Trichoderma spp. genome and gene structure

G.H.GOLDMAN, C.H.PELLIZZON, M.MARINS*, J.O.MCINERNEY† AND M.H.S.GOLDMAN‡

*Faculdade de Ciências Farmacêuticas, * Faculdade de Medicina and ‡ Faculdade de Odontologia de Ribeirão Preto, Universidade de São Paulo, Ribeirão Preto, SP, Brazil*
† Natural History Museum, Cromwell Road, London, UK

9.1 Introduction

Fungi in the genus *Trichoderma* have been used for the production of lytic enzymes and to control a wide range of plant pathogenic fungi. In the last years, much progress has been made in elucidating the molecular biology of *Trichoderma* spp. The objective of this chapter is to provide a summary of data concerning size and organization of the genome and gene structure in *Trichoderma* spp., mainly translation control sequences and codon usage. This summary will aid in the investigation of the molecular genetics of this genus.

9.2 Size and organization of the genome

Filamentous fungi usually contain haploid nuclei and a relatively small genome, frequently about 25 to 50 Mb (for a review, see Skinner *et al.,* 1991). The development of pulse-field gel electrophoresis (PFGE) has allowed electrophoretic karyotyping of several yeasts and filamentous fungi. The use of PFGE and molecular karyotyping technology has led to the assignment of cloned genes to chromosomal locations. New understandings can arise through the utilization of this technology; for example, molecular karyotyping can aid in the detection of translocations and variations in chromosome number and can be used to generate chromosome-specific sublibraries. Chromosomal DNA from *Trichoderma* spp. has been separated by using different PFGE techniques, e.g. contour-clamped homogeneous electric field, rotary electrode, and transverse-alternating field electrophoreses (Gilly and Sands, 1991; Hayes *et al.,* 1993; Herrera-Estrella *et al.,* 1993; Mäntylä *et al.,* 1992). The estimated genome sizes and chromosome numbers of *Trichoderma* spp. range from 31 to 39 Mb and from 3 to 7, respectively (see Volume 1, Chapter 11). Chromosomes differed substantially in size. The sizes of the individual chromosomes indicate significant variation between the cellulolytic *T. reesei* and those

Table 9.1 *Trichoderma* spp. gene sequence database

Gene[a]	Product	Organism	Reference	Accession[b]
hfb1	hydrophobin I	*T. reesei*	Nakari-Setela *et al.* (1996)	Z68124
cre1	CREI protein	*T. harzianum*	Ilmen *et al.* (1996)	X95369
pkc1	protein kinase C	*T. reesei*	Morawetz *et al.* (1996)	U10016
tef1	translocation elongation factor 1α	*T. reesei*	Nakari *et al.* (1993)	Z23012
cbh II	cellobiohydrolase II	*T. reesei*	Chen *et al.* (1987)	M55080
prb1	alkaline proteinase	*T. harzianum*	Geremia *et al.* (1993)	M87518 and M87519
pgk1	3-phosphoglycerate kinase	*T. reesei*	Vanhanen *et al.* (1991)	M81623 and M61878
bg11	β-D-glucoside glucohydrolase	*T. reesei*	Mach (1994)	U09580
ech-42	chitinase	*T. harzianum*	Hayes *et al.* (1994); Carsolio *et al.* (1994)	X79381
cbh1	cellulose 1,4-β-cellobiosidase	*T. koningii*	Wey *et al.* (1994)	X69976
tub1	beta-tubulin	*T. viride*	Goldman *et al.* (1993)	Z15054
tub2	beta-tubulin	*T. viride*	Goldman *et al.* (1993)	Z15055
egIII	endoglucanase III	*T. reesei*	Saloheimo *et al.* (1988)	M19373
eg15	endo-1-4-β-glucanase V	*T. reesei*	Saloheimo *et al.* (1994)	Z33381

Table 9.1 *(Cont)*

Gene[a]	Product	Organism	Reference	Accession[b]
egl1	cellulase	*T. longibrachiatum*	Perez-Gonzalez (unpublished)	X60652
pk11	pyruvate kinase	*T. reesei*	Schindler *et al.* (1993)	L07060
chit33	chitinase	*T. harzianum*	Limon *et al.* (1995)	X80006
pyr4	orotidine-5′-phosphate decarboxylase	*T. harzianum*	Heidenreich and Kubicek (1994)	U05192
axe1	acetyl xylan esterase	*T. reesei*	Margolles-Clark *et al.* (1996b)	Z69256
18s rRNA	18S rRNA	*T. harzianum*	Schlick *et al.* (1994)	Z48812
5.8s rRna	5.8S rRNA	*T. harzianum*	Schlick *et al.* (1994)	
b16-2	glucan endo-1,6-β-glucosidase	*T. harzianum*	Lora *et al.* (1995)	X79196
ura5	orotidine-5′-phosphate decarboxylase	*T. reesei*	Berges *et al.* (1990)	X55879
pkt1	serine/threonine protein kinase	*T. reesei*	Morawetz *et al.* (1994)	U05811
pgk-49	phosphoglycerate kinase	*T. viride*	Goldman *et al.* (1990)	X54284
chi42, pc1ch1	endochitinase	*T. harzianum*	Draborg *et al.* (1996)	U49455
5s rRNA	5S ribosomal RNA	*T. harzianum*	Ospina-Giraldo *et al.* (unpublished)	U58631
crea	DNA-binding protein	*T. reesei*	Takashima *et al.* (unpublished)	D63514

