Percy: You know, they do say that the Infanta's eyes are more beautiful than the famous Stone of Galveston.

.....

Edmund: And what's that, exactly?

Percy: Well, it's a famous blue stone, and it comes (points dramatically) from Galveston.

Edmund: I see. And what about it?

Percy: Well, My Lord, the Infanta's eyes are bluer than it, for a start.

Edmund: I see. And have you ever seen this stone?

Percy: (nods) No, not as such, My Lord, but I know a couple of people who have, and they say it's very very blue indeed.

Edmund: And have these people seen the Infanta's eyes?

Percy: No, I shouldn't think so, My Lord.

Edmund: And neither have you, presumably.

Percy: No, My Lord.

Edmund: So, what you're telling me, Percy, is that something you have never seen, is, slightly, less blue than something else you have never seen.

"The Black Adder", episode 4

The study of a novel zinc finger gene cluster *TZF* and a genomic region flanking the Histone H4 replacement gene H4r of *Drosophila melanogaster*

D i s s e r t a t i o n
zur Erlangung des Grades
„**D o k t o r**
d e r N a t u r w i s s e n s c h a ft“

am Fachbereich Biologie
der Johannes Gutenberg-Universität
in Mainz

Vorverlegt von Wenli Gu
Geb. in Jiangsu, China
Mainz, 2001

Contents

Part I:

The study of a zinc finger gene cluster in *Drosophila melanogaster*

Introduction

Synaptonemal complexes (SCs) are structures found between paired homologous chromosomes in the meiotic prophase nuclei (Moses, 1968). They consist of two compact proteinaceous axes, the lateral elements, each one associating with one of the homologous chromosomes. The lateral elements are connected by thin transversal filaments. Between the two lateral elements another longitudinal structure is formed, the central element, which crosses the transversal filaments. The lateral elements and the central element make up the tripartite structure of the SCs. In most eukaryotes analysed so far, the condensation, pairing, recombination and segregation of chromosomes is accompanied by the assembly and disassembly of the SCs. Several rat SC proteins have been identified and the cDNA clones for some of them have been analysed (Heyting et al., 1985, 1987, 1988, 1989; Lammers et al. 1994, Smith and Benavente 1992). *Drosophila* is an exception because this structure is only formed during the meiotic prophase in females but not in males (Rasmussen 1973, 1974). At the same time, recombination does not take place in *Drosophila* males. On the other hand, there exist specialized structures in *Drosophila* primary spermatocytes, the lampbrush loops, which are formed by the actively transcribed fertility genes located on the Y chromosome.

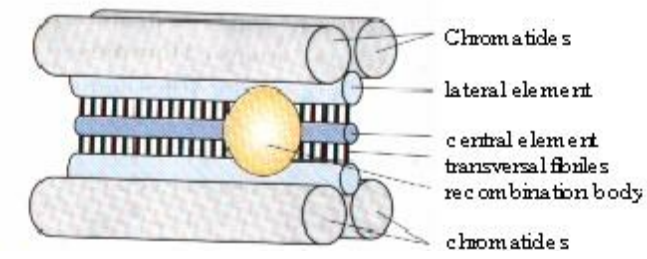


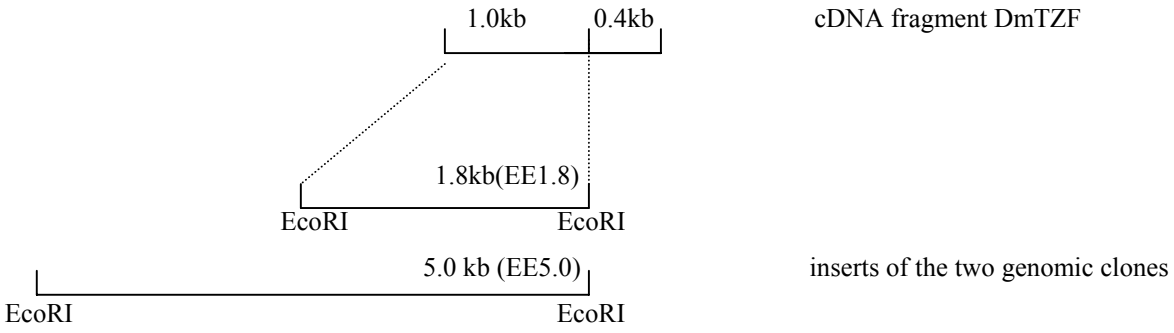
Fig 1: Schema of a synaptonemal complex (SC) in pachytene phase.

It was observed that antibodies against rat SC proteins cross react with Y chromosome lampbrush loops present in *Drosophila melanogaster* (Hennig & Heyting, unpublished data). An immunoscreening of *Drosophila* testis cDNA libraries with an antiserum against a rat SC protein (Lammers et al., 1994) was carried out to identify the structural component of this cross immunoreaction (Sun, 1994). A part of a novel gene was cloned which encodes a putative zinc finger protein. It was named *Tzf* gene. The *Tzf* gene is a single copy gene with particularly strong expression in embryos, testes and ovaries (Sun, 1994). It was the starting point of this work to complete the cDNA sequence of this gene and to find out its genomic organization.

Results

Genomic fragment containing *Tzf* gene

The 1.4 kb cDNA fragment of the *Tzf* gene, DmTZF (Sun, 1994), was used as a probe to screen a *Lambda* - DASH2 genomic DNA library. Two clones, with respectively a 5.0 kb and a 1.8 kb EcoRI-EcoRI insert were identified (Xiao, Xu & Hennig, unpublished data). The inserts were further subcloned, and the subclones were sequenced. The entire sequences of the EE5.0 and EE1.8 fragments were reconstructed from the subcloned sequences. The 1.0 kb fragment at the 5'part of the cDNA fragment DmTZF was contained in the sequence of genomic fragment EE1.8. The 0.4 kb 3' fragment was still missing.



To look for the genomic segment containing the 3'end of *Tzf* gene, a lawrist 4 cosmid library of *D. melanogaster* genomic DNA (kindly provided by Dr. Hoheisel, DKFZ) was screened with the cDNA fragment DmTZF. Three cosmid clones 89B4, 89B5 and 46A gave strong signals. 89B4 and 89B5 proved to be identical in their digestion patterns with several different enzymes. Only 89B4 was analysed in detail. The hybridisations of EcoRI and SalI digests of 89B4 DNA with the missing 0.4 kb cDNA fragment have shown that this fragment was contained in this genomic clone. It reacts with a 2.2 kb EcoRI fragment and a 9.0 kb SalI fragment. Both fragments were cloned into pBluescript and sequenced. A subclone of the 2.2kb EcoRI fragment, called eb2, contained the 0.4 kb cDNA fragment of

Tzf. Another subclone from the 9.0 kb SalI fragment, ab1, overlapped with the fragment EE1.8, the fragment EE5.0 from the λ DASH-2 library and the 2.2 kb EE fragment from the lawrist 4 cosmid library. This allowed deducing the relative localization of the three fragments (see Fig 2). Their organization was confirmed by sequencing the PCR products obtained with primers CX1 and DM5.

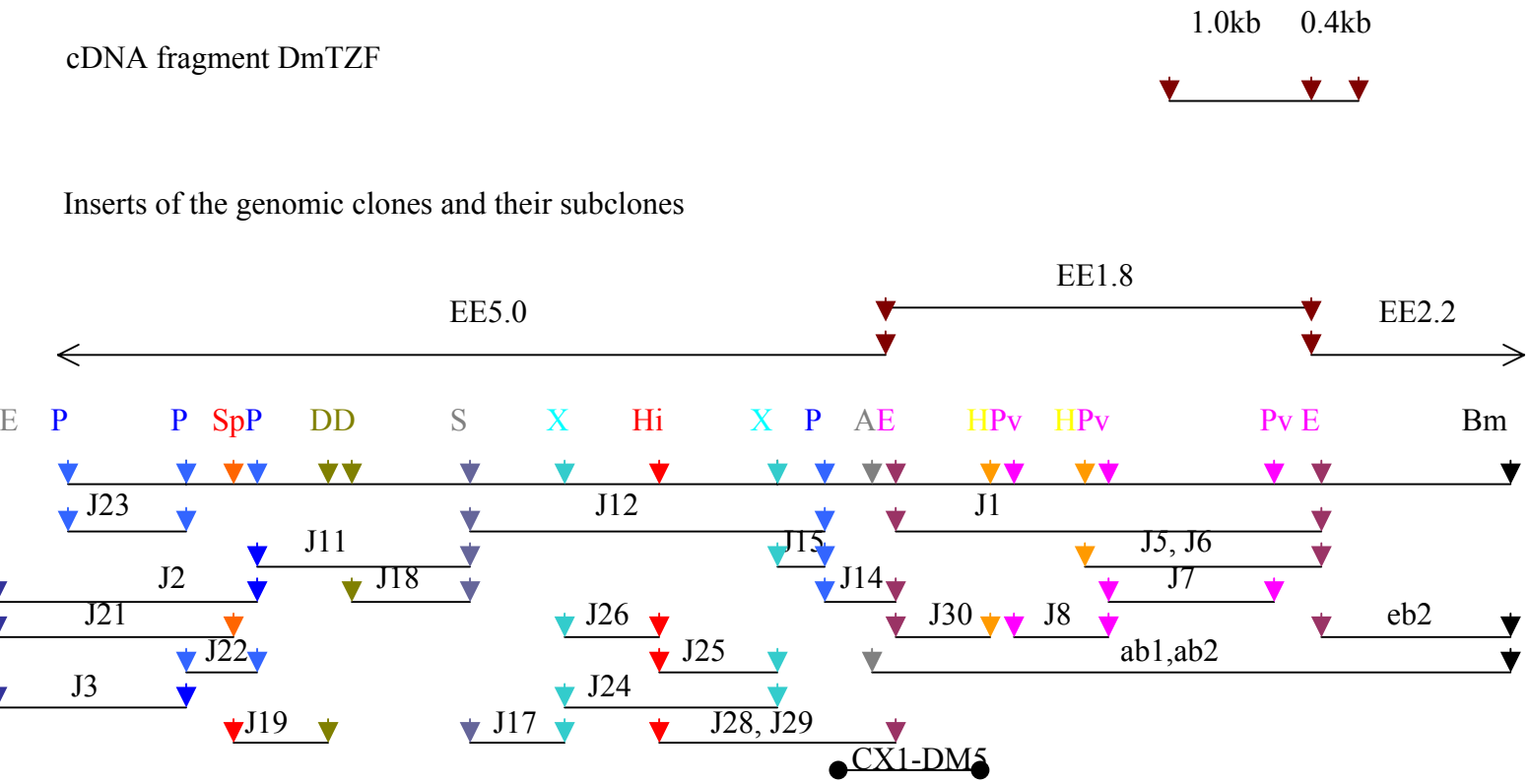


Fig 2: The organization of the subclones from the genomic fragments containing the *Tzf* and *Tzf*2 genes. The names of the restriction enzymes are abbreviated according to the abbreviation list in Appendix I.

Cloning of *Tzf*2 gene

Sequencing of EE1.8 and EE5.0 has proved that only EE1.8 contains the *Tzf* sequence. But EE5.0 also hybridised with the probe DmTZF which suggested a sequence similarity between EE5.0 and DmTZF. Analysis of the sequence of EE5.0 showed that parts of the sequence could also code for segments of a

zinc-finger protein. To see if these sequences are indeed expressed, RT-PCR experiments were performed with the primer pair GW1 and GW2, designed according to the EE5.0 sequence. Two different total RNA preparations were used as template. They were obtained from *D.melanogaster* ovaries and embryos.

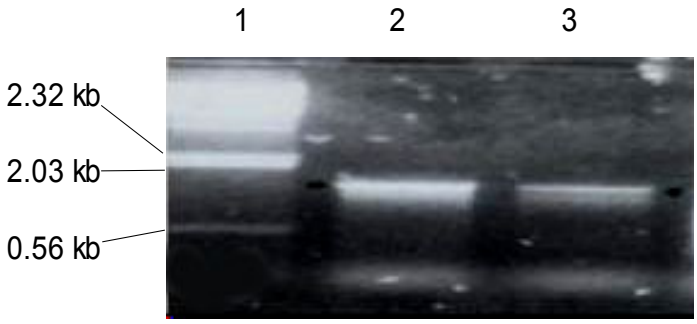


Fig 3: RT-PCR product amplified with primer GW1 and GW2

Lane1: DNA marker λ-Hind III Lane2, 3: RT-PCR product with embryo and ovary RNA

A major signal could be seen as product of the RT-PCR reactions with both embryo and ovary RNA as template. The DNA was recovered from the gel and digested with BglII and XhoI, which were designed into the 5'ends of GW1 and GW2 as integrated restriction sites. The digested fragment was ligated into pBluescript, digested with BamHI (compatible end with BglII) and XhoI. Two different clones, CX1 and CX9, were obtained, of which CX1 contains the whole RT-PCR product, and CX9 only part of it because of the existence of an internal BglII site in this fragment (see Fig 4). Different subclones were made from these two clones and the sequence of the whole fragment was obtained. The alignment of this sequence and the sequence of the genomic clone EE5.0 showed that the entire cDNA fragment GW1-GW2 is indeed contained in this segment of the genomic sequence. The alignment also showed the positions of the introns on the genomic sequence, which excluded the possibility that this product was an artefact caused by genomic DNA contamination in the RNA.

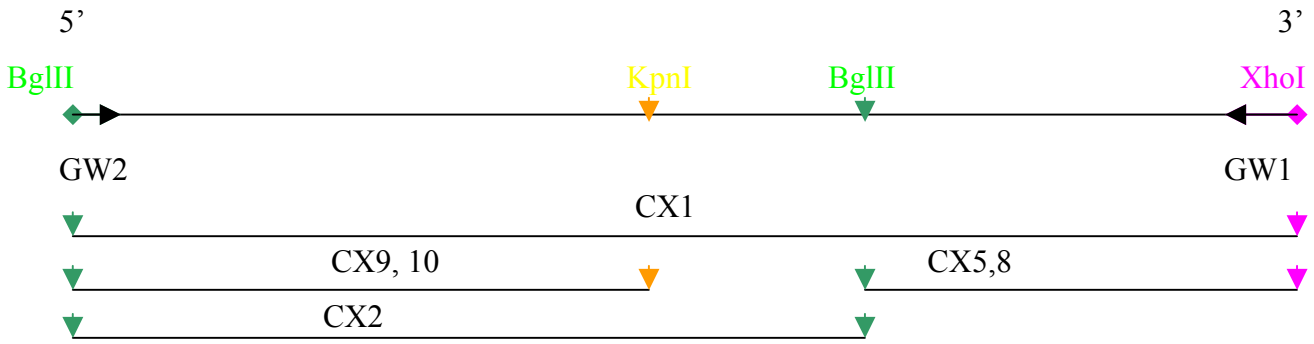
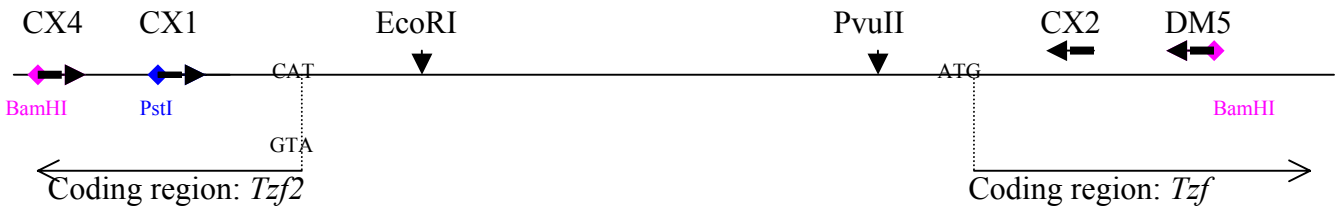


Fig 4: Subclones from the cDNA clone of *Tzf2*

An in-frame stop codon was found at the 3' end of this cDNA fragment. The 5'end of the cDNA was obtained through RACE reaction (see below). The start codon could be verified through an in-frame stop codon 50 bp upstream of it. This gene was named *Tzf2*.

These experiments have revealed two similar genes, *Tzf* and *Tzf2*. They exist in opposite orientation. Their start codons are only separated by 729 bp.

RACE Reaction



To obtain the 5' ends of the two genes, *Tzf* and *Tzf2*, RACE experiments were carried out for both cDNAs.

The *Tzf2* gene

8µg total RNA of *Drosophila melanogaster* embryo was reverse transcribed with primer CX4 to get a gene - specific cDNA pool for the *Tzf2* gene. 1/20 of the cDNA pool was tailed with poly-dC at the 3' end. The PCR reaction was then carried out with the primer CX1 and an adaptor. CX1 is an internal gene-specific primer downstream to the primer CX4, and the adapter anneals to the poly-dC end. The PCR product was too weak to be seen on an agarose gel.

The southern blot of an aliquot of the PCR product was therefore hybridised to the PCR product of primer CX1 and DM5, labelled through hot PCR. This probe recognizes the intergenic region between both *Tzf* genes. Another aliquot of the CX1-adapter PCR product was loaded on another agarose gel and the region corresponding to the signal in the hybridisation was cut out using the low-melting agarose gel method.

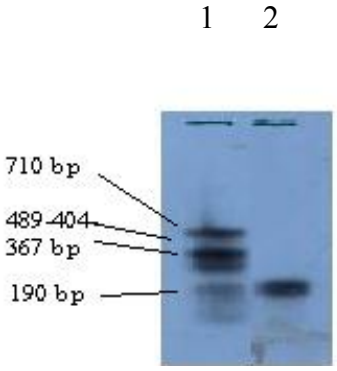


Fig 5: Hybridisation result of the first PCR blot with the intergenic probe CX1-DM5. The blot was once more hybridised with labelled pBluescript, so that the marker could be seen on the blot. The second hybridization did not change the signal position in the probe lane.
Lane 1: DNA marker: HpaII digestion of pBluescript.
LaneII: CX1-adapter PCR product.

1/20 of the LM agarose gel mixture was directly used as template for a reamplification with the same primer pair. CX1 and adapter now gave a PCR product visible through EtBr staining. It was cloned into pBluescript after digestion with PstI and HindIII, the restriction sites integrated respectively in both primers. Fifty clones were picked and their inserts checked by hybridising with the labelled intergenic probe CX1-DM5. Four clones gave positive signals and were thus sequenced. Among them, C1, C4 and C7 start from the same G_{3327} (Fig 6). This makes it likely that this G is the transcriptional initiation site of the *Tzf2* Gene.

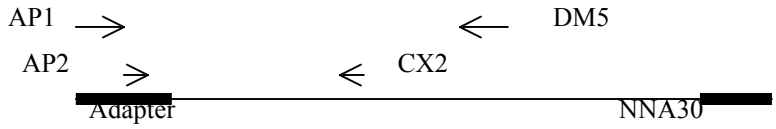
```
3201  CCGAGTCTCT CATTGTCGCG AATAGGAGAT GGATCTCTTC CAGGAGGCTA
TZF2 gene ←
3251  CTTACCCTCC GCTATAATAT TTATAATAAC ACGCTCGCTG GACGAAGACG
3301  ACGCCAAAAA AATAGATTGT AGTCTGGAGC GAAGTGCACG CGTTGCAATG
3351  CGAATGTCGC CACACAGGGT GACCACAGTC GACGCTAAGG GTATTCCATT
```

Fig 6: Transcription initiation of *Tzf2* gene. The nucleotides are numbered according to Appendix I.

The *Tzf* gene

The same strategy as for *Tzf2* was also performed for the *Tzf* gene, with DM5 as the primer to establish a cDNA pool. Primer CX2 and the same adapter as for the *Tzf* gene were used to amplify from this pool. They failed to give reasonable product even after several reamplifications. The Marathon RACE Kit from CLONTECH was then tried to reveal the transcription start of the *Tzf* gene in a different way.

Poly A^+ mRNA was obtained from total RNA made from *D.melanogaster* embryos with Oligotex™ Kit from Qiagen. First strand cDNA was obtained from 1 µg mRNA, using the AMV Reverse Transcriptase and a modified oligo (dT) primer provided in the Marathon RACE kit with two degenerate nucleotides positioned at the 3'end. A second-strand Enzyme Cocktail containing RNase H, *E.coli* DNA polymerase I and *E.coli* DNA ligase was then added to the first strand mixture to yield a double-strand cDNA pool, which was subsequently made blunt ended with T4 DNA polymerase. The Marathon cDNA adapter was ligated to the blunt ends of the double-stranded cDNA molecules. The first PCR was performed with the *Tzf*-specific primer DM5 and adapter-specific primer AP1. A nested PCR was performed using an aliquot of the first PCR product, with the primer CX2 down stream from DM5 and the primer AP2 downstream from AP1. The product was cloned into the PCR 2.1 vector (TA cloning kit, Invitrogen). Sequencing results have shown that eight clones contained the upstream sequence of *Tzf*, five of which stopped at the same nucleotide G^{3864}, which was identified as the transcription start of *Tzf*.



Tzf cDNA with adapter at both ends

```
3851  TTTCGCTGTG AATGGCCAGT GGCCGTGGAC GCTTAAATTC AGCTGTTTGA
3901  ATCTTCTGCA ACAACCGAAG TGATTGTTTT ATCGAATTAG CCATGAAAAC
                                                  Tzf gene
3951  TGAGTCCAAC GAGAAGTGGG TGGTGTGCCG CGTTTGCCTG AACAATCCCA
```

Fig 7 : Transcription initiation of the *Tzf* gene. The nucleotides are numbered according to Appendix I.

The structure of the *Tzf* and *Tzf2* genes

The cDNA sequences of *Tzf* and *Tzf2* genes were compared with the genomic sequence, and the genomic organization of both genes could be deduced (see Fig 8).

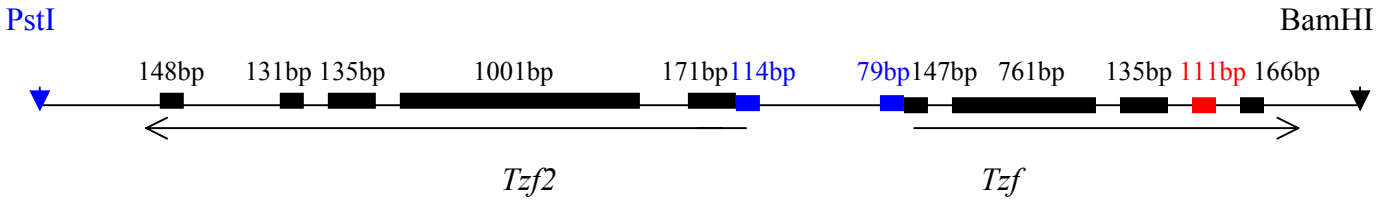


Fig 8: The gene structure of *Tzf* and *Tzf2*. The long thin line stands for the genomic sequence. Thick bars stand for the exons. Blue bars for the 5' UTR and black ones for coding regions. The red block in the *Tzf* gene stands for the exon that is possibly subject to alternative splicing (see below).

Tzf and *Tzf2* genes and the *Drosophila* genome project

After the characterizing of *Tzf* (Sun, 1994) and *Tzf2*, the complete *Drosophila melanogaster* genome was sequenced and published in the Flybase. The genomic and cDNA sequences of *Tzf* and *Tzf2* was used to search the Flybase and to compare our data and the genome sequencing data. The genomic sequence of *Tzf-Tzf2* region (listed in Appendix I) was identical with the celera sequence gb / AE003728.1 (*Drosophila* genomic scaffold 142000013386035 Section 53/105). Two genes identified for this genomic clone, Gene CG4413 and Gene CG4936 correspond to our *Tzf* and *Tzf2,* respectively.

GC4936 is identical with *Tzf2*. Comparison of the *Tzf* cDNA sequence and the cDNA sequence of the gene CG4413 recorded in the Flybase showed that the CG4413 transcript lacked the complete 111bp exon (see Fig 8 & Fig 9), which encodes the C-terminal end of the third zinc finger and the complete fourth finger.

951 ⟶ *Tzf* gene transcription
GCACATGAACACTCACACGGGGAACCGACCATACAAGTGCAACTACTGTCCAGCTGCCTTCGCCGATCGATCCACGAAAACCA
AACATCATAGgtgagaaatattcgaaacagttgaaagtaaattaaactgacattatttctctccag*AATTCACACTAAGGAGC*
3rd exon
GTCCCTACGTGTGCGACGTTTGCTCCAGAACGTTTACCTACTCGGACAACCTGAAGTTCCACAAGATGATTCACACGGGGGAG
4th exon
*AAGCCGCATGT*gtaagcatcaacatattttactacttcattatcctgcaataatagctattctctttcagCTGTGATCTTTGT
5th exon
GGCAAAGGATTTGTGAAGGCCTACAAATTGCGTTTGCATCGGGAAACGCATAATAGACGTATCACCTGGAGAAATGACGCAGA
1250

Fig 9: Genomic fragment containing part of the 3rd, 5th and the complete 4th exon (italic letters) of *Tzf* gene. Lower case letters are from the introns. The first gt and last ag of the introns are underlined. *Tzf* cDNA contains all three exons, whilst in CG4413 transcript, the third exon and the fifth exon are directly ligated together, skipping the fourth exon.

The *Tzf3* and *Tzf4* genes: two additional genes of a gene family

Several hundred base pairs away from the *Tzf* and *Tzf2* genes, two other zinc finger genes were found in the database, which also encode zinc finger proteins similar to the *Tzf* and *Tzf2* proteins. The protein products of all four genes have the same modular structure, and they are located next to each other.

Names of *Tzf* Genes	Names in the Flybase
Tzf	CG4413
Tzf2	CG4936
Tzf3	CG4854
Tzf4	CG4424

Alignment and Comparison of *Tzf, Tzf2, Tzf3* and *Tzf4*:

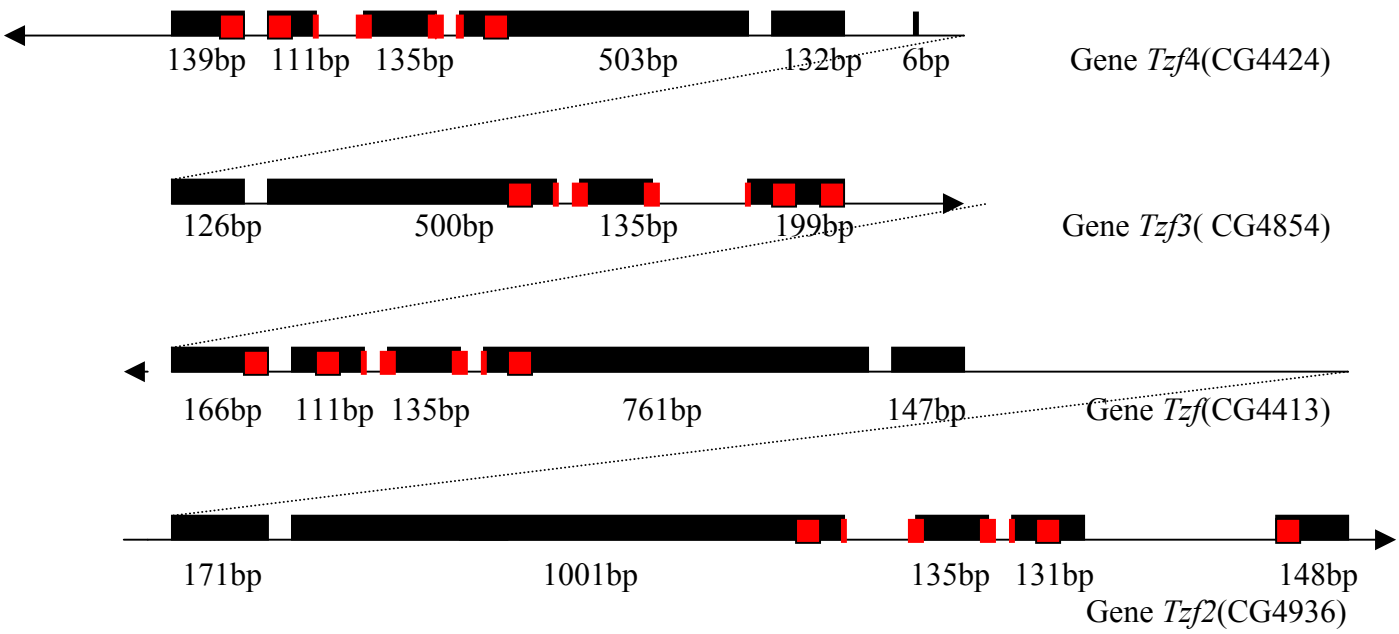


Fig 10: Genomic organization of the four *Tzf* genes. The thick bars represent the coding exons, the thin lines the introns. Only the exon sizes are presented, the introns are all shown in proportion to their length. The red boxes represent DNA segments coding for the zinc finger motifs. They are sometimes separated in two adjacent exons. The arrowheads show the direction of transcription, and the dashed lines connect the interrupted genomic sequence.

The *Tzf4* and *Tzf3* genes are localized in a head-to-head orientation, which corresponds to the orientation of *Tzf* and *Tzf2*. The arrangement of the five zinc fingers is conserved in all four genes.

```
      1                                                  50
TZF2  MRDSAAHASP AAAATSTQKW IVCRVCLQQP ...KEPMASI FNDDSE....
TZF3  ~~~~~~~~~~ ~MHTNVDSRD LKCRICLVQP ....KDESLM PTEP......
TZF4  ~~~~~~~~~~ ~~~~~~MAMM TLCRTCLQDG ...EAHMVSI FQTADDRLPG
TZF   ~~~~~~~~~~ ~MKTESNEKW VVCRVCLNNP SEGEELLHDI FSETAS....

      51                                                100
TZF2  .KDLTHMIRE CGGVPIKQ.. .FDHYPDKIC EKCFKVLKMA FKFRETCQRS
TZF3  ..DFPDKIKR CTGVELSE.. .SPDWPNRIC TSCALLLRAA LKLRSLCQQT
TZF4  GVSLCDKIES LSGIQIRATA KEEVLPTRIC LRCKAFLTLA HKFRQICQRS
TZF   .TRLDQMLHI CAGIPVSL.. .DDNFPDKMC SKCVRCLRLC YKFRLTCQRS

      101                                               150
TZF2  YGHLRQFVGP VEVEQRPPEK KGSETATKLE PDVDPDEAEQ EPEHDEEDED
TZF3  EKDLKEQK.. .......... .........L QEINIEIVHD E...QETKKK
TZF4  NEFLREYVIK .......... ....D..AVE QGVVKEVVQQ TR..PSTPPP
TZF   HQHIMDML.. .DREASNANA AGEGDLLSIA EDLSVESVLK S....WEDYA
```

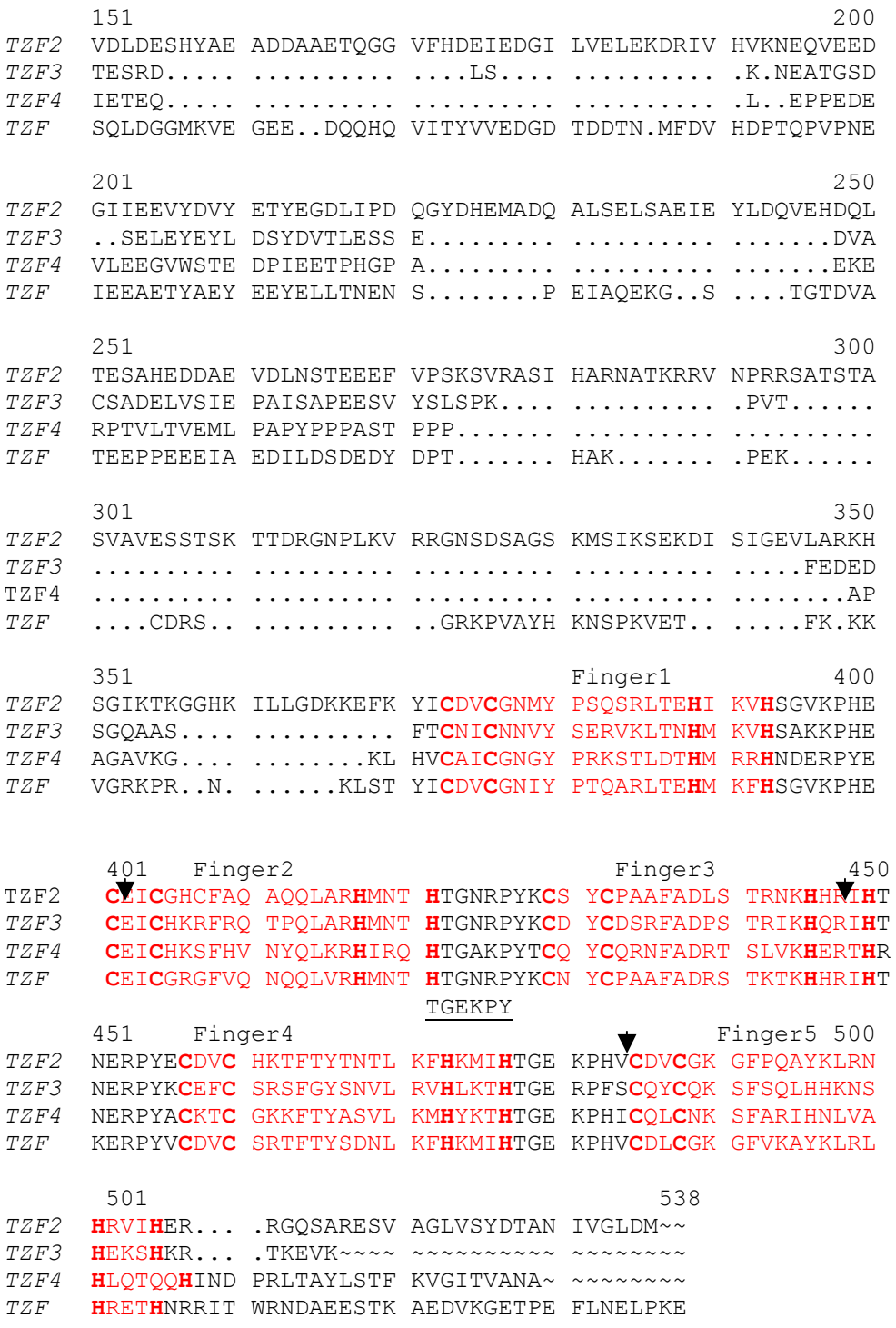


Fig 11: Amino acid sequence alignment of the four *TZF* proteins. The zinc fingers are highlighted in red. The sections with similarity to zinc fingers at the N-terminal ends of the proteins are highlighted in blue. Intron positions are shown with arrowheads.

