Mol Gen Genomics (2005) 273: 20–32
DOI 10.1007/s00438-005-1114-0

ORIGINAL PAPER

**Rays H. Y. Jiang · Angus L. Dawe · Rob Weide
Marjo van Staveren · Sander Peters · Donald L. Nuss
Francine Govers**

Elicitin genes in *Phytophthora infestans* are clustered and interspersed with various transposon-like elements

Received: 12 August 2004 / Accepted: 21 December 2004 / Published online: 9 February 2005

Abstract Sequencing and annotation of a contiguous stretch of genomic DNA (112.3 kb) from the oomycete plant pathogen *Phytophthora infestans* revealed the order, spacing and genomic context of four members of the elicitin (*inf*) gene family. Analysis of the GC content at the third codon position (GC3) of six genes encoded in the region, and a set of randomly selected coding regions as well as random genomic regions, showed that a high GC3 value is a general feature of *Phytophthora* genes that can be exploited to optimize gene prediction programs for *Phytophthora* species. At least one-third of the annotated 112.3-kb *P. infestans* sequence consisted of transposons or transposon-like elements. The most prominent were four Tc3/gypsy and Tc1/copia type retrotransposons and three DNA transposons that belong to the Tc1/mariner, Pogo and PiggyBac groups, respectively. Comparative analysis of other available genomic sequences suggests that transposable elements are highly heterogeneous and ubiquitous in the *P. infestans* genome.

Keywords Class I element · Class II element · Late blight · CHROMO domain

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00438-005-1114-0

Communicated by W.R. McCombie

R. H. Y. Jiang · R. Weide · F. Govers (✉)
Plant Sciences Group, Laboratory of Phytopathology,
Graduate School of Experimental Plant Sciences,
Wageningen University,
Binnenhaven 5, 6709 PD, Wageningen, The Netherlands
E-mail: Francine.Govers@wur.nl
Tel.: +31-317-483138
Fax: +31-317-483412

A. L. Dawe
Biology Department,
New Mexico State University,
234 Foster Hall, Las Cruces, NM 88003, USA

M. Staveren · S. Peters
Greenomics,
Plant Research International,
P.O. Box 16, 6700 AA, Wageningen, The Netherlands

D. L. Nuss
Center for Biosystems Research,
University of Maryland Biotechnology Institute,
College Park, MD 20742-4450, USA

Introduction

Phytophthora infestans is the causative agent of potato late blight and one of the most devastating plant pathogens known today. *Phytophthora* belongs to the class Oomycetes, and the genus comprises over 60 species, all of which are notorious pathogens of crop plants, trees and ornamentals (Erwin and Ribeiro 1996). Their growth morphology and dispersal strategy resemble those of fungi and the weaponry that oomycetes and fungi use to attack plants appears to be comparable (Latijnhouwers et al. 2003). In the eukaryotic phylogenetic tree, however, oomycetes are classified as heterokonts together with brown algae and diatoms, and are positioned on a branch completely separate from fungi (Baldauf 2003). The distinct phylogenetic positions of oomycetes and fungi are manifest in, among others, differences in intracellular structures, cell wall composition, physiological and biochemical processes and ploidy level (Erwin and Ribeiro 1996; Kamoun 2003).

It is very likely that their evolutionary history has also shaped the genes and genomes of oomycetes. Within oomycetes, *Phytophthora* is the most extensively studied genus and data on *Phytophthora* genes, gene structure, gene expression, and repeat elements are steadily accumulating (Kamoun 2003). The EST databases for *P. infestans* (Kamoun et al. 1999; Randall et al., in press) and *P. sojae* (Qutob et al. 2000) are the most advanced (https://xgi.ncgr.org/spc; http://www.pfgd.org) and various other smaller scale EST projects are ongoing (e.g. in *P. nicotianae*; Skalamera et al. 2004). In addition, the genomes of two *Phytophthora* species have been

sequenced, that of *P. sojae* to nine times coverage (genome.jgi-psf.org/sojae1/sojae1.home.html) and that of *P. ramorum* to seven times coverage (genome.jgi-psf.org/ramorum1/ramorum1.home.html). *Phytophthora* spp. show a wide range of genome sizes, which generally exceed those of fungi and other microorganisms; the *P. ramorum* genome is 65 Mb long whereas the *P. infestans* genome is estimated to comprise 240 Mb (Kamoun 2003).

To fully explore the available genome and EST databases requires annotation tools and gene prediction programs specifically trained for *Phytophthora*. A first step toward the establishment of such tools is the detailed analysis and annotation of relatively small genomic regions. The discovery of typical features of coding regions, patterns of repeat distribution, and diversity and number of transposable elements will be vital for the design of automated gene prediction programs to be used to scan whole genomes. Here, we present the annotation of the 112.3-kb sequence of a bacterial artificial chromosome (BAC) obtained from a genomic library of *P. infestans*. This BAC was selected from a physical contig that spans a number of elicitin genes.

Elicitins belong to a particular class of extracellular proteins produced by *Phytophthora* species. They were first characterized on the basis of their ability to induce defense responses in plants, in particular in *Nicotiana* species, and are thought to act as species-specific avirulence factors, and thus, as determinants of the host range for selected plant-*Phytophthora* interactions (Ricci et al. 1992; Kamoun et al. 1998). Elicitins can act as a sterol-carrier (Mikes et al. 1998), a biological function that seems to be essential, since *Phytophthora* itself cannot synthesize sterols and must retrieve them from external sources (Hendrix and Guttman 1970).

In *P. infestans*, elicitins are encoded by a complex multigene family (Kamoun et al. 1999). All *inf* elicitin genes encode putative extracellular proteins that share the 98 amino-acid elicitin domain corresponding to the mature canonical INF1 protein, the most abundant extracellular protein. Five *inf* genes encode proteins with an extended C-terminal domain (Kamoun et al. 1999). Preliminary data based on genomic Southern hybridizations suggested that several members of the *inf* elicitin gene family are clustered in the genome. Studies in two other *Phytophthora* species also demonstrated clustering of elicitin genes. In both, *P. cinnamomi* and *P. cryptogea*, a 6-kb genomic region encompasses four elicitin genes (Panabieres et al. 1995; Duclos et al. 1998). In *P. infestans* the *inf1* gene is highly expressed in mycelium but not in sporangia or cysts (Kamoun et al. 1997). Most other *inf* genes show a similar expression pattern but the expression levels vary. In an EST library from mycelium, *inf5* and *inf6* were among the four most abundant cDNA clusters, suggesting that these *inf* genes are highly active (Kamoun et al. 1999).

The aims of this study were (1) to sequence and annotate a long stretch of *P. infestans* genomic DNA, (2) to investigate the order and spacing of the *inf* elicitin genes located in this stretch, and (3) to examine the genomic context of the elicitin gene cluster. Annotation of the sequence revealed many repeats and showed that the elicitin gene cluster is interspersed with transposons and transposon-like elements representing various classes and groups. Analysis of the GC content and codon usage of coding sequences revealed characteristic features of *Phytophthora* genes that will be instrumental for gene prediction.

Materials and methods

BAC library screening

The *P. infestans* BAC library used for screening has been described by Whisson et al. (2001). Screening was done by colony hybridization according to standard procedures, and ^{32}P-labelled probes were prepared by the random hexamer method with a random primer labelling kit (Prime-a-Gene; Gibco-BRL). The *inf* elicitin probes were prepared from EST clones from the *P. infestans* MY EST library described by Kamoun et al. (1999). Alkaline lysis was used to isolate plasmid DNA and inserts were released by digestion with *Eco*RI and *Bam*HI and subsequently purified from the gel after electrophoresis. The *inf1* probe was derived from EST clone MY18D10, *inf2A* from MY05C05, *inf2B* from MY02D01, *inf3* from MY19C07, *inf4* from MY11E04, *inf5* from MY01C05 and *inf6* from MY01D04. BACs that hybridized to the *inf* probes were picked and grown in LB containing chloramphenicol (12.5 μg/ml). BAC DNA was isolated by alkaline lysis and digested with restriction enzymes. Fragments were size separated by electrophoresis on agarose gels and transferred to Hybond N^+ membranes. Hybridization with the individual *inf* probes was performed to confirm the identity of the *inf* genes located on the BACs.

BAC fingerprinting, contig building and insert size determination

To obtain BAC fingerprint patterns, 1-μg aliquots of BAC DNA were digested with 10 U of *Hin*dIII in 100-μl reactions for 4 h at 37°C; the digestion products were then precipitated by isopropanol and dissolved in 10-μl of TE for gel electrophoresis. For contig building, fragments from different BACs that were of identical length were considered to be common fragments.

To determine the BAC insert sizes, BAC DNA was digested with *Not*I, which separates the vector pBeloBAC11 from the insert. The digested DNA was analyzed on CHEF (Contour-clamped Homogeneous Electric Field) gels using a CHEF-DR II Pulse Field Gel Electrophoresis Apparatus (Bio-Rad, CA, USA). The CHEF gels consisted of 1% agarose in 0.5 times TBE, and electrophoresis was performed for 18 h in five times TBE

buffer at 13°C, with a 5–15 s switch time (linear ramping) and at a constant voltage of 220 V.

