so are discarded as a bias source (Schulz *et al.* 2012; http://soap.genomics.org.cn/SOAPdenovo-Trans.html last access April 11th 2014). Poorly optimized assembly parameters, can lead to less effective use of the data. The most important parameter in de Bruijn graph assemblers is the hash length, or k-mer length (Schulz *et al.* 2012). The perfect value for this parameter marks a trade-off between sensitivity and specificity: longer k-mers bring more specificity but lower coverage, while smaller k-mers allows locating more overlapping sequences (i.e. higher sensitivity) while increasing the number of ambiguous repeats (Zerbino & Birney 2008). In other words, an assembly with a small k-mer increases the probability of two reads assembly together which results in longer, but less contigs, whereas an assembly with a large k-mer decreases that same probability which

produce more, but shorter contigs. However, due to the complexity of the assembly process, the size and quantity of contigs are not directly proportional to the size of the k-mers used, and can drift according to the data.

When comparing the two assemblies, it is possible to speculate about the k-mer value used in the ARK genomics assembly (k-mer=21). This value seems to be a poor choice, because it does not result in more neither in bigger contigs than those obtained with the present Velvet/Oases assembly. Regardless, the present assemblies (both for transcriptome reconstruction and for sequence identification), might not also have the perfect k-mer value, due to our RAM restrictions. Another method usually applied in these cases is the multiple k-mer assembly (Surget-Groba & Montoya-Burgos 2010).This method consists in assembling the data with different k-mers and then merge them into one, non-redundant transcriptome (Surget-Groba & Montoya-Burgos 2010). This strategy is based on the theory that the assemblies with longer k-mer values perform best on high expression genes, but poorly on low expression genes (Surget-Groba & Montoya-Burgos 2010). So, merging the different assemblies would cover genes at different expression levels (Surget-Groba & Montoya-Burgos 2010; Schulz *et al.* 2012), as applied in the assembly of *Nicotiana benthamiana* (Nakasugi *et al.* 2013) or *Sphenodon punctatus* (Miller *et al.* 2012).

Currently, it is known that *de novo* transcriptome assemblies entail a certain level of redundancy, due to the assembly of different isoforms of the same gene or even potential sequence variations among individuals (Duan *et al.* 2012). Plus, since the plant sequences and those considered potentially plant, from the plant/fungus sequence identification step, were joined together to take the maximum information possible from the data, and the k-mer value was not ideal, the redundancy level of the transcriptome increased. In this case, it is usual to take a clustering step, which removes contigs with a high similarity, keeping the longer contigs. Since this step was not sufficient, leaving behind several redundant contigs, the redundancy blast pipeline was used. In the absence of a reference genome, it is practically impossible to separate redundancy introduced by the assembly and the presence of different isoforms. In this way, some of these redundancy steps could have compromised the actual composition and size of the transcriptome. Also, the selection of the longer contigs as the most representatives, could lead to the same conclusions. To overcome this problem, a

reference genome would had to be available, where we could map the reads back to it, identifying the real isoforms and excluding the redundancy.

The study of *Vitis vinifera* transcriptome, one of the closest species to *Coffea* spp. with a genome assembled, showed a *de novo* transcriptome assembly with only one k-mer (41), which resulted in 106670 contigs after a simple step of redundancy cleaning. After mapping the contigs against the reference genome, the reference transcriptome was reduced to 60075 contigs (Venturini *et al.* 2013). This confirms the common existence of redundant contigs in the assembly which were discarded with the utilization of the reference genome. Despite the absence of a reference genome for *Coffea* spp. to confirm the current results, the approximated statistical results of both coffee and grapevine transcriptomes suggests the good quality of the present assembly.

Due to the differences between the two transcriptome assemblies highlighted above, it is impossible to compare directly the results of differentially expressed transcripts. So, an analysis of differential expressed ESTs was made, to compare with the results of the ARK genomics analysis which used the same ESTs as reference.

In addition to differences related with software and parameters used, and probably also due to a lack of proper pre-processing in the ARK genomics analysis, as shown above, differences between the read mapping softwares could also have influenced the results. Comparisons between read alignment softwares showed little differences between BWA (used for mapping in the ARK genomics analysis) and Bowtie (used in the current analysis) making the differences of each one negligible (Bao *et al.* 2011; Ruffalo *et al.* 2011). In this case, what seems to make all the difference between the two mappings is the RSEM algorithm. RSEM takes into account the uncertainty of mapping, especially when we are dealing with small reads, which can be mapped in several places. The idea is to allow RSEM to decide which alignments are most likely to be correct, rather than giving the aligner this responsibility. RSEM receive all the possible alignments for each read from Bowtie and uses a model based on maximum likeliwood to calculate the probability of a read belonging to a certain transcript, giving a more accurate estimation of expression (Li & Dewey 2011). This could be especially important if we take into account the redundancy of the reference ESTs. The estimation of expression by RSEM may have dilute the expression of some

redundant transcripts, and so, influenced their counting and consequently the number of differential expressed ESTs (Hiller *et al.* 2009; Trapnell *et al.* 2012).

Besides that, the differential expression software could be involved in the differences detected between analyses. Both EdgeR, the software used in the ARK genomics analysis, and EBSeq used in the current analysis, assume a negative binomial distribution of the data. EdgeR was developed to enable analysis of experiments with small numbers of replicates applying an empirical Bayes procedure to moderate the degree of overdispersion across genes (Robinson *et al.* 2009), whereas EBSeq estimates the posterior likelihoods of differential and equal expression by the aid of empirical Bayesian methods (Leng *et al.* 2013). According to Seyednasrollah *et al.* 2013, the different methods of normalization and parameters used in different software have little or no influence in the final results. Whereby, the influence of such topics will be discarded.

Comparing the results of the commonly assign differentially expressed ESTs it was possible to conclude that EdgeR is somehow liberal on DE calling, corroborating the results of other studies (Soneson & Delorenzi 2013; Seyednasrollah *et al.* 2013). EBSeq seems to be more conservative as it is possible to see when the counts from the Ark genomics analysis are used with EBSeq, where only 17 of the 46 ESTs tested, passed the false discovery rate cutoff (PPDE > 0.95). This difference may be related with EdgeR problem with outliers. In the presence of outliers (i.e. values extremely high or low in both replicates or just one replicate) EdgeR become much more liberal, both in fold change and p-value calculation (Soneson & Delorenzi 2013). On the other hand, EBSeq have a more restrictive DE calling, independent of the number of replicates (as much as possible), and unaffected by outliers (Soneson & Delorenzi 2013). Besides that, EBSeq is much more user friendly, especially when used together with RSEM.

In the study of a pathosystem, such as Coffee – *C. kahawae*, both the pathogen and the host are of extreme importance. In this way, the use of deep-sequencing to study the pathosystem can retrieve data ffrom both players of the interaction. In the ARK genomics analysis, the genetic information of the fungus was not recovered, unlike the current analysis.

4.2. Plant-Fungus Separation

Previous studies showed the successful separation of genetic information from a plant and pathogen when an infected host is sequenced (Sebastiana *et al.* 2009; Fernandez *et al.* 2012; Zhuang *et al.* 2012)

Two methods were used for plant-fungus contig identification, using the Plant-fungus transcriptome: MIPS-EST3 and a blast pipeline. Based in concordant results provided by both methods , 198036 plant contigs and 653 fungus contigs were identified, however 2208 contigs remained unclassified as both methods gave contradictory results. The blast search of this pipeline was made using a reduced number of *C. kahawae* and other *Colletotrichum* species sequences. Thus, it is possible that several fungus sequences did not get a hit due to the reduced database information available. The same may have occurred relatively to the plant sequences. Some of the contigs assembled in the plant-fungus transcriptome may not exist in the control transcriptome, and so, no homologies were found. On the other hand, EST3 method does not account for the existence of other contaminants beyond the *C. kahawae* itself. As the assembly for this pipeline was made with non cleaned reads (to make sure that we were not losing any of the fungus genetic information for the assembly, which could lead to poor separation of sequences due to missassemblies of the fungal data) the presence of other commune contaminants is possible. Thus, as EST3 can only classify sequences as fungus or plant, some of its classifications can be incorrect or missing. The previous arguments reinforce the necessity of using the two methods, and the two concordant results to clearly identify sequence origin.

4.3. Differential expression analysis

Great differences between the susceptible and resistant genotypes responses to the fungal infection were identified, both in number of differentially expressed unigenes and the functional categories to which they belong.

Comparisons between time-points and control vs inoculated showed more differentially expressed unigenes in the resistant genotype, relatively to the susceptible genotype. Such is admissible considering a probable higher number of processes that must be activated by the resistant genotype to stop the fungal growth.

These results are corroborated by other studies with plant-pathogen interactions, such as *Pinus monticola* - *Cronartium ribicola* (Liu *et al.* 2013), where 562 differentially expressed genes were found in the compatible interaction and 789 in the incompatible interaction.

Between control and inoculated samples, the absence of unigenes expressed only at 48 hpi, in both genotypes suggests the continuity of the response to the fungus infection, since the differential expressed unigenes at 48hpi were already activated at 24hpi or were still activated at 72hpi. The distribution of the unigenes by the profiles, corroborates this idea of continuity, since in both genotypes, the most representative of them included the time-points 48hpi and 72hpi.

Also, the predominance of differently expressed unigenes at 72 hpi, and the high representativeness of the j) profile (increases from 48hpi to 72hpi) may correspond to the switch from the biotrophic to the necrotrophic phase in the susceptible genotype, and to the accumulation of phenols and display of HR in 50% of infection sites in the resistant genotype (Loureiro *et al.* 2012a). This stimulates a more intense response for both resistant and susceptible genotypes, which explains the higher activation at 72hpi.

The characterization of the differentially expressed unigenes by KOG and KEGG annotations, made by time-point and genotype or by expression profile, revealed relevant categories related to the plant-fungus interaction.

The KOG annotation in all comparisons presented in this study, may be considered incomplete due to the great percentage of annotations being inconclusive (other functions, function unknown and general function prediction only). This is a common fact occuring in several other studies that include plant sequence annotation, such as the transcriptome study of *Youngia japonica (Peng et al. 2014)*, or the transcriptomic analysis of Paulownia infected by Paulownia witches'-broom *Phytoplasma (Mou et al.* 2013*),* reflecting the yet low coverage of gene databases regardless of the high advances on knowledge obtained in the last few years.

The categories “Signal transduction mechanisms” and “Post-translational modification, protein turnover, chaperones” are consistently the most representative of the differentially expressed unigenes between control and inoculated samples. These two categories are indicative that coffee is transcriptionally very active during the

infection of *C. kahawae*, triggering several signalling mechanisms and increasing protein biosynthesis. Other study on the interaction of coffee with leaf rust (*Hemileia vastatrix*), showed in the KOG annotation, the same categories as the most representative (Fernandez *et al.* 2012) . This may suggest that these two categories are related with a general defense response of coffee to pathogen infection, since infection takes place in different plant tissues. .

On the other hand, the annotation of the different profiles, unraveled differences on these same categories. For the resistant genotype, the "Post-translational modification, protein turnover, chaperones" and "Signal transduction mechanisms" categories are represented in stable and increased along time profiles, while in the susceptible genotype, these only appear in the stable profiles. This may indicate a different triggering and activation pattern of the infection response by the resistant genotype.