All four gene products have at the C-terminal end a cluster of five C2H2-type zinc fingers with the consensus sequence Φ-X-C-$X_{2,4,5}$-C-X_3-Φ-X_5-Φ-X_2-H-$X_{3,4}$-H, where X represents any amino acid, Φ a hydrophobic residue. At the N-terminal end, some short sequences remind of zinc fingers but are

incomplete. This is typical for proteins including many copies of zinc finger domains, where incomplete or degenerate copies of finger domain are often found (Rosenfeld & Margalit, 1993). The inter-finger regions between fingers 2-3, 3-4 and 4-5 agree approximately with the TGEKPY consensus sequence named H-C link (Schuh et al., 1986), whilst the inter-1-2-region seems to be more degenerated than other inter-finger regions. MOTIF and PROSITE function from the HUSAR DNA analysis package were used to analyse all four sequences for further details. No other motifs were found, except for a cytochrome C domain in *TZF2*, 3 and 4 (C_{401}-K_{406} of *TZF3* and *TZF*4; C_{457}-K_{462} of *TZF2*). But this is probably of no relevance, because they all have one zinc finger domain with a H residue after the second C, which happens to be part of a cytochrome C consensus sequence (Mathews, 1985). PSORT function from the same package showed that all four *TZF* products are likely to be nuclear proteins.

Discussion

The four *Tzf* genes are paralogues

As shown in Fig 10, *Tzf* and *Tzf2* have the same gene organization, with the exon 2 containing the finger 1 and part of finger 2, exon 3 containing part of finger 2 and 3, exon 4 part of finger 3 and finger 4, and exon 5 the finger 5. *Tzf3* and *Tzf4* also obey this rule, although with small deviations. *Tzf4* has an additional small exon of 6 bp (named exon 0 to maintain the exon terminology of all *Tzf* genes) at the N-terminal of exon 1. In *Tzf3*, exon 4 and exon 5 are merged into one exon containing finger 4, finger 5 and part of finger 3. With these two exceptions, the all intron positions are completely conserved among the four genes. As shown in Fig 11 they are located in the following positions:
Intron between exon 1 and 2: between P(E,Q)65 and I(L,V)66,
Intron between exon 2 and 3: in E2 of the finger 2, between the 2nd and the 3rd position of the codon,
Intron between exon 3 and 4: in R(L,K)14 of finger 3, between the 2nd and the 3rd position of the codon,
Intron between exon 4 and 5: in V(S,I) 2 of finger 5, between the 2nd and the 3rd position of the codon.

Therefore, the four *Tzf* genes have an almost identical genomic organization. Also, their products share the same modular structure with five canonical zinc fingers at the C-terminal end and possible remnants of two fingers at the N-terminal end. Sun (Sun, 1994) divided the *TZF* protein into five domains according to their isoelectric values (pI values) and their homology with the corresponding segments of the *Drosophila hydei* homologue of *Tzf* (Dh*Tzf*) (Sun, 1994). This domain system was extended to all four *TZF* proteins. The following two tables show the pI value and size of the five domains, as well as their similarity and identity among the four *TZF* proteins.

	Domain I (pI / size)	Domain II (pI / size)	Domain III (pI / size)	Domain IV (pI / size)	Domain V (pI / size)
TZF	7.22 / 85	3.38 / 147	11.26 / 41	9.47 / 133	4.51 / 33
TZF2	8.22 / 93	3.74 / 178	11.59 / 90	8.63 / 133	4.62 / 27
TZF3	6.95 / 78	3.97 / 85	3.33 / 16	10.16 / 133	11.07 / 7
TZF4	7.95 / 85	3.99 / 87	10.80 / 12	10.26 / 135	9.55 / 22

	Domain I (identity similarity)		Domain II		Domain III		Domain IV		Domain V	
TZF vs TZF2	**59.756**	**42.683**	37.931	20.000	37.500	25.000	**81.203**	**79.699**	44.444	22.222
TZF2 vs TZF3	46.154	32.051	37.037	25.926	25.000	25.000	62.406	57.143	42.857	14.286
TZF3 vs TZF4	38.158	27.623	29.268	21.951	14.286	14.286	56.391	48.120	28.571	14.286
TZF vs TZF4	51.316	35.526	**50.000**	**33.333**	25.000	25.000	57.143	48.872	83.333	33.333
TZF vs TZF3	38.462	29.487	35.484	22.581	33.333	26.667	62.406	55.639	57.143	28.571
TZF2 vs TZF4	46.753	35.065	36.047	27.907	33.333	25.000	52.632	47.368	26.316	21.053

Domain IV, which contains the five zinc fingers, is shared with the highest similarity among the *TZF* proteins. Domain I with the N-terminal finger remnants also has a considerable inter-*TZF*s similarity. For both domains, the proteins *TZF* and *TZF2* are most similar among the four *TZF*s. Domain II is the largest domain for *TZF* and *TZF2*, but it is much smaller for the other two proteins. In spite of the length heterogeneity, domain II of all four proteins is very acidic with the pI value between 3.38 and 3.99. The highest amino acid sequence similarity is between *TZF* and *TZF4*. The a.a. sequences of domain III and domain V are considerably diverged, as is their lengths.

Unlike most gene clusters generated through unequal crossover, the *Tzf* genes do not have a tandem orientation. Small chromosomal inversions could account for their opposite orientations. The simplest scenario with the least number of chromosomal rearrangements would be as following:
Tzf and *Tzf2* were generated from the ancestral zinc finger gene. A small chromosomal inversion of one of both copies caused their opposite orientations. Subsequently *Tzf* and *Tzf2* were again duplicated by an unequal crossover, creating *Tzf3* and *Tzf*4. Alternatively, *Tzf* and *Tzf2* could be a duplication of *Tzf3* and *Tzf*4.

The pattern of silent (synonymous) nucleotide changes within the zinc finger coding region provides a measure of evolutionary distances between the paralogues (Li&Graur, 1991). The amino acids conserved among all four *TZF* proteins were sorted out and their codons were compared in the following tables:

	Finger1							Fingerlink 1-2	
	C_{373}	C_{376}	N_{378}	Y_{380}	L_{386}	H_{389}	H_{393}	P_{398}	E_{400}
TZF	tgc	tgc	aat	tat	ctc	cac	cat	cca	gag
TZF2	tgc	tgc	aac	tat	ctt	cac	cac	ccg	gag
TZF3	tgc	tgc	aat	tac	ttg	cac	cac	cca	gaa
TZF4	tgc	tgt	aat	tat	ctg	cac	cat	cct	gag

	Finger2									Fingerlink2-3			
	C_{401}	E_{402}	I_{403}	C_{404}	Q_{413}	L_{414}	R_{416}	H_{417}	H_{421}	T_{422}	G_{423}	P_{426}	Y_{427}
TZF	tgc	gag	atc	tgc	cag	ctg	cgg	cac	cac	acg	ggg	cca	tac
TZF2	tgc	gag	atc	tgt	cag	ctg	cgc	cac	cac	acc	gga	ccg	tac
TZF3	tgc	gaa	atc	tgt	cag	ttg	agg	cac	cac	acc	ggt	ccc	tac
TZF4	tgc	gag	att	tgc	cag	ctg	cgc	cac	cac	acg	gga	cca	tat

	Finger3										Fingerlink3-4			
	C_{429}	Y_{431}	C_{432}	F_{436}	A_{437}	D_{438}	K_{444}	H_{445}	R_{447}	H_{449}	E_{452}	R_{453}	P_{454}	Y_{455}
TZF	tgc	tac	tgt	ttc	gcc	gat	aaa	cat	aga	cac	gag	cgt	ccc	tac
TZF2	tgc	tat	tgc	ttc	gcc	gac	aag	cac	aga	cac	gag	cga	ccc	tac
TZF3	tgt	tat	tgc	ttc	gcc	gat	aag	cat	agg	cac	gaa	cga	ccg	tac
TZF4	tgc	tat	tgc	ttc	gcg	gat	aag	cat	aga	cat	gag	cgt	cct	tat

	Finger4							fingerlink4-5		
	C_{457}	C_{460}	F_{464}	Y_{466}	L_{470}	H_{473}	H_{477}	T_{478}	G_{479}	E_{480}
TZF	tgc	tgc	ttt	tac	ctg	cac	cac	acg	ggg	gag
TZF2	tgc	tgc	ttc	tac	ttg	cac	cat	acg	gga	gag
TZF3	tgc	tgc	ttt	tac	ctc	cat	cat	acc	ggt	gaa
TZF4	tgc	tgc	ttc	tat	ctt	cac	cac	acg	ggc	gaa

	Finger5				
	C_{485}	C_{488}	K_{490}	F_{492}	H_{501}
TZF	tgt	tgt	aaa	ttt	cat
TZF2	tgc	tgc	aag	ttc	cac
TZF3	tgc	tgc	aag	ttc	cac
TZF4	tgc	tgc	aaa	ttc	cac

For each two different genes, the number of nucleotide substitutions was counted and listed in the following table:

	TZF/TZF2	*TZF3/TZF4*	*TZF/TZF3*	*TZF2/TZF4*	*TZF/TZF4*	*TZF2/TZF3*
Synonymous substitutions	24	28	31	24	23	24

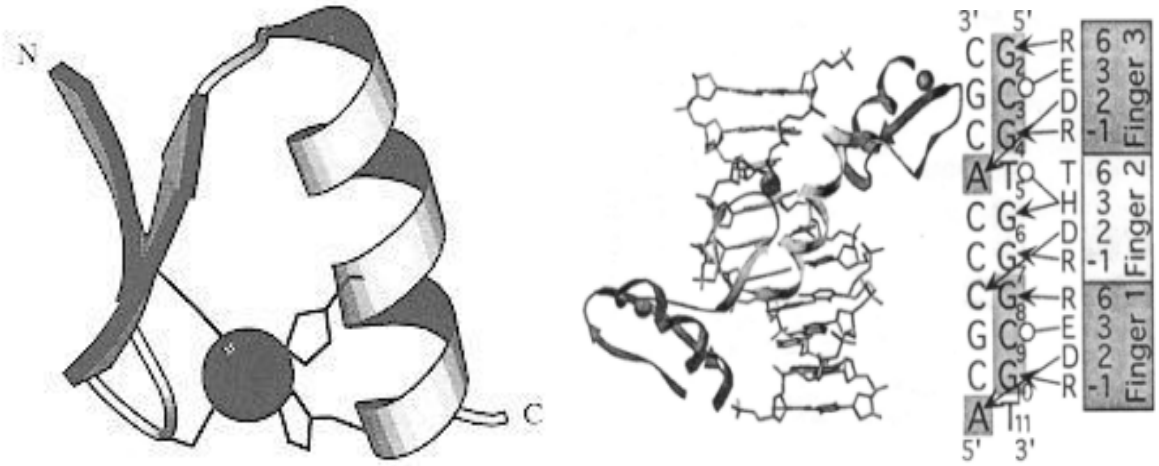
Between *TZF*/*TZF4* are the least synonymous substitutions, between *TZF2*/*TZF3* the second least. This result is not contradictory to the crossover-inversion-crossover hypothesis, although a definite proof of the hypothesis still demands further data.

Possible functions of the *TZF* proteins

Current estimates suggest that genes encoding C2H2 type zinc fingers account for 0,7%-2% of eukaryotic genomes (Hoovers et al., 1992; Shannon et al., 1998; Clarke&Berg, 1998; Böhm, unpublished data). Present evidence suggests that most of these genes encode sequence specific nucleic acid binding proteins. Although some of these proteins are known to participate in pattern formation, cellular proliferation and tumorigenesis (Wieschaus et al., 1984; Boulay et al., 1987; Schuh et al., 1986; Call et al., 1990; Supp et al., 1996), the biological function of the vast majority of zinc finger proteins is still unknown.

DNA-binding?

Most of C2H2 type zinc finger proteins are DNA-binding transcription factors. They fold in the presence of zinc to form a ββα domain. Each finger binds a single zinc ion that is sandwiched between the two-stranded antiparallel β-sheet and the α-helix. The zinc ion is tetrahedrally coordinated between two cysteines at one end of the β-sheet and the two histidines in the C-terminal portion of the α-helix (Michael et al., 1992; Shi & Berg, 1995). X-ray studies of zinc finger-DNA complexes (Pavletich & Pabo, 1991) showed that the α-helical portion of each finger fits in the major groove of DNA, and that the binding of successive fingers causes the protein to wrap around the DNA. The majority of base contacts occurs among three base pair segments along one DNA strand with four amino acids at the position -1, 2, 3, and 6 (see Fig 12).



(a) (b)

-1 12 3 4

Finger 1 PYA**C**PVESCDRRFSRSDELTR**H**IRI**H**TGQK

Finger 2 PFQ**C**RI--- CMRNFSRSDHLTT**H**IRT**H**TGEK

Finger 3 PFA**C**DI---CGRKFARSDERKR**H**TKI**H**LRQK

\+ ! + φ φφ+ φ ! !

(c)

Fig 12 **(a)**The ββα domain of zinc finger protein Zif268. **(b)** Structure of the three fingers of Zif 268 bound to DNA. Base contacts made from position –1, 2, 3, and 6 of each helix are indicated schematically to the right of the structure. Arrows indicate contact mediated by hydrogen bonds; open circles indicate hydrophobic interactions (Eltrod-Erickson et al., 1998) **(c)** The sequence of the three fingers is shown with the H and Cs participating in the zinc coordination in bold. The positions of amino acids are numbered according to convention. *Filled squares* below the sequence indicate the positions of the conserved hydrophobic residues. *Filled circles* and *stars* indicate residue positions that are involved in phosphate and base contacts (respectively) in most of the fingers.

Different zinc fingers could bind to different DNA segments. Because of the binding pattern of the zinc finger proteins to the DNA molecules, the specificity of the DNA recognition of zinc finger is mostly only dependent on the four amino acids at the position -1, 2, 3, and 6 (Pavletich & Pabo, 1991; Elrod-Erickson et al., 1998). As the data from zinc finger selection continues to grow, especially from the experiments to "design" zinc fingers to recognize certain DNA sequences, it becomes possible to predict the zinc finger specificity according to its amino acid sequence (for review see Wolfe et al., 2000; Choo & Isalan, 2000).

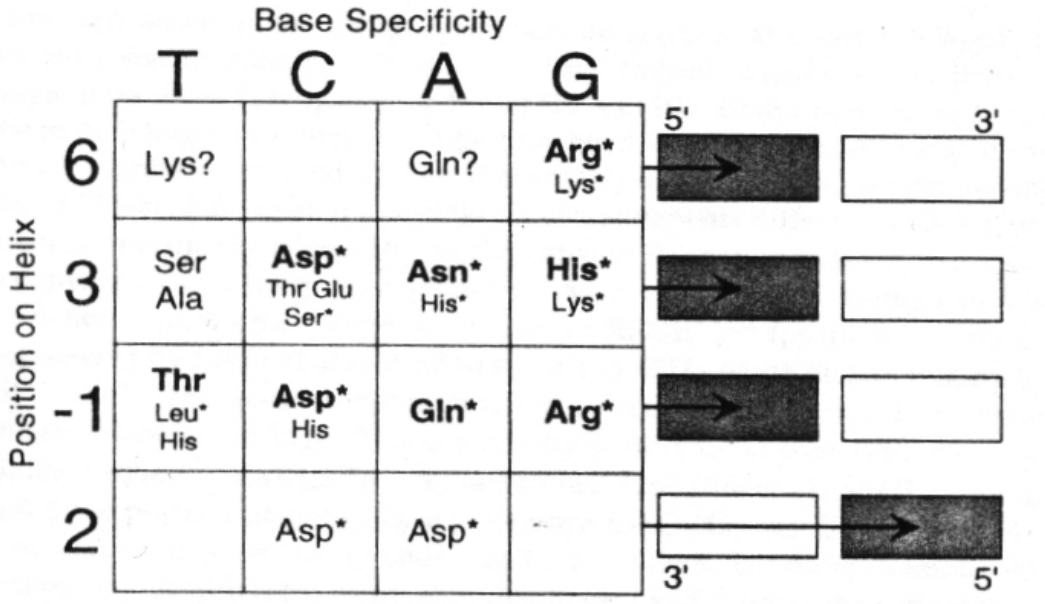


Fig 13: Pattern of side-chain base interactions that provide an approximate "recognition" code of zinc fingers that have a canonical binding mode. This chart describes contacts between residues at key positions in the α-helix (-1, 2, 3 and 6) and bases at the corresponding positions in the canonical subsite (see Fig 12b). Boldface type highlights amino acids that occur most frequently in phage display "designing" selections when a particular base specificity is desired, and an asterisk indicates contacts that have been observed in structural studies. Question marks indicate that the specificity of the respective amino acid/base contact is uncertain. Positions for which base specificity is largely undefined are left blank. Adapted from Wolfe et al., 2000.

These rules were used to predict the DNA-binding specificity of the four *TZF* proteins, giving following result for each of them:

TZF proteins	Predicted binding-sequence
TZF1	xxTGxATCCxAxxGx
TZF2	xxxGxATCCxCxxGA
TZF3	xGxGxATCCxxxxGA
TZF4	xCGGxxTTCxxxxAG

According to this prediction, the four *TZF*s could bind to very similar sequences, especially as the middle parts of their recognition sequences are almost identical (GxATCC). They could either compete with each other for the same DNA fragment, or they could bind to a larger DNA sequence consisting of

of several binding motifs in tandem. Or, an alternative possibility is that they have no function in DNA binding but that they bind to RNA or participate in protein interactions with their zinc finger domains.

RNA-binding?

Zinc fingers are potentially nucleic acid-binding proteins, which could be DNA-, RNA-, or DNA- and RNA-binding (for review see Leon & Roth, 2000). It was proposed that the conserved H-C link sequence is only present in DNA-binding proteins, but not in RNA-binding ones (Darby & Joho, 1992). This might be taken as an indication that the *Tzf* genes code for DNA-binding proteins. This prediction is, however, uncertain. The protein XFG5-1, for example, is a specific RNA-binding protein in *in vitro* experiments, but contains the conserved H-C link (Köster et al., 1991). The Wilm's tumor suppressor, WT1, is another good example, which was initially characterised as a DNA-binding protein. Now it is more likely to function in gene regulation at the RNA-level as well (Larsson et al., 1995; Bruening et al., 1996; Bordeesy & Pelletier, 1998). A number of C2H2 zinc fingers have been identified that bind RNA, but, aside from TFIIIA and p43, the biological significance of these interactions requires further study (Friesen & Darby, 1997; 1998;).

Protein binding?

Recently, zinc fingers were shown to be able to perform not only DNA-binding function, but also to participate in protein-protein interactions. Among the C2H2 fingers, zinc finger proteins Ikaros and Aiolos were identified to have both DNA-binding fingers and protein-protein interaction fingers responsible for homo- or heterodimerization (Sun et al., 1996; Georgoporos et al., 1997; Morgan et al 1997). Considering the clustered localization of the *Tzf* genes which provides a structural basis for a co-regulation of their expression, some fingers of the *TZF* proteins could very well also function in forming homopolymers from one *TZF* subunit, or heteropolymers from two or more different *TZF*s. Different combinations of different *TZF*s with the varying DNA sequence they recognize could bind different target DNA sequences. They could also control the interaction of the *TZF* proteins with other proteins.

The four *Tzf* genes are tightly clustered

Zinc finger protein clusters in mammals

A considerable body of evidence has emerged to suggest that a large fraction of human zinc finger genes (ZNF genes) are arranged in clusters (with two or more genes located in the same interval) scattered throughout many chromosomes, including 3,7,8,10,11,12,16,17,19,20,21,22, and X. (Aubry et al., 1992; Hoovers et al., 1992; Huebner et al., 1991; Rousseau-Merck et al., 1993; Lichter et al., 1992; Saleh et al., 1992; Tommerup &Vissing, 1995). Especially the ZNF91 and ZNF45 gene families, located at q13.2 and p12-p13.1 of human chromosome 19 (H19), were intensively studied and provided informations on evolutionary aspects of zinc finger genes (Constantihou-Deltas et al., 1992; Shannon et al., 1996; Bellefroid et al., 1993). In both cases, the zinc finger genes are arranged in "head-to-tail" tandem arrays with relatively even spacing between neighbouring genes. This observation was consistent with the idea that familial gene clusters have arisen primarily through multiple, *in situ* tandem duplication events of single progenitor loci (Ohno, 1970). Members of a tandem gene family are more similar in sequence to one another than they are to ZNF genes located elsewhere in the chromosome. Less is known about the their genomic organization of other ZNF gene clusters, but there is evidence that they are also mostly arranged as "head-to-tail" tandem arrays (Shannon et al., 1998; Calabro et al., 1995; Aubry et al., 1992; Derry et al., 1995).

Comparative studies of region H19 in human and the homologous mouse chromosome region have provided evidence that many, and perhaps all, of the clustered ZNF families located in H19 are represented by similar clusters located in syntenic regions of mouse chromosomes 7, 8, 9, 10, and 17 (Shannon et al., 1998). The mouse genes are highly similar to their human counterparts in overall structure and patterns of expression. But the human genes possess more repeats than their murine orthologs. The extent of conservation of individual repeat units varies among the zinc fingers encoded by orthologous genes, with amino acid identities of individual zinc finger repeats ranging from 68%-96% (Shannon et al., 1998).

Zinc finger protein clusters in *Drosophila*: *Tzf*1-4 as the tightest ZF gene cluster

The distribution of zinc finger protein coding genes was well studied for yeast *S. cerevisiae* (Böhm et al., 1997) and *C. elegans* (Clarke & Berg, 1998). Neither organism reveals clusters in genome. In yeast, the majority of zinc finger proteins contain exactly two finger domains; only 10% have more than two. In contrast, *C.elegans has* more zinc finger proteins with three or more fingers than those with only two. Comparing the data from *S.cerevisiae*, *C.elegans*, mouse and human, a clear evolutionary trend

enhancing the complexity in the zinc finger protein domain structure can be seen. First, the number of fingers in a protein increases. Subsequently, the number of genes was increased by duplication. This appears to reflect an increasing demand in regulatory proteins for more complex organisms.

In *Drosophila*, zinc finger gene clusters have 2-5 members (S. Böhm, personal communication). One example of well-studied clusters is the one including *odd-skipped (odd), sob* and *bowel (bowl)* (Hart et al., 1996). *Odd* and *sob* are located at 24A, in the same orientation, with 24kb of DNA between them. *Bowl* is located further away, at 24C2-5. *Sob* and *bowl* have five C2H2 type zinc fingers and *odd* has four. The fingers of *odd* and the N-terminal four fingers of *sob* and *bowl* share an extremely high identity (97,3% between sob and *bowl*, 86,6% between *sob* and *odd* and 87,5% between *bowl* and *odd*). Beyond the zinc fingers, no homology was indicated. The expression of *sob* and *odd* are strikingly similar, while *bowl* is expressed in a different pattern. It was suggested that all three genes have a common ancestor, and arose by at least two independent duplication events (Hart et al., 1996).

Two other zinc finger genes *spalt (sal)* and *spalt-related (salr*) at region 32F/33A are located head to head, in opposite orientations, approximately 65 kb from each other (Reuter et al., 1996; Celis et al., 1996). They have at least partially redundant functions, share high sequence homology and a late expression pattern. s*alr* does not have the early expression pattern of *sal*. Another gene, *spalt adjacent (sala*), was identified in a tail-to-tail orientation with *sal*. Although not sharing any amino acid sequence homology with *sal* and *salr*, it shares the early expression pattern of *sal*. It was suggested that an ancestral *sal/salr* gene underwent complete duplication followed by a chromosomal rearrangement, which separated the *salr* gene from its early cis-regulatory elements. These elements came into the vicinity of *sal* and were appropriated by the unrelated *sala* gene, resulting in the observed similarities between early *sal* and *sala* expression patterns (Reuter et al., 1996).

Up to now, the *Tzf* cluster is the most compact *Drosophila* zinc finger gene cluster with the distance between the adjacent members of not more than 1 kb. They appear to have the same ancestor, considering their high homology in finger region and the completely conserved intron positions. Because they are not orientated in tandem, it is unlikely that all four genes were direct products of multiple duplications of a single gene. *Tzf* is mainly expressed in testis, ovary and embryo; while *Tzf*2 is homogenously expressed everywhere (Xu & Hennig, unpublished data). From the different expression pattern, it seems unlikely that they share all their cis-regulating elements.

A splicing variant of *Tzf*?

The difference between the *Tzf* and the CG4413 transcripts can be explained in different ways:

1. A mistake in the sequencing CG4413 or cloning of a falsely spliced RNA molecule.
2. The CG4413 transcript corresponds to the real splicing pattern and the *Tzf* transcript represents an incompletely spliced molecule. The *Tzf* cDNA was cloned from a testis cDNA library, whilst the CG4413 transcript was obtained from RNA of whole flies. The two forms could be both ubiquitously expressed, or the *Tzf* transcript discovered by us is testis-specific. In this case, it would be of interest to see whether the other *Tzf* tanscripts are also subject to alternative splicing.

```
  1 MKTESNEKWVVCRVCLNNPSEGEELLHDIFSETASTRLDQMLHICAGIPV 50
    ||||||||||||||||||||||||||||||||||||||||||||||||||
  1 MKTESNEKWVVCRVCLNNPSEGEELLHDIFSETASTRLDQMLHICAGIPV 50
            .         .         .         .         .
 51 SLDDNFPDKMCSKCVRCLRLCYKFRLTCQRSHQHIMDMLDREASNANAAG 100
    ||||||||||||||||||||||||||||||||||||||||||||||||||
 51 SLDDNFPDKMCSKCVRCLRLCYKFRLTCQRSHQHIMDMLDREASNANAAG 100
            .         .         .         .         .
101 EGDLLSIAEDLSVESVLKSWEDYASQLDGGMKVEGEEDQQHQVITYVVED 150
    ||||||||||||||||||||||||||||||||||||||||||||||||||
101 EGDLLSIAEDLSVESVLKSWEDYASQLDGGMKVEGEEDQQHQVITYVVED 150
            .         .         .         .         .
151 GDTDDTNMFDVHDPTQPVPNEIEEAETYAEYEEYELLTNENSPEIAQEKG 200
    ||||||||||||||||||||||||||||||||||||||||||||||||||
151 GDTDDTNMFDVHDPTQPVPNEIEEAETYAEYEEYELLTNENSPEIAQEKG 200
            .         .         .         .         .
201 STGTDVATEEPPEEEIAEDILDSDEDYDPTHAKPEKCDRSGRKPVAYHKN 250
    ||||||||||||||||||||||||||||||||||||||||||||||||||
201 STGTDVATEEPPEEEIAEDILDSDEDYDPTHAKPEKCDRSGRKPVAYHKN 250
            .         .         .         .         .
251 SPKVETFKKKVGRKPRNKLSTYICDVCGNIYPTQARLTEHMKFHSGVKPH 300
    ||||||||||||||||||||||||||||||||||||||||||||||||||
251 SPKVETFKKKVGRKPRNKLSTYICDVCGNIYPTQARLTEHMKFHSGVKPH 300
            .         .         .         .         .
301 ECEICGRGFVQNQQLVRHMNTHTGNRPYKCNYCPAAFADRSTKTKHHS.. 348
    ||||||||||||||||||||||||||||||||||||||||||||||||
301 ECEICGRGFVQNQQLVRHMNTHTGNRPYKCNYCPAAFADRSTKTKHHRIH 350
            .         .         .         .         .
349 .................................CDLCGKGFVKAYKLR 363
                                     ||||||||||||||||
351 TKERPYVCDVCSRTFTYSDNLKFHKMIHTGEKPHVCDLCGKGFVKAYKLR 400
            .         .         .
364 LHRETHNRRITWRNDAEESTKAEDVKGETPEFLNELPKE 402
    |||||||||||||||||||||||||||||||||||||||
401 LHRETHNRRITWRNDAEESTKAEDVKGETPEFLNELPKE 439
```

Fig 14: Comparison of the putative protein products of *Tzf* cDNA (the lower line) and CG4413 cDNA (the upper line). The comparison was made using the function GAP of the HUSAR DNA analyse package. The zinc fingers are highlighted in red.

Alternative splicing in zinc finger genes was recently reported for human ZNF41 (Rosati et al., 1999) and mouse KRC (Mak et al., 1998) genes. ZNF41 gene encodes at the C-terminal zinc fingers and at N-terminal a KRAB/FPB domain. Exon skipping at the N-terminus leads to selective usage of two different KRAB/FBP modules, encoding peptides differing in C-terminus and expressed in different tissues. Zinc fingers themselves were in ZNF41 genes constantly expressed. In contrast, the product of KRC gene contains at both N-terminal and C-terminal DNA-binding domains, of which the zinc fingers are essential components. Multiple differentially spliced transcripts were identified in brain and thymus, skipping different zinc finger domains.

Do *TZF* proteins correspond to SCP proteins of the rat?

The cDNA clones encoding the *TZF* protein were obtained by the immunoscreening of the testis cDNA expression libraries with an antiserum against the rat SC protein, SCP3 (Lammers et al., 1994). SCP3 has been located on the lateral elements of the rat SC from zygotene- up to late diplotene-phase. The amino acid sequence deduced from SCP3 does not contain any known nucleic acid binding motif.

It was demonstrated that the N-terminal part of *TZF* protein (DmP3 peptide, Sun, 1994) does react with the antiserum against SCP3. DmP3 contains the first 80 amino acids and is highly divergent compared to the other *TZF* proteins. The amino acid sequence of this peptide does not have any detectable similarity with SCP3, which, however, could not completely exclude the possibility that *TZF* is the *Drosophila* “homologue“ of SCP3 in the sense that it takes over the same responsibility as SCP3 in rat. However, it is more likely that the immuncrossreaction was just a coincidence caused by the local structural similarity between rat SCP3 and the *Drosophila TZF*.

Summary

A zinc finger gene *Tzf1* was cloned by screening a λ-DASH2 cDNA expression library with an anti-Rat SC antibody. A λ-DASH2 genomic DNA library and cosmid lawrist 4 genomic DNA library were screened with the cDNA fragment of *Tzf1* to determine the genomic organization of *Tzf1*. Another putative zinc finger gene *Tzf2* was found about 700 bp upstream of *Tzf1*.

RACE experiment was carried out for both genes to establish the whole length cDNA. The cDNA sequences of *Tzf* and *Tzf2* were used to search the Flybase (Version Nov, 2000). They correspond to two genes found in the Flybase, CG4413 and CG4936. The CG4413 transcript seems to be a splicing variant of *Tzf* transcripts. Another two zinc finger genes *Tzf3* and *Tzf4* were discovered *in silico*. They are located 300 bp away from *Tzf* and *Tzf2*, and a non-tandem cluster was formed by the four genes. All four genes encode proteins with a very similar modular structure, since they all have five C2H2 type zinc fingers at their c-terminal ends. This is the most compact zinc finger protein gene cluster found in *Drosophila melanogaster.*

Introduction

Synaptonemal complexes (SCs) are structures found between paired homologous chromosomes in the meiotic prophase nuclei (Moses, 1968). They consist of two compact proteinaceous axes, the lateral elements, each one associating with one of the homologous chromosomes. The lateral elements are connected by thin transversal filaments. Between the two lateral elements another longitudinal structure is formed, the central element, which crosses the transversal filaments. The lateral elements and the central element make up the tripartite structure of the SCs. In most eukaryotes analysed so far, the condensation, pairing, recombination and segregation of chromosomes is accompanied by the assembly and disassembly of the SCs. Several rat SC proteins have been identified and the cDNA clones for some of them have been analysed (Heyting et al., 1985, 1987, 1988, 1989; Lammers et al. 1994, Smith and Benavente 1992). *Drosophila* is an exception because this structure is only formed during the meiotic prophase in females but not in males (Rasmussen 1973, 1974). At the same time, recombination does not take place in *Drosophila* males. On the other hand, there exist specialized structures in *Drosophila* primary spermatocytes, the lampbrush loops, which are formed by the actively transcribed fertility genes located on the Y chromosome.

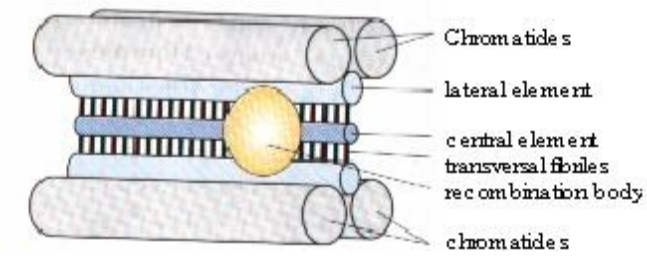


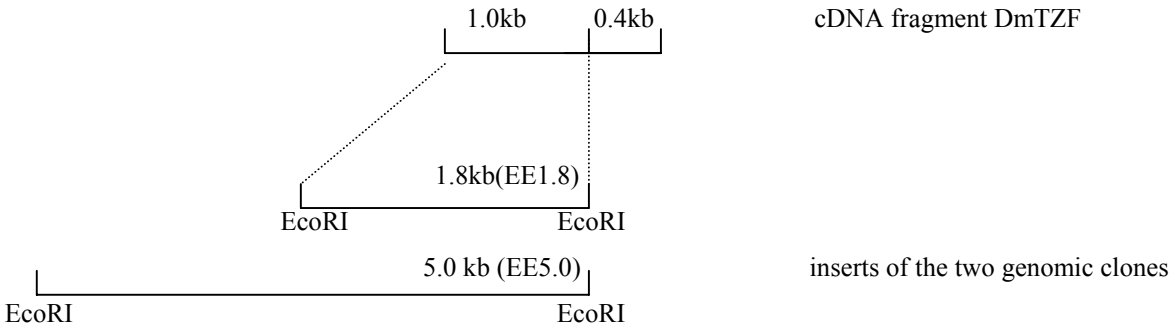
Fig 1: Schema of a synaptonemal complex (SC) in pachytene phase.

It was observed that antibodies against rat SC proteins cross react with Y chromosome lampbrush loops present in *Drosophila melanogaster* (Hennig & Heyting, unpublished data). An immunoscreening of *Drosophila* testis cDNA libraries with an antiserum against a rat SC protein (Lammers et al., 1994) was carried out to identify the structural component of this cross immunoreaction (Sun, 1994). A part of a novel gene was cloned which encodes a putative zinc finger protein. It was named *Tzf* gene. The *Tzf* gene is a single copy gene with particularly strong expression in embryos, testes and ovaries (Sun, 1994). It was the starting point of this work to complete the cDNA sequence of this gene and to find out its genomic organization.

