J Mol Evol (2006) 63:458–472
DOI: 10.1007/s00239-005-0211-3

JOURNAL OF MOLECULAR EVOLUTION

Nonneutral GC3 and Retroelement Codon Mimicry in *Phytophthora*

Rays H. Y. Jiang, Francine Govers

Laboratory of Phytopathology, Plant Sciences Group, and Graduate School of Experimental Plant Sciences, Wageningen University, Binnenhaven 5, NL-6709 PD Wageningen, The Netherlands

Received: 5 September 2005 / Accepted: 20 May 2006 [*Reviewing Editor:* Dr. Yves Van de Peer]

Abstract. *Phytophthora* is a genus entirely comprised of destructive plant pathogens. It belongs to the Stramenopila, a unique branch of eukaryotes, phylogenetically distinct from plants, animals, or fungi. *Phytophthora* genes show a strong preference for usage of codons ending with G or C (high GC3). The presence of high GC3 in genes can be utilized to differentiate coding regions from noncoding regions in the genome. We found that both selective pressure and mutation bias drive codon bias in *Phytophthora*. Indicative for selection pressure is the higher GC3 value of highly expressed genes in different *Phytophthora* species. Lineage specific GC increase of noncoding regions is reminiscent of whole-genome mutation bias, whereas the elevated *Phytophthora* GC3 is primarily a result of translation efficiency-driven selection. Heterogeneous retrotransposons exist in *Phytophthora* genomes and many of them vary in their GC content. Interestingly, the most widespread groups of retroelements in *Phytophthora* show high GC3 and a codon bias that is similar to host genes. Apparently, selection pressure has been exerted on the retroelement's codon usage, and such mimicry of host codon bias might be beneficial for the propagation of retrotransposons.

Key words: GC3 — *Phytophthora* — Codon bias — Retrotransposon

Introduction

In nearly all organisms, the 20 amino acids are specified in universal sets of nucleotide triplets called codons. The redundancy of codons enables species to systematically use certain synonymous codons, and in many organisms codon bias is observed. Two major mechanisms are considered responsible for such biases in codon usage: selection pressure and mutation bias (Sharp et al. 1993).

Mutational bias is a global force acting on all sequences. The base composition is constrained by genome-wide mutational processes. In some organisms the whole genome has shifted to extreme GC or AT content by mutational bias. GC content in the green alga *Chlamydomonas reinhardtii* is higher than 60% based on cesium chloride estimate (Scala et al. 2002) (*C. reinhardtii* genome sequencing project at DOE-JGI; http://www.jgi.doe.gov/). The highest AT content was found in the malaria parasite *Plasmodium falciparum*, with about 80% AT (Bowman et al. 1999). A shift of the whole genome to an extreme AT content is also observed in the free-living protist *Dictyostelium discoideum* (Eichinger et al. 2005) and the bacterium *Borrelia burgdorferi* (Fraser et al. 1997) that causes Lyme disease. These genomes with extreme nucleotide composition have evolved independently because the organisms belong to entirely different phylogenetic groups. Codon usage can also be driven by global mutational bias, which creates different trends between species. For example, in a number of species, the overall GC content of the genome shows correlation with species-specific codon bias (Chen et al. 2004).

Correspondence to: Francine Govers; *email:* Francine.Govers@wur.nl

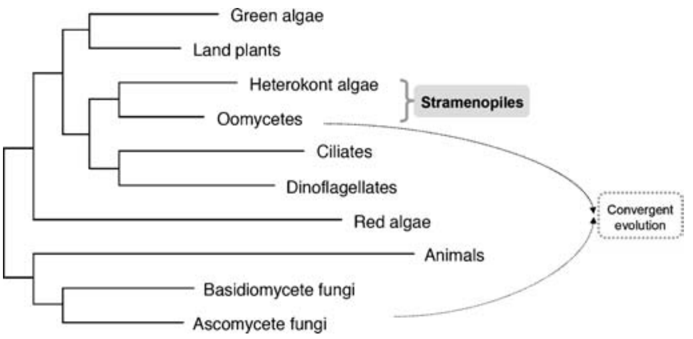


Fig. 1. Phylogenetic tree showing the evolutionary relationships between the major eukaryotic groups. The oomycetes and the (ascomycetous and basidiomycetous) fungi are placed in distinct phylogenetic groups. Their morphological resemblance is presumably due to convergent evolution as indicated by arrows. Reproduced and adapted from Latijnhouwers et al. (2003).

The other major mechanism is selective pressure that works only on coding sequences via selection exerted on translation processes. Selection on synonymous codon positions could lead to a co-adaptation of codon usage and tRNA content. Efficient protein expression can be established via this mechanism. For genes that are expressed at high levels, such selective pressure to optimize translation is expected to be stronger. In such cases, codon bias is expected to correlate with the quantity of tRNA and expression levels. The abundance of tRNAs has been shown to correspond to preference of codons in organisms such as *Escherichia coli* and *Saccharomyces cerevisiae* (Kanaya et al. 1999). A clear correlation between codon usage and gene expression levels was found in *Caenorhabditis eleg*ans, *Drosophila melanogaster*, and *Arabidopsis thaliana* (Duret and Mouchiroud 1999).

The third codon position is rather unique in its GC content. Because of the redundancy of codons, GC3 (GC content at the third codon position) could be largely neutral and representative for whole-genome mutational bias, a phenomenon found in various bacterial species (Sueoka 1995; Sueoka 1999). However, any codon position can be subjected to translation-coupled selection, and GC3 is no exception. So multiple forces can be exerted on the third codon position, and GC3 is therefore determined by a combination of mutational bias, translation-coupled selection, and possibly other selection forces.

The notorious plant pathogen *Phytophthora infestans* is one of the organisms that clearly shows codon bias (Randall et al. 2005; Jiang et al. 2005). *Phytophthora* is a genus comprised of over 65 phytopathogenic species which cause severe damage in agriculture, forestry, and natural habitats. *P. infestans* (causing potato late blight) and *Phytophthora sojae* (causing soybean stem and root rot) are economically important crop pathogens (Erwin and Ribeiro 1996). *Phytophthora ramorum* is a recently discovered species destroying woody shrubs and trees including oaks along the west coast of the United States (Rizzo et al. 2005) and in Europe. *Phytophthora* belongs to the Stramenopiles, heterokonts that evolved distant from plants, animals, and fungi (Baldauf 2003; Baldauf et al. 2000; Margulis and Schwarts 2000) (Fig. 1). *Phytophthora* morphologically resembles fungi and convergent evolution has shaped these two major groups of plant pathogens with similar weaponry to attack plants (Latijnhouwers et al. 2003). Assembled draft genome sequences of *P. sojae* and *P. ramorum* were released in 2004 (http://genome.jgi-psf.org and http://phytophthora.vbi.vt. edu) and genome sequencing of *P. infestans* is in progress (http://www.broad.mit.edu). *P. sojae* and *P. infestans*, the two species for which extensive EST data sets are available (Randall et al. 2005; Gajendran et al. 2006; Qutob et al. 2000), have an estimated GC content of about 50% overall in the genome and about 60% in coding regions (Jiang et al. 2005). In *P. infestans*, high GC3 has been reported to correlate with a high GC content and was found to be related to codon bias (Jiang et al. 2005; Randall et al. 2005). An elevated GC content is not found in the marine diatom *Thalassiosira pseudonana*, the only other heterokont that has been sequenced to date. *T. pseudonana* has coding regions with only 48% GC (Armbrust et al. 2004). The finding that in *Phytophthora* GC3 (> 70%) is high compared to the average genome GC content (ca. 50%) (Jiang et al. 2005) raises the possibility that translation-coupled selection plays a major role in determining GC3 in *Phytophthora*.

Interestingly, some mobile elements in *P. infestans* also show a high GC content (Jiang et al. 2005). Mobile elements constitute the most abundant genetic material in higher eukaryotes. The relationship between transposons and hosts may be a continuum from extreme parasitism to mutualism (Kidwell and Lisch 2001). On the one hand, mobile elements are parasitic. They enrich their abundance by using the cellular apparatus and at the cost of host energy. On the other hand, mobile elements play a central role in the structure, function, and evolution of eukaryotic genomes (Bennetzen 2000; Kazazian 2004) such as providing cis-regulatory sequences, assembling a kinetochore, and speeding up protein diversification (Nekrutenko and Li 2001; Topp et al. 2004). To enlarge genome sizes, retroelements are particularly effective because they transpose via an mRNA intermediate synthesized by a reverse transcriptase. The amplification of the mRNA intermediate may rapidly increase retroelement copy number. The genome sizes of *Phytophthora* species vary. The two sequenced genomes *P. sojae* and *P. ramorum* are 95 and 65 Mb, respectively. The *P. infestans* genome is much larger, 245 Mb, and this seems to be primarily due to insertions of transposable elements (Jiang et al. 2005). *P. infestans* has various retrotransposons and heterogeneous DNA transposons in its genome, some of which are transcribed, indicating they are active (Ah Fong and Judelson 2004; Jiang et al. 2005; Judelson 2002). To understand the relationship between transposible elements and host genomes at a genetic level, investigation of their codon choices can be informative.