Shotgun cloning, sequencing and sequence assembly

BAC DNA was purified using Plasmid-Safe ATP-Dependent DNase (Epicenter) to remove contaminating *E. coli* genomic DNA, subsequently sheared, fractionated and cloned into TOPO vector (TOPO Shotgun Subcloning kit) as described by the manufacturer (Invitrogen, CA, USA). Average insert sizes of the shotgun clones were between 2.5 and 3 kb. Plasmids were manually prepared from cultures of stored colonies using the Qiaprep 8 system (Qiagen) with a vacuum manifold. Each preparation was checked for yield by electrophoresis, and then submitted to the DNA Sequencing Core Facility at the Center for Biosystems Research for analysis on ABI 3100 or 377 (Applied Biosystems, CA, USA) machines. Dual (5′ and 3′) sequence reads of the cloned fragments were obtained using the M13 forward and reverse primers. The resulting sequence files were scanned using the SeqMan unit of DNAStar (DNAStar Inc.) running on a Macintosh G4 computer to check for the presence of known vector sequences and assess the quality of data prior to further analyses. Shotgun sequences were base-called using the PHRED base caller and assembled using the Gap4 assembler in the Staden2003 package. Using the PREGAP4 interface, GAP4-assembled sequences were parsed into the GAP4 assembly database (Bonfield et al. 1995). The GAP4 interface and its features were then used for editing and sequence finishing. Consensus calculations with a quality cut-off value of 40 were performed from within GAP4 using a probabilistic consensus algorithm based on expected error rates output by PHRED. By sequencing PCR products using custom-designed primers and bridging the ends of contiguous fragments the remaining sequence gaps were closed. Most of the gap-closure sequencing was performed at Greenomics, PRI (Wageningen) using the ABI PRISM Big Dye Terminator Cycle Sequencing Ready reaction kit with FS AmpliTaq DNA polymerase (Perkin Elmer, MA, USA) and analyzed on an ABI 3730XL DNA Analyzer. To verify the assembly, read-pairs were analyzed on direction and size using a maximum size spacing of 2.5 kb.

Programs for sequence annotation

Sequences were analyzed with Vector NTI 8. For BLAST searches we used the NCBI BLAST program and the Standalone-BLAST Version 2.2.3 (Altschul et al. 1990). Repeat analysis was done with PiPMaker (Schwartz et al. 2000) and for multiple sequence alignment ClustalX 1.0 was used (Jeanmougin et al. 1998). Phylogenetic tree construction was performed using Molecular Evolutionary Genetic Analysis (MEGA) Version 2.1 (Kumar et al. 1994). Calculation scripts were written in Python2.2 (http://www.python.org).

Genome databases and EST databases

P. infestans and *P. sojae* EST databases are accessible at https://xgi.ncgr.org/spc [Syngenta *Phytophthora* Consortium (SPC) EST sequence databases; Randall et al., in press] and http://www.pfgd.org [*Phytophthora* Functional Genomics Database; previously the *Phytophthora* Genome Consortium database (https://xgi.ncgr.org/pgc); Kamoun et al. 1999; Qutob et al. 2000)]. The sequences from *Fusarium graminearum* and *Aspergillus nidulans* were downloaded from the Broad Institute website (http://www.broad.mit.edu/annotation) and the *Blumeria graminis* EST database was downloaded from the Phytopathogenic Fungi and Oomycete EST Database Version 1.4 (Soanes et al. 2002). Random genomic sequences of *P. infestans* and *P. sojae*, produced by the Broad Institute (Cambridge, MA, USA) and the DOE Joint Genome Institute (Walnut Creek, CA, USA), respectively, were retrieved from the NCBI trace file archive (http://www.ncbi.nlm.nih.gov/Traces).

Results and discussion

Clustering of elicitin genes in the *P. infestans* genome

One-third of the BAC colonies from a library representing 10-fold coverage of the *P. infestans* genome (Whisson et al. 2001) was screened with cDNA clones for seven members of the *inf* elicitin gene family— *inf1*, *inf2A*, *inf2B*, *inf3*, *inf4*, *inf5* and *inf6*. Genomic Southern analysis had previously shown that all seven are single-copy genes. At least seven BACs hybridized to two or more of the *inf* probes, suggesting some degree of clustering of the *inf* elicitin genes in the genome. Insert sizes were determined by analyzing *Not*I digests on CHEF gels. By fingerprinting *Hin*dIII digests and hybridization of fingerprint blots with the *inf* probes one physical contig spanning the seven *inf* elicitin genes could be assembled (Fig. 1). BAC11A5 carrying four elicitin genes including the canonical *inf1* gene was chosen for further analysis.

Sequencing and annotation of BAC clone 11A5

BAC11A5 was sequenced using a combination of shotgun and directed approaches. A total number of 1031 sequence reads was assembled and edited using PHRED to yield one continuous 112.3-kb sequence contig with an average of 5.38-fold coverage (GenBank Accession No. AY830090). Annotation of the assembled sequence revealed the four expected elicitin genes, two putative genes and a large number of transposon-like sequences (Fig. 2,

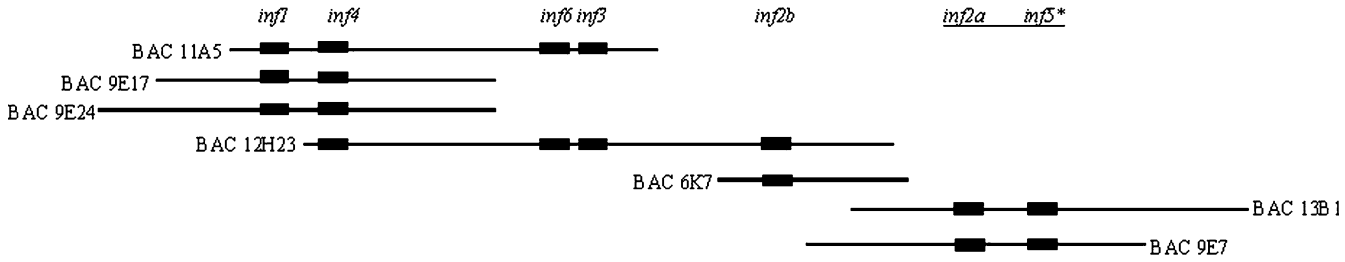

Fig. 1 A *P. infestans* BAC contig of approximately 250 kb containing seven *inf* elicitin genes. The *filled boxes* represent the *inf* genes. Note that the order of *inf2A* and *inf5* has not been determined. BAC insert sizes are: BAC6K7, 45 kb; BAC9E7, 50 kb; BAC11A5, 130 kb, BAC9E24, 150 kb; BAC12H23, 120 kb; BAC13B1, 120 kb

Table 1). One of the two putative genes, named *ORF457*, has an ORF of 1374 bp with no homology to known genes. Its annotation however, is supported by a perfect match (*E* value 0) with EST contig CON_001_15864 compiled from one EST sequence present in a library of mating cultures (Randall et al., in press). The other putative gene was designated *kre6-like* because its deduced amino acid sequence shows homology to the yeast protein KRE6, a protein involved in the synthesis of β-glucan (P32486). However, *kre6-like* seems to be a pseudogene, having two frameshift mutations that disrupt the ORF.

The order of the four elicitin genes on the sequence contig is *inf1*, *inf6*, *inf3* and *inf4*, and this is in agreement with the order deduced from the physical BAC contig (Fig. 1). All four *inf* genes have perfect EST hits (*E* values 0 with the *P. infestans inf* ESTs mentioned in Materials and methods). Nevertheless, like *kre6-like*, *inf3* seems to be a pseudogene: it has one frameshift mutation and lacks a start codon. In the *P. infestans* EST database, which now consists of ~75,000 ESTs (Randall et al., in press) there are only four *inf3* ESTs, suggesting that *inf3* is expressed at a lower level than the other six elicitin genes in the contig with 13–352 ESTs. The intergenic region between *inf3* and *ORF457* is rather short (221 bp from the TAA stop codon of *inf3* to the ATG start codon of *ORF457*) and *inf3* and *ORF457* are located in the most gene dense region of the BAC, together with *inf6*. *ORF457* is 29.2 kb away from its other neighboring gene, *kre6-like* (Fig. 2). The presence of such a gene island illustrates the uneven gene density in the *P. infestans* genome and explains the discrepancy between a projected genomic coding capacity of 13,000 genes based on the overall gene density on BAC11A5 (this study) and the recently predicted 18,000 unigenes based on EST analysis (Randall et al., in press). Surveys of other genomic regions in *P. infestans* and of the assembled *P. sojae* and *P. ramorum* genome sequences (accessible via http://www.jgi.doe.gov/) revealed that gene islands are very common in *Phytophthora* species (data not shown).

