RAFAELA LEITE PRADO ROCHA

CHARACTERIZATION OF LTR-RETROTRANSPOSONS ON *Hemileia vastatrix* GENOME

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Fitotecnia, para a obtenção do título de *Magister Scientiae.*

VIÇOSA
MINAS GERAIS - BRASIL
2017

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T

R672c
2017

Rocha, Rafaela Leite Prado, 1991-
Characterization of LTR-retrotransposons on *Hemileia vastatrix* Genome / Rafaela Leite Prado Rocha. – Viçosa, MG, 2017.
vi, 21f. : il. (algumas color.) ; 29 cm.

Orientador: Ney Sussumu Sakiyama.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.18-21.

1. *Hemileia vastatrix*. 2. Café - Doenças e pragas. 3. Cafeeiro. 4. Café - Genética. 5. Genoma. I. Universidade Federal de Viçosa. Departamento de Fitotecnia. Programa de Pós-graduação em Fitotecnia. II. Título.

CDD 22 ed. 633.732

RAFAELA LEITE PRADO ROCHA

CHARACTERIZATION OF LTR-RETROTRANSPOSONS ON *Hemileia vastatrix* GENOME

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Fitotecnia, para a obtenção do título de *Magister Scientiae.*

APROVADA: 31 de março de 2017.

Eveline Teixeira Caixeta
(Coorientadora)

Marisa Vieira de Queiroz
(Coorientadora)

Tiago Antônio de Oliveira Mendes

Ney Sussumu Sakiyama
(Orientador)

DEDICATÓRIA

À minha mãe, Stael, por ser exemplo de mulher e minha guia,

Dedico.

AGRADECIMENTOS

Primeiramente, agradeço a Deus pelo dom da vida e pelas benções concedidas.

À Universidade Federal de Viçosa, pela excelente infraestrutura e ensino oferecido durante todo o tempo que estive aqui.

Ao Programa de Pós Graduação em Fitotecnia pela oportunidade de realizar meu estudos nesta Universidade e pelo excelente suporte oferecido a nós alunos.

Ao meu orientador, Dr. Ney Sussumu Sakiyama, por todas as conversas e conselhos oferecidos e pela confiança na realização do trabalho.

À Dr.ª Eveline Teixeira Caixeta, pela amizade, suporte, confiança, conselhos e ensinamentos de vida professional ofertados.

À Dr.ª Marisa Vieira de Queiroz, pelas contribuições oferecidas ao trabalho e confiança depositada em mim.

Ao Dr. Tiago Mendes, pelos ensinamentos em Bioinformática, contribuições ofertadas ao nosso trabalho.

Ao Dr. Aluizio Borém, por todos os conselhos e confiança depositada durante minha tragetória nesta Universidade.

À Familia Biocafé, por todo o suporte e amizade.

Aos amigos que fiz durante toda minha tragetória.

À minha familia, por acreditar nos meus sonhos.

Muito Obrigada!

SUMÁRIO

ABSTRACT

ROCHA, Rafaela Leite Prado, Univerisdade Federal de Viçosa, March, 2017. **Characterization of LTR-Retrotransposons on *Hemileia vastatrix* Genome**. Advisor: Ney Sussumu Sakiyama. Co-Advisors: Eveline Teixeira Caixeta and Marisa Vieira de Queiroz.

Brazil is the biggest producer and exporter of coffee in the world. The country, as the rest of coffee growing regions, suffers with coffee rust disease. Rust, one of the earliest coffee diseases studied scientifically, is caused by the obligate biotrophic fungus *Hemileia vastatrix*. This disease is the most harmful that affects coffee trees, which may cause drastic drops in productivity if not controlled. The fungus infects coffee leaves penetrating through stomata and develops powdery orange pustules on the abaxial surface. The pathogen displays high levels of genetic variability leading to appearance of new physiological races and supplanting the resistance of coffee varieties obtained in breeding programs. The mechanism that causes such variability is yet not known, since it is accepted that the fungus relies on asexual reproduction and the sexual stage has not been observed in nature. The genome of *H. vastatrix* has been reported as one of the largest known rust genome. A very large fraction of the published rust genomes has shown a high percentage of repetitive elements, such as transposable elements (TEs). In this sense, TEs activity is suggested as an important source for generation of variability. Thus, there is a chance that the transposable DNA elements are responsible for increasing genetic variability of *H. vastatrix*. Therefore, this study aimed to identify presence, frequency and location of transposable elements, as well as verify the relationship between transposable elements and the high variability found on genome of *H. vastatrix*. We found 6,516 gene-coding proteins and 1,109 complete LTR retrotransposons in the genome. From these gene-coding proteins, 65 were close to LTR retrotransposons. The results suggested that LTR retrotransposon generally insert itself 5,000bp from a gene-coding protein. Also, LTR retrotransposons identified insert itself on AT-rich conserved regions. Thus, the found data suggest that insertion of LTR retrotransposons in *H. vastatrix* genome could be targeting its integration site based on nucleotide composition. Moreover, we identified LTR retrotransposons located close to coding regions, which could be contributing with gene expression modulation and hence affecting pathogen variability.

RESUMO

ROCHA, Rafaela Leite Prado, Univerisdade Federal de Viçosa, março de 2017. **Characterization of LTR-Retrotransposons on *Hemileia vastatrix* Genome**. Orientador: Ney Sussumu Sakiyama. Coorientadoras: Eveline Teixeira Caixeta e Marisa Vieira de Queiroz.

O Brasil é o maior produtor e exportador de café do mundo. O país, como o resto das regiões produtoras de café, sofre com a doença da ferrugem do cafeeiro. A ferrugem é causada pelo fungo biotrófico *Hemileia vastatrix*. Esta doença pode levar a quedas drásticas na produtividade quando não controlada. Esse patógeno apresenta altos níveis de variabilidade genética, o que resulta em aparecimento de novas raças fisiológicas e, consequentemente, a suplantação da resistência de variedades de café obtidas pelos programas de melhoramento genético. O mecanismo que causa essa variabilidade ainda não é conhecido, uma vez que o fungo tem reprodução assexuada e o estádio sexual não foi observado na natureza. Tem sido relatado que os genomas das ferrugens apresentam alta porcentagem de elementos repetitivos, incluido os elementos transponíveis (ET). Como a atividade de ET tem sido sugerida como uma fonte importante de geração de variabilidade, existe a possibilidade de que os ETs sejam um dos responsáveis pela grande variabilidade genética encontrada em *H. vastatrix*. Portanto, este estudo teve como objetivo identificar a presença, a frequência e a localização de LTR retrotransposons no genoma de *H. vastatrix*, bem como verificar a relação entre LTR retrotransposons e a variabilidade encontrada nesse patógeno. Foram identificados no genoma analisado 6.516 genes codificadores de proteínas e 1.109 retrotransposons do tipo LTR completos. Dos genes codificadores de proteínas, 65 estavam próximos de retrotransposons do tipo LTR. Observou-se que os retrotransposons LTR geralmente se inserem a 5.000 pb de um gene codificador de proteínas no genoma desse patógeno. Os resultados encontrados demonstraram que retrotransposons LTR inserem-se em regiões conservadas ricas em AT. Assim, os dados sugerem que a inserção dos retrotransposons LTR em *H. vastatrix* podem se direcionar com base na composição nucleotídica de uma região. Além disso, foram identificados retrotransposons próximos de regiões codificadoras, o que poderia contribuir para modulação da expressão gênica e, consequentente, afetar a variabilidade do patógeno.