Lerat et al. (2002) reported that a lowered GC content is a host-independent characteristic common to all mobile elements. In *C. elega*ns, *A. thaliana*, *D. melanogaster, S. cerevisia*, and *Homo sapiens*, mobile elements exhibit overall AT-richness despite the fact that the host genomes vary in GC content. It was found that mobile elements exhibit codon usage bias similar to weakly expressed host genes in AT-rich genomes like *C. eleg*ans but show no similarity in GC-rich genomes such as *D. melanogaster* (Lerat et al. 2002). However, in *P. infestans*, the sharing of codon bias appears to be related to the copy number of retrotransposons. Two Ty3/Gypsy retrotransposons occur frequently in the genome and share similar codon bias as host genes, whereas other retrotransposons with a lower copy number do not share such codon bias (Jiang et al. 2005). This raises the question whether high GC3 is a general feature for the most widespread *Phytophthora* retrotransposons.

With the availability of the whole-genome sequence of *P. sojae* and *P. ramorum* and large EST data sets of *P. sojae* and *P. infestans*, it is now feasible to analyze GC3 and codon usage in detail on a genome-wide scale, and to correlate expression levels with codon bias. The aim of this study is to search for evidence for whole-genome mutational bias and/or selection pressure in *Phytophthora*. For that purpose we (1) analyzed the relative synonymous codon usage (RSCU) and GC3 in the two sequenced *Phytophthora* genomes, (2) investigated whether the nucleotide composition differs between coding regions and intergenic regions, and (3) analyzed the relationship of codon bias between high-copy retrotransposons and host genes.

Materials and Methods

Genome Databases and EST Databases

The *P. infestans* EST databases are accessible at http://www.pfgd.org and http://staff.vbi.vt.edu/estap and most *Phytophthora* EST sequences are available through GenBank (Kamoun et al. 1999; Qutob et al. 2000; Randall et al. 2005) and http://phytophthora.vbi.vt.edu/EST. *Blumeria graminis* and *P. sojae* EST databases were downloaded from Phytopathogenic Fungi and Oomycete EST Database Version 1.4 (Soanes et al. 2002); http://cogeme.ex.ac.uk. The genomic sequences and annotated protein sequences of *P. sojae* and *P. ramorum* were obtained from the website of the DOE Joint Genome Institute; http://www.jgi.doe.gov/. The annotated genes from *Aspergillus nidulans, Fusarium graminearum, Magnaporthe grisea, Neurospora crassa, Stagonospora nodorum*, and *Ustilago maydis* were downloaded from the Broad institute website; http://www.broad.mit.edu/annotation.

Sequence data from this study have been submitted to GenBank under accession nos. DQ645740 (*GypsyPs-1B*), DQ645741 (*GypsyPs-2*), DQ645742 (*GypsyPs-1A*), DQ645743 (*CopiaPr-1*), DQ645744 (*GypsyPr-2*) and DQ645745 (*GypsyPr-0*).

Bioinformatics Tools

Sequences were analyzed in the Vector NTI 8 package. For BLAST searches we used the NCBI BLAST program and Standalone-BLAST version 2.2.3 (Altschul et al. 1997). Multiple sequence alignment was performed by ClustalX 1.8, and for phylogenetic tree construction Molecular Evolutionary Genetic Analysis 2.1 (MEGA) (Kumar et al. 2001) was used. Phylogeny reconstruction of reverse transcriptase domains of retrotransposons was performed by neighbor-joining analysis. Poisson correction (PC) was chosen as the distance parameter as specified in the program MEGA. Other methods like minimum evolution (ME) and maximum parsimony (MP) gave similar results for clades with bootstrap values higher than 60. Close-Neighbour-Interchage with search level 2 was used for ME and the heuristic search method was used for MP as specified in the program MEGA. The inferred phylogeny was tested by 1000 bootstrap replicates. The calculation scripts were written in Python 2.2 (http://www.python.org) and are available from the authors upon request.

GC Content and Codon Analysis

GC contents of the entire sequence, first, second, and third codon positions (GCaverage, GC1, GC2, and GC3, respectively) were calcuated for each gene. P_3 is the GC content of the third codon position with the exception of ATG, TGG, and three stop codons. These codons are also excluded from the calculations of the GC contents of the first codon position (P_1) and the second codon position (P_2). Therefore, P_3 is slightly different from GC3. For the GC frame plot, GC content was calculated with a 300-bp sliding window for all six reading frames similar to the methods used in the program FramePlot (Ishikawa and Hotta 1999). Relative synonymous codon usage (RSCU) values were calculated as performed by Sharp and Li (1986). For the correlation of codon usage between different gene sets, Trp, Met, and three stop codons were excluded from analysis. The one-tailed significance of Pearson correlation coefficient was calculated for each linear correlation. For noncoding region GC analysis, sequences between −100 and −200 bp, and between −1 and −100 upstream of the start codons, were used. Neighboring coding sequences were excluded from the data set. From each genome, a set of 10,000 sequences was randomly selected to test the differences.

Statistical Analysis of the Differences Between Data Sets

Statistical analysis was performed with the SPSS 12.0.1 package according to the program instructions. In order to make the data set a better approximation of normal distribution, the GC percentage was converted into an arcsin value before performing the *F*-test and *t*-test. The *F*-test was conducted to determine whether the two samples had different variances. The one-tailed probability that the variances were not significantly different was calculated. If two samples showed equal variances, *t*-test was

Table 1. GC content and GC3 in *Phytophthora,* diatom, and several fungal species: Data sets derived from ESTs are shaded; data derived from *Phytophthora* species are in boldface

Data set	Total ORFs	% ORFs with GC3 max[a]	% ORFs with GC >60[b]	GC (%)[c]	GC3 (%)[d]	GC (%) non-coding regions
EST						
Blumeria graminis EST-derived ORFs	283	11.7	17.0	45.4	43.1	-
***P. infestans* EST-derived ORFs**	**1,000**	**81.1**	**98.6**	**57.8**	**70.7**	**-**
***P. sojae* EST-derived ORFs**	**1,000**	**98.6**	**99.9**	**62.3**	**83.5**	**-**
Genomic						
***P. sojae* ORFs**	**10,000**	**93.7**	**99.6**	**60.0**[e]	**76.0**[e]	**52.4**[e]
***P. ramorum* ORFs**	**10,000**	**90.3**	**99.0**	**58.6**[e]	**73.1**[e]	**51.8**[e]
Aspergillus nidulans ORFs	1,000	53.4	69.6	53.5	58.7	47.2
Fusarium graminearum ORFs	1,000	41.9	55.4	51.8	55.6	45.2
Magnaporthe grisea ORFs	1,000	82.2	91.0	57.9	68.5	47.7
Neurospora crassa ORFs	1,000	82.2	92.1	56.2	65.7	45.3
Stagonospora nodorum ORFs	1,000	66.9	81.8	54.8	61.6	46.2
Ustilago maydis ORFs	1,000	62.6	86.2	55.6	61.5	50.7
Thalassiosira pseudonana ORFs	1,000	22.8	26.7	47.5	47.7	-

[a] Percentage of genes having GC3 higher than GC1 or GC2

[b] Percentage of genes having GC1, GC2 or GC3 higher than 60%

[c] The average GC content

[d] The average GC3

[e] These differences between *P. sojae* and *P. ramorum* are significant. Kolmogorov-Smirnov *Z* tests and Mann-Whitney *U* tests showed that two groups of data are significantly different ($p < 0.001$), and Wald-Wolfowitz median tests showed significant differences in the median values ($p < 0.001$)

performed to determine whether the two data sets had the same mean. The probability that two samples came from data sets with the same mean was calculated. If two samples showed unequal variances, Kolmogorov-Smirnov *Z* test and Mann-Whitney *U* test were used to determine whether the variance in each of two independent samples came from the same underlying population. Wald-Wolfowitz median test was conducted to determine whether there was a difference in median values between the two samples.

Results

High GC3 Causes a High GC Content in Phytophthora *Genes*

Two sets of 10,000 ORFs were taken from the whole-genome sequences of *P. sojae* and *P. ramorum*, respectively. High GC content was found in the ORFs as previously reported (Jiang et al. 2005; Randall et al. 2005). The average GC percentage is 60.0% for *P. sojae* and 58.6% for *P. ramorum*. In particular, high GC3 was found: GC3 is higher than GC2 or GC1 in more than 90% of the ORFs in both *Phytophthora* species (Table 1). *P. sojae* and *P. ramorum* genes show a GC3 of 76.0% and 73.1%, respectively. For comparison, GC3 was also calculated for one other straminopile of which the genome has been sequenced and several fungi, five ascomycetes, and one basidiomycete. In the stramenopile *T. pseudonana*, genes do no show high GC3 but in several fungal species an elevated GC3 content was found in genes compared to the noncoding sequences. The level of GC3 increase differs in different species. In *Magnaporthe grisea* and *Neurospora crassa*, genes have a high GC3, with an average value above 65%, and in more than 80% of the genes in both species GC3 is higher than GC1 or GC2. However, in *Blumeria graminis* and *Fusarium graminearum*, less than 50% of the genes have GC3 higher than GC1 or GC2 (Table 1).

To investigate whether or not high GC3 is a specific feature for coding regions, a set of 10,000 randomly selected genomic sequences of 1 kb in size, similar to the typical gene size (1–2 kb), was retrieved from *P. sojae*. These genomic sequences and the 10,000 ORFs were used for GC analysis and the distribution of GC content was plotted. ORFs have a GC content peak around 60%, whereas genome sequences show a peak around 50%. These two distinct peaks indicate that coding regions have a higher GC content than average random genomic regions (Fig. 2A). The GC3 of the majority of ORFs is even higher, with a peak around 70% (Fig. 2A).