The four elicitin genes and the two putative genes all have the 19-bp core promoter consensus sequence that spans the transcription start site in several oomycete genes (Table 1) (McLeod et al. 2004). Neither the elicitin

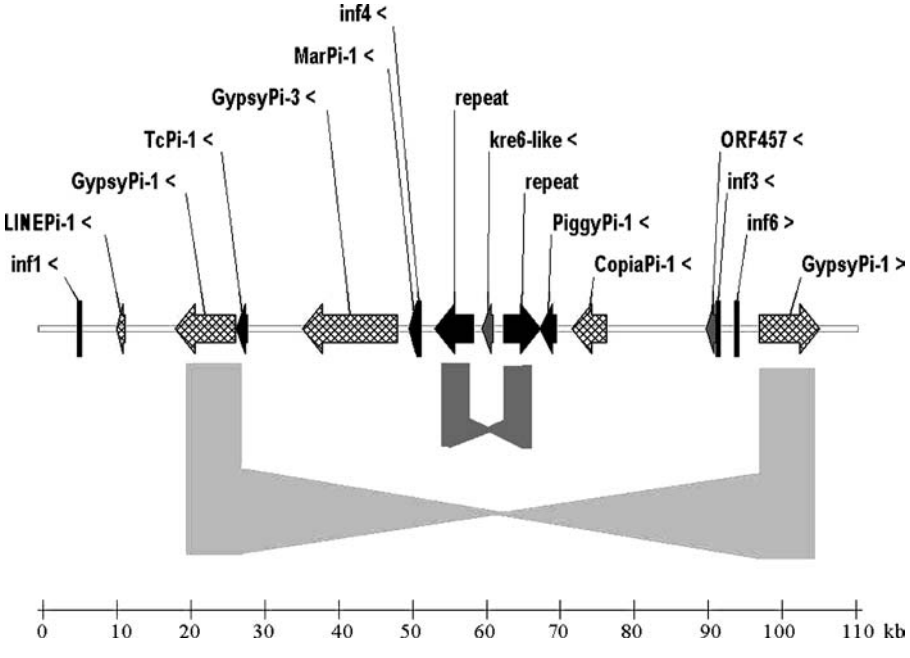

Fig. 2 Genes and mobile elements present on BAC11A5. In this schematic drawing of the 112.3-kb sequence, all elements are shown to scale (see *scale bar*). The symbols > and < following the codes show the orientation of the genes and mobile elements, which are indicated by the *arrowed boxes*. The 8.3-kb and 5.3-kb repeats are connected by the *light* and *dark gray arrowed blocks*, respectively

Table 1 Genes present on BAC11A5 and the GC content in each reading frame

Genes[a]	Reading frame	Protein length (aa)	ORF location (bp)	GC content(%)[b]			EST hits	TI distance (bp)[c]	Transcriptional initiation site[d]
				Frame 1	Frame 2	Frame 3			
Inf1	1	118	5505–5862	43.70	54.24	83.90	318	37	**TTCCATTGTGCAATTTGCT**
Inf4	2	118	51885–52242	40.34	48.31	67.80	13	38	**CCTCATTCCGCAATTTCCA**
kre6-like[e]	3	475	60681–62112	45.38	49.89	50.74	None	94	**TAGCCCACTCTAATTTTCG**
ORF457	1	457	91284–92658	55.90	44.64	56.24	1	16	**TTCACAGCTCAAACTTGTG**
Inf3[f]	2	188	92880–93447	40.21	68.09	72.34	4	30	**TCTCACTCTGCAATCTGCT**
Inf6	2	183	95517–96069	59.24	59.56	65.57	352	44	**TGCCATTCTCCAATTTGCT**

[a] The six genes and ORFs are ordered according to their position in the 112.3 kb sequence contig. GenBank Accession Nos. AY830094 (*inf1*), AY830095 (*inf4*), AY830093 (*kre6-like*), AY830097 (*ORF457*), AY830092 (*inf3*), AY830096 (*inf6*)
[b] The highest percentage is shown *underlined*
[c] Distance from the predicted transcription initiation site to the start codon
[d] As described by Pieterse et al. (1994) and McLeod et al. (2004); the residues that match the consensus (TNSCAWTCTS-CAATTTGCW) are shown in *bold face*
[e] Pseudogene; length is calculated after frameshift correction
[f] Pseudogene; length is calculated after frameshift correction and after adding a start codon

genes nor the two putative genes have introns, in agreement with the observation that the majority of identified *Phytophthora* genes lack introns (Kamoun 2003).

High GC content in the third codon position as a general feature of coding sequences in *Phytophthora*

Phytophthora genes generally show a high GC content (Qutob et al. 2000; Hraber and Weller 2001). Due to the redundant nature of the codons, this feature is expected to be more pronounced at the third positions of codons (referred to as GC3) (Kamoun and Styer 2000). The coding regions of the six genes identified in BAC11A5 have an average GC content of 55.67% and this is slightly higher than the overall average GC content of (51.60%) of the 112.3-kb sequence contig. In contrast, the GC3 in the coding regions is on average 66.10%, and in all six genes the GC3 is higher than GC1 and GC2 (Table 1). When the GC content of the BAC11A5 sequence, calculated using a sliding window of 300 bp, is plotted against position, a high GC3 in a coding region is visualized as a "GC peak" against the average GC content of 51.60%. This is shown in Fig. 3 with a GC plot of the first 12.5 kb of the BAC sequence containing *inf1*. The *inf1* ORF resides in frame –1, thus the third position of the codon is in GC frame +1/−3, which gives a GC content peak that precisely correlates with the position of the ORF of *inf1* (positions 5507–5864). Scanning the remaining 100 kb of the sequence in a similar way revealed GC peaks at the positions of the *inf3*, *inf4*, *inf6* and *kre6-like* genes. In addition, a number of other distinctive peaks were found that correspond to several transposon and retrotransposon-like sequences, such as *MarPi-1*, *GypsyPi-1*, *GypsyPi-2* and *CopiaPi-1* as described below.

To determine whether a high GC3 is a general feature of *P. infestans* genes, we retrieved the sequences of 79 full-length *P. infestans* genes from GenBank and calculated their GC3 values. They represent a heterogeneous set of genes involved in various biological processes such as metabolism, cellular organization, energy, signal transduction and plant-pathogen interaction. The average GC content of the coding sequences of the genes is 58.15%, whereas the average GC3 is 73.05%. With three exceptions, the highest GC content is always at the third position of the codons (Fig. 4). Even some genes that have an average GC content of less than 52% have a significantly high GC3 (>67%), e.g., the G-protein α subunit gene *pigpa1* (AY050536), a microtubule binding protein gene (gi23394381) and the elicitin gene *inf4* (Table 1). The exception is the *ipiB* gene family, a cluster of three linked genes located on a 5.4-kb genomic fragment with very short intergenic regions (Pieterse et al. 1994). The predicted IPI-B proteins have a high glycine content (36.82%), with 51.64% of the glycine residues being encoded by GGT and 18.85% by GGA, resulting in a high average GC content of ~66%. Hence, the GC3 value in *ipiB* is lower (~48%) than GC1 (~78%) and GC2 (~71%).

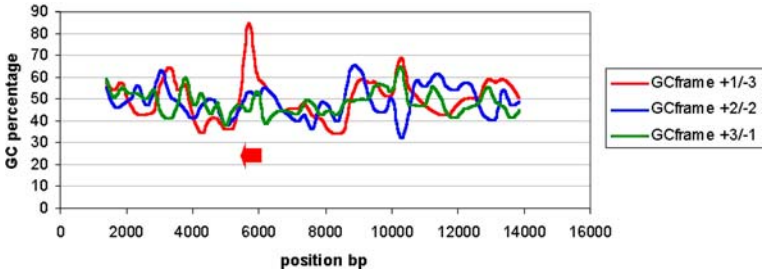

Fig. 3 Pattern of GC content distribution in a 12.5-kb sequence of BAC11A5 in three frames. The percentage GC is plotted against position in the sequence contig. The calculation was done with a sliding window size of 300 bp. The *arrow* indicates the position of *inf1*. The *inf1* ORF resides in frame-1

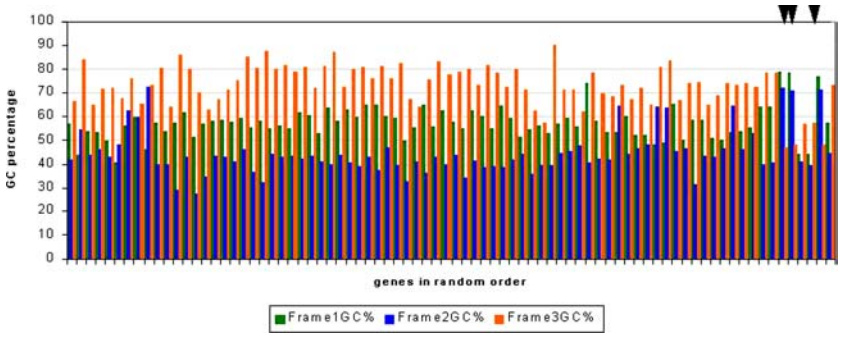

Fig. 4 Visualization of the GC content in three reading frames for 79 *P. infestans* genes retrieved from GenBank. For each gene the percentage GC in each of the three reading frames is plotted in three *bars* (*green, blue and red*) within one *column*. The *filled arrowheads* indicate the columns representing the three *ipiB* genes

To assess the utility of the high GC3 for gene prediction in *Phytophthora* we compared the GC3 of random genomic sequences with that of EST sequences. From the NCBI trace file archive we retrieved a thousand randomly selected genome sequences with an average size of 600 bp from two *Phytophthora* species, *P. infestans* and *P. sojae*. GC3 was calculated for randomly selected frames. For each of the two species we also retrieved 1000 randomly selected, good-quality EST contigs from EST databases. For these EST sequences the GC3 was calculated from the putative ORFs. As shown in Fig. 5, the four sets of sequences give three distinctive peaks, the majority of the genomic sequences of both species have a GC3 lower than 55%, whereas in the *P. infestans* EST sequences the GC3 in most cases exceeds 60% and in *P. sojae* ESTs even 80%. To compare this with the situation in other organisms, a similar calculation was performed on 1000 randomly selected ORFs and genome sequences from two fungi, *F. graminearum* and *A. nidulans*. In both fungi, the GC3 peak of the ORFs is only slightly higher (around 55%) than the GC3 peak of the genome sequences (around 50%).