CHARACTERIZATION OF LTR-RETROTRANSPOSONS ON *Hemileia vastatrix* GENOME

Introduction

Coffee is one of the most traded commodities in the world (Vega et al. 2003). It is important for economy of more than 60 countries and is the main source of income for more than 100 million people (ICO 2016). Brazil is the biggest producer and exporter of coffee. The country has 2,22 million hectares of planted area and ten producer states (CONAB 2016).

The genus *Coffea* has more than 100 species, being only *C. arabica* and *C. canephora* cultivated. All predominant growing species of the genus *Coffea* are susceptible to *Hemileia vastatrix,* causal agent of coffee leaf rust disease. *C. arabica*, the main cultivated species, is the most susceptible. However, some varieties of *C. canephora* has shown desired resistance to rust (Waller 1982). Although arabica is highly susceptible to *H. vastatrix,* the species is known for its high yielding and cupping quality under optimal wheatear and soil conditions, which could explain the preference for this specie at a world level (Van Der Vossen et al. 2015).

Coffee rust disease is believed to have begun with wild coffee in Ethiopia but it was first reported in Sri Lanka and, it has been spread to all coffee producer countries ever since (Kushalappa 1989). In Brazil, the disease was first reported in 1970 in Bahia state and around four months later it was found all over Brazil's coffee growing regions (Zambolim 2016). The disease was feared for many years until it was proved the effectiveness of chemical control and in virtue of the limited damage caused by *H. vastatrix*, especially at high altitudes. Technical authorities and farmers considered the disease manageable even though significant losses were caused by the pathogen since then. These losses went unnoticed as a result of an interaction between coffee tree phenology and this disease (Avelino et al. 2015).

In the recent years, however, the disease has once again drawn the attention of farmers and technical authorities due to an epidemy that has been affecting Central and

South America. Yield losses on the most affected countries as Ecuador, El Salvador, Honduras, Nicaragua and Panama where estimated around 30 to 90% (Zambolim 2016). The chemical control, in some ways, has efficiently controlled the pathogen when the recommendation of the product has been precisely followed by farmers (Chalfoun & Carvalho 1999). However, factors as cost of production, risk of intoxication of both workers and harvested berries and yet emergence of new physiologic races lead to research of more rational disease control practices, highlighting the importance of breeding for resistance (Garçon et al. 2004).

Resistant cultivars have shown to be a more suitable option to disease control, since it is a more sustainable and efficient way to repress rust disease. However, the high levels of genetic variability observed on *H. vastatrix* races (Cabral et al. 2016; Maia et al. 2013) has contributed to pathogen overcome resistance of coffee cultivars obtained on breeding programs (Varzea & Marques 2005).

There were more than 50 races of *H. vastatrix* identified by Coffee Rust Research Center (CIFC) in Portugal, fifteen of it has already been found in Brazil's plantations. Because sexual stage of *H. vastatrix* have not been seen in nature, the mechanisms which lead to high emergence of *H. vastatrix* new physiological races are still unknown (Zambolim 2016). Although sexual stage has not been observed on nature, a study showed that *H. vastatrix* has a hidden sexual reproduction, also known as cryptosexuality. This study provided compelling evidence that meiosis occurs in the uredineospores (Carvalho et al. 2011). The fact could contribute but does not fully explain such genetic variability found on the pathogen.

Since the introduction of the pathogen in Latin America, coffee rust has evolved due to the selection pressure exerted by resistant genotypes (Avelino et al. 2015). The hypothesis that mutation and selection pressure could be leading to variability of *H. vastatrix* has been explored, but none of them were able to fully explain such mechanisms. Studies found no direct link between such high phenotypic and molecular variability (Cristancho & Escobar 2008). Although the use of molecular markers as SSR and sequencing of the rDNA-ITS regions failed to provide informative results, the use of random amplification of polymorphic DNAs (RAPDs) and amplified fragment length polymorphisms (AFLPs) showed to be the only conclusive markers up to date. Thus, with the development of sequencing and genome assembly techniques it is believed that more comprehensive results will soon be obtained (Talhinhas et al. 2017).

H. vastatrix has one of the largest fungal genomes reported, with an average size of approximately 760 Mbp, which has been estimated by flow cytometry. In comparison, other rust species have a genome average size of 225.23 Mbp (Carvalho et al. 2014; Tavares et al. 2014; Talhinhas et al. 2017). This size variation could be explained by differential expansions of transposable elements, which seems to be related with the fungal lifestyle (Castanera et al. 2016).

Transposable elements (TEs) are genetic units which can move from place to place among the genome. Those units colonize eukaryotic and prokaryotic genomes and are responsible to contribute with increasing genetic variability. The genome percentage occupied by TEs varies among species but it can reach up to 50% on mammals, 80% some plants and 40% on fungi (Castanera et al. 2016; Wicker et al. 2007).

TEs can be classified into two distinct groups based on mechanistic and enzymatic criteria. This groups are based on the presence and absence of RNA transposition intermediate. Class I elements, also known as retrotransposons, have a RNA transposition intermediate and are not divided into subclasses. Elements of class I are LTR, DIRS, PLE, LINE and SINE. Class II elements are known as DNA transposons and are divided into two subclasses distinguished by the number of DNA strands that are cut during transposition. Neither class II subclasses have a RNA transposition intermediate. Elements class II subclasse I comprehend TIR and Crypton while subclasse II comprehend Helitron and Maverik elements (Wicker et al. 2007).

Previous studies have shown that TEs can influence regulation of genes adjacent to them at a transcriptional and post-transcriptional level. Also, DNA transposons and their transposases could be considered great sources of basic components for emerging new transcription factors (Feschotte 2008). This demonstrates how the integration of transposable elements among the genome could contribute to increase pathogen variability. Additionally, it was previously reported that plant pathogen genome size could have an impact on their pathogenicity (D’Hondt et al. 2011). Thus, as the movement of transposable elements can contribute in genome size variation, they could be impacting pathogen virulence.

The activity of transposable elements among the genome can also be harmful and create genome instability. Consequently, some organisms have developed defense mechanisms which can inhibit its movement. There are three main fungal silencing mechanism which are RIP (Repeat-induced point mutation), MIP (methylation induced premeiotically) and Quelling (RNA interference). RIP is a gene silencing mechanism that mutates transposable elements and control its migration among fungal genome. This mechanism results on C to T and G to A changeover which can be accompanied by the methylation of remaining cytosine (Wostemeyer & Kreibich 2002). The absence of RIP or its low efficiency also appears responsible with TEs accumulation and consequently increased size genomes (Amselem et al. 2015). MIP leads to methylation of cytosine residues without mutations and different from RIP, this mechanism is reversible. Both RIP and MIP occurs during sexual reproduction (Wostemeyer & Kreibich 2002). Quelling involves RNA interference machinery that suppress TE expression (Amselem et al. 2015). It is acknowledged that mechanisms could affect regions nearby TEs, perhaps interfering on coding regions.

Transposable elements activity can also contribute to chromosomal rearrangements. This TE-induced event, also known as ectopic recombination, can cause chromosomal inversions due to abnormal alignment and can directly impact on genome variability. The inversions can be found in natural populations and are geographically widespread and polymorphic, which suggests that they are selectively advantageous (Feschotte & Pritham 2007).