Results

Genomic fragment containing *Tzf* gene

The 1.4 kb cDNA fragment of the *Tzf* gene, DmTZF (Sun, 1994), was used as a probe to screen a *Lambda* - DASH2 genomic DNA library. Two clones, with respectively a 5.0 kb and a 1.8 kb EcoRI-EcoRI insert were identified (Xiao, Xu & Hennig, unpublished data). The inserts were further subcloned, and the subclones were sequenced. The entire sequences of the EE5.0 and EE1.8 fragments were reconstructed from the subcloned sequences. The 1.0 kb fragment at the 5'part of the cDNA fragment DmTZF was contained in the sequence of genomic fragment EE1.8. The 0.4 kb 3' fragment was still missing.



To look for the genomic segment containing the 3'end of *Tzf* gene, a lawrist 4 cosmid library of *D. melanogaster* genomic DNA (kindly provided by Dr. Hoheisel, DKFZ) was screened with the cDNA fragment DmTZF. Three cosmid clones 89B4, 89B5 and 46A gave strong signals. 89B4 and 89B5 proved to be identical in their digestion patterns with several different enzymes. Only 89B4 was analysed in detail. The hybridisations of EcoRI and SalI digests of 89B4 DNA with the missing 0.4 kb cDNA fragment have shown that this fragment was contained in this genomic clone. It reacts with a 2.2 kb EcoRI fragment and a 9.0 kb SalI fragment. Both fragments were cloned into pBluescript and sequenced. A subclone of the 2.2kb EcoRI fragment, called eb2, contained the 0.4 kb cDNA fragment of

Tzf. Another subclone from the 9.0 kb SalI fragment, ab1, overlapped with the fragment EE1.8, the fragment EE5.0 from the λ DASH-2 library and the 2.2 kb EE fragment from the lawrist 4 cosmid library. This allowed deducing the relative localization of the three fragments (see Fig 2). Their organization was confirmed by sequencing the PCR products obtained with primers CX1 and DM5.

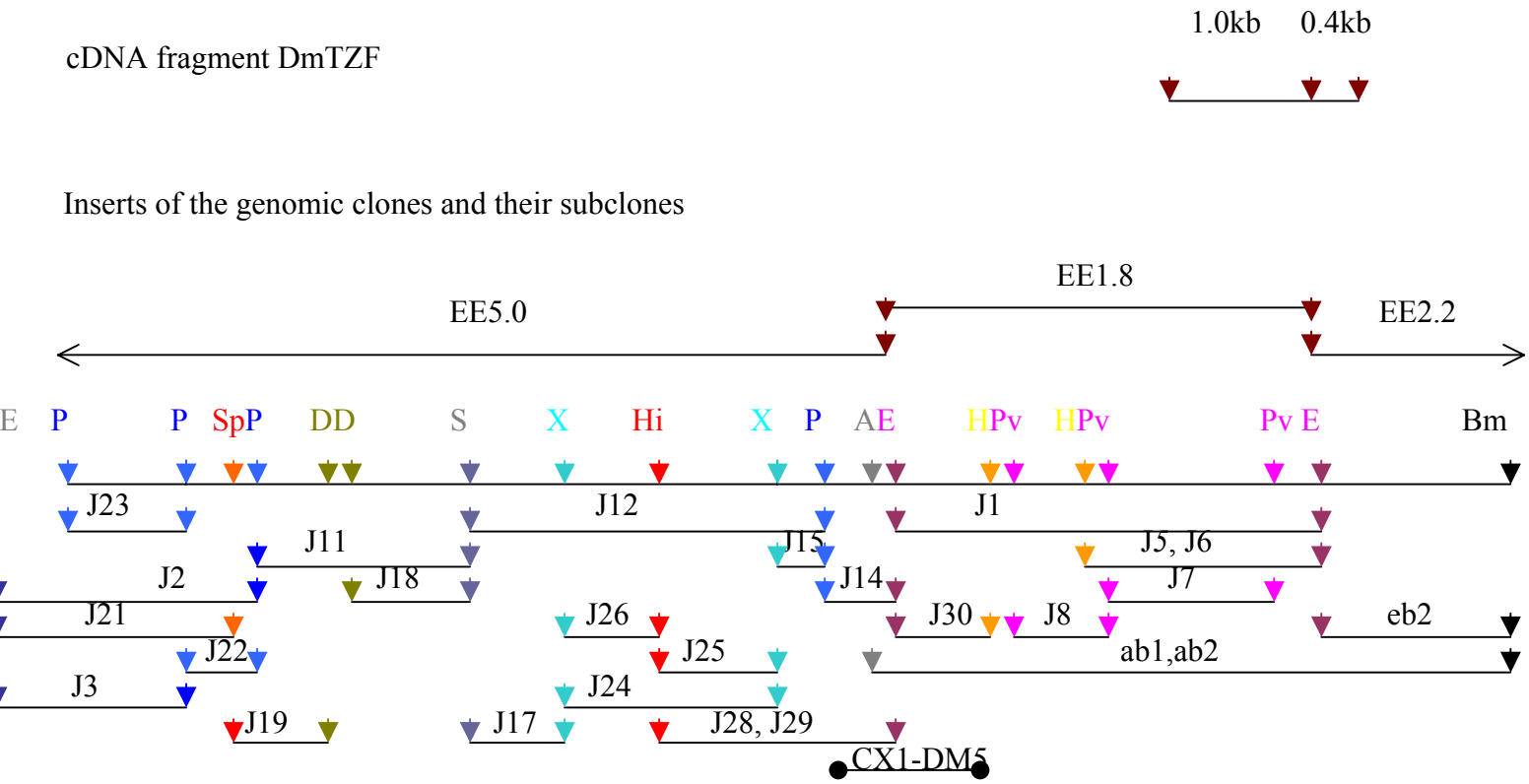


Fig 2: The organization of the subclones from the genomic fragments containing the *Tzf* and *Tzf*2 genes. The names of the restriction enzymes are abbreviated according to the abbreviation list in Appendix I.

Cloning of *Tzf*2 gene

Sequencing of EE1.8 and EE5.0 has proved that only EE1.8 contains the *Tzf* sequence. But EE5.0 also hybridised with the probe DmTZF which suggested a sequence similarity between EE5.0 and DmTZF. Analysis of the sequence of EE5.0 showed that parts of the sequence could also code for segments of a

zinc-finger protein. To see if these sequences are indeed expressed, RT-PCR experiments were performed with the primer pair GW1 and GW2, designed according to the EE5.0 sequence. Two different total RNA preparations were used as template. They were obtained from *D.melanogaster* ovaries and embryos.

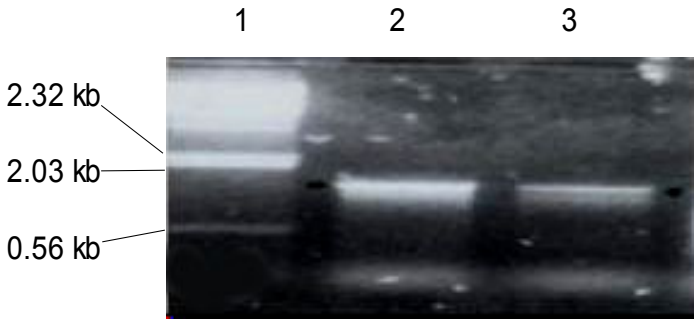


Fig 3: RT-PCR product amplified with primer GW1 and GW2

Lane1: DNA marker λ-Hind III Lane2, 3: RT-PCR product with embryo and ovary RNA

A major signal could be seen as product of the RT-PCR reactions with both embryo and ovary RNA as template. The DNA was recovered from the gel and digested with BglII and XhoI, which were designed into the 5'ends of GW1 and GW2 as integrated restriction sites. The digested fragment was ligated into pBluescript, digested with BamHI (compatible end with BglII) and XhoI. Two different clones, CX1 and CX9, were obtained, of which CX1 contains the whole RT-PCR product, and CX9 only part of it because of the existence of an internal BglII site in this fragment (see Fig 4). Different subclones were made from these two clones and the sequence of the whole fragment was obtained. The alignment of this sequence and the sequence of the genomic clone EE5.0 showed that the entire cDNA fragment GW1-GW2 is indeed contained in this segment of the genomic sequence. The alignment also showed the positions of the introns on the genomic sequence, which excluded the possibility that this product was an artefact caused by genomic DNA contamination in the RNA.

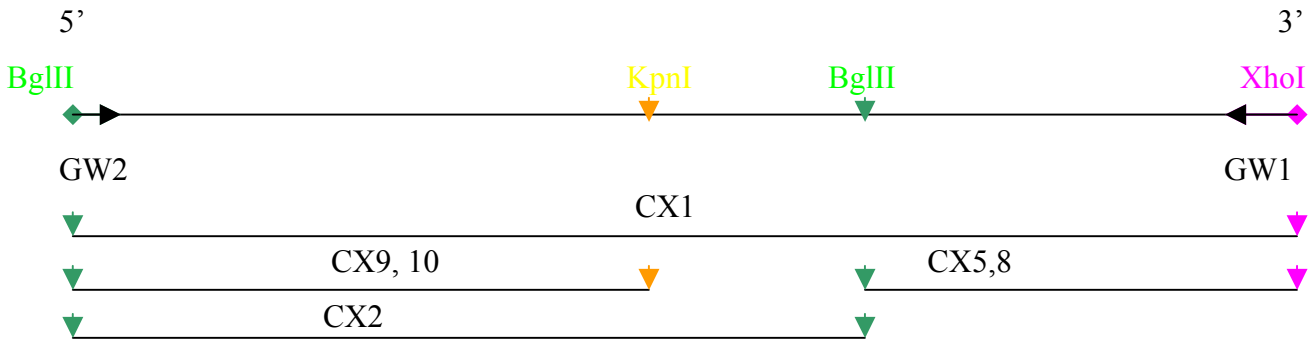
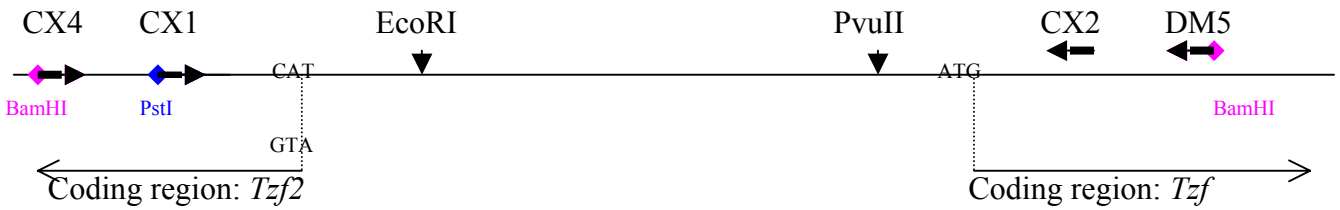


Fig 4: Subclones from the cDNA clone of *Tzf2*

An in-frame stop codon was found at the 3' end of this cDNA fragment. The 5'end of the cDNA was obtained through RACE reaction (see below). The start codon could be verified through an in-frame stop codon 50 bp upstream of it. This gene was named *Tzf2*.

These experiments have revealed two similar genes, *Tzf* and *Tzf2*. They exist in opposite orientation. Their start codons are only separated by 729 bp.

RACE Reaction



To obtain the 5' ends of the two genes, *Tzf* and *Tzf2*, RACE experiments were carried out for both cDNAs.

The *Tzf2* gene

8µg total RNA of *Drosophila melanogaster* embryo was reverse transcribed with primer CX4 to get a gene - specific cDNA pool for the *Tzf2* gene. 1/20 of the cDNA pool was tailed with poly-dC at the 3' end. The PCR reaction was then carried out with the primer CX1 and an adaptor. CX1 is an internal gene-specific primer downstream to the primer CX4, and the adapter anneals to the poly-dC end. The PCR product was too weak to be seen on an agarose gel.

The southern blot of an aliquot of the PCR product was therefore hybridised to the PCR product of primer CX1 and DM5, labelled through hot PCR. This probe recognizes the intergenic region between both *Tzf* genes. Another aliquot of the CX1-adapter PCR product was loaded on another agarose gel and the region corresponding to the signal in the hybridisation was cut out using the low-melting agarose gel method.

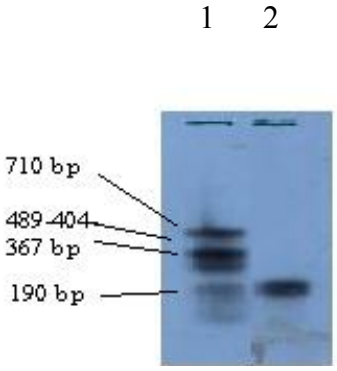


Fig 5: Hybridisation result of the first PCR blot with the intergenic probe CX1-DM5. The blot was once more hybridised with labelled pBluescript, so that the marker could be seen on the blot. The second hybridization did not change the signal position in the probe lane.
Lane 1: DNA marker: HpaII digestion of pBluescript.
LaneII: CX1-adapter PCR product.

1/20 of the LM agarose gel mixture was directly used as template for a reamplification with the same primer pair. CX1 and adapter now gave a PCR product visible through EtBr staining. It was cloned into pBluescript after digestion with PstI and HindIII, the restriction sites integrated respectively in both primers. Fifty clones were picked and their inserts checked by hybridising with the labelled intergenic probe CX1-DM5. Four clones gave positive signals and were thus sequenced. Among them, C1, C4 and C7 start from the same G_{3327} (Fig 6). This makes it likely that this G is the transcriptional initiation site of the *Tzf2* Gene.

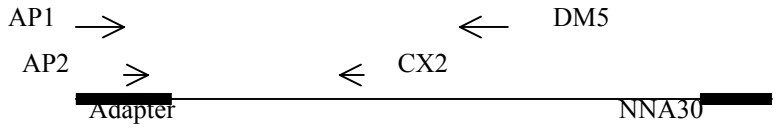
```
3201  CCGAGTCTCT CATTGTCGCG AATAGGAGAT GGATCTCTTC CAGGAGGCTA
TZF2 gene ←
3251  CTTACCCTCC GCTATAATAT TTATAATAAC ACGCTCGCTG GACGAAGACG
3301  ACGCCAAAAA AATAGATTGT AGTCTGGAGC GAAGTGCACG CGTTGCAATG
3351  CGAATGTCGC CACACAGGGT GACCACAGTC GACGCTAAGG GTATTCCATT
```

Fig 6: Transcription initiation of *Tzf2* gene. The nucleotides are numbered according to Appendix I.

The *Tzf* gene

The same strategy as for *Tzf2* was also performed for the *Tzf* gene, with DM5 as the primer to establish a cDNA pool. Primer CX2 and the same adapter as for the *Tzf* gene were used to amplify from this pool. They failed to give reasonable product even after several reamplifications. The Marathon RACE Kit from CLONTECH was then tried to reveal the transcription start of the *Tzf* gene in a different way.

Poly A^+ mRNA was obtained from total RNA made from *D.melanogaster* embryos with Oligotex™ Kit from Qiagen. First strand cDNA was obtained from 1 µg mRNA, using the AMV Reverse Transcriptase and a modified oligo (dT) primer provided in the Marathon RACE kit with two degenerate nucleotides positioned at the 3'end. A second-strand Enzyme Cocktail containing RNase H, *E.coli* DNA polymerase I and *E.coli* DNA ligase was then added to the first strand mixture to yield a double-strand cDNA pool, which was subsequently made blunt ended with T4 DNA polymerase. The Marathon cDNA adapter was ligated to the blunt ends of the double-stranded cDNA molecules. The first PCR was performed with the *Tzf*-specific primer DM5 and adapter-specific primer AP1. A nested PCR was performed using an aliquot of the first PCR product, with the primer CX2 down stream from DM5 and the primer AP2 downstream from AP1. The product was cloned into the PCR 2.1 vector (TA cloning kit, Invitrogen). Sequencing results have shown that eight clones contained the upstream sequence of *Tzf*, five of which stopped at the same nucleotide G^{3864}, which was identified as the transcription start of *Tzf*.



Tzf cDNA with adapter at both ends

```
3851  TTTCGCTGTG AATGGCCAGT GGCCGTGGAC GCTTAAATTC AGCTGTTTGA
3901  ATCTTCTGCA ACAACCGAAG TGATTGTTTT ATCGAATTAG CCATGAAAAC
                                                  Tzf gene
3951  TGAGTCCAAC GAGAAGTGGG TGGTGTGCCG CGTTTGCCTG AACAATCCCA
```

Fig 7 : Transcription initiation of the *Tzf* gene. The nucleotides are numbered according to Appendix I.

The structure of the *Tzf* and *Tzf2* genes

The cDNA sequences of *Tzf* and *Tzf2* genes were compared with the genomic sequence, and the genomic organization of both genes could be deduced (see Fig 8).

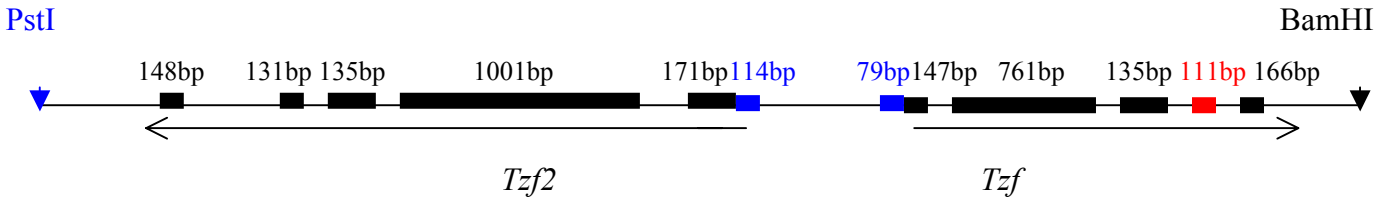


Fig 8: The gene structure of *Tzf* and *Tzf2*. The long thin line stands for the genomic sequence. Thick bars stand for the exons. Blue bars for the 5' UTR and black ones for coding regions. The red block in the *Tzf* gene stands for the exon that is possibly subject to alternative splicing (see below).

Tzf and *Tzf2* genes and the *Drosophila* genome project

After the characterizing of *Tzf* (Sun, 1994) and *Tzf2*, the complete *Drosophila melanogaster* genome was sequenced and published in the Flybase. The genomic and cDNA sequences of *Tzf* and *Tzf2* was used to search the Flybase and to compare our data and the genome sequencing data. The genomic sequence of *Tzf-Tzf2* region (listed in Appendix I) was identical with the celera sequence gb / AE003728.1 (*Drosophila* genomic scaffold 142000013386035 Section 53/105). Two genes identified for this genomic clone, Gene CG4413 and Gene CG4936 correspond to our *Tzf* and *Tzf2,* respectively.

GC4936 is identical with *Tzf2*. Comparison of the *Tzf* cDNA sequence and the cDNA sequence of the gene CG4413 recorded in the Flybase showed that the CG4413 transcript lacked the complete 111bp exon (see Fig 8 & Fig 9), which encodes the C-terminal end of the third zinc finger and the complete fourth finger.

951 ⟶ *Tzf* gene transcription
GCACATGAACACTCACACGGGGAACCGACCATACAAGTGCAACTACTGTCCAGCTGCCTTCGCCGATCGATCCACGAAAACCA
AACATCATAGgtgagaaatattcgaaacagttgaaagtaaattaaactgacattatttctctccag*AATTCACACTAAGGAGC*
3rd exon
GTCCCTACGTGTGCGACGTTTGCTCCAGAACGTTTACCTACTCGGACAACCTGAAGTTCCACAAGATGATTCACACGGGGGAG
4th exon
*AAGCCGCATGT*gtaagcatcaacatattttactacttcattatcctgcaataatagctattctctttcagCTGTGATCTTTGT
5th exon
GGCAAAGGATTTGTGAAGGCCTACAAATTGCGTTTGCATCGGGAAACGCATAATAGACGTATCACCTGGAGAAATGACGCAGA
1250

Fig 9: Genomic fragment containing part of the 3rd, 5th and the complete 4th exon (italic letters) of *Tzf* gene. Lower case letters are from the introns. The first gt and last ag of the introns are underlined. *Tzf* cDNA contains all three exons, whilst in CG4413 transcript, the third exon and the fifth exon are directly ligated together, skipping the fourth exon.

The *Tzf3* and *Tzf4* genes: two additional genes of a gene family

Several hundred base pairs away from the *Tzf* and *Tzf2* genes, two other zinc finger genes were found in the database, which also encode zinc finger proteins similar to the *Tzf* and *Tzf2* proteins. The protein products of all four genes have the same modular structure, and they are located next to each other.

Names of *Tzf* Genes	Names in the Flybase
Tzf	CG4413
Tzf2	CG4936
Tzf3	CG4854
Tzf4	CG4424

Alignment and Comparison of *Tzf, Tzf2, Tzf3* and *Tzf4*:

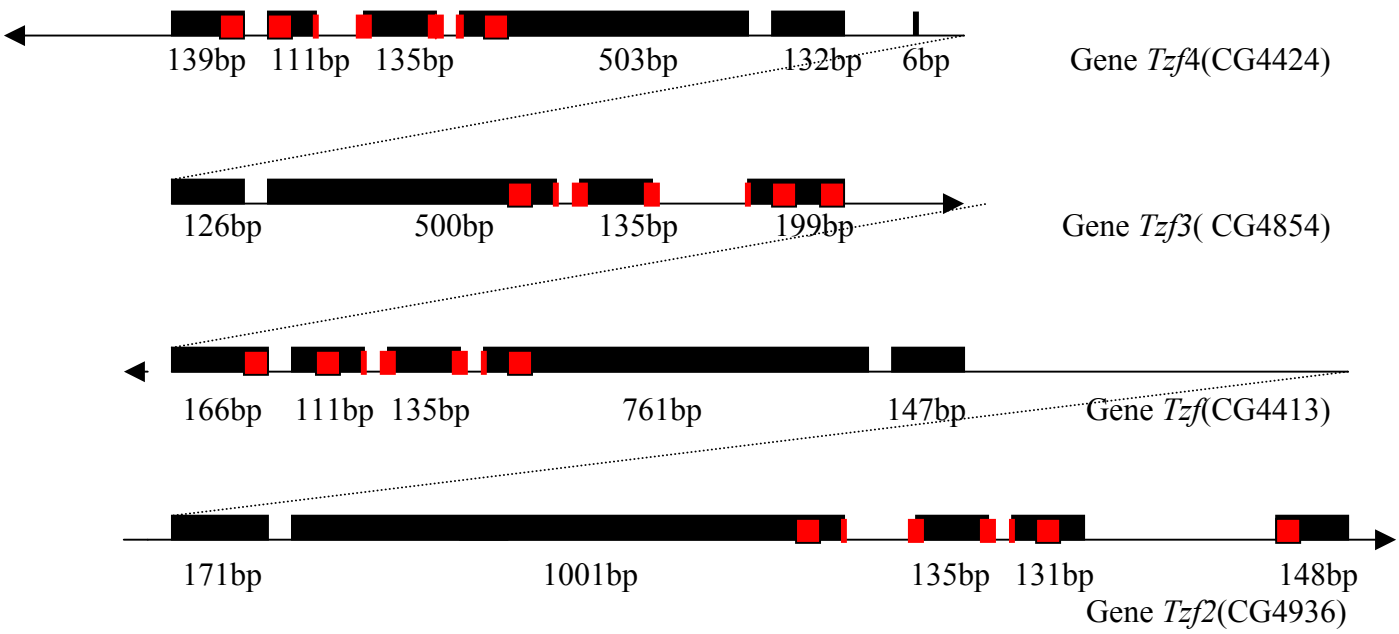


Fig 10: Genomic organization of the four *Tzf* genes. The thick bars represent the coding exons, the thin lines the introns. Only the exon sizes are presented, the introns are all shown in proportion to their length. The red boxes represent DNA segments coding for the zinc finger motifs. They are sometimes separated in two adjacent exons. The arrowheads show the direction of transcription, and the dashed lines connect the interrupted genomic sequence.

The *Tzf4* and *Tzf3* genes are localized in a head-to-head orientation, which corresponds to the orientation of *Tzf* and *Tzf2*. The arrangement of the five zinc fingers is conserved in all four genes.

```
      1                                                  50
TZF2  MRDSAAHASP AAAATSTQKW IVCRVCLQQP ...KEPMASI FNDDSE....
TZF3  ~~~~~~~~~~ ~MHTNVDSRD LKCRICLVQP ....KDESLM PTEP......
TZF4  ~~~~~~~~~~ ~~~~~~MAMM TLCRTCLQDG ...EAHMVSI FQTADDRLPG
TZF   ~~~~~~~~~~ ~MKTESNEKW VVCRVCLNNP SEGEELLHDI FSETAS....

      51                                                100
TZF2  .KDLTHMIRE CGGVPIKQ.. .FDHYPDKIC EKCFKVLKMA FKFRETCQRS
TZF3  ..DFPDKIKR CTGVELSE.. .SPDWPNRIC TSCALLLRAA LKLRSLCQQT
TZF4  GVSLCDKIES LSGIQIRATA KEEVLPTRIC LRCKAFLTLA HKFRQICQRS
TZF   .TRLDQMLHI CAGIPVSL.. .DDNFPDKMC SKCVRCLRLC YKFRLTCQRS

      101                                               150
TZF2  YGHLRQFVGP VEVEQRPPEK KGSETATKLE PDVDPDEAEQ EPEHDEEDED
TZF3  EKDLKEQK.. .......... .........L QEINIEIVHD E...QETKKK
TZF4  NEFLREYVIK .......... ....D..AVE QGVVKEVVQQ TR..PSTPPP
TZF   HQHIMDML.. .DREASNANA AGEGDLLSIA EDLSVESVLK S....WEDYA
```

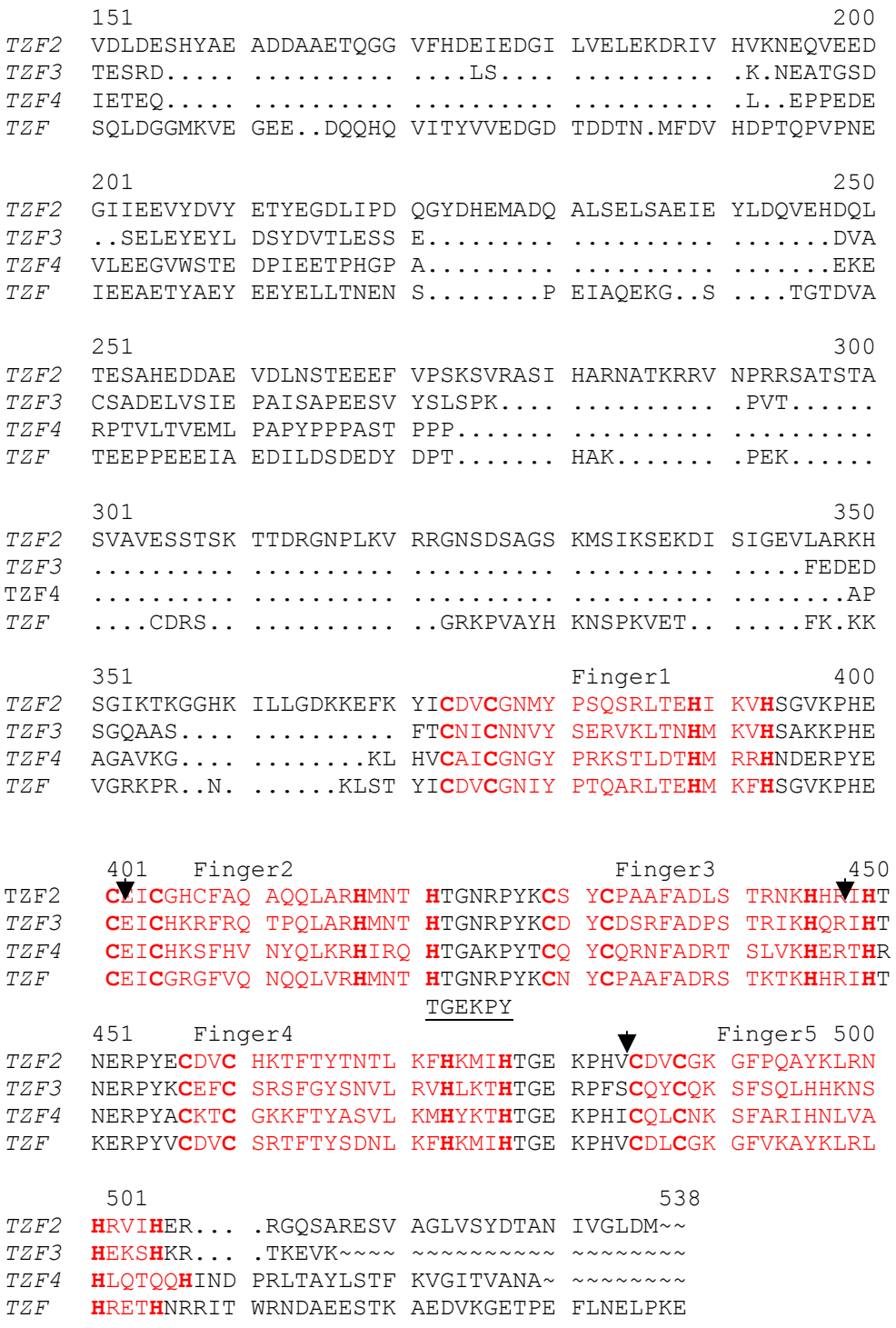


Fig 11: Amino acid sequence alignment of the four *TZF* proteins. The zinc fingers are highlighted in red. The sections with similarity to zinc fingers at the N-terminal ends of the proteins are highlighted in blue. Intron positions are shown with arrowheads.

All four gene products have at the C-terminal end a cluster of five C2H2-type zinc fingers with the consensus sequence Φ-X-C-$X_{2,4,5}$-C-X_3-Φ-X_5-Φ-X_2-H-$X_{3,4}$-H, where X represents any amino acid, Φ a hydrophobic residue. At the N-terminal end, some short sequences remind of zinc fingers but are

incomplete. This is typical for proteins including many copies of zinc finger domains, where incomplete or degenerate copies of finger domain are often found (Rosenfeld & Margalit, 1993). The inter-finger regions between fingers 2-3, 3-4 and 4-5 agree approximately with the TGEKPY consensus sequence named H-C link (Schuh et al., 1986), whilst the inter-1-2-region seems to be more degenerated than other inter-finger regions. MOTIF and PROSITE function from the HUSAR DNA analysis package were used to analyse all four sequences for further details. No other motifs were found, except for a cytochrome C domain in *TZF2*, 3 and 4 (C_{401}-K_{406} of *TZF3* and *TZF*4; C_{457}-K_{462} of *TZF2*). But this is probably of no relevance, because they all have one zinc finger domain with a H residue after the second C, which happens to be part of a cytochrome C consensus sequence (Mathews, 1985). PSORT function from the same package showed that all four *TZF* products are likely to be nuclear proteins.

Discussion

The four *Tzf* genes are paralogues

As shown in Fig 10, *Tzf* and *Tzf2* have the same gene organization, with the exon 2 containing the finger 1 and part of finger 2, exon 3 containing part of finger 2 and 3, exon 4 part of finger 3 and finger 4, and exon 5 the finger 5. *Tzf3* and *Tzf4* also obey this rule, although with small deviations. *Tzf4* has an additional small exon of 6 bp (named exon 0 to maintain the exon terminology of all *Tzf* genes) at the N-terminal of exon 1. In *Tzf3*, exon 4 and exon 5 are merged into one exon containing finger 4, finger 5 and part of finger 3. With these two exceptions, the all intron positions are completely conserved among the four genes. As shown in Fig 11 they are located in the following positions:
Intron between exon 1 and 2: between P(E,Q)65 and I(L,V)66,
Intron between exon 2 and 3: in E2 of the finger 2, between the 2nd and the 3rd position of the codon,
Intron between exon 3 and 4: in R(L,K)14 of finger 3, between the 2nd and the 3rd position of the codon,
Intron between exon 4 and 5: in V(S,I) 2 of finger 5, between the 2nd and the 3rd position of the codon.

Therefore, the four *Tzf* genes have an almost identical genomic organization. Also, their products share the same modular structure with five canonical zinc fingers at the C-terminal end and possible remnants of two fingers at the N-terminal end. Sun (Sun, 1994) divided the *TZF* protein into five domains according to their isoelectric values (pI values) and their homology with the corresponding segments of the *Drosophila hydei* homologue of *Tzf* (Dh*Tzf*) (Sun, 1994). This domain system was extended to all four *TZF* proteins. The following two tables show the pI value and size of the five domains, as well as their similarity and identity among the four *TZF* proteins.

	Domain I (pI / size)	Domain II (pI / size)	Domain III (pI / size)	Domain IV (pI / size)	Domain V (pI / size)
TZF	7.22 / 85	3.38 / 147	11.26 / 41	9.47 / 133	4.51 / 33
TZF2	8.22 / 93	3.74 / 178	11.59 / 90	8.63 / 133	4.62 / 27
TZF3	6.95 / 78	3.97 / 85	3.33 / 16	10.16 / 133	11.07 / 7
TZF4	7.95 / 85	3.99 / 87	10.80 / 12	10.26 / 135	9.55 / 22

	Domain I (identity similarity)		Domain II		Domain III		Domain IV		Domain V	
TZF vs TZF2	**59.756**	**42.683**	37.931	20.000	37.500	25.000	**81.203**	**79.699**	44.444	22.222
TZF2 vs TZF3	46.154	32.051	37.037	25.926	25.000	25.000	62.406	57.143	42.857	14.286
TZF3 vs TZF4	38.158	27.623	29.268	21.951	14.286	14.286	56.391	48.120	28.571	14.286
TZF vs TZF4	51.316	35.526	**50.000**	**33.333**	25.000	25.000	57.143	48.872	83.333	33.333
TZF vs TZF3	38.462	29.487	35.484	22.581	33.333	26.667	62.406	55.639	57.143	28.571
TZF2 vs TZF4	46.753	35.065	36.047	27.907	33.333	25.000	52.632	47.368	26.316	21.053

Domain IV, which contains the five zinc fingers, is shared with the highest similarity among the *TZF* proteins. Domain I with the N-terminal finger remnants also has a considerable inter-*TZF*s similarity. For both domains, the proteins *TZF* and *TZF2* are most similar among the four *TZF*s. Domain II is the largest domain for *TZF* and *TZF2*, but it is much smaller for the other two proteins. In spite of the length heterogeneity, domain II of all four proteins is very acidic with the pI value between 3.38 and 3.99. The highest amino acid sequence similarity is between *TZF* and *TZF4*. The a.a. sequences of domain III and domain V are considerably diverged, as is their lengths.

