frontiers in
PLANT SCIENCE

ORIGINAL RESEARCH ARTICLE
published: 31 October 2014
doi: 10.3389/fpls.2014.00594

Annotation of a hybrid partial genome of the coffee rust (*Hemileia vastatrix*) contributes to the gene repertoire catalog of the Pucciniales

Marco A. Cristancho[1]*, David Octavio Botero-Rozo[1,2], William Giraldo[1], Javier Tabima[1,2], Diego Mauricio Riaño-Pachón[2†], Carolina Escobar[1], Yomara Rozo[1], Luis F. Rivera[1], Andrés Durán[1], Silvia Restrepo[2], Tamar Eilam[3], Yehoshua Anikster[3] and Alvaro L. Gaitán[1]

[1] Plant Pathology, National Center for Coffee Research – CENICAFÉ, Chinchiná, Colombia
[2] Departamento de Ciencias Biológicas, Universidad de los Andes, Bogotá, Colombia
[3] Institute for Cereal Crops Improvement, Tel Aviv University, Tel Aviv, Israel

Edited by:
Sébastien Duplessis, National Institute of Agronomic Research, France

Reviewed by:
Guus Bakkeren, Agriculture and Agri-Food Canada, Canada
Diana Fernandez, Institut de Recherche Pour le Développement, France

***Correspondence:**
Marco A. Cristancho, Department of Plant Pathology, National Center for Coffee Research – CENICAFÉ, Km 4 vía a Manizales, Chinchiná 2427, Colombia
e-mail: marco.cristancho@cafedecolombia.com

† Present address:
Diego Mauricio Riaño-Pachón, Labóratorio Nacional de Ciência e Tecnologia do Bioetanol, Centro Nacional de Pesquisa em Energia e Materiais, Campinas, São Paulo, Brazil

Coffee leaf rust caused by the fungus *Hemileia vastatrix* is the most damaging disease to coffee worldwide. The pathogen has recently appeared in multiple outbreaks in coffee producing countries resulting in significant yield losses and increases in costs related to its control. New races/isolates are constantly emerging as evidenced by the presence of the fungus in plants that were previously resistant. Genomic studies are opening new avenues for the study of the evolution of pathogens, the detailed description of plant-pathogen interactions and the development of molecular techniques for the identification of individual isolates. For this purpose we sequenced 8 different *H. vastatrix* isolates using NGS technologies and gathered partial genome assemblies due to the large repetitive content in the coffee rust hybrid genome; 74.4% of the assembled contigs harbor repetitive sequences. A hybrid assembly of 333 Mb was built based on the 8 isolates; this assembly was used for subsequent analyses. Analysis of the conserved gene space showed that the hybrid *H. vastatrix* genome, though highly fragmented, had a satisfactory level of completion with 91.94% of core protein-coding orthologous genes present. RNA-Seq from urediniospores was used to guide the de novo annotation of the *H. vastatrix* gene complement. In total, 14,445 genes organized in 3921 families were uncovered; a considerable proportion of the predicted proteins (73.8%) were homologous to other Pucciniales species genomes. Several gene families related to the fungal lifestyle were identified, particularly 483 predicted secreted proteins that represent candidate effector genes and will provide interesting hints to decipher virulence in the coffee rust fungus. The genome sequence of Hva will serve as a template to understand the molecular mechanisms used by this fungus to attack the coffee plant, to study the diversity of this species and for the development of molecular markers to distinguish races/isolates.

Keywords: genome, coffee rust, coffee, plant pathogens diversity, RNA-seq, genetic variants

INTRODUCTION

Coffee rust caused by the fungus *Hemileia vastatrix* (Hva) leads to widespread damage to crops worldwide. The disease develops polycyclic epidemics in a season, which means there is an overlapping succession of infection cycles. Under tropical conditions and in semi-perennial plants, such as coffee, Hva poses a permanent threat for producers. In the absence of control methods, the pathogen has been reported to cause losses of up to 30% in susceptible varieties of the species *Coffea arabica* during mild epidemics (Monaco, 1977; Rivillas et al., 2011). The fungus, which is a biotroph that targets *Coffea* as the single known host genus, spread from Africa and was responsible for the collapse of coffee production in India and Ceylon by the mid XIX century. It arrived in America in 1970, and since then, it has quickly disseminated to all the other coffee-producing areas of the continent. In Colombia, Hva was reported for the first time in 1983 in the central coffee-producing zone of the country (Leguizamón et al., 1984). The presence of Hva in almost every coffee plantation in the world has been one of the main drivers for plant breeders to release rust-resistant varieties. Recent outbreaks of the disease have affected major areas of coffee production in Colombia (Rozo et al., 2012) and Central America (Cressey, 2013), and the evidence linked these new epidemics to changes in weather patterns, including rainfall distribution and quantity (Cristancho et al., 2012).

Multiple Hva races have been reported throughout the world (Rodrigues et al., 1975; Carvalho et al., 1987), and studies from the CIFC (Coffee Rust Research Center) in Portugal have identified over 30 races of the pathogen using a series of more than 40 differential coffee genotypes (Rodrigues et al., 1993). Historically, race II has been predominant in most countries, and it attacks all cultivated varieties of the species *C. arabica* that have not been bred for disease resistance (Rodrigues et al., 1975). In addition to race II, 6 other physiological races have been identified in Colombia using a set of differential plants developed at CIFC (Oeiras, Portugal), attacking some lines of the resistant cultivars (Castillo and Leguizamón, 1992; Gil and Ocampo, 1998; Alvarado and Moreno, 2005; Rozo et al., 2012). At least 10 more isolates not differentiated by CIFC differential plants, remain to be characterized in Colombia, and several other unknown isolates have also been detected elsewhere (Gouveia et al., 2005).

The emergence of new races and more aggressive isolates in plant pathogens threaten agriculture worldwide as recently observed with the wheat stem rust fungus and the new epidemics of coffee rust, which clearly indicate that further detailed studies and continuous monitoring are needed to improve integrated disease management strategies that mitigate their destructive effect (Aime et al., 2006). Being obligate pathogens, the study of rust fungi biology is particularly challenging and needs substantial investments given the fact that a large set of differential plants have to be employed for the classification of races. The development of novel tools for the identification of isolates is critical to study the biology of these major plant pathogens and genomics might offer such tools. Differential Hva genes expressed in urediniospore, appressoria, and haustoria have already been identified (Fernandez et al., 2012; Talhinhas et al., 2014).

Genomic studies of plant pathogens have provided insights into their evolution, the mechanisms that generate genetic variability and the repertoire of genes that are involved in pathogenesis. These studies have shown that rust fungi exhibit very large genome sizes [*Melampsora lini* = Mli (Nemri et al., 2014) is the largest fungal genome so far] compared to other fungi, containing very large numbers of genes, over 16,000 in most cases, compared to other fungi groups such as Ustilaginomycotina [*U. maydis* = 6786 protein coding genes (Schirawski et al., 2010)] or other non-rust Pucciniomycotina such as *Mixia osmundae* (Toome et al., 2014). Rusts also show a large content in transposable elements (i.e., nearly 50%, in the genomes analyzed so far, Duplessis et al., 2014). All these features indicate that rust genomes in general are complex to sequence and assemble due to the repetitive content and large genome size.

The discovery of predicted secreted virulence determinants in plant pathogens has also been possible through genomic analysis. Secreted proteins have been linked to the virulence of plant-pathogenic fungi (Spanu, 2012) and many have been predicted in *Melampsora* spp. (Joly et al., 2010; Hacquard et al., 2012), *P. striiformis* (Cantu et al., 2011, 2013), *P. graminis* (Duplessis et al., 2011), and *H. vastatrix* (Fernandez et al., 2012). Thus, the discovery of the secreted protein genes (Saunders et al., 2012) and the functional demonstration of their decisive role in the infection process help in unraveling previously unknown mechanisms of pathogenicity that operate in biotrophic fungi.

We have obtained genome and transcriptome sequences of the coffee rust fungus, and we expect these data to allow the identification of potential molecular markers for the study of rust isolates/races. The knowledge of the Hva genome and particularly of its secretome is a critical point for understanding the mechanisms used by the fungus during the colonization of coffee tissues and allows for comparisons of pathogenesis processes in other rust fungus-plant interactions. The chimeric genome assembly obtained was further used to define polymorphism between isolates and to analyse its basic contents such as the gene complement and TE families. The predicted proteome was additionally supported by a transcriptome analysis of Hva urediniospores. Within the predicted gene complement, a more precise analysis was performed on predicted secreted proteins, likely containing Hva candidate effectors.