Therefore, this work aimed to identify and characterize LTR retrotransposons in *H. vastatrix* genome, which has been shown as the most abundant TE, as well as verify the relationship between those elements and the high variability found on the pathogen to better understand this important coffee disease.

Methodology

Database

The DNA sequence of the reference genome of *H. vastatrix* was used. The sequences were obtained in partnership with the Universidade Federal de Viçosa (UFV) – Viçosa (Brazil), Universidade Federal de Lavras (UFLA) – Lavras (Brazil), and the

University of Delaware (UDEL) - Newark - USA. The sequencing of *H. vastatrix* genomic DNA race XXXIII was held by two platforms: Pacific Biosciences (PacBio RS II) e Illumina (HiSeq 2500). The genome used is a second draft assembled and is available for the group to work.

The assembled genome contains 58,535 contigs with N50 and N90 of 11,385 and 3,762 bp, respectively. The size of the genome is 480Mbp and CG content is 33.6%.

Identification of genes and validation

The identification of genes was performed using the genome as model of the species closest to *H. vastatrix* available on National Center for Biotechnology Information - NCBI database (https://www.ncbi.nlm.nih.gov/). To perform that we used *Scipio* software (http://www.webscipio.org/) (Keller et al. 2008) and a given set of protein sequences, more precisely sequences of the closest genome selected, to determine the precise gene structures on *H. vastatrix* genome.

Augustus software *v2.5.5* (http://bioinf.uni-greifswald.de/augustus/) (Stanke et al. 2004) was used to predict genes through their alignment with ESTs (Expressed Sequence Tags) of *Ustilago maydis,* as related species. This strategy benefits the identification of coding sequences (Open Reading Frames – ORFs). A novel strategy was performed to verify the quality of the gene assembly in the genome, using *CEGMA* software (http://korflab.ucdavis.edu/datasets/cegma/) (Parra et al. 2007).

The distance between genes and LTR retrotransposons was calculated and possible gene interference caused by LTR integration was analyzed using custom Perl scripts to verify if LTR integration could be modulating gene expression.

Molecular Function Analysis of Proteins by Functional Annotation

Aiming the assignment of GO (gene ontology) categories, the tool Goanna (http://www.agbase.msstate.edu/cgi-bin/tools/GOanna.cgi) was used (Mccarthy et al. 2006). The parameters were set using an expect value of 10^{-5}, word size of 3, low-complexity filter turned on, and the BLOSUM62 matrix. The GO annotation was based on SwissProt database. The results obtained with Goanna were recategorized using the tool GOSlim Viewer (http://www.agbase.msstate.edu/cgi-bin/tools/goslimviewer_select.pl) using the generic set (Mendes et al. 2013).

Identification, classification and analysis of transposable elements

To search for Class I complete transposons we used the software LTR-Finder (http://tlife.fudan.edu.cn/ltr_finder/) (Xu & Wang 2007), which search for patterns that represent the typical structure of a LTR retrotransposon. Thus, the software reports possible models of LTR retrotransposons at distinct levels of trust, according to the signs and domains present on the sequence.

After the search for complete LTR retrotransposons, sequences of approximately 1000 bp downstream and upstream of each element was analyzed using custom Perl scripts. A distance matrix was calculated between the sequences upstream of the LTR integration site using the software Clustal Omega (http://www.ebi.ac.uk/Tools/msa/clustalo/) (Li et al. 2015). The same procedure was performed for sequences downstream of the LTR integration site. Also, we searched for patterns on those sequences which could suggest an integration preference.

Results and Discussion

Gene prediction and functional characterization

To perform *Augustus* software training, we searched on National Center for Biotechnology Information (NCBI) - Taxonomy for the cataloged species closest to *H. vastatrix*, which had described their genetic structures. The species close to *H. vastatrix* available on the database are related to it on level of order, more precisely they are classified as *Puccinialis*. Of the 14 available genomes from *Puccinialis* only 5 have their genetic structures published. The species are *P. graminis, P. sorghi, P. triiticina, P. striiformis* and *Melampsora larici-populina.*

In order to decide which of those species are closest to *Hemileia*, a local alignment (BLAST) was performed. Protein sequences from those species were alignment against *H. vastatrix* genome. To filter the results, we used two different cut-off scores, the first one being 40% identity and 70% coverage and the second 40% identity and 40% coverage. Table 1 shows the number of orthologous genes found on *H. vastatrix*.

Table 1 - Results from alignment between *H. vastatrix* genome and five other species

		70% de cobertura		**40 % de cobertura**	
Species name	Total number of proteins	Number of orthologous on *H. vastatrix*	Percentage of shared proteins	Number of orthologous on *H. vastatrix*	Percentage of shared proteins
Melampsora larici-populina	16,372	1,360	8.31	3,175	19.39
Puccinia graminis	15,979	963	6.03	2,796	17.50
Puccinia sorghi	21,078	1,645	7.80	4,173	19.80
Puccinia striiformis	20,502	1,373	6.70	3,499	17.07
Puccinia triticina	15,685	1,232	7.85	3,164	20.17

M. larici-populina and *P. triticina* held a higher number of proteins shared with *H. vastatrix*. For 70% coverage, *M. larici-populina* has a higher percentage of shared proteins with *H. vastatrix*, being 8.31% and for 40% coverage *P. triticina* showed a higher percentage of shared proteins with the pathogen, being 20.17%.

In this way, *M. larici-populina* was selected as a model for the prediction of proteins on *H. vastatrix* genome by the software *Scipio* (Keller et al. 2008), using its protein sequences available on NCBI. The output files were generated but no genes were predicted. Thus, the second closest species, *P. triticina* was chosen and at the end of the analysis the output files were created. Only 1,482 genes were predicted, which is a smaller number than expected when compared with the average number of genes on Basidiomycota phylum, which is around 15,000 genes (Mohanta & Bae 2015).

Looking for a more adequate analysis, gene prediction were done using ESTs of species available in *Augustus* software *v3.2.2* (Stanke et al. 2004) closest to *H. vastatrix*. *Ustilago maydis* is the closest specie available on *Augustus* which belongs to the same phylum as *H. vastatrix*, this being Basidiomycota. The use of *U. maydis* can also be justified once the fungus has a well-annotated genome and also is considered one of the best model to study host-pathogen interactions (Sonah et al. 2016). When using *U. maydis* as model to train *Augustus* software, 6,516 protein-coding genes were identified. Previous study predicted 6,500 protein-coding genes present on *U. maydis* (Duplessis et al. 2011), which indicates that the gene prediction was successful, since *U. maydis* is well-annotated genome (Sonah et al. 2016). However, the same study predicted 16,399 and 17,773 protein-coding genes in *M. larici-populina* and *P. graminis.* This lower number of genes found in *H. vastatrix* genome, when comparing with other rust genomes, could be explained by fact that the *H. vastatrix* assembled genome is still fragmented and displays about 58,000 contigs.

In the interest of validate gene prediction process, the software *CEGMA* was used, which searches for a set of 248 conserved genes on eukaryotic genomes. The results showed that 239 (96.37%) of those conserved eukaryotic genes are present in the annotated genome of *H. vastatrix*, indicating a high level of integrity during the assembly and gene prediction process of this genome (Parra et al. 2007). Note that despite the small number of gene-coding proteins, when comparing with results found for *P. graminis* and M. *larici-populina*, the assembled genome displays high level of integrity.