Unlike most gene clusters generated through unequal crossover, the *Tzf* genes do not have a tandem orientation. Small chromosomal inversions could account for their opposite orientations. The simplest scenario with the least number of chromosomal rearrangements would be as following:
Tzf and *Tzf2* were generated from the ancestral zinc finger gene. A small chromosomal inversion of one of both copies caused their opposite orientations. Subsequently *Tzf* and *Tzf2* were again duplicated by an unequal crossover, creating *Tzf3* and *Tzf*4. Alternatively, *Tzf* and *Tzf2* could be a duplication of *Tzf3* and *Tzf*4.

The pattern of silent (synonymous) nucleotide changes within the zinc finger coding region provides a measure of evolutionary distances between the paralogues (Li&Graur, 1991). The amino acids conserved among all four *TZF* proteins were sorted out and their codons were compared in the following tables:

	Finger1							Fingerlink 1-2	
	C_{373}	C_{376}	N_{378}	Y_{380}	L_{386}	H_{389}	H_{393}	P_{398}	E_{400}
TZF	tgc	tgc	aat	tat	ctc	cac	cat	cca	gag
TZF2	tgc	tgc	aac	tat	ctt	cac	cac	ccg	gag
TZF3	tgc	tgc	aat	tac	ttg	cac	cac	cca	gaa
TZF4	tgc	tgt	aat	tat	ctg	cac	cat	cct	gag

	Finger2									Fingerlink2-3			
	C_{401}	E_{402}	I_{403}	C_{404}	Q_{413}	L_{414}	R_{416}	H_{417}	H_{421}	T_{422}	G_{423}	P_{426}	Y_{427}
TZF	tgc	gag	atc	tgc	cag	ctg	cgg	cac	cac	acg	ggg	cca	tac
TZF2	tgc	gag	atc	tgt	cag	ctg	cgc	cac	cac	acc	gga	ccg	tac
TZF3	tgc	gaa	atc	tgt	cag	ttg	agg	cac	cac	acc	ggt	ccc	tac
TZF4	tgc	gag	att	tgc	cag	ctg	cgc	cac	cac	acg	gga	cca	tat

	Finger3										Fingerlink3-4			
	C_{429}	Y_{431}	C_{432}	F_{436}	A_{437}	D_{438}	K_{444}	H_{445}	R_{447}	H_{449}	E_{452}	R_{453}	P_{454}	Y_{455}
TZF	tgc	tac	tgt	ttc	gcc	gat	aaa	cat	aga	cac	gag	cgt	ccc	tac
TZF2	tgc	tat	tgc	ttc	gcc	gac	aag	cac	aga	cac	gag	cga	ccc	tac
TZF3	tgt	tat	tgc	ttc	gcc	gat	aag	cat	agg	cac	gaa	cga	ccg	tac
TZF4	tgc	tat	tgc	ttc	gcg	gat	aag	cat	aga	cat	gag	cgt	cct	tat

	Finger4							fingerlink4-5		
	C_{457}	C_{460}	F_{464}	Y_{466}	L_{470}	H_{473}	H_{477}	T_{478}	G_{479}	E_{480}
TZF	tgc	tgc	ttt	tac	ctg	cac	cac	acg	ggg	gag
TZF2	tgc	tgc	ttc	tac	ttg	cac	cat	acg	gga	gag
TZF3	tgc	tgc	ttt	tac	ctc	cat	cat	acc	ggt	gaa
TZF4	tgc	tgc	ttc	tat	ctt	cac	cac	acg	ggc	gaa

	Finger5				
	C_{485}	C_{488}	K_{490}	F_{492}	H_{501}
TZF	tgt	tgt	aaa	ttt	cat
TZF2	tgc	tgc	aag	ttc	cac
TZF3	tgc	tgc	aag	ttc	cac
TZF4	tgc	tgc	aaa	ttc	cac

For each two different genes, the number of nucleotide substitutions was counted and listed in the following table:

	TZF/TZF2	*TZF3/TZF4*	*TZF/TZF3*	*TZF2/TZF4*	*TZF/TZF4*	*TZF2/TZF3*
Synonymous substitutions	24	28	31	24	23	24

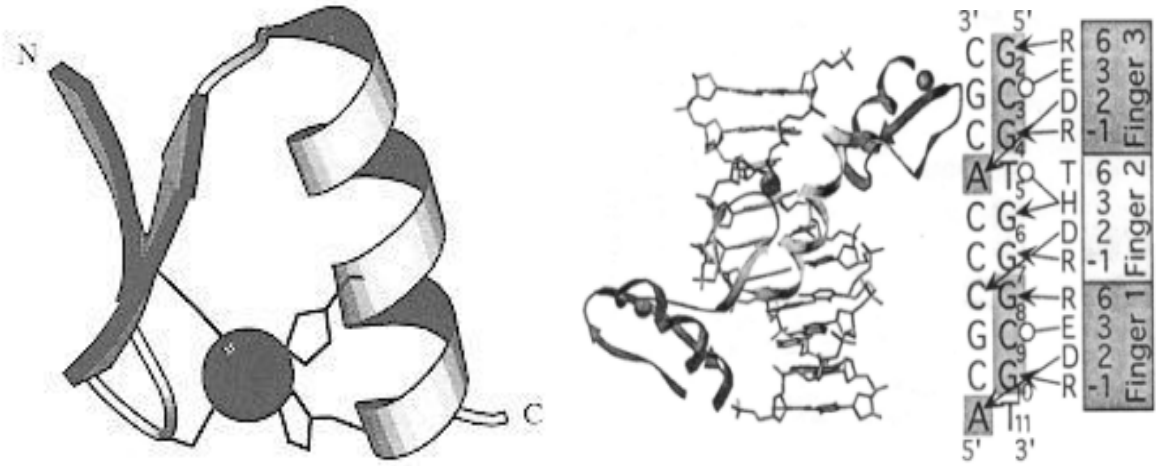
Between *TZF*/*TZF4* are the least synonymous substitutions, between *TZF2*/*TZF3* the second least. This result is not contradictory to the crossover-inversion-crossover hypothesis, although a definite proof of the hypothesis still demands further data.

Possible functions of the *TZF* proteins

Current estimates suggest that genes encoding C2H2 type zinc fingers account for 0,7%-2% of eukaryotic genomes (Hoovers et al., 1992; Shannon et al., 1998; Clarke&Berg, 1998; Böhm, unpublished data). Present evidence suggests that most of these genes encode sequence specific nucleic acid binding proteins. Although some of these proteins are known to participate in pattern formation, cellular proliferation and tumorigenesis (Wieschaus et al., 1984; Boulay et al., 1987; Schuh et al., 1986; Call et al., 1990; Supp et al., 1996), the biological function of the vast majority of zinc finger proteins is still unknown.

DNA-binding?

Most of C2H2 type zinc finger proteins are DNA-binding transcription factors. They fold in the presence of zinc to form a ββα domain. Each finger binds a single zinc ion that is sandwiched between the two-stranded antiparallel β-sheet and the α-helix. The zinc ion is tetrahedrally coordinated between two cysteines at one end of the β-sheet and the two histidines in the C-terminal portion of the α-helix (Michael et al., 1992; Shi & Berg, 1995). X-ray studies of zinc finger-DNA complexes (Pavletich & Pabo, 1991) showed that the α-helical portion of each finger fits in the major groove of DNA, and that the binding of successive fingers causes the protein to wrap around the DNA. The majority of base contacts occurs among three base pair segments along one DNA strand with four amino acids at the position -1, 2, 3, and 6 (see Fig 12).



(a) (b)

-1 12 3 4

Finger 1 PYA**C**PVESCDRRFSRSDELTR**H**IRI**H**TGQK

Finger 2 PFQ**C**RI--- CMRNFSRSDHLTT**H**IRT**H**TGEK

Finger 3 PFA**C**DI---CGRKFARSDERKR**H**TKI**H**LRQK

\+ ! + φ φφ+ φ ! !

(c)

Fig 12 **(a)**The ββα domain of zinc finger protein Zif268. **(b)** Structure of the three fingers of Zif 268 bound to DNA. Base contacts made from position –1, 2, 3, and 6 of each helix are indicated schematically to the right of the structure. Arrows indicate contact mediated by hydrogen bonds; open circles indicate hydrophobic interactions (Eltrod-Erickson et al., 1998) **(c)** The sequence of the three fingers is shown with the H and Cs participating in the zinc coordination in bold. The positions of amino acids are numbered according to convention. *Filled squares* below the sequence indicate the positions of the conserved hydrophobic residues. *Filled circles* and *stars* indicate residue positions that are involved in phosphate and base contacts (respectively) in most of the fingers.

Different zinc fingers could bind to different DNA segments. Because of the binding pattern of the zinc finger proteins to the DNA molecules, the specificity of the DNA recognition of zinc finger is mostly only dependent on the four amino acids at the position -1, 2, 3, and 6 (Pavletich & Pabo, 1991; Elrod-Erickson et al., 1998). As the data from zinc finger selection continues to grow, especially from the experiments to "design" zinc fingers to recognize certain DNA sequences, it becomes possible to predict the zinc finger specificity according to its amino acid sequence (for review see Wolfe et al., 2000; Choo & Isalan, 2000).

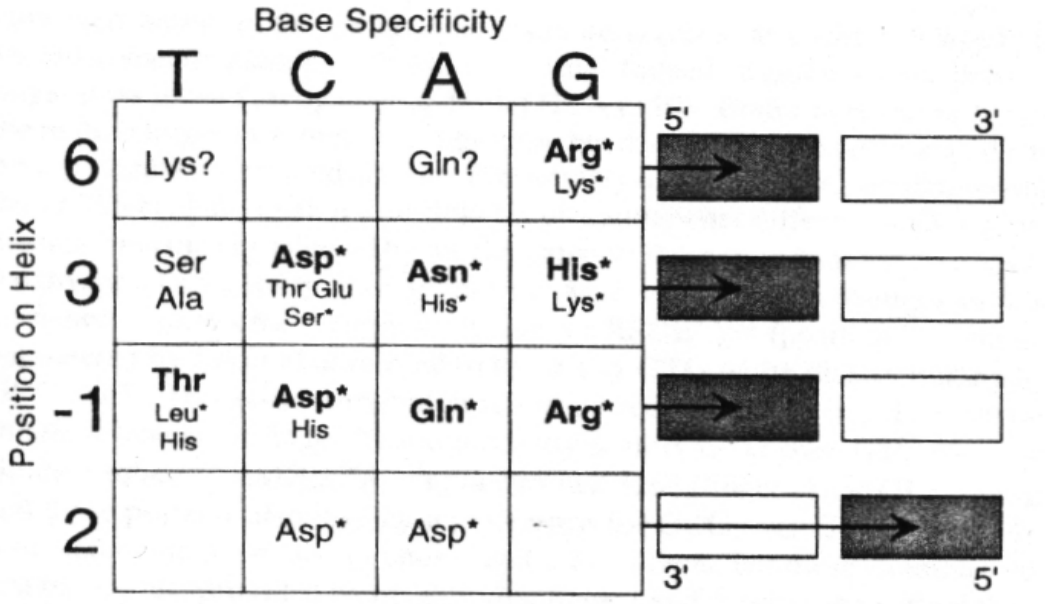


Fig 13: Pattern of side-chain base interactions that provide an approximate "recognition" code of zinc fingers that have a canonical binding mode. This chart describes contacts between residues at key positions in the α-helix (-1, 2, 3 and 6) and bases at the corresponding positions in the canonical subsite (see Fig 12b). Boldface type highlights amino acids that occur most frequently in phage display "designing" selections when a particular base specificity is desired, and an asterisk indicates contacts that have been observed in structural studies. Question marks indicate that the specificity of the respective amino acid/base contact is uncertain. Positions for which base specificity is largely undefined are left blank. Adapted from Wolfe et al., 2000.

These rules were used to predict the DNA-binding specificity of the four *TZF* proteins, giving following result for each of them:

TZF proteins	Predicted binding-sequence
TZF1	xxTGxATCCxAxxGx
TZF2	xxxGxATCCxCxxGA
TZF3	xGxGxATCCxxxxGA
TZF4	xCGGxxTTCxxxxAG

According to this prediction, the four *TZF*s could bind to very similar sequences, especially as the middle parts of their recognition sequences are almost identical (GxATCC). They could either compete with each other for the same DNA fragment, or they could bind to a larger DNA sequence consisting of

of several binding motifs in tandem. Or, an alternative possibility is that they have no function in DNA binding but that they bind to RNA or participate in protein interactions with their zinc finger domains.

RNA-binding?

Zinc fingers are potentially nucleic acid-binding proteins, which could be DNA-, RNA-, or DNA- and RNA-binding (for review see Leon & Roth, 2000). It was proposed that the conserved H-C link sequence is only present in DNA-binding proteins, but not in RNA-binding ones (Darby & Joho, 1992). This might be taken as an indication that the *Tzf* genes code for DNA-binding proteins. This prediction is, however, uncertain. The protein XFG5-1, for example, is a specific RNA-binding protein in *in vitro* experiments, but contains the conserved H-C link (Köster et al., 1991). The Wilm's tumor suppressor, WT1, is another good example, which was initially characterised as a DNA-binding protein. Now it is more likely to function in gene regulation at the RNA-level as well (Larsson et al., 1995; Bruening et al., 1996; Bordeesy & Pelletier, 1998). A number of C2H2 zinc fingers have been identified that bind RNA, but, aside from TFIIIA and p43, the biological significance of these interactions requires further study (Friesen & Darby, 1997; 1998;).

Protein binding?

Recently, zinc fingers were shown to be able to perform not only DNA-binding function, but also to participate in protein-protein interactions. Among the C2H2 fingers, zinc finger proteins Ikaros and Aiolos were identified to have both DNA-binding fingers and protein-protein interaction fingers responsible for homo- or heterodimerization (Sun et al., 1996; Georgoporos et al., 1997; Morgan et al 1997). Considering the clustered localization of the *Tzf* genes which provides a structural basis for a co-regulation of their expression, some fingers of the *TZF* proteins could very well also function in forming homopolymers from one *TZF* subunit, or heteropolymers from two or more different *TZF*s. Different combinations of different *TZF*s with the varying DNA sequence they recognize could bind different target DNA sequences. They could also control the interaction of the *TZF* proteins with other proteins.

The four *Tzf* genes are tightly clustered

Zinc finger protein clusters in mammals

A considerable body of evidence has emerged to suggest that a large fraction of human zinc finger genes (ZNF genes) are arranged in clusters (with two or more genes located in the same interval) scattered throughout many chromosomes, including 3,7,8,10,11,12,16,17,19,20,21,22, and X. (Aubry et al., 1992; Hoovers et al., 1992; Huebner et al., 1991; Rousseau-Merck et al., 1993; Lichter et al., 1992; Saleh et al., 1992; Tommerup &Vissing, 1995). Especially the ZNF91 and ZNF45 gene families, located at q13.2 and p12-p13.1 of human chromosome 19 (H19), were intensively studied and provided informations on evolutionary aspects of zinc finger genes (Constantihou-Deltas et al., 1992; Shannon et al., 1996; Bellefroid et al., 1993). In both cases, the zinc finger genes are arranged in "head-to-tail" tandem arrays with relatively even spacing between neighbouring genes. This observation was consistent with the idea that familial gene clusters have arisen primarily through multiple, *in situ* tandem duplication events of single progenitor loci (Ohno, 1970). Members of a tandem gene family are more similar in sequence to one another than they are to ZNF genes located elsewhere in the chromosome. Less is known about the their genomic organization of other ZNF gene clusters, but there is evidence that they are also mostly arranged as "head-to-tail" tandem arrays (Shannon et al., 1998; Calabro et al., 1995; Aubry et al., 1992; Derry et al., 1995).

Comparative studies of region H19 in human and the homologous mouse chromosome region have provided evidence that many, and perhaps all, of the clustered ZNF families located in H19 are represented by similar clusters located in syntenic regions of mouse chromosomes 7, 8, 9, 10, and 17 (Shannon et al., 1998). The mouse genes are highly similar to their human counterparts in overall structure and patterns of expression. But the human genes possess more repeats than their murine orthologs. The extent of conservation of individual repeat units varies among the zinc fingers encoded by orthologous genes, with amino acid identities of individual zinc finger repeats ranging from 68%-96% (Shannon et al., 1998).

Zinc finger protein clusters in *Drosophila*: *Tzf*1-4 as the tightest ZF gene cluster

The distribution of zinc finger protein coding genes was well studied for yeast *S. cerevisiae* (Böhm et al., 1997) and *C. elegans* (Clarke & Berg, 1998). Neither organism reveals clusters in genome. In yeast, the majority of zinc finger proteins contain exactly two finger domains; only 10% have more than two. In contrast, *C.elegans has* more zinc finger proteins with three or more fingers than those with only two. Comparing the data from *S.cerevisiae*, *C.elegans*, mouse and human, a clear evolutionary trend

enhancing the complexity in the zinc finger protein domain structure can be seen. First, the number of fingers in a protein increases. Subsequently, the number of genes was increased by duplication. This appears to reflect an increasing demand in regulatory proteins for more complex organisms.

In *Drosophila*, zinc finger gene clusters have 2-5 members (S. Böhm, personal communication). One example of well-studied clusters is the one including *odd-skipped (odd), sob* and *bowel (bowl)* (Hart et al., 1996). *Odd* and *sob* are located at 24A, in the same orientation, with 24kb of DNA between them. *Bowl* is located further away, at 24C2-5. *Sob* and *bowl* have five C2H2 type zinc fingers and *odd* has four. The fingers of *odd* and the N-terminal four fingers of *sob* and *bowl* share an extremely high identity (97,3% between sob and *bowl*, 86,6% between *sob* and *odd* and 87,5% between *bowl* and *odd*). Beyond the zinc fingers, no homology was indicated. The expression of *sob* and *odd* are strikingly similar, while *bowl* is expressed in a different pattern. It was suggested that all three genes have a common ancestor, and arose by at least two independent duplication events (Hart et al., 1996).

Two other zinc finger genes *spalt (sal)* and *spalt-related (salr*) at region 32F/33A are located head to head, in opposite orientations, approximately 65 kb from each other (Reuter et al., 1996; Celis et al., 1996). They have at least partially redundant functions, share high sequence homology and a late expression pattern. s*alr* does not have the early expression pattern of *sal*. Another gene, *spalt adjacent (sala*), was identified in a tail-to-tail orientation with *sal*. Although not sharing any amino acid sequence homology with *sal* and *salr*, it shares the early expression pattern of *sal*. It was suggested that an ancestral *sal/salr* gene underwent complete duplication followed by a chromosomal rearrangement, which separated the *salr* gene from its early cis-regulatory elements. These elements came into the vicinity of *sal* and were appropriated by the unrelated *sala* gene, resulting in the observed similarities between early *sal* and *sala* expression patterns (Reuter et al., 1996).

Up to now, the *Tzf* cluster is the most compact *Drosophila* zinc finger gene cluster with the distance between the adjacent members of not more than 1 kb. They appear to have the same ancestor, considering their high homology in finger region and the completely conserved intron positions. Because they are not orientated in tandem, it is unlikely that all four genes were direct products of multiple duplications of a single gene. *Tzf* is mainly expressed in testis, ovary and embryo; while *Tzf*2 is homogenously expressed everywhere (Xu & Hennig, unpublished data). From the different expression pattern, it seems unlikely that they share all their cis-regulating elements.

A splicing variant of *Tzf*?

The difference between the *Tzf* and the CG4413 transcripts can be explained in different ways:

1. A mistake in the sequencing CG4413 or cloning of a falsely spliced RNA molecule.
2. The CG4413 transcript corresponds to the real splicing pattern and the *Tzf* transcript represents an incompletely spliced molecule. The *Tzf* cDNA was cloned from a testis cDNA library, whilst the CG4413 transcript was obtained from RNA of whole flies. The two forms could be both ubiquitously expressed, or the *Tzf* transcript discovered by us is testis-specific. In this case, it would be of interest to see whether the other *Tzf* tanscripts are also subject to alternative splicing.

```
  1 MKTESNEKWVVCRVCLNNPSEGEELLHDIFSETASTRLDQMLHICAGIPV 50
    ||||||||||||||||||||||||||||||||||||||||||||||||||
  1 MKTESNEKWVVCRVCLNNPSEGEELLHDIFSETASTRLDQMLHICAGIPV 50
            .         .         .         .         .
 51 SLDDNFPDKMCSKCVRCLRLCYKFRLTCQRSHQHIMDMLDREASNANAAG 100
    ||||||||||||||||||||||||||||||||||||||||||||||||||
 51 SLDDNFPDKMCSKCVRCLRLCYKFRLTCQRSHQHIMDMLDREASNANAAG 100
            .         .         .         .         .
101 EGDLLSIAEDLSVESVLKSWEDYASQLDGGMKVEGEEDQQHQVITYVVED 150
    ||||||||||||||||||||||||||||||||||||||||||||||||||
101 EGDLLSIAEDLSVESVLKSWEDYASQLDGGMKVEGEEDQQHQVITYVVED 150
            .         .         .         .         .
151 GDTDDTNMFDVHDPTQPVPNEIEEAETYAEYEEYELLTNENSPEIAQEKG 200
    ||||||||||||||||||||||||||||||||||||||||||||||||||
151 GDTDDTNMFDVHDPTQPVPNEIEEAETYAEYEEYELLTNENSPEIAQEKG 200
            .         .         .         .         .
201 STGTDVATEEPPEEEIAEDILDSDEDYDPTHAKPEKCDRSGRKPVAYHKN 250
    ||||||||||||||||||||||||||||||||||||||||||||||||||
201 STGTDVATEEPPEEEIAEDILDSDEDYDPTHAKPEKCDRSGRKPVAYHKN 250
            .         .         .         .         .
251 SPKVETFKKKVGRKPRNKLSTYICDVCGNIYPTQARLTEHMKFHSGVKPH 300
    ||||||||||||||||||||||||||||||||||||||||||||||||||
251 SPKVETFKKKVGRKPRNKLSTYICDVCGNIYPTQARLTEHMKFHSGVKPH 300
            .         .         .         .         .
301 ECEICGRGFVQNQQLVRHMNTHTGNRPYKCNYCPAAFADRSTKTKHHS.. 348
    ||||||||||||||||||||||||||||||||||||||||||||||||
301 ECEICGRGFVQNQQLVRHMNTHTGNRPYKCNYCPAAFADRSTKTKHHRIH 350
            .         .         .         .         .
349 .................................CDLCGKGFVKAYKLR 363
                                     ||||||||||||||||
351 TKERPYVCDVCSRTFTYSDNLKFHKMIHTGEKPHVCDLCGKGFVKAYKLR 400
            .         .         .
364 LHRETHNRRITWRNDAEESTKAEDVKGETPEFLNELPKE 402
    |||||||||||||||||||||||||||||||||||||||
401 LHRETHNRRITWRNDAEESTKAEDVKGETPEFLNELPKE 439
```

Fig 14: Comparison of the putative protein products of *Tzf* cDNA (the lower line) and CG4413 cDNA (the upper line). The comparison was made using the function GAP of the HUSAR DNA analyse package. The zinc fingers are highlighted in red.

Alternative splicing in zinc finger genes was recently reported for human ZNF41 (Rosati et al., 1999) and mouse KRC (Mak et al., 1998) genes. ZNF41 gene encodes at the C-terminal zinc fingers and at N-terminal a KRAB/FPB domain. Exon skipping at the N-terminus leads to selective usage of two different KRAB/FBP modules, encoding peptides differing in C-terminus and expressed in different tissues. Zinc fingers themselves were in ZNF41 genes constantly expressed. In contrast, the product of KRC gene contains at both N-terminal and C-terminal DNA-binding domains, of which the zinc fingers are essential components. Multiple differentially spliced transcripts were identified in brain and thymus, skipping different zinc finger domains.

Do *TZF* proteins correspond to SCP proteins of the rat?

The cDNA clones encoding the *TZF* protein were obtained by the immunoscreening of the testis cDNA expression libraries with an antiserum against the rat SC protein, SCP3 (Lammers et al., 1994). SCP3 has been located on the lateral elements of the rat SC from zygotene- up to late diplotene-phase. The amino acid sequence deduced from SCP3 does not contain any known nucleic acid binding motif.

It was demonstrated that the N-terminal part of *TZF* protein (DmP3 peptide, Sun, 1994) does react with the antiserum against SCP3. DmP3 contains the first 80 amino acids and is highly divergent compared to the other *TZF* proteins. The amino acid sequence of this peptide does not have any detectable similarity with SCP3, which, however, could not completely exclude the possibility that *TZF* is the *Drosophila* “homologue“ of SCP3 in the sense that it takes over the same responsibility as SCP3 in rat. However, it is more likely that the immuncrossreaction was just a coincidence caused by the local structural similarity between rat SCP3 and the *Drosophila TZF*.

Summary

A zinc finger gene *Tzf1* was cloned by screening a λ-DASH2 cDNA expression library with an anti-Rat SC antibody. A λ-DASH2 genomic DNA library and cosmid lawrist 4 genomic DNA library were screened with the cDNA fragment of *Tzf1* to determine the genomic organization of *Tzf1*. Another putative zinc finger gene *Tzf2* was found about 700 bp upstream of *Tzf1*.

RACE experiment was carried out for both genes to establish the whole length cDNA. The cDNA sequences of *Tzf* and *Tzf2* were used to search the Flybase (Version Nov, 2000). They correspond to two genes found in the Flybase, CG4413 and CG4936. The CG4413 transcript seems to be a splicing variant of *Tzf* transcripts. Another two zinc finger genes *Tzf3* and *Tzf4* were discovered *in silico*. They are located 300 bp away from *Tzf* and *Tzf2*, and a non-tandem cluster was formed by the four genes. All four genes encode proteins with a very similar modular structure, since they all have five C2H2 type zinc fingers at their c-terminal ends. This is the most compact zinc finger protein gene cluster found in *Drosophila melanogaster.*

Part II:

Subcloning and sequencing of the cosmid 19G11

Introduction

Histones are highly conserved small basic proteins that constitute the elementary units of chromatin in the nuclei of all eukaryotic cells, the nucleosomes. In higher eukaryotes histones are encoded by multigene families, containing members of two types: replication dependent, or cell-cycle regulated histone genes and replication independent, or replacement histone genes (Schümperli, 1986).

Cell-cycle regulated histones are expressed only in S-phase cells. Whilst the replacement histones genes are also expressed in non-S-phase cells, albeit at a low level. The distinction between the two types of histone genes is also based on their structure, genomic organization, mode of regulation and the type of mRNA (Osley, 1991; Schümperli, 1988).

The cell-cycle regulated histone genes are present in multiple copies, contain no introns and their transcripts are not polyadenylated (Wells, 1986; Wells &McBride, 1989). In contrast, the replacement histone genes display no strict regulation in relation to the cell-cycle. They are single copy genes and resemble the rest of the protein-coding genes by the presence of introns and polyA tails in their mRNAs (Brush et al, 1985; Schümperli, 1986).

The *Drosophila* histone H4 replacement gene H4r was cloned in our Lab in 1996 (Akhmanova et al, 1996). It displays all properties of a histone replacement gene: It contains two introns, generates poly-adenylated mRNA, represents the predominant H4 transcript in non-dividing tissues and is present in the genome as a single copy. The encoded polypeptide is, however, identical to the *Drosophila* cell-cycle regulated histone H4.

H4r is localized in the region 88C of chromosome *3*, in very near neighbourhood of the *punt* gene (Ruberte et al. 1995). To study the regulation pattern of H4r gene and thus to gain an insight into the significance of all histone replacement genes, it is necessary to clone the flanking region of H4r gene. H4r cDNA was hybridised as a probe to a lawrist 4 cosmid library of *D. melanogaster* genomic DNA (Hoheisel et al, 1991). One positive clone, 19G11, was obtained and it was verified that it contains H4r. We decided to sequence this cosmid to get detailed information about the flanking region of the H4r gene, and also to search for possible related genes, which are located in the vicinity of H4r and are possibly co-regulated with H4r.

Results

BamHI was chosen to subclone Cosmid 19G11

To obtain the flanking region of H4r gene, H4r cDNA was hybridised as a probe to a lawrist 4 cosmid library of *D. melanogaster* genomic DNA (Hoheisel et al, 1991). One positive clone, 19G11, was obtained and it was verified that 19G11 contains H4r. To subclone the insert of cosmid 19G11, its DNA was digested with several enzymes.

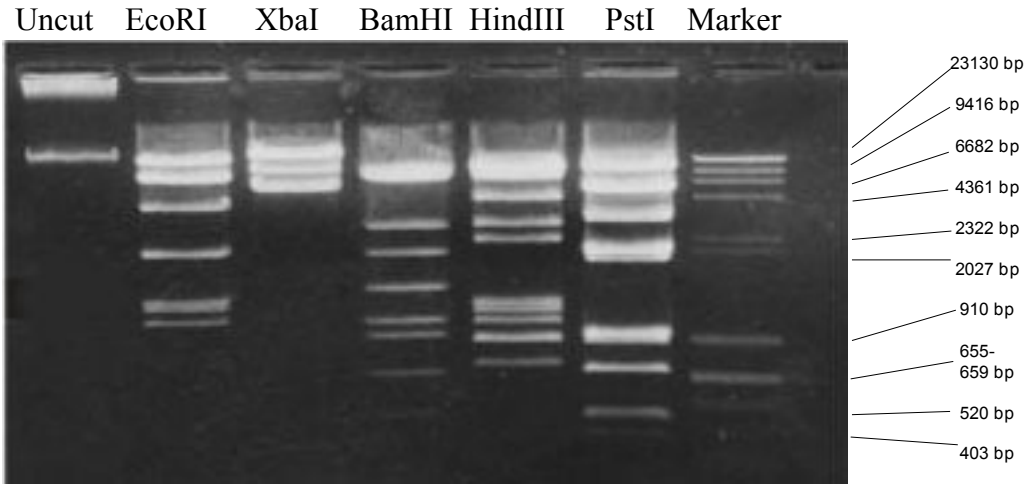


Fig 1: Digestion pattern of Cosmid 19G11with several different enzymes.

The largest fragments of BamHI digests were smaller than 9 kb. Therefore they could be relatively easily cloned into plasmid vectors. BamHI was thus chosen for subcloning 19G11.

The BamHI fragments were named Bam1 to Bam8. After further running the digestion gel, it turned out that the signal Bam8 actually consisted of 4 fragments with similar sizes. They were named 8S1, 8S2, 8M and 8B.

Subcloning of the Bam fragments

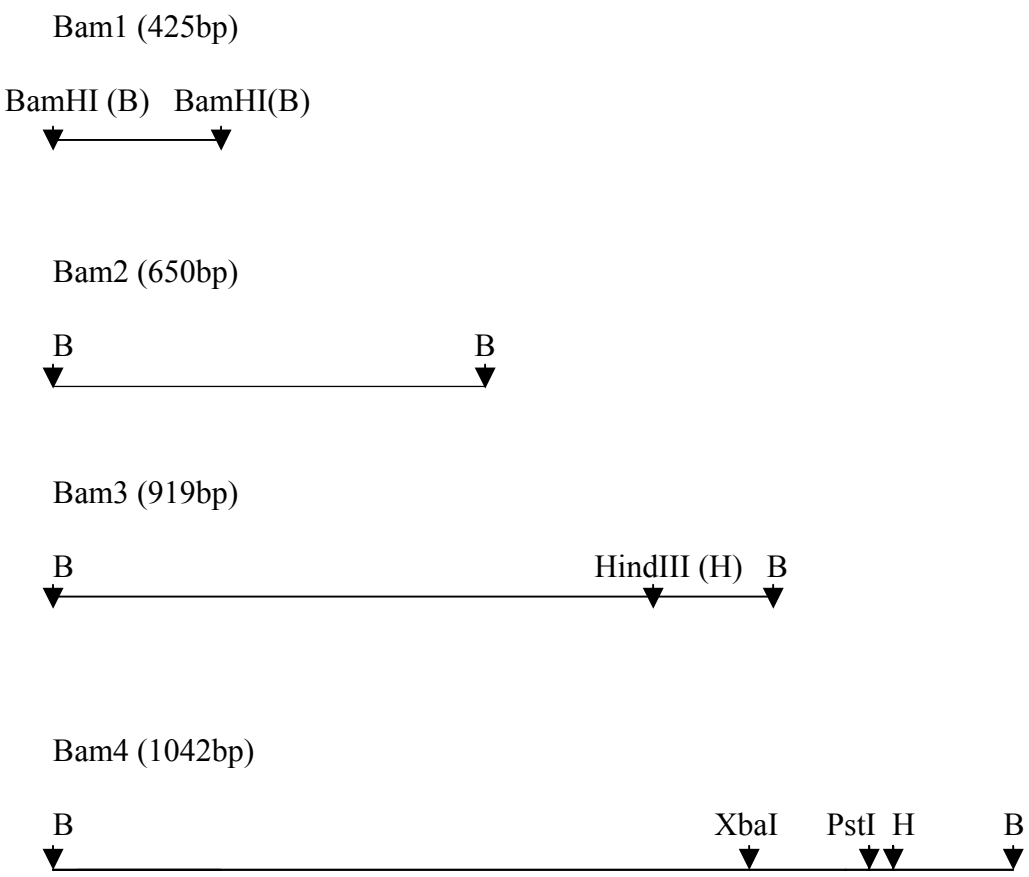
All fragments were recovered from agarose gel and ligated into pBluescript vector that was already digested with BamHI and dephosphorylated. The ligation mixture was used to transform competent cells of the *E.coli* strain XL1Blue. After each transformation, minipreps were made from 6 to 12

separate white colonies, and the recovered plasmid DNA was digested with BamHI. The plasmids with an insert of the same size as the original fragment were further analysed.

Their inserts were again recovered from agarose gel and digested with several other enzymes. When all digestion patterns were as same as the digestion patterns of the original fragment, the plasmids were considered to be the correct subclones.

Mapping and subcloning of Bam1-Bam4

For fragments smaller than 1200 bp (Bam1 to Bam 4), at least two correct subclones were sequenced from both ends. These fragments are illustrated in the following regarding their sizes and digestion patterns with some enzymes.

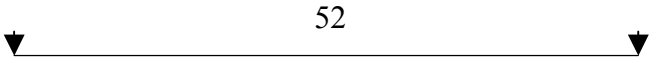


Mapping and subcloning of Bam5, Bam6 and Bam7

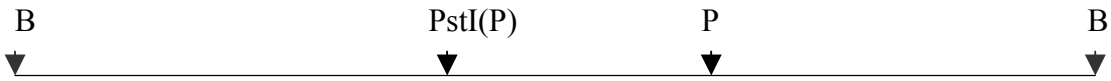
The bigger fragments that could not be covered by sequencing from both ends were further digested to smaller fragments and then subcloned. The subclones were then sequenced from both ends.