RESULTS

NUCLEAR DNA CONTENT

The nuclear DNA content of 10 Hva samples was measured by Flow cytometry analysis (**Table 1**). Two groups could be distinguished among the 10 samples: one contains four samples that showed a lower content of 1.17–1.29 pg of DNA and a second with the six remaining samples showed a higher content of 1.55–1.76 pg of DNA per urediniospore (~30% more). Based on these results and compared to the genome size of *P. triticina* used as a control (135 Mb, Puccinia Group Genomes Database, Broad Institute), we estimate the genome size of Hva to be 243–324 Mb.

Table 1 | Nuclear DNA content of Hva urediniospore samples measured by FCM.

***Coffea* species and genotypes**	**Geographical location**	**DNA Content (pg)**	
		Group 1	**Group 2**
C. arabica var. Caturra	La Alcancía, Antioquia	–	1.63
C. arabica var. Caturra	El Cedral, Pereira, Risaralda	–	1.55
C. arabica var. Caturra	Santa María, Antioquia	1.29	–
C. arabica var. Caturra (Acc. 1421)	Chinchiná, Caldas	1.17	–
C. arabica BA-13	Chinchiná, Caldas	–	1.76
C. arabica × *C. canephora*—Timor Hybrid H-419/2	Chinchiná, Caldas	–	1.62
C. arabica × *C. canephora*—Timor Hybrid H-584	Chinchiná, Caldas	1.21	–
C. arabica var. Tipica	Chinchiná, Caldas	1.18	–
C. arabica var. Mundo Novo	Chinchiná, Caldas	–	1.70
C. liberica	Chinchiná, Caldas	–	1.70
	Mean	1.21	1.66
	Standard deviation	0.05	0.08
	CV%	5.87	6.41

Hva urediniospore samples were stained with Propidium Iodide for Flow Cytometry fluorescence measures following the protocol described by Eilam et al. (1994).

GENOME ASSEMBLIES

We sequenced the genomes of the following isolates: HvCat, Hv387, Hv949, HvDQ952, HvH179, HvH569, HvH701, and

HvMar; for isolate HvCat, Illumina and 454 sequencing were combined. We obtained a total of 412 million short-reads from Illumina and 5.8 million reads from 454. The fraction of reads that passed quality filtering was over 85% in all Hva Illumina sequenced samples but only 52% for the 454 Hva sequenced sample (Table S1, see methods section for filtering parameters). A hybrid 454-Illumina assembly was obtained, combining all genomic Hva sequences with the script clc_novo_assembly from the assembler suite CLC Assembly Cell v4.0.1 (CLC bio, Aarhus, Denmark); we have also performed separate assemblies of the genomes of each isolate. Unfortunately due to inherent characteristics and composition of the genomes (low GC content and richness in TE, detailed in the following sections), we were only able to obtain partial genome assemblies. In order to improve the overall genome assembly, our strategy was to generate an hybrid assembly that takes into account all reads obtained from the 8 isolates together to produce a unique sequence that is a chimera of the sequenced isolates. We were able to define a genome sequence of 333 Mb (129X sequencing depth) composed of 396,264 contigs and 302,466 scaffolds.

We assessed the completeness of the hybrid and individual genomes by running CEGMA with a set of 248 ultra-conserved Core Eukaryotic Genes (CEGs) (Parra et al., 2007) (**Table 2**). Statistics for the hybrid assembly are shown in **Table 3**; based on the Hva chimeric assembly data, and using Jellyfish (Marçais and Kingsford, 2011) to compute the number of distinct k-mers of different lengths and their relationship to coverage, we estimated the size of the Hva genome to be 333 Mb. Considering that the HvaHybrid genome sequence was the only one with a fairly good coverage of conserved core genes, we decided to use it as the reference genome for further analysis.

A total of 23.2% of the paired-end reads mapped in the same contig of the HvaHybrid genome assembled. Most unpaired reads (66.8%) matched two different contigs (useful for scaffolding, data not shown). Several contigs displayed coverage greater than 100X, but most of the contigs exhibited low coverage (**Figure 1**); however, over-coverage was pronounced in the short contigs (**Figure 2**). We also illustrated the range of coverage of the contigs and its association to the contigs size (**Figure 3**). The largest contig, Hvcontig_23458 (85 Kbp) showed good coverage. However, the next two contigs in size, Hvcontig_171 (45 kbp) and Hvcontig_161 (71 kbp), showed over-coverage and belong to the Hva mitochondria (see below). Bacterial contamination was very low representing less than 1% of the sequences (**Figure 4**).

Table 3 | Summary of the Hva genome hybrid assembly.

N° Contigs assembled		396,264
N° Scaffolds assembled		302,466
Total residues assembled		333,481,311
Length	Max	85,126
	Average	841.56
	N50	1,59
Reads	Total	336,649,188
	Unassembled	197,88,611
	Assembled	316,860,577
	Multihit	37,520,793
	Potential pairs	
	Paired	78,105,740
	Not Paired	255,469,308

The 454 and Illumina clean reads were assembled and the same reads were mapped against this set of assembled contigs using the software CLC Assembly Cell v4.0.1.

Table 2 | Statistics gathered from the genome assemblies of Hva individual isolates and the hybrid assembly.

Sample IDs	*Coffea* species and genotypes	Raw reads	Clean reads[a]	Contigs	Assembly size	Completeness[b]
HvHybrid[c]		412,417,464	359,076,496	396,264	333,258,024	91.94%
HvCat 454[d]		5,860,446				
HvCat Illumina	*C. arabica* var. Caturra	48,396,016	43,704,716	254,645	122,820,521	57.26%
Hv387	*C. canephora* CIl56	58,593,986	50,782,526	211,495	150,707,107	44.35%
Hv494	H89: *C. arabica* var. Bourbon resistant × *C. arabica* CaRCV3	55,326,774	47,738,686	211,728	138,293,025	35.08%
HvDQ952	F2 – *C. arabica* var. Caturra × HdT 1343 (*C. arabica* × *C. canephora*)	43,875,056	38,844,164	197,927	121,119,448	31.85%
HvH_179	H3101: (*C. arabica* CaCV1 × Hdt (*C. arabica* × *C. canephora*) 1343 574CV2) × *C. arabica* CtyR	49,025,718	42,033,780	203,770	131,574,289	31.05%
HvH_569	H3041: (*C. arabica* × HarrarR3) × HdT(*C. arabica* × *C. canephora*) 1343 Africa 1386	51,960,392	45,000,606	202,168	133,358,100	37.50%
HvH_701	H2094: (*C. arabica* MundoNovo) × F502 (*C. arabica* accession from Tanganica)	60,634,018	53,080,264	215,628	158,292,515	56.86%
HvMar_1	*C. arabica* var. Caturra	44,605,504	39,408,690	203,360	125,814,765	22.18%

[a] *Clean reads were obtained after quality trimming and removal of duplicates.*

[b] *Completeness of the genome was calculated running the software CEGMA with a set of 248 ultra-conserved Core Eukaryotic Genes (CEGs) (Parra et al., 2007).*

[c] *The hybrid assembly was generated by the combination of all short reads from the eight isolates.*

[d] *Eight and a half plates were sequenced with 454 technology.*

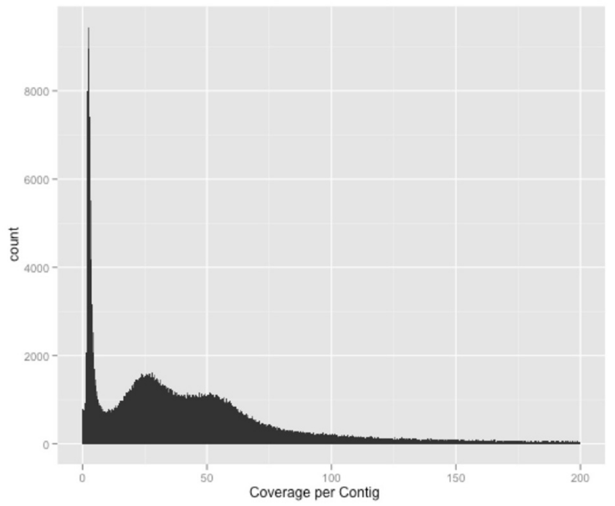

FIGURE 1 | Plot representing the reads coverage per contig in the HvHybrid assembly. Low read coverage was identified in most contigs.

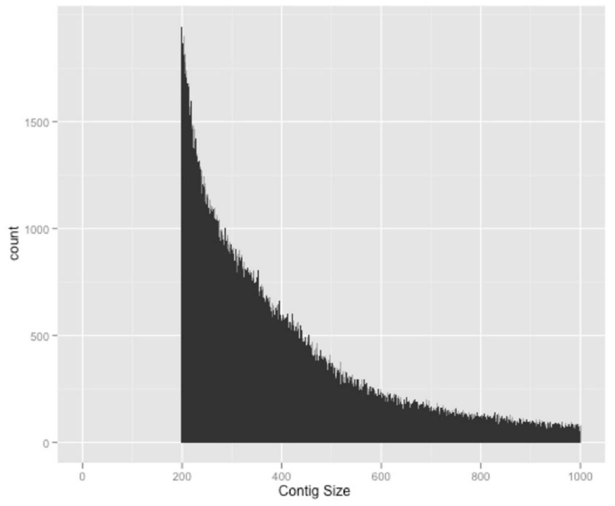

FIGURE 2 | Association between number of reads and size of contigs in the HvHybrid Assembly. Over-coverage was identified in short contigs.