Molecular function categorization

Aiming the functional categorization analyzes of the 6,516 gene coding proteins found on *H. vastatrix* genome, the tool Goanna was used. The proteins were categorized by their molecular function in 39 distinct categories (Figure 1). Some of the proteins were unable to be characterized. DNA binding is one of the most abundant categories and play important rule in DNA repair (Duplessis et al. 2011). It is important to highlight the abundance of proteins related with kinases activity. Protein kinases plays an important role in pathogen development and adaptation to different environmental (Cristancho et al. 2014) and could be contributing to pathogen variability.

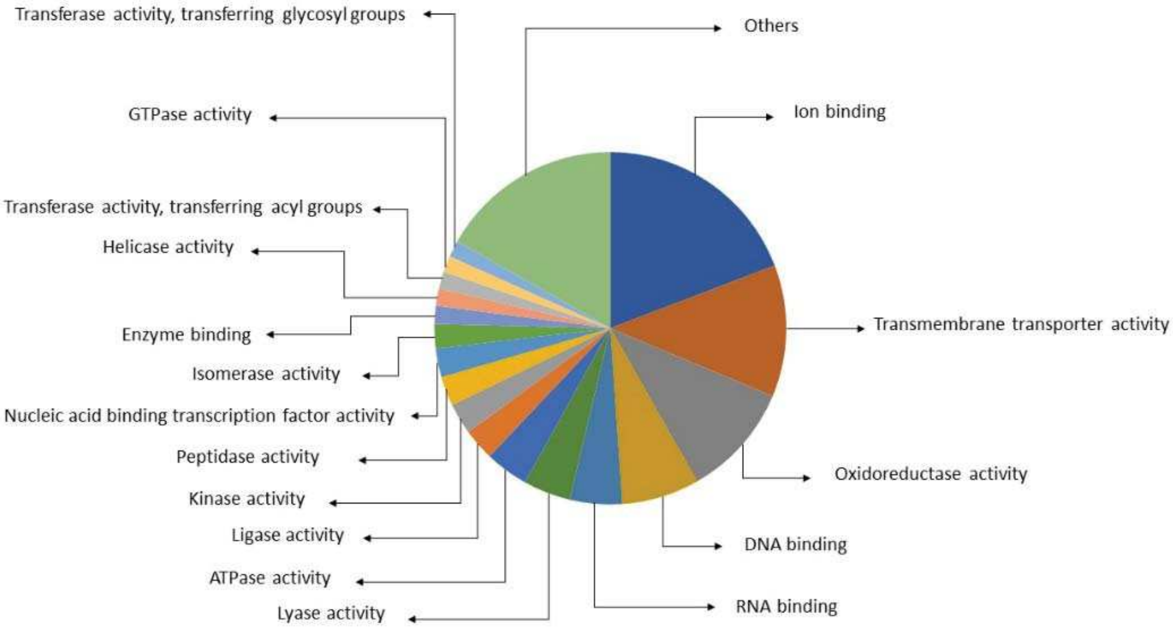

Figure 1. Funtional categorization of *H. vastatrix* proteins

Figure 2 shows functional categorization based on its biological process. The proteins were categorized on 69 distinct functions, being cellular nitrogen compound metabolic process the most abundant function. A considerable amount of the gene-coding proteins is related with stress response. This genetic feature could be contributing to increase *H. vastatrix* pathogenicity. Also, the integration of transposable elements close to genes which play a role on stress response process could be the key to better understand transposons evolutionary potential.

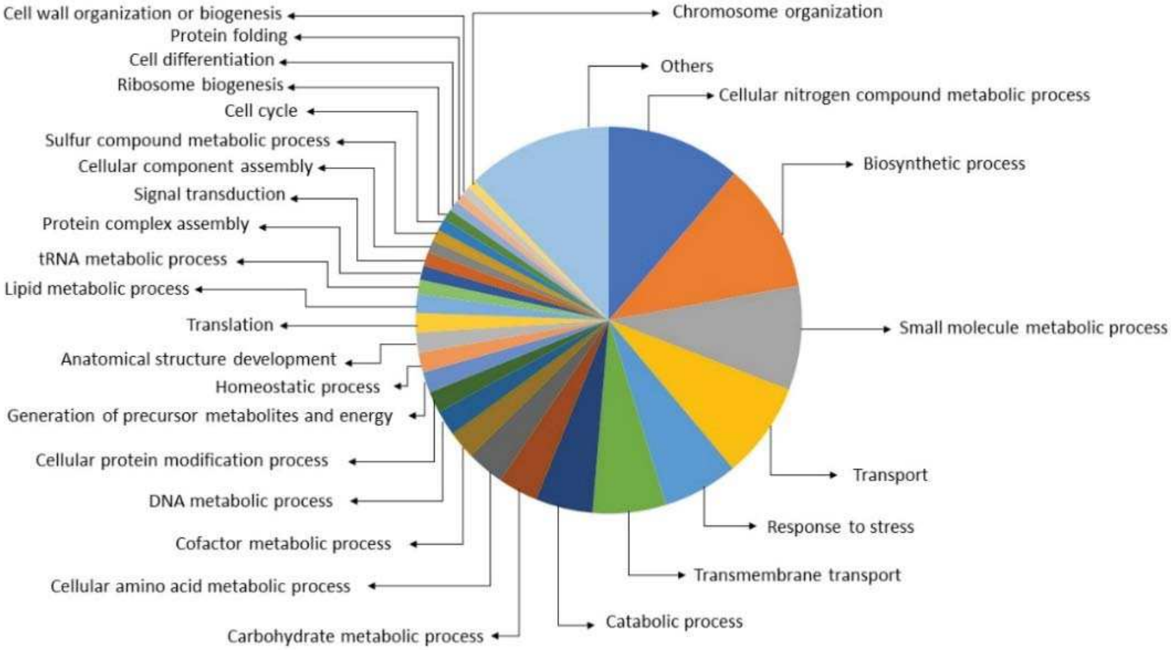

Figure 2. Funtional categorization of *H. vastatrix* proteins

LTR retrotransposons prediction

LRT-finder software (Xu & Wang 2007) available online was able to predict a total of 1,109 LTR (Long terminal repeats) retrotransposons. The LTR retrotransposons were found in different contigs with a density ranging from 1 to 12 LTR per contig (Figure 3). Most contigs showed no retrotransposons.

Previous studies showed that yeast retrotransposons have a mechanism for site selection integration. The elements insert on regions of genome which would tolerate integration without causing negative effects (Bushman 2003). Another study shows that transposon integration chemistry, in mammals, share many similarities with retroviral integration. However, the mechanism to select a target site is distinct among the genome (Yant et al. 2005). Thus, the fact that many LTR retrotransposons were found in a few contigs suggests that insertion and fixation of them could be not random along *H. vastatrix* genome.