The higher percentage GC content in the coding region is caused by high GC3. When GC1, GC2, and GC3 of the 10,000 ORFs were plotted, GC1 and GC2 form two distinct peaks with a lower GC content than the peak of GC3 (Fig. 2B). GC1 has a peak around 40% and GC2 has a peak around 60%. When the same analysis was performed with *P. ramorum* genome sequences and ORFs, similar results were

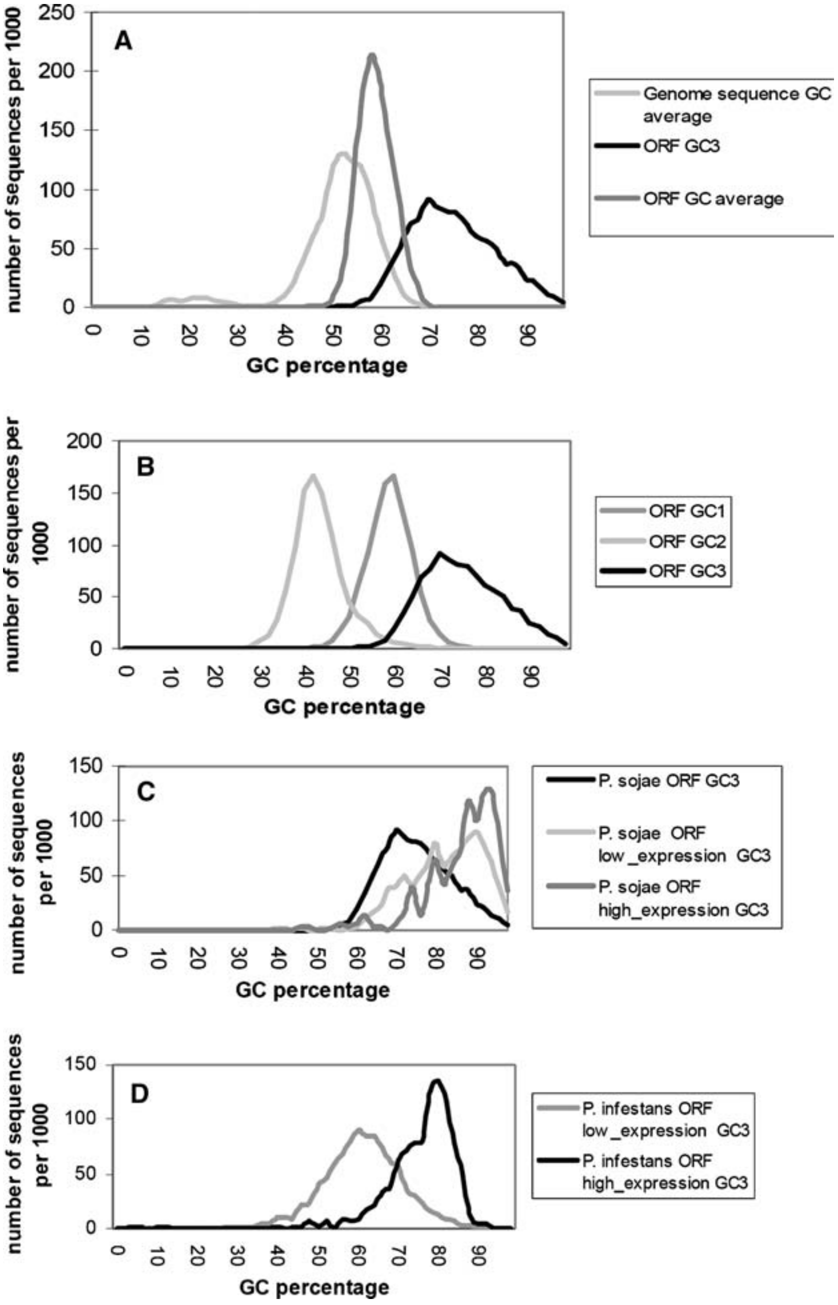


Fig. 2. **A** GC plot of coding regions vs. randomly selected genomic regions in *P. sojae*. Two sets of 10,000 sequences were used for the analysis. **B** GC plot of GC1, GC2, and GC3 of 10,000 *P. sojae* ORFs. **C** GC plot of annotated genes vs. weakly/highly expressed genes in *P. sojae*. Nine hundred thirty-six weakly expressed genes and 305 highly expressed genes were used for the analysis. **D** GC plot of weakly vs. highly expressed genes in *P. infestans*. Three thousand weakly expressed genes and 700 highly expressed genes were used for the analysis. In each graph, the values do not show equal variance in different data sets. Kolmogorov-Smirnov *Z* tests and Mann-Whitney *U* tests showed that any pair of data is significantly different ($p < 0.001$) and Wald-Wolfowitz median tests showed significant differences in the median values for any pair of data ($p < 0.001$).

obtained (data not shown). In *Phytophthora*, the median P_3 value (a value that is slightly different from GC3; for details see Materials and Methods) is much higher than the median GC value of genomic sequences. In *P. sojae*, the median P_3 value is 75.1% whereas the median GC content of the genome is 53.5%. For any pair of data depicted in Fig. 2, Kolmogorov-Smirnov *Z* tests and Mann-Whitney *U* tests showed that two groups are significantly different ($p < 0.001$) and Wald-Wolfowitz median tests showed significant differences in the median values ($p < 0.001$). We can conclude that on a whole-genome scale, *P. sojae* and *P. ramorum* genes on average show a higher GC content than noncoding regions, and this increase in GC content is mainly due to a high GC3 value.

High GC3 and Codon Bias

The high GC3 of *Phytophthora* genes is a result of the codon bias as previously reported from a small set of *P. infestans* genes (Jiang et al. 2005). To investigate the relationship between GC3 and codon bias in the whole genome, RSCU values were calculated for codons from two sets of 10,000 genes derived from *P. sojae* and *P. ramorum*. RSCU values are defined as the number of times that a particular codon is observed, relative to the number of times that the codon would be observed in the absence of any codon usage bias (Sharp and Li 1986). In the absence of any codon usage bias, the RSCU value would be 1. A codon that is used less frequently than expected will have a RSCU value lower than 1, and vice versa, higher than 1 for a codon that is used more frequently than expected. In both genomes, with the exception of TGA (stop codon), all codons with a third position A or T have a RSCU value lower than 1. Except for GGG (Gly), TTG (Leu), and AGG (Arg), all codons ending with C or G have a RSCU value higher than 1 (Table 2).

The percentage of codons ending with A or T and that of codons ending with C or G is plotted to show the cause of high GC3 of genes in *P. sojae* and *P. ramorum* (Fig. 3). Except for the stop codon, all the other redundant amino acid codons have a preference for the degenerate third position ending with C or G. A slightly lower GC% was found in *P. ramorum* codons. The preference of high GC3 codons in *P. sojae* and *P. ramorum* agrees with the results from *P. infestans* (Hraber and Weller 2001; Jiang et al. 2005; Randall et al. 2005). Therefore it can be concluded that the high GC3 of *Phytophthora* genes is due to biased usage of codons ending with C or G.

Whole-Genome Mutation Bias

The two *Phytophthora* species show a different degree of codon bias. On average *P. sojae* genes show higher GC3 than *P. ramorum*. Furthermore, *P. sojae* also shows a more biased RSCU value (Table 2). For almost every codon with a RSCU value >1, *P. so*jae has a higher value than *P. ramorum*, and for codons with a RSCU value <1, *P. sojae* has a lower value than *P. ramorum*. Statistical tests showed that these differences are significant. Moreover, a higher percentage of codons ending with G or G is used in *P. sojae* compared to *P. ramorum* (Fig. 3). These observations suggest a higher GC bias in *P. sojae* than in *P. ramorum*.

Whole-genome mutation bias was found to lead to species-specific codon biases (Chen et al. 2004) and it may have caused the differences between *P. sojae* and *P. ramorum*. To detect the mutation bias, 500-bp noncoding sequences representing intergenic regions were used for GC3 content analysis. Because the characteristic transcription initiation site of oomycetes is typically found within 100 bp upstream of the start codon (McLeod et al. 2004), noncoding regions (5'UTR and partial promoter) were extracted from upstream promoter sequences ranging from -1 to -100 bp in front of 10,000 ORFs. Intergenic regions were extracted from -100 to -200 bp in front of the same 10,000 ORFs. The average GC of *P. sojae* noncoding regions is 54.1%, and that of *P. ramorum* noncoding regions, 53.0%. The GC plot shows two overlapping peaks; the peak of *P. sojae* has a slight shift toward a high GC content compared to that of *P. ramorum* (data not shown). Because only the data set of -1 to -100 bp shows equal variance ($p > 0.001$), a *t*-test was conducted and the result showed that the difference of means is significant (Table 3). Three additional tests which do not presume equal variance of data sets were conducted to compare the -100 to -200 bp intergenic regions. Kolmogorov-Smirnov *Z* test and Mann-Whitney *U* test showed a significant difference between the *P. sojae* and the *P. ramorum* data sets. Wald-Wolfowitz median test showed significant difference in median values (Table 3). We also used a set of 5000 sequences of 1000-bp noncoding regions, and a statistically significant difference between *P. sojae* (53.6%) and *P. ramorum* (53.1%) was obtained. When two sets of 5000 randomly chosen noncoding sequences of *P. ramorum* were compared, no significant difference was detected using various test methods (Table 3). Compared to *P. ramorum*, the increased GC content in the noncoding regions in *P. sojae* is in line with the higher GC3 value of coding regions. *P. sojae* shows an increase in GC content in coding regions, 5'UTR and promoter regions, and intergenic regions of 1.4%, 1.1%, and 0.6%, respectively. The shift of base composition in coding regions (GC3), noncoding regions, and intergenic regions in *P. sojae* suggests a global effect of mutation bias in this species.