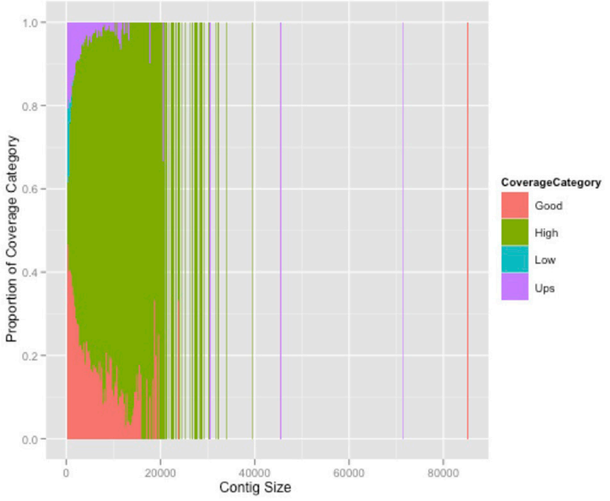

FIGURE 3 | Association between contig size and coverage. Contigs were clustered by color by running an R script: 0–5 low coverage, 5–45 good coverage, 45–100 high coverage and more than 100 over-coverage.

MITOCHONDRIAL GENOME ANNOTATION

We used the *P. graminis* f.sp. *tritici* (PGT) (Puccinia Group Genomes Database, Broad Institute) mitochondrial genome to find homologs in the *HvHybrid* assembly by BLAST (*e* = 1e-5). BLAST sequence similarities were pictured using the visualizing tool Circoletto (Darzentas, 2010). As shown in **Figure 5**, Hvcontig_161 (71,379 bp, %GC = 33%) and Hvcontig_171 (45,581 bp, %GC = 35%) cover over 70% of the mitochondrial genome of PGT. We also found that 7 contigs of the HvCat assembly entirely covered the mitochondrial genomes of PGT (79,748 bp) and the soybean rust *Phakopsora pachyrhizi* (31,825 bp; Stone et al., 2010) (results not shown). From this analysis we estimate that the Hva mitochondrial genome it is at least of the size of the PGT mitochondrial genome.

DE NOVO IDENTIFICATION OF TRANSPOSONS

The genomes of rust fungi sequenced to date all contain large numbers of transposable elements, which is a major problem for proper assembly (Duplessis et al., 2011; Zheng et al., 2013). A careful annotation of TEs was performed in the chimeric assembly and it showed that the Hva genome also contained a large proportion of repeated sequences. Interspersed repeats were identified in 74.4% of the Hva assembled hybrid contigs using the algorithm RepeatMasker. A similar fraction of repeats was identified in the individual assemblies (71–74%). A large proportion of repeats were classified as LTR elements (38.7%), a smaller proportion was classified as DNA elements (7.2%), only 2.1% were classified as LINEs, and 26.3% of repeats were unclassified.

The sequences were annotated for the presence of LTRs, direct repeats and inverted repeats as well as sequence similarities to repeat sequences from Pucciniales. Four novel retrotransposon families were identified in the Hva genome (Table S2).

PREDICTING PROTEIN-CODING GENES

In order to capture the gene space of the coffee rust genome, we performed a transcriptome analysis of freshly harvested urediniospores based on Illumina RNA-Seq. A total of 44,297–64,752 transcripts could be identified in the three RNA-seq based libraries with the program Trinity, with the HvCatNor normalized library holding the smallest number of genes (Table S3). We aligned those transcripts onto the chimeric and individual samples assemblies and showed that the normalized library contains the largest fraction of Hva expressed mapped genes (Table S4). The normalization approach decreases the prevalence of high abundance transcripts and equalizes transcript concentrations in a cDNA sample, thereby increasing the discovery of low abundance transcripts. The level of contamination of the Hva

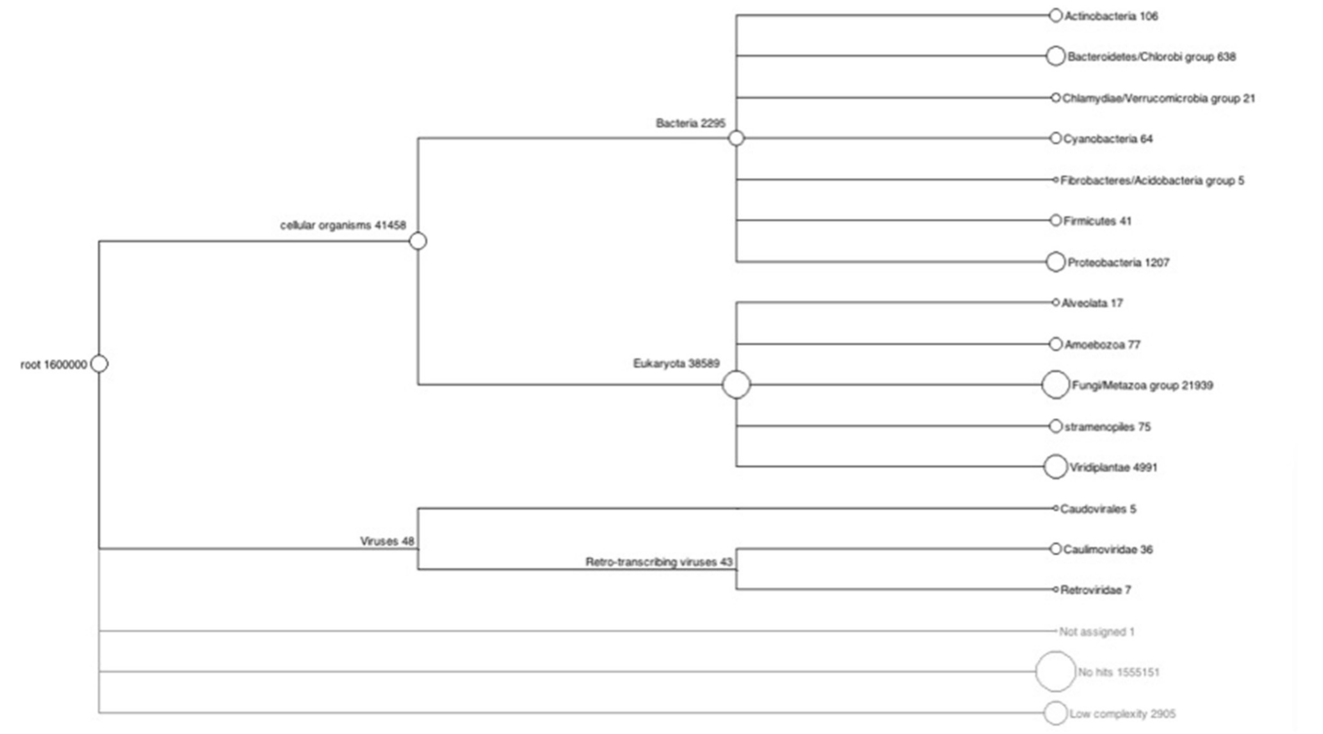

FIGURE 4 | Analysis of sequence contamination in the HvHybrid assembly. The HvHybrid assembled contigs were searched for contaminants with the program Megan4 (Huson et al., 2001). Each node is labeled by a taxon and the number of reads assigned to the taxon. The size of a node (circle size) is scaled logarithmically to represent the number of assigned reads. A low bacterial and plant sequence contamination was identified in the assembly.

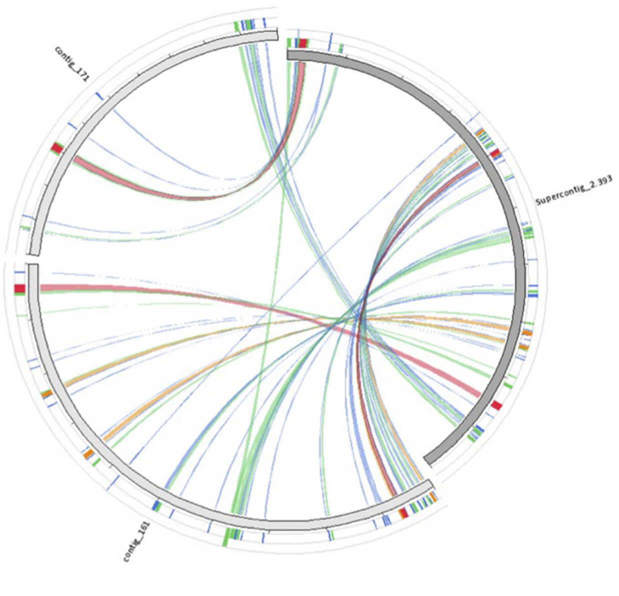

FIGURE 5 | Homology of Hva mitochondrial sequences to the mitochondrial genome of PGT. Clone_161 and clone_171 sequences from HvHybryd assembly were compared to the mitochondrial genome of *P. graminis* f.sp *tritici* (Supercontig_2.393; 79,2Kb) by running BLAST sequence similarities (e = 1e-5). Homologous regions are illustrated with Circoletto (Darzentas, 2010).