Each species systematically uses certain synonymous codons in coding sequences. A biased usage of GC rich codons is expected in *Phytophthora* genes. To investigate the relationship between GC3 and codon usage, codon usage was calculated for 79 *P. infestans* genes (Supplementary Table S1). Except for the stop codon with a preference for TAA, and arginine with a slight preference for codons ending with A or T, *P. infestans* prefers to use codons with a G or C at the third position. Codon usage analysis on ORFs derived from 1000 *P. infestans* ESTs and 1000 *P. sojae* ESTs gave similar results: in both species there is a clear preference for codons with a G or C at the third position including the codons for arginine.

The unequal GC distribution combined with a high GC3 in coding regions is a fascinating feature of *Phytophthora* genomes. The finding that GC3 peaks in GC genome scans can reveal the positions of genes (as shown for *inf1* in Fig. 3) justifies exploitation of this feature as a gene discovery tool for *Phytophthora*.

Repeat distribution in BAC11A5

A dot plot analysis of the BAC11A5 sequence revealed repeats of different lengths and in different orientations (data not shown). Most prominent were two large repeat units of 8.2 and 5.3 kb, respectively (Fig. 2). The 8.2-kb repeat unit is a retrotransposon of the Ty3-Gypsy family that we named *GypsyPi-1*. The two *GypsyPi-1* elements, which are 99% similar, are in inverted orientations and located 71.8 kb apart. Since each *GypsyPi-1* element has a pair of long terminal repeats (LTRs), there are four copies of the 345-bp LTR in the sequence contig. In the other large repeat unit of 5.3 kb no transposon-related elements could be identified. The 5.3-kb repeats are also in inverted orientation and separated by a 4.1-kb stretch that contains the *kre6-like* pseudogene.

In addition to the larger repeats the sequence contig contains a number of smaller tandem repeats and inverted repeats. One 33-bp fragment is repeated twice in tandem with 100% similarity, and five inverted repeats

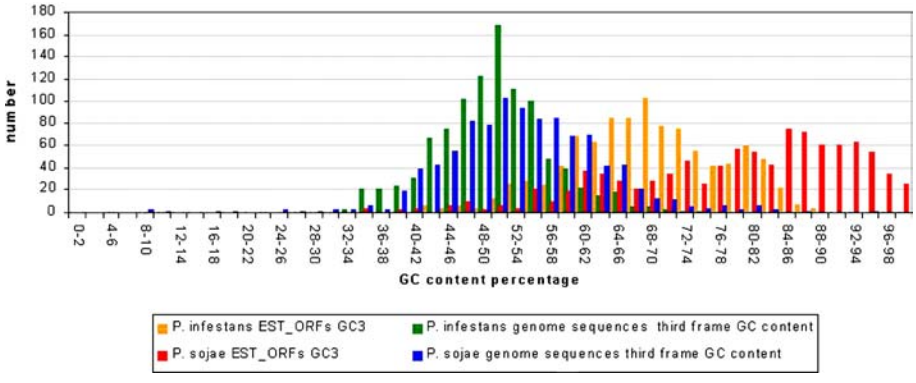

Fig. 5 Mean GC3 values based on 1000 randomly selected EST ORFs and 1000 randomly selected genomic sequences from *P. infestans* and *P. sojae*. The *y-axis* shows the number of EST ORFs or genomic fragments that have the GC3 value indicated on the *x-axis*

ranging in length from 35 to 76 bp display pairwise identity levels of 86–100%. Interestingly, three of the five inverted repeats are located in promoter regions. A perfect 35-bp inverted repeat (interrupted by a 1-bp spacer) is located 508 bp upstream of the *inf1* ORF, and similarly, a 45-bp perfect inverted repeat (including a 1-bp spacer), is located 675 bp upstream of the *inf4* ORF. Also the promoter region of the *kre6-like* pseudogene contains an inverted repeat. It is located 900 bp upstream of the ORF, and is 76 bp in length; the inverted repeats are 87% identical. It is conceivable that these palindrome-like sequences have functions in regulating gene expression but this remains to be determined.

BAC11A5 contains three different Class II transposons

Transposable elements are mobile DNA sequences that can move from one genomic location to the other. They are classified into two groups according to their transposition intermediate. Class I elements transpose via RNA, and a reverse transcriptase is needed to convert the RNA intermediate into DNA for transposition. Class II elements transpose directly as DNA molecules and no RNA intermediate is needed (Feschotte et al. 2002). They are characterized by Terminal Inverted Repeats (TIRs), and by a transposase gene located between the TIR borders.

Within the BAC11A5 sequence contig there are three class II elements. They range in size from 1.5 to 2.5 kb and account for 5.7% of the BAC sequence (Fig. 3, Table 2). Their transposase sequences are highly diverged, and based on BLAST homology and phylogenetic analysis these three transposons belong to three different families, i.e., Tc1/mariner, Pogo and PiggyBac. The Tc1/mariner type of transposons (Plasterk 1996) together with Pogo transposons (Tudor et al. 1992; Smit and Riggs 1996) is probably the most widespread class II elements. Figure 6 shows a phylogenetic tree constructed on the basis of an alignment of the transposase regions of various characterized Tc1/mariner and Pogo elements and two of the three class II elements identified in BAC11A5, named *MarPi-1* and *Tc1Pi-1*, respectively. *MarPi-1* groups with the Mariner like transposons (Fig. 6) and the 200-amino acid transposase displays the highest BLASTP homology with a putative transposase from rice (AC093017_21; *E* value 2e-08). No evidence for the TIRs expected to flank the *MarPi-1* transposase could be found. *Tc1Pi-1* is a Tc1 like transposon (Fig. 6). The 375-residue transposase sequence, obtained after correcting for a frameshift mutation, shows the highest BLASTP homology to a putative transposase from *Anopheles gambiae* (XP_310448; E value 1e-08). The TIRs of *Tc1Pi-1* are inverted repeats of 120 bp flanking the transposase; the TIRs show 61.3% similarity to each other. The third class II element found in BAC11A5 belongs to the PiggyBac family and was named *PiggyPi-1*. Members of this family of transposases are mainly found in animals and are related to the transposase of the canonical piggyBac transposon from the moth *Trichoplusia ni* (Sarkar et al. 2003). They show no obvious homology to other transposon families. The deduced 724 amino acid sequence of the *PiggyPi-1* transposase shows the highest BLASTP homology to that of the piggyBac transposable element of *Homo sapiens* (NP 689808.2; *E* value 9e-15). As in the case of *MarPi-1* no TIRs could be detected.

BAC11A5 harbors a diverse group of retrotransposons

Class I elements are also referred to as retroelements or retrotransposons, because they transpose via an mRNA intermediate synthesized by an indispensable reverse transcriptase activity. Class I elements are classified into LTR retrotransposons, Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear elements (SINEs) (Baltimore 1985; Echalier 1989; Flavell et al. 1992). BAC11A5 contains three LTR retrotransposons and one LINE.

LTR retrotransposons are specified by their direct long terminal repeats. The LTRs typically bracket several genes, among which the two major genes called *gag* and *pol*. A number of proteins such as protease (PR), reverse transcriptase (RT), integrase (INT) and RNaseH (RH) can be encoded by the *pol* gene. Based on the sequence divergence of reverse transcriptases and also the order of RT and INT coding domains, LTR

Table 2 Characteristics of three different class II elements present on BAC11A5

Element[a]	Group	Size (kb)	TIR (bp)	Number of copies in BAC11A5	TIR pair similarity (%)	GC content (%)[b]			EST hits[c]	Homologue in *P. sojae*[d]	Homologue in *P. ramorum*[d]
						Frame 1	Frame 2	Frame 3			
MarPi-1	Mariner	~1.5		1		58.00	44.72	58.29	36	No	No
Tc1Pi-1	Tc1	~1.5	120	1	63.1	56.68	43.09	63.82	0	No	No
PiggyPi-1	Piggy	~2.5		1		50.00	44.40	43.02	0	No	No

[a] GenBank Accession Nos. AY830109 (*MarPi-1*), AY830110 (*Tc1Pi-1*), AY830111 (*PiggyPi-1*)

[b] The values for GC content refer to the transposase coding regions. The highest percentage is shown in *underlined*

[c] Based on BLASTN searches against 72,904 *P.infestans* ESTs (*E* value less than e^{-100} and percentage identity higher than 95%)

[d] Based on BLASTN search against *P. sojae* and *P. ramorum* genome sequences; *E* value less than e^{-100} and percentage identity higher than 80%