High GC3 Is Associated with High Levels of Expression

For *P. sojae*, both the whole-genome sequence and EST sequences are available, and this offers the opportunity to analyze the association between high GC3 and expression level. The GC3 analysis showed that EST-derived ORFs show higher GC3 (83.5%) than ORFs derived from whole-genome sequences (76.0%) (Table 1). This difference suggests that expressed genes have higher GC3 than the average genes annotated from the genome. To further analyze the relationship between GC3 and expression, *P. sojae* ESTs were divided into two groups: one containing highly expressed genes and the other containing weakly expressed genes. A total of 1602 unique EST contigs derived from zoospores, mycelium, and infection tissues from the Phytopathogenic Fungi and Oomycete EST Database (Soanes et al. 2002) were used. Of this public EST data set, which consists of contigs representing roughly 5%–10% of all *P. sojae* genes, 936 contigs with fewer than 3 ESTs were defined as more weakly expressed and contigs with more than 5 ESTs were defined as more highly

Table 2. Relative Synonymous Codon Usage (RSCU) and 3rd position GC frequency in 10,000 *P. sojae* genes and 10,000 *P. ramorum* genes. A total of 10,818,654 codons were counted in *P. sojae* and 10,040,752 in *P. ramorum.*

Species[a]	Amino acid	RSCU[b]					
	Stop	TAA[d]	TAG[d]	TGA[d]			
ps/pr		**0.9/1.0**	**1.0/1.1**	**1.1/1.0**			
	A	GCA[c]	GCC	GCG[c]	GCT[c]		
ps/pr		**0.5/0.7**	1.2/1.2	**1.4/1.3**	**0.8/0.9**		
	C	TGC	TGT				
ps/pr		1.5/1.5	0.5/0.5				
	D	GAC[c]	GAT[c]				
ps/pr		**1.5/1.4**	**0.5/0.6**				
	E	GAA[c]	GAG[c]				
ps/pr		**0.5/0.6**	**1.5/1.4**				
	F	TTC[d]	TTT[c]				
ps/pr		**1.6/1.5**	**0.4/0.5**				
	G	GGA	GGC[d]	GGG[c]	GGT[c]		
ps/pr		0.8/0.8	**1.9/1.8**	**0.7/0.6**	**0.7/0.8**		
	H	CAC[d]	CAT[d]				
ps/pr		**1.6/1.5**	**0.4/0.5**				
	I	ATA	ATC[c]	ATT[c]			
ps/pr		0.1/0.1	**2.1/1.9**	**0.8/1.0**			
	K	AAA[c]	AAG[c]				
ps/pr		**0.3/0.4**	**1.7/1.6**				
	L	CTA[d]	CTC[d]	CTG[d]	CTT[d]	TTA[d]	TTG[c]
ps/pr		**0.3/0.4**	**1.6/1.5**	**2.5/2.4**	**0.6/0.7**	**0.1/0.2**	**0.9/1.0**
	M	ATG					
ps/pr		1.0/1.0					
	N	AAC[d]	AAT[d]				
ps/pr		**1.6/1.5**	**0.4/0.5**				
	P	CCA[d]	CCC[d]	CCG	CCT[d]		
ps/pr		**0.6/0.7**	**1.0/0.9**	1.6/1.6	**0.7/0.8**		
	Q	CAA[d]	CAG[d]				
ps/pr		**0.5/0.6**	**1.5/1.4**				
	R	AGA	AGG[c]	CGA[d]	CGC[d]	CGG[d]	CGT[c]
ps/pr		0.4/0.4	**0.5/0.4**	**0.9/1.0**	**2.4/2.3**	**1.0/0.9**	**0.9/1.1**
	S	AGC[c,d]	AGT[d]	TCA	TCC	TCG[d]	TCT
ps/pr		**1.6/1.5**	**0.6/0.7**	0.5/0.5	1.0/1.0	**1.8/1.7**	0.6/0.6
	T	ACA[c]	ACC[c]	ACG	ACT		
ps/pr		**0.5/0.6**	**1.1/1.0**	1.8/1.8	0.6/0.6		
	V	GTA[d]	GTC[d]	GTG	GTT[d]		
ps/pr		**0.2/0.3**	**1.2/1.0**	2.1/2.1	**0.5/0.6**		
	W	TGG					
ps/pr		1.0/1.0					
	Y	TAC	TAT				
ps/pr		1.6/1.6	0.4/0.4				

[a] ps, *P. sojae*; pr, *P. ramorum*.

[b] RSCU values differing in *P. sojae* and *P. ramorum* are in bold.

[c] The RSCU values of codons have equal variance in *P. sojae* and *P. ramorum* as indicated by F-test, and one-tailed T-test showed that the difference of means is significant ($p < 0.001$).

[d] The RSCU values of these codons do not show equal variance in *P. sojae* and *P. ramorum*; Kolmogorov-Smirnov Z tests and Mann-Whitney U tests showed that two groups of data are significantly different ($p < 0.001$) and Wald-Wolfowitz median tests showed significant differences in the median values ($p < 0.001$).

expressed. On the GC plot, average genes give a GC3 peak around 70%, the more weakly expressed genes show a GC3 peak round 80%–90%, and the more highly expressed genes show a GC3 peak above 90% (Fig. 2C). The distinct peaks show that expressed genes have a higher GC3 than the annotated genes of the genome, and in particular, the high GC3 feature is more pronounced in genes with high expression levels.

A similar analysis was carried out with a large EST data set generated from a wide range of life stages and culture conditions in *P. infestans*. This data set consists of 18,256 unigenes representing at least 50% of all *P. infestans* genes and probably more (Randall et al. 2005). Three thousand unigene contigs consisting of a single EST were defined as weakly expressed genes and 700 unigene contigs with more than 10 ESTs were defined as highly expressed. The

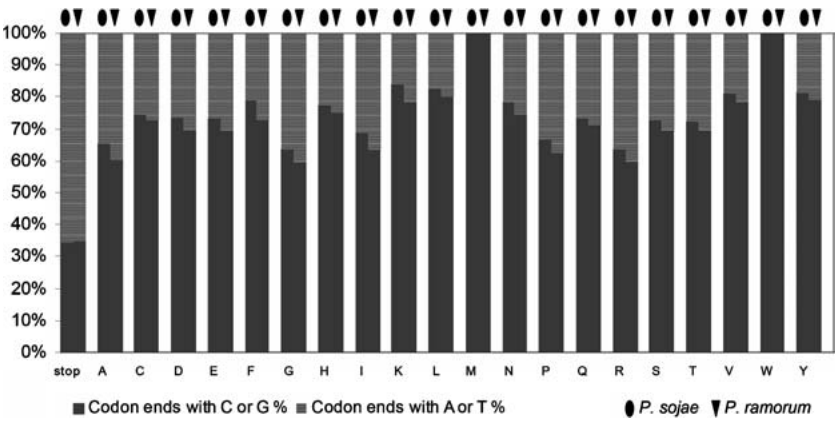


Fig. 3. The percentage of codons ending with C/G and A/T in *P. sojae* and *P. ramorum*. On the *X*-axis, single-letter amino acid codes are used.

GC3 of weakly expressed genes shows a peak around 60%, whereas that of the highly expressed genes gives a peak around 80% (Fig. 2D). Statistical tests showed that the differences are significant. Therefore the highly expressed genes in *P. infestans* also show a higher GC3, which agrees with the results obtained with *P. sojae*.

Genes Can Be Visualized as Peaks on a GC Plot

Because high GC3 is a feature for *Phytophthora* genes and because random genomic sequences on average have a lower GC content, it should be possible to utilize GC3 values to differentiate coding and non-coding regions in the genome. Previously, *inf1* was shown to appear as a peak in a 12-kb region of *P. infestans* on a GC frame plot (Jiang et al. 2005). To investigate this feature on a larger scale, scaffold 52 (JGI genome release version 1) of *P. ramorum* was used as a whole for a GC frame plot to visualize the position of genes and GC peaks. Scaffold 52 is 350 kb in size and contains 80 annotated genes with various predicted functions such as (de)phosphorylation, transport, and DNA repair. On average, genes are 2 kb in size and 2–3 kb apart from each other. The genes are rather evenly distributed except for two 20-kb gene-poor regions located at about 80 and 130 kb from the start. On a genomic GC frame plot, the positions containing the majority of the genes can be shown as GC peaks (Fig. 4A). At a GC threshold of 75%, only two GC peaks are found in the intergenic regions (Fig. 4B). Further analysis of these two peak regions revealed that one 9-kb gene is missing in the initial annotation and also one fragmented retrotransposon was not annotated. At a GC threshold of 65%, only one annotated gene lacks a GC peak, and this is a retrotransposon-like element (Fig. 4A). Chi-square tests were conducted to calculate the probability that peaks occur randomly or in genes. This showed that above a GC threshold of 70%, the chance that the peaks contain genes is highly significant ($p < 0.001$). Therefore, we conclude that on this GC frame plot of a 350-kb genomic *P. ramorum* sequence, most GC peaks above 70% represent genes. Analysis of a 110-kb region containing elicitin genes in *P. infestans* also showed similar results, with peaks representing genes and some high GC retrotransposons (Jiang et al. 2005). *P. infestans* has a larger genome than *P. ramorum* and the peaks derived from retrotransposons appear more frequently.