Table 9.1 *(Cont)*

Gene[a]	Product	Organism	Reference	Accession[b]
5.8s rRNA	5.8S ribosomal RNA	*T. longibrachiatum*	Ruiz-Sala *et al.* (1993)	L07957
bgn3.1	endo-1,3(4)-β-glucanase	*T. harzianum*	de la Cruz *et al.* (1995)	X84085
ind-a1	INDA1	*T. harzianum*	Vasseur *et al.* (1995)	Z22594
ind-c11	INDC11	*T. harzianum*	Vasseur *et al.* (1995)	Z22221
25s rRNA	25S ribosomal RNA	*T. reesei*	Vanhanen and Penttilä (unpublished)	X77580
5.8s rRNA	5.8S ribosomal RNA	*T. reesei*	Vanhanen and Penttilä (unpublished)	X77579
18s rRNA	18S ribosomal RNA	*T. reesei*	Vanhanen and Penttilä (unpublished)	X77581
actin	actin	*T. reesei*	Matheucci *et al.* (1995)	Z75421
cbh2	cellobiohydrase II	*T. reesei*	Stangl *et al.* (1993)	X70232 and S54964
endo51	endoglucanase I	*T. reesei*	Penttilä *et al.* (1986)	M15665
xy11	arabinofuranosidase/β-xylosidase	*T. koningii*	Huang *et al.* (unpublished)	U38661
cre154	Cre1	*T. reesei*	Strauss *et al.* (1995)	U27356
18s rRNA	18S ribosomal RNA	*T. longibrachiatum*	Kuhls (unpublished)	Z31019
cons-b4	serine + alanine-rich protein	*T. harzianum*	Goldman *et al.* (1994)	Z22229

Table 9.1 *(Cont)*

Gene[a]	Product	Organism	Reference	Accession[b]
xln2	endoxylanase II	*T. reesei*	Saarelainen *et al.* (1993)	S67387
tham-ch	endochitinase	*T. hamatum*	Fekete *et al.* (unpublished)	Z71415
imid	imidazoleglycerol-phosphate	*T. harzianum*	Goldman *et al.* (1992)	Z11528 and S47086
qid3	putative catabolite-repressed protein	*T. harzianum*	Lora *et al.* (1994)	X71913
glucu1	α-glucuronidase	*T. reesei*	Margolles-Clark *et al.* (1996a)	Z68706
cell1	1,4-β-D-glucan cellobiohydrolase	*T. viride*	Cheng *et al.* (1990)	X53931
xyn1	endo-β-1,4-xylanase I	*T. reesei*	Torronen *et al.* (1992)	S51973
xyn2	endo-β-1,-4-xylanase I	*T. reesei*	Torronen *et al.* (1992)	S51975
trpl32	ribosomal protein L32	*T. harzianum*	Lora *et al.* (1993)	X71914
th1433	14.3.3.protein	*T. harzianum*	Harman and Hayes (unpublished)	U24158

[a] Most of the gene names were derived from the original articles. Some of them were assigned by us.
[b] Accession number for the GenBank/EMBL DNA Sequence data library.

Trichoderma spp. active in biocontrol (Herrera-Estrella *et al.,* 1993). From data based on gene location and DNA homology (as deduced from hybridization signals), the same authors have shown that *T. harzianum* and *T. viride* are closely related and could have evolved in the same phylogenetic branch, whereas *T. reesei* would most probably have derived from an independent branch. In another study, Mäntylä *et al.* (1992) determined molecular karyotypes of strains of *T. reesei* that had undergone mutagenesis and screening to produce strains that are hyperproducers of cellulase. These authors showed that rather extensive alterations in genome organization occurred in these strains.

9.3 Gene cloning

A large number of *Trichoderma* genes have been cloned (Table 9.1). These genes have been cloned using differential hybridization (Goldman *et al.,* 1994; Vasseur *et al.,* 1995), synthetic probes based on protein sequence data (Geremia *et al.,* 1994), heterologous gene probes (Heindenreich and Kubicek, 1994), a combination of synthetic oligonucleotides and PCR-based amplification (Hayes *et al.,* 1994) or complementation utilizing adequate expression vectors in *Saccharomyces cerevisiae* (Goldman *et al.,* 1992).

9.4 Translation control sequences

Kozak (1978) proposed a model for the initiation of translation in eukaryotes in which the ribosomal subunits can scan the messenger RNA from the 5' end and initiate translation at the first AUG triplet encountered. The context of the triplet is important, and indeed there is a high degree of conservation of the sequence around the initiator codon, GCC$^{A}/_{G}$CC*AUG*G being the consensus in mammalian mRNAs (Kozak, 1987), $^{A}/_{Y}$A$^{A}/_{U}$A*AUG*UCU in *Saccharomyces cerevisiae* (Cigan and

Table 9.2 Frequency of bases around the translation initiation codon[a]

	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	+4[b]	+5	+6
G	8	7	9	6	13	7	2	5	4	5	7	11	8
A	11	13	8	8	11	7	6	30	18	20	13	6	11
T	10	10	12	5	13	15	8	1	10	1	11	15	15
C	14	13	14	24	6	14	27	7	11	17	12	11	9
Consensus sequence[c]						$^{T}/_{C}$	C	A	A	$^{A}/_{C}$	$^{A}/_{Y}$	$^{T}/_{G}$	$^{T}/_{A}$

[a] Data were compiled from the sequences listed in Table 9.1.
[b] The start codon AUG represents +1 to +3.
[c] The consensus sequence was assigned according to the following criteria: If the frequency of a single nucleotide is greater than or equal to 50% and greater than twice that of the second most abundant nucleotide, it is assigned as the consensus nucleotide and given in upper case. If the second criterion is satisfied but not the first, then the nucleotide is shown in lower case. If the sum of the frequencies is greater than 75% (but neither satisfies the above criteria), they are jointly assigned the status of the consensus.