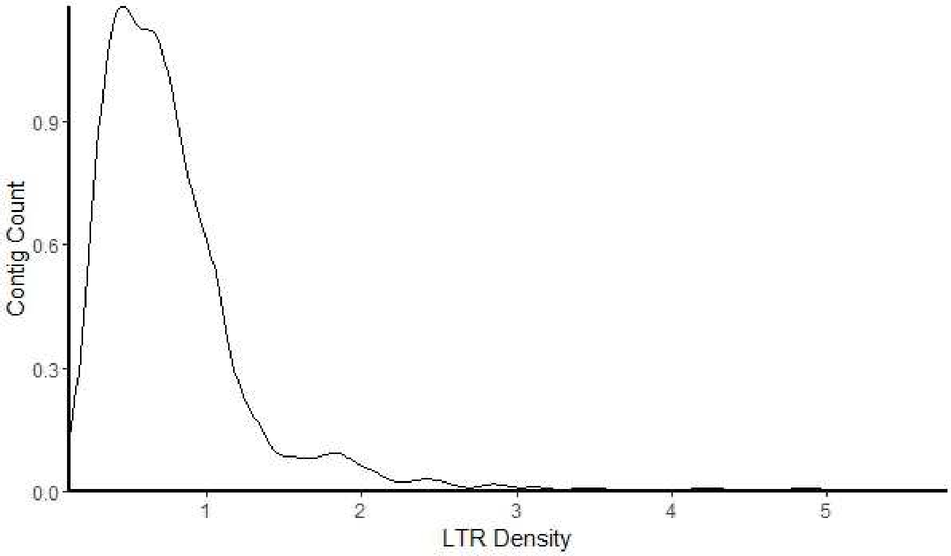

Figure 3 –LTR density on *Hemileia vastatrix* genome

To analyze the similarities between all LTR retrotransposon sequences identified in *H. vastatrix* genome, the sequences were aligned using *Clustal Omega* software (Li et al. 2015). After, we calculated a distance matrix between all LTR retrotransposons using the results obtained from the alignment performed by *Clustal Omega.* The distance matrix results were clustered using *R* software (Figure 4). Generally, LTR retrotransposons present in fungi can be classified into one of two different groups, the *copia* or *gypsy* superfamily, based on their reverse transcriptase sequence and other structural features (Goodwin & Poulter 2000). Our results suggest that the sequences of LTR retrotransposons found in *H. vastatrix* genome are conserved and could be clustered into three major groups (Figure 4). However, our analysis compared the complete sequence of all LTR retrotransposon and not just the structural features, which can explain the additional group found.

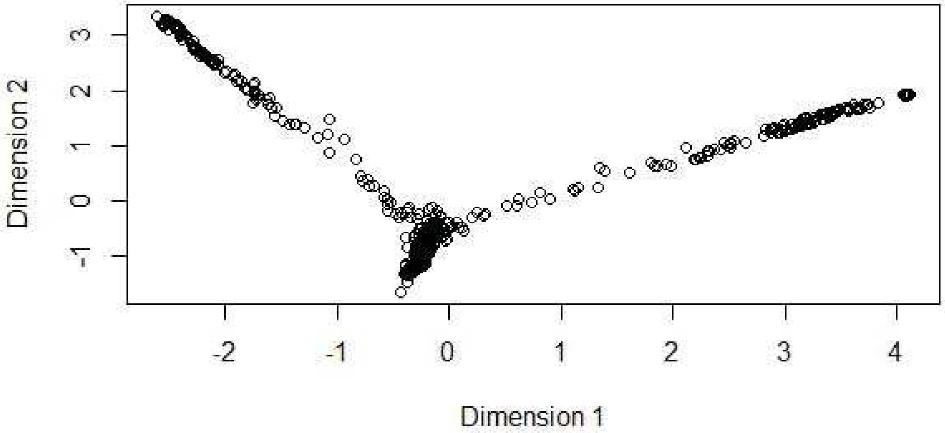

Figure 4 –Clustering of distance matrix data calculated between all LTR retrotransposons sequences extracted from *Hemileia vastatrix* genome

Integration site analyses suggest preference for integration

To analyze the LTR retrotransposon integration site, a thousand base pairs sequence upstream and downstream of its integration site were analyzed using custom Perl scripts. Matrix distance were calculated by *Clustal Omega* software and the results were clustered using R. Clustering of distance matrix data (Figure 5) calculated between the sequences upstream and downstream of the integration site showed that the integration site is very conserved.

Given the found results, we decided to analyze nucleotide composition of those sequences. Comparison between upstream and downstream sequences of retrotransposons site (Figure 6) showed that there is nucleotide conservation on both LTR integration sites, since for all the positions within the thousand base pairs sequence extracted, all nucleotides was more frequent than expected (25%). Thus, this data could suggest preference for site integration of retrotransposons.

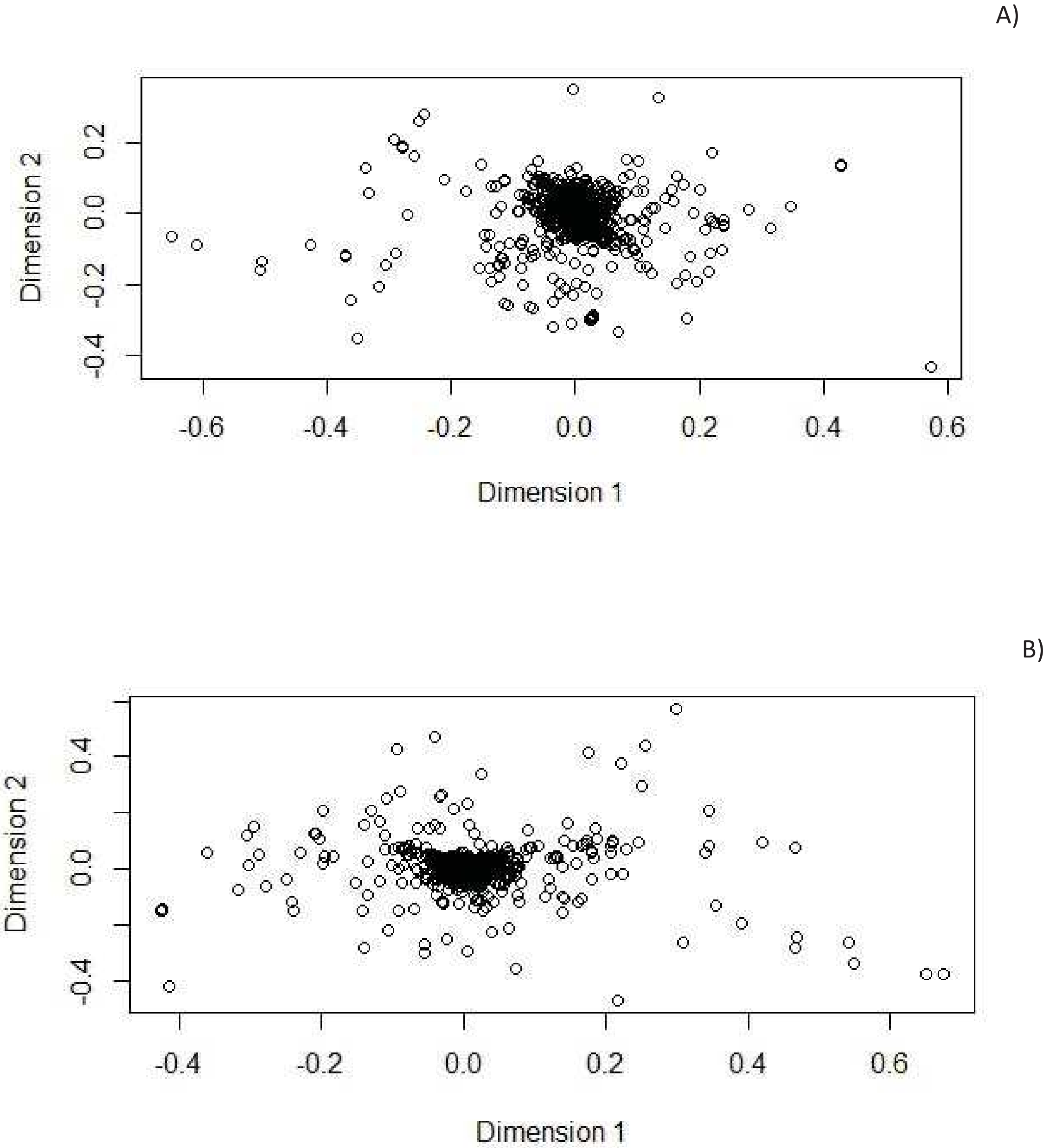

Figure 5 – Clustering of distance matrix data of the sequences upstream of LTR integration site B) Clustering of distance matrix data of the sequences downstream of LTR integration site