The KEGG annotation sustain this idea, since the phenylalanine and phenylpropanoid pathways are mainly represented in the resistant genotype (control vs inoculated), and covers all the sampled time-points. Also, stable and "expression increasing" profiles show a predominance of these pathways in the resistant genotype, while in the susceptible genotype, all the profile categories are related with plant growth and development . Phenylalanine and phenylpropanoid pathways are known to be related with the defense response in different pathosystems, namely in tobacco - *Phytophthora megasperma*, or coffee - *Hemileia vastatrix,* since enzymes like phenylalanine ammonia-lyase(PAL) are activated in the early stages of these interactions, (Dixon & Paiva 1995; Dorey *et al.* 1997; Silva *et al.* 2002). PAL was already identified as having an important role in resistance, being involved in the production of several compounds associated with fungal invasion, like phenylpropanoids and suberin or other phenolic compounds normally deposited in the host cell walls at the point of fungal invasion (Silva *et al.* 2006). PAL was also implicated in the production of salicylic acid, another defense-related compound (Mauch-Mani & Slusarenko 1996; Silva *et al.* 2006) In coffee, studies revealed an early increase of PAL and SA, associated with the resistance response to *H.vastatrix* (Silva *et al.* 2002, Sa *et al.* 2014).

"Amino sugar and nucleotide sugar metabolism "and "starch and sucrose metabolism" are well known for being directly related with biosynthesis and metabolism of

nucleotides of fundamental importance in plant growth and development (Winter & Huber 2000; Stasolla *et al.* 2003). The representativeness of these categories are justified by the model used in this study (hypocotyls), which is a structure in constant growth and development.

Using both inoculated genotypes as a pair of conditions for differential expression analysis allowed the comparisons of level of expression between resistant and susceptible inoculated libraries, giving us the idea of the unigenes more expressed in each genotype instead of the unigenes that are down or upregulated relatively to the control.

Since in this step the comparison was made between two different genotypes, an exhaustive comparison between them could lead to false conclusions, since the differences of expression could not be related to the plant reaction to the fungus infection but rather to genotype-specific responses. By selecting the categories related with defense, we intended to focus on the conditions that were varied (pathogen infection), minimizing the differences derived from other causes.

Analyzing the patterns of the defense-related unigenes, it was possible to see a higher and early response to the infection by the susceptible genotype. Also, the categories related with reactive oxygen species (Van Breusegem & Dat 2006) and response to stress indicates that the cell at 48 hpi is still reacting , while the resistant genotype only reacts in a similar way at 72hpi Vargas *et al.* (2012) showed that in the interaction of the hemibiotrophic fungus *Colletotricum graminicola* with maize, a strong induction of defense mechanisms occurs at early stages of infection. Moreover, these authors hypothesized that the switch to necrotrophic growth (occurring in our pathosystem at 72 hpi) enables the fungus to evade the effects of the plant immune system and allows for full fungal pathogenicity (Vargas *et al.* 2012). Allied to this, a KEGG annotation of the same unigenes was made, and the pathways of phenylalanine and phenilpropanoid were only identified in the resistant genotype, showing that these compounds may have a main role in resistance. Previous studies on transcriptional responses to downy mildew infection in a susceptible (*Vitis vinifera*) and a resistant (*V. riparia*) grapevine species showed a similar profile, where about 75% of the transcripts were more strongly expressed in *V. vinifera* and about 25% were more strongly expressed in *V. riparia*. Also, the resistant genotype showed to have more

specific transcripts that were absent in the susceptible genotype (Polesani *et al.* 2010).

5. Conclusions

RNA-seq data analysis for differential expression is still a technique in development, and so, there is no ideal method or tool for all the analyses. The lack of a reference genome, make the task even more difficult: transcriptome assembly entails several problems, such as redundancy and the need of high computer power. In the mapping phase, the use of short reads can lead to miss-mappings if measures are not taken into account. Additionally, several software were developed for differential expression analysis, with more restrictive or more liberal parameters for differential expression calling. However, there is no clear consensus about the best practices yet: it is up to the user to choose the software more adapted to their data and purpose. The comparison of the current and ARK genomics analysis showed, not only the importance of an appropriate data treatment but also the differences between the results when taking different approaches. The use of appropriate parameters and software and the control between multiple phases is of the utmost importance in a proper bioinformatics analysis. Relatively to the biological results, new insights were provided relatively to the differences of expression associated to coffee susceptibility and resistance responses. The resistant genotype showed a more intense response of gene expression to the infection. A peak of expression at 72hpi was identified for both genotypes, possibly related with the the switch from the biotrophic to necrotrophic phase in the susceptible genotype and the accumulation of phenols and display of HR in 50% of infection sites in the resistant genotype. Annotation of diffentially expressed transcripts showed a high biological transcriptomic activity of both genotypes and functional categories already identified as related with defense response , such as phenylalanine and phenilpropanoid biosynthesis which were only identified in the resistant response to *C. kahawae.* Also, the resistant genotype showed a higher effective response to infection in all the time points and the susceptible genotype an early stress-leaded response to the infection. Through this study, the first steps were taken into the better understanding of coffee resistance to C. kahawae, potentially applicable to similar pathosystems. Further analysis is needed to obtain more biochemical and metabolic information from the results.

References

Bao S, Jiang R, Kwan W, Wang B, Ma X, Song YQ (2011) Evaluation of next-generation sequencing software in mapping and assembly. *Journal of human genetics*, **56**, 406–414.

Boetzer M, Henkel C V, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, **27**, 578–579.

Calduch-Giner J a, Bermejo-Nogales A, Benedito-Palos L, Estensoro I, Ballester-Lozano G, Sitjá-Bobadilla A, Pérez-Sánchez J (2013) Deep sequencing for *de novo* construction of a marine fish (*Sparus aurata*) transcriptome database with a large coverage of protein-coding transcripts. *BMC genomics*, **14**.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC bioinformatics*, **10**, 421.

Charrier A, Berthaud J (1985) Botanical Classification of Coffee. In: *Coffee* (eds Clifford MN, Willson KC), pp. 13–47. Springer US, Boston, MA.

Coker JS, Davies E (2004) Identifying adaptor contamination when mining DNA sequence data. *BioTechniques*, **37**, 194-198.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

Dixon RA, Paiva NL (1995) Stress-Induced Phenylpropanoid Metabolism. *The Plant cell*, **7**, 1085–1097.

Dorey S, Baillieul F, Pierrel M-A Saindrenan P, Fritig B, Kauffmann S (1997) Spatial and temporal induction of cell death, defense genes, and accumulation of salicylic acid in tobacco leaves reacting hypersensitively to a fungal glycoprotein. *Molecular plant-microbe interactions*, **10**, 646–655.

Duan J, Xia C, Zhao G, Jia J, Kong X (2012) Optimizing *de novo* common wheat transcriptome assembly using short-read RNA-Seq data. *BMC genomics*, **13**, 392.

Emmersen J, Rudd S, Mewes H-W, Tetko IV (2007) Separation of sequences from host-pathogen interface using triplet nucleotide frequencies. *Fungal genetics and biology : FG & B*, **44**, 231–241.

Falgueras J, Lara AJ, Fernández-Pozo N, Cantón FR, Pérez-Trabado G, Claros MG (2010) SeqTrim: a high-throughput pipeline for pre-processing any type of sequence read. *BMC bioinformatics*, **11**, 38.

Fernandez D, Tisserant E, Talhinas P, Azinheira H, Vieira A, Petitot A-S, Poulain J, Da Silva C, Silva MC, Duplessis S (2012) 454-pyrosequencing of *Coffea arabica* leaves infected by the rust fungus *Hemileia vastatrix* reveals in planta-expressed pathogen-secreted proteins and plant functions. *Molecular plant pathology*, **13**, 17–37.

Figueiredo A, Loureiro A, Batista D, Monteiro F, Várzea V, Pais MS, Gichuru EK, Silva MC (2013) Validation of reference genes for normalization of qPCR gene expression data from *Coffea spp.* hypocotyls inoculated with *Colletotrichum kahawae*. *BMC research notes*, **6**.

Gichuru EK (1997) Resistance mechanisms in Arabica coffee to coffee berry disease (*Colletotrichum kahawae Sp. Nov.*); a review. *Kenya Coffee (Kenia)*, **67**, 2441–2444.

Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou Y-C, Pugh TJ, Robertson G, Chittaranjan S, Ally A, Asano JK, Chan SY, Li HI, McDonald H, Teaguel K, Zhao Y, Zeng T, Delaney A, Hirst M, Morin GB, Jones SJM, Tai IT, Marra MA (2010) Alternative expression analysis by RNA sequencing. *Nature methods*, **7**, 843-847.

Hiller D, Jiang H, Xu W, Wong WH (2009) Identifiability of isoform deconvolution from junction arrays and RNA-Seq. *Bioinformatics*, **25**, 3056–3059.

Van Hilten H, Fisher P, Wheeler M, Wagner B (2011) *The Coffee Exporter's Guide*. International Trade Centre UNCTAD/GATT, Geneva.

Hindorf H, Omondi CO (2011) A review of three major fungal diseases of *Coffea arabica* L. in the rainforests of Ethiopia and progress in breeding for resistance in Kenya. *Journal of Advanced Research*, **2**, 109–120.

Kanehisa M, Araki M, Goto S, Hattoril M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokinatsu T, Yamanishi Y (2008) KEGG for linking genomes to life and the environment. *Nucleic acids research*, **36**, 480–484.

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

Lashermes P, Anthony F (2007) Coffee. In: *Genome Mapping and Molecular Breeding Plants, Technical Crops* (ed Kole C), pp. 108–118. Springer-Verlag, Berlin.

Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, Haag JD, Gould MN, Stewart RM, Kendziorski C (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, **29**, 1035–1043.

Li B, Dewey C (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, **12**.

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

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Liu L, Li Y, Li S, Hu N,, He Y, Pong R, Lin D, Lu L, Law M (2012) Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology*, **2012**.

Liu J-J, Sturrock RN, Benton R (2013) Transcriptome analysis of *Pinus monticola* primary needles by RNA-seq provides novel insight into host resistance to *Cronartium ribicola*. *BMC genomics*, **14**, 884.

Loureiro A, Figueiredo A, Batista D, Baraldi T, Várzea V, Azinheira HG, Talhinhas P, Pais MS, Gichuru EK, Silva MC (2012a) New cytological and molecular data on coffee – *Colletotrichum kahawae* interactions. In: *Proceedings of the 24th International Conference on Coffee Science (ASIC)* . Costa Rica.

Loureiro A, Nicole MR, Várzea V, Moncadac P, Bertrandb B, Silva MC (2012b) Coffee resistance to *Colletotrichum kahawae* is associated with lignification, accumulation of phenols and cell death at infection sites. *Physiological and Molecular Plant Pathology*, **77**, 23–32.

Masaba DM, van der Vossen HAM (1992) Coffee berry disease: the current status. In: *Colletotrichum: Biology, Pathology and Control* (eds Bailey JA,Jeger MJ), pp. 237–249.

Mauch-Mani B, Slusarenko AJ (1996) Production of Salicylic Acid Precursors Is a Major Function of Phenylalanine Ammonia-Lyase in the Resistance of Arabidopsis to *Peronospora parasitica*. *The Plant cell*, **8**, 203–212.