RNA-seq datasets with plant, bacterial and other contaminant sequences was moderate; the fraction of contamination for each sample was 13.6% for the normalized library HvCatNor, 12,3% for the HvH420_701 library, and 18,9% for the HvCat955 library (Table S5). We carried out homology annotation with BLAST against Hva germinating urediniospore transcripts dataset and other rusts predicted proteins datasets (**Table 4**). Our Hva urediniospores dataset is as expected very similar to the Hva germinating urediniospores transcripts identified by Talhinhas et al. (2014).

The predictions of the gene coding space was performed using TopHat (Trapnell et al., 2009) for mapping of RNA-seq data against the HvHybrid genome assembly and proteins were predicted with Augustus (Stanke and Waack, 2003; size filter = 70 amino-acids), using the RNA-Seq data as a guide. We identified a total of 21,345 contigs that matched the RNA-seqs and we predicted a total of 18,234 protein sequences with an average length of 1047 bp for the gene models. We identified 13,796 Hva protein homologs (73.86%) in the Pucciniales order (67,118 sequences) using blastp (e = 1e-3).

The total set of sequences was filtered to remove repeats identified in RepBase Release 17.01, and this resulted in a final set containing 14,445 predicted protein-coding gene sequences. Over 96% of this set of gene models was identified in the individual assemblies (Table S6). We explored this set of predicted proteins searching for KOGs in the NCBI Conserved Domain Database; 8458 Hva protein-coding genes having a KOG homolog were

Table 4 | Homology of Hva transcript datasets to Pucciniales predicted proteins.

		Hva urediniospore transcripts		
Hva samples[a]	**Assembled transcripts**	**HvCatNor**	**HvH420_701**	**HvCat955**
I				
Hva (gU)	4267	91.0% (3884)	93.1% (3973)	93.9% (4007)
H. vastatrix (Ap)	3627	59.2% (2147)	63.1% (2290)	63.8% (2315)
H. vastatrix (H)	4465	50.0% (2232)	49.1% (2202)	49.6% (2229)
Rust species	**Predicted proteins**			
II				
Pt	11,630	42.2% (18,703)	40.0% (22,240)	36.6% (23,705)
Pgt	15,979	43.8% (19,411)	41.5% (23,129)	37.8% (24,480)
Pst	22,815	43.5% (19,257)	40.4% (22,544)	36.9% (23,874)
Mlp	16,694	39.2% (17,385)	40.8% (22,765)	37.8% (21,302)

I. Homology sequence analysis between Hva transcriptomes datasets (this study) and Hva germinating urediniospores, appresoria, and haustoria transcript sequences (gU, Ap, H, Talhinhas et al., 2014) was performed with the program BLASTn and an $E = 1e^{-20}$. II. BLASTx sequence similarity analysis of H. vastatrix RNA-seq sequences and other rust predicted protein datasets ($E = 1e^{-3}$). Numbers represent fraction (%) of hits found.

[a] *Hva samples described in Fernandez et al. (2012).*

functionally annotated (Table S7). A total of 3921 gene families with 2–66 gene members were identified with OrthoMCL (Table S8); we also identified 2103 orphan genes.

SECRETOME ANNOTATION

We predicted 659 secreted proteins using PProwler and 775 secreted proteins with the SignalP algorithm. A total of 180 proteins in our Hva set presented homologs with the secreted proteins already predicted in Hva (Fernandez et al., 2012). A Venn diagram (**Figure 6**) showed shared and unique coincidences between the three sets of data, including 44 proteins extracted by comparison with the dataset predicted by Fernandez et al. (2012). Most of the secreted proteins predicted are organized in gene families and they were mapped with tblastx to at least one contig from each of the individual assemblies; a final set of 28 predicted proteins was obtained after filtering those belonging to the same gene family and they were functionally annotated with blastp against swissprot, RefSeq, Uniref100 and the non-redundant protein sequences databases (Table S9). Only five sequences did not have a homolog sequence with other Pucciniales fungi. Six sequences had a homolog already identified as a secreted protein in *M. larici-populina*. We did not identify in the genome of Hva a homolog of ps87 of *P. striiformis* f.sp. *tritici*, a conserved secreted protein in several fungal plant pathogens (Gu et al., 2011). However, an identical copy of the Hva RTP1 gene (GenBank: FR851895), a transferred protein belonging to the family of effectors in rusts (Pretsch et al., 2013) was identified, suggesting that some effectors are very well conserved between different rust species (Spanu, 2012).

We identified homologs of the predicted secreted proteins in all but one of the Hva individual assemblies; a homolog of protein KF018005 was not identified in the assembly of HvCat. The secreted proteins mapping to the individual assemblies showed that 6 proteins were identical in every Hva sample. The remaining 22 predicted proteins displayed polymorphism in at least two isolates (Table S10).

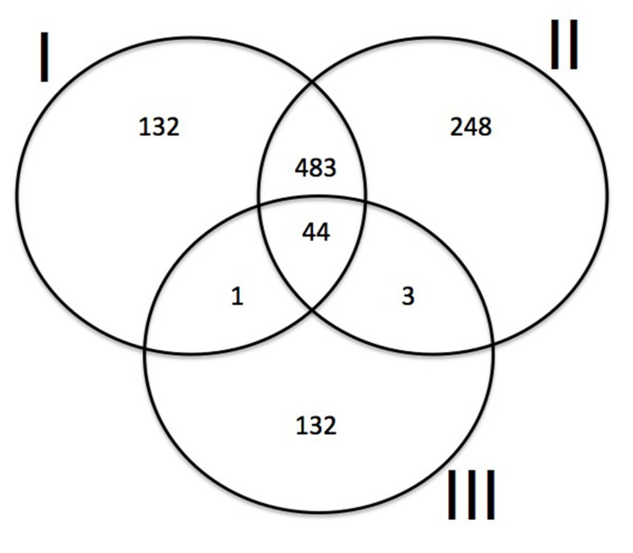

FIGURE 6 | Venn diagram showing the consensus set of Hva secreted predicted proteins. PProwler (0.9 probability cut-off) and SignalP were used to predict secreted proteins. A set of 14,445 putative proteins was used for classification into secreted and non-secreted proteins. The results were compared with Hva secreted proteins predicted by Fernandez et al. (2012). Hva secreted proteins predictions: **(I)** PProwler (this study). **(II)** SignalP (this study). **(III)** Secreted proteins predicted by Fernandez et al. (2012).

For Hva proteins KF018008, KF018015, KF018016, KF018020, KF018028 we did not find evidence of predicted homologs in other rusts, representing unique genes of the coffee-rust interaction not represented in other pathosystems. Interestingly, protein KF018020 had no homolog in any database and we could not detect polymorphisms of this protein-coding gene in the Hva individual assemblies, but the Hva genome holds 32 copies of this gene. Protein KF018028, represented by a single copy in the genome, is surprisingly the most diverse of the Hva secreted proteins. It is worth noting that it does not have homologs in any other organism.

PROTEIN KINASES (PKs)

Given the fact that PKs are involved in essential pathways related to development and adaptation to different environments (Miranda and Barton, 2007), we determined differences between PKs families present in pathogen and non-pathogen basidiomycetes. A total of 210 PKs were identified within the set of Hva predicted proteins using HMM3 (Eddy, 2011).

This set of sequences was compared against the predicted PKs of other Pucciniales showing that most protein kinases, including gene families coding for signal transduction pathways, are shared between rust genomes (Table S11). We wanted to know the protein kinases exclusively found in pathogenic basidiomycetes and for that we run a BLASTp homology search of PKs of five pathogenic species, *H. vastatrix, P. graminis* f.sp *tritici*, *P. triticina*, *M. larici-populina*, and *U. maydis* against predicted PKs of the non-pathogenic basidiomycetes *Coprinopsis cinerea* (Stajich et al., 2010) and *Laccaria bicolor* (Martin et al., 2008). There are 18 PKs unique to the genomes of plant pathogenic fungi (Table S12) and we identified 236 PKs sequences present in *C. cinerea* and *L. bicolor* but not in pathogenic fungi. In the later group we highlight functions TKL/TKL, PKL/ccin9, PKL/CAK/Fmp29, FunK1, AgaK1, atypical/PIKK/TRRAP, and atypical/HisK PKs families because they were not identified in any of the pathogen species although they are expanded in unique families in *C. cinerea* (Stajich et al., 2010) and *L. bicolor* (Martin et al., 2008).