Retrotransposons Have a More Variable GC Content Than Genes

The genome of *P. infestans* has heterogeneous groups of retrotransposons. Retrotransposons can exhibit either similar or low GC content compared to average *Phytophthora* genes. On the GC frame plot of scaffold 52 in *P. ramorum*, several high GC peaks (above 75%) are derived from retrotransposon-like elements and also the only annotated gene without a GC peak above 65% is a retrotransposon. In the whole genome of *P. sojae*, annotated genes with an exceptionally low GC content are often retrotransposons. Of the 10,000 *P. sojae* ORFs, 16 with a GC content less than 50% were selected for further analysis. Except for a few fragmented pseudo-genes, most of them are retrotransposon-like elements (R.H.Y, Jiang, unpublished results).

The previously characterized retrotransposons *GypsyPi-1* (AY830091) and *GypsyPi-3* (AY830104) show high GC3 and a similar codon usage as *P. infestans* genes (Jiang et al. 2005). The RSCU values of all 64 codons were used for correlation with the exception of ATG, TGG, and three stop codons. Each dot in the graph represents one codon. The level of correlation that is represented by the coefficient of determination (R^2) ranges in value from 0 to 1. If R^2 is 1, there is a perfect correlation; if R^2 is 0, there is no correlation. The correlation of codon usage between *GypsyPi-1*/*GypsyPi-3* and *P. infestans* genes gave an R^2 value of 0.78, which shows a high level of correlation ($p < 0.001$) (Fig. 5A). Such a correlation is not found between *P. sojae* genes and

Table 3. GC content in the non-coding regions in *P. sojae* and P. *ramorum*

Non-coding region from	GC% average	Standard deviation of GC%	GC median	*p* value of F-test[c]	*p* value of T-test[d]	Kolmogorov-Smirnov Z test[e]	Mann-Whitney U test[f]	Wald-Wolfowitz Median test[g]
***P. ramorum* (−100–200 bp)[a]**	51.8	7.6	52	< 0.001	–	< 0.001	< 0.001	< 0.001
***P. sojae* (−100–200 bp)[a]**	52.4	8.4	53					
***P. ramorum* (−1–100 bp)[a]**	53.0	7.3	54	0.003	< 0.001	< 0.001	< 0.001	< 0.001
***P. sojae* (−1–100bp)[a]**	54.1	7.6	55					
***P. ramorum* (−100–200 bp)[b]**	51.9	7.6	52	0.38	0.79	0.61	0.6	
***P. ramorum* (−100–200 bp)[b]**	51.8	7.6	52					
***P. ramorum* (−1–100 bp)[b]**	52.9	7.3	54	0.44	0.85	0.42	0.51	0.95
***P. ramorum* (−1–100 bp)[b]**	53.0	7.3	54					

[a] Regions between −100 to −200 bp and between −1 to −100 upstream of the start codons were used for the analysis. From each genome, a set of 10,000 sequences was randomly selected.

[b] Regions between −100 to −200 bp and between −1 to −100 upstream of the start codons were used for the analysis. 5,000 randomly chosen *P. ramorum* sequences were compared to the other 5,000 randomly chosen *P. ramorum* sequences.

[c] To make the sample a better approximation of Normal distribution, GC percentages were converted into arcsin values before the F-tests. F-test was conducted to determine whether the two samples have different variances, *p* value is the one-tailed probability that the variances are NOT significantly different ($p > 0.001$).

[d] GC percentages were converted into arcsin values before the T-tests. T-test was performed to determine whether the two samples have the same mean, *p* value is the probability that two samples come from data sets with the same mean. Two-tailed T-tests were used for comparing the -1-100 bp regions between *P. sojae* and *P. ramorum,* and for comparing two randomly chosen *P. ramorum* data sets.

[e] [f] Kolmogorov-Smirnov *Z* test and Mann-Whitney U test were used to determine whether the variable in each of two independent samples comes from the same underlying population. Equal variance was not presumed in these tests, *p* value is the probability that two samples come from the same underlying population.

[g] Wald-Wolfowitz Median test was conducted to determine whether there is a difference in median values between the two samples. Equal variance was not presumed in the test, *p* value is the probability that two samples have the same median values.

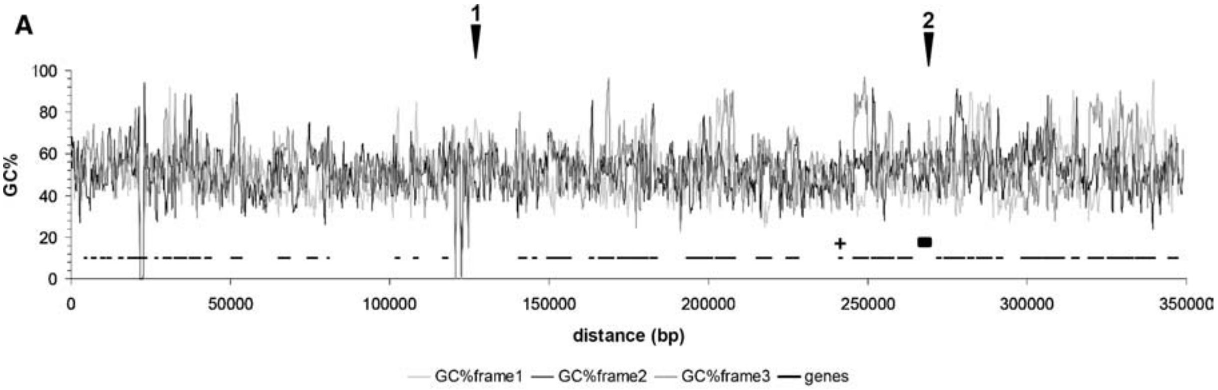
T. pseudonana genes ($p = 0.08$) (Fig. 5C). However, not all retrotransposons show a high correlation between RSCU values and host genes. Another *P. infestans* retrotransposon, *CopiaPi-2* (AY830099), does not have high GC3 or similar codon usage as *P. infestans* genes. The regression between *CopiaPi-2* and *P. infestans* genes gives a negative slope value and an R^2 value of 0.15, and the correlation is statistically not significant ($p = 0.01$) (Fig. 5B). Among the retrotransposon fragments with an unusually low GC content in *P. sojae*, five sequences with a predicted ORF longer than 200 bp were used for RSCU correlation analysis with *P. sojae* ORFs. The regression gives an R^2 value of 0.14 and the correlation is not significant ($p = 0.01$) (Fig. 5D).

The Most Abundant Retrotransposons Have a Similar Codon Bias to Phytophthora *Genes*

GypsyPi-1 and *GypsyPi-3* of *P. infestans*, which both have a high GC3, may have undergone recent transposition events. This assumption is based on the high level of sequence similarity between the LTR pairs, but also on the fact that among seven described retrotransposons, *GypsyPi-1* and *GypsyPi-3* have the highest copy number in 500-kb sequences derived from several regions in the genome. In contrast, the retrotransposon *CopiaPi-2*, with a low GC3, has fewer copies than *GypsyPi-1* and *GypsyPi-3* in the regions analyzed (Jiang et al. 2005).

To investigate whether the higher copy number of retrotransposons is associated with a high GC3, several of the most widely spread retrotransposons in the genomes of *P. sojae* and *P. ramorum* were retrieved for analysis. A set of 100 sequences encoding reverse transcriptase domains was extracted from the whole-genome sequence *of P. sojae* and sequences showing BLAST identity higher than 95% were considered to be derived from the same type of retrotransposon. A total of 23 unique sequences (BLAST identity among each other, < 95%) were found. The copy number was determined by the number of genomic positions with BLAST hit identity of more than 95%. Three sequences with the highest copy number (> 90) were selected, and we assume that these three reverse transcriptases represent three high-copy retrotransposons in *P. sojae*. Using a similar approach, three high-copy retrotransposons were identified in the *P. ramorum* genome.

These six high-copy retrotransposons were subsequently characterized based on BLAST homology. They are all Gypsy-like elements with the exception of one Copia-like retrotransposon, *CopiaPr-1* (ca. 70 copies). The Gypsy-like elements of *P. sojae* were named *GypsyPs-1A* (ca. 100 copies), *GypsyPs-1B* (ca. 100 copies), and *GypsyPs-2* (ca. 300 copies), and those of *P. ramorum, GypsyPr-2* (ca. 300 copies) and *GypsyPr-0* (ca. 100 copies). A phylogenetic tree based on the reverse transcriptase domain was constructed to visualize their relationship (Fig. 6A). *CopiaPr-1*



B

GC threshold	Genes having peaks above GC threshold	Peaks above GC threshold outside genes	Genes without peaks above GC threshold	Total peaks
65%	79	59	1	139
70%	68	18	12	98*
75%	50	2	31	83*
80%	36	1	44	81*
85%	25	0	55	80*

Fig. 4. **A** GC frame plot of scaffold 52 of *P. ramorum*. The GC content was calculated from a scanning window of 300 bp. The positions of 80 annotated genes are indicated underneath the GC graphs. Two sequence gaps gave 0% GC in the plot. Peak 1 and peak 2 are above 75% but no genes were annotated. Peak 1 corresponds to a retroelement. Peak 2 corresponds to an ORF of 9156 bp that was missing in the initial annotation (indicated by the black rectangle). The retroelement indicated by + shows no GC peak above 65%. **B** Relationship between GC peaks and 80 genes of scaffold 52 in *P. ramorum*. Chi-square tests were conducted to determine whether the peaks significantly appear more in genes. The probability that peaks occur randomly (50% in genes, 50% outside of genes, because the total length of genes and intergenic regions are about equal) was calculated. *These peaks appear significantly more frequently in genes ($p < 0.001$).

does not fall in the highly supported Ty1-Copia group. However, it shows significant BLAST homology with Copia-like sequences, and its reverse transcriptase domain is downstream of the integrase domain. These features strongly suggest that it is a Copia-like element. Several of the high-copy retrotransposons derived from different *Phytophthora* species form one clade. *GypsyPi-1* and *GypsyPi-3* from *P. infestans* and *GypsyPs-1A* and *GypsyPs-1B* from *P. sojae* all belong to clade-1. Within clade-1, they typically share 60% BLAST similarity with each other in the reverse transcriptase domain. The other two Gypsy elements, *GypsyPs-2* and *GypsyPr-2*, from *P. sojae* and *P. ramorum*, respectively, fall in clade-2. Clade-1 and clade-2 therefore represent the most widespread retrotransposons in the *Phytophthora* genomes. The copy number of the retroelements was also estimated by BLASTN in the *P. sojae* and *P. ramorum* genomes. Homologous elements were detected with a hit with a BLAST *E* value < 1e-30 and identity > 80%. Each of the retrotransposons has a homologue in the other genome and different expansion patterns were found. *GypsyPs-2, Gypsy-Pr2*, and *CopiaPr-1* show similar expansions in both genomes, whereas *GypsyPs-1A* and *GypsyPs-1B* are less expanded in *P. sojae* than in *P. ramorum*. The homologue of *GypsyPr-0* is a low-copy element in *P. sojae* (Fig. 6B).