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

Miller H, Biggs P, Voelckel C, Nelson N (2012) *de novo* sequence assembly and characterisation of a partial transcriptome for an evolutionarily distinct reptile, the tuatara (*Sphenodon punctatus*). *BMC genomics*, **13**.

Mou H-Q, Lu J, Zhu S-F, Lin C-L, Tian G-Z, Xu X, Zhao W-J (2013) Transcriptomic analysis of *Paulownia* infected by Paulownia witches'-broom *Phytoplasma*. *PloS One*, **8**.

Muñoz L (2010) Perspectivas de la caficultura colombiana., In: *III Conferencia Mundial del Café*, Ciudad de Guatemala

Nagalakshmi U, Waern K, Snyder M (2010) RNA-Seq: a method for comprehensive transcriptome analysis. *Current protocols in molecular biology*, **4**.

Nakasugi K, Crowhurst RN, Bally J, Wood CC, Hellens RP, Waterhouse PM (2013) *de novo* transcriptome sequence assembly and analysis of RNA silencing genes of *Nicotiana benthamiana*. *PloS one*, **8**.

Peatman E, Li C, Peterson BC, Straus DL, Farmerc BD, Beck BH (2013) Basal polarization of the mucosal compartment in *Flavobacterium columnare* susceptible and resistant channel catfish (*Ictalurus punctatus*). *Molecular immunology*, **56**, 317–27.

Peng Y, Gao X, Li R, Cao G (2014) Transcriptome sequencing and *de novo* analysis of *Youngia japonica* using the illumina platform. (CA Ouzounis, Ed,). *PloS one*, **9**.

Polesani M, Bortesi L, Ferrarini A, Zamboni A, Fasoli M, Zadra C, Lovato A, Pezzotti M, Delledonne M, Polverari A (2010) General and species-specific transcriptional responses to downy mildew infection in a susceptible (*Vitis vinifera*) and a resistant (*V. riparia*) grapevine species. *BMC genomics*, **11**.

Robinson MD, McCarthy DJ, Smyth GK (2009) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* **26**, 139-140.

Rodrigues CM, de Souza AA, Takita MA, Kishi LT, Machado MA (2013) RNA-Seq analysis of *Citrus reticulata* in the early stages of *Xylella fastidiosa* infection reveals auxin-related genes as a defense response. *BMC genomics*, **14**.

Ruffalo M, LaFramboise T, Koyutürk M (2011) Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics (Oxford, England)*, **27**, 2790–2796.

Sá M, Ferreira JP, Queiroz VT, Villas-Boas L, Silva MC, Almeida MH, Guerra-Guimarães L, Bronze MR (2014). A liquid chromatography-electrospray tandem mass spectrometry method for the simultaneous quantification of salicylic, jasmonic, and abscisic acid in *Coffea arabica* leaves". *Journal of the Science of Food and Agriculture* **94**, 529-536

Santos D (2011) Comparative analysis of 454 pyrosequencing data from coffee transcriptomes.

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–1092.

Sebastiana M, Figueiredo A, Acioli B, Sousa L, Pessoa F, Baldé A, Pais MS (2009) Identification of plant genes involved on the initial contact between ectomycorrhizal symbionts (*Castanea sativa* – European chestnut and *Pisolithus tinctorius*). *European Journal of Soil Biology*, **45**, 275–282.

Seluja GA, Farmer A, McLeod M, Harger C, Schad PA (1999) Establishing a method of vector contamination identification in database sequences. *Bioinformatics*, **15**, 106–10.

Seyednasrollah F, Laiho A, Elo LL (2013) Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in bioinformatics*.

Silva MC, Nicole M, Guerra-Guimarães L, Rodrigues Jr. CJ. (2002). Hypersensitive cell death and post-haustorial defense responses arrest the orange rust (*Hemileia vastatrix*) growth in resistant coffee leaves. *Physiological and Molecular Plant Pathology* **60**, 169-183.

Silva MC, Várzea V, Azinheira HG, Fernandez D, Petitot A-S, Bertrand B, Lashermes P, Nicole M (2006) Coffee resistance to the main diseases: leaf rust and coffee berry disease. *Brazilian Journal of Plant Physiology*, **18**, 119–147.

Soneson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics*, **14**.

Stasolla C, Katahira R, Thorpe T a, Ashihara H (2003) Purine and pyrimidine nucleotide metabolism in higher plants. *Journal of Plant Physiology*, **160**, 1271–95.

Surget-Groba Y, Montoya-Burgos JI (2010) Optimization of *de novo* transcriptome assembly from next-generation sequencing data. *Genome research*, **20**, 1432–1440.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf Y, Yin JJ, Natale DA (2003) The COG database: an updated version includes eukaryotes. *BMC bioinformatics*, **4**.

Trapnell C, Hendrickson DG, Sauvageau M, Goff , Rinn JL, Pachter L (2012) Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology*, **31**, 46–53.

Venturini L, Ferrarini A, Zenoni S, Tornielli GB, Fasoli M, Dal Santo S, Minio A, Buson G, Tononi P, Zago ED, Zamperin G, Bellin D, Pezzotti M, Delledonne M (2013) *de novo* transcriptome characterization of *Vitis vinifera* cv. Corvina unveils varietal diversity. *BMC genomics*, **14**.

Vieira LGE, Andrade AC, Colombo CA, Carazolle MF, Pereira GAG (2006) Brazilian coffee genome project: an EST-based genomic resource. *Brazilian Journal of Plant Physiology*, **18**, 95–108.

Van der Vossen H, Cook R, Murakaru G (1976) Breeding for resistance to coffee berry disease caused by *Colletrotrichum coffeanum* Noack (Sensu Hindorf) in *Coffea arabica* LI Methods of preselection for resistance. *Euphytica*, **25**, 733–745.

Van der Vossen H, Walyaro DJ (1980) Breeding for resistance to coffee berry disease in *Coffea arabica* L. II. Inheritance of the resistance. *Euphytica*, **29**, 777–791.

Van der Vossen H, Walyaro DJ (2009) Additional evidence for oligogenic inheritance of durable host resistance to coffee berry disease (*Colletotrichum kahawae*) in Arabica coffee (*Coffea arabica* L.). *Euphytica*, **165**, 105–111.

Vargas WA, Sanz Martín JM, Rech GE, Rivera LP, Benito EP, Díaz-Mínguez JM, Thon MR, Sukno SA (2012) Plant defense mechanisms are activated during biotropic and necrotrophic development of *Colletotrichum grainicola* in maize. Plant Physiology, **158**, 1342-1358.

Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, Haddock SHD, Huff KD, Mitchell IM, Plumbley MD, Waugh B, White EP, Wilson P (2014) Best Practices for Scientific Computing. *PLoS Biology*, **12**.

Winter H, Huber SC (2000) Regulation of sucrose metabolism in higher plants: localization and regulation of activity of key enzymes. *Critical Reviews in Plant Sciences*, **19**, 31–67.

Yang Y, Smith SA (2013) Optimizing *de novo* assembly of short-read RNA-seq data for phylogenomics. *BMC genomics*, **14**.

Yazawa T, Kawahigashi H, Matsumoto T, Mizuno H (2013) Simultaneous transcriptome analysis of Sorghum and *Bipolaris sorghicola* by using RNA-seq in combination with *de novo* transcriptome assembly. *PloS one*, **8.**

Zerbino DR, Birney E (2008) Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

Zhao Y, Tang H, Ye Y (2012) RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics*, **28**, 125–126.

Supplementary material

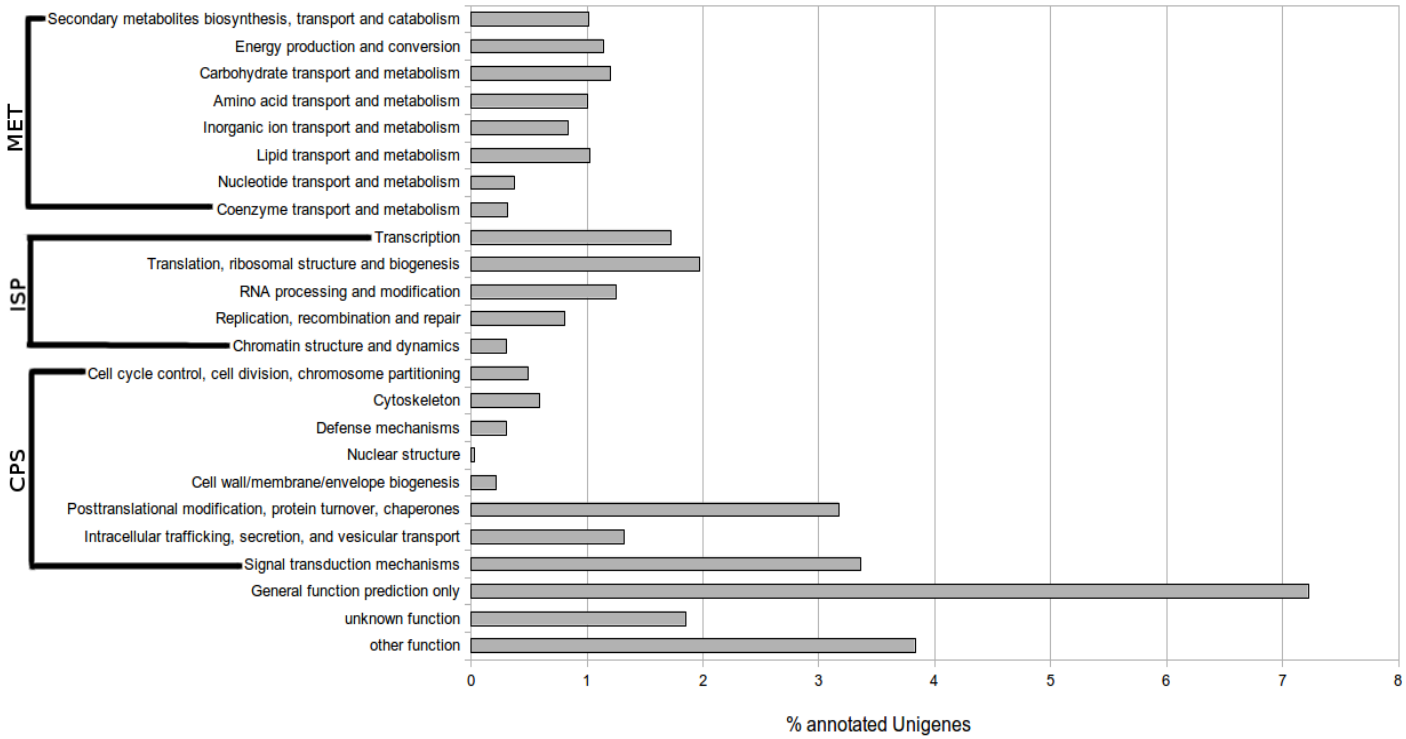


Figure S1 - KOG annotation of the transcriptome. The representativity of each term is showed by the percentage of the total transcriptome. The terms are divided in three major categories: CPS – Cellular process and signalling; ISP – Information storage and processing; MET - metabolism