Bam5 (1391 bp)

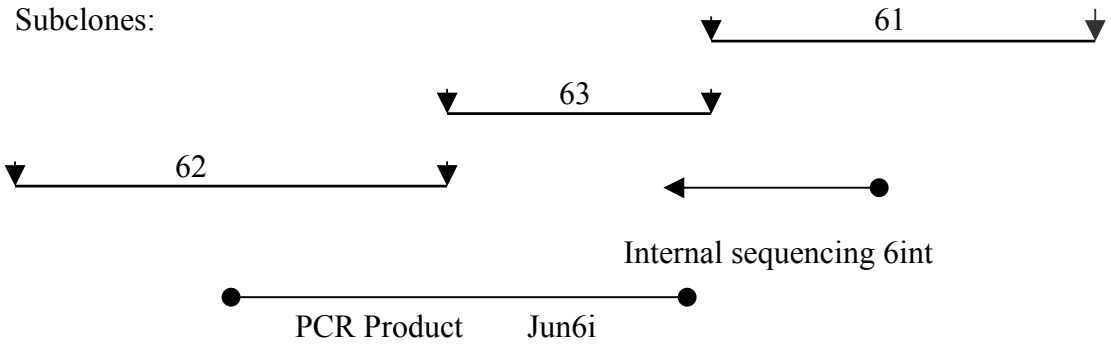
Subclones:



Bam6 (1962 bp)



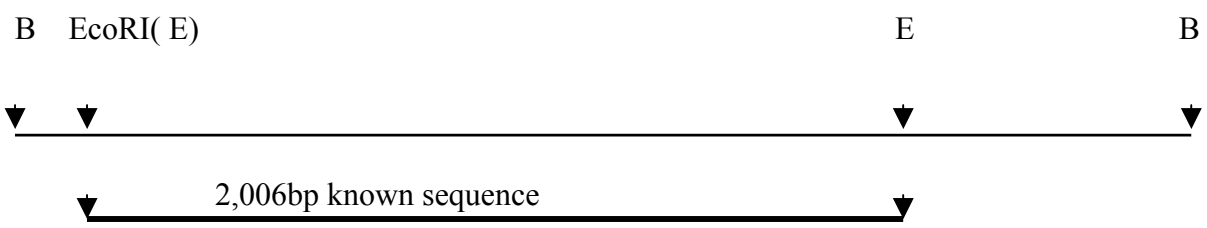
Subclones:



The arrangement of the fragments digested from Bam6 was decided with an internal sequencing with primer 6int, which covered the junction between fragment 63 and 61, and a PCR product of primer B61 and B63, which covered the junction between 62 and 63.

Primer 6int: 5'-GCCGGTGGAATGTTTCTGTGGA-3'

Fragment Bam7 (2702 bp)



A 2,006 bp EcoRI-EcoRI fragment in the middle of the Bam7 fragment was already sequenced (Akhmanova, 1997). Sequencing from both ends could cover the unknown parts and overlapped with the know EcoRI fragment.

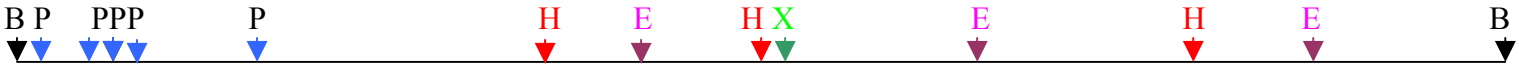
Mapping, subcloning and sequencing of fragments 8S1, 8S2, 8M and 8B

The biggest BamHI fragments 8S1, 8S2, 8M and 8B were sequenced using combinations of different strategies. They were further digested, subcloned and sequenced from both ends. Also, primers were made according to the partial sequence of the subclones and PCRs were performed, their products cloned and sequenced from both ends. Finally, some internal sequencing was also carried out, which was at the beginning of this work not possible because of technical limitations. The final sequence of each BamHI fragment could be determined as the contig of all these overlapping sequence segments.

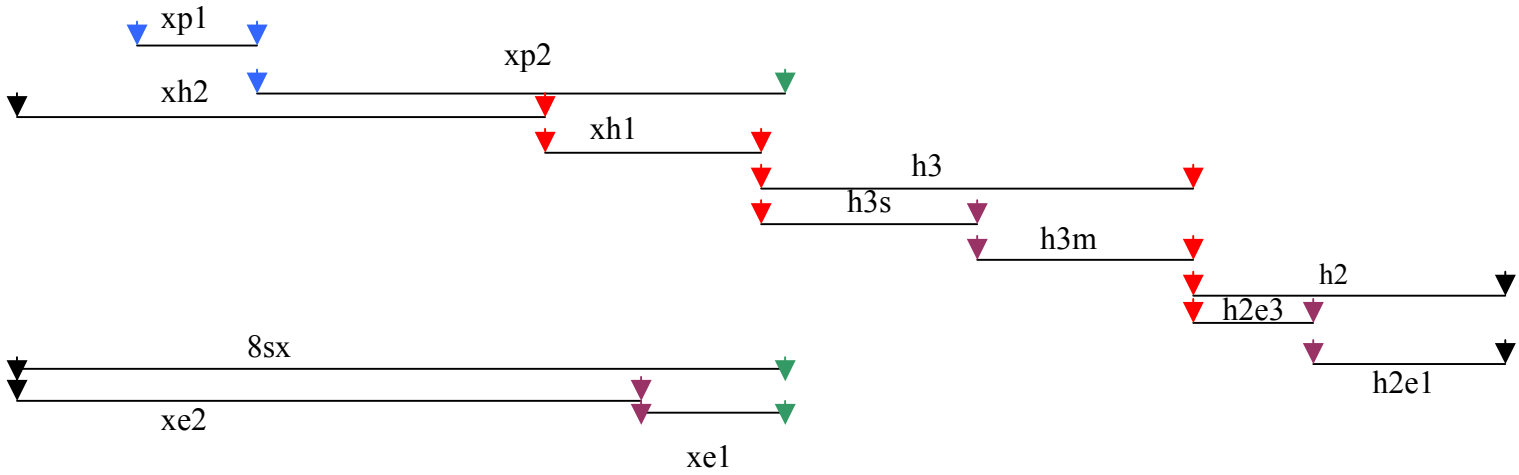
For each BamHI fragment, the localization of subclones and PCR products are as follows.

Mapping and subcloning of fragment 8S1(6903 bp)

a.

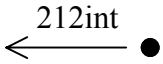


b.



c.

d.



a. Restriction map of fragment 8S1. The enzymes used for subcloning are shown in different colours.

b. Subclones originated from fragment 8S1. They were digested from 8S1 with enzyme combinations that are marked at the ends of each fragment, and then ligated into pBluescript vector, which was prepared by digestion with the same enzymes.

c. PCR products with 8S1 or its subclones as template. They were all cloned into PCR 2.1 vector (TA cloning kit, Invitrogen).

Primers	Template	Product
L1 +L2	h2frag	L1
L3 +L4	h3frag	L3
L5 +L6	8sxfrag	L5
WG210+WG211	8s1	j1
WG214+WG215	8s1	j3

L1 5' GGA CGG TTC GGT GGA CAA GAC-3' (21mer)
L2 5' TTT AGC GGC TCT CAG CTG CCT GCG GT-3' (26mer)
L3 5' CGC CAC ATG GCA CTC AAT CAA CTT AGC-3' (27mer)
L4 5' CAT AGA CTC CCA TTC GTT T-3' (19mer)
L5 5'CCA GAT CGT GCT CCT CCA GCA AG-3' (23mer)
L6 5'GAT CTT CTT GGC CAG GAC-3' (18mer)

WG210 5'CGT GGC CAC ACA GGA TAT GCA -3' (21mer)
WG211 5'GCG CTA TGA CTC TTT CCC GC-3' (20mer)
WG214 5'GTC GAT TGC TGC TGC TGA-3' (18mer)
WG215 5'GTC CAT TGC CAG TAA TTC GC-3' (20mer)

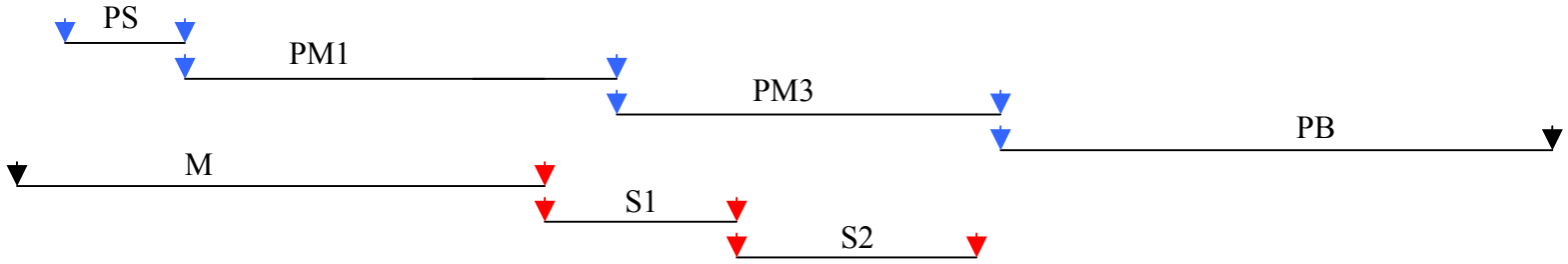
d. Internal sequencing with 8S1 as template.

WG212 5' GGG TTG TAG GTT CAT CGC-3' (18mer)

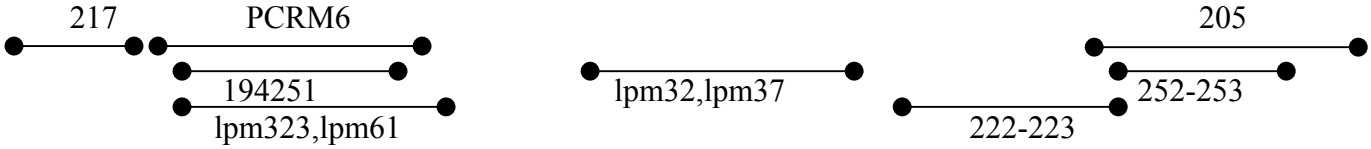
Mapping and subcloning of fragment 8M

a.

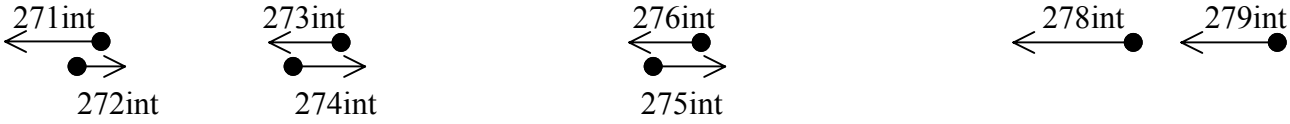
b.



c.



d.



a. Restriction map of fragment 8M.

b. Subclones originated from fragment 8M.

c. PCR products with 8M or its subclones as template. They were all cloned into PCR 2.1 vector (TA cloning kit, Invitrogen).

Primers	Template	Product
WG218+ WG219	8M	217
WG192	M	PCRM6
WG193+WG194	PM1	lpm61, lpm323
WG251+WG194	PM1	194251
WG195+WG196	PM3	lpm32, lpm37
WG 223+ V4	8M	222-223
WG205+WG206	PB	205

Primers:

WG192 5'-GTG ATT TCT TGG TGC-3’ (15mer)

WG193 5'-CAT CCG CAT GAT AGA TCC-3’ (18mer)

WG194 5'-CGT CAT CAG ACG GTT GGG T-3' (19mer)

WG195 5'-GCC TTC AGT TTC CGG TTA TG-3' (20mer)

WG196 5'-GGA ATG GCT GTA TTA TAC -3' (18mer)

WG218 5'-GGC TGT AGC TTA CGA CAC TCC TTG-3' (24mer)

WG219 5'-GCG CAG CGT ACT CAA GAC GAT ATC C3’ (25mer)

WG205 5'-CGC CTA GGC CTT AGC CTT-3’ (18mer)

WG206 5'-GGC ACG CCG CCG TTA ATA-3’ (18mer)

WG223 5'-GTC TGA CGC CTA TTA TCT CG-3’ (20mer)

V4 5'-GGT ATC ATT CGC ACA CTC CCC AG-3’ (23mer)

WG251 5’-CGA ATA TCT CGG TTT CTC ACG GAC-3’ (24mer)

WG252 5’-GTG AAG GGA ACT GCT GGC AA-3’ (20mer)

WG253 5’-ATT TGT TCG CTG ACG AAG CGC G-3’ (22mer)

d Internal sequencing with 8M as template.

WG271 5’-CAC ACT CAC ATA AAG GCT CAC ACA C-3’ (25mer)

WG272 5’-CCC AAA AAA CCG CAT TAT TCT GCC-3’ (24mer)

WG273 5’-CCA AGG CTT ACG ATC ACT-3’ (18mer)

WG274 5’-CGT GTG CAT GTG TGT GTT-3’ (18mer)

WG275 5’-GCA GCT ATC CAC AGC TCG TTT-3’ (21mer)

WG276 5’-CTC GAT GTG AAA CGC ACA AGC AAT GG-3’ (26mer)

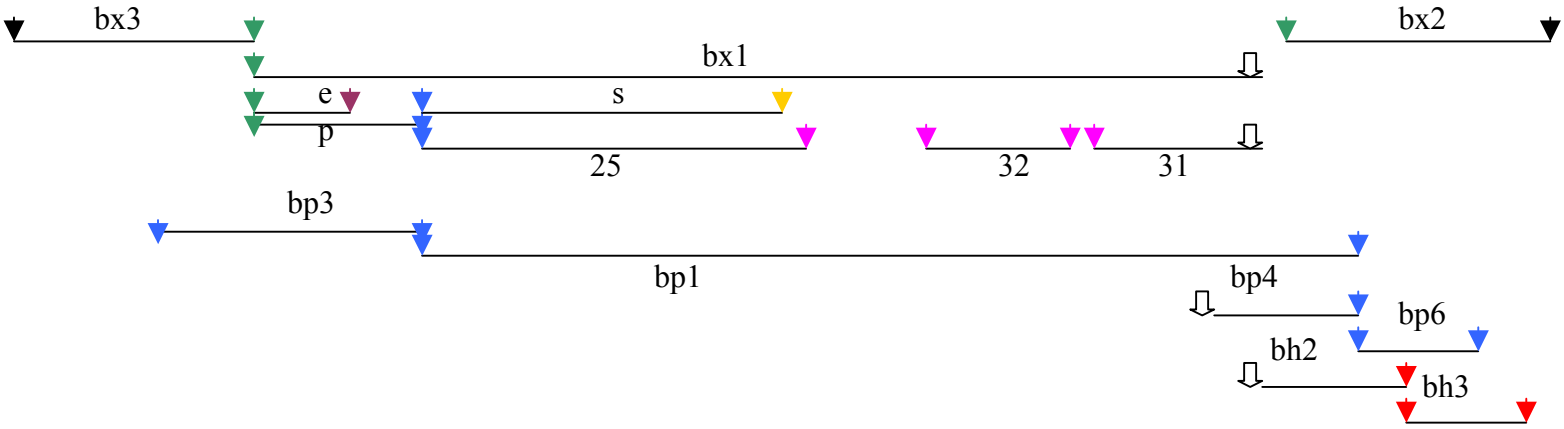
WG278 5’-GCC CAA ATC GGG GGC GAA TT-3’ (20mer)

WG279 5’-CGC TTC TCC GCC TTA CAC CTC-3’ (21mer)

Mapping and subcloning of fragment 8B

a

b.



c.

d.

a. Restriction map of fragment 8B.

b. Subclones originated from fragment 8B. One end of subclones bx1, 31, bp4, bh2 was not a real restriction site. They were probably digested by the star activity of BamHI.

c. PCR products with 8B as template. They were all cloned into PCR 2.1 vector (TA cloning kit, Invitrogen).

Primers	Template	Product
WG220+WG221	BX3	a
WG224+WG225	S	2245
WG231+WG234	BX1	g
WG232+WG233	p	2g

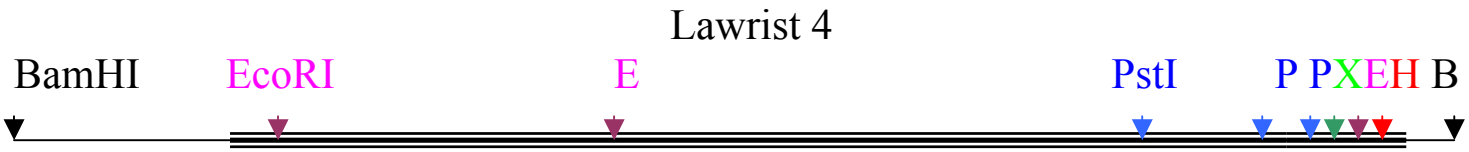
WG220 5'-GGA CAA CGA TCC GCT GAC CTT-3' (21mer)
WG221 5'-TTG GGA CGA TCG ATA GGC AGG-3' (21mer)
WG224 5'-CGG GAA ACT CGT CAC AGC GGC AGA CAA A-3' (28mer)
WG225 5'-CAA CAC CCA TAC GCC TTT GC-3' (20mer)
WG231 5'-GCT CAA CTG GAA TCG AGT TAT CGC -3'(24mer)
WG232 5'-GTC GGG CAG TTT AGA AAA CG-3' (20mer)
WG233 5'-AGA ATT GGC CAG TGG ATT GGC CGA-3' (24mer)
WG234 5'-GCG ATG GAA AGC TCA ACC TTC-3' (21mer)

d. Internal sequencing with 8B DNA as template.

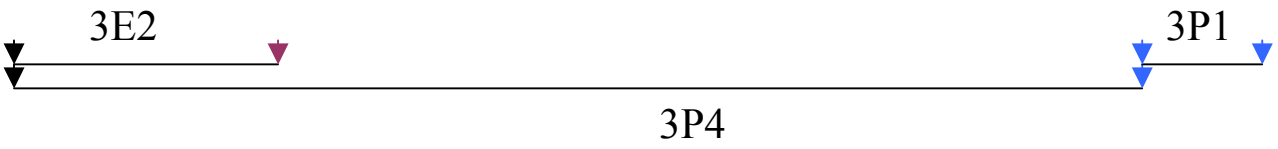
WG227 5'-GTT CGC GAG TTC AAC TCG GTT AGG C-3' (25mer)
WG229 5'-GGC GTT CAG GCG TTC CGA GTA CTC CAA TTT-3' (30mer)
WG230 5'-ATT CCT GTC CTG GCA TGC GAC TTT CCT TCG-3' (30mer)

Mapping and subcloning of fragment 8S2

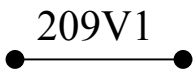
a.



b.



c.



Sequencing of 8S2 ends showed that 8S2 was the fragment containing the Lawrist 4 cosmid vector. The two segments flanking the vector were mapped and sequenced.

a. Restriction map of fragment 8S2.

b. Subclones originated from fragment 8S2.

c. PCR product with 8S2 as template. It was cloned into PCR 2.1 vector (TA cloning kit, Invitrogen).

Primer	template	product
V1 + WG209	3E2	209V1

Primer:

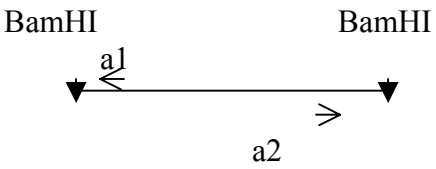
V1 5'-GCG ATG ACC CTG CTG ATT GGT TCG-3' (24mer)

WG209 5'- GGC TCC ACT TAA GAT GTT CGG C-3' (22mer)

Arrangement of BamH1 fragments in cosmid 19G11

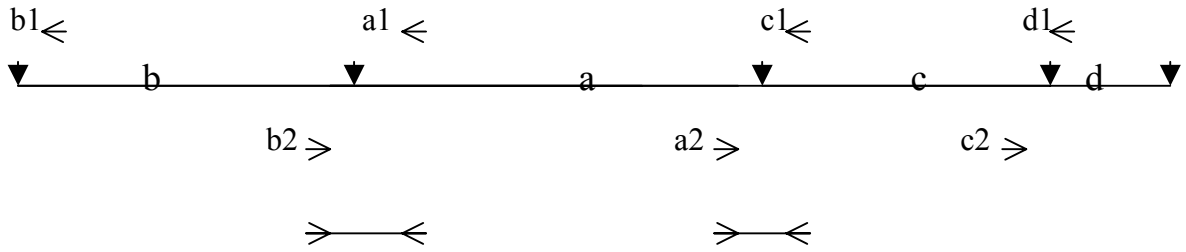
The arrangement of all the BamHI fragments in cosmid 19G11 was studied using a PCR test concept.

Two primers were designed for each fragment. They were located few hundred base pairs away from the both ends, and oriented towards outside.



A series of PCR was performed using the intact cosmid DNA as template. Each primer had the combination with every other primer except the one from the same fragment.

When two fragments are located adjacent to each other, PCR products should be expected in the combination of the two primers oriented against each other, one from each fragment, both towards the overlapping BamHI restriction site end.



As for primer a1 in the figure, it would only give a PCR product with primer b2 from fragment b. The primers from other fragments, even when they are in same orientation like b2, are just too far away from a1 to give a reasonable PCR product with a1. So it can be concluded that fragment b is located adjacent to fragment a, with the orientation shown in figure.

When the adjacent fragment is very small, like fragment c in the figure, primer d1 from fragment d is still in a reasonable distance from a2 so that not only the combination a2-c1, but also a2-d1 can give a

PCR product. In this case, the sizes of PCR products were compared, and the arrangement of all three fragments can be concluded.

All the junction PCR products were cloned and sequenced, to be sure of the arrangement, and also to make sure that there are no very small fragment between the two "adjacent" fragments, which was too small to be seen in digestion gel.

The primers designed for all the fragments to deduce their arrangement are shown in the following:

Fragment Bam1
B11 5'-GCG GCC GCC GTG CCA CTG AGA ATT G-3' (25mer)
B12 5'-GCG GCC GCC AAT TCT CAG TGG CAC G-3' (25mer)
Fragment Bam2
B21 5'-GCG GCC GCC CAT GTG CAT GTC GCC GAC T-3' (28mer)
B22 5'-GCG GCC GCC GAC CAA AAT GTC ATA TCC CCC GG-3' (32mer)
Fragment Bam3
B31 5'-GCG GCC GCA AGC TTA CCT CCA TCG ACT GGG C-3' (31mer)
B32 5'-GCG GCC GCG CTG GAG AAT CAG GCT GAG-3' (27mer)
Fragment Bam4
B41 5'-GCG GCC GCG GGT GCT GAT CCA TCT CGA-3' (27mer)
B42 5'-GCG GCC GCC TGT TGA GGA TCT GGC TGC-3' (27mer)
Fragment Bam5
B51 5'-GCG GCC GCC CTG GAG AGT GGC TAC TCT GGT-3' (30mer)
B52 5'-GCG GCC GCG GTG AGA TAC AGA CTC AAC CCC-3' (30mer)
Fragment Bam6
B61 5'-GCG GCC GCT TGC AGA AAT CGA GAG CCC GC-3' (29mer)
B62 5'-GCG GCC GCC GGT TGT GTA CAT AAG GCG AGC G -3' (31mer)
B63 5'-GCG GCC GCT TGC TGA TGC TGC TGC TGG A-3' (28mer)
B64 5'-GCG GCC GCC CCG CTT CAA CGA ACT CAC TGC-3' (30mer)
Fragment Bam7
B71 5'-GCG GCC GCA GGC TCG TGG CGA ATA ATC G-3' (28mer)
B72 5'-GCG GCC GCA GAT GGC AGG ACG TGC-3' (24mer)
Fragment Bam8S1
B8s1 5'-GCG TAC GTT CTC CAA TCC CTC GG-3' (23mer)

B8s2 5'-TGG CCA GCG TGA TCT TGG TAC-3' (21mer)

Fragment Bam8M

B8M1 5'-CGG CAA CGG CTT CAC ATT CGA-3' (21mer)

B8M2 5'-CTA GTT CGC ACT CCT ATT AAC G-3' (22mer)

Fragment Bam8S2

B8V2 5'-GAC GCA GGT ATC GTA T-3' (16mer)

B8V3 5'-GGT ATC ATT CGC ACA CTC CCC AG-3' (23mer)

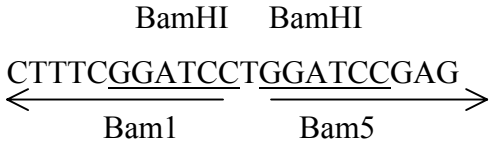
Fragment Bam8B

B8B1 5'-CGA GTT GTC CAA GTC GAA TAC CC-3' (23mer)

B8B2 5'-ATC CCT CTG TGC AAT TCG AGA CG-3' (23mer)

Sequencing result:

Sequencing of the PCR product of B12 and B52 shows a duplicate BamHI site at the junction.



All the other junction PCR products have shown sequences from the two fragments overlapping at the BamHI site. There is no other fragment in between.

Genomic arrangement of Bam fragments

To exclude the possible rearrangement of the genomic DNA fragments in establishing the cosmid library, all the junction PCRs (except the two with B8V2 and B8V3 from the lawrist vector) were carried out once more, using genomic DNA of *D. melanogaster* as template. For all the reactions, the same result was obtained as with cosmid 19G11 DNA as template. It was concluded that the sequence of the 19G11 insert does correspond to the 34 kb genomic fragment around H4r gene. It is not a cloning

artefact during the establishing of the cosmid library resulted from artificial ligation of several fragments, which are actually not adjacent to each other in genome.

Conclusion

This work was done between 1997 and 1998. The *Drosophila* genome was yet far from completed at that time. This sequence was submitted to the Genebank under the Accesion Nr: AJ007334.

Methods for Part I and II

Cosmid DNA Preparation

The cosmid glycerol stab (kept at -80°C) was streaked onto LB plate with 30μg/ml kamamycin. Separate colonies were inoculated into 100ml LB medium with 30μg/ml Kanamycin. Cells from an overnight culture with an a OD_{600} of more than 1.2 were collected in corex tubes and centrifuged at 4,000rpm for 10min with a HB4 rotor. The supernatant was discarded as completely as possible. 1ml Sol I was added to the pellet and vigorously vortexed until the pellet was completely resuspended. 2ml freshly made Sol II was then added, and the tubes were moderately shaken for 5min at room temperature. 1.5ml ice-cold Sol III was then added, and the tubes were briefly vortexed to be well mixed. The mixture was let stand on ice for 5min, and then centrifuged at 10,000 rpm for 10min with a HB4 rotor. The supernatant was carefully recovered into new tubes, and extracted with once phenol/chloroform (1:1) and twice chloroform. The purified supernatant was precipitated with 0.6 volume isopropanol. The pellet was washed with 70% ethanol and dried in vacuum. The pellet was dissolved in water.

Solution I:	Solution II:	Solution III:
50mM glucose	0.2N NaOH	5M potassium acetate (60ml)
25mM Tris.Cl (pH8.0)	1%SDS	glacial acetic acid (11.5ml)
10mM EDTA (pH8.0)		H_2O (28.5ml)

Mini-prep of Plasmid DNA

Single bacteria colonies were picked in 2ml liquid LB-medium with 50μg/ml Ampicillin, and incubated overnight at 37°C. The cultures were transferred into Eppendorf tubes, and spun down for 30sec at max speed. The pellets were dissolved completely in 100μl Sol I. 200μl freshly made Sol II was then added followed by moderate shaking, and afterwards 150μl ice-cold Sol III with brief vortexing. The mixture was then centrifuged for 10min at max speed. The supernatant was transferred to new tubes and then precipitated with 1ml absolute ethanol. The precipitation took 10min at RT, the tubes were then centrifuged for 15min at max speed. The DNA pellets were washed by shaking in 70% ethanol and spinning down once more. The pellets were dried in vacuum and then dissolved in 30μl H_2O.

The solutions are as same as those for the cosmid preparation.

DNA Digestion

Proper amount of DNA was digested with restriction enzyme of at least 3 Units/μg DNA. The volume of enzyme never exceeded 1/10 of the total volume. The digestion system was completed with the corresponding buffer provided by the manufacturer. In the case of double digestion, the one-phor-all buffer (Pharmacia) with a right concentration for both enzymes was chosen. The reaction was carried out at 37°C for 1-2 hr.

Agarose Gel Electrophoresis

The gel concentration was chosen according to the size of the fragments to be separated on the gel. 0.7% agarose was used when the major fragments were bigger than 5kb, 2% agarose when they were smaller than 1kb. For fragments with sizes in between or of unknown size, 1% gel was used as standard.

The probes to be analysed were mixed with 6x loading buffer (0.25% bromophenol blue, 0.25% xylene cyanol FF and 15% Ficoll), and the electrophoresis was carried out in 1x TBE with 0.08 μg/ml EtBr.

Two different kinds of marker were used, for standard and small fragments respectively. The standard marker consists of 2 parts: *lambda* DNA digested with Hind III, and plasmid pYH48 DNA digested with AluI. The fragments sizes are

23130, 9416, 6682, 4361, 2322, 2027 bp (*lambda* part),

and 910, 655-659, 520, 403, 317, 281, 257, 226, 187 bp (pYH48 part).

A marker for smaller molecules was obtained from a HpaII digestion of pBluescript DNA. The fragments are

710, 489-404-367, 242, 190, 147, 118-110, 67-57-40-26 bp.

Fragment Recovery from the Gel

Low-melting agarose method:

A block of gel was dug out in front of the fragment to be recovered. Low-melting (LM) agarose gel solution of the same concentration as the whole agarose gel was poured into it. After that the low-melting agarose had congealed, the gel was let run further till the wanted fragment had completely run into the LM agarose piece. This piece with the fragment was then cut out and melted at 65°C for 10min. After adding an equal volume of TE and 1/20 volume of 3M NaOAC to the gel solution, it was extracted twice with phenol and once with chloroform. The DNA was precipitated with 2.5 volume ethanol and dissolved in water.

Macherey-Nagel method:

NucleoSpin Extract 2in1 from company Macherey-Nagel was used to recover some fragments. The process was carried out according to the protocol of the manufacturer.

Vector Preparation

When a vector is needed which was only cut by one restriction enzyme, it was dephosphorylated before ligation to reduce the vector self-ligation background. Dephosphorylation was carried out with the alkaline phosphatase from Boehringer-Mannheim (Now: Roche Molecular Biochemicals) according to the protocol of the manufacturer.

Ligation

100-200 ng vector and at least 3 times the molar amount of insert were mixed together. 5U T4 DNA ligase (MBI-Fermentas), 1/10 from the total volume of ligation buffer was added and the final volume was adjusted with water.

Or the same molar amounts of vector and insert were mixed and the total volume adjusted to 20µl with water. This 20µl mixture was then added to one aliquot of Ready-to-Go DNA ligase (Pharmacia).
The ligation mixture was incubated in a 14°C water bath over night (Fermentas ligase) or for 2-5 hr (Ready-to-Go ligase) and was then ready for transformation.

Transformation

With XL1-Blue competent cells

One single colony of XL1-Blue was inoculated into 5ml LB and incubated overnight at 37°C. 0.5ml from this starter culture was transferred into 100ml fresh LB and was let grown for about 3 hours until the OD_{600} of 0.3-0.5 was reached. The cells were harvested by 10min centrifugation at 4°C for 2000rpm with rotor HB4, and afterwards gently resuspended in 20ml ice cold 0.1M $CaCl_2$. They were harvested again by centrifugation and again resuspended in 1.5ml ice cold $CaCl_2$. After at least one-hour incubation on ice, the cells could be used as competent cells.

Each 100µl of cells was mixed with 2µl ligation mix and incubated on ice for 30min to 1 hour. They were then heat shocked at 42°C for 90sec, and placed back onto ice for at least 2min. 400 µl LB medium were then added to the cells and they were gently shaken at 37°C for 1 hour. This culture was then plated on LB-Amp plates (50µg/ml Ampicillin) with 40µl 2% IPTG and 40µl 20mg/ml X-Gal. The plates were incubated at 37°C incubator overnight.

With commercial competent cells

Some transformations were carried out with Competent TOP10F‘ cells from Invitrogen. The procedure was done according to the instructions from the manufacturer.

PCR

The proper primers were designed and they were synthesized by PudongGen. In each 50µl PCR reaction, there were 50-100 pmole of each primer, 100ng template DNA, 1 Unit Taq polymerase (Gibco), 10nmol dNTP, 5µl 10x PCR buffer (Gibco).
PCR was carried out under the following conditions: denaturation at 94°C for 3min, then 30 cycles with denaturation at 94°C for 45sec, annealing at 68°C for 40sec, and extension at 72°C for 100sec till 45sec (depending on the size of the product to be amplified). The annealing temperature varied between 55°C and 70°C according to the Tm of the primers (Sambrook et al 1989). The reaction was stopped after a final extension step at 72°C for 10min.

PCR Product Cloning

With restriction enzymes

The primers for PCRs were designed in this way that restriction enzyme sites were attached at their 5' parts, normally 3-4 nucleotides away from the very 5' end. The product was digested with these enzymes, and ligated into proper vector.

With PCR 2.1 vector (Invitrogen)

PCR products were cloned into the PCR 2.1 vector (TA cloning Kit, Invitrogen) according to the information of the manufacturer.

Southern Transfer

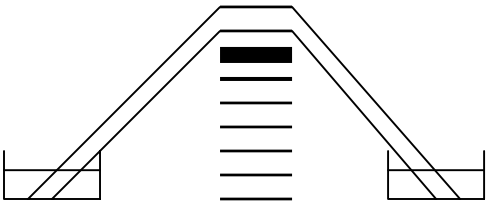
Downward Blotting was carried out according to Koetsier et al, 1993.

A pile of tissue paper was set on the table. 5 pieces of 3MM filter paper and one piece of Hybond-N+ Nylon membrane (Boeringer-Mannheim) was cut into exactly the same size as the gel. 2 pieces of dry 3MM filter were put on the pile of tissue paper, then in the following

sequence 1 piece of 3MM filter paper prewetted in NaOH, the prewetted H+ Nylon membrane, the gel, 2 pieces of prewetted 3MM filter paper. Finally, 2 long sheets of prewetted 3MM covered the pile as „bridge“ connecting two tanks filled with 0.4N NaOH. Blotting was finished in 1.5- 2.5 hours, the membrane was marked and crosslinked in UV light for 5min.

Figure: Downward southern blotting.