Donahue, 1987), and $CA^{C}/_{A}{}^{A}/_{C}AUGC$ in filamentous fungi (Ballance, 1991). Table 9.2 shows the frequency of bases around the translation initiation codon for genes from *Trichoderma* spp. Based on these data, the consensus sequence for mRNA in *Trichoderma* spp. is $^{T}/_{C}CAA^{A}/_{C}AUG^{A}/_{Y}{}^{T}/_{G}{}^{T}/_{A}$. The consensus sequence around the initiator codon in Table 9.2 shows that the *AUG* environment in *Trichoderma* spp. is also highly conserved. The most important position would appear to be -3 (the A of *AUG* being +1 and the preceding base being -1), at which a purine is nearly always present (97%) (Kozak, 1987) and usually as an A.

9.5 Codon usage

The knowledge of the pattern of codon usage in a genome has a number of practical applications in the investigation of the molecular genetics of that species, e.g. in interpreting sequence data and in designing oligonucleotide probes. Table 9.3 shows the codon frequencies in *Trichoderma* spp. Codon usage in *Trichoderma* spp. was evaluated in terms of two statistics. Firstly, codon usage was added up every time the codon was used. A more sophisticated measure is the Relative Synonymous Codon Usage (RSCU) measure of Sharp and Devine (1989). This is an expression of the number of times a particular codon is used relative to how often it is expected to be used if codon usage bias does not exist. RSCU values that are close to 1.00 indicate that the particular codon is being used at about the unbiased frequency. As the RSCU value moves further away from 1.00, either there is a bias for more frequent use of the codon (RSCU values greater than 1.00) or there is a bias against the use of this codon (RSCU values less than 1.00). Using RSCU values has the advantage of normalising codon usage for each codon. If a particular amino acid is used frequently in a dataset, the number of times that the encoding triplets are used will seem quite high (the converse is true for amino acids with low frequency of use). RSCU values are independent of amino acid usage and so looking at these values can give a better estimate of codon preference. In this dataset, for instance, the UUC codon for Phe and the CUC codon for Leu are both used approximately the same number of times (431 in the case of Phe and 436 in the case of Leu). However, the UUC codon is used 1.37 times more often than expected, whereas the CUC codon is used 2.17 times more frequently than in a situation where no bias exists. The converse is true for the UGU codon for Cys and the GGG codon for Gly. While the latter is used more frequently, its RSCU value is further from 1.00, so although UGU is not used very frequently, the pressure against using GGG is greater. The frequency of use of GGG is related to the frequency with which Gly appears in the proteins in this dataset. In general, the codons that end with a strong-bonding nucleotide (G or C) appear to be favored. The average G + C composition of the dataset is approximately 58%, while the average G + C base composition at the third position of codons for which there is a synonymous alternative (all codons except those encoding Met, Trp and the three termination codons) is 70%. It is obvious that mutational pressure towards an elevated G + C-content genome has a considerable effect on codon usage. The exceptions to this rule appear to be when the middle nucleotide of the triplet is strongly bonding. In these cases, there is discrimination against a G in the third position. This situation does not seem to exist for these C-ending codons and in all cases the C-ending codons are used more frequently than expected.

Table 9.3 Codon frequencies in *Trichoderma* spp. genes[a]

AA	Codon	N	RSCU	AA	Codon	N	RSCU	AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	200	0.63	Ser	UCU	327	1.28	Tyr	UAU	163	0.51	Cys	UGU	62	0.42
	UUC	431	1.37		UCC	398	1.56		UAC	472	1.49		UGC	232	1.58
					UCA	109	0.43								
Leu	UUA	6	0.03		UCG	246	0.96	STOP	UAA	21	0.00	Trp	UGG	258	1.00
	UUG	127	0.63		AGU	71	0.28	STOP	UAG	14	0.00				
	CUU	193	0.96		AGC	381	1.49	STOP	UGA	6	0.00	Arg	CGU	146	1.27
	CUC	436	2.17										CGC	221	1.92
	CUA	35	0.17	Pro	CCU	267	1.08	His	CAU	75	0.40		CGA	112	0.97
	CUG	408	2.03		CCC	425	1.72		CAC	298	1.60		CGG	67	0.58
					CCA	107	0.43						AGA	53	0.46
Ile	AUU	252	0.96		CCG	191	0.77	Gln	CAA	140	0.39		AGG	91	0.79
	AUC	508	1.94						GAG	586	1.61				
	AUA	27	0.10	Thr	ACU	271	0.90					Gly	GGU	374	0.93
					ACC	553	1.83	Asn	AAU	169	0.36		GGC	896	2.23
Met	AUG	354	1.00		ACA	127	0.42		AAC	764	1.64		GGA	241	0.60
					ACG	255	0.85						GGG	95	0.24
Val	GUU	249	0.97					Lys	AAA	54	0.16				
	GUC	563	2.18	Ala	GCU	451	1.13		AAG	639	1.84	Glu	GAA	120	0.37
	GUA	34	0.13		GCC	758	1.90						GAG	529	1.63
	GUG	185	0.72		GCA	166	0.42	Asp	GAU	301	0.68				
					GCG	219	0.55		GAC	580	1.32				

[a] This table was compiled from a total of 17 109 codons (41 genes).