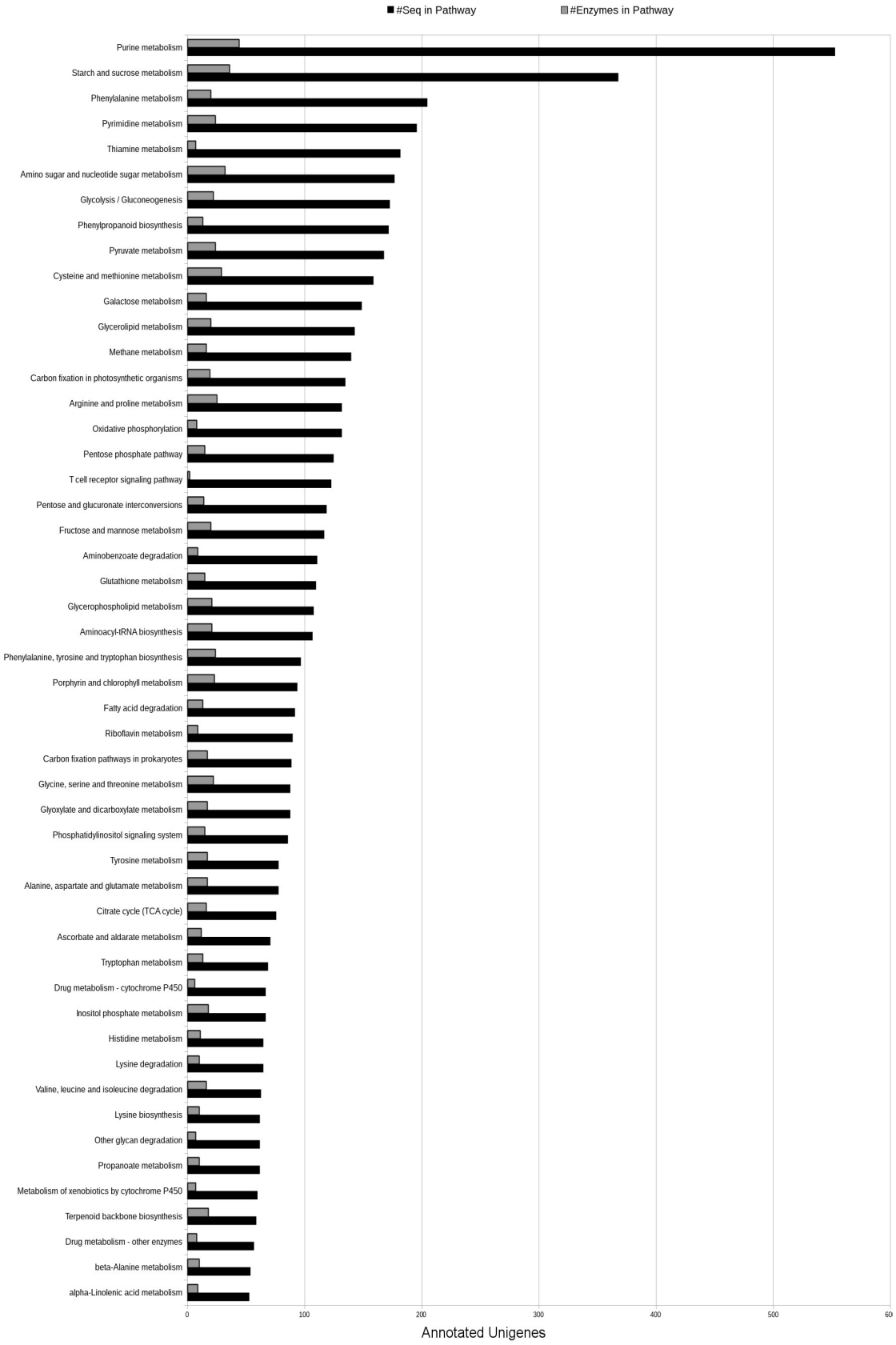
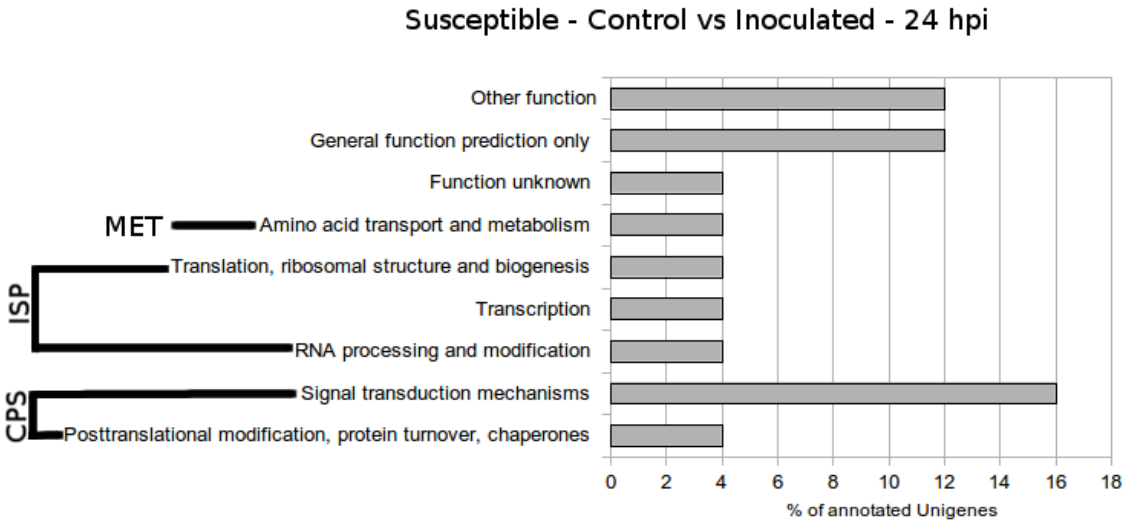


Figure S2 – Top 50 KEGG annotation of the transcriptome.



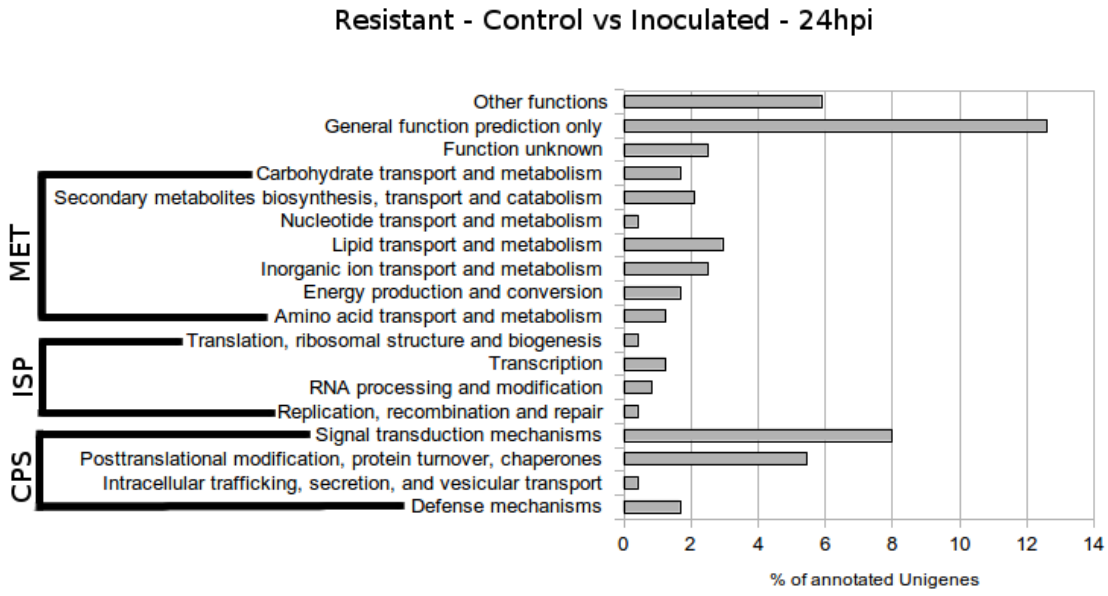
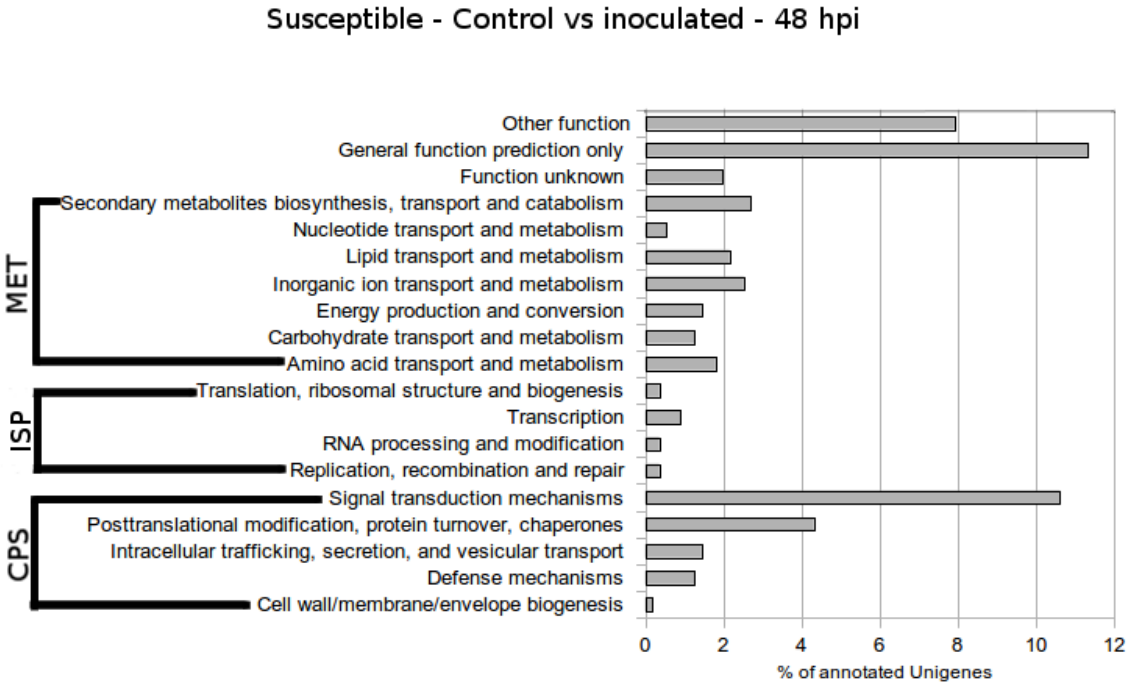


Figure S3 - KOG annotation of the differentially expressed unigenes of the comparisons between control and inoculated at 24hpi for both genotypes . The representativity of each term is showed by the percentage of the total transcriptome. The terms are divided in three major categories: CPS – Cellular process and signalling; ISP – Information storage and processing; MET - Metabolism