3MMpaper Nylon membrane gel piece NaOH tank



Labelling

Reaction

Nick-translation

DNA fragment was mixed with 2µl DNaseI solution (10^{-5} mg/ml in water)(Gibco), 10 U (2µl) *E.coli* DNA polymerase I (Gibco), 2µl 10x NT Buffer, 2µl 20mM dNTP, 3µl [α-^{32}P] dCTP (50 µCi, 3000Ci/mmole) and the system was filled up to 20µl with H_2O. The mixture was incubated at 16°C for 2 hours.

10x NT Buffer consists of 500mM Tris pH7.8, 500 mM $MgCl_2$, 1 mM dithiothreitol and 500µg/ml bovine serum albumin.

Hot-PCR

100ng DNA template, 33µM α-^{32}P dCTP, 2,5µM dATP, dTTP, dGTP, 50pmole each primer, 1x PCR buffer (Gibco) and 1U Taq polymerase (Gibco) was mixed in 20µl final volume. The amplification condition was same with non-radioactive-PCR.

Purification

After the labelling reaction, the labelled fragment was separated from free nucleotides by passing the reaction mixture through a Sephadex G-100 column. The first radioactivity peak (50-100cps) was collected and used as probe for hybridisation.

Hybridisation

Denhardt system (for self-made blots)

Prehybridisation was carried out in 2x SSC, 0,1%SDS, 5x Denhardt's at 68°C for 0.5 hours with continues shaking. The probe was denatured by boiling for 5 minutes and cooling on ice. The denatured probe was added to the hybridisation solution that was identical with prehybridisation solution and shaken overnight in a 68°C water bath. The blot was then washed with 2x SSC, 0.1% SDS and 0.5x SSC, 0.1%SDS at 65°C for 0.5 hour, respectively. Afterwards, the blot was once more briefly rinsed with 2x SSC at room temperature and was subsequently exposed to X-ray film (X-omat, Kodak).

Phosphate system (for cosmid blots)

Hybridisation was carried out in 0.5M Na-phosphate, pH7.2, 7% SDS, 1mM EDTA, 0.1mg/ml yeast tRNA at 65°C over night. The blots were rinsed twice at room temperature in 40 mM sodium phosphate, pH7.2, 0.1% SDS, 1.5 litre of the same buffer was added to up to four filters and slowly rocked in a water bath of 65°C for 15 to 30 minutes. The filters were then exposed to X-ray film using two intensifying screens.

RNA Preparation

GTC method

A 2% agar plate was made with apple juice. The flies were let lay eggs on the plates at 22°C over night. The eggs were collected and washed extensively with tap water and then with 0.9% NaCl. The embryos were treated with 1% Triton X-100 in 0.9% NaCl for 5 min at room temperature. After being washed five times with 0.9% NaCl, the embryos were suspended in 3% $NaClO_3$ solution for 3 min. Finally the embryos were again washed with 0.9% NaCl and frozen in liquid nitrogen. 0.3g embryos (wet weight) were lyophilised and then suspended in 2.4ml guadinium thiocyanate solution (50% guanidinium thiocyanate, 0.5% sarcosyl and 25 mM sodium citrate, pH7.0) and homogenized. The homogenate was then loaded on the top of 0.9ml CsCl solution (5.7M CsCl and 0.1mM EDTA, pH7.0) and centrifuged in an IEC SB405 rotor at 32k rpm for 16 hr at 15°C. The RNA pellet was dissolved in 200µl TES buffer (10mM Tris-Cl, pH7.4, 5mM EDTA and 1% SDS) and extracted with 1 volume of chloroform/butanol-1 (4:1). The RNA was then recovered by ethanol precipitation and dissolved in water. The RNA concentration was determined with a spectrophotometer (Sambrook et al. 1989).

With RNeasy Kit (Qiagen)

Some total RNA preparations were carried out with the RNeasy Kit according to the manual of manufacturer.

Genomic DNA Preparation

Genomic DNA preparation was basically carried out according to Ashburner (Ashburner, 1989) with modifications. Each 5 flies were fully homogenized in 50µl homogenizing buffer. 50µl lysis buffer was added and incubated at 65°C for 20min. After cooling to room temperature, 33µl ice cold 4M KOAc was added and incubated on ice for 30min. The supernatant was transferred to new tubes after a centrifugation at 12K (Eppendorf table

centrifuge) for 15min. It was afterwards extracted twice with a 1/1 mixture of phenol/chloroform, once with chloroform, and then precipitated in ethanol.

RACE

Tailing method

Reverse Transcription (first strand cDNA synthesis)

1µl (1pmol/µl) primer was annealed to 2µl (4mg/ml) *Drosophila melanogaster* embryo total RNA, with 10µl formamide (with 3µM EDTA), 12.4µl pure formamide, 1.6µl PIPES, and 3µl 4M NaCl overnight at 45°C after 10min incubation at 85°C.The mixture was precipitated with EtOH and the pellet was dried in air and dissolved in 4µl 5x first strand buffer (Gibco), 0.5µl 20mM dNTP, 0.5µl 40 U/ml RNAsin, 2µl DTT, 12.5µl H_2O, and incubated at 37°C for 15 min. 0.5 µl M-MLV Reverse transcriptase (200U/µl, GIBCO) was then added to this mixture and the reverse transcription was carried out at 37°C for 2 hours.

Tailing

0.5µl of the first strand cDNA pool was diluted with 12µl H_2O, boiled for 5 minutes and chilled on ice. 4µl 5x TdT Buffer (Gibco), 3µl 5mM $CoCl_2$, 1µl 10mM dCTP are then added, and the tailing mixture was prewarmed at 37°C before 1µl terminal deoxynucleotidyl transferase (TdT) (25U, GIBCO) was added. Tailing was performed at 37°C for 30 minutes.

PCR

1µl tailing mixture was used as template for the following PCR amplification. A second internal primer downstream to the primer used in cDNA synthesis was used as one of the two primers. An adapter with a poly G stretch and multiple restriction sites was used as another primer. 10µl PCR reaction was loaded with 2µl 6x loading buffer onto a 2% agarose gel. The southern blot was hybridised to detect the position of the weak product. The signal was cut out from the gel and reamplified. The product was cloned into pBluescript and sequenced.

With marathon RACE kit (CLONTECH)

All procedures were carried out according the manual of the manufacturer.

Part III:

Analysis of 19G11 sequence

Computer analysis based on sequence viewing

Nucleotide composition

The nucleotide composition of sequence 19G11 was analysed with the program COMPOSITION and NUCWEIGHT of HUSAR DNA analysis package from DKFZ, Heidelberg (http://genius.embnet.dkfz-heidelberg.de).

19G11 consists of

A: 9,420 27.7%
C: 7,778 22.8%
G: 7,309 21.5%
T: 9,547 28.0%

This composition is not especially GC-rich. A considerable proportion of coding sequences might be expected.

Repeats

To check for repetitive sequences in the 19G11 fragment, the TANDEM function of HUSAR package and a Repeat Masker2 program (http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker) were used for analysis. 916 bp (2.69% of the whole length of 19G11 fragment) were recognized as repetitive sequences. These repeats are listed in Table 1.

The Pao element was originally identified in the silkworm *Bombyx mori* (Xiong et al., 1993). Pao-like elements belong to neither the *gypsy*-Ty3 group nor to the Ty1-*copia* group of retrotransposons. They were proposed to constitute a third group of retrotransposons. In contrast to the other groups, few Pao-like elements have been isolated and characterised (Abe et al, 2001). *ninja* was the first PAO-like element identified in *Drosphila simulans* and also the first one in *Drosophila* (Ogura et al, 1996). The sequence of the Pao-like element of 19G11 is the first evidence for the existence of PAO-like elements in *Drosophila melanogaster* genome.

DNAREP1_DM is one of the most prominent repetitive sequences in the *Drosophila melanogaster* genome. It carries several thousand copies of DNAREP1_DM. Noticeable conservation of the termini and multiple internal deletions were observed for

DNAREP1_DM (Kapitonov & Jurka, 1999).

Classification	numbers	length	percent
Retroelements	1	64 bp	0.29 %
8702 8765 PAO-type			
Simple repeats	8	377 bp	1.11 %
8329 8383 (TG)n 13717 13767 (CA)n 13810 13863 (TA)n 9714 9765 (CAG)n 12092 12141 (CAA)n 12869 12905 (CAG)n 29871 29909 (GGA)n 33345 33383 (CTG)n			
Low complexity	3	134 bp	0.39 %
3907 3927 AT rich 15033 15123 A rich 25564 25585 AT rich			
DNA elements	1	241 bp	1.11%
1557 1761 DNAREP1_DM DNA 1825 1860 DNAREP1_DM DNA			

Table1: Repetitive sequences in the 19G11 fragment.

Three di-nucleotides repeats and five tri-nucleotides repeats were identified in 19G11 fragment. The repeat density is 0.088 di-nucleotide repeat/kb and 0.147 tri-nucleotide repeat/kb. According to the statistics of Katti et al. (Katti et al, 2001), there are 2923 di-nucleotide repeats and 2367 tri-nucleotide repeats in chromosome arm *3R* of *D.melanogaster* (27.86 Mb). The average density would be 0.105 repeat/kb and 0.085 repeat/kb for the di-necleotide and tri-nucleotide repeats, respectively. The 19G11 fragment has, therefore, approximately the average distribution of the simple DNA repeats of the *D.melanogaster* chromosome 3R.

Frames searching

Possible open reading frames in the 19G11 sequence were searched with the program FRAMES of HUSAR. Open reading frames are plotted as boxes bordered by potential start and stop codons. Potential start codons are shown as short lines that extend above the box and potential stop codons as short lines that extend below the box.

Xpound

The structure of nucleotide sequences in exons, introns and between genes can be usefully modelled using a conventional and straightforward probabilistic model as in XPound (A. Thomas & Skolnick 1994). The 19G11 sequence was analysed with the XPOUND function of HUSAR, the following result in Fig 2 is presented in terms of the probability that each base in a sequence is coding.

The shortcoming of this program is that the training set used to estimate the parameters that are needed for exon prediction consists of only 159 sequences of annotated human genomic DNA. Therefore the prediction of sequences for *Drosophila* are almost certainly even less reliable than for human ones. However, as more substantiated programs are not available, the application of this program provides a useful method for analysis.

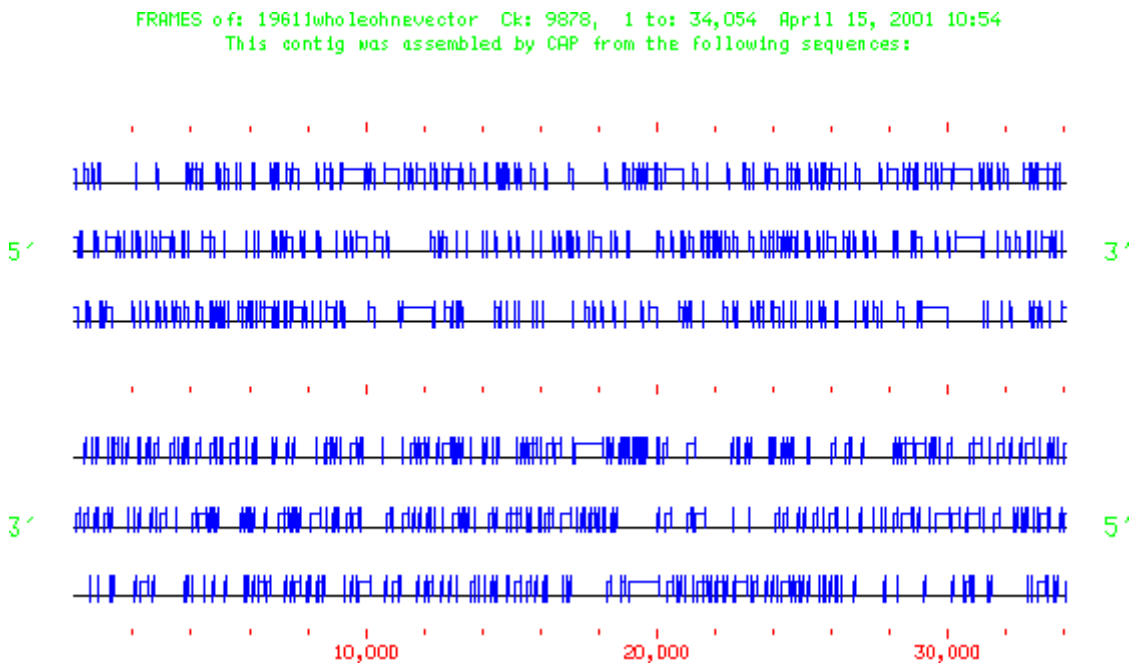


Fig1: FRAMES analysis of 19G11 sequence.

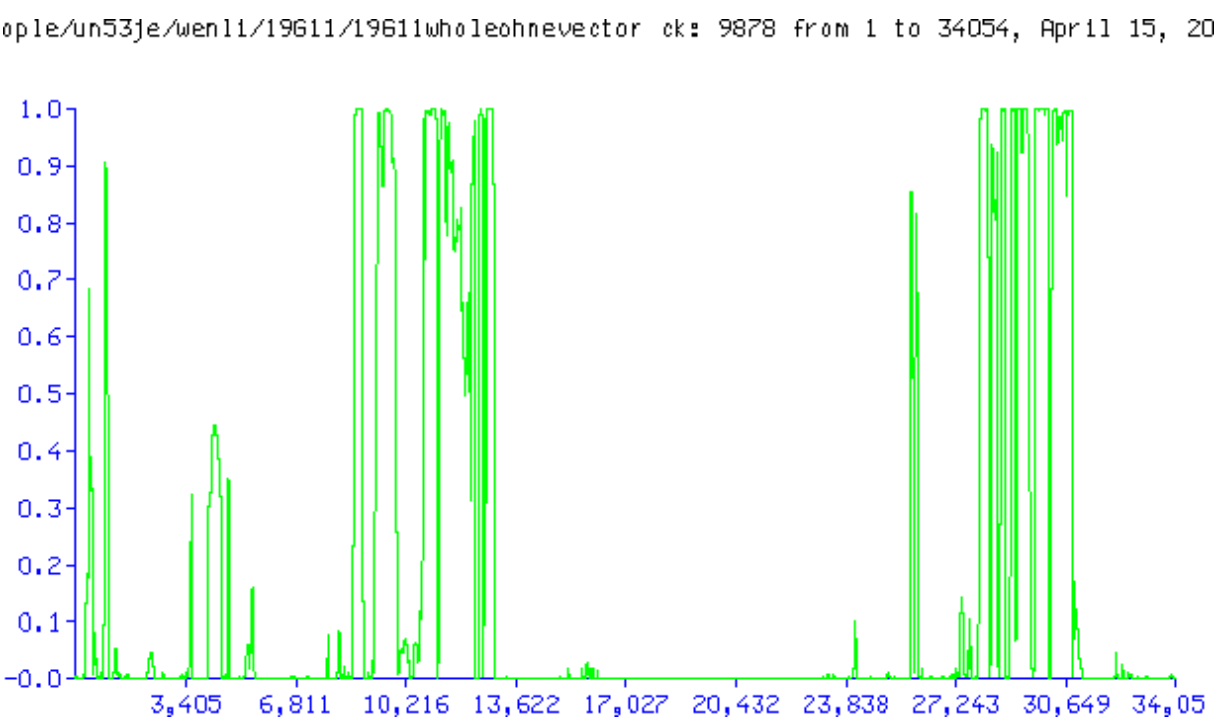


Fig2: Xpound analysis of the 19G11 sequence.

Grail (Gene Recognition and Assembly Internet Link)

Grail is the most widely used program to analyse genome data in searching for coding exons and genes. The Grail link from the site http://compbio.ornl.gov/Grail-1.3 was applied for the 19G11 sequence.

The following list summarizes the exons predicted by Grail. They are illustrated and compared to the EST fragments at the section: sequence analysis methods reviewing.

Exons predicted by Grail:

foward (same orientation as the numbered 19G11 strand in AppendixII)

Start	End
4180	4687
9313	9963
10621	11156
12476	12577
12610	12938
27919	28256
28330	28520
28589	28765
28916	29090
29661	30099
30178	30855
33773	34053
22866	22980

```
reverse (opposite orientation as the numbered 19G11 strand in AppendixII)
   33813   33991
   17595   17793
   16930   17324
   15802   16872
   15065   15246
   13947   15010
   12822   13022
   10935   11178
   28851   28868
   21091   21226
    2954    3329
    2484    2615
    2240    2419
```

Computer analysis based on EST-searching

Grail and Xpound are both gene-forecasting programs based on consensus pattern searching. They only predict the most PROBABLE genes from a genomic sequence, but cannot identify the REAL genes.

In contrast, Blast search (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) uses the growing EST (Expressed Sequence Tag) database, and looks for parts of genomic sequence that are indeed expressed *in vivo*. The genes identified with Blast are more probable to be real genes as the DNA sequences are at least expressed at the RNA-level. As the EST database has already been developed since years, a huge amount of RNA fragments from different tissue sources and developmental stages are sequenced and documented. One would expect that most physiologically expressed genes could be at least partially found back in the EST database. It cannot be excluded, of course, that some exceptional genes may be not yet represented. Some genes may be very rarely expressed in a very narrow developmental window or under certain stimuli so that their RNAs could not yet be identified.

BlastN search against the whole *Drosophila melanogaster* EST database was performed using the complete sequence of 19G11. A number of exons and parts of exons could be identified. Those exons included in the same EST piece were considered to belong to the same gene. In this way, 9 putative genes could be identified. Their structure and products are recorded in the following part. The nucleotides are numbered as in the complete 19G11 sequence (AppendixII).

Gene1

ESTs identified:

AI296812 from larva
AI295014 larva
AI260827 larva

All three ESTs contain the following two exons. Exon1 is upstream to Exon2. Exon2 ends downstream at nucleotide 64, followed by a canonical intron splicing donor site *gt*. This gene extends most probably further into the genomic region outside of 19G11.

Genomic organization:

```
     GGATCAAGTT GTGGAAGTTC GAGTTCAATT GGGTAGGGAG AGATACTGTG

51   GTGAATCACC CAC**CTGGACG TCGTTGCCCC GCACTCCGCA CTGATTATCA**
               tg
**101  CGGAGCAGGT TGCGATCGCT CTGGCTGAGA TTGCCGGATA GCAGGATTCT**
                                                          Exon2
**151  CATGAAGTAG TCACACTCCC GGATCGAAAT GCAGTGACCG GTGACACGGC**

**201  CACTCGGAAT TTTTGCAACA CAATTGATGG GCGGGAGTTG AG**CTGGAATA
                                                  ga
251  TCGGAGATCG GGGATCGTCC CACGTTAACT TAAGTTCACA GCCAGACTTG

301  TTTACGATCC CGTATTCCGT ACACACTCAC **CATTCGCGCC CAATGCCAGC**
                                   tg                     Exon1
**351  GAGCCGACAA CAACAAGTAG AGCTGGAAAA CTTCCCATGC CGGACCGGAT**
                                        S  G  M
**401  GAGCGGAACG AATGAGACTG CGATCGAAGC CGATCAGATA CGATACA**GCT
       *
451  TGAGTCTGAG ATTAACATTA AGTTCGGCCG GTGCCTGTGA ATTACCAGAG
```

Gene Product

The sequence of the both exons was translated in three frames, one of them contains a long open reading frame extending from the middle part of Exon1 to the end of Exon2. The first Methionin and an in-frame upstream stop codon are indicated in upper section (genomic organization). The putative product has the following sequence:

```
> putative product of gene 1

    1  MGSFPALLVV VGSLALGANA QLPPINCVAK IPSGRVTGHC ISIRECDYFM

   51  RILLSGNLSQ SDRNLLRDNQ CGVRGNDVQV
```

This segment could be the N-terminal part of the whole gene product, as there seems to be more downstream exons of the same gene outside of cosmid 19G11. BlastP search against Swissprot database was performed with this sequence. The segment from a.a. 31 to a.a. 80 has 30% identity and 44% similarity with the *Drosophila* protease *Easter* precursor from the trypsin family (Misra et al. 1998). To check the significance of this similarity, the complete contig of the three identified ESTs was translated into the following peptide:

```
> translation of the complete EST contig

        M GSFPALLVVV GSLALGANAQ LPPINCVAKI PSGRVTGHCI
SIRECDYFMR ILLSGNLSQS DRNLLRDNQC GVRGNDVQVC CPSTAGLGAL
THPLLPSDCG KVRWQRSNDT DTRIREFPWL ALIEYTRGNQ EKIHACGGVL
ISDRYVLTAA HCVAQAATSN LQITAVRLGE WDTSTNPDCQ YHEDSKVADC
APSYQDIAIE ELLP
```

This may be not yet the complete protein product of this gene, as the stop codon was not yet identified. Another BlastP search was done with this sequence and a trypsin consensus domain (Rawlings & Barret, 1994) was identified. This gene codes therefore most likely for a new member of the trypsin family in *Drosophila.* It has 38% identity and 53% similarity with *Easter* precursor.

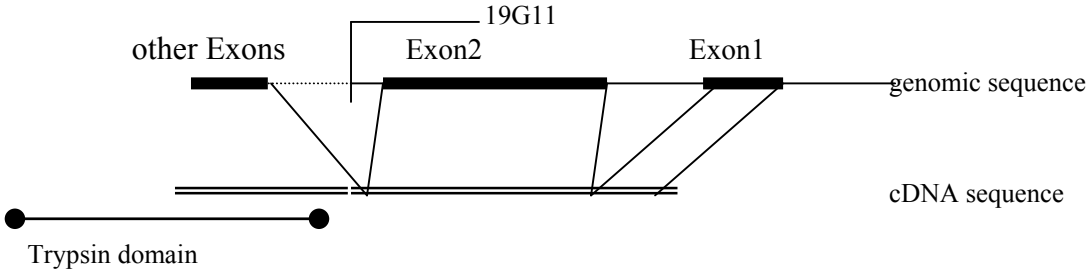


Fig3: genomic organization of the putative gene 1 of 19G11.

Gene2

ESTs identified:

BF504294	from	adult testis
BF497782		adult testis
AA440145		embryo
AA949568		embryo
AA440941		embryo
AA536609		embryo

The four ESTs from embryos (those ESTs with AA in their terminology) and the two ESTs from adult testis (their terminologies start with BF) seem to represent two different transcripts. They share the common 5'exon (Exon1), and their second 5'exons (Exon2) start at the same nucleotide G_{8516}. Exon2AA (revealed by embryo ESTs) ends at nucleotide 8873, and Exon3AA starts at position 9313. But Exon2BF (represented by the testis ESTs) ends at around the position 8682-8692, and the next exon - Exon3BF - starts at a position around 9380. As no canonical donor and acceptor splicing sites could be found in either region, the exon/intron junctions could not be exactly identified. Unconventional splicing sites might be used in this case.

These two transcripts are probably created from alternative splicing of the same gene (the putative gene2 of 19G11). Considering the fact that both BF ESTs were obtained from adult testis and all AA ESTs from embryo, there might be two tissue-specific transcripts.

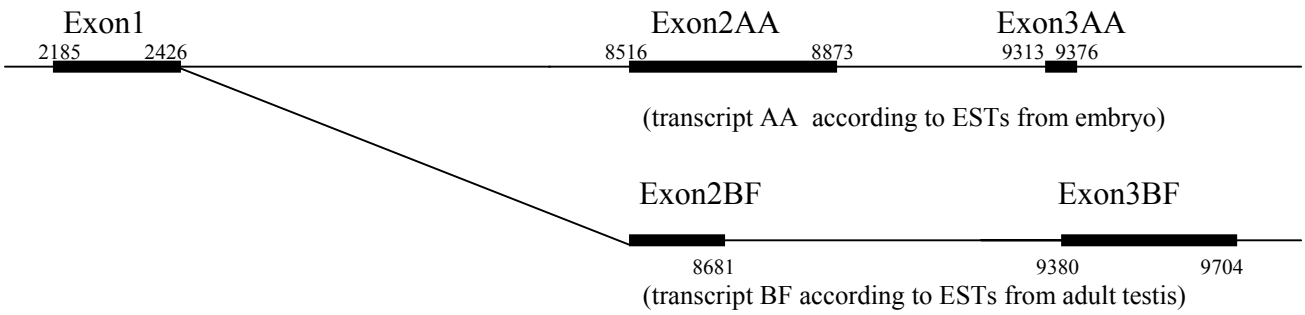


Fig4: genomic organization of the identified exons which belong to the putative gene2.

Gene product

Both transcripts according to the BF ESTs and AA ESTs were translated in all possible reading frames and the product peptides searched with BlastP program. No homology with known proteins could be identified.

Genomic organization:

2051 GTTTGTATAA ATAGCAGCAA AATAGTGATA TTCGATAGAT CATGTGAATA

2101 AATATTTATT TTTCTTCAAA AATATCGATA TATTGAAATG TAGATAAATT

2151 TGTCACATCC CTAGTTCGCC CGAATGGCTG CGTG**TGTGTC TGTGCGCGCG**

2201 CCTGTATTGT CCGCCATCTT GTCAGCCCGA CTTCATCGAA AATTACAAAT

2251 TTAATCGTTT AACGCGTTTT ATGCCCACTT AACACACCAG AAAGTGCTGC

2301 AGTACACATT TTCCCACAAA AAGGATATCG TCTTGAGTAC GCTGCGCTCA Exon1

2351 GCAAAGGGGG ATAAAATTGC ATTTGAAAGT GGAATTGTTG GTGCGGAGAA

2401 AAAATTGTGC AGCAAAAAAT TCCCAG*gt*CT GTGCTGTATG TGTGTGTGAG

2451 AGGCAGGCCA GGGTTGCCGA TCCTCCTATT TATATGTGCA TAAAATAGAT

AA ESTs:

8451 ACGATGACGT CTTCCTTTGA ACTTTAAACC
CATTGGTTTT GTAACGCTTT

8501 TCTTTTCGAT TGC*ag***GTCAA GTAACGAGAG ATAACAATAG AGCAACAAGA**

8551 GCAGCAGCAA AAACAACAGG AGCCGAAAGC ACTGGAAACA AAAGCGGCAA Exon2AA

8601 CGGCTTCACA TTGGACATGT CATGTCAGCA AGCCTCCAGA TTCCATTCAG

8651 CCACCGCTAC AGCCACATTA GCCAAGAGCA CAGCCACCAG AAGGATTGCC

8701 ACAGCAGCAG CAGCCGCAAC AGCAACAGCG ATAGCCGCAG CAACAGCCGC

8751 AGCAGCAGCA GCAACGTGAC GCCCGTGGAG AGCATTGCCG GCAAGACGAC

8801 GTCCGAGGAC TCGGATCCCT ATGCCTTCAC CGAGACTGTG GCCGTCACAC

8851 CACCCATTCT ATTCAATGCA CAG*gt*AAAGA AGCAGTCGGA AATTTCCTAA

8901 ACCCCGTCCA TATCAGGATT TGCATAGATC AAAAAATTGT AGTATTTCTG

8951 TAAGAAAACT GTATACATAT GGATGGGGGT TTCTTTGTAG ATAAGATCAT

9001 CTGATTTTTA TACATAAAAC TAGGTTCTTT GTCAACCGTA GTATTTAGCT

9051 TACGTTACCA CAGTTTTACC ACCATTATTT TGAAACTTGT TATTTGTGAG

9101 CCTTTCAAAA CACTTTCAAG TGTATGCTAA TCACATGGTA AATAAATTCT

9151 GGAATTTTTA TTGCAAAAGA AATTGGTGAT AATTTCAGAA CCTGAACTTC

9201 AATATGAACA GGTTCCAACT TTTGATATAT GGTATTATAT TACGCTCGCT

9251 TTATGTACAC AACCGTATTC TATGAAATTC ACTTACCCAA CTGTTTTGCT

9301 TACATATTGC *ag***AAATCGAG AGCCCGCCTA ACCGACAGCA ATAGAGGCAG Exon3AA**

9351 CAACAAGAGG CAGACGGCAG CAACAGCTGC GGCCAACAGA AAGGCGAACC

9401 TGGTGGCCCA ACTGAGTGTC ACAGAGGCAG CAAAGGCGCA GGCGTCTTTG

BF ESTs:

8451 ATGACGT CTTCCTTTGA ACTTTAAACC
CATTGGTTTT GTAACGCTTT

8501 TCTTTTCGAT TGC*ag***GTCAA GTAACGAGAG ATAACAATAG AGCAACAAGA**

8551 GCAGCAGCAA AAACAACAGG AGCCGAAAGC ACTGGAAACA AAAGCGGCAA Exon2BF

8601 CGGCTTCACA TTGGACATGT CATGTCAGCA AGCCTCCAGA TTCCATTCAG

8651 CCACCGCTAC AGCCACATTA GCCAAGAGCA CAGCCACCAG AAGGATTGCC

8701 ACAGCAGCAG CAGCCGCAAC AGCAACAGCG ATAGCCGCAG CAACAGCCGC

8751 AGCAGCAGCA GCAACGTGAC GCCCGTGGAG AGCATTGCCG GCAAGACGAC

8801 GTCCGAGGAC TCGGATCCCT ATGCCTTCAC CGAGACTGTG GCCGTCACAC

8851 CACCCATTCT ATTCAATGCA CAGGTAAAGA AGCAGTCGGA AATTTCCTAA

8901 ACCCCGTCCA TATCAGGATT TGCATAGATC AAAAAATTGT AGTATTTCTG

8951 TAAGAAAACT GTATACATAT GGATGGGGGT TTCTTTGTAG ATAAGATCAT

9001 CTGATTTTTA TACATAAAAC TAGGTTCTTT GTCAACCGTA GTATTTAGCT

9051 TACGTTACCA CAGTTTTACC ACCATTATTT TGAAACTTGT TATTTGTGAG

9101 CCTTTCAAAA CACTTTCAAG TGTATGCTAA TCACATGGTA AATAAATTCT

9151 GGAATTTTTA TTGCAAAAGA AATTGGTGAT AATTTCAGAA CCTGAACTTC

9201 AATATGAACA GGTTCCAACT TTTGATATAT GGTATTATAT TACGCTCGCT

9251 TTATGTACAC AACCGTATTC TATGAAATTC ACTTACCCAA CTGTTTTGCT

9301 TACATATTGC AGAAATCGAG AGCCCGCCTA ACCGACAGCA ATAGAGGCAG

9351 CAACAAGAGG CAGACGGCAG CAACAGCTG**C GGCCAACAGA AAGGCGAACC**

9401 TGGTGGCCCA ACTGAGTGTC ACAGAGGCAG CAAAGGCGCA GGCGTCTTTG

9451 GCAAGCAACA ACACAACGAA TTTCCATCAT GTCACGCAAT CTCAGAGACA

9501 GTCGACGGCG CTGCAGTTGC AATTGCCACT GCAATCCCAG TCACAGTCGC Exon3BF

9551 AGGCCTCGCC GAAGCGGGCC ACCAACGTGT GCATAGTCCG CCCGCAGCAA

9601 CAGCAGCTGG AGAAGATAGC CACCTCGGAG TCCTGCCAGT CGCCGGCAGC

9651 ACCACCACCG CTTTACGCCC ACACTCCATC GCTGTGGCAG ACGCCGCTGC

9701 TCATAGACAA TGGGCAAAAG CAACAGCTCC TCCAGCAGCA GCATCAGCAA

Gene3

ESTs identified:

AA941394 from embryo
AA263422 embryo

The two ESTs overlap with each other and revealed a single exon of over 800bp.

Genomic organization

```
 9651  ACCACCACCG CTTTACGCCC ACACTCCATC GCTGTGGCAG ACGCCGCTGC
 9701  TCATAGACAA TGGGCAAAAG CAACAGCTCC TCCAGCAGCA GCATCAGCAA
 9751  CCGCAACAGC AACAGTCCGT TGCTATTGCG TTGGTCAGTC CGCCCACATC
 9801  GCCCGCCTCA TTACCTTCGC CCACTCTGCC GCCTGCCACC GCTGCAAGTG
 9851  ACCGCCATGG TGGCACCGAT TTCCGTATCG CCCAAGGGTG GATTACCTTT
 9901  GCCGCCATCG AAGTTCCATC ACACCACACC TGGCGCAACA TCTGCAGAAG
 9951  GTGGAGTGCT TGAAAAAAAA GAAATCCTTG CCACTGGCCT GCCAAAACAA
10001  CAACAATAAC AGCAATTTGC CGAATAACAA CAATGTGGAG TCGCTTAAGA
10051  AACCGGTGGT GCAGGGAACG AGCTACAATC AGACTCATCC GCCACCGCTG
10101  ATGGTTTTCA ATACGGGAAC AGTTGCAGTT CCCGCGCAGA GTCCGCAGAC
10151  TGCTGCTCCA CAGAAACATT CCACCGGCAA CAGCGTAGAT GACAGCGATC
10201  TCAACGAGAT ACCCGTCAAT GTTATCTTCA GAAAGCCGCA AGAGGCAGGC
10251  GGACGGCGAA AACAGGTGGA CCGGGAGGAT TAAGTGCACC TGTTTCGGGA
10301  ACGCCGCAAA CTCGTCCAGC TGAAGTGAAA ATGGTGACTC CTCTCACGCC
10351  GCCCACTCCA CCAGAGATGA GCGCACCGCC CCCTGTAGCG CAAATGCAAC
10401  CCCCGCAGAT ACCCACGTCT TGTGTTCCAG CTTTAGCTCC CAGCTTCAAA
10451  GTGTCGTCAC CAGCAGTTCT CAGCCCGAAG GTGATCTCAC CAGCTCCTGC
10501  AAGCCCAAAG CTCTTGTGTC CGACAGCACC CGCTTCAACG AACTCACTGC
10551  AAATTGCGCC AAAAGTGTTC CAGCCACTAC AACCTCATCT GCACCAGCAT
10601  TAGCCACGAA ATCAGAACAA ATGTCTTCCA AAGTGGCCAA TTTAAACGCC
```

Gene product

The exon sequence was translated in all 6 reading frames as no orientation information could be obtained from the intron sequence. An open reading frame exists, extending through the entire exon.