Sharp and Devine (1989) identified a small number of codons that appear to be "universally" preferred. These include the WWC codons (W = U or A): UUC (Phe), UAC (Tyr), AUC (Ile) and AAC (Asn). It appears that in *Trichoderma,* these codons are also used at a greater frequency than is expected and thus constitute preferred codons. The "universally optimum" UUC codon for Phe is used more than twice as often as the UUU codon. For Leu, the CUC and CUG codons are both used preferentially with both "A"-ending codons being rarely used. Of the three Ile codons, the AUA codon is only used one-tenth as often as expected, the AUU codon is used at about the expected frequency, and the AUC codon is preferred. Again the "A"-ending codon is rarely used to encode valine. The GUC codon is preferred and the GUA codon is used less than expected. This is probably due to the necessity for an optimum hydrogen-bonding interaction between the codon and its cognate amino-acyl tRNA. The GUG (strong-weak-strong) codon may involve a set of bonds that are too strong.

The codons that possess a "C" in the middle position and third position are used more often than expected. The codons that have a "C" in the middle position and a "G" in the third position do not appear to be favored (although their usage is only slightly less than the expected). The "A"-ending codons are used less than half as often as would be expected and the "U"-ending codons are used about as often as would be expected. When the strong-bonding "G" residue is found in the middle of a codon, there is reduced usage of the "G"-ending codons and a strong preference for the "C"-ending codons. These data suggest that the identity of the middle nucleotide of a triplet has a dramatic effect on the usage of the "G"-ending codons. When the middle nucleotide is weak-bonding (used two hydrogen bonds during duplex formation), there is a strong preference for the usage of "G"- and "C"-ending codons. When the middle nucleotide is strong-bonding (either an "A" or a "U"), then only the "C"-ending codons are preferentially used. The explanation for this probably lies either in steric hindrance or selection for more rapid translation of the codon. In cases where the codon-anticodon interaction is too strong, translation may be slowed down.

Of course the information in Table 9.3 does not take into consideration the variation within the dataset. The table merely presents a composite picture of the codon usage pattern for the dataset as a whole. It is necessary to use correspondence analysis to identify the major sources of variation in codon usage in the dataset. Correspondence analysis of a molecular dataset (usually carried out on the RSCU values) seeks to identify the major source of variation within the dataset. Each gene is assigned a position on a 59-dimensional axis, so constructed because there are 59 codons for which there is an asynonymous alternative (excluding the three STOP codons and Trp, which is encoded by UGG). If there is no codon usage bias, the "cloud" formed by the points representing the genes will appear spherical. If there is a codon usage trend (from high GC to low GC; or from high Effective Number of Codons (ENC) values to low ENC values; or from a high abundance of A in the first position to a high abundance of T in the first position;…or whatever), then the "cloud" will no longer look spherical but will assume a sausage-like shape. The axis that goes through the middle of this "sausage" is the axis that "corresponds" to the major source of variation in the dataset (the most important factor of dispersion). At one end of the axis are the genes with high codon bias and the genes with low codon bias are at the other end. The computer programs do scaling according to gene length and other considerations, but these are not of

Table 9.4 Correspondence analysis to identify the major sources of variation in *Trichoderma* spp. codon usage

Gene name	AX1	Laa	GC	GC3s	ENC
xyn1	−47	222	0.62	0.88	33.7
xln2	−47	223	0.62	0.89	30.7
ura5	−43	236	0.62	0.86	32.6
eg11	−41	463	0.65	0.90	33.1
hfb1	−41	97	0.65	0.78	29.6
imid	−38	208	0.66	0.85	34.7
indc11	−35	339	0.62	0.83	34.4
pk11	−34	538	0.61	0.82	31.9
eg15	−33	242	0.65	0.77	41.2
endo51	−29	459	0.63	0.86	37.2
tef1	−18	460	0.59	0.79	26.8
th1433	−15	262	0.59	0.77	34.9
tub2	−10	446	0.58	0.75	33.1
pgk49	−7	423	0.59	0.79	35.5
cre154	−7	402	0.62	0.73	41.8
pkt1	−7	662	0.59	0.76	43.8
cbh1	−6	513	0.59	0.73	40.7
crea	−6	402	0.62	0.73	41.9
pkc1	0	1139	0.60	0.73	45.8
b16-2	1	60	0.57	0.64	46.6
inda1	3	573	0.56	0.67	39.8
glucu1	5	847	0.57	0.71	48.0
celll	7	513	0.59	0.70	44.3
chit33	7	321	0.56	0.67	44.5
actin	9	366	0.56	0.69	33.6
cre1	9	409	0.60	0.68	46.1
pyr	9	379	0.58	0.67	50.5
tub1	11	446	0.55	0.64	40.6
xyn2	12	229	0.55	0.63	46.6
bgl1	15	744	0.58	0.67	49.0
xyl1	16	500	0.58	0.62	50.2
chi42, pc1ch1	23	424	0.54	0.67	41.8
endchi1	29	428	0.53	0.62	44.9
ech2	31	424	0.53	0.65	43.9
cbhII	32	471	0.55	0.56	55.1
consb-4	32	170	0.66	0.58	38.4
trp 132	34	137	0.58	0.66	46.5
eg3	40	418	0.53	0.56	55.9
prb1	42	409	0.52	0.49	44.8
axe1	43	302	0.58	0.58	57.0
bgn3	52	762	0.51	0.49	54.0

AX1: Position on the axis of greatest dispersion.
Laa: The number of amino acids in the gene.
GC: The G + C base composition summed over all positions.
GC3s: The G + C base composition at the third position of codons that have a synonymous alternative.
ENC: Effective number of codons.