DISCUSSION

We have generated a de novo hybrid genome assembly of the coffee rust from the sequence of 8 rust samples. We also assembled transcript sequences obtained from normalized and non-normalized RNA-seq libraries representing the urediniospore stage of the fungus. The hybrid genome assembly was the most comprehensive in terms of capturing the largest proportion of the gene space in Hva, therefore offering a picture of a chimeric genome of this species. There appears to be a large extent of repetitive sequences in this chimeric genome, which was evident in our hybrid assembly, as shown by over-covered contigs. For example, two of the three largest contigs showed over-coverage and this event was also found in most contigs shorter than 400 bp. Contamination analysis showed the presence of Bacteria and Viridiplantae sequences in the sequencing reads but the fraction of these sequences was very low. The limitation of having to collect spore samples from plant tissue renders it impossible to have samples free of other organisms; consequently, finding bacterial, plant and other DNA sequences was not unexpected.

We estimated the Hva genome size to be 243–324 Mb by FCM and the assembled scaffolds size was close to the larger figure (333 Mb). Differences between the relative amounts of DNA measured by FCM might be due to the presence of a mixture of different cells containing different number of nuclei in the Hva samples tested; Hva urediniospores in coffee leaves carry out meiosis giving rise to spores at different stages of development containing unequal numbers of nuclei in a process referred as cryptosexuality (Carvalho et al., 2011). A similar mechanism of parasexual recombination has been described in *P. triticina* (Wang and McCallum, 2009). The haplophase has not been recognized in Hva and only urediniospores and teliospores representing the dikaryophase have been identified (De Castro et al., 2009).

Our Hva genome size estimates fall remarkably short of recent measurements of 733 Mbp (Carvalho et al., 2014) and 796.8 Mbp (Tavares et al., 2014) obtained by FCM of Hva nuclei isolated from urediniospores. Highly repetitive genomes such as the Hva genome are complex to sequence and analyze and genomes with a high content of repeats are difficult to sequence completely (Sun et al., 2003). Overall, the Hva genome size is larger compared with other fungal genomes, including the Basidiomycetes *M. larici-populina* (101.1 Mb), *P. graminis* f.sp. *tritici* (88.6 Mb) (Duplessis et al., 2011), and *Laccaria bicolor* (68.9 Mb) (Martin et al., 2008), the arbuscular mycorrhizal fungus *Rhizophagus irregularis* (153 Mb) (Tisserant et al., 2013), and the Ascomicota *Tuber melanosporum* (125 Mb) (Martin et al., 2010), and *Blumeria graminis* f.sp. *tritici* (174 Mb) (Parlange et al., 2011).

Our analysis indicated that the Hva mitochondrial genome is at least the size of the *P. graminis* mitochondria. The different genome size estimates obtained so far make imperative the assembly of an Hva genome from a single Hva isolate to clearly elucidate the real nuclear genome size, mitochondrial genome size and fraction of repetitive sequences for this fungus.

We identified a large fraction of repetitive sequences in the hybrid genome; 74.4% of the assembled contigs contain repetitive sequences, with most of them representing transposable elements. Because of the hybrid nature of the assembly, this might be and over-estimate of the real fraction of repetitive sequences present in the genome. However, given the large estimates for the Hva genome size, we expect the genome sequence to contain a large proportion of repeats. A high proportion of transposable elements have also been identified in the genomes of other rusts (Duplessis et al., 2011; Zheng et al., 2013) and the plant pathogen *Blumeria graminis* (Spanu et al., 2010). Genome expansion caused by the replication of TEs has been shown to occur in filamentous fungal and oomycete pathogens of plants, and some expansion of virulence-related genes are associated with their large genome size (Kemen and Jones, 2012). The high diversity of many *Avr*-genes in the rice blast fungus *Magnaporthe grisea* is related to their association with repeated sequences (Huang et al., 2014). On the other hand, the non-pathogen basidiomycetes *L. bicolor* (Martin et al., 2008) and *C. cinerea* (Stajich et al., 2010) harbor a much-reduced proportion of repeated sequences. Whether the genome of Hva has suffered an expansion of virulence-related genes mediated by transposition events should be investigated in further detail. Given the fact that a hybrid genome might contain an over-representation of the fraction of repetitive elements present in single genomes, there is still need to be cautious about the final proportion of repeats in the Hva genome.

The assembly exhibited a high level of fragmentation as shown by the large number of scaffolds obtained in the final assembly and the low N50 value. This fragmentation can be explained by the highly repetitive nature of the Hva genome. It should be possible in the future to improve this assembly by sequencing large insert libraries that will aid in resolving the repetitive nature and to enlarge scaffolds of the Hva assembly (Raffaele and Kamoun, 2012). An additional approach that might be implemented to improve our current hybrid assembly would be to use a "fosmid-to-fosmid" strategy as that followed by Zheng et al. (2013), who significantly improved an earlier assembly of the *P. striiformis* f.sp. *triticina* genome (Cantu et al., 2011). The GC content of the Hva genome (33%) was lower than *M. larici-populina* (41%), and *P. graminis* f.sp *tritici* (43.3%) (Duplessis et al., 2014). This difference could be explained by GC repetitive sequences collapsing into contigs, therefore yielding a GC content reduction because GC sequences are underrepresented. Though the hybrid Hva genome assembly was highly fragmented, the CEGMA analysis indicated that a significant amount of the genome's gene space was revealed and we consider the current hybrid assembly to be representative of the gene space of a chimeric Hva genome.

Comparative genomics showed considerable similarities between Hva and other rust fungal genomes; over 73% of Hva predicted proteins had homologs among Pucciniales protein datasets. Although rust genomes vary in size, they are very similar in gene content suggesting the presence of a large core set of rust fungus specific genes needed for their pathogenicity. It will also be significant to study the function and specificity of Hva predicted proteins not found in other rusts and study their virulence species-specific adaptations. All in all this set of Hva predicted proteins represent a valuable resource that contributes to the Pucciniales gene repertoire.

In order to capture the gene space of the coffee rust genome, we performed a transcriptome analysis of freshly harvested urediniospores based on Illumina RNA-Seq. The number of secreted proteins predicted in this study is smaller than the number found in *M. larici-populina* (1184 SSPs) and *P. graminis* f.sp. *tritici* (1106 SSPs) genomes (Duplessis et al., 2011), perhaps reflecting the Hva partial genome assembled. Also, it has to be considered that *M. larici-populina* and *P. graminis* f.sp. *tritici* SSPs were predicted with the SignalP, TargetP, and TMHMM algorithms while we did not include TMHMM in our predictions. The non-inclusion of TMHMM transmembrane protein predictions in some way renders our current set of secreted proteins incomplete. On the whole, this set of Hva predicted secreted-proteins is a basic tool for the identification of pathogenicity-related genes as shown for other rusts (Joly et al., 2010; Cantu et al., 2011). It is possible that avirulence elicitors be present among the set of predicted secreted proteins, as it has been found in flax rust (Catanzariti et al., 2006). For Hva proteins KF018008, KF018015, KF018016, KF018020, KF018028 we did not find evidence of predicted homologs in other rusts, representing unique genes of the coffee-rust interaction not represented in other pathosystems; secreted proteins have been found to be lineage-specific in other rusts as well (Duplessis et al., 2011). Overall analysis of the Hva predicted secretome shows that secreted proteins are well conserved among plant rusts and that they include functions most likely involved in the pathogenesis of the fungus. Therefore, this group of annotated secreted proteins suits well as prime candidates for functional testing.

The Hva genome contains most of the gene families coding for signal transduction pathways identified in the genomes of other rust fungi. It is assumed that these gene families are involved in signal perception mechanisms of rust urediniospores (Duplessis et al., 2011) and gives them a highly specialized mechanism for the detection of stomata (Kemen and Jones, 2012). We have grouped the candidate PKs with signal perception roles related to pathogenesis in 18 gene families, those identified in pathogen basidiomycetes but absent in non-pathogenic species.