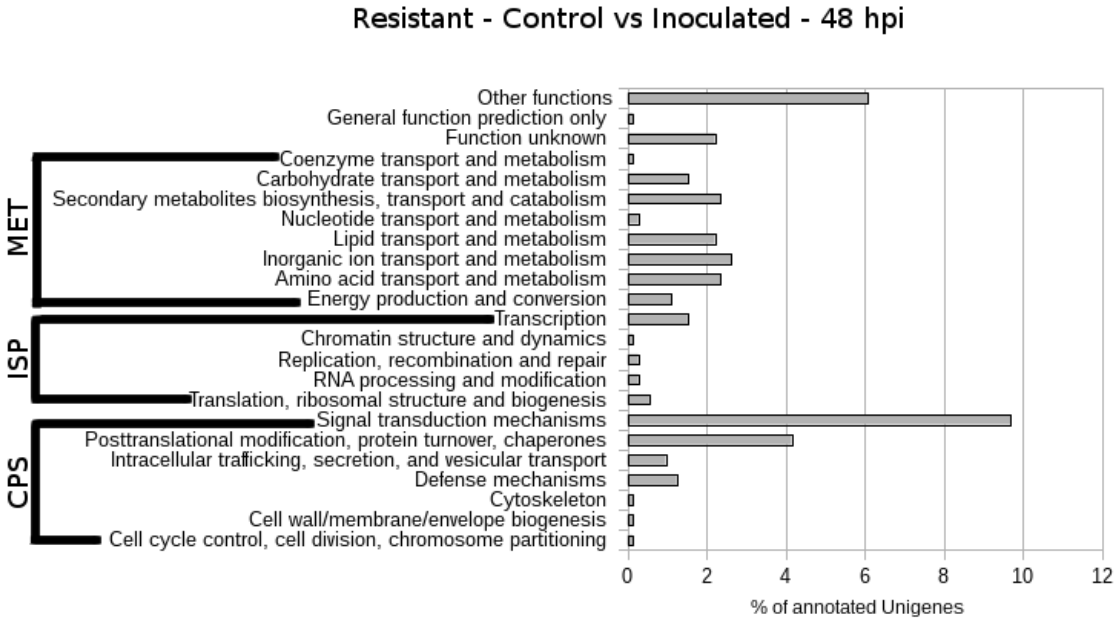
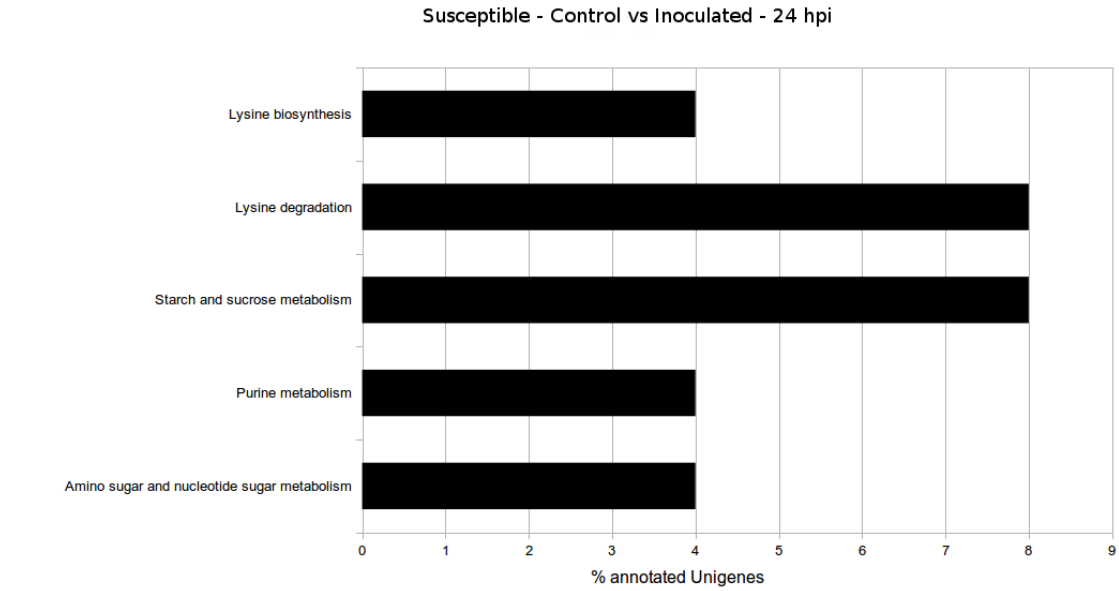


Figure S4 - KOG annotation of the differentially expressed unigenes of the comparisons between control and inoculated at 24hpi for both genotypes . The representativity of each term is showed by the percentage of the total transcriptome. The terms are divided in three major categories: CPS – Cellular process and signalling; ISP – Information storage and processing; MET - Metabolism



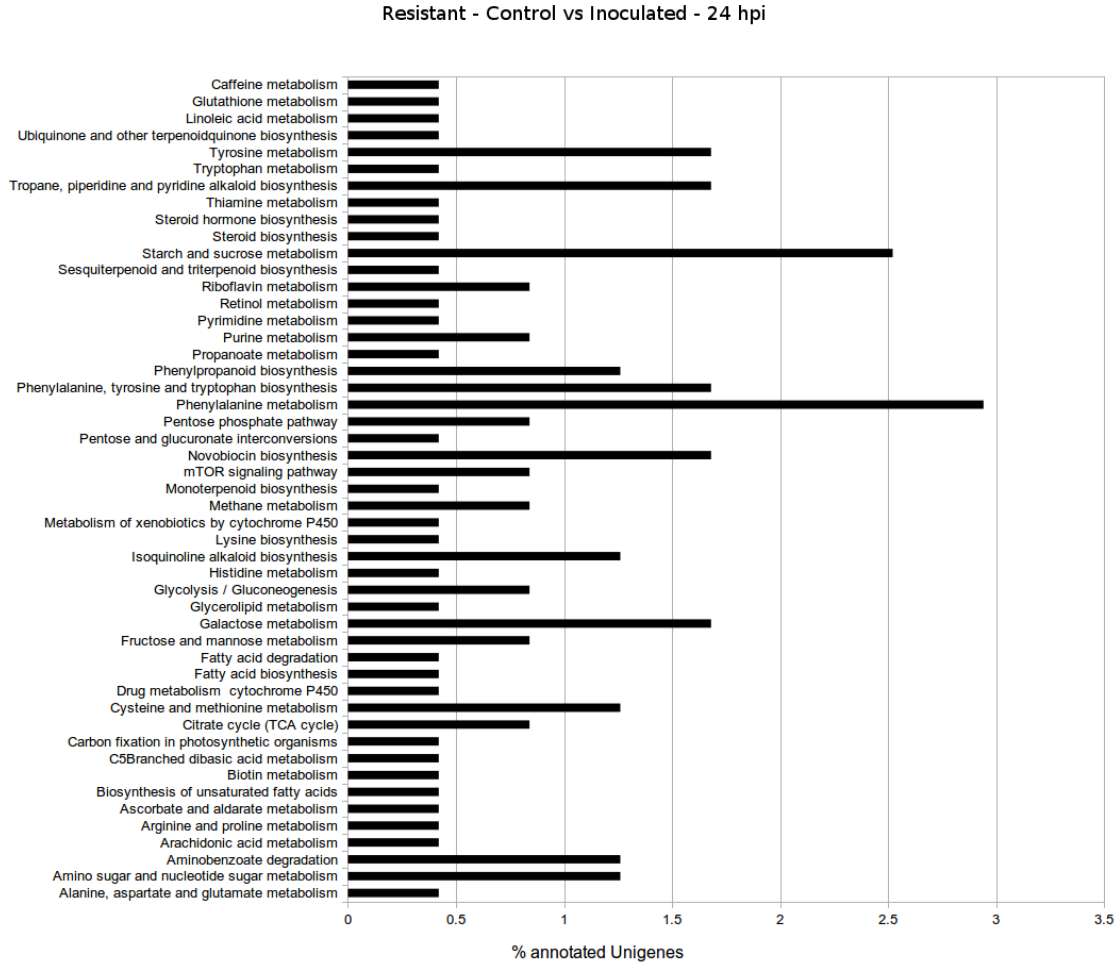
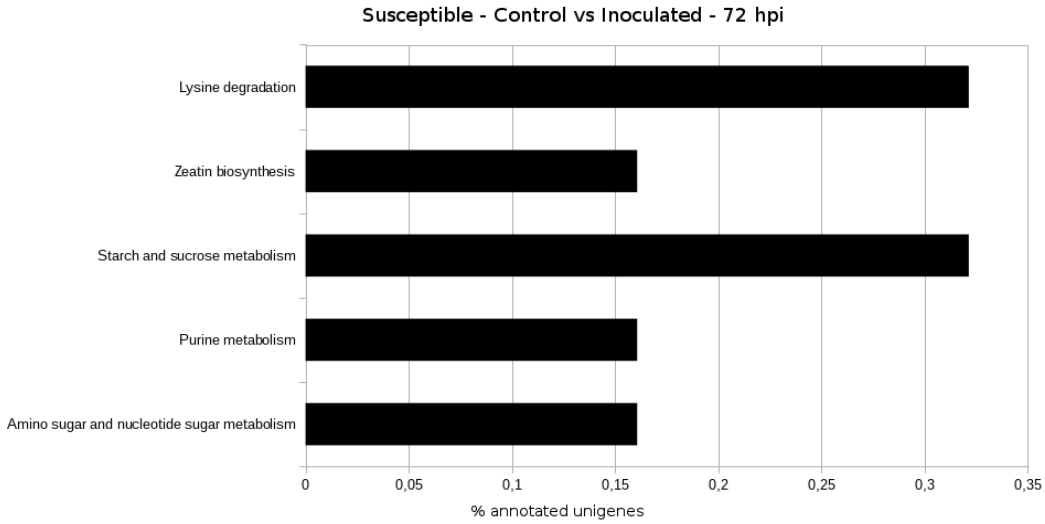


Figure S5 - KEGG annotation of the differentially expressed unigenes of the comparisons between control and inoculated at 24hpi for both genotypes . The representativity of each term is showed by the percentage of the total transcriptome.