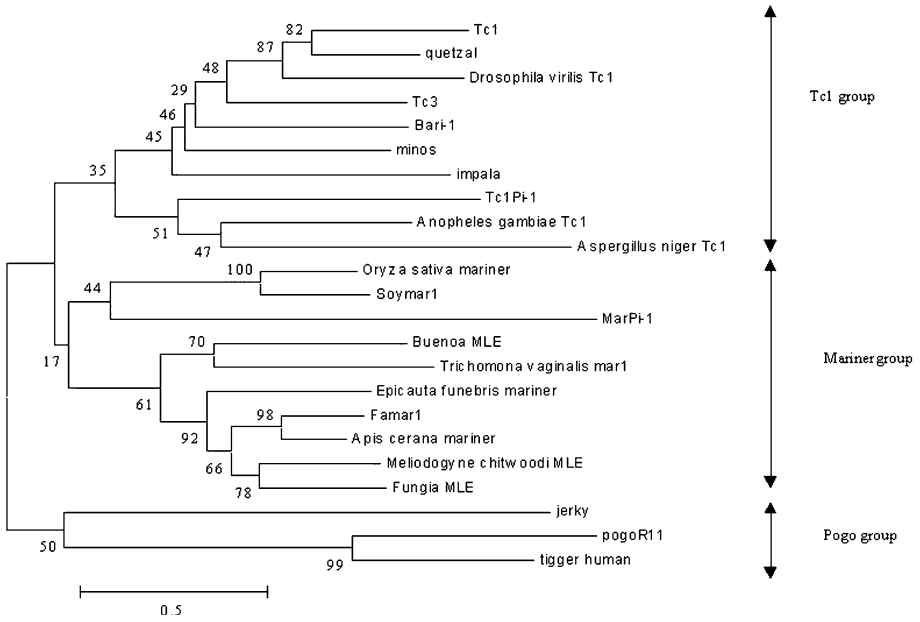

Fig. 6 Phylogenetic tree of class II elements belonging to the Tc1/mariner and Pogo groups. The transposase protein sequences including the conserved DDE regions were used to construct the unrooted tree based on Neighbor-Joining analysis. The robustness of groupings was estimated by using 1000 bootstrap replicates; the *numbers* next to branching points indicate the percentage of replicates supporting each branch. The sequences used for the phylogenetic tree, their GenBank Accession Nos. and their species sources were the following: *A. gambiae Tc1*, XP_310448.1 [*A. gambiae*]; *Apis cerana mariner*, BAB86288.1 mariner transposase [*A. cerana*]; *Aspergillus niger Tc1*, AAB50684.1 putative Tc1-mariner class transposase [*A. niger*]; *Bari-1*, S33560 transposon-like element Bari-1 [*Drosophila melanogaster*]; *Buenoa MLE*, AAC28142.1 mariner transposase [*Buenoa* sp.]; *Drosophila virilis Tc1*, AAA88882.1 Tc1-like transposase [*D. virilis*]; *Epicauta funebris mariner*, AAC28145.1 mariner transposase [*E. funebris*]; *Famar1*, AAO12863.1 Famar1 transposase [*Forficula auricularia*]; *Fungia MLE*,BAB32436.1 transposase [*Fungia* sp. *Kusabiraishi*]; *impala*, AAB33090.2 transposase [*Fusarium oxysporum*]; *jerky*, NP_03241.3jerky [*Mus musculus*]; *Meliodogyne chitwoodi MLE*, CAD26968.1 transposase [*M. chitwoodi*]; *minos*, S26856 transposon Minos [*Drosophila hydei*]; *pogoR11*, S20478 transposon pogoR11 [*D. melanogaster*]; *quetzal*, AAB02109.1 transposase [*Anopheles albimanus*]; *O. sativa mariner*, AC093017_21putative transposase [*O. sativa*]; *Soymar1*, AAC28384.1 mariner transposase [*Glycine max*]; *Tc1*, P03934 TC1A_CAEELTransposable element TC1 transposase [*Caenorhabditis elegans*]; *Tc3*, P34257 TC3A_CAEEL Transposable element TC3 transposase [*C. elegans*]; *tigger human*, AAH37869.1 Tigger transposable element derived 4 [*H. sapiens*] and *Trichomonas vaginalis mar1*, AAP45328.1 mar1 putative transposase [*T. vaginalis*]. *MarPi-1* and *Tc1Pi-1* are two of the three *P. infestans* class II elements identified in this study

retrotransposons are further divided into two groups, the Ty1/copia and the Ty3/gypsy group (Xiong and Eickbush 1990). Figure 7 shows a phylogenetic tree constructed on the basis of the RT domains of various identified retroelements and the three LTR elements identified in BAC11A5. Two appear to belong to the Ty3/gypsy group, while the third one belongs to the Ty1/copia group. One of the Ty3/gypsy like elements, called *GypsyPi-1*, is located on the 8.2 kb repeat unit described above, and hence two *GypsyPi-1* copies are present in BAC11A5. These two copies share 99.4% similarity at the nucleotide level. One copy has a frameshift mutation in the *pol* gene; the other copy still possesses intact *gag* and *pol* genes. A GAG protein of 328 amino acids, together with a 1517-residue POL protein comprised of PR, RT, RH and INT domains are encoded by two ORFs flanked by a pair of LTRs. The second Ty3/gypsy like element, named *GypsyPi-2*, is 13.0 kb in length with LTRs of 530 bp. The Ty1/copia like retrotransposon, named *CopiaPi-1*, is 4.7 kb in length with LTRs of 220 bp. *CopiaPi-1* shows the domain order RT-INT which is for typical Ty1/copia elements, in contrast to the INT-RT domain order in the two Tc3/gypsy like elements.

Due to their sequence divergence, LINEs are considered to be the most ancient group of transposable elements (Xiong and Eickbush 1990). LINEs lack the terminal direct repeat and possess a polyadenylation signal in the 3′end of the sequence. One LINE of 3.2 kb was identified in the BAC11A5 sequence and was named *LINEPi-1*. Like other identified LINEs (Noma et al. 1999; Schmidt 1999), *LINEPi-1* codes for a RT but lacks INT. A protein with endonucleolytic activity is encoded upstream of the RT gene. The 3′ polyadenylation signal A_8 is found 80 bp downstream of the stop codon of the *pol* gene.

The four more-or-less intact retroelements and the LINE make up 33.2% (37.3 kb) of the sequence contig. This figure underestimates the total fraction of class I elements in this BAC, since a number of fragmented elements have not been taken into account.

The distribution of transposable elements in *Phytophthora*

To investigate whether other regions of the genome contain a similar distribution of transposable elements, a number of BAC sequences were retrieved from