Illumina and 454 sequencing was used to generate a draft genome in different Hva isolates. Due to the complexity of the genome sequence—similar to other rust fungi- a minimal draft chimeric genome was defined by considering the genome of the different isolates altogether. The genome sequence is a novel resource in Pucciniales, a group that includes many species that are economically major diseases of several crops. It provides data to study the evolution of this important group of plant pathogens. The draft genome sequence of Hva will serve as a template for future assemblies of isolates of this fungus and to understand the molecular mechanisms used by this pathogen to attack the coffee plant and to study its diversity. It will also be the basis for the development of molecular markers to distinguish races/isolates given the enormous difficulties of trying to identify coffee rust races by the use of differential plants. The genomic data of the coffee leaf rust presented here are a reference to track changes in field populations, to characterize the decline in sensitivity against widely used fungicides such as triazoles and strobilurins that are used in coffee rust disease management, and to preserve genotype identity in fungal collections. The increased use of coffee rust-resistant varieties will increase the selective pressure to favor complex fungal genotypes, and resources such as the secretome set is the cornerstone for the development of innovative resistance mechanisms to control this pathogen.

MATERIALS AND METHODS

NUCLEAR DNA CONTENT ESTIMATED BY FCM

Flow cytometry (FCM) was used to estimate nuclear DNA content in urediniospores of *H. vastatrix*, following the protocol described by Eilam et al. (1994), modified for the uredinial stage. Urediniospores were suspended in 0.1% Tween 20 in water for 20 min. The suspension was incubated for 2 min. in a 1000-watt microwave on 50% power level, adding Propidium Iodide and RNase to final concentrations of 4 μg/ml and 50 μg/ml respectively, and incubated for 1 h at 37°C. The urediniospores samples were run on a Bacton Dickinson FACS IV flow cytometer. Urediniospores of *P. triticina* were used as a control. Data from the flow cytometer were analyzed using the Flowing Software v2.5 at the Centre for Biotechnology University of Turku, Finland. Urediniospores of Hva and *P. triticina* were stained and analyzed simultaneously, with the standard control positioned on channel 200. The *C*-Value (pg) was converted to base pairs (bp), considering that 1pg = 978 Mb (Dolezel et al., 2003). *H. vastatrix* samples used for FCM are described in **Table 1**.

GENOME AND TRANSCRIPTOME SEQUENCING AND ASSEMBLY

Hva urediniospore samples were scraped from infected coffee leaves taking care to sample very young pustules with no evidence of the presence of the hyperparasitic fungus *Lecanicillium lecanii*. The coffee genotypes sampled for Hva and the sequencing technologies used are described in **Table 2**. DNA was extracted using the DNeasy Plant Mini-Kit (Qiagen, Hilden, Germany); *H. vastatrix* DNA samples were used to construct 100 bp paired-end libraries and sequenced by Illumina™ HiSeq 2000 at BGI in China. Single-end libraries were sequenced by ROCHE™ 454 GS FLX Titanium method at Macrogen in Korea.

Reads were subjected to quality control checks using FastQC (Babraham Bioinformatics, Babraham Institute), trimmed using the CLC quality_trim script (CLC bio, Aarhus, Denmark), masked or filtered by low complexity end regions, and exclusion of reads shorter than 70 nucleotides. Mdust and SeqClean were used for the cleaning process (The Gene Index Project, Harvard University—http://sourceforge.net/projects/seqclean/files/). Trimmed and filtered reads were assembled with the CLC Assembly Cell v4.0.1 (CLC bio, Aarhus, Denmark) with the following parameters: deletions penalty = 3, no global alignment, remove duplicates, min contig length = 200 bp, paired-end distance = 200–400 bp. The quality of the assembly was assessed with CLC tools (clc_assembly_viewer, assembly_info) and in-house R scripts available at (http://bioinformatics.cenicafe.org/index.php/wiki/Third_Hybrid_Assembly_of_454_and_Illumina_data_with_CLC).

The hybrid assembly was analyzed using MEGAN 4 (Huson et al., 2001) to assess the level of possible contamination and to perform a first approximation of the biological communities associated with Hva on the coffee leaf. Blastx was performed using the contigs from the hybrid assembly (Illumina + 454 short reads) (396,264 contigs) against the NCBI non-redundant protein database. An *E*-value of $10e^{-3}$ was used as a cut-off following the recommendation from the MEGAN developers. MEGAN was used to map and visualize the Low Common Ancestor (LCA) in the NCBI tree taxonomy for each contig. With the aim of filtering out putative contaminated sequences, contigs that presented similarities to reported fungal sequences were extracted to form a reliable set of Hva genome contigs. The reliable set of *H. vastatrix* genome contigs was compared against the *P. graminis* f.sp. *tritici* and *P. pachyrhizi* mitochondrial genomes using Blastn (with an *E*-value threshold of $1E^{-5}$).

For RNA-seq sample preparation, Hva urediniospores were scraped from infected coffee leaves of the coffee genotypes described in Table S3, taking care to sample very young pustules with no evidence of the presence of the hyperparasitic fungus *Lecanicillium lecanii*. RNA was extracted from urediniospores using the RNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA). Normalized library construction was performed at Evrogen, Moscow, Russia using Kamchatka crab duplex-specific nuclease (Zhulidov et al., 2004). First-strand cDNA was prepared from poly(A)+ *H. vastatrix* urediniospores RNA using a SMART™ PCR cDNA Synthesis Kit (Clontech), according to the manufacturer's protocol. SMART™ Oligo II and CDS primers (Clontech) were used for first-strand cDNA synthesis. A 1.5 ml aliquot of a 100 ng/ml of the first-strand cDNA solution was incubated for normalization with 0.25 Kunitz units of duplex-specific nuclease from kamchatka crab and amplified by PCR. Sequencing of amplified cDNA products was performed on an Illumina™ HiSeq 2000 system (BGI, Shenzhen, 518083, China).

RNA-seq data were filtered before assembly. The quality of the transcripts was measured using the FASTX-Toolkit, reads were trimmed by quality and duplicates were removed. Clean reads longer than 200bp were assembled using the Trinity package (Grabherr et al., 2011). First, the reads were run through Trinity's Inchworm module, which assembles the read data set into different pools of reads, and the Chrysalis module was used to construct de Brujin graphs for all the read pools obtained using Inchworm. We used the module Butterfly that reconciles de Brujin graphs using the read pools from the former modules and output assembled contigs. We mapped Hva transcript datasets to the HvHybrid 454-Illumina assembly and we also compared transcripts against the NR database to identify plant, bacterial and other contaminant sequences.

TRANSPOSABLE ELEMENTS PREDICTION

We surveyed the frequency and classes of TE-like elements present in the HvHybrid assembly using the algorithm RepeatMasker (Smit et al., 1996-2004) and the RepBase12.12 and fngrep.ref databases. The fngrep.ref database included 1726 transposable elements identified in fungi. Novel retrotransposon families were manually annotated from the gene families identified with OrthoMCL (see below).

GENE PREDICTION

For the prediction of gene models, we followed the "align then assemble" approach (Martin and Wang, 2011). We mapped RNA-seq short reads to the genome using TopHat (Trapnell et al., 2009), and we identified putative transcriptional units using Augustus (Stanke and Waack, 2003). Protein sequences were computationally deduced from the transcriptional units. Gene families from predicted proteins larger than 70 amino acids were identified with OrthoMCL using a default MCL inflation value of 1.5 and a blastp *e*-value of $10e^{-5}$ (Li et al., 2003). We explored the set of predicted proteins, searching for KOGs using the CD-Search Tool and the Conserved Domain Database (www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml).

COMPARATIVE GENOMICS

The Hva genome contigs were aligned against the genomes of *P. graminis*, *M. larici-populina* and *U. maydis* using Mauve (Darling et al., 2004). The Low Collinear Block (LCB) values were set through visual inspection by searching the best block size for each pair of alignments (largest coverage of both genomes). Finally, values used for LCB were as follows: *P. graminis* 12,154, *M. larici-populina* 10,409, and *U. maydis* 1203.

For genome annotation, we used custom Perl scripts and basic bioinformatics software such as BLAST (Altschul et al., 1990). The databases we used for comparisons corresponded to 67,118 Pucciniales sequences comprising 16,694 protein-coding genes from *M. larici-populina* (Duplessis et al., 2011), 22,815 *P. striiformis* f.sp. *tritici* sequences (Cantu et al., 2011), 15,979 *P. graminis* f.sp. *tritici* sequences (Duplessis et al., 2011), and 11,630 *P. triticina* sequences (Xu et al., 2011).

For homology searches of protein kinases (PKs) we run BLASTp (version 2.2.28) with an $e = 1e^{-10}$. We searched *H. vastatrix* predicted proteins against 131 predicted PKs from *Sacharomyces cerevisiae* and then we compared the coffee rust PKs against *M. laricis-populina*, *P. striiformis* f.sp *tritici*, *P. graminis* f.sp *tritici*, *P. triticina,* and *U. maydis* predicted PKs.