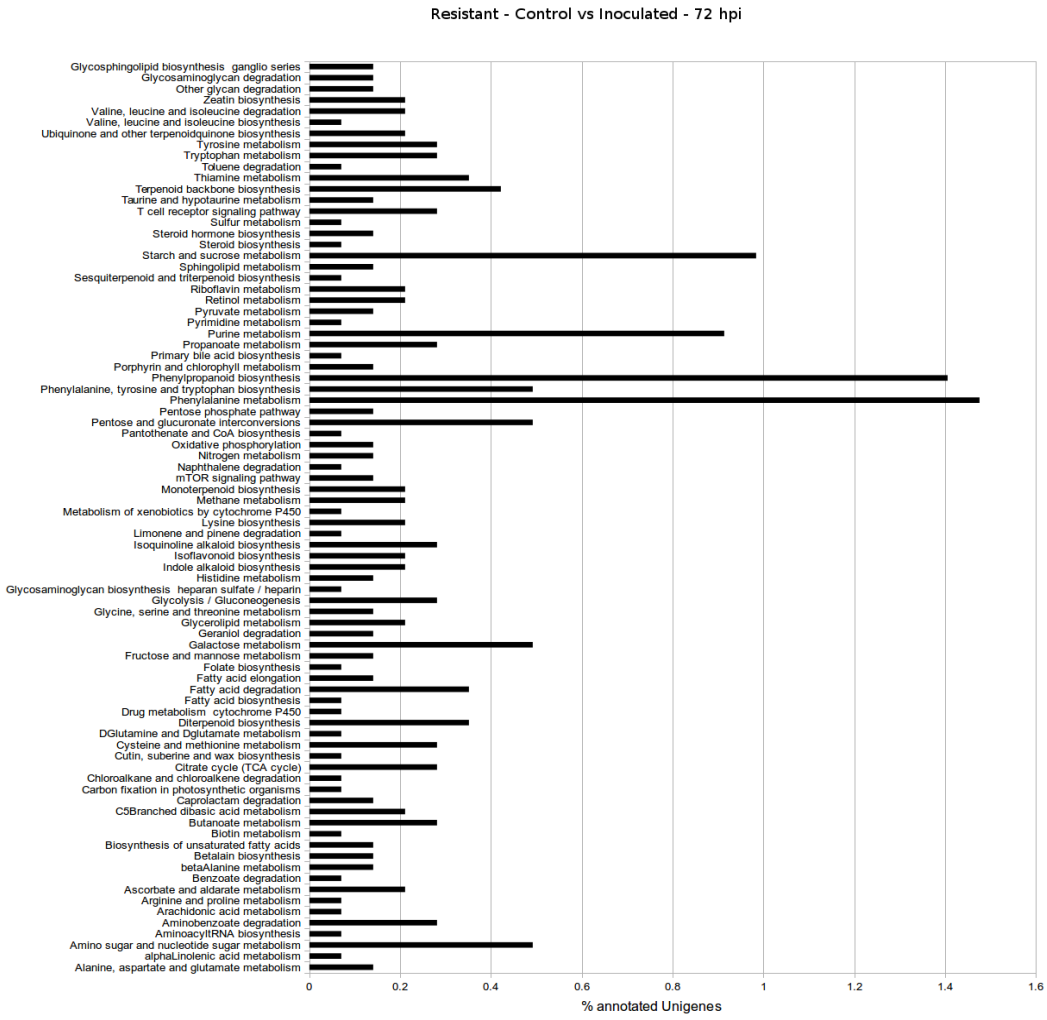


Figure S6 - KEGG annotation of the differentially expressed unigenes of the comparisons between control and inoculated at 72hpi for both genotypes . The representativity of each term is showed by the percentage of the total transcriptome.

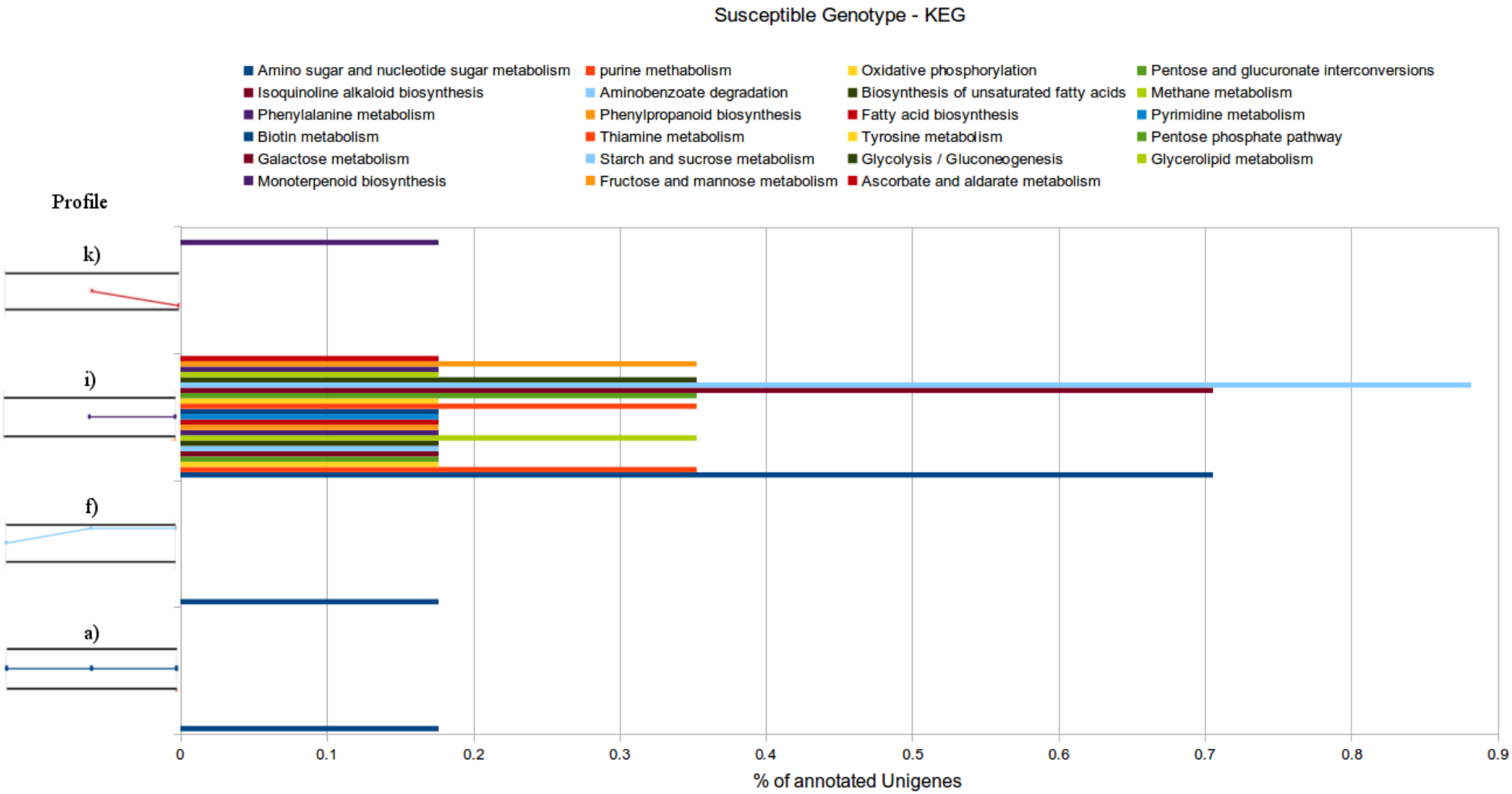


Figure S7a – KEGG annotation per profile of the susceptible genotype. The percentage of unigenes is relatively to the total of differentially expressed unigenes of the susceptible comparisons – Control vs Inoculated

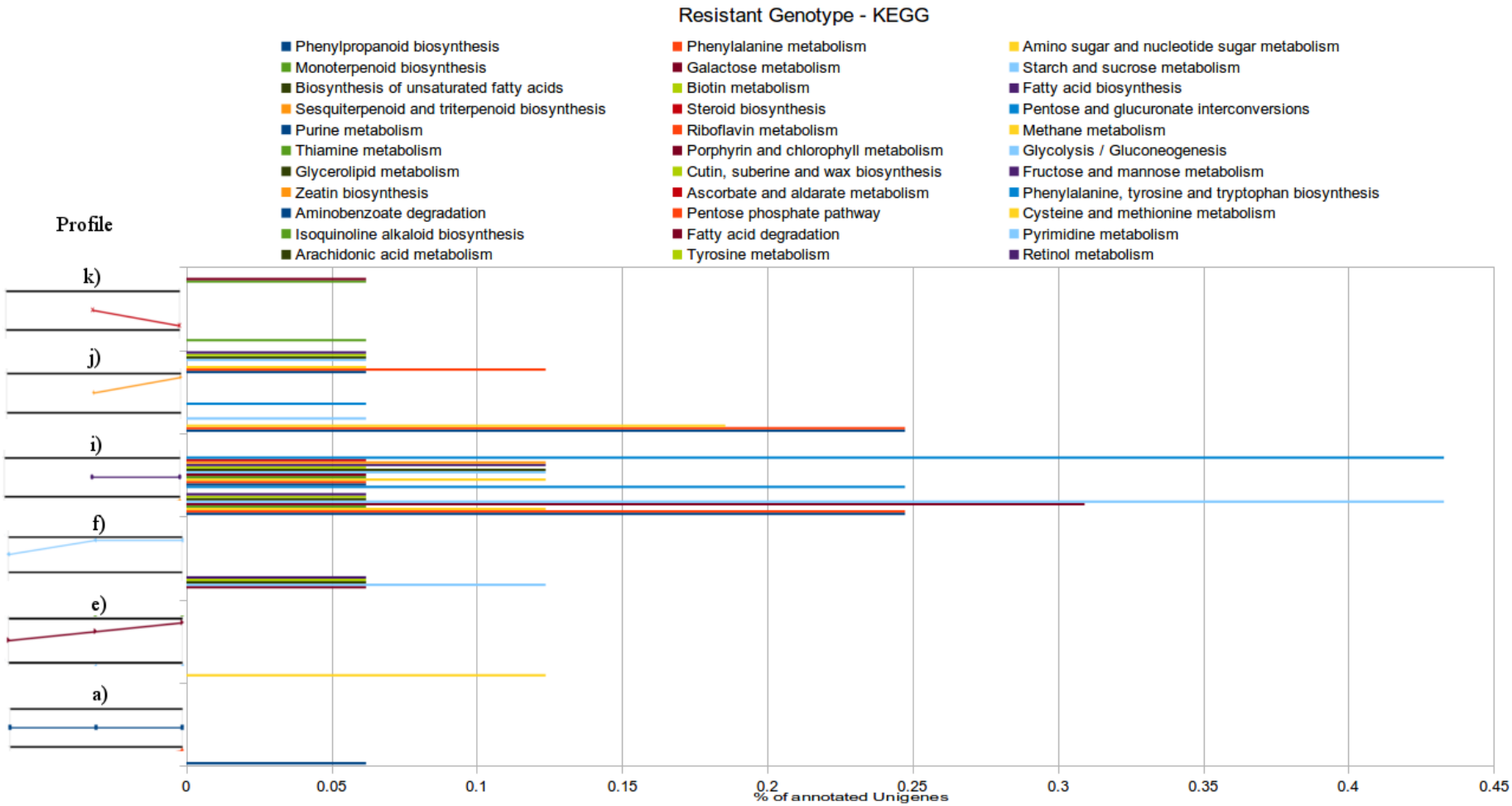


Figure S7b – KEGG annotation per profile of the resistant genotype. The percentage of unigenes is relatively to the total of differentially expressed unigenes of the susceptible comparisons – Control vs Inoculated

Table S1 - Comparison of the current and ARK genomics analysis.