From the *P. sojae* genome, the set of 23 sequences coding for reverse transcriptases (3000 bp) was used for GC3 analysis and 21 of them show a higher GC3 value (> 60%) than the average GC content of the genome. Although the increase in GC3 value does not correlate with higher copy numbers (Fig. 7), the three high-copy retrotransposons do have a high GC3 value. In *P. sojae*, *GypsyPs-1A, GypsyPs-1B*, and *GypsyPs-2* have a GC3 value of 71%, 67%, and 69%, respectively. In *P. ramorum*, *GypsyPr-2, GypsyPr-0*, and *CopiaPr-1* have a GC3 value of 65%, 65%, and 80%, respectively.

P. sojae sequences (3000 bp) encoding the reverse transcriptases of the three high-copy retrotransposons (*GypsyPs-1A, GypsyPs-1B*, and *GypsyPs-2*) were used for codon analysis. A high level of correlation of codon usage was found between the high-copy retrotransposons and *P. sojae* genes, with an R^2 value of 0.84 (Fig. 5F). A similar high level of correlation was found between the high-copy retrotransposons in *P. ramorum* (*GypsyPr-2, GypsyPr-0*, and *CopiaPr-1*) and *P. ramorum* genes, with an R^2 of 0.88 (Fig. 5E), and this agrees with the high level of correlation found in *P. infestans* (Fig. 5A). Using the entire ORFs of the retrotransposons (4–6 kb) gave a very similar GC content and correlation in RSCU value to host genes. We can conclude that in the three *Phytophthora* genomes, high-copy retrotransposons share similar codon biases as host genes.

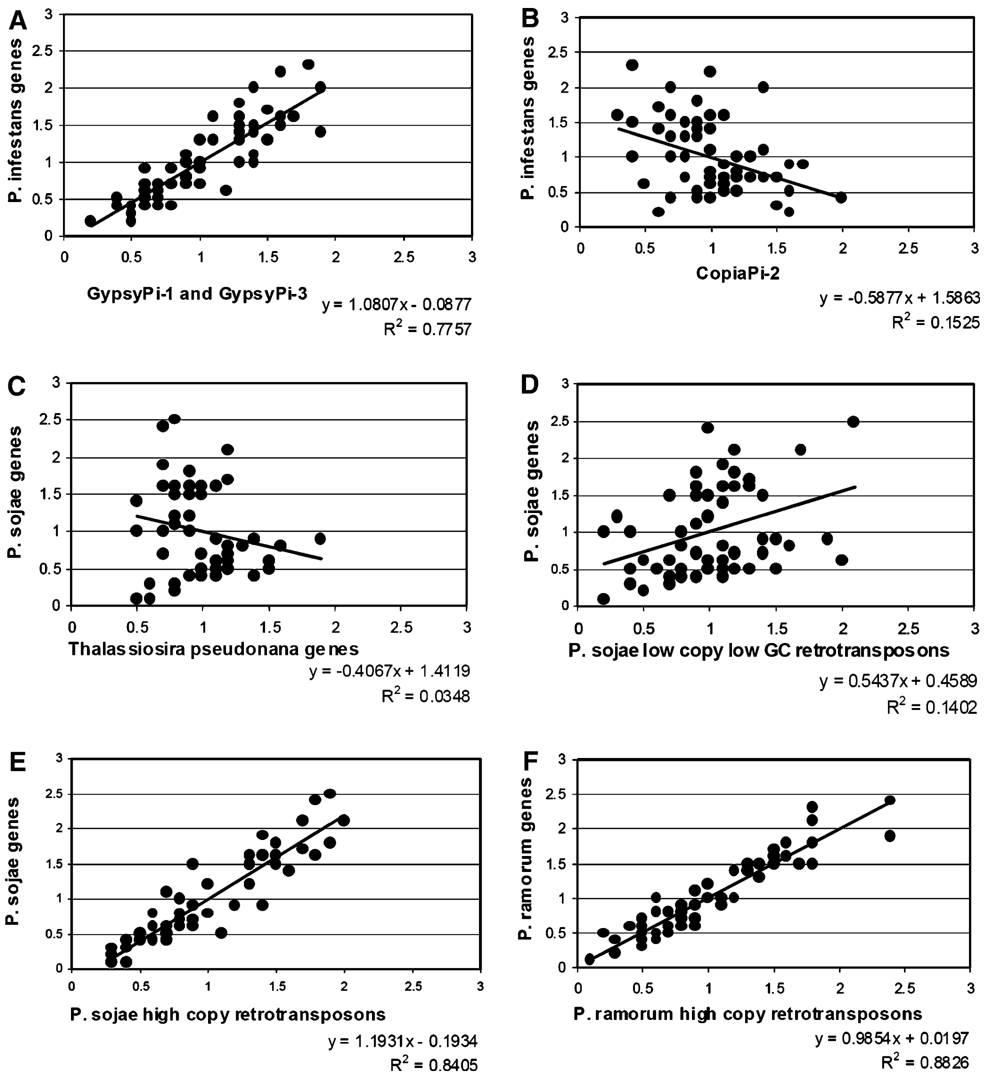


Fig. 5. Correlation of codon usage between different sets of genes. The RSCU values of all codons were used for correlation with the exception of ATG, TGG, and three stop codons. Each dot in the graph represents one codon. A set of 1000 ORFs from *P. infestans* ESTs was correlated with (**A**) *GypsyPi-1* and *GypsyPi-3* elements and (**B**) a *CopiaPi-2* element. *P. sojae* genes were correlated with (**C**) *T. pseudonana* genes and (**D**) *P. sojae* low-copy, low-GC retrotransposon fragments. **E** Correlation between *P. sojae* genes and *P. sojae* high-copy retrotransposons. **F** Correlation between *P. ramorum* genes and *P. ramorum* high-copy retrotransposons. In the regression equation, x represents RSCU values of codons in one group and y represents RSCU values of codons in the other group. The regression coefficient (the slope), y-intercept at $x = 0$, and squared correlation coefficient (R^2) are also shown. The correlation of A, E and F is significant ($p < 0.001$), whereas the correlation of B ($p = 0.01$), C ($p = 0.08$), and D ($p = 0.01$) is not significant.

Discussion

Mutational bias is a global force to change base composition, whereas selection pressure is a local force acting on coding sequences. In this study we show that both forces participate in shaping codon bias in *Phytophthora*. The majority of the *Phytophthora* species are monophyletic and have probably evolved rather recently (Cooke et al. 2000). The close phylogenetic relationship may explain the similar elevated GC content in coding regions in several *Phytophthora* species. However, lineage-specific increases in GC content have also occurred. *P. sojae* shows a significantly higher GC content in the noncoding region compared to *P. ramorum*, and so we infer that whole-genome mutation bias has shifted the *P. sojae* base composition toward a higher GC content. At the same time, selection pressure can be clearly detected by the correlation of high expression levels and GC content in *P. sojae* and *P. infestans*. Therefore, both mutation bias and selection pressure are at work in the *Phytophthora* genomes, and they drive the codon bias with different emphasis: mutation bias gives rise to differences between species and selection pressure tunes the codon usage to expression levels.