Table 9.5 Results of the chi-squared test for significant differences between the RSCU values for the highly biased dataset and the lowly biased dataset (asterisk indicates a codon that is used significantly more frequently)

AA	Codon	N	RSCU	N	RSCU	AA	Codon	N	RSCU	N	RSCU
Phe	UUU	15	0.70	27	0.75	Ser	UCU	10	0.48	43	1.34
	UUC	28	1.30	45	1.25		UCC*	43	2.08	38	1.19
							UCA	1	0.05	34	1.06
Leu	UUA	1	0.09	3	0.15		UCG*	30	1.45	21	0.66
	UUG	1	0.09	26	1.27		AGU	2	0.10	13	0.41
	CUU	3	0.26	30	1.46		AGC	38	1.84	43	1.34
	CUC*	36	3.13	30	1.46						
	CUA	2	0.17	5	0.24	Pro	CCU	7	0.43	42	1.53
	CUG*	26	2.26	29	1.41		CCC*	33	2.03	29	1.05
							CCA	2	0.12	22	0.80
Ile	AUU	14	0.74	44	1.21		CCG*	23	1.42	17	0.62
	AUC*	42	2.21	58	1.60						
	AUA	1	0.05	7	0.19	Thr	ACU	11	0.44	59	1.27
							ACC*	56	2.22	71	1.53
Met	AUG	17	1.00	23	1.00		ACA	3	0.12	30	0.65
							ACG*	31	1.23	26	0.56
Val	GUU	6	0.31	54	1.69						
	GUC*	54	2.81	43	1.34	Ala	GCU	18	0.71	85	1.73
	GUA	0	0.00	11	0.34		GCC*	63	2.50	54	1.10
	GUG	17	0.88	20	0.62		GCA	2	0.08	34	0.69
							GCG	18	0.71	23	0.47
Tyr	UAU	2	0.06	32	1.05						
	UAC*	64	1.94	29	0.95	Cys	UGU	1	0.06	19	0.97
							UGC*	31	1.94	20	1.03
stop	UGA	2	0.00	1	0.00						
stop	UAA	2	0.00	3	0.00	Trp	UGG	19	1.00	38	1.00
stop	UAG	1	0.00	1	0.00						
						Arg	CGU	3	0.48	24	1.80
His	CAU	2	0.29	19	0.93		CGC*	26	3.63	13	0.98
	CAC*	12	1.71	22	1.07		CGA	0	0.00	14	1.05
							CGG	8	1.12	10	0.75
Gln	CAA	4	0.17	39	0.83						
	CAG*	44	1.83	55	1.17	Arg	AGA	0	0.00	8	0.60
							AGG	6	0.84	11	0.83
Asn	AAU	6	0.14	42	0.62						
	AAC*	79	1.86	94	1.38	Gly	GGU	9	0.25	62	1.12
							GGC*	115	3.22	96	1.74
Lys	AAA	3	0.13	10	0.36		GGA	7	0.20	52	0.94
	AAG	42	1.87	45	1.64		GGG	12	0.34	11	0.20
Asp	GAU	11	0.42	42	1.01	Glu	GAA	3	0.15	18	0.88
	GAC*	42	1.58	41	0.99		GAG*	36	1.85	23	1.12

Chi-squared values follow for optimal codons:

Codon UCC = 8.758	Codon CGC = 25.250	Codon ACG = 11.490
Codon UAC = 39.507	Codon CUG = 4.302	Codon GUC = 25.746
Codon UGC = 18.058	Codon CCG = 9.204	Codon GCC = 33.858
Codon UCG = 9.782	Codon CAG = 16.545	Codon GAC = 12.116
Codon CUC = 15.126	Codon AUC = 6.550	Codon GGC = 48.731
Codon CCC = 10.638	Codon ACC = 7.917	Codon GCA = 13.537
Codon CAC = 4.543	Codon AAC = 17.461	

major importance. The function of the analysis for any dataset is to identify why deviations from the spherical cloud occur, such as base and so on.

In every organism that has been examined to date, it has been shown that not all codons are used with equal frequency in all of the genes of the organism. Correspondence analysis finds the major source of variation in a dataset; at one end of this axis are the genes in which the greatest codon bias occurs. This amount of selectivity might be trivial, so a chi-squared test is used to see if there is a significant difference between the usage of codons in genes where less bias exists. Hereafter, "highly biased" will indicate that codon usage in that gene is more strongly biased than average. The reason for its position must be investigated to see if it is related to another statistic such as GC3s or ENC or position on the chromosome, etc. The significance of this phenomenon is that if a species exhibited a large long-term effective population size and is not subject to appreciable random genetic drift, then it will have had enough time to streamline its codon usage into an efficient means of rapidly translating mRNA. It will evolve a more biased codon usage pattern, which gives it a better chance of incorporating the correct tRNA (the population of which will also have reduced diversity) into the growing chain more quickly. The genes that benefit most from this kind of behavior are the highly expressed genes which exert a stronger selective pressure on the organism. In most prokaryotes and yeast we see the greatest bias in the highly expressed genes. In mammals, for instance, which have small, long-term effective population sizes and are subject to the vagaries of random genetic drift and frequent extinction, the codon usage is merely a reflection of the GC content of the region of DNA in which the gene resides. If we know which pattern a particular organism is likely to have, we can predict what the codon usage pattern for a particular (unknown) gene might be.