Condition	EST	Current analysis		ARK genomics analysis		Current analysis read counts				ARK genomics analysis read counts				ARK genomics analysis read counts + Ebseq analysis	
		PPDE	Log2FC	p-value	Log2FC	1 C	2C	1Q	2Q	1 C	2C	1Q	2Q	PPDE	Log2FC
RCRQ48	Ca_HDT832-2_M05513	1.000	1.757	0.000	3.764	155.830	465.480	256.760	734.170	46.000	132.000	310.000	851.000	1.000	2.279
RCRQ48	Ca_HDT832-2_M08666	1.000	-1.527	0.000	-3.878	97.400	781.040	48.760	16.600	8.000	32.000	1.000	0.000	1.000	-5.743
RCRQ48	Ca_HDT832-2_M11179	1.000	2.918	0.000	4.921	3.000	2.000	15.000	20.000	3.000	9.000	51.000	79.000	0.000	0.791
RCRQ48	Ca_HDT832-2_M14285	1.000	3.690	0.000	1.433	1.980	18.120	19.620	75.470	69.000	413.000	186.000	284.000	0.986	-1.415
RCRQ48	Ca_HDT832-2_M54300	1.000	1.999	0.000	28.887	124.430	1989.350	610.840	1447.800	0.000	0.000	8.000	24.000	0.000	2.818
RCRQ48	Ca_HDT832-2_M54550	1.000	2.402	0.001	4.953	4.130	32.220	15.170	52.080	0.000	1.000	0.000	11.000	0.720	0.613
RCRQ48	Ca_HDT832-2_M54599	1.000	5.229	0.000	3.210	0.000	42.430	104.960	208.790	1.000	7.000	7.000	19.000	0.000	0.400
RCRQ48	Ca_HDT832-2_M55940	0.998	3.921	0.000	7.020	0.000	1.350	0.000	26.010	0.000	1.000	0.000	49.000	0.854	2.444
RCRQ48	Ca_HDT832-2_M57320	0.998	2.401	0.000	4.607	1.680	5.010	3.130	21.270	0.000	3.000	9.000	16.000	0.000	0.651
RCRQ72	Ca_H147-1_N00621	1.000	1.947	0.000	4.144	13.000	38.820	87.050	80.980	4.000	9.000	117.000	74.000	0.665	1.290
RCRQ72	Ca_H147-1_N00912	0.998	2.844	0.000	30.016	3.000	2.000	24.060	12.950	0.000	0.000	36.000	27.000	0.000	4.481
RCRQ72	Ca_H147-1_N03343	0.987	4.345	0.000	6.739	0.970	1.000	30.000	26.020	2.000	0.000	123.000	54.000	0.000	3.129
RCRQ72	Ca_HDT832-2_M00373	1.000	1.259	0.000	3.310	1660.100	2706.000	4604.210	4086.080	198.000	333.000	2021.000	2404.000	0.593	1.455
RCRQ72	Ca_HDT832-2_M00447	0.997	2.963	0.000	6.220	356.150	391.940	3116.550	1710.210	40.000	42.000	3168.000	1904.000	1.000	3.200
RCRQ72	Ca_HDT832-2_M04739	1.000	4.440	0.000	6.603	6.630	3.280	126.220	78.690	0.000	1.000	42.000	40.000	0.978	3.478
RCRQ72	Ca_HDT832-2_M11179	1.000	4.984	0.000	8.481	1.000	1.000	38.000	52.000	1.000	0.000	158.000	143.000	0.997	4.603
RCRQ72	Ca_HDT832-2_M14285	1.000	4.578	0.000	1.682	2.090	8.970	120.510	123.500	141.000	138.000	339.000	424.000	0.979	-0.959
RCRQ72	Ca_HDT832-2_M14990	1.000	-7.299	0.000	-1.568	139.260	75.220	0.000	0.000	167.000	234.000	63.000	50.000	1.000	-4.501
RCRQ72	Ca_HDT832-2_M15595	1.000	5.806	0.002	-3.080	0.000	0.000	12.520	54.220	10.000	10.000	2.000	0.000	1.000	-4.733
RCRQ72	Ca_HDT832-2_M22730	1.000	3.895	0.000	7.214	1.250	0.890	15.540	28.470	1.000	0.000	30.000	99.000	0.955	3.216
RCRQ72	Ca_HDT832-2_M24255	1.000	4.781	0.000	1.486	0.000	3.160	50.860	50.340	49.000	48.000	122.000	107.000	0.991	-1.564
RCRQ72	Ca_HDT832-2_M25152	1.000	6.437	0.000	32.482	0.000	0.000	50.000	50.010	0.000	0.000	172.000	178.000	0.007	5.748
RCRQ72	Ca_HDT832-2_M32644	1.000	3.798	0.000	1.762	2.510	4.820	46.290	54.130	23.000	23.000	47.000	87.000	0.982	-1.361

RCRQ72	Ca_HDT832-2_M33411	1.000	6.473	0.000	3.970	0.000	0.000	51.000	51.620	15.000	17.000	213.000	210.000	0.750	0.919
RCRQ72	Ca_HDT832-2_M37463	0.959	2.747	0.000	5.779	1.000	2.000	15.880	7.000	0.000	3.000	76.000	62.000	0.927	3.089
RCRQ72	Ca_HDT832-2_M42026	1.000	4.741	0.000	28.891	0.000	3.210	53.560	45.460	0.000	0.000	16.000	13.000	0.000	3.425
RCRQ72	Ca_HDT832-2_M53490	1.000	1.492	0.000	3.755	488.090	467.430	1282.800	965.530	61.000	70.000	933.000	531.000	0.303	0.766
RCRQ72	Ca_HDT832-2_M53657	1.000	5.058	0.000	29.649	80.890	146.830	3276.490	3066.630	0.000	0.000	13.000	37.000	0.000	3.979
RCRQ72	Ca_HDT832-2_M54300	1.000	1.758	0.000	30.193	391.230	602.890	1312.100	1511.710	0.000	0.000	43.000	28.000	0.000	1.374
RCRQ72	Ca_HDT832-2_M54599	1.000	5.844	0.000	3.282	3.080	9.190	385.740	254.370	1.000	3.000	14.000	19.000	0.000	2.134
RCRQ72	Ca_HDT832-2_M55019	1.000	2.004	0.000	4.196	21.010	25.250	67.560	93.250	13.000	24.000	250.000	324.000	0.953	2.523
SCSQ48	Ca_H147-1_N06250	1.000	5.681	0.000	1.701	0.000	0.000	27.980	37.020	42.000	70.000	218.000	223.000	0.898	-0.721
SCSQ48	Ca_HDT832-2_M00373	1.000	1.384	0.000	3.845	1135.900	1130.370	3621.220	3500.840	54.000	99.000	1393.000	1262.000	0.304	0.336
SCSQ48	Ca_HDT832-2_M05513	1.000	2.786	0.000	4.789	44.630	81.230	494.910	501.500	7.000	11.000	262.000	339.000	0.270	-0.007
SCSQ48	Ca_HDT832-2_M14285	1.000	3.784	0.000	1.416	2.000	4.240	42.760	69.280	35.000	64.000	169.000	151.000	1.000	-2.509
SCSQ48	Ca_HDT832-2_M42971	0.996	4.054	0.000	1.706	0.000	1.000	12.060	24.040	13.000	27.000	65.000	91.000	0.000	-0.719
SCSQ48	Ca_HDT832-2_M50633	1.000	4.011	0.000	1.741	10.610	22.780	264.120	355.920	2.000	12.000	25.000	31.000	0.000	-0.428
SCSQ48	Ca_HDT832-2_M54300	1.000	1.226	0.000	28.573	306.890	295.560	994.780	719.540	0.000	0.000	10.000	8.000	0.000	3.380
SCSQ48	Ca_HDT832-2_M55019	1.000	2.378	0.000	4.768	6.000	24.110	87.610	85.600	2.000	9.000	186.000	176.000	1.000	1.233
SCSQ72	Ca_H147-1_N06250	0.990	4.167	0.000	1.965	0.000	3.000	12.000	69.490	66.000	52.000	94.000	316.000	0.596	-1.260
SCSQ72	Ca_HDT832-2_M05918	1.000	5.731	0.000	7.889	1.000	0.000	30.400	69.460	1.000	0.000	65.000	135.000	1.000	3.896
SCSQ72	Ca_HDT832-2_M08452	1.000	5.083	0.000	2.602	84.480	59.580	822.590	3686.270	4.000	1.000	3.000	23.000	0.656	-0.892
SCSQ72	Ca_HDT832-2_M25116	0.958	3.730	0.000	29.392	1.310	0.000	4.440	25.080	0.000	0.000	4.000	20.000	0.000	2.960
SCSQ72	Ca_HDT832-2_M25152	1.000	6.279	0.000	32.258	0.000	0.000	28.600	59.540	0.000	0.000	55.000	119.000	1.000	6.853
SCSQ72	Ca_HDT832-2_M54300	1.000	1.708	0.000	5.343	429.410	248.660	528.730	1365.580	1.000	0.000	5.000	30.000	0.000	4.665
SCSQ72	Ca_HDT832-2_M54599	1.000	4.922	0.001	28.410	10.460	11.000	156.050	471.200	0.000	0.000	2.000	10.000	0.000	-0.764

Chapter III

Concluding Remarks

CBD is a disease of hard and expensive control with a high economic and social impact. It was already recognized that a durable control of this disease has to come from a better knowledge about the pathogen infection process and the plant's resistance mechanisms. As such, there is a need for a continued effort on studying the host-pathogen relationship.

Joining different research fields to unravel the molecular mechanisms of coffee resistance, like informatics and molecular biology, can lead to significant advances on the knowledge needed to implement strategies towards the improvement of coffee resistance durability to CBD, taking into account the environmental needs and the development of a sustainable coffee economy.

Currently, with the constant innovation in NGS and the large and complex choices of methods and analysis software, it is difficult to keep up with the best approaches to address the problems that still arise. Actually, with recent methods like RNA-seq, there is still no consensus about the best practices of analysis for all kinds of data. Therefore, it is up to the user to adjust and keep track of the quality of analysis, in order to bring the best possible approximation of the data to reality.

In this work, the importance of such care is evidenced by the comparison of two different analysis of the same data. It was shown that the indiscriminate use of a standard approach independently of the data source can lead to not so realistic results. In addition, it was possible to recognize the importance of proper computational power. Insufficient CPU and RAM power can be a limiting factor, restricting the analysis possibilities.

Furthermore, it was possible to infer the different proprieties of two methods for host-pathogen sequence separation. Neither of the methods was perfect for sequence separation, especially if the presence of contaminants other than the pathogen could not be excluded. Ideally, the use of more than one method for separation is advised.

Finally, a first look at the annotation results, already showed unigenes differentially expressed related with pathways involved in defense responses and the recognition of the different reactions of susceptible and resistant genotypes to fungus infection. A closer look to the results, will lead to new relevant knowledge able to support and improve coffee breeding for resistance to CBD.

References

Anthony F, Combes C, Astorga C, Bertand B, Graziosi G, Lashemes P (2002) The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. *Theoretical and Applied Genetics*, **104**, 894–900.

Bigger M (2006) The dissemination of coffee cultivation throughout the world. *Tropical Agriculture Association Newsletter*, **26**, 15–19.

Bohnert R, Behr J, Ratsch G (2009) Transcript quantification with RNA-seq data. *BMC bioinformatics*, **10**.

Van den Bosch F, Gilligan C (2008) Models of Fungicide Resistance Dynamics. *Annual review of phytopathology*, **46**, 123–147.

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

Butler G, Rasmussen MD, Lin MF, Santos MAS, Sakthikumar S, Monro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL, Kellis M, Cuomo CA *et al.* (2009) Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature*, **459**, 657–62.

Charrier A, Berthaud J (1985) Botanical Classification of Coffee. In: *Coffee* (eds Clifford MN, Willson KC), pp. 13–47. Springer US, Boston, MA.

Chen Z, Nunes MA, Silva MC (2004) Appressorium turgor pressure of *Colletotrichum kahawae* might have a role in coffee cuticle penetration. *Mycologia*, **96**, 1199–1208.

Chung W-H, Ishii H, Nishimura K, Fukaya M, Yano K, Kajitani Y (2006) Fungicide Sensitivity and Phylogenetic Relationship of Anthracnose Fungi Isolated from Various Fruit Crops in Japan. *Plant Disease*, **90**, 506–512.

Davis A (2013) A New Combination in *Psilanthus* (Rubiaceae) for Australasia, and Nomenclatural Notes on *Paracoffea*. *Novon*, **13**, 182–184.