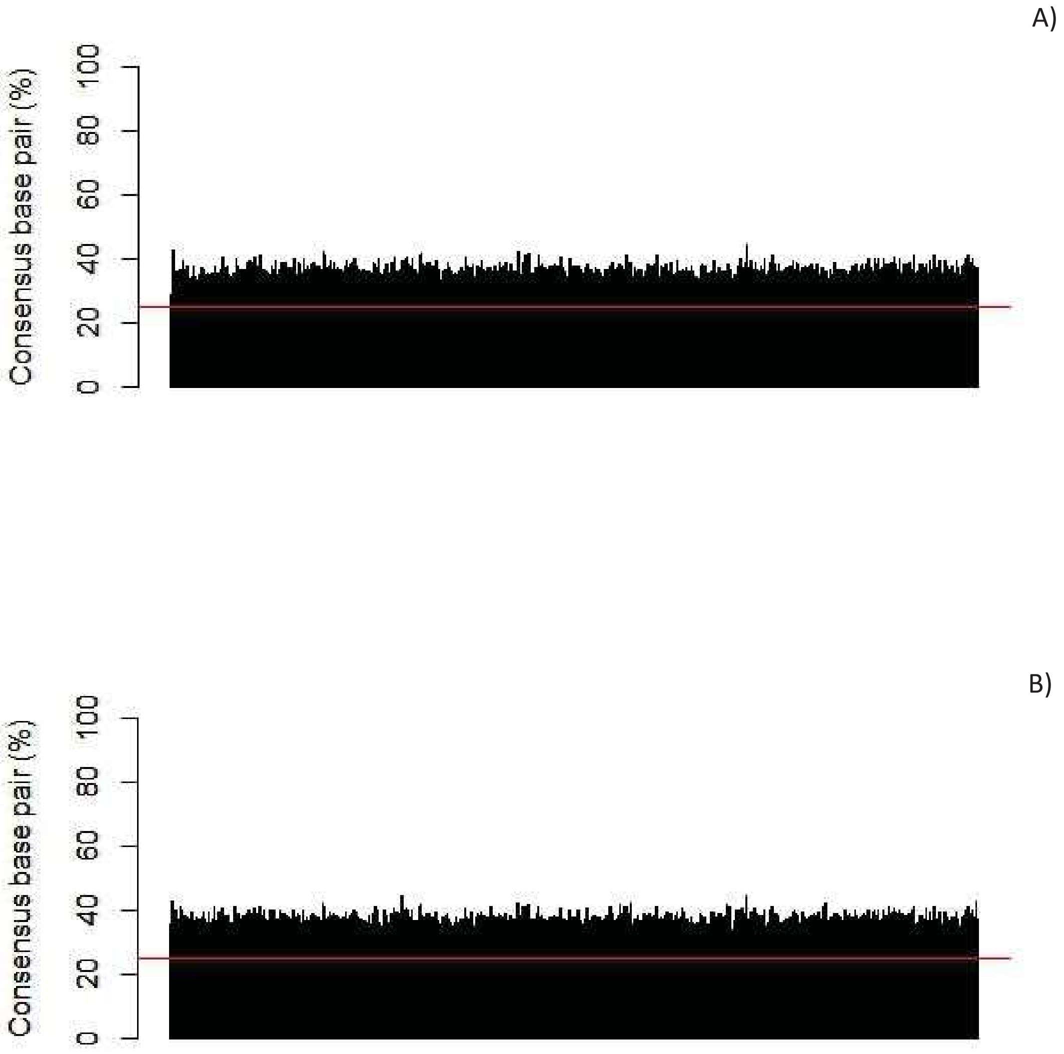

Figure 6 – A) Consensus base pair frequency upstream of LTR integration site B) Consensus base pair frequency downstream of LTR integration site

We looked the nucleotide composition of the consensus sequence upstream and downstream of LTR retrotransposon integration site (Figure 7). Custom Perl scrips was used to calculate the percentage of each nucleotide on the consensus sequence. The results showed that the integration site is a AT-rich region, which supported the results for such preference for site integration. The analyses showed that proportion of A and T at the consensus sequence upstream and downstream are the same, even though the sequences are different.