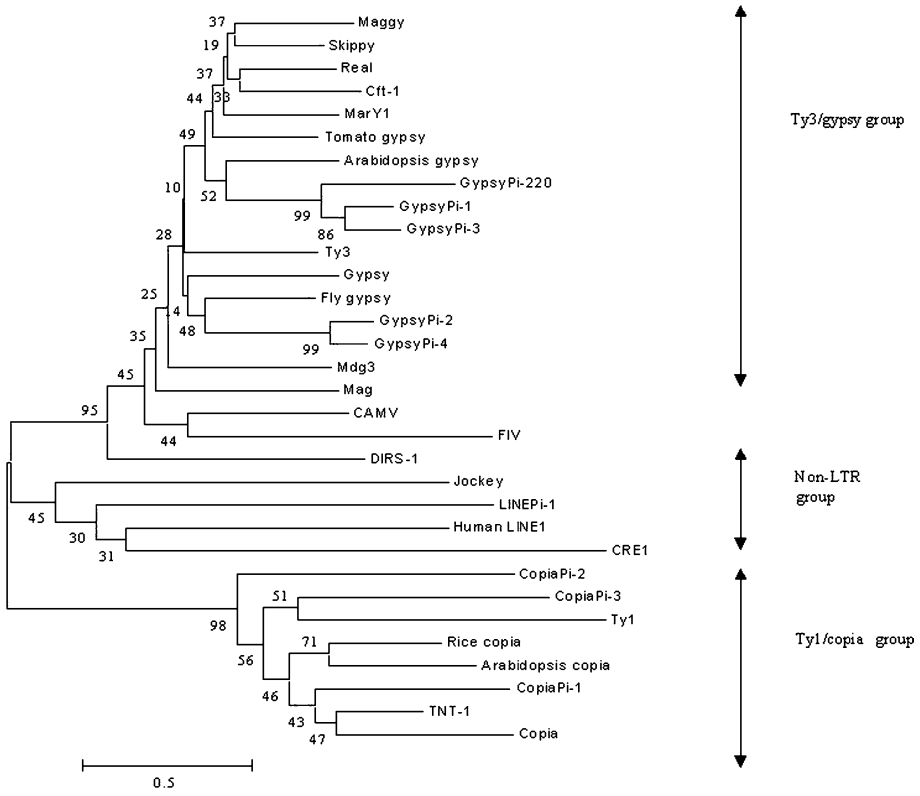

Fig. 7 Phylogenetic tree of retroelements. The reverse transcriptase protein sequences were used to construct the unrooted tree based on Neighbor-Joining analysis. The robustness of groupings was estimated using 1000 bootstrap replicates; the *numbers* next to branching points indicate the percentage of replicates supporting each branch. On the *right* different groups of retrotransposons are indicated. The sequences used and their source species were the following: Arabidopsis copia, BAB84015.1 polyprotein [*A. thaliana*]; Arabidopsis gypsy, AF128395 retrotransposon [*A. thaliana*]; CAMV, M90543 reverse transcriptase [Cauliflower Mosaic Virus]; *Cft-1*, AAF21678 pol polyprotein [*C. fulvum*]; *Copia*, P04146 copia protein [*D. melanogaster*]; *Cre1*, A34728 transposon CRE1 [*Crithidia fasciculata*]; *DIRS-1*, C24785 DIRS-1element [*D. discoideum*]; *FIV*, S23820 pol polyprotein [Feline Immunodeficiency Virus]; *Gypsy*, AAB50148, polyprotein [*D. melanogaster*]; *GypsyPi-220*, AF490339 strain 220 gypsy-like retrotransposon [*P. infestans*]; *human LINE1*, P08547 HUMAN LINE-1 HOMOLOG [*H. sapiens*]; *Jockey*, P21328 mobile element jockey [*D. melanogaster*]; *Mag*, S08405 silkworm transposon mag [*Bombyx mori*]; *Maggy*, AAA33420 polyprotein [*Magnaporthe grisea*]; *MarY1*, BAA78625 polyprotein [*Tricholoma matsutake*]; *Mdg3*, T13798 retrotransposon mdg3 [*D. melanogaster*]; *REAL*, BAA89272 polyprotein Pol [*Alternaria alternata*]; *Rice copia*, AAR88589.1 putative copia-like retrotransposon protein [*O. sativa*]; *Skippy*, S60179 retrotransposon skippy [*F. oxysporum*]; *Tnt1*, P10978 Tnt-1element [*Nicotiana tabacum*]; *Tomato gypsy*, T17459 Gypsy-like polyprotein [*Lycopersicon esculentum*]; *Ty1*, B2267 retrotransposon Ty9121 [*Saccharomyces cerevisiae*] and *Ty3*, S69842 Ty3 protein [*S. cerevisiae*]. *GypsyPi-1*, *GypsyPi-2*, *GypsyPi-3*, *GypsyPi-4*, *CopiaPi-1*, *CopiaPi-2* and *CopiaPi-3* are the seven *P. infestans* retrotransposons identified in this study

GenBank and analyzed for the presence of retrotransposons. In total, 500 kb of genomic DNA sequence composed of fragments from five partially sequenced BACs was compiled and fragments without unordered gaps were used for the analysis. The BACs which are derived from the same library and the same strain as BAC11A5 are PI-BAC-14M19 (AC146943), PI-BAC-14P22 (AC146983), PI-BAC-21G17 (AY497062), PI-BAC-25C5 (AC147181) and PI-BAC-26O7 (AC147180) and cover randomly selected regions containing genes with significant homology to a variety of known sequences, e.g., β-glucosidase, deoxyribose-phosphate aldolase and elongation factor. In the 500-kb fragments two further copies of *GypsyPi-1* were found, each sharing 97% homology at the nucleotide level with the copy found in BAC11A5. In addition, two new Ty3/gypsy like retrotransposons, *GypsyPi-3* and *GypsyPi-4*, and two new Ty1/copia like retrotransposons, *CopiaPi-2* and *CopiaPi-3*, were identified. The features of the seven retroelements and one LINE are summarized in Table 3 and Fig. 8.

The retroelements identified in this study are not identical to any of the previously described Ty1/copia and Ty3/gypsy elements in *Phytophthora*. Tooley and Garfinkel (1996) identified a number of Ty1/copia like elements in *P. infestans* by degenerate PCR but the partial sequences are not suitable for phylogenetic analysis. By using a similar approach, Judelson (2002) obtained partial Ty3/gypsy sequences from several *Phytophthora* species, and one of these, *P. infestans GypsyPi-220*, was analyzed in more detail. *GypsyPi-220* is closely related to *GypsyPi-1* and *GypsyPi-3* with 56 and 60% similarity, respectively and in the phylogenetic tree these three retroelements seem to form a subclass of *P. infestans* specific Tc3/gypsy like retrotransposons (Fig. 7).

To investigate whether the transposable elements found in BAC11A5 also exist in other *Phytophthora*

Table 3 Characteristics of the different retrotransposons described in this study

Element[a]	Group	Size (kb)	LTR (bp)	Number of copies in BAC11A5	Number of copies in random 500-kb segment	LTR pair similarity (%)	GC content (%)[b]			EST hits[c]	Homologue in *P. sojae*[d]	Homologue in *P. ramorum*[d]
							Frame 1	Frame 2	Frame 3			
GypsyPi-1	Gypsy	8.2	350	2	4	99.7; 99.7; 100.0; 100.0	57.55	41.88	72.92	4	Yes	Yes
GypsyPi-2	Gypsy	13.0	530	1	1	97.7	54.61	38.01	63.47	1	No	No
GypsyPi-3	Gypsy	7.6	314	0	2	94.2/99.4	53.60	42.24	71.48	0	Yes	Yes
GypsyPi-4	Gypsy	~12[e]	540	0	1	98.0	54.51	40.23	72.56	1	Yes	Yes
CopiaPi-1	Copia	4.7	220	1	1	91.4	52.41	39.66	64.01	0	Yes	No
CopiaPi-2	Copia	5.7	240	0	1	81.3	50.89	43.42	37.86	1	Yes	Yes
CopiaPi-3	copia	5.5	238	0	1	98.3	67.94	36.24	45.80	0	No	No
LINEPi-1	LINE	3.2	None	1	ND	None	50.17	44.07	52.04	1	No	No

[a] The retrotransposons listed are found either on BAC11A5 or in a random set (500 kb) of genomic sequences (see Materials and methods). GenBank Accession Nos. AY830091 (*GypsyPi-* 1), AY830106 (*GypsyPi-2*), AY830104 (*GypsyPi-3*), AY830107 (*GypsyPi-4*), AY830098 (*CopiaPi-1*), AY830099 (*CopiaPi-2*), AY830100 (*CopiaPi-3*), AY830108 (*LINEPi-1*)

[b] The GC content calculation was performed on the reverse transcriptase regions of the retrotransposons. The highest percentage is shown *underlined*

[c] Based on BLASTN searches against 72,904 *P. infestans* ESTs (*E* value less than e^{-100} and percentage identity higher than 95%)

[d] Based on BLASTN search against *P. sojae* and *P. ramorum* genome sequences (*E* value less than e^{100} and percentage identity higher than 80%)

[e] Due to a gap in the GenBank sequence the size of *GypsyPi-4* can not be determined exactly

genomes, NCBI trace files containing 1,533,511 *P. sojae* genome sequences and 898,494 *P. ramorum* sequences were searched using BLASTN. Hits with more than 80% homology and *E* values lower than e-100 were considered to be homologous. Neither the LINE element nor any of the DNA transposons was found in *P. sojae* or *P. ramorum* (Tables 2, 3). In contrast, several LTR retrotransposons of both the Tc3/gypsy and Tc1/copia classes were detected (Table 3), indicating that these elements are more widely distributed in the genus and invaded the *Phytophthora* genome before speciation. Phylogenetically *P. infestans* is not closely related to *P. sojae* or *P. ramorum*; the tree species fall into three different clades (Kroon et al. 2004). The class II elements could well be clade-specific and, hence, be present in more closely related species within the *P. infestans* clade, such as *P. mirabilis* and *P. phaseoli*. This is true for *DodoPi*, a recently described *P. infestans* hAT-like DNA

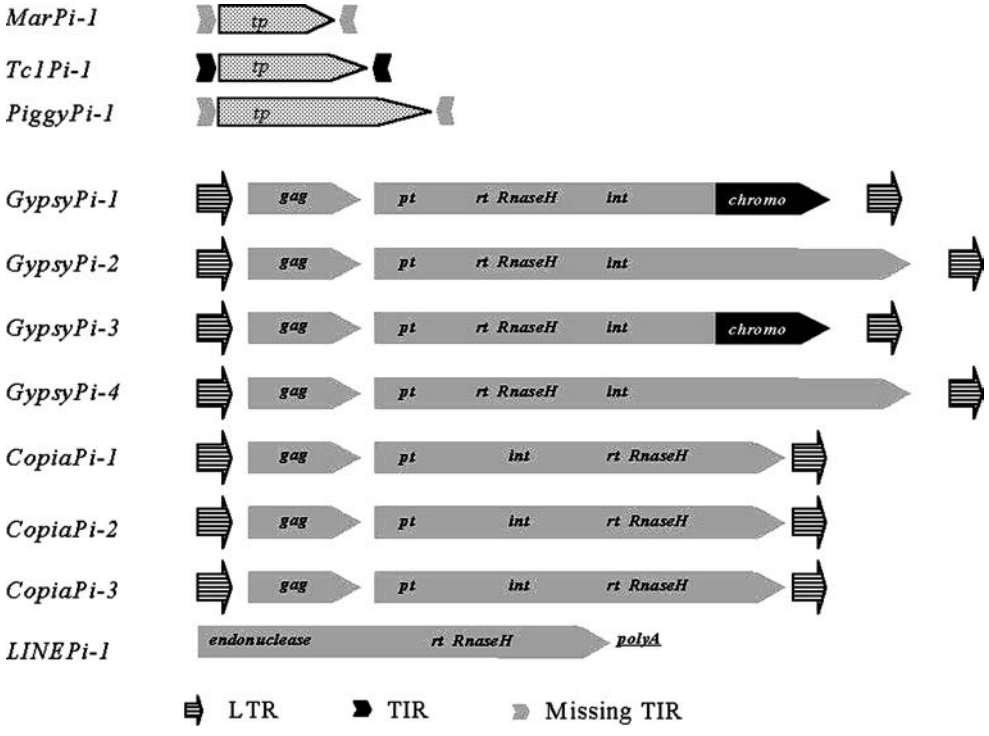

Fig. 8 Schematic representation of mobile elements described in this study. The sizes of the elements are not shown to scale, but are listed in Tables 2 and 3. *TIR* Terminal Inverted Repeat; *LTR* Long Terminal Repeat; *tp* transposase; *pt* protease; *rt* reverse transcriptase; *int* integrase; *chrom* CHROMO domain

transposon that is not related to the class II transposons described in this study (Ah Fong and Judelson 2004).