Davis A, Govaerts R, Bridson DM, Stoffelen P (2006) An annotated taxonomic conspectus of the genus *Coffea* (Rubiaceae). *Botanical Journal of the Linnean Society*, **152**, 465–512.

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

Ferrão J (1993) Café. In: *A aventura das plantas e os descobrimentos Portugueses*, pp. 261–262.

Fitt B, McCartney H, Walklate P (1989) The Role of Rain in Dispersal of Pathogen Inoculum. *Annual Review of Phytopathology*, **27**, 241–270.

Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, Faris JD, Rasmussen JB, Solomon PS, McDonald BA, Oliver RP (2006) Emergence of a new disease as a result of interspecific virulence gene transfer. *Nature Genetics*, **38**, 953–956.

Garber M, Grabherr M, Guttman M, Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods*, **8**, 469–477.

Gibbs J (1969) Inoculum sources for coffee berry disease. *Annals of Applied Biology*, **64**, 515–522.

Gichuru EK (1997) Resistance mechanisms in Arabica coffee to coffee berry disease (*Colletotrichum kahawae* Sp. Nov.); a review. *Kenya Coffee (Kenia)*, **67**, 2441–2444.

Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin JF (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *BMC genomics*, **12**.

Griffiths E, Gibbs JN, Waller JM (1971) Control of coffee berry disease. *Kenya Coffee,***36**,307–28.

Gururani MA, Venkatesh J, Upadhyaya CP, Nookaraju A, Pandey SK, Park SW (2012) Plant disease resistance genes: Current status and future directions. *Physiological and Molecular Plant Pathology*, **78**, 51–65.

Hammond-Kosack K, Jones J (1996) Resistance gene-dependent plant defense responses. *The Plant cell*, **8**, 1773–91.

Van Hilten H, Fisher P, Wheeler M, Wagner B (2011) *The Coffee Exporter's Guide*. International Trade Centre UNCTAD/GATT, Geneva.

Jeffries P, Koomen I (1992) Strategies and prospects for biological control of diseases caused by *Colletotrichum*. In: *Colletotrichum: biology, pathology and control* (eds. Bailey JA, Jeger MJ, British Society for Plant Pathology), pp 337-357, C.A.B. International, Wallingford.

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

Lashermes P, Combes MC, Robert J, Trouslot P, D'Hont A, Anthony F, Charrier A (1999) Molecular characterization and origin of the *Coffea arabica* L. genome. *Molecular & general genetics,* **261**, 259–266.

Lashermes P, Anthony F (2007) Coffee. In: *Genome Mapping and Molecular Breeding Plants, Technical Crops* (ed Kole C), pp. 108–118. Springer-Verlag, Berlin.

Lécolier A, Besse P, Charrier A, Tchakaloff T-N, Noirot M (2009) Unraveling the origin of *Coffea arabica* "Bourbon pointu" from La Réunion: a historical and scientific perspective. *Euphytica*, **168**, 1–10.

Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, Haag JD, Gould MN, Stewart RM, Kendziorski C (2013) EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, **29**, 1035–1043.

Li B, Dewey C (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics*, **12**.

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Liu L, Li Y, Li S, Hu N,, He Y, Pong R, Lin D, Lu L, Law M (2012) Comparison of next-generation sequencing systems. *Journal of Biomedicine & Biotechnology*, **2012**.

Loman NJ, Misra R V, Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, **30**, 434–439.

Loureiro A, Nicole MR, Várzea V, Moncadac P, Bertrandb B, Silva MC (2012b) Coffee resistance to *Colletotrichum kahawae* is associated with lignification, accumulation of phenols and cell death at infection sites. *Physiological and Molecular Plant Pathology*, **77**, 23–32.

Manga B (1997) Observations sur la diversité de la population de *Colletotrichum kahawae* agent de l'anthracnose des baies du caféier Arabica. Implications pour l'amélioration génétique. In: *Proceedings of 17th International Conference on Coffee Science (ASIC),* Kenya

Manuel L, Talhinhas P, Várzea V, Neves-Martins J (2010) Characterization of *Colletotrichum kahawae* Isolates Causing Coffee Berry Disease in Angola. *Journal of Phytopathology*, **158**, 310–313.

Mardis ER (2008) Next-generation DNA sequencing methods. *Annual review of genomics and human genetics*, **9**, 387–402.

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

Martin JA, Wang Z (2011) Next-generation transcriptome assembly. *Nature reviews Genetics*, **12**, 671–682.

Masaba DM, Van der Vossen HAM (1992) Coffee berry disease: the current status. In: *Colletotrichum: Biology, Pathology and Control* , pp. 237–249.

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

Metzker ML (2010) Sequencing technologies - the next generation. *Nature reviews. Genetics*, **11**, 31–46.

Miller H, Biggs P, Voelckel C, Nelson N (2012) De novo sequence assembly and characterisation of a partial transcriptome for an evolutionarily distinct reptile, the tuatara (*Sphenodon punctatus*). *BMC Genomics*, **13**, 12.

Mortazavi A, Williams B, McCue K (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, **5**, 621–628.

Mou H-Q, Lu J, Zhu S-F, Lin C-L, Tian G-Z, Xu X, Zhao W-J (2013) Transcriptomic analysis of *Paulownia* infected by Paulownia witches'-broom *Phytoplasma*. *PloS One*, **8**.

Mulinge SK (1971) Effect of altitude on the distribution of the fungus causing coffee berry disease in Kenya. *Annals of Applied Biology*, **67**, 93–98.

Nagalakshmi U, Waern K, Snyder M (2010) RNA-Seq: a method for comprehensive transcriptome analysis. *Current protocols in molecular biology*, **4**.

Ntahimpera N, Wilson LL, Ellis M a, Madden L V (1999) Comparison of rain effects on splash dispersal of three *Colletotrichum* species infecting strawberry. *Phytopathology*, **89**, 555–563.

Nutman FJ (1970) Coffee berry disease. *Tropical Pest Management*, **16**, 277–286.

Nutman FJ, Roberts FM (1960) Investigations on a disease of *Coffea arabica* caused by a form of *Colletotrichum coffeanum* Noack II Some Factors afecting infection by the Pathogen. *Transactions of the British Mycological Society*, **43**, 643–659.

Oshlack A, Robinson MD, Young MD (2010) From RNA-seq reads to differential expression results. *Genome biology*, **11**.

Osorio N (2002) The global coffee crisis: a threat to sustainable development. In: *World summint on sustainable Development* . Johannesburg.

Ozsolak F, Milos P (2010) RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, **12**, 87–98.

Parchman TL, Geist KS, Grahnen J a, Benkman CW, Buerkle CA (2010) Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. *BMC genomics*, **11**.

Peatman E, Li C, Peterson BC, Straus DL, Farmer BD, Beck BH (2013) Basal polarization of the mucosal compartment in *Flavobacterium columnare* susceptible and resistant channel catfish (*Ictalurus punctatus*). *Molecular immunology*, **56**, 317–27.

Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, Bertoni A, Swerdlow AHP, Gu Y (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, **13**, 341.

Raina SN, Mukai Y, Yamamoto M (1998) In situ hybridization identifies the diploid progenitor species of *Coffea arabica* (Rubiaceae). *Theoretical and Applied Genetics*, **97**, 1204–1209.

Rodrigues CM, de Souza AA, Takita MA, Kishi LT, Machado MA(2013) RNA-Seq analysis of *Citrus reticulata* in the early stages of *Xylella fastidiosa* infection reveals auxin-related genes as a defense response. *BMC Genomics*, **14**, 676.

Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M *et al.* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–52.

Schroth G, Krauss U, Gasparotto L, Aguilar JAD, Vohland K (2000) Pests and diseases in agroforestry systems of the humid tropics. *Agroforestry Systems*, **50**, 199–241.

Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–1092.

Sebastiana M, Vieira B, Lino-Neto T, Monteiro F, Figueiredo A, Sousa L, Pais MS, Tavares R, Paulo OS (2014) Oak Root Response to Ectomycorrhizal Symbiosis Establishment: RNA-Seq Derived Transcript Identification and Expression Profiling. *PloS one* **9**.

Seyednasrollah F, Laiho A, Elo LL (2013) Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in bioinformatics*.

Silva DN, Talhinhas P, Cai L, Manuel L, Gichuru EK, Loureiro A, Várzea V, Paulo OS, Batista D (2012) Host-jump drives rapid and recent ecological speciation of the emergent fungal pathogen *Colletotrichum kahawae*. *Molecular Ecology*, **21**, 2655–2670.

Silva MC, Várzea V, Guerra-Guimarães L, Azinheira HG, Fernandez D, Petitot A-S, Bertrand B, Lashermes P, Nicole M (2006) Coffee resistance to the main diseases: leaf rust and coffee berry disease. *Brazilian Journal of Plant Phydiology*, **18**, 119–147.

Soneson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. *BMC bioinformatics*, **14**, 18.

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

Surget-Groba Y, Montoya-Burgos JI (2010) Optimization of de novo transcriptome assembly from next-generation sequencing data. *Genome Research*, **20**, 1432–1440.

Topik S (2004) The world coffee market in the eighteenth and nineteenth centuries, from colonial to national regimes.

Ukers WH (1935) *All About Coffee*. Tea and Coffee Trade Journal Company, New York

Varzea VM, Rodrigues Jr CJ, Medeiros E (1993) Different pathogenicity of CBD isolates on coffee genotypes. In: *Proceedings of 15th International Conference on Coffee Science (Montpellier),* Kenya

Vega FE, Rosenquist E, Collins W (2003) Global project needed to tackle coffee crisis. *Nature*, **425**, 343.

Vega FE, Simpkins A, Aime MC, Posada F, Peterson SW, Rehner SA, Infante F, Castillo A, Arnold AE (2010) Fungal endophyte diversity in coffee plants from Colombia, Hawai'i, Mexico and Puerto Rico. *Fungal Ecology*, **3**, 122–138.

Vermeulen H (1970) Coffee berry disease in Kenya. I. *Colletotrichum* spp. colonizing the bark of *Coffea arabica*. *Netherlands Journal of Plant Pathology*, **76**, 277–284.

Van Der Vossen H (2009) The cup quality of Diesease_Resistant cultivars of Arabica coffee (*Coffea arabica*). *Experimental Agriculture*, **45**, 323.

Van der Vossen H, Cook R, Murakaru G (1976) Breeding for resistance to coffee berry disease caused by *Colletrotrichum coffeanum* Noack (Sensu Hindorf) in *Coffea arabica* LI Methods of preselection for resistance. *Euphytica*, **25**, 733–745.

Van der Vossen H, Walyaro DJ (2009) Additional evidence for oligogenic inheritance of durable host resistance to coffee berry disease (*Colletotrichum kahawae*) in Arabica coffee (*Coffea arabica* L.). *Euphytica*, **165**, 105–111.

Waller JM (1972) Water-borne spore dispersal in coffee berry disease and its relation to control. *Annals of Applied Biology*, **71**, 1–18.

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

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 JM, Masaba DM (2006) The microflora of coffee surfaces and relationships to coffee berry disease. *International Journal of Pest Management*, **52**, 89–96.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, **10**, 57–63.

Weir BS, Johnston PR, Damm U (2012) The *Colletotrichum gloeosporioides* species complex. *Studies in mycology*, **73**, 115–80.

Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bähler J (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–43.

Wrigley G (1988) *Coffee (Tropical Agriculture Series)*. Longman Sc & Tech, London