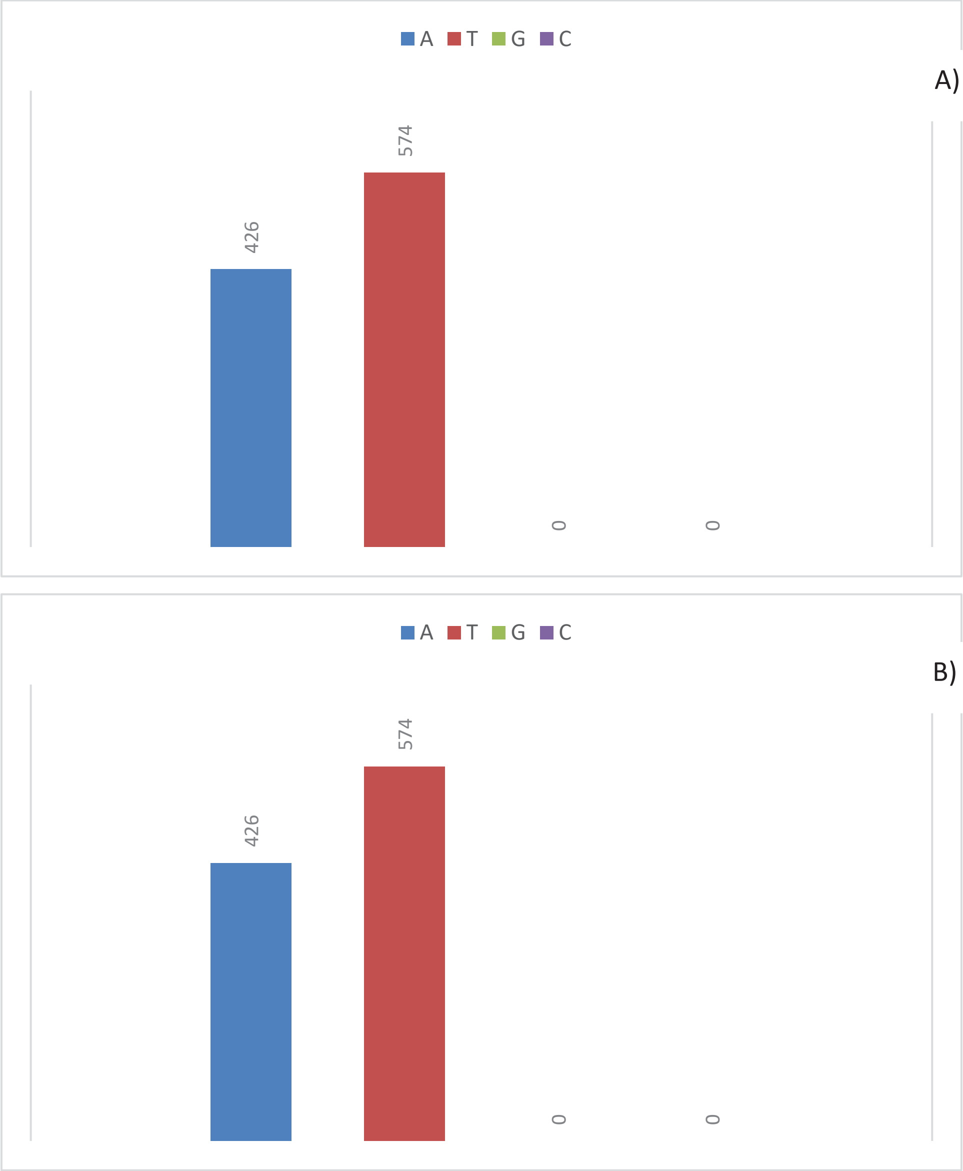

Figure 7 – A) Nucleotide composition of the consensus sequence upstream of LTR integration site B) Nucleotide composition of the consensus sequence downstream of LTR integration site

AT-rich regions are more easily opened for LTR retrotransposon integration when compared with CG-rich regions. Also, AT-rich regions may be involved with gene promoters or enhancers. Another event responsible for AT-rich regions is RIP (repeat-induced point mutation), which is a defense mechanism present in fungi that silence transposable elements movement. Studies have been shown that AT-rich regions are formed by transposable elements invasion followed by mutation by RIP (Testa et al. 2016). In this way, the prominent level of AT on the sequences upstream and downstream of LTR retrotransposons could be an evidence of mutation triggered by RIP. It is important to emphasize that RIP occurs prior to meiosis. Thus, the found results could support the cryptosexuality evidence found on *H. vastatrix* (Carvalho et al. 2011).

Gene and LTR distance analyses

We also looked for LTR retrotransposons which could be inserting itself close and on coding regions. We analyzed, using both genes and LTR retrotransposons coordinates, if any predicted genes were overlapping with LTR retrotransposons sequences. Table 2 show that seven predicted genes are overlapping with LTRs.

Table 2 - Predicted genes which overlapped with LTR retrotransposons

Contig ID	**Gene ID**	**Match (Genbank ID)**
4510	g477	XP_007412705.1
18706	g1105	Mutator protein (ABE02631.1)
20563	g1235	Putative gag-pol precursor (AAT75246.1)
24681	g1522	Putative retrotransposon protein (AAX96316.1)
28554	g1764	No hit
28661	g1767	Transposase (ABA97626.1) / Chromossom segregation protein SMC (TIGR02169)
43013	g2380	Putative retrotransposon protein (ABA98116.1)

Many of the genes found overlapping with LTR retrotransposons sequences are hallmarks of retrotransposons structure. Those proteins are responsible for reverse transcription, transposition and integration. However, two proteins should be highlighted. The proteins g1105 showed significant similarity with ABE02631.1, also known as mutator protein (Knox et al. 2010). Mutator transposons are the most mutagenic and active plant transposon discovered. These transposons can increase mutation rates 50 times (Lisch 2002) and could be contributing to increase of *H. vastatrix* genetic variability. Also, it was not found a hit for g1764 protein. Many proteins which plays role on host resistance is believed to be specie specific. Thus, the detailed study of both proteins is important to better understand such elevated levels of genetic variability.

We calculated the distance between LTR retrotransposons and predicted genes which were present on the same contig (Figure 8). We found 65 proteins located on the same contig as LTR retrotransposons. We also performed the functional categorization of those 65 genes based on assignment of Gene Orthology (GO) using AgBase (Mccarthy et al. 2006). None of those proteins could be classified once there were no hits found. This result suggests that those proteins could be specie specific and thus could play an important rule on pathogen virulence. We also found that most of the genes were located around 5000 base pairs distant of LTR retrotransposons, that is, when we move away from a gene there is a possibility to find a LTR retrotransposon at 5000 base pairs of it. This integration pattern was also observed on species such yeast and human. Genomic analysis demonstrated that integration of transposons on yeast and human genome are preferable and occurs in a specific distance of a given genetic structure (Bushman 2003).

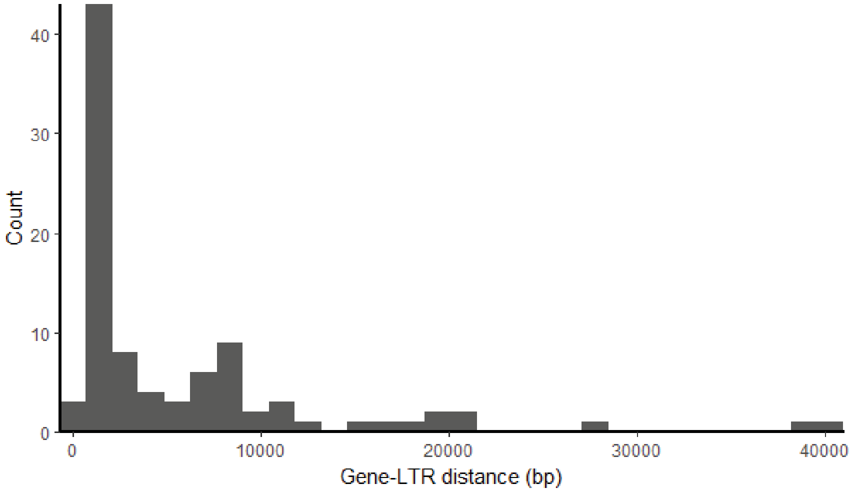

Figure 8 – Distance between LTR retrotransposons and genes present on the same contig

Proteins related with pathogenicity, in general, are secreted proteins. To check if any of those 65 proteins could potentially be an effector, we first looked for the presence of signal peptide, using Signal P 4.0 online tool. Three of the 65 proteins showed presence of signal peptide. Although the presence of signal peptide suggested that the protein is secreted, we used Target P online tool and only two of those three is part of secretory pathway. Thus, these two proteins have potential to be effectors and once these proteins are close to transposable elements, they could be affected by them. Transposable elements contribute with important sites for ectopic recombination (Goldfarb et al. 2016). Thus, if such type of recombination occurs, that could be a major source of variability, resulting perhaps in increased fungus pathogenicity. The presence of LTR retrotransposons close to those proteins could be regulating gene expression and consequently leading the pathogen overcome host resistance. Thus, further studies to will check if those proteins are up or downregulated will help us to better understand if they could be leading to appearance of new races of *H. vastatrix.* We will also investigate other ways in which the retrotransposons could be affecting gene expression such as interference in promoter and regulatory regions.