GypsyPi-1 and *GypsyPi-3* belong to a class of retrotransposons that code for POL proteins with a CHROMO domain

GypsyPi-1 and *GypsyPi-3* are closely related: the similarity at the protein sequence level between their reverse transcriptase regions is 71%. The C-terminal portions of the POL proteins deduced from the *GypsyPi-1* and *GypsyPi-3* sequences both contain a CHROMO (CHRromatin Organization MOdifier) domain (PF00385). The closely related *GypsyPi-220* is truncated at the C-terminus and also lacks the LTRs (Judelson 2002). Domain searches using these protein sequences against the Pfam database (Bateman et al. 1999, 2000) gave hits with this CHROMO domain with an *E* value of 1e-08. CHROMO domains are associated with alteration of the structure of chromatin to the condensed morphology of heterochromatin (Cavalli and Paro 1998). Proteins with CHROMO domains can often modify the structure of chromatin. Examples include the *Drosophila* and human heterochromatin protein Su (HP1) (Aasland and Stewart 1995), the *Drosophila* protein Polycomb (Pc) (Paro and Hogness 1991) and mammalian DNA-binding/helicase proteins (Koonin et al. 1995). A POL protein including a CHROMO domain is not unprecedented. A CDART (Conserved Domain Architecture Retrieval Tool) (Geer et al. 2002) search revealed several other retroelements with a CHROMO domain at the C-terminal end of POL, and a similar protein architecture to *GypsyPi-1* and *GypsyPi-3*; among these are *CfT-1* from the fungus *Cladosporium fulvum* (AAF21678), *Skipper* from *Dictyostelium discoideum* (T14598), a retrotransposon from the fish *Takifugu rubripes* (AAC33526) and two plant retrotransposons (*Oryza sativa* NP_920591and *Arabidopsis thaliana* NP_683628). It will be interesting to discover to what extent this CHROMO domain influences the retrotransposition events, and whether they modify chromatin structure in the host.

Activity of the transposable elements

The genomic region covered by BAC11A5 seems to be a hotspot for retrotransposon and DNA transposon insertions. However, several of the mobile elements identified in BAC11A5 and in the 500 kb of random genomic DNA sequence, carry numerous mutations and deletions indicating that the majority of the transposons are inactive. Nevertheless, EST database searches show the occurrence of transcripts of a diverse group of mobile elements in various developmental stages of the life cycle of *P. infestans*, demonstrating that at least some transposons are actively transcribed. From *GypsyPi-1*, *GypsyPi-2*, *CopiaPi-2* and *LINEPi-1* up to four transcripts are found in the EST database with a homology in the range of 95–99% at nucleotide level (Table 3). *MarPi-1* is represented by36 transcripts with more than 95% homology and distributed over EST libraries from different developmental stages (Table 2).

As described above, the GC3 of the ORFs in *P. infestans* genes is generally high. To determine whether the ORFs present in the transposable elements have the same characteristic we performed a GC analysis of the 800-bp ORF encoding RT in *GypsyPi-1*, *GypsyPi-2*, *GypsyPi-3*, *GypsyPi-4*, *CopiaPi-1*, *CopiaPi-2*, *CopiaPi-3* and *LINEPi-1* and the transposase ORF of *MarPi-1*, *Tc1Pi-1* and *PiggyPi-1*. The results are shown in Tables 2 and 3. All four Tc3/gypsy elements, one Tc1/copia like element (*CopiaPi-1*) and one DNA transposon (*Tc1Pi-1*) have a high GC3 value like most *P. infestans* genes. The GC3 content is above 60% and is higher than GC1 and GC2. In contrast, *CopiaPi-2*, *CopiaPi-3*, *LINEPi-1*, *MarPi-1* and *PiggyPi-1* either do not show a high GC3 or the GC3 value is not the highest of the three.

GypsyPi-1 and *GypsyPi-3* not only show the high GC3 feature, but also have a similar codon usage to *Phytophthora* genes. Codon usage was calculated for the deduced GAG and POL proteins of *GypsyPi-1* and *GypsyPi-3* and the same calculation was performed on several sets of ORFs deduced from 1000 randomly selected ESTs each from *P. infestans* and *P. sojae*. A high correlation was found for the codon usage in *P. infestans* and *P. sojae* ORFs (regression line $y = 0.96\ x + 1.38$; $R^2 = 0.92$) and in *P. infestans* ORFs and *GypsyPi-1*/*GypsyPi-3* ORFs (regression line $y = 0.88\ x + 3.78$; $R^2 = 0.73$) (data not shown).

The LTRs of LTR retrotransposons are generated during the replication and integration process as a pair of identical sequences (Boeke and Corces 1989). The divergence of this pair of sequences indicates the time elapsed since the last transposition event: the more divergent the LTR pair is, the longer ago the transposition event occurred. The sequence similarity was calculated for the LTR pairs of the four Tc3/gypsy elements and three Tc1/copia elements. The highest sequence similarity is found for the pairs of the four *GypsyPi-1* copies, ranging from 99.4 to 100.0% whereas the lowest similarity (81.3 %) is found in *CopiaPi-2* (Table 3). This indicates that *GypsyPi-1* transposed more recently then the other retrotransposons identified in this study. Like *GypsyPi-1*, *GypsyPi-3* is probably a relatively 'young' retroelement. Both seem to be widespread because four and two copies of *GypsyPi-1* and *GypsyPi-3*, respectively, were found in 500 kb of random genomic fragments and the sequences of the copies located at different positions show high similarity in their LTRs as well as coding regions.

It is remarkable that both the 'young' retroelements *GypsyPi-1* and *GypsyPi-3* show a high GC3 and a codon usage that is similar to that of other *P. infestans* genes. Maybe these elements have already resided in *P. infestans* since the early stage of *P. infestans* evolution and

gradually acquired the characteristics of host genes, allowing a more efficient use of host cellular machinery during replication and transposition. Indeed *GypsyPi-1* is transcribed, as demonstrated by the identification of *GypsiPi-1* ESTs.

Conclusions

Physical mapping of BACs, BAC sequencing and annotation of a long contiguous stretch of genomic DNA of *P. infestans* showed that members of the elicitin gene family are clustered in the genome but yet dispersed over a large region of 200–250 kb that harbours repeats and numerous transposable elements. Two of the four *inf* genes, *inf3* and *inf6*, reside on a gene island with one other gene of unknown function, whereas the two other *inf* genes, *inf1* and *inf4*, are 46 kb apart, and 86 and 40 kb, respectively, away from the gene island. Comparison of the coding and non-coding sequences showed that the GC content of the coding regions is slightly higher. More significant was the high GC content of the third base of a codon in an ORF, the GC3, a characteristic feature that can be used in gene prediction programs. In the promoter regions a few putative regulatory elements were found, but as yet the relevance of these elements is unknown.

Transposons and retrotransposons are ubiquitous in various kingdoms, such as fungi, plants, ciliates and animals (Kim et al. 1998; Daboussi and Capy 2003). Due to the difference in their mode of transposition, class I and class II elements are thought to contribute differently to genome size (Kumar and Bennetzen 1999). Class II DNA transposons use a "copy/cut-paste" mode of transposition, while class I retroelements transpose via an RNA intermediate step and can therefore potentially be propagated in large numbers. One-third of the BAC11A5 sequence contig consists of class I retroelements, and other genomic regions in *P. infestans* also contain numerous retroelements (this study; Judelson 2002; Tooley and Garfinkel 1996). Without a genome sequence it is, as yet, not possible to calculate the overall percentage of transposon sequences in the *P. infestans* genome, but it seems likely that transposons are, at least in part, responsible for its large size of 240 Mb.

With two *Phytophthora* genomes sequenced (http://www.jgi.doe.gov/), comparative genomic studies of the genus are now within reach. Efforts are currently focused on finding distinctive features in the genomes of various *Phytophthora* species and developing gene annotation tools. This present study provides a small-scale inventory of genome organization and genome structure in *P. infestans*, and represents a first step into the annotation process. With a large EST repository (Randall et al., in press) and a draft genome sequence of *P. infestans* in hand (O'Neill et al. 2004) *P. infestans* is ready to enter the genomics arena.