SECRETED PROTEINS

The *H. vastatrix* predicted proteins were classified into secreted and non-secreted proteins. For this task, the programs SignalP 4.0 (Petersen et al., 2011) and PProwler (Hawkins and Boden, 2006) were used to predict putatively secreted proteins. A 0.9 probability cut-off was used for PProwler predictions. A set of secreted proteins predicted previously for *H. vastatrix* by Fernandez et al. (2012) was used for comparison with our predictions. Briefly, a Blastp was performed between our set of *H. vastatrix* proteins and the predictions by Fernandez et al. (2012). Finally, a set of proteins that showed similarity (Blastp $e = 1e^{-5}$) with the secreted proteins predicted by Fernandez et al. (2012) was obtained. Reciprocal comparisons of the three sets of secreted proteins were performed (SignalP, PProwler and

Fernandez-Blastp) to establish the proteins shared by the three predictions.

AVAILABILITY

Raw data and metadata for the Genome project is available at NCBI, BioProject ID: PRJNA188788 and the Transcriptome project ID: PRJEB2960. Predicted and secreted proteins are available at http://bioinformatics.cenicafe.org/index.php/wiki/CoffeeRustPredictedProteins.

The hybrid reference assembled genome contigs are available for download at: http://bioinformatics.cenicafe.org/index.php/wiki/CoffeeRustHybridDraftAssembly_Contigs.

ACKNOWLEDGMENTS

The authors wish to thank the National Federation of Coffee Growers of Colombia, the Ministry of Agriculture and Rural Development and the Colombian Administrative Department of Science, Technology and Innovation—Colciencias for financial support. Authors would also like to thank Perttu Terho from the Centre for Biotechnology, University of Turku, Finland for data analysis from the flow cytometer experiment. Diego Mauricio Riaño-Pachón acknowledges funding provided by the Faculty of Sciences at Universidad de los Andes through the Assistant Professor Support Program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fpls.2014.00594/abstract

REFERENCES

Aime, C., Matheny, P. B., Henk, D. A., Frieders, E. M., Nilsson, R. H., Piepenbring, M., et al. (2006). An overview of the higher-level classification of Pucciniomycotina based on combined analyses of nuclear large and small subunit rDNA sequences. *Mycologia* 98, 896–905. doi: 10.3852/mycologia.98.6.896

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Alvarado, G., and Moreno, G. (2005). Cambio de la virulencia de *Hemileia vastatrix* en progenies de Caturra x Híbrido de Timor. *Cenicafe* 56, 110–126.

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M., Chen, X., Kojima, K., et al. (2011). Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f.sp. *tritici*, the causal agent of wheat stripe rust. *PLoS ONE* 6:e24230. doi: 10.1371/journal.pone.0024230

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f.sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14:270. doi: 10.1186/1471-2164-14-270

Carvalho, A., Eskes, A. B., Castillo, J., Sreenivasan, M., Echeverri, J., Fernandez, C., et al. (1987). "Breeding programs," in *Coffee Rust: Epidemiology, Resistance, and Management*, eds A. C. Kushalappa and A. B. Eskes (Boca Raton, FL: CRC Press), 293–336

Carvalho, C. R., Fernandes, R. C., Carvalho, G. M. A., Barreto, R. W., and Evans, H. C. (2011). Cryptosexuality and the genetic diversity paradox in coffee rust, *Hemileia vastatrix*. *PLoS ONE* 6:e26387. doi: 10.1371/journal.pone.0026387

Carvalho, G. M. A., Carvalho, C. R., Barreto, R. W., and Evans, H. C. (2014). Coffee rust genome measured using flow cytometry: does size matter? *Plant Pathol.* 63, 1022–1026. doi: 10.1111/ppa.12175

Castillo, J., and Leguizamón, J. (1992). Virulencia de *Hemileia vastatrix* determinada por medio de plantas diferenciales de café en Colombia. *Cenicafe* 43, 114–124

Catanzariti, A. M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *Plant Cell* 18, 243–256. doi: 10.1105/tpc.105.035980

Cressey, D. (2013). Coffee rust regains foothold. *Nature* 493, 587. doi: 10.1038/493587a

Cristancho, M. A., Rozo, Y., Escobar, Y., Rivillas, C. A., and Gaitán, A. L. (2012). Outbreak of coffee leaf rust (*Hemileia vastatrix*) in Colombia. *New Dis. Rep.* 25, 2044–0588. doi: 10.5197/j.2044-0588.2012.025.019

Darling, A. C. E., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* 14, 1394–1397. doi: 10.1101/gr.2289704

Darzentas, N. (2010). Circoletto: visualizing sequence similarity with Circos. *Bioinformatics* 26, 2620–2621. doi: 10.1093/bioinformatics/btq484

De Castro, R., Evans, H. C., and Barreto, R. W. (2009). Confirmation of the occurrence of teliospores of *Hemileia vastatrix* in Brazil with observations on their mode of germination. *Trop. Plant Pathol.* 34, 108–113. doi: 10.1590/S1982-56762009000200005

Dolezel, J., Bartos, J., Voglmayr, H., and Greilhuber, J. (2003). Nuclear DNA content and genome size of trout and human. *Cytometry A* 51, 127–128. doi: 10.1002/cyto.a.10013

Duplessis, S., Bakkeren, G., and Hamelin, R. (2014). Advancing knowledge on biology of rust fungi through genomics. *Adv. Bot. Res.* 70, 173–209. doi: 10.1016/B978-0-12-397940-7.00006-9

Duplessis, S., Cuomo, C. A., Lin, Y. C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108, 9166–9171. doi: 10.1073/pnas.1019315108

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLoS Comp. Biol.* 7:e1002195. doi: 10.1371/journal.pcbi.1002195

Eilam, T., Bushnell, W. R., and Anikster, Y. (1994). Relative nuclear DNA content of rust fungi estimated by flow cytometry of Propydium iodide-stained pycniospores. *Phytopathology* 84, 728–735. doi: 10.1094/Phyto-82-1212

Fernandez, D., Tisserant, E., Talhinhas, P., Azinheira, H., Vieira, A., Petitot, A. S., et al. (2012). 454-pyrosequencing of *Coffea arabica* leaves infected by the rust fungus *Hemileia vastatrix* reveals in planta-expressed pathogen-secreted proteins and plant functions in a late compatible plant–rust interaction. *Mol. Plant Pathol.* 13, 17–37. doi: 10.1111/j.1364-3703.2011.00723.x

Gil, L. F., and Ocampo, J. D. (1998). Identificación de la raza XXII (V5-6) de *Hemileia vastatrix* Berk. y Br. en Colombia. *Cenicafé* 49, 340–344

Gouveia, M., Ribeiro, A., Várzea, V., and Rodrigues, C. J. Jr. (2005). Genetic diversity in *Hemileia vastatrix* based on RAPD markers. *Mycologia* 97, 396–404. doi: 10.3852/mycologia.97.2.396

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29, 644–652. doi: 10.1038/nbt.1883

Gu, B., Shiv, D., Kale, Q. W., Dinghe, W., Qiaona, P., Hua, C., et al. (2011). Rust secreted protein Ps87 is conserved in diverse fungal pathogens and contains a RXLR-like motif sufficient for translocation into plant cells. *PLoS ONE* 6:e27217. doi: 10.1371/journal.pone.0027217

Hacquard, S., Joly, D. L., Lin, Y. G., Tisserant, E., Feau, N., Delaruelle, C., et al. (2012). A comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (Poplar Leaf Rust). *Mol. Plant Microbe Interact.* 25, 279–293. doi: 10.1094/MPMI-09-11-0238

Hawkins, J., and Boden, M. (2006). Detecting and sorting targeting peptides with recurrent networks and support vector machines. *J. Bioinform. Comput. Biol.* 4, 1–18. doi: 10.1142/S0219720006001771

Huang, J., Si, W., Deng, Q., Li, P., and Yang, S. (2014). Rapid evolution of avirulence genes in rice blast fungus *Magnaporthe oryzae*. *BMC Genet.* 15:45. doi: 10.1186/1471-2156-15-45

Huson, D. H., Mitra, S., Weber, N., Ruscheweyh, H., and Schuster, S. C. (2001). Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* 21, 1552–1560. doi: 10.1101/gr.120618.111

Joly, D. L., Feau, N., Tanguay, P., and Hamelin, R. C. (2010). Comparative analysis of secreted protein evolution using expressed sequence tags from four poplar leaf rusts (*Melampsora* spp.). *BMC Genomics* 11:422. doi: 10.1186/1471-2164-11-422

Kemen, E., and Jones, J. D. (2012). Obligate biotroph parasitism: can we link genomes to lifestyles? *Trends Plant Sci.* 17, 448–457. doi: 10.1016/j.tplants.2012.04.005