Thus, correspondence analysis was performed to identify the major sources of variation in *Trichoderma* spp. codon usage (Table 9.4). The genes are arranged in order of their appearance on the axis of greatest dispersion. The genes at the top of the table are the more biased genes and the genes towards the bottom of the table are less biased. In order to examine whether there was a difference in the usage of codons in the genes from either end of the axis of greatest dispersion, a total of five genes were selected from either end (1246 codons from one end and 2033 from the other). The cumulative RSCU values for each set were compared and a chi-squared test for heterogeneity within amino acid groups was carried out to the level of $P < 0.01$. A total of 20 codons were used significantly more frequently in the highly biased set than in the lowly biased set. This is an indication of the considerable amount of variation in codon usage within the dataset. The results of this analysis are shown in Table 9.5, with an asterisk denoting the codons that are used significantly more often in the highly biased dataset.

9.6 Conclusions

For *Trichoderma* spp., the estimated genome sizes range from 31 to 39 Mb and chromosome numbers range from 3 to 7. This large variation can be explained by assuming the hypothesis that variation in numbers and sizes of chromosomes is tolerated in imperfect fungi because meiosis does not occur and so chromosome pairing is unnecessary (Harman *et al.,* 1993; Kistler and Miao, 1992). Well over 50 genes from *Trichoderma* spp. have now been isolated and characterized. Their DNA

sequences have revealed the presence of a number of common sequence elements that might be important in the expression of these genes. We have shown an initial summary of their gene structure. Our analyses put more emphasis on translational rather than transcriptional signals. Further research on transcriptional signals will need more functional analysis *in vivo* and *in vitro*. Other points of interest for analysis are RNA splicing signals (intron splice junctions and internal consensus sequences), presence of signal peptides and DNA regions important for gene regulation. Some of these topics are currently under investigation in our laboratory.

References

BALLANCE, J.D. 1991. Transformation systems for filamentous fungi and an overview of fungal gene structure. In Leon, S.A. and Berka, R.M. (eds), Molecular Industrial Mycology Systems and Applications for Filamentous Fungi. Marcel Dekker, New York, pp. 1–29.

BERGES, T., PERROT, M., and BARREAU, C. 1990. Nucleotide sequences of the *Trichoderma reesei ura3* (OMPdecase) and *ura5* (OPRTase) genes. *Nucl. Acids Res.* **18:** 7183.

CARSOLIO, C., GUTIERREZ, A., JIMENEZ, B., VAN MONTAGU, M., and HERRERA-ESTRELLA, A. 1994. Characterization of *ech-42,* a *Trichoderma harzianum* endochitinase gene expressed during mycoparasitism. *Proc. Nat. Acad. Sci. USA* **91:**10903–10907.

CHEN, C.M., GRITZALI, M., and STAFFORD, D.W. 1987. Nucleotide sequence and deduced primary structure of cellobiohydrolase II from *Trichoderma reesei*. *Biotechnology* **5:**274–278.

CHENG, C., TSUKAGOSHI, N., and UDAKA, S. 1990. Nucleotide sequence of the cellobiohydrolase gene from *Trichoderma viride*. *Nucl. Acids Res.* **18:**5559.

CIGAN, A.M. and DONAHUE, T.F. 1987. Sequence and structural features associated with translational initiator regions in yeast—a review. *Gene* **59:**1–18.

DE LA CRUZ, J., PINTOR-TORO, J.A., BENITEZ, T., LLOBELL, A., and ROMERO, L.C. 1995. A novel endo-*ß*-1,3-glucanase, *BGN13.1,* involved in the mycoparasitism of *Trichoderma harzianum*. *J. Bacteriol.* **177:**6937–6945.

DRABORG, H., CHRISTGAU, S., HALKIER, T., RASMUSSEN, G., DALBOGE, H., and KAUPPINEN, S. 1996. Secretion of an enzymatically active *Trichoderma harzianum* endochitinase by *Saccharomyces cerevisiae*. *Curr. Genet.* **29:**404–409.

GEREMIA, R.A., GOLDMAN, G.H., JACOBS, D., VILA, S.B., ARDILES, W., VANMONTAGU, M., and HERRERA-ESTRELLA, A. 1994. Molecular characterization of the proteinase-encoding gene, *prb1,* related to mycoparasitism by *Trichoderma harzianum*. *Mol. Microbiol.* **8:**603–613.

GILLY, J.A. and SANDS, J.A. 1991. Electrophoretic karyotype of *Trichoderma reesei*. *Biotechnol. Lett.* **13:**477–482.

GOLDMAN, G.H., VASSEUR, V., CONTRERAS, R., and VAN MONTAGU, M. 1994. Sequence analysis and expression studies of a gene encoding a novel serine + alanine-rich protein in *Trichoderma harzianum*. *Gene* **144:**113–117.

GOLDMAN, G.H., VILLARROEL, R., VAN MONTAGU, M., and HERRERA-ESTRELLA, A. 1990. Sequence of the *Trichoderma viride* phosphoglycerate kinase gene. *Nucl. Acids Res.* **18:**6717.

GOLDMAN, G.H., TEMMERMAN, W., JACOBS, D., CONTRERAS, R., VAN MONTAGU, M., and HERRERA-ESTRELLA, A. 1993. A nucleotide substitution in one of the beta-tubulin genes of *Trichoderma viride* confers resistance to the antimitotic drug methyl benzimidazole-2-yl-carbamate. *Mol Gen. Genet.* **240:**73–80.