Acknowledgments We are grateful to Sharmili Mathur for expert technical assistance, Steve Whisson for providing the BAC library and filters, Grardy van den Berg for screening the BAC library, and Pierre de Wit for critically reading the manuscript. This work was financially supported by NWO-Aspasia Grant No. 015.000.057 and USDA Cooperative Agreement No. 58-8230-6-081. The authors acknowledge Syngenta for access to the Syngenta *Phytophthora* Consortium EST Database, and the Broad Institute and the DOE Joint Genome Institute for depositing random genomic sequences of *P. infestans* and *P. sojae*, respectively, in the NCBI trace file archive.

References

Aasland R, Stewart AF (1995) The chromo shadow domain, a second chromo domain in heterochromatin-binding protein 1, HP1. Nucleic Acids Res 23:3168–3173

Ah Fong AM, Judelson HS (2004) The hAT-like DNA transposon *DodoPi* resides in a cluster of retro- and DNA transposons in the stramenopile *Phytophthora infestans*. Mol Genet Genomics 271:577–585

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Baldauf SL (2003) The deep roots of eukaryotes. Science 300:1703–1706

Baltimore D (1985) Retroviruses and retrotransposons—the role of reverse transcription in shaping the eukaryotic genome. Cell 40:481–482

Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer ELL (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 27:260–262

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer ELL (2000) The Pfam protein families database. Nucleic Acids Res 28:263–266

Boeke JD, Corces VG (1989) Transcription and reverse transcription of retrotransposons. Annu Rev Microbiol 43:403–434

Bonfield JK, Smith K, Staden R (1995) A new DNA sequence assembly program. Nucleic Acids Res 23:4992–4999

Cavalli G, Paro R (1998) Chromo-domain proteins: linking chromatin structure to epigenetic regulation. Curr Opin Cell Biol 10:354–360

Daboussi MJ, Capy P (2003) Transposable elements in filamentous fungi. Annu Rev Microbiol 57:275–299

Duclos J, Fauconnier A, Coelho AC, Bollen A, Cravador A, Godfroid E (1998) Identification of an elicitin gene cluster in *Phytophthora cinnamomi*. DNA Seq 9:231–237

Echalier G (1989) Drosophila retrotransposons—interactions with genome. Adv Virus Res 36:33–105

Erwin DC, Ribeiro OK (1996) *Phytophthora* diseases worldwide. The American Phytopathological Society, St. Paul, Minn.

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nature Rev Genetics 3:329–341

Flavell AJ, Smith DB, Kumar A (1992) Extreme heterogeneity of Ty1-Copia group retrotransposons in plants. Mol Gen Genetics 231:233–242

Geer LY, Domrachev M, Lipman DJ, Bryant SH (2002) CDART: protein homology by domain architecture. Genome Res 12:1619–1623

Hendrix JW, Guttman SM (1970) Sterol or calcium requirement by *Phytophthora parasitica* var. *nicotianae* for growth on nitrate. Mycologia 62:195–198

Hraber PT, Weller JW (2001) On the species of origin: diagnosing the source of symbiotic transcripts. Genome Biol 2:37

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci 23:403–405

Judelson HS (2002) Sequence variation and genomic amplification of a family of Gypsy-like elements in the oomycete genus *Phytophthora*. Mol Biol Evol 19:1313–1322

Kamoun S (2003) Molecular genetics of pathogenic oomycetes. Eukaryot Cell 2:191–199

Kamoun S, Styer A (2000) An improved codon usage table for *Phytophthora infestans*. http://www.oardc.ohio-state.edu/phytophthora/codon.htm

Kamoun S, van West P, de Jong AJ, de Groot KE, Vleeshouwers VGAA, Govers F (1997) A gene encoding a protein elicitor of *Phytophthora infestans* is down-regulated during infection of potato. Mol Plant Microbe Interact 10:13–20

Kamoun S, van West P, Vleeshouwers VGAA, de Groot KE, Govers F (1998) Resistance of *Nicotiana benthamiana* to *Phytophthora infestans* is mediated by the recognition of elicitor protein INF1. Plant Cell 10:1413–1426

Kamoun S, Hraber P, Sobral B, Nuss D, Govers F (1999) Initial assessment of gene diversity for the oomycete pathogen *Phytophthora infestans* based on expressed sequences. Fungal Genet Biol 28:94–106

Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF (1998) Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res 8:464–478

Koonin EV, Zhou S, Lucchesi JC (1995) The chromo superfamily: new members, duplication of the chromo domain and possible role in delivering transcription regulators to chromatin. Nucleic Acids Res 23:4229–4233

Kroon LP, Bakker FT, Van Den Bosch GB, Bonants PJ, Flier WG (2004) Phylogenetic analysis of *Phytophthora* species based on mitochondrial and nuclear DNA sequences. Fungal Genet Biol 41:766–782

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Annu Rev Genet 33:479–532

Kumar S, Tamura K, Nei M (1994) MEGA—Molecular Evolutionary Genetics Analysis software for microcomputers. Comput Appl Biosci 10:189–191

Latijnhouwers M, de Wit PJGM, Govers F (2003) Oomycetes and fungi: similar weaponry to attack plants. Trends Microbiol 11:462–469

McLeod A, Smart CD, Fry WE (2004) Core promoter structure in the oomycete *Phytophthora infestans*. Eukaryot Cell 3:91–99

Mikes V, Milat ML, Ponchet M, Panabieres F, Ricci P, Blein JP (1998) Elicitins, proteinaceous elicitors of plant defense, are a new class of sterol carrier proteins. Biochem Biophys Res Commun 245:133–139

Noma K, Ohtsubo E, Ohtsubo H (1999) Non-LTR retrotransposons (LINEs) as ubiquitous components of plant genomes. Mol Gen Genet 261:71–79

O'Neill K, Zody MC, Karlsson E, Govers F, van der Vondervoort P, Weide R, Whisson S, Birch P, Ma L, Birren B, Fry W, Judelson H, Kamoun S, Nusbaum C (2004) Sequencing the *Phytophthora infestans* genome: preliminary studies. Abstracts of the Annual Meeting of the NSF *Phytophthora* Molecular Genetics Network. New Orleans, May 21–23, 2004, p. 5

Panabieres F, Marais A, LeBerre JY, Penot I, Fournier D, Ricci P (1995) Characterization of a gene cluster of *Phytophthora cryptogea* which codes for elicitins, proteins inducing a hypersensitive-like response in tobacco. Mol Plant Microbe Interact 8:996–1003

Paro R, Hogness DS (1991) The Polycomb protein shares a homologous domain with a heterochromatin-associated protein of *Drosophila*. Proc Natl Acad Sci USA 88:263–267

Pieterse CMJ, Van West P, Verbakel HM, Brasse P, Van den Berg-Velthuis GCM, Govers F (1994) Structure and Genomic Organization of the *ipib* and *ipio* gene clusters of *Phytophthora infestans*. Gene 138:67–77

Plasterk RH (1996) The Tc1/mariner transposon family. Curr Top Microbiol Immunol 204:125–143

Qutob D, Hraber PT, Sobral BWS, Gijzen M (2000) Comparative analysis of expressed sequences in *Phytophthora sojae*. Plant Physiol 123:243–253

Randall TA et al (2004) Large-scale gene discovery in the oomycete *Phytophthora infestans* reveals likely components of phytopathogenicity shared with true fungi. Mol Plant Microbe Interact, in press

Ricci P, Trentin F, Bonnet P, Venard P, Moutonperronnet F, Bruneteau M (1992) Differential production of parasiticein, an elicitor of necrosis and resistance in tobacco, by isolates of *Phytophthora Parasitica*. Plant Pathol 41:298–307

Sarkar A, Sim C, Hong YS, Hogan JR, Fraser MJ, Robertson HM, Collins FH (2003) Molecular evolutionary analysis of the widespread *piggyBac* transposon family and related "domesticated" sequences. Mol Genet Genomics 270:173–180

Schmidt T (1999) LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. Plant Mol Biol 40:903–910

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W (2000) PipMaker—a web server for aligning two genomic DNA sequences. Genome Res 10:577–586

Skalamera D, Wasson AP, Hardham AR (2004) Genes expressed in zoospores of *Phytophthora nicotianae*. Mol Genet Genomics 270:549–557

Smit AF, Riggs AD (1996) *Tiggers* and DNA transposon fossils in the human genome. Proc Natl Acad Sci USA 93:1443–1448

Soanes DM, Skinner W, Keon J, Hargreaves J, Talbot NJ (2002) Genomics of phytopathogenic fungi and the development of bioinformatic resources. Mol Plant Microbe Interact 15:421–427

Tooley PW, Garfinkel DJ (1996) Presence of Ty1-copia group retrotransposon sequences in the potato late blight pathogen *Phytophthora infestans*. Mol Plant Microbe Interact 9:305–309

Tudor M, Lobocka M, Goodell M, Pettitt J, O'Hare K (1992) The *pogo* transposable element family of *Drosophila melanogaster*. Mol Gen Genet 232:126–134

Whisson SC, van der Lee T, Bryan GJ, Waugh R, Govers F, Birch PRJ (2001) Physical mapping across an avirulence locus of *Phytophthora infestans* using a highly representative, large-insert bacterial artificial chromosome library. Mol Genet Genomics 266:289–295

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse-transcriptase sequences. EMBO J 9:3353–3362