> putative product of Gene 3

```
     QHQQPQQQQS VAIALVSPPT SPASLPSPTL PPATAAVTAM VAPISVSPKG
  51 GLPLPPSKFH HTTLAQHLQK VECLKKKKSL PLACQNNNNN SNLPNNNNVE
 101 SLKKPVVQGT SYNQTHPPPL MVFNTGTVAV PAQSPQTAAP QKHSTGNSVD
 151 DSDLNEIPVN VIFRKPQEAG GAPKTGGPGE LSAPVSGTPQ TRPAEVKMVT
 201 PLTPPTPPEM SAPPLLAQMQ PPQIPTSCVP ALAPSFKVSS PAVLSPKVIS
 251 PAPASPKLLC PTAPASTNSL QIAPK
```

BlastP search was performed with this peptide sequence, no homology with known protein could be identified. The PROSITE function of HUSAR was applied to analyse for known protein motifs in this peptide. It has five potential Protein kinase C phosphorylation (Woodget et al., 1986; Kishimoto et al., 1985) sites, three potential N-meristoylation (Toeler et al., 1988; Grand, 1989) sites, two potential Casein kinase II phosphorylation (Pinna,1990) sites, two potential N-glycosylation (Miletich and Broze, 1990; Gavel & von Heijne, 1990) sites and one potential cAMP-dependent protein kinase phosphorylation (Glass et al, 19986; Glass & Smith, 1983) site.

```
                                                  PKC PHOSPHO SITE
QHQQPQQQQSVAIALVSPPTSPASLPSPTLPPATAAVTAMVAPISVSPKGGLPLPPSKFH
                                                                 ASN GLYCOSYLATION
                           ASN GLYCOSYLATION   MYRISTYL
               CAMP PHOSPHO SITE         PKC PHOSPHO SITE
HTTLAQHLQKVECLKKKKSLPLACQNNNNNSNLPNNNNVESLKKPVVQGTSYNQTHPPPL

                           CK2 PHOSPHO SITE       MYRISTYL
MVFNTGTVAVPAQSPQTAAPQKHSTGNSVDDSDLNEIPVNVIFRKPQEAGGAPKTGGPGE

      MYRISTYL           CK2 PHOSPHO SITE             PKC PHOSPHO SITE
LSAPVSGTPQTRPAEVKMVTPLTPPTPPEMSAPPLLAQMQPPQIPTSCVPALAPSFKVSS
            PKC PHOSPHO SITE
    PKC PHOSPHO SITE
PAVLSPKVISPAPASPKLLCPTAPASTNSLQIAPK
```

Gene4

ESTs identified:

AA803855 from ovary
A I512756 ovary

Both ESTs revealed two exons, with the intron between them containing canonical donor- and acceptor- sites.

Genomic organization

```
10851  CTGCAGACGC AGCTTGGACA CTCTGAGTGA GAACGAGTCC TTTTCAGTGG
10901  ATTCGCCCTA CTGCCTACAA CAGCATTGGC TGCACTCCGG CTTTAATAAT
10951  AAGTCCCACG ACGCGCCACT ATCGCAAAGT AACCGGCGTG AGGATCGAAT
11001  TGCTGTTCGT AAAGGAGCCC TGAGGCGACA AGCGCTGCAG CTACTGTCTA   Exon1
11051  CGCGCTCCCT GCAGGAGCTT CCCATGCGAG CGGCCAAACA GCGGCTGCAG
11101  TGCGTCCAGA ACATGCTGAT CAAGTACCAA GATCACGCTG GCCAGCAGCA
11151  GGGAATgtAA GTTTAATGTA GTGTAATCAT TACCCAATTT ATACTAATTG
11201  TACTGTTTTC CGTTAAagTG GAATTGGCAA CCACTGTCTG GTGGCCAGCT
11251  GCAAGCAGCC CACGCTTAGT ATGGCAGCGC ACTGCGAGCG TCACATCGTG
11301  AACAACAGCA CCCAGCAGCT TTTCCAGCCT TGCGTCGCCT GGCGGATGGA
11351  TGGAACCGCT TGCCAGGCTC CTGTCTTCGA TGTGCTGCAC ACATTGGCGC  Exon2
11401  TCTGCAAGGT GCACAGCCAC CTGCGTTCTG GTATGGATGG AGCCCGTCCA
11451  GCATCGAAAC AGCCGCCCGT TAGTTTACCA GTCGCTGGAG TGGCGACTCT
11501  TTATGTGCCT GTGAAGCAGC AGCGAAAGCG CAAGGCTAAT ACGAATGCCG
11551  TGGCTCGTCC TCAGAAACGT GGTAGGAAGC CGGCGAATGA ACCGATTGCT
```

Gene Product

One main open reading frame extends through the entire cDNA sequence. No similar peptides could be identified by the Blast P program. The putative protein sequence is:

> putative product of Gene 4

```
        QQHWLHSGFN NKSHDAPLSQ SNRREDRIAV RKGALRRQAL QLLSTRSLQE
  51    LPMRAAKQRL QCVQNMLIKY QDHAGQQQGI GIGNHCLVAS CKQPTLSMAA
 101    HCERHIVNNS TQQLFQPCVA WRMDGTACQA PVFDVLHTLA LCKVHSHLRS
 151    GMDGARPASK QPPVSLPVAG VATLYVPVKQ QR
```

The PROSITE function of HUSAR was again applied to analyse for known protein motifs in this peptide. It has four potential N-meristoylation sites, three potential Protein kinase C phosphorylation sites, three potential N-glycosylation sites and two potential Casein kinase II phosphorylation sites.

```
                    PKC PHOSPHO SITE          CK2 PHOSPHO SITE
          ASN GLYCOSYLATION                PKC PHOSPHO SITE
QQHWLHSGFNNKSHDAPLSQSNRREDRIAVRKGALRRQALQLLSTRSLQELPMRAAKQRL
                   MYRISTYL                        ASN GLYCOSYLATION
               MYRISTYL       PKC PHOSPHO SITE  ASN GLYCOSYLATION
QCVQNMLIKYQDHAGQQQGIGIGNHCLVASCKQPTLSMAAHCERHIVNNSTQQLFQPCVA
                                    MYRISTYL
                                MYRISTYL
                               CK2 PHOSPHO SITE
WRMDGTACQAPVFDVLHTLALCKVHSHLRSGMDGARPASKQPPVSLPVAGVATLYVPVKQ
QR
```

Gene5

ESTs identified:

AW942247 from embryo
A I 544155 embryo

Both ESTs revealed two exons from one gene. AW942247 shows ten extra nucleotides between 12823 and 12828 in Exon2 compared to the genomic DNA sequence. This is probably a sequencing artefact of AW942247, especially as AI 544155 overlaps completely in this region with the genomic sequence 19G11.

Genomic organization

```
 12401  TGTATATATA GAAGGAAGCT GGACAAATTC CCTAAGATTT TTTCAGTATT
12451  TTGACACGTT ATTTCTCTCT TCC
```
AGATGAT ATAGCATTGA ATGGCGCCCA

```
12501
```
CTTGCTGGAG GAGCACGATC TGGTAAATGT GTTCGACACG CTGTCAGACG Exon1

12551 **ATGCCTTCAA CGAGCTGTTC CAATCCG***gt*G TGTATTTAAC CACAATTTTA

12601 CTTGTTCAGC TTTTCATTGT ACTAATCCAT GTGTGCCTTC GCTGCCAACG

12651 CCACCGCCAT CGCCTTGGAC CTGTGCCTAT GC*ag***TGCAAC AAGCCGAGTG**

12701 **CGAGGCTATG GACCGGGCTT TGGACCGGGC CTTACAGCAG ACAATGGGCG**

12751 **GCTCGGCGGA CAGCGCCTTT CTTAACGATT TCCTGGACGT CGGCGACGAT**

12801 **CTGCTGGCCG ATGCTGTGAT GCACTCACCA AACACGTCCG GCATCGATGC**

12851 **TCCTCCCCTC TTTGGGGACA GCAGCAGCGG CGGTGGCAAT AGCAGCAGCA**

12901 **ACGGCGCCTC CGACATCCGG GGCTTGGTGC AGACCTAATT CCGGCATCAG Exon2**

12951 **GTCGGATTTA TGGCCTACAG AATTTACAGA TTCATTTAGA GAGACAGAGA**

13001 **GAGAGATTTT CAAGCTTGAT TTCCCATTCA CTTTTAGGCA GTTTTCGCTT**

13051 **AGAATTTCGG TATTTTCTTT TTGGCCATTT CTTACCTGCG ATTCTAGTTT**

13101 **GGCACAATGT TTCTATATGC AGCTTCAATG TTATGCATGC ATTCACCAGC**

13151 **AGATATGCAT GAATAAAATA GCATTCAAAA ACCATATAGT TATGCGTTAA**

13201 **TGTG**AAAAAC AAAAAAAAAA AACACACACA AAAAGTAAAA AATTTGAAAA

13251 AATGTCCAAA AATATTAAAT TCAACTTTAA ATCAACATTA AGAATGAAAT

Gene Product

Possible products from all three reading frames, which all include multiple stop codons, were searched with the BlastP program. No similarity was found with any peptide in SWISSPROT Database.

Gene6

ESTs identified:

BF487868	from	adult testis
BF490646		adult testis
BF500182		adult testis
BF502886		adult testis
BF503154		adult testis
BF500662		adult testis
BF497922		adult testis
BF487536		adult testis
BF488085		adult testis
BF493348		adult testis
BF503529		adult testis
BF499841		adult testis
BF491655		adult testis
BF488733		adult testis
AI134580		head
BF496020		adult testis

The ESTs revealed 7 exons from a putative Gene 6. All introns are surrounded by canonical splicing sites.

Genomic organization

15301 GACTCCCATT **CGTTTAACAA CATAATTTTC TCCTAGGAAA CAGTTAGTTT**

15351 GCCTTGACAC ATACTCCGAG TAACTCCCGA AAACAAAATA CAAAAAACAT

15401 CAGCAGAACC GCCGAATTAA GAAACCCGCT AATCCTTCCC AGGATTCTTT

15451 CAGTGCGTTT CGTCGAGTTG TGGTCGAGGA TTCAAATTCA AAGGTTATAT Exon1

15501 TGAAAATTAT TATTTTCTAT TTTGTTTTCC TTGCTCGACC ACCAACCCAA

15551 TCGCATCTAA TCGCAAGGAG CATTCAGTCC AGTGCAAAAG AGACAAAAAC

15601 TGACCAGATC TGGTCCGGAT TATCCCCGTT TTGCTAA*G* *gt* GATTGAGTGC

15651 CATGTGGCGG CATTTCCTAC TCTGGGTTGT CCGTATGCAG CGATTATTAC

15701 ATCATATCCT AGC*ag***GCGTG TCGTTTGCTG TCGGTTGTAA TGCATACAAA**

15751 TGCCTATCTG AGCTGCCGTC TCCCCGATTA GTCACTTTTT TTGTACGTTT Exon2

15801 GTAAGCTGCC GCCAGTTTTC AGAGTGGCGC CACGGGGATA CGTGGAATAG

15851 CGTG*gt*AAGT GGGCGCCACA TGCTCCACCA TCGACCCCAC TAACCGACTC

15901 CTCACCACC*a g***TTGTGCTCT ACGTATATTT TTATATCATC ATTGCGGAAC Exon3**

15951 CACAAAGCTC TCGACTACTT TCTAACTGAG GAACTGAATC AAAG*gt*GAGT

16001 TCAATTCAGC ATATTCTGTA TATTTGCGCT ATGACTCTTT CCCGCATAAA

.
.
.

22801 TGATATTTCA CATTTCTGCT TAGCTTTTGA AATAATTTTC TTTTTTTGTA

22851 ATATATTAAC TCTAGACTGG ACTTACCCAC TTTCGTTTC*a g***GACTCTCCT**

22901 TCAGACGCAG TACAATCTCC GACTGGCCGT CGATAGCGAA ACGCGGCGGT Exon4

22951 CGATTTGAGG CAACCAAGCT GATTAGTGTG *gt*GAGTATTA TGACTTGGAT

23001 GGATATGGAG CTTATATAAC TAGAGAAACT TCGCCGCCTT TGTCCTTTTA

.
.
.

27851 GTCATTAATT TAATCCAGTG ATTTGCCATA TCATATTGAC AGATACGTAA

27901 CGTCATTATT TTTCCT*ag***AA TAATATGTCC GTGAGTCGAG TGACTATGAT**

27951 GCGAAAGGGC CACTCCGGGG AGGTAGCACG CAAGCCCAAC ACTGTGGTGG

28001 TGTCGGTTCC ACCGCTGGTG AAGAAGTCCA GCAAGAGCCG CTCGTTCCAC Exon5

28051 TTCCGCTATC TGGAGCTGTG CCGGGCCAAG AATCTGACGC CGGTGCCGGA

28101 AATCCGCAGC AAGTCGAATG CGACCACCAC CTTTCTGGAG CTGTGCGGCG

28151 ATAAGCTGGC GGTCAGCGAT TGGCAGCTCC TAACCGAGGC GCTCCACTAT

28201 GATCTCGTGC TCCAGCATCT GGTGGTGCGC CTGCGACGCA CATATCCACA

28251 AA*gt*AGGTGA TCTTTGGTAG TCTGCTACTC TGTATGGAAT GTGAATTATA

```
28301  CCATTTCGTT TATTCATTTT TTTCGGCag C CAACATTGAT CCCATTGACA
28351  CCGAGAAACG AGCCCGACTT TTTCGCCAGC GGCCAGTGAT CTATACTCGC
28401  TTCATATTCA ACAGTTTGGT CCAGGCGATT GCCAACTGTG TTTCGAGCAA Exon6
28451  CAAAAATCTA AGTGTGTTGA AGCTGGAGGG ATTGCCATTG CAGGATGGAT
28501  ATATCGAGAC CATTGCCAAG gtGCGTTGAA AAGTTTTGCG GTCGGGATCA
28551  CGTTTCCAAG TACACACATA TATCCAATCC ATTCACag GC ACTGGCAGAC
28601  AACGAATGCC TCGAAACAGT GAGTTTTCGC AAATCCAACA TTGGCGATAA Exon7
28651  GGGCTGCGAG GTGGTGTGCA ACACAGCCAA ATACCTGAAT CGCATCGAAG
```

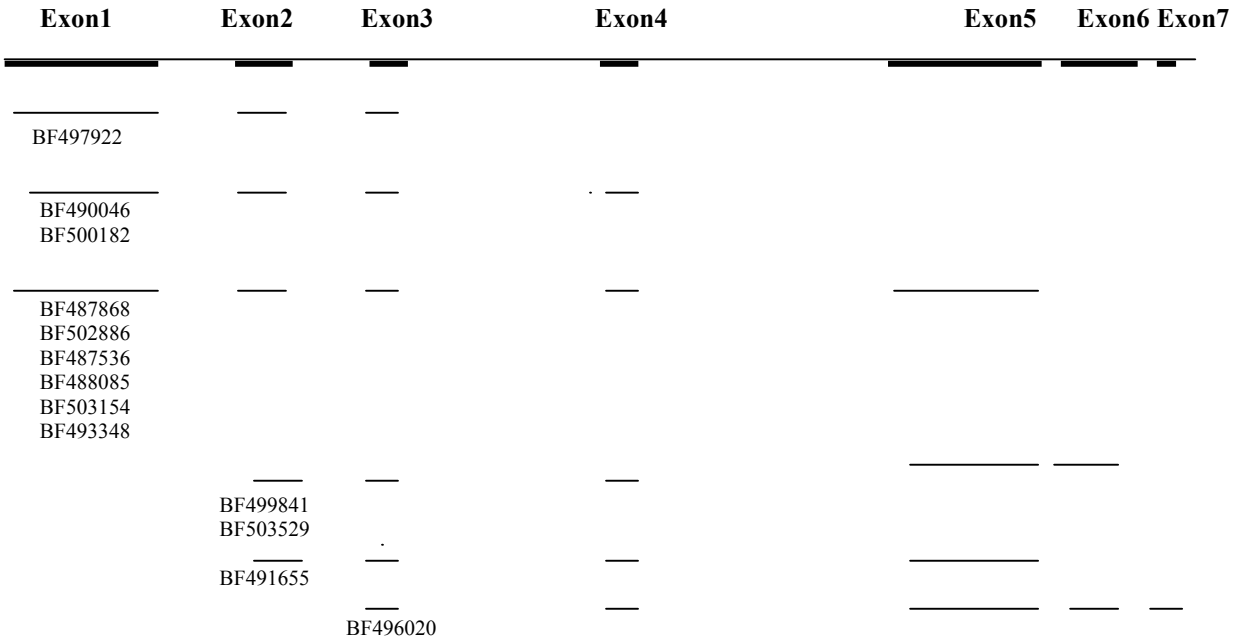


Fig5: Distribution of ESTs as proof for exons of Gene 6.

Gene Product

The cDNA fragment was translated in all three frames. All include in-frame stop codons in their sequences. Homologous peptides were searched for all translated peptide sequences with BlastP program, no homology was found.

Gene7

ESTs identified:

BF491687 from adult testis
AI946574 adult testis

Four exons were revealed from these two overlapping ESTs. The last intron between Exon3 and Exon4 has no canonical splicing sites, so the exact junction between them could not be definitely deduced.

Genomic organization

```
   28951  GAAGTCCCTC CGCTACCGTA GCGTCGATGT GAACACGATT GGCGGTCTGC
   29001  GCACGGTTTT GTTGGCTGAC AACCCGGAGA TTGGCGACGT GGGCATCCGG Exon1
   29051  TGGATAACCG AGGTGTTGAA AGAGGATGCT TGGATAAAAA gtACGTAGAG
   29101  TCCGAATAGT GCGATCCATA CAGTTCAAAT ACCCCACagA AATCGACATG
   29151  GAGGGCTGCG GCCTGACGGA TATCGGGGCA AATCTAATTC TCGATTGCCT
   29201  GGAGCTGAAC ACGGCCATTA CGGAGTTCAA TGTGCGAAAC AACGAAGGAA
   29251  TCAGTAAGTT CCTGCAGCGA AGTATCCATG ATCGTCTTGG CTGTTTACCA Exon2
   29301  GAGGAGAAAC AGGAGCCAGA GTATGATCTC AGTTGCGTCA ACGGGCTACA
   29351  GAGCCTGCCC AAGAACAAGA AGGTCACCGT CTCTCAACTG CTGTCCCACA
   29401  CCAAAGCATT GGAGGAGCAG CTCTCCTTCG AGCGAACGTT GCGCAAGAAG
   29451  GCCGAGAAGC TGAATGAGAA GCTTAGCCAC CAGCTCATGA GACCCGACTC
   29501  CAATCACATG GTTCAAGAGA AGGCCATGGA GGGAGGATCA CAAACAAACA
   29551  TTTCGAGGGA ATATGTGGCG CGGAATGATG TTATGCCAGA AGTCATCAAA
   29601  AAgtGGGCTA GCTCTCAATT CTCAATTCAA TGCCTATCTG ATTTGAGTAT
   29651  TTCTTCACag TTCCCAAAGC TACCGCCAGT CGCACTTCAA CCGGCTGGTC
   29701  AACAGTGCGG CCACCAGTCC CGAAGTCACA CCCCGCAGCG AGATTGTCAC Exon3
   29751  ATTGCGCAAG GAGCAGCAGC TGCAACGTCA ACAACCCCCA CCAATGGAGG
   29801  TCAAGCATCT TTCCTTGGAG CAGCAAATCC GAAATCTGCG CGACGTGCAG
   29851  AAAAAGGTGG ACTTGGACGT GGAGGAAGAG GAGGAGGAGG AGGAACAGCA
.
.
.
   30851  CCTAGGGGTC GTGACTTTAG GTTCCTCCAT TCAAAATCCC CGCAGCCATT
   30901  CGCAGCCGGA GCAGCGACAC CACAGCTGGC TGATACGCCG AGCAGTAACA
```

```
30951  ACAACAACAC CACTACCACC ACCATCACAC CCACAACAAA GCAGCCAACT Exon4
31001  CGATTCGATA GGATCCCCAC TGGGAGAGCA AAATTTGCAG TCATACTCCC
31051  GGGCAGTAGC AGTAGAAAAG GGAACCTGCT CCTTTTTGCA CGTCCTGCCA
31101  TCTGCGAATT CAAGCTGCAC CAAAAATGTA CAAAAATATA CTTTGGTCTT
```

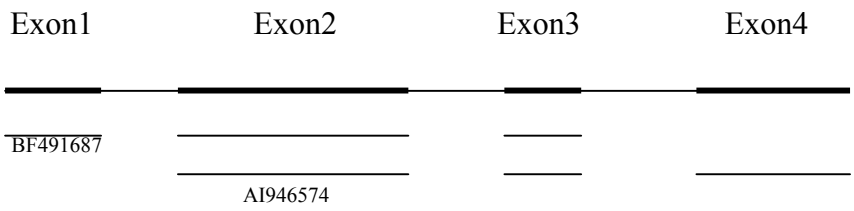


Fig6: distribution of ESTs as proof for the putative gene7.

Gene Product

The exons were translated in three reading frames and their product searched with BlastP program. No similar peptide was found.

Gene8

ESTs identified:

BE977916 from adult testis
BF491390 adult testis

Genomic organization

```
29801  TCAAGCATCT TTCCTTGGAG CAGCAAATCC GAAATCTGCG CGACGTGCAG
29851  AAAAAGGTGG ACTTGGACGT GGAGGAAGAG GAGGAGGAGG AGGAACAGCA
29901  GGCGGAGGAA AGTCAATCCG AGTCGGAGCT GCAGAACGAG GAGCAACAGC
29951  ATTACGAACA GCAAATGCAG GTCCAACGCA AACATCTCCA GGTGCGCAAG
30001  GTTCGCAGTG AGATTAAGTA TGTGGAAAAC AATCCCAAGG AGGCAGCCAA Exon1
```

30051 AAAGAATCGC GAGTCCAAGT CGGACCATGA GTTTGCCAAC GAGAGAGAT*g*

30101 *t*GAGTAGTAT CCACAAGATA AATTAACCAG TACTAGGAAT TTATTCTCTT

30151 GTACGATAAC CCATCTTATC CCCAT*ag***TTC AAGCTTAATC CCTCTGTGCA**

30201 ATTCGAGACG GACATTGGCG ACAATTTGAT GGTCAATCCT GGCCACCGAT

30251 ACGAGGGCGG CGGGGGCGAT ACGGGCTATG TCTACAACTA CGAGCATGAG

30301 CAGCAGCAGC AGCCAGTCAA GCGGGGCTAC GAGCACGGCT ATGTGGTGGG

30351 CGTGGGTGAC GGATCCCACA GGAGGCAGAG GCAATCTCAA CTGGTCGAGG Exon2

30401 CTTTGGTGCA AAAACGTGTC CCAGGCGCCA GCGATGGACA TGTGGCGCAG

30451 TTCGTTAGCA ATCTGGAACG ACAAGCGAAT GCTGGTAAAA CGGGGAAAAA

30501 GCGCCTTAAA CCTCGGCCTG AGGACGATCT TCAGGTACCA GTCGGCGACA

30551 TGCACATGGA GTCGTCGTAT ATGTCCCGCT CCGAAGAACT CTCCTCAACG

30601 GACGTTACGC TGGAGAACTC AGACTACGAG ACGGAGGCGA CGGACTCCAC

30651 GTTACTTAGT AGCTCGAAAT ACTCCTCCAT GCATGTCTTT GTGCGGCGCA

Gene product

The exons were translated in all three forward frames, one of them contains an open reading frame, which encodes a product with partial similarity with two known proteins.

> putative product of Gene8

AEESQSESEL QNEEQHHYEQ QMQVQRKHLQ VRMVRSEIKY VENNPKEASK

51 KNRESKSDHE FANERDFKLN PSVQFETDIG DNLMVNPGHR YEGGGGDTGY

101 VYNYEHEQQQ QPVKRGYEHG YVVGVGDGSH RRQRQSQLVE ALVQKRVPGA

151 SDGHVAQFVS NLERQANAGK TGKKRLKPLP EDDLQVPVGD MHMESSYMSR

201 SEELSSTDVT LENSDYE

The first underlined segment (aa 26 to aa83) has 28% identity and 48% similarity with an ATP-dependent protease from *Helicobacter pylori* J99 (Alm et al., 1999), while the second underlined segment (aa 119 to aa167) has 30% identity and 60% similarity with a transcriptional repressor CYTR (Valentin-Hansen et al., 1986) from *E.coli*.

Gene9

ESTs identified:

AI945488 from adult testis
BE977239 adult testis

These two ESTs revealed a single exon fragment, which overlaps with the Exon4 of gene7 and Exon2 of gene8. Probably, gene7, gene8 and gene9 are actually different splicing forms of the same gene, whose alternative exons span at least from the nucleotide 29001 to 31191.

Genomic organization

```
30401  CTTTGGTGCA AAAACGTGTC CCAGGCGCCA GCGATGGACA TGTGGCGCAG
30451  TTCGTTAGCA ATCTGGAACG ACAAGCGAAT GCTGGTAAAA CGGGGAAAAA
30501  GCGCCTTAAA CCTCGGCCTG AGGACGATCT TCAGGTACCA GTCGGCGACA
30551  TGCACATGGA GTCGTCGTAT ATGTCCCGCT CCGAAGAACT CTCCTCAACG
30601  GACGTTACGC TGGAGAACTC AGACTACGAG ACGGAGGCGA CGGACTCCAC
30651  GTTACTTAGT AGCTCGAAAT ACTCCTCCAT GCATGTCTTT GTGCGGCGCA
30701  AGCAATCGGA GTCCATGTCA CTCACAGAAG AGGCCGGCGA CGGAGATGCC
30751  GGCGGTGGTG GAGGCTCTGG CGATTTCGGC GACCAAAATG TCATATCCCC
30801  GGCCAATGTC TACATGTCCC TGCAGCTCCA GAAGCAGCGG GAGCAGAGCG
30851  CCTAGGGGTC GTGACTTTAG GTTCCTCCAT TCAAAATCCC CGCAGCCATT
30901  CGCAGCCGGA GCAGCGACAC CACAGCTGGC TGATACGCCG AGCAGTAACA
30951  ACAACAACAC CACTACCACC ACCATCACAC CCACAACAAA GCAGCCAACT
31001  CGATTCGATA GGATCCCCAC TGGGAGAGCA AAATTTGCAG TCATACTCCC
31051  GGGCAGTAGC AGTAGAAAAG GGAACCTGCT CCTTTTTGCA CGTCCTGCCA
31101  TCTGCGAATT CAAGCTGCAC CAAAAATGTA CAAAAATATA CTTTGGTCTT
31151  ACTAATTTCC ATTAAAGATT TATTATTTGT GTGCTCCCTA AGCGAGTTTG
```

Gene product

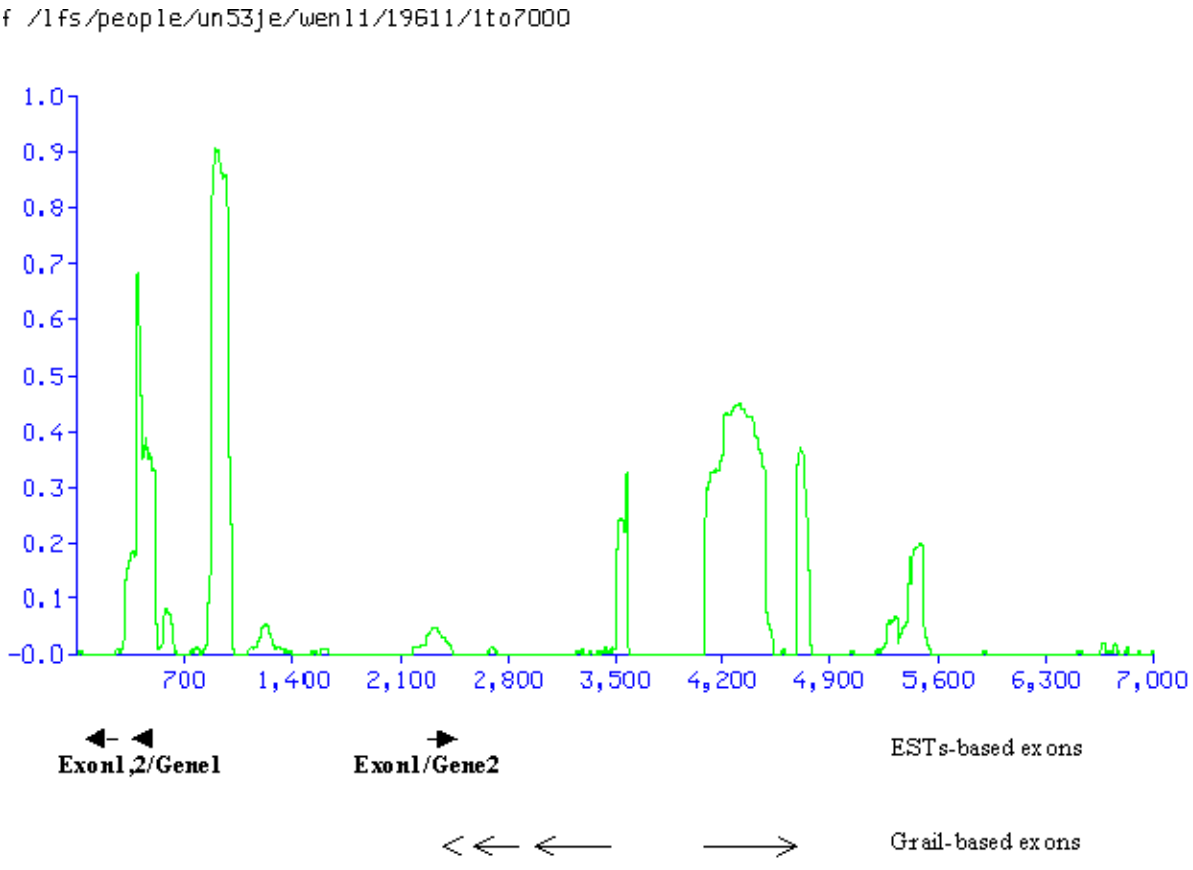
The exon was translated in all six forward reading frames, all possible products were searched in the Swissprot Database. No homology could be found with known peptides.

H4r and *Punt*

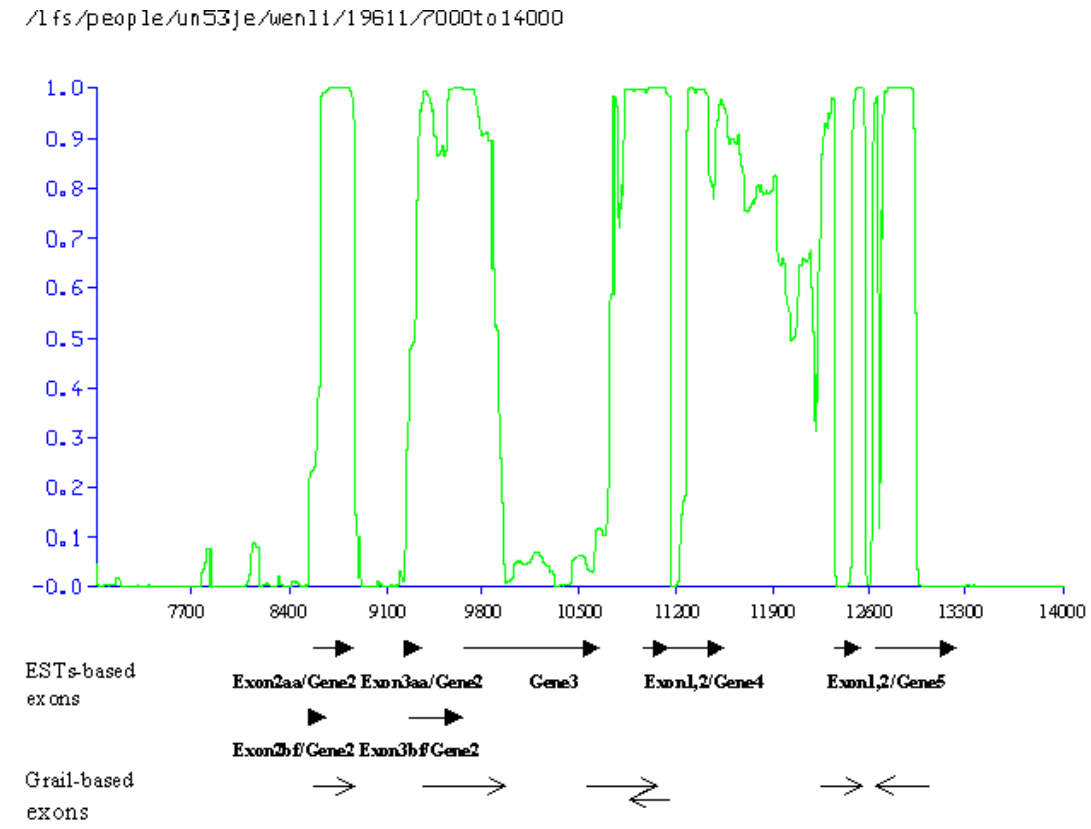
The H4r histone replacement gene and the *punt* gene are recorded in the EST-database as known genes. Their genomic structure was already fully characterized (Akhmanova et al., 1996).

Sequence analysis methods reviewing

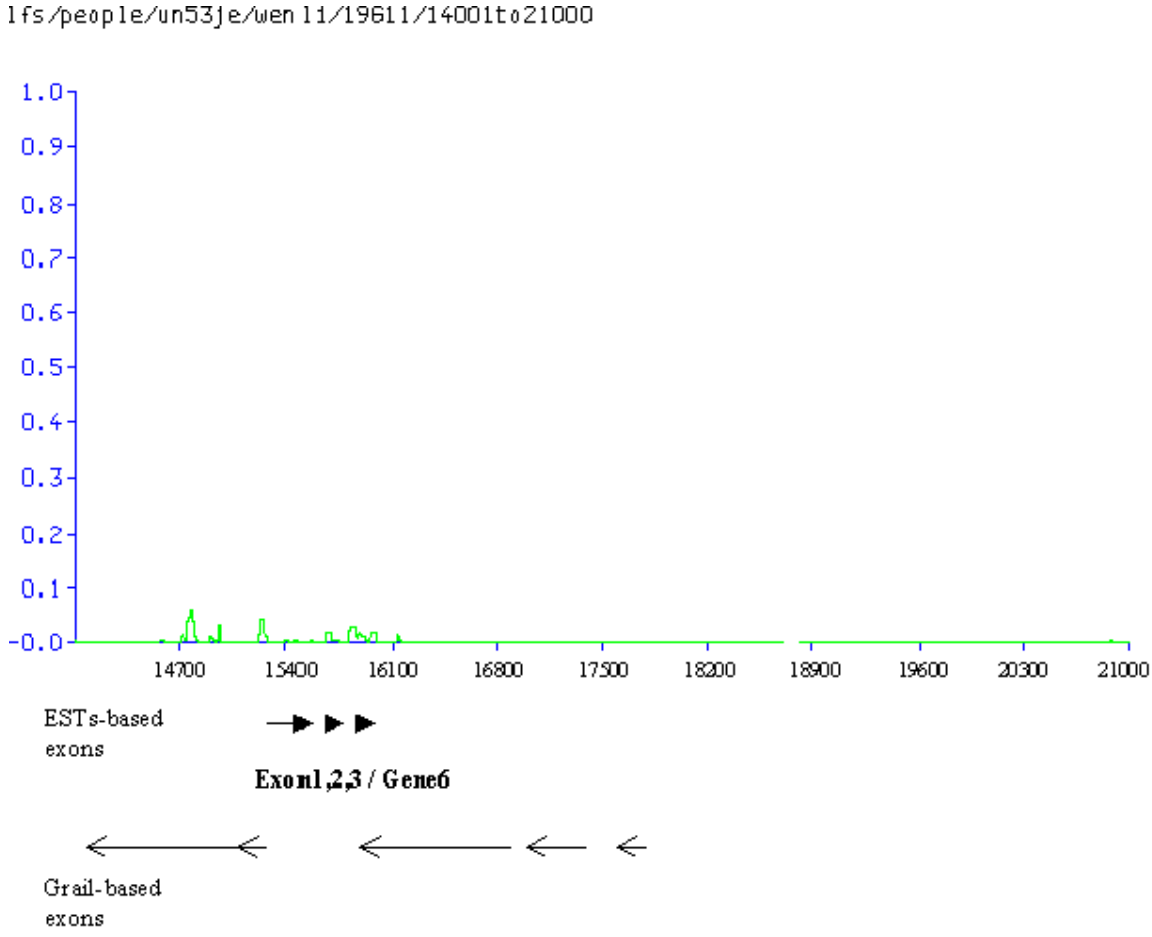
The results of EST search, Xpound and Grail analysis were compared with each other. The EST results were used to evaluate the other two programs in this case. Filled arrowheads represent the exons revealed by EST search and the open ones are those predicted by Grail (only the exons evaluated as excellent or good were considered). The whole sequence was divided into 5 parts to make the exon symbols long enough to be clearly visible.