Leguizamón, J. E., Baeza, C. A., Fernández, O., Moreno, G., Castillo, Z. J., and Orozco, F. J. (1984). Identification of race II of *Hemileia vastatrix* Berk y Br. in Colombia. *Cenicafé* 35, 26–28

Li, L., Stoeckert, C. J. Jr., and Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* 13, 2178–2189. doi: 10.1101/gr.1224503

Marçais, G., and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* 27, 764–770. doi: 10.1093/bioinformatics/btr011

Martin, F., Aerts, A., Ahrén, D., Brun, A., Danchin, E. G., Duchaussoy, F., et al. (2008). The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature* 452, 88–92. doi: 10.1038/nature06556

Martin, F., Kohler, A., Murat, C., Balestrini, R., Coutinho, P. M., Jaillon, O., et al. (2010). Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature* 464, 1033–1038. doi: 10.1038/nature 08867

Martin, J. A., and Wang, Z. (2011). Next-generation transcriptome assembly. *Nat. Rev. Genet*. 12, 671–682. doi: 10.1038/nrg3068

Miranda, D., and Barton, G. J. (2007). Classification and functional annotation of eukaryotic protein kinases. *Proteins* 68, 893–914. doi: 10.1002/prot.21444

Monaco, L. C. (1977). Consequences of the introduction of coffee rust into Brazil. *Ann. N.Y. Acad. Sci*. 287, 57–71. doi: 10.1111/j.1749-6632.1977.tb34231.x

Nemri, A., Saunders, D. G., Anderson, C., Upadhyaya, N. M., Win, J., Lawrence, G. J., et al. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. *Front. Plant Sci.* 5:98. doi: 10.3389/fpls.2014.00098

Parlange, F., Oberhaensli, S., Breen, J., Platzer, M., Taudien, S., and Šimková, H., et al. (2011). A major invasion of transposable elements accounts for the large size of the *Blumeria graminis* f.sp. *tritici* genome. *Funct. Integr. Genomics* 11, 671–677. doi: 10.1007/s10142-011-0240-5

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067. doi: 10.1093/bioinformatics/btm071

Petersen, T., Brunak, S., von Heijne, G., and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786. doi: 10.1038/nmeth.1701

Pretsch, K., Kemen, A., Kemen, E., Geiger, M., Mendgen, K., and Voegele, R. (2013). The rust transferred proteins-a new family of effector proteins exhibiting protease inhibitor function. *Mol. Plant Pathol.* 14, 96–107. doi: 10.1111/j.1364-3703.2012.00832.x

Puccinia Group Genomes Database. Available online at: http://www.broadinstitute.org/annotation/genome/puccinia_group/GenomeDescriptions.html#P_triticina_1_1_V1.

Raffaele, S., and Kamoun, S. (2012). Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* 10, 417–430. doi: 10.1038/nrmicro2790

Rivillas, C., Serna, C., Cristancho, M., and Gaitán, A. (2011). *Roya del Cafeto en Colombia: Impacto, Manejo y Costos del Control.* Chinchiná: Boletín Técnico No. 36, Cenicafe.

Rodrigues, C. J. Jr., Bettencourt, A. J., and Rijo, L. (1975). Races of the pathogen and resistance to coffee rust. *Ann. Rev. Phytopathol.* 13, 49–70. doi: 10.1146/annurev.py.13.090175.000405

Rodrigues, C. J. Jr., Várzea, V., Godinho, I. L., Palma, S., and Rato, R. C. (1993). "New physiologic races of *Hemileia vastatrix*," in *Proceedings of the 15th International Conference on Coffee Science* (Montpellier), 318–321.

Rozo, Y., Escobar, C., Gaitán, A. L., and Cristancho, M. A. (2012). Aggressiveness and genetic diversity of *Hemileia vastatrix* during an epidemic in Colombia. *J. Phytopathol*. 160, 732–740. doi: 10.1111/jph.12024

Saunders, D. G. O., Win, J., Cano, L. M., Szabo, L. J., Kamoun, S., and Rafaelle, S. (2012). Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. *PLoS ONE* 7:e29847. doi: 10.1371/journal.pone.0029847

Schirawski, J., Mannhaupt, G., Münch, K., Brefort, T., Schipper, K., Doehlemann, G., et al. (2010). Pathogenicity determinants in smut fungi revealed by genome comparison. *Science* 330, 1546–1548. doi: 10.1126/science.11 95330

Smit, A., Hubley, R., and Green, P. (1996-2004). *RepeatMasker Open-3.0.* Available online at: http://www.repeatmasker.org.

Spanu, P. D. (2012). The genomics of obligate (and nonobligate) biotrophs. *Annu. Rev. Phytopathol.* 50, 91–109. doi: 10.1146/annurev-phyto-081211-173024

Spanu, P. D., Abbott, J. C., Amselem, J., Burgis, T. A., Soanes, D. M., Stüber, K., et al. (2010). Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* 330, 1543–1546. doi: 10.1126/science.1194573

Stajich, J. E., Wilke, S., Ahrén, D., Au, C. A., Birren, B. W., Borodovsky, M., et al. (2010). Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom *Coprinopsis cinerea* (*Coprinus cinereus*). *Proc. Natl. Acad. Sci. U.S.A.* 107, 11889–11894. doi: 10.1073/pnas.1003391107

Stanke, M., and Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19(Suppl. 2), ii215–ii225. doi: 10.1093/bioinformatics/btg1080

Stone, C. L., Buitrago, M. L., Boore, J. L., and Frederick, R. D. (2010). Analysis of the complete mitochondrial genome sequences of the soybean rust pathogens *Phakopsora pachyrhizi* and *P. meibomiae. Mycologia* 102, 887–897. doi: 10.3852/09-198

Sun, X., Le, H. D., Wahlstrom, J. M., and Karpen, G. H. (2003). Sequence analysis of a functional Drosophila centromere. *Genome Res.* 13, 182–194. doi: 10.1101/gr.681703

Talhinhas, P., Azinheira, H., Vieira, B., Loureiro, A., Tavares, S., Batista, D., et al. (2014). Overview of the functional virulent genome of the coffee leaf rust pathogen *Hemileia vastatrix* with an emphasis on early stages of infection. *Front. Plant Sci.* 5:88. doi: 10.3389/fpls.2014.00088

Tavares, S., Ramos, A. P., Pires, A. S., Azinheira, H. G., Caldeirinha, P., Link, T., et al. (2014). Genome size analyses of Pucciniales reveal the largest fungal genomes. *Front. Plant Sci.* 5:422. doi: 10.3389/fpls.2014.00422

Tisserant, E., Malbreil, M., Kuo, A., Kohler, A., Symeonidi, A., Balestrini, R., et al. (2013). Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proc. Natl. Acad. Sci. U.S.A.* 110, 20117–20122. doi: 10.1073/pnas.1313452110

Toome, M., Ohm, R. A., Riley, R. W., James, T. Y., Lazarus, K. L., Henrissat, B., et al. (2014). Genome sequencing provides insight into the reproductive biology, nutritional mode and ploidy of the fern pathogen *Mixia osmundae*. *New Phytol.* 202, 554–564. doi: 10.1111/nph.12653

Trapnell, C., Pachter, L., and Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111. doi: 10.1093/bioinformatics/btp120

Wang, X., and McCallum, B. (2009). Fusion body formation, germ tube anastomosis, and nuclear migration during the germination of urediniospores of the wheat leaf rust fungus, *Puccinia triticina*. *Phytopathology* 99, 1355–1364. doi: 10.1094/PHYTO-99-12-1355

Xu, J., Linning, R., Fellers, J., Dickinson, M., Zhu, W., Antonov, I., et al. (2011). Gene discovery in EST sequences from the wheat leaf rust fungus *Puccinia triticina* sexual spores, asexual spores and haustoria, compared to other rust and corn smut fungi. *BMC Genomics* 12:161. doi: 10.1186/1471-2164-12-161

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nat. Commun.* 4:2673. doi: 10.1038/ncomms3673

Zhulidov, P. A., Bogdanova, E. A., Shcheglov, A. S., Vagner, L. L., Khaspekov, G. L., Kozhemyako, V. B., et al. (2004). Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res.* 32:e3. doi: 10.1093/nar/gnh031

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 30 May 2014; accepted: 11 October 2014; published online: 31 October 2014.
Citation: Cristancho MA, Botero-Rozo DO, Giraldo W, Tabima J, Riaño-Pachón DM, Escobar C, Rozo Y, Rivera LF, Durán A, Restrepo S, Eilam T, Anikster Y and Gaitán AL (2014) Annotation of a hybrid partial genome of the coffee rust (Hemileia vastatrix) contributes to the gene repertoire catalog of the Pucciniales. Front. Plant Sci. ***5****:594. doi: 10.3389/fpls.2014.00594*
This article was submitted to Plant-Microbe Interaction, a section of the journal Frontiers in Plant Science.