In all organisms the third codon position GC content is under mutational bias, translation-coupled selection, and possibly other selection forces, but the relative contribution of each mechanism differs in

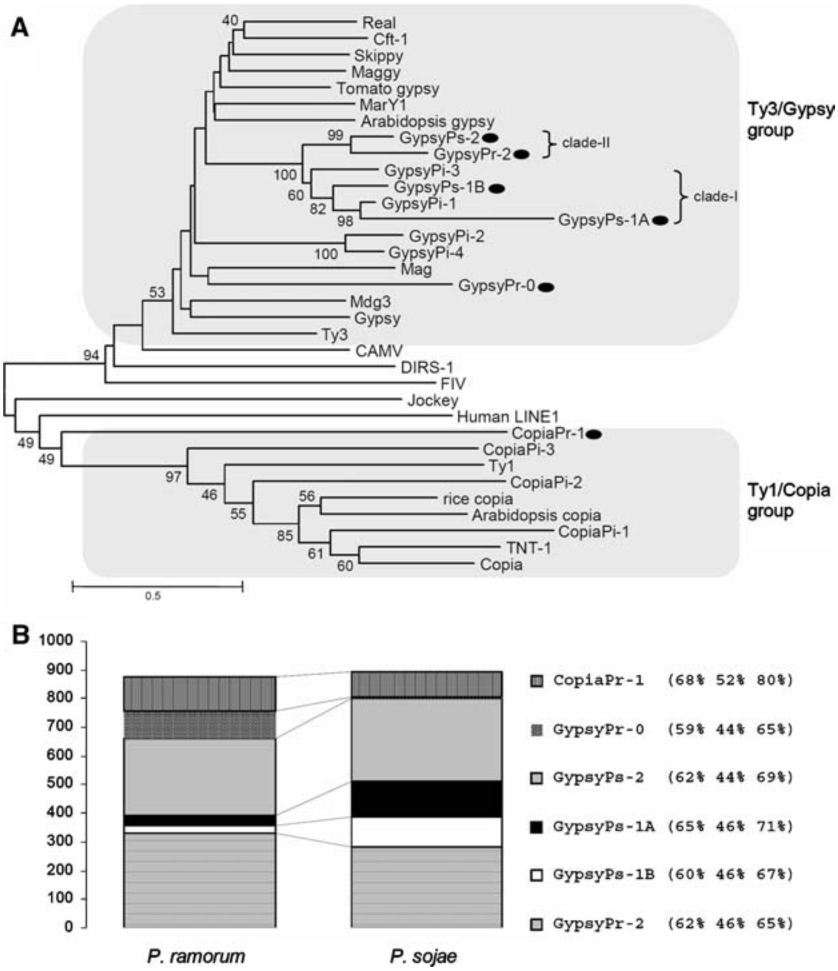


Fig. 6. **A** Phylogenetic tree of high-copy retrotransposons of *Phytophthora* and other organisms. The reverse transcriptase domains were used to construct the unrooted phylogram based on neighbor-joining analysis. Confidence of groupings was estimated by using 1000 bootstrap replicates; numbers next to the branching point indicate the percentage of replicates supporting each branch. On the right different groups of retrotransposons are indicated. *Arabidopsis copia*, BAB84015.1 polyprotein (*Arabidopsis thaliana*); *Arabidopsis gypsy*, AF128395 retrotransposon (*Arabidopsis thaliana*); CAMV, M90543 reverse transcriptase (cauliflower mosaic virus); *Cft-1*, AAF21678 pol polyprotein (*Cladosporium fulvum*); *Copia*, P04146 copia protein (*Drosophila melanogaster*); *CopiaPi-1*, AY830098 copia-like retrotransposon (*Phytophthora infestans*); *CopiaPi-2*, AY830099 copia-like retrotransposon (*Phytophthora infestans*); *CopiaPi-3*, AY830100 copia-like retrotransposon (*Phytophthora infestans*); *DIRS-1*, C24785 DIRS-1element (*Dictyostelium discoideum*); *FIV*, S23820 pol polyprotein (feline immunodeficiency virus); *Gypsy*, AAB50148, polyprotein (*Drosophila melanogaster*); *GpsyPi-1*, AY830091 gypsy-like retrotransposon (*Phytophthora infestans*); *GpsyPi-2*, AY830106 gypsy-like retrotransposon (*Phytophthora infestans*); *GpsyPi-3*, AY830104 gypsy-like retrotransposon (*Phytophthora infestans*); *GpsyPi-4*, AY830107 gypsy-like retrotransposon (*Phytophthora infestans*); *human LINE1*, P08547 human line-1 homologue (*Homo sapiens*); *Jockey*, P21328 mobile element jockey (*Drosophila melanogaster*); *Mag*, S08405 silkworm transposon mag (*Bombyx mori*); *Maggy*, AAA33420 polyprotein (*Magnaporthe grisea*); *MarY1*, BAA78625 polyprotein (*Tricholoma matsutake*); *Mdg3*, T13798 retrotransposon mdg3 (*Drosophila melanogaster*); *Real*, BAA89272 polyprotein Pol (*Alternaria alternata*); *Rice copia*, AAR88589.1 putative copia-like retrotransposon protein (*Oryza sativa*); *Skippy*, S60179 retrotransposon skippy (*Fusarium 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 (*Saccharomyces cerevisiae*). *GypsyPs-1B, GypsyPs-1A, GypsyPs-2, GypsyPr-2, GypsyPr-0*, and *CopiaPr-1* are the six high-copy retrotransposons identified in this study and indicated by black dots. **B** The presence of the six high-copy retrotransposons in the genome of *P. sojae* and *P. ramorum*. The *Y*-axis represents the estimated copy number. Copy number is obtained from BLAST hit with identity percentage >80% and *E* value <1e-30. In parentheses the GC content at each codon position (GC1, GC2, and GC3 respectively) is listed.

different phylogenetic groups. In various bacterial species, the large variation in GC3 was found to be mainly due to directional mutation pressure, whereas translation-coupled selection plays an insignificant role (Sueoka 1995; Sueoka 1999). In vertebrate genomes, the DNA composition of a chromosome varies in segments termed isochores and the GC content of embedded genes correlates with that of the isochores (Bernardi et al. 1985). The selective advantage of increased thermodynamic stability of proteins and DNA in the higher body temperature appears to be responsible for the GC3 variation in vertebrate genes, whereas mutational bias plays a minor role (Alvarez-Valin et al. 2002). In eukaryotes with lower cellular complexity, such as *M. grisea* (Dean et al. 2005), *N. crassa* (Broad Institute, *N.*

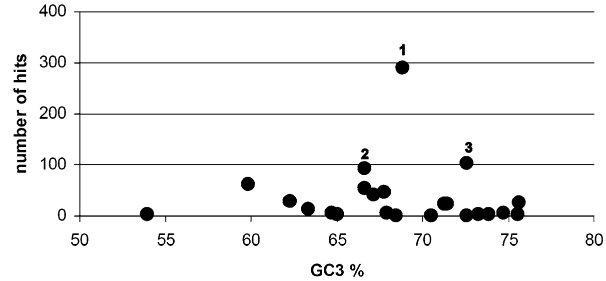


Fig. 7. The relationship between GC3 and copy number of a set of retrotransposons in *P. sojae*. The three high-copy retrotransposons are marked with numbers: 1 (*GypsyPs-2*), 2 (*GpysyPs-1A*), and 3 (*GpysyPs-1B*).

crassa genome sequencing project), and *Phytophthora*, GC content is elevated in coding regions. This suggests that in these organisms translation-coupled selection plays a major role in determining GC3.

Based on our analysis we conclude that *Phytophthora* has a nonneutral GC3. In *P. sojae* P_3 represents only about 10% of the genome (one-third of the genome codes for proteins). If P_3 is neutral, then the rest of the genome (90%) will be selected, which is a highly unlikely scenario. Moreover, extremely low-GC content ORFs are mostly fragmented pseudogenes or retrotransposons (R.H.Y. Jiang et al., unpublished data), supporting the suggesting that the elevated P_3 is a result of selection. Noncoding regions and intergenic regions show a much lower GC content than GC3. A neutral GC3 would suggest selection constraints imposed on the noncoding regions. Note that the possible function of this selection constraint on noncoding regions is different from the translation-coupled selection pressure discussed above. The intergenic regions have a similar GC content as the average genome, so they represent more of a "neutral" state than P_3. Therefore we conclude that P_3 is primarily the result of translation-coupled selection, and that directional mutation bias plays only a minor role in increasing P_3 values within a genome. Because of the nonneutrality of GC3 in *Phytophthora*, the lower GC content in the noncoding regions is mainly due to lack of translation-coupled selection pressure.

Translation-coupled selection has unequal effects in changing GC3 in different eukaryotes. In eukaryotes with high cellular complexity like mammals, the role of translation-coupled selection in causing a heterogeneous GC content within a genome seems to be very limited (Sueoka and Kawanishi 2000), whereas in eukaryotes with lower cellular complexities it seems to play a major role. This difference could be due to the differences in effective population sizes as mentioned by Duret et al. (1999). Because random drift overcomes selection in small populations, a mutation that is advantageous in a species with large effective population sizes may be neutral in a small population. In general, single-celled organisms have much larger effective population sizes than long-lived animals and reductions in population size are usually associated with increased cellular complexity (Lynch and Conery 2003). In microbes like fungi or oomycetes, propagation through spores can produce massive numbers of progeny from a single individual within one generation. It is therefore likely that fitness differences among codons can be strongly selected and this may explain why in these eukaryotic microbes translation-coupled selection has a significant impact on the GC3. In *Phytophthora*, the influence of translation-coupled selection on the GC3 is even strong enough to be detected in retrotransposons. However, only intact, high-copy number retrotransposons that seem to be active mobile elements have a high GC3. Most intact low-GC retrotransposons have low copy numbers and are probably the result of recent horizontal gene transfer. The mutated and fragmented low-GC retrotransposons have lost activity and this might have resulted from lack of translation-coupled selection pressure.

Invasion and replication of retrotransposons can change genome sizes within a short time frame. The difference in genome size between wheat and rice is mainly due to amplification of retrotransposons in the gene-poor regions (Sandhu and Gill 2002). In the last 3 million years, the maize genome has increased from 1200 to 2400 Mb due to retrotransposon activity (SanMiguel et al. 1998). In *P. infestans*, heterogeneous retrotransposons were suggested to be largely responsible for the large genome size of 245 Mb (Jiang et al. 2005). Since several of the most abundant retroelements characterized in this study are present in different *Phytophthora* species, they probably invaded the *Phytophthora* genome before speciation. On top of that, lineage specific expansion has occurred. *GpysyPs-1A* and *GypsyPs-1B* appear to have expanded in *P. sojae* but to a lesser extent in *P. ramorum* (Fig. 6B). *GyspyPr-0* is expanded in *P. ramorum* but its homologue has only a few copies in *P. sojae*. These expansion patterns may contribute to the different sizes of the *Phytophthora* genomes. A recent burst of *GypsyPi*-1and *GypsyPi-3* activity may have largely contributed to the increase in size of the *P. infestans* genome.