GOLDMAN, G.H., DEMOLDER, J., DEWAELE, S., HERRERA-ESTRELLA, A., GEREMIA, R.A., VAN MONTAGU, M., and CONTRERAS, R. 1992. Molecular cloning of the

imidazoleglycerolphosphate dehydratase gene of *Trichoderma harzianum* by genetic complementation in *Saccharomyces cerevisiae* using a direct expression vector. *Mol. Gen. Genet*. **234**:481–488.

HARMAN, G.E., HAYES, C.K., and LORITO, M. 1993. The genome of biocontrol fungi: modification and genetic components for plant disease management strategies. In Lumsden, R.D. and Vaughn, J.L. (eds), Pest Management Biologically Based Technologies: Proceedings of Beltsville Symposium XVIII, American Chemical Society, Washington, D.C., pp. 347–354.

HAYES, C.K., HARMAN, G.E., WOO, S.L., GULLINO, M.L., and LORITO, M. 1993. Methods for electrophoretic karyotyping of filamentous fungi in the genus *Trichoderma*. *Anal. Biochem.* **209:**176–182.

HAYES, C.K., KLEMSDAL, S., LORITO, M., DI PIETRO, A., PETERBAUER, C., NAKAS, J.P., TRONSMO, A., and HARMAN, G.E. 1994. Isolation and sequence of an endochitinase-encoding gene from a cDNA library of *Trichoderma harzianum*. *Gene* **138:**143–148.

HEIDENREICH, E.J. and KUBICEK, C.P. 1994. Sequence of the *pyr4* gene encoding orotidine-5'-phosphate decarboxylase from the biocontrol fungus *Trichoderma harzianum*. *Gene* **147:**151–152.

HERRERA-ESTRELLA, A., GOLDMAN, G.H., VAN MONTAGU, M., and GEREMIA, R.A. 1993. Electrophoretic karyotype and gene assignment to resolved chromosomes of *Trichoderma* spp. *Mol. Microbiol.* **7:**515–521.

ILMEN, M., THRANE, C., and PENTTILÄ, M. 1996. The glucose represser gene *cre1* of *Trichoderma:* Isolation and expression of a full length and a truncated mutant form. *Mol. Gen. Genet.* **251:**451–460.

KISTLER, H.C. and MIAO, V.P.W. 1992. New modes of genetic change in filamentous fungi. *Ann. Rev. Phytopathol.* **30:**131–152.

KOZAK, M. 1978. How do eukaryotic ribosomes select initiation regions in messenger RNA? *Cell* **15:**1109–1123.

KOZAK, M. 1987. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. *Nucl. Acids Res.* **15:**8125–8133.

LIMON, M.C., LORA, J.M., GARCIA, I., DE LA CRUZ, J., LLOBELL, A., BENITEZ,T., and PINTOR-TORO, J.A. 1995. Primary structure and expression pattern of the 33-kDa chitinase gene from the mycoparasitic fungus *Trichoderma harzianum*. *Curr. Genet.* **28:**478–483.

LORA, J.M., DE LA CRUZ, J., BENITEZ, T., LLOBELL, A., and PINTOR-TORO, J.A. 1994. A putative catabolite-repressed cell wall protein from the mycoparasitic fungus *Trichoderma harzianum*. *Mol. Gen. Genet.* **242:**461–466.

LORA, J.M., DE LA CRUZ, J., LLOBELL, A., BENITEZ, T., and PINTOR-TORO, J.A. 1995. Molecular characterization and heterologous expression of an endo-*ß*-1,6-glucanase gene from the mycoparasitic fungus *Trichoderma harzianum*. *Mol. Gen. Genet.* **247:** 639–645.

LORA, J.M., GARCIA, I., BENITEZ, T., LLOBELL, A., and PINTOR-TORO, J.A. 1993. Primary structure of *Trichoderma harzianum* ribosomal protein L32. *Nucl. Acids Res.* **21:** 3319.

MACH, R.L. 1994. Direct submission EMBL accession number U09580.

MÄNTYLÄ, A.L., ROSSI, K.H., VANHANEN, S.A., PENTTILÄ, M.E., SUOMINEN, P.L., and NEVALAINEN, H. 1992. Electrophoretic karyotyping of wild-type mutant *Trichodermalongibrachiatum (reesei)* strains. *Curr. Genet.* **21:**471–477.

MARGOLLES-CLARK, E., SALOHEIMO, M., SIIKA-AHO, M., and PENTILLÄ, M. 1996a. The *a*-glucuronidase-encoding gene of *Trichoderma reesei*. *Gene* **172:**171–172.

MARGOLLES-CLARK, E., TENKANEN, M., SODERLUND, H., and PENTTILÄ, M. 1996b. Acetyl xylan esterase from *Trichoderma reesei* contains an active-site serine residue and a cellulose-binding domain. *Eur. J. Biochem.* **237:**553–560.

MATHEUCCI, E. JR, HENRIQUE-SILVA, F., EL-GOGARY, S., ROSSINI, C.H., LEITE, A., VERA, J.E., URIOSTE, J.C., CRIVELLARO, O., and EL-DORRY, H. 1995. Structure, organization and promoter expression of the actin-encoding gene in *Trichoderma reesei*. *Gene* **161:**103–106.

MORAWETZ, R., MISCHAK, H., GOODNIGHT, J., LENDENFELD, T., MUSHINSKI, J.F., and KUBICEK, C.P. 1994. A protein kinase-encoding gene, *pkt1,* from *Trichoderma reesei,* homologous to the yeast *YPK1* and *YPK2 (YKR2)* genes. *Gene* **146:** 309–310.