Conclusion

The found data suggest that insertion of LTR retrotransposons in *H. vastatrix* genome could be targeting its integration site based on nucleotide composition. We identified 6,516 gene coding proteins, some of them close to LTR retrotransposons integration site, that could be affected by TEs transposition. Moreover, we identified LTR retrotransposons located close to coding regions, which could be contributing with gene expression modulation and hence affecting pathogen variability.

References

Amselem, J., Lebrun, M. & Quesneville, H., 2015. Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. *BMC Genomics*, 16(141), pp.1–14.

Avelino, J. et al., 2015. The coffee rust crises in Colombia and Central America (2008 – 2013): impacts , plausible causes and proposed solutions. *Food Security*, 7(2), pp.303–321.

Bushman, F.D., 2003. Targeting survival: Integration site selection by retroviruses and LTR-Retrotransposons. *Cell*, 115(2), pp.135–138.

Carvalho, C.R. et al., 2011. Cryptosexuality and the Genetic Diversity Paradox in Coffe Rust, Hemileia vastatrix. *PLOS one*, 6(11), pp.1–7.

Carvalho, G.M.A. et al., 2014. Coffee rust genome measured using flow cytometry : does size matter ? *Plant Pathology*, 63, pp.1022–1026.

Castanera, R. et al., 2016. Transposable Elements versus the Fungal Genome : Impact on Whole-Genome Architecture and Transcriptional Profiles. *PLOS Genetics*, pp.1–27.

Chalfoun, S.M. & Carvalho, V.L., 1999. Controle químico da ferrugem (hemileia vastatrix berk & br.) do cafeeiro através de diferentes esquemas de aplicação. *Pesquisa Agropecuária Brasileira*, 34(3), pp.363–367.

CONAB, 2016. Acompanhamento da safra brasileira.

Cristancho, M.A. et al., 2014. Annotation of a hybrid partial genome of the coffee rust (Hemileia vastatrix) contributes to the gene repertoire catalog of the Pucciniales. *Frontiers in Plant Science*, 5(October), pp.1–11.

Cristancho, M. & Escobar, C., 2008. Transferability of SSR markers from related Uredinales species to the coffee rust Hemileia vastatrix. *Genetics and Molecular Research*, 7(4), pp.1186–1192.

D'Hondt, L. et al., 2011. Applications of flow cytometry in plant pathology for genome size determination, detection and physiological status. *Molecular Plant Pathology*, 12(8), pp.815–828.

Duplessis, S. et al., 2011. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *PNAS*, 108(22), pp.9166–9171.

Feschotte, C., 2008. Transposable elements and the evolution of regulatory networks. *Nature Reviews - Genetics*, 9, pp.397–405.

Feschotte, C. & Pritham, E.J., 2007. DNA Transposons and the Evolution of Eukaryotic Genomes. *Annual Review of Genetics*, 41(35), pp.331–368.

Garçon, C.L.P. et al., 2004. Controle da Ferrugem do Cafeeiro com Base no Valor de Severidade. *Fitopatologia Brasileira*, 29(31), pp.489–491.

Goldfarb, M. et al., 2016. Evidence of ectopic recombination and a repeat-induced point (RIP) mutation in the genome of Sclerotinia sclerotiorum , the agent responsible for white mold. *Genetics and Molecular Biology*, 39(3), pp.426–430.

Goodwin, T.J.D. & Poulter, R.T.M., 2000. Multiple LTR-Retrotransposon Families in the Asexual Yeast Candida albicans. *Genome Research*, 10, pp.174–191.

ICO, 2016. World Coffee Production. *International Coffee Production*.

Keller, O. et al., 2008. Scipio : Using protein sequences to determine the precise exon / intron structures of genes and their orthologs in closely related species. *BMC Bioinformatics*, 12, pp.1–12.

Knox, A.K. et al., 2010. CBF gene copy number variation at Frost Resistance - 2 is associated with levels of freezing tolerance in temperate-climate cereals. *Theoretical and Applied Genetics*, 121, pp.21–35.

Kushalappa, A.C., 1989. Advances in coffee rust research. *Annual Review of Phytopathology*, 27(57), pp.503–531.

Li, W. et al., 2015. The EMBL-EBI bioinformatics web and programmatic tools framework. *Nucleic Acids Research*, 43(April), pp.580–584.

Lisch, D., 2002. Mutator transposons. *Trends in Plant Science*, 7(11), pp.498–504.

Mccarthy, F.M. et al., 2006. AgBase : a functional genomics resource for agriculture. *BMC Genomics*, 13, pp.1–13.

Mendes, T.A.O. et al., 2013. Repeat-Enriched Proteins Are Related to Host Cell Invasion and Immune Evasion in Parasitic Protozoa. *Molecular Biology and Evolution*, 30(4), pp.951–963.

Mohanta, T.K. & Bae, H., 2015. The diversity of fungal genome. *Biological Procedures Online*, 17(8), pp.1–9.

Parra, G., Bradnam, K. & Korf, I., 2007. Genome analysis CEGMA : a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9), pp.1061–1067.

Sonah, H., Deshmukh, R.K. & Bélanger, R.R., 2016. Computational Prediction of Effector Proteins in Fungi : Opportunities and Challenges. *Frontiers in Plant Science*, 7(February), pp.1–14.

Stanke, M. et al., 2004. AUGUSTUS : a web server for gene finding in eukaryotes. *Nucleic Acids Research*, 32, pp.309–312.

Talhinhas, P. et al., 2017. The coffee leaf rust pathogen Hemileia vastatrix : one and a half centuries around the tropics. *Molecular Plant Pathology*, pp.1–13.

Tavares, S. et al., 2014. Genome size analyses of Pucciniales reveal the largest fungal genomes. *Frontiers in Plant Science*, 5(422), pp.1–11.

Testa, A.C., Oliver, R.P. & Hane, J.K., 2016. OcculterCut : A Comprehensive Survey of AT-Rich Regions in Fungal Genomes. *Genome Biology and Evolution*, 8(6), pp.2044–2064.

Varzea, V.M.P. & Marques, D.V., 2005. Population variability of *Hemileia vastatrix* vs coffee durable resistence. In *Durable Resistence to Coffee Leaf Rust*. pp. 53–74.

Vega, F.E., Rosenquist, E. & Collins, W., 2003. Global project needed to tackle coffee crisis. *Nature*, 425(September), p.2003.

Van Der Vossen, H., Bertrand, B. & Charrier, A., 2015. Next generation variety development for sustainable production of arabica coffee (Coffea arabica L .): a review. *Euphytica*, (204), pp.243–256.

Waller, J.M., 1982. Coffee rust epidemiology and control. *Crop Protecrion*, 1(4), pp.385–404.

Wicker, T. et al., 2007. A unified classification system for eukaryotic transposable elements. *Nature reviews. Genetics*, 8(12), pp.973–982.

Wostemeyer, J. & Kreibich, A., 2002. Repetitive DNA elements in fungi (Mycota) : impact on genomic architecture and evolution. *Current Genetics*, 41, pp.189–198.

Xu, Z. & Wang, H., 2007. LTR _ FINDER : an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research*, 35, pp.265–268.

Yant, S.R. et al., 2005. High-Resolution Genome-Wide Mapping of Transposon Integration in Mammals. *Molecular and Cellular Biology*, 25(6), pp.2085–2094.

Zambolim, L., 2016. Current status and management of coffee leaf rust in Brazil. *Tropical Plant Pathology*, 41, pp.1–8.