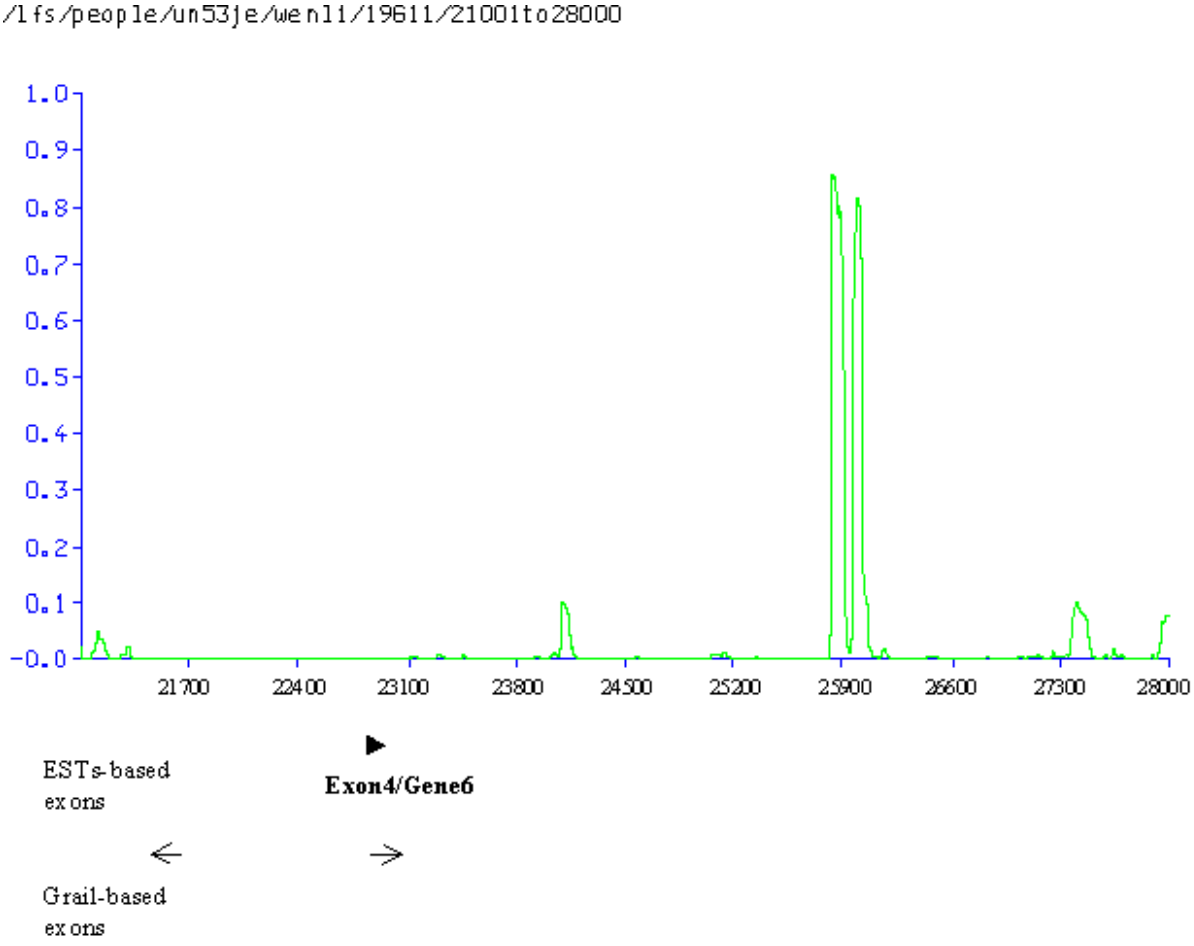
PartI: Base 1 to 7000.



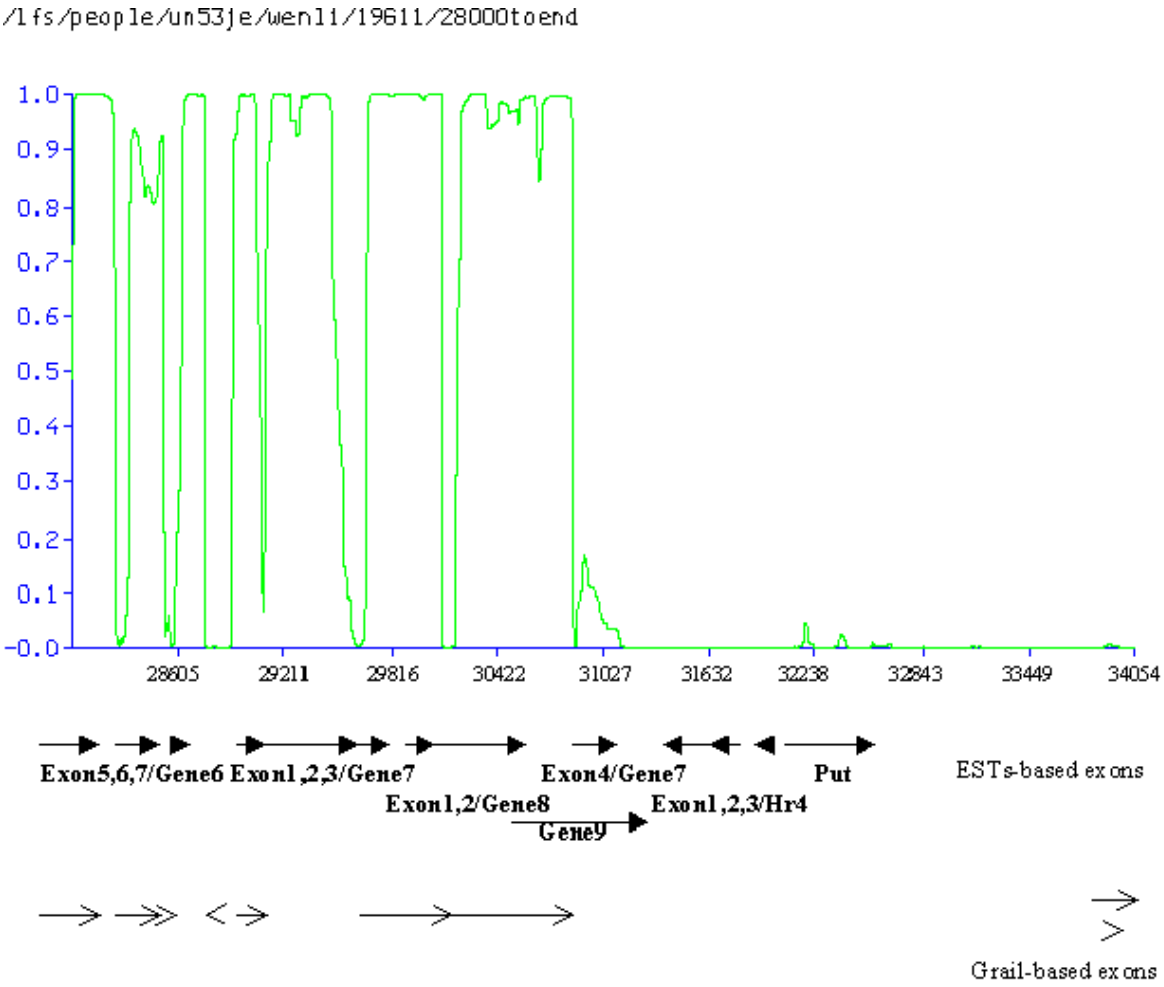
Part2: Base 7001 to 14,000.



Part III: Base 14001 to 21000.



Part IV: Base 21001 to 28000.



PartV: Base 28001 to the end (34054).

Conclusion

The already experimentally characterized genes histone replacement H4r and *punt* were used as calibrators to check the program Grail and Xpound. Grail didn't succeed to predict either of them; Xpound gave the *punt* exons a probability of c.a. 8%. H4r could not be predicted, either. These results reflect the present status of our methodology to analyse genomic DNA sequences. It must hence be concluded that the assessment for potential genes as made in this paper is rather preliminary, additional or other genes might finally be identified in the region of the genomic 19G11 sequence.

From the comparison of the predictions of the two programs and the EST data, it was also proved once more clearly that neither Xpound nor Grail could well predict possible exon

structures from a genomic sequence. But both managed to predict some exons correctly, keeping in mind that for most of the cases no experimental proof has yet been provided. They are useful choices for exon prediction when there's no or little EST data.

Also the EST data are of limited value:

1. ESTs are just segments of the cDNAs, whose 5' and 3' ends are often lost. Also some parts of the transcripts with more complicating secondary structures might be not represented in the EST database as they could not be properly reverse transcribed under the normal conditions to establish the EST libraries.

2. Some genes may be not yet represented in EST database because they may be very rarely expressed in a very narrow developmental stage or they are only expressed as a reaction to some special external stimuli so that their RNAs could not yet be identified.

Part IV:

P-element mediated excision at the *punt*-H4r locus, 88C8-10

Introduction

P-elements are widely used in *Drosophila* genetics. Since the identification of their existence and the clarification of their transposition mechanisms, several genetic tools have been developed taking advantage of P-elements (Engels, 1997; O'Kane & Gehring, 1987). Especially the possibility to achieve site-directed mutagenesis with P-elements makes *Drosophila* an excellent model system to study gene functions.

P-elements are bounded by 31-bp inverted terminal repeats and make an 8-bp target site duplication upon their insertion (O'Hare & Rubin, 1983). P-elements are DNA-intermediate transposons that move via cut-and-paste transposition (Kaufman & Rio, 1992). Their excision generates a double-strand DNA break in the chromosome (Engels et al, 1990; Gloor et al, 1991; Kaufman & Rio, 1992). The DNA breaks are repaired by host cells, which gives a variety of excision products. They can be repaired either by homologous recombination or by non-homologous end-joining. Homologous recombination leads to gene conversions when the double-strand break is made in pre-meiotic germ line cells (Geyer et al, 1988; Engels et al, 1990; Gloor et al, 1991; Nassif et al, 1994). Most often, this conversion produces the replacement of the excised P-element with the other P element from the same site on the sister chromatid, yielding an exact regeneration of the P-element at the excision site (Johnson-Schlitz and Engels, 1993). Gene conversion could also result in the precise loss of the P element if the allelic template site does not have a P-element (Engels et al, 1990; Nassif & Engels, 1993).

Although most of the transpositions cause the so-called "precise" loss of the P-element with a full restoration of the original allele, in some cases an "imprecise" loss may happen. In these instances, a small part of P-element is left at the original insertion site while the major part is cut away. In some other cases, P-element transposition caused also deletion of the sequences flanking the insertion site (Engels et al, 1990; Nassif & Engels, 1993).

To study the function of the H4r gene, P-element excision was chosen as the method for the mutagenesis of H4r gene in *Drosophila melanogaster*. A "put" strain with an insertion of a P-element at the 5'-UTR of the *punt* gene in the neighbourhood of H4r gene (Akamanova, 1996) was crossed with a jumpstarter strain offering the transposase (Robertson et al, 1988).

The P-element in “put” strain could thus be excised with the help of the transposase. In the case of an “imprecise” loss, mutant strains could be obtained with deletion or insertion around the initial insertion position of the P-element.

Methods und Materials

Containers

All transparent plastic fly containers were obtained from the company Greiner Labortechnik. They were closed with proper foam tops.

Size	Name	Diameter in (mm)	Height in (mm)	Volume in (ml)
big	PS-Dosen-U Teil, RD	53	100	175
middle	PS-Röhrchen, FB	36	83	68
small	PS-Röhrchen, FB	22	63	16

Fly medium

Cooked fly medium

216 g agar
10 l water — cooked for ca. 20 Minutes, till the agar is completely dissolved.

480 g dry yeast
130 g soja flour
2420 g corn
5 l water — mixed in a container then given to the cooking agar solution.
2160 g malt extract

1080 g sirup	added
10.5 l water	added; part of this water was used to rinse all the containers for the rest of medium components. Under constant mixing, the medium was heated to 90°C, and then kept cooking for about 30 minutes.
50 g Nipagin	
2 l water	Nipagin was dissolved in hot (but not cooking) water.
120 ml Propionic acid	Nipagin und Propionic acid were given into the mixture about 10 minutes before filling the fly containers with it. It's important that these two components were not cooked for too long, but well mixed into the medium.

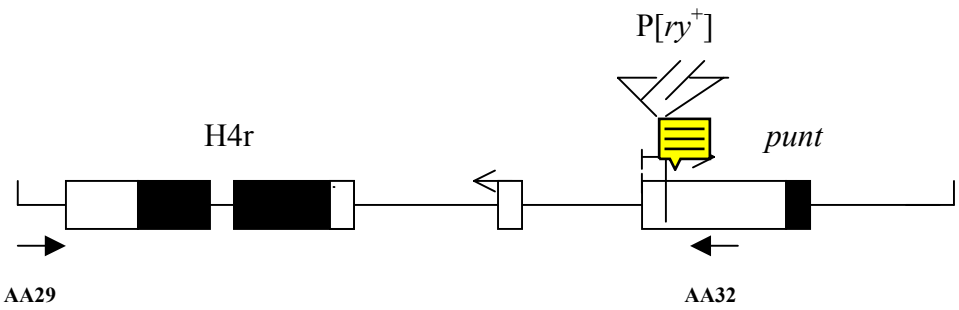
"Nipagin" (4-Hydroxibenzoesäuremethylester) is a Fungicide. It is sometimes added into bread to prevent moulding.

Instant *Drosophila* medium

„Formular 4-24 blue" Instant *Drosophila* Medium (Carolina Biological Supply Company, Burlington, NC)

Put Strain

A „put" strain containing a P-element insertion at 88C8-10, in the 5'-UTR of the *punt* gene was used as the starting strain. *punt* is a housekeeping gene, which encodes a protein of the type II receptor STK family closely related to the vertebrate activin receptor. *Punt* and the H4r gene are located on opposite DNA-strands, with 118 base pairs between their transcription start points. The P-element insertion is only 148 base pairs away from the initiation of H4r gene.



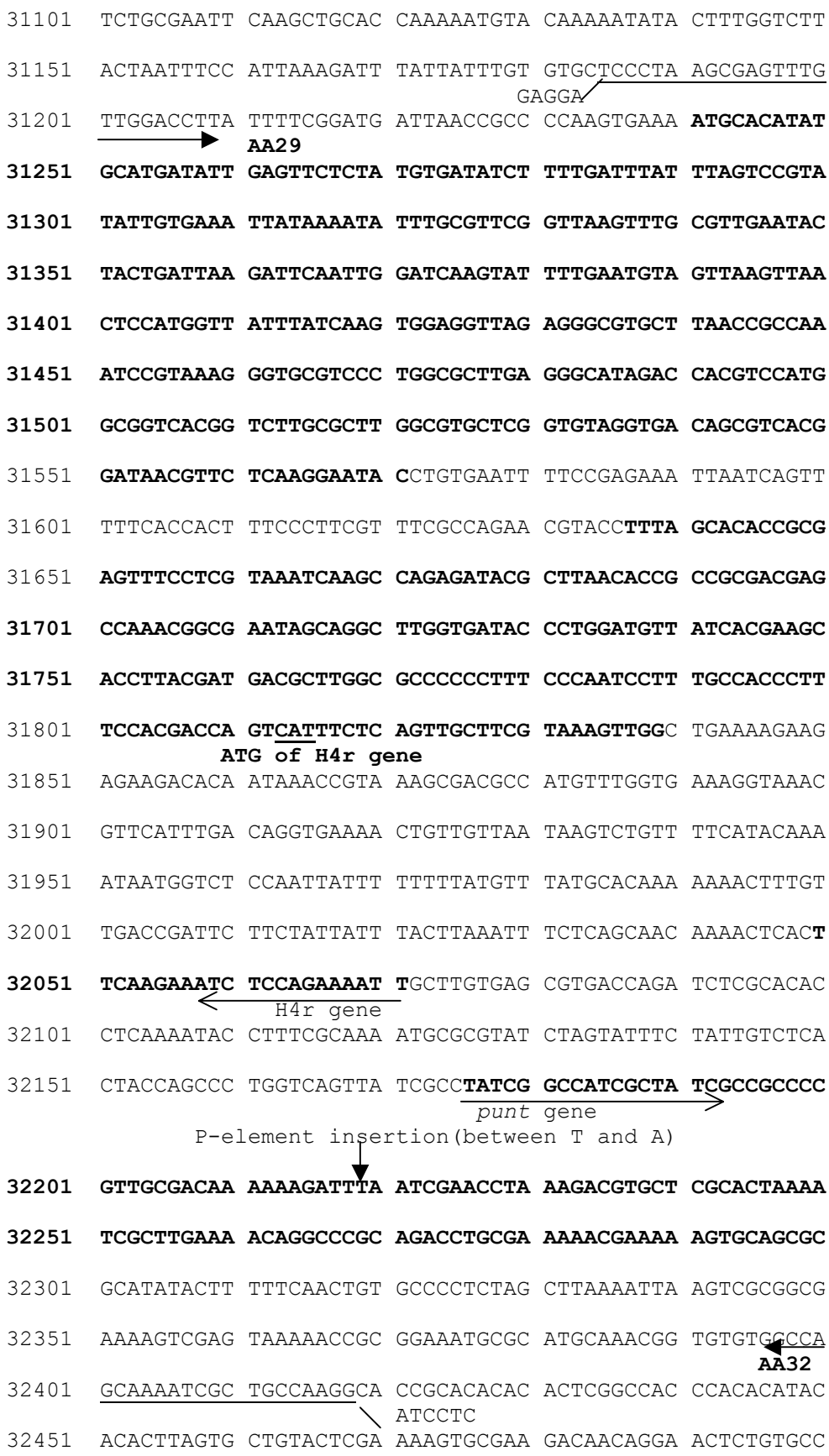

31101 TCTGCGAATT CAAGCTGCAC CAAAAATGTA CAAAAATATA CTTTGGTCTT
31151 ACTAATTTCC ATTAAAGATT TATTATTTGT GTGCTCCCTA AGCGAGTTTG
GAGGA
31201 TTGGACCTTA TTTTCGGATG ATTAACCGCC CCAAGTGAAA ATGCACATAT
AA29
31251 GCATGATATT GAGTTCTCTA TGTGATATCT TTTGATTTAT TTAGTCCGTA
31301 TATTGTGAAA TTATAAAATA TTTGCGTTCG GTTAAGTTTG CGTTGAATAC
31351 TACTGATTAA GATTCAATTG GATCAAGTAT TTTGAATGTA GTTAAGTTAA
31401 CTCCATGGTT ATTTATCAAG TGGAGGTTAG AGGGCGTGCT TAACCGCCAA
31451 ATCCGTAAAG GGTGCGTCCC TGGCGCTTGA GGGCATAGAC CACGTCCATG
31501 GCGGTCACGG TCTTGCGCTT GGCGTGCTCG GTGTAGGTGA CAGCGTCACG
31551 GATAACGTTC TCAAGGAATA CCTGTGAATT TTCCGAGAAA TTAATCAGTT
31601 TTTCACCACT TTCCCTTCGT TTCGCCAGAA CGTACCTTTA GCACACCGCG
31651 AGTTTCCTCG TAAATCAAGC CAGAGATACG CTTAACACCG CCGCGACGAG
31701 CCAAACGGCG AATAGCAGGC TTGGTGATAC CCTGGATGTT ATCACGAAGC
31751 ACCTTACGAT GACGCTTGGC GCCCCCCTTT CCCAATCCTT TGCCACCCTT
31801 TCCACGACCA GTCATTTCTC AGTTGCTTCG TAAAGTTGGC TGAAAAGAAG
ATG of H4r gene
31851 AGAAGACACA ATAAACCGTA AAGCGACGCC ATGTTTGGTG AAAGGTAAAC
31901 GTTCATTTGA CAGGTGAAAA CTGTTGTTAA TAAGTCTGTT TTCATACAAA
31951 ATAATGGTCT CCAATTATTT TTTTTATGTT TATGCACAAA AAAACTTTGT
32001 TGACCGATTC TTCTATTATT TACTTAAATT TCTCAGCAAC AAAACTCACT
32051 TCAAGAAATC TCCAGAAAAT TGCTTGTGAG CGTGACCAGA TCTCGCACAC
H4r gene
32101 CTCAAAATAC CTTTCGCAAA ATGCGCGTAT CTAGTATTTC TATTGTCTCA
32151 CTACCAGCCC TGGTCAGTTA TCGCCTATCG GCCATCGCTA TCGCCGCCCC
punt gene
P-element insertion(between T and A)
32201 GTTGCGACAA AAAAGATTTA ATCGAACCTA AAGACGTGCT CGCACTAAAA
32251 TCGCTTGAAA ACAGGCCCGC AGACCTGCGA AAAACGAAAA AGTGCAGCGC
32301 GCATATACTT TTTCAACTGT GCCCCTCTAG CTTAAAATTA AGTCGCGGCG
32351 AAAAGTCGAG TAAAAACCGC GGAAATGCGC ATGCAAACGG TGTGTGCCCA
AA32
32401 GCAAAATCGC TGCCAAGGCA CCGCACACAC ACTCGGCCAC CCACACATAC
ATCCTC
32451 ACACTTAGTG CTGTACTCGA AAAGTGCGAA GACAACAGGA ACTCTGTGCC

Fig 1: genomic organization of *punt*, H4r genes and the P-element insertion in "put" strain. The genomic structure of the H4r gene is shown with exons in bold letters.

Crosses*:

P $\quad \dfrac{ry^{506}\ \text{put P}\ [ry^{+}]}{TM3,\ ry\ Sb^{1}}$ X $\dfrac{TM3,\ Sb,\ ry^{\text{Ru}}\ [\Delta 2\text{-}3\ \ ry^{+}]}{Ly}$

+ eyes, Sb bristles, + wings ↓ + eyes, Sb bristles, Ly wings

F1 $\dfrac{ry^{506}\text{put P}[ry^{+}]}{TM3,\ ry^{\text{Ru}}\ Sb\ \text{P}[\Delta 2\text{-}3\ \ ry^{+}]}$ X $\dfrac{TM3,\ ry\ Sb^{1}}{Ly}$

+ eyes, Sb bristles, + wings + eyes, Sb bristles, Ly wings

F2 $\dfrac{ry^{\ 506}}{TM3,\ ry\ Sb^{1}}$ $\dfrac{ry^{506}\text{put P}[ry^{+}]}{TM3,\ ry\ Sb^{1}}$

ry eyes, Sb bristles, + wings + eyes, Sb bristles, + wings

*: Only the genotype of the third chromosome is shown.

TM3: balancer for the third chromosome, homozygous lethal *ry*: rosy eyes allele, recessive *Ly*: Lyra wings, dominant *Sb*: Stubble bristles, dominant (Abbreviations according to Linsley & Zimm, 1992)

Result

P generation

In the P generation cross, 30 "put" strain ry^{506} put P [ry^{+}]/ TM3, *ry* Sb^{1} males were crossed with 30 *TM3*, *ry* ^{Ru} [Δ2-3 ry^{+}]/ *Ly* females.

ry^{506} put P [ry^{+}]/ *TM3*, *ry* Sb^{1} flies have the P-element insertion in 5'UTR of the *punt* gene. This P-element contains a copy of the mini-*rosy* gene which could rescue a *rosy* genotype background.

TM3, *ry* ^{Ru} [Δ2-3 ry^{+}]/ *Ly* flies are the so-called jump starters with the insertion of a jump starter P-element which could mobilise other P-elements, but not themselves.

F1 generation

The two genotypes of flies in the F1 cross can be distinguished by their phenotypes.

In ry^{506}put P[ry^{+}] / *TM3*, ry^{Ru} *Sb* P[Δ2-3 ry^{+}] flies, the two different P-elements were brought together. In gametogenesis of these flies, the transposase of the jump starter P-elements can mobilise the "*ry*" P-elements in the third chromosome originating from the *punt* flies, so that part of their gametes would have a mutated "*ry*" chromosome.

F2 generation

In the F2 generation, 56 flies with rosy eyes were observed among ca. 5,000 flies with red eyes. They had normal wings and *Sb* bristles so they could only have the combination of the *TM3* chromosome from the balancer strain and the mutated *ry* chromosome from "put". The *ry* chromosome from "put" had originally a P-element with the *ry* gene, which rescued their *rosy* phenotype. In these *ry* flies, either the P-elements were excised, or they had an internal recombination so that the *ry* allele was no more intact. Further experiments were concentrated on the *rosy* flies. They were backcrossed with their maternal strain "put". 5 of them died before leaving any offspring.

F3 and F4 generation

Type I

From 43 F2 flies, three kinds of offspring were observed in the F3 generation: *Sb ry, Sb* ry^+, and the Sb^+ ry^+ flies which indicates that these flies had a mutated *ry* chromosome which was viable over the original put chromosome.

F3:

$\frac{ry^{506}}{ry^{506}\text{put P}[ry^+]}$	$\frac{ry^{506}}{\text{TM3, } ry\ Sb^1}$	$\frac{ry^{506}\text{put P}[ry^+]}{\text{TM3, } ry\ Sb^1}$
+ eyes, + bristles	*ry* eyes, *Sb* bristles	+ eyes, *Sb* bristles

x

↓

F4:

$\frac{ry^{506}}{ry^{506}}$	$\frac{ry^{506}}{\text{TM3, } ry\ Sb^1}$
ry eyes, + bristles	*ry* eyes, *Sb* bristles

Males and females from the ry^{506} / *TM3*, *ry* Sb^1 flies were crossed. Heterozygote ry^{506} / *TM3*, *ry* Sb^1 flies and homozygote ry^{506} / ry^{506} flies were obtained in the F4 generation, which indicate that all these *ry* chromosomes were also viable in a homozygous constitution.
Stable strains were maintained with the homozygote ry^{506} / ry^{506}.

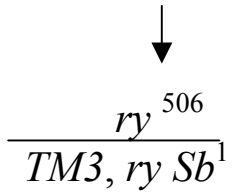
<u>Type II</u>

From 8 F2 flies, two kinds of offspring, *Sb ry* and *Sb* ry^+, were observed in the F3 generation. Sb^+ ry^+ flies did not exist, indicating that these flies had a mutated *ry* chromosome, which was not viable over the original put chromosome.

F3:

$\frac{ry^{506}\text{ put P}[ry^+]}{TM3, ry\ Sb^1}$	$\frac{ry^{506}}{TM3, ry\ Sb^1}$
+ eyes, *Sb* bristles	*ry* eyes, *Sb* bristles

x

F4: 

ry eyes, *Sb* bristles

Males and females from the ry^{506} / *TM3*, *ry* Sb^{1} flies were crossed. Only heterozygote ry^{506} / *TM3*, *ry* Sb^{1} flies were obtained in F4 generation, which indicates that all these *ry* chromosomes were also homozygos lethal.
Stable strains were maintained as heterozygotes of the constitution ry^{506} / *TM3*, *ry* Sb^{1}.

Molecular examination

It was reported that in some cases of P-element transposition internal excisions or rearrangements of P-elements occur, which could also fit the observation of the existence of internally deleted P-elements (Engels et al, 1990; and references therein). If the P-element in "put" strain had an internal rearrangement, which distorted the *rosy* gene, the F2 flies would also have *rosy* eyes, as in the case of a real excision. To exclude this possibility and to check if the P-element was really excised, PCR reactions with primers AA29 and AA32 were carried out for each F4 strain. AA32 and AA29 would only give a product when the P-element was, at least partly, excised, otherwise they were too far away from each other to yield PCR products under the conditions of the reaction.

Genomic DNA was prepared from 3 to 5 homozygote flies for type I strains. For all flies, AA29 and AA32 gave a product of the size expected if no insert is present, i.e. 1.2 kb. The PCR products were cloned and sequenced, they all correspond to the wild-type genomic sequence between AA29 and AA32.

For type II flies, genomic DNA was prepared from 3 heterozygote flies. Another primer P was designed which has the identical sequence as the end repeat of P-element. PCR with AA29 + P and AA32 + P gave no signal. So it was concluded that at least the end repeats of the P-element were excised.

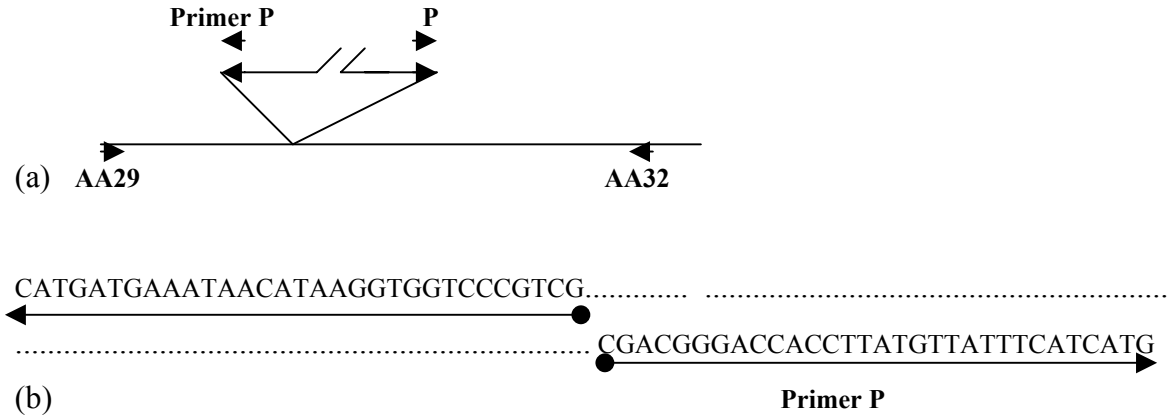


Fig 2: (a): The genomic structure of the P element inserted in „put" strain. A 31-base pairs repeat flanks both ends of the P-element in reversed orientation. (b) The sequence of the 31-base pairs repeat. Primer P was designed according to the sequence of this repeat. The arrowheads show the orientation of the primers.

The PCR reaction with AA29 and AA32 gave similar results as for type I strains. Only one major product at 1.2kb was seen. All the products were cloned and at least 5 subclones were sequenced for each product. They all corresponded to the wild type sequence between AA29 and AA32. Because of the presence of the balancer chromosome, it cannot be excluded that this product was only amplified from the balancer.

To exclude the effect of the balancer chromosome, it was intended to cross the F4 flies with another double balancer. The *TM3* balancer could be exchanged with another balancer marked with Green Florescence Protein. One could thus check the embryos coming out from the intercrossing of these flies and pick out those without the fluorescent balancer, i.e. the homozyotes for the *ry* chromosome. PCR experiments could be done with these embryos. Unfortunately, all the 8 type II F3 strains survived badly compared to type I flies, and they all died out before the planned balancer exchange. A definite conclusion could not be made about these flies.

Discussion

From c.a. 5,000 F2 flies, we obtained 56 flies that have undergone either a P-transposition or an internal recombination of P-elements. Among 51 of them that survived long enough to leave offspring, 43 had a new chromosome 3, which is viable over the original chromosome 3 of put flies. Molecular data could show that they all had a complete and precise loss of P-element from the original insertion. The other 8 of the 56 F2 flies had a new chromosome 3, which was not viable over the put chromosome 3. At least the end repeats of their P-elements were transposed from the original position.

We did not succeed to obtain H4r mutations through P-element excision. There are two possible reasons:

1. H4r may have an indispensable function so that mutant embryo in F2 did not survive to adults. It could also be that some of the 8 type II flies contained the mutation, but did not survive long enough to be completely analysed because of their bad viability. Considering the importance of histones for chromatin constitution, it would be not surprising if the H4r mutants are haploinsufficient.
2. The number of screened excisions was too low to recover excisions extending into the H4r gene. A possible enhancement of the sensitivity of the screening is to introduce a section of mutated H4r into the genome of put flies with the help of P-elements, so that the double strand breaks left by the transposable P-element in the F1 germ line could be repaired with the aid of the mutated partial H4r copy as template.

References

Abe, H., Ohbayashi, F., Sugasaki, T., Kanehara, M., Terada, T., Shimada, T., Kawai, S., Mita, K., Kanamori, Y., Yamamoto, M.T. and Oshiki, T. 2001 Two novel Pao-like retrotransposons (Kamikaze and Yamato) from the silkworm species Bombyx mori and B. mandarina: common structural features of Pao-like elements. Mol Genet Genomics 265(2):375-85.

Akhmanova, A., Miedema, K. and Hennig, W. 1996. Identification and characterization of the Drsophila histone H4 replacement gene. FEBS Lett. 388:219-222.

Alm, R.A., Ling, L.-S.-L., Moir, D.T. et al. and Trust, T.J. 1999. Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature 397(6715):176-180.

Ashburner, M. 1989. Drosophila: a laboratory manual. Cold Spring Harbor Laboratory Press.

Aubry, M., Marineau, C., Zhang, F.R., Zahed, L., Figlewicz, D., Delattre, O., Thomas, G., deJong, P.J., Julien, J.P. and Rouleau