MORAWETZ, R., LENDENFELD, T., MISCHAK, H., MUHLBAUER, M., GRUBER, F., GOODNIGHT, J., DE GRAAFF, L.H., VISSER, J., MUSHINSKI, J.F., and KUBICEK, C.P. 1996. Cloning and characterisation of genes (*pkc1* and *pkcA*) encoding protein kinase C homologues from *Trichoderma reesei* and *Aspergillus niger*. *Mol. Gen. Genet.* **250:**17–28.

NAKARI, T., ALATALO, E., and PENTTILÄ, M.E. 1993. Isolation of *Trichoderma reesei* genes highly expressed on glucose-containing media: characterization of the *tef1* gene encoding translation elongation factor 1 alpha. *Gene* **136:**313–318.

NAKARI-SETALA, T., ARO, N., KALKKINEN, N., ALATALO, E., and PENTTILÄ, M. 1996. Genetic and biochemical characterization of the *Trichoderma reesei* hydrophobin *HFBI*. *Eur. J. Biochem.* **235:**248–255.

PENTTILÄ, M., LEHTOVAARA, P., NEVALAINEN, H., BHIKHABHAI, R., and KNOWLES, J. 1986. Homology between cellulase genes of *Trichoderma reesei:* complete nucleotide sequence of the endoglucanase I gene. *Gene* **45:**253–263.

RUIZ-SALA, P., PEREZ-GONZALES, J.A., and RAMON-VIDAL, D. 1993. Nucleotidesequence of a *Trichoderma longibrachiatum* DNA fragment encoding the 5.8S rRNA gene. *Nucl. Acids Res.* **21:**741.

SAARELAINEN, R., PALOHEIMO, M., FAGERSTROM, R., SUOMINEN, P.L., and NEVALAINEN, K.M. 1993. Cloning, sequencing and enhanced expression of the *Trichoderma reesei* endoxylanase II (pI 9) gene *xln2*. *Mol. Gen. Genet.* **241:**497–503.

SALOHEIMO, A., HENRISSAT, B., HOFFREN, A.M., TELEMAN, O., and PENTTILÄ, M. 1994. A novel, small endoglucanase gene, *egl5,* from *Trichoderma reesei* isolated by expression in yeast. *Mol. Microbiol.* **13:**219–228.

SALOHEIMO, A., LEHTOVAARA, P., PENTTILÄ, M., TEERI, T.T., STAHLBERG, J., JOHANSSON, G., PETTERSSON, G., CLAEYSSENS, M., TOMME, P., and KNOWLES, J.K.C. 1988. *EGIII,* a new endoglucanase from *Trichoderma reesei:* The characterization of both gene and enzyme. *Gene* **63:**11–22.

SCHINDLER, M., MACH, R.L., VOLLENHOFER, S.K., HODITS, R., GRUBER, F., VISSER, J., DE GRAAFF, L., and KUBICEK, C.P. 1993. Characterization of the pyruvate kinase-encoding gene *(pki1)* of *Trichoderma reesei*. *Gene* **130:**271–275.

SCHLICK, A., KUHLS, K., MEYER, W., LIECKFELDT, E., BORNER, T., and MESSNER, K. 1994. Fingerprinting reveals gamma-ray induced mutations in fungal DNA: implications for identification of patent strains of *Trichoderma harzianum*. *Curr. Genet.* **26:** 74–78.

SHARP, P.M. and DEVINE, K.M. 1989. Codon usage and gene expression level in *Dictyostelium discoideum:* highly expressed genes do "prefer" optimal codons. *Nucl. Acids Res.* **17:**5029–5039.

SKINNER, D.Z., BUDDE, A.D., and LEONG, S.A. 1991. Molecular karyotype analysis of fungi. In Benett, J.W. and Lasure, L.L. (eds), More Gene Manipulations in Fungi. Academic Press, San Diego-New York-Boston, pp. 87–105.

STANGL, H., GRUBER, F., and KUBICEK, C.P. 1993. Characterization of the *Trichoderma reesei cbh2* promoter. *Curr. Genet.* **23:**115–122.

STRAUSS, J., MACH, R.L., ZEILINGER, S., HARTLER, G., STOFFLER, G., WOLSCHEK, M., and KUBICEK, C.P. 1995. *Cre1,* the carbon catabolite represser protein from *Trichoderma reesei*. *FEBS Lett.* **376:**103–107.

TORRONEN, A., MACH, R.L., MESSNER, R., GONZALEZ, R., KALKKINEN, N., HARKKI, A., and KUBICEK, C.P. 1992. The two major xylanases from *Trichoderma reesei:* characterization of both enzymes and genes. *Biotechnology* **10:**1461–1465.

VANHANEN, S., SALOHEIMO, A., KNOWLES, J.K.C., PENTTILÄ, M., and ILMEN, M. 1991. Promoter structure and expression of the 3-phosphoglycerate kinase-encoding gene *(pgk1)* of *Trichoderma reesei*. *Gene* **106:**129–133.

VASSEUR, V., VAN MONTAGU, M., and GOLDMAN, G.H. 1995. *Trichoderma harzianum* genes induced during growth on *Rhizoctonia solani* cell walls. *Microbiology* **141:**767–774.

WEY, T.T., HSEU, T.H., and HUANG, L. 1994. Molecular cloning and sequence analysis of the cellobiohydrolase I gene from *Trichoderma koningii* G-39. *Curr. Microbiol.* **28:** 31–39.