The replication of these genetic parasites involves transcription of the retroelement and synthesis of the reverse transciptase. If a retrotransposon is only under whole-genome mutation bias, GC3 should have a similar value as the mean GC content of the genome. Our results show that high-copy retrotransposons have an increased GC content similar to coding sequences, which indicates that their codon usage has been driven by selection pressure to optimize protein translation. The transposable elements of *C. elegans,* *S. cerevisiae, A. thaliana, D. melanogaster*, and *H. sapiens* appear to be AT-rich regardless of the base composition of their host genomes (Lerat et al. 2002). The AT bias of retrotransposons was suggested to reflect the AT-rich characteristics of the reverse transcriptase of retroviruses. For example, in some lentiviruses the error-prone reverse transcriptase may lead to the preference of G-to-A and C-to-T transition (Zsiros et al. 1999). However, in the case of the three *Phytophthora* genomes, high-copy retrotransposons are likely to be primarily under selection pressure similar to that imposed on host genes.

Retrotransposon activity can be disastrous to hosts because it is able to cause massive deleterious mutations. Tight host control of retrotransposon mobilization is needed in any viable species. Multiple steps can be used to control retrotransposon activity, such as reduction of expression and degradation of transcripts by gene silencing and limiting integration (Labrador and Corces 1997). Facing host surveillance systems, successful retrotransposons must be able to handle these control steps in order to spread in the genome. In *Phytophthora*, mimicry of host codon usage can be beneficial for genetic parasitic elements to optimize their production of reverse transcriptase. Utilizing the host cellular machinery efficiently may be one of the strategies for retrotransposons to propagate successfully in the genome.

Acknowledgments. We thank Martijn Staats for his advice concerning the statistical analyses, Brett Tyler for helpful suggestions and discussions, and the DOE-Joint Genome Institute (http://www.jgi.doe.gov/index.html), in particular, Brett Tyler, Jeff Boore, and Dan Rokhsar, for *P. sojae* and *P. ramorum* genome sequences. This work was financially supported by an Aspasia grant from the Netherlands Organisation for Scientific Research (NWO-Aspasia 015.000.057).

References

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, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Alvarez-Valin F, Lamolle G, Bernardi G (2002) Isochores, GC3 and mutation biases in the human genome. Gene 300:161–168

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou SG, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kroger N, Lau WWY, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306:79–86

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

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972–977

Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42:251–269

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Bowman S, Lawson D, Basham D, Brown D, Chillingworth T, Churcher CM, Craig A, Davies RM, Devlin K, Feltwell T, Gentles S, Gwilliam R, Hamlin N, Harris D, Holroyd S, Hornsby T, Horrocks P, Jagels K, Jassal B, Kyes S, McLean J, Moule S, Mungall K, Murphy L, Oliver K, Quail MA, Rajandream MA, Rutter S, Skelton J, Squares R, Squares S, Sulston JE, Whitehead S, Woodward JR, Newbold C, Barrell BG (1999) The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. Nature 400:532–538

Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH (2004) Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci USA 101:3480–3485

Cooke DE, Drenth A, Duncan JM, Wagels G, Brasier CM (2000) A molecular phylogeny of *Phytophthora* and related oomycetes. Fungal Genet Biol 30:17–32

Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu JR, Pan H, Read ND, Lee YH, Carbone I, Brown D, Oh YY, Donofrio N, Jeong JS, Soanes DM, Djonovic S, Kolomiets E, Rehmeyer C, Li W, Harding M, Kim S, Lebrun MH, Bohnert H, Coughlan S, Butler J, Calvo S, Ma LJ, Nicol R, Purcell S, Nusbaum C, Galagan JE, Birren BW (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature 434:980–986

Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci USA 96:4482–4487

Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q, Tunggal B, Kummerfeld S, Madera M, Konfortov BA, Rivero F, Bankier AT, Lehmann R, Hamlin N, Davies R, Gaudet P, Fey P, Pilcher K, Chen G, Saunders D, Sodergren E, Davis P, Kerhornou A, Nie X, Hall N, Anjard C, Hemphill L, Bason N, Farbrother P, Desany B, Just E, Morio T, Rost R, Churcher C, Cooper J, Haydock S, Van Driessche N, Cronin A, Goodhead I, Muzny D, Mourier T, Pain A, Lu M, Harper D, Lindsay R, Hauser H, James K, Quiles M, Babu MM, Saito T, Buchrieser C, Wardroper A, Felder M, Thangavelu M, Johnson D, Knights A, Loulseged H, Mungall K, Oliver K, Price C, Quail MA, Urushihara H, Hernandez J, Rabbinowitsch E, Steffen D, Sanders M, Ma J, Kohara Y, Sharp S, Simmonds M, Spiegler S, Tivey A, Sugano S, White B, Walker D, Woodward J, Winckler T, Tanaka Y, Shaulsky G, Schleicher M, Weinstock G, Rosenthal A, Cox EC, Chisholm RL, Gibbs R, Loomis WF, Platzer M, Kay RR, Williams J, Dear PH, Noegel AA, Barrell B, Kuspa A (2005) The genome of the social amoeba *Dictyostelium discoideum*. Nature 435:43–57

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

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, Gwinn M, Dougherty B, Tomb JF, Fleischmann RD, Richardson D, Peterson J, Kerlavage AR, Quackenbush J, Salzberg S, Hanson M, vanVugt R, Palmer N, Adams MD, Gocayne J, Weidman J, Utterback T, Watthey L, McDonald L, Artiach P, Bowman C, Garland S, Fujii C, Cotton MD, Horst K, Roberts K, Hatch B, Smith HO, Venter JC (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390:580–586

Gajendran K, Gonzales MD, Farmer A, Archuleta E, Win J, Waugh ME, Kamoun S (2006) *Phytophthora* functional genomics database (PFGD): functional genomics of *Phytophthora*-plant interactions. Nucleic Acids Res 34:D465–D470

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

Ishikawa J, Hotta K (1999) FramePlot: a new implementation of the frame analysis for predicting protein-coding regions in bacterial DNA with a high G + C content. FEMS Microbiol Lett 174:251–253

Jiang RHY, Dawe AL, Weide R, Van Staveren M, Peters S, Nuss DL, Govers F (2005) Elicitin genes in *Phytophthora infestans* are clustered and interspersed with various transposon-like elements. Mol Genet Genomics 273:20–32

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, 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

Kanaya S, Yamada Y, Kudo Y, Ikemura T (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238:143–155

Kazazian HH Jr (2004) Mobile elements: drivers of genome evolution. Science 303:1626–1632

Kidwell MG, Lisch DR (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. Evolut Int J Org Evolut 55:1–24

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17:1244–1245

Labrador M, Corces VG (1997) Transposable element-host interactions: regulation of insertion and excision. Annu Rev Genet 31:381–404

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

Lerat E, Capy P, Biemont C (2002) Codon usage by transposable elements and their host genes in five species. J Mol Evol 54:625–637

Lynch M, Conery JS (2003) The origins of genome complexity. Science 302:1401–1404

Margulis L, Schwarts KV (2000) Five kingdoms: an illustrated guide to the phyla of life on earth. W.H. Freeman, New York

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

Nekrutenko A, Li W-H (2001) Transposable elements are found in a large number of human protein-coding genes. Trends Genet 17:619–621

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

Randall TA, Dwyer RA, Huitema E, Beyer K, Cvitanich C, Kelkar H, Fong AM, Gates K, Roberts S, Yatzkan E, Gaffney T, Law M, Testa A, Torto-Alalibo T, Zhang M, Zheng L, Mueller E, Windass J, Binder A, Birch PR, Gisi U, Govers F, Gow NA, Mauch F, van West P, Waugh ME, Yu J, Boller T, Kamoun S, Lam ST, Judelson HS (2005) Large-scale gene discovery in the oomycete *Phytophthora infestans* reveals likely components of phytopathogenicity shared with true fungi. Mol Plant Microbe Interact 18:229–243

Rizzo DM, Garbelotto M, Hansen EM (2005) *Phytophthora ramorum:* integrative research and management of an emerging pathogen in California and Oregon forests. Annu Rev Phytopathol 43:309–335

Sandhu D, Gill KS (2002) Gene-containing regions of wheat and the other grass genomes. Plant Physiol 128:803–811

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nature Genet 20:43–45

Scala S, Carels N, Falciatore A, Chiusano ML, Bowler C (2002) Genome properties of the diatom *Phaeodactylum tricornutum*. Plant Physiol 129:993–1002

Sharp PM, Li WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24:28–38

Sharp PM, Stenico M, Peden JF, Lloyd AT (1993) Codon usage: Mutational bias, translational selection, or both? Biochem Soc Trans 21:835–841

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

Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325

Sueoka N (1999) Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C. J Mol Evol 49:49–62

Sueoka N, Kawanishi Y (2000) DNA G + C content of the third codon position and codon usage biases of human genes. Gene 261:53–62

Topp CN, Zhong CX, Dawe RK (2004) Centromere-encoded RNAs are integral components of the maize kinetochore. Proc Natl Acad Sci USA 101:15986–15991

Zsiros J, Jebbink MF, Lukashov VV, Voute PA, Berkhout B (1999) Biased nucleotide composition of the genome of HERV-K related endogenous retroviruses and its evolutionary implications. J Mol Evol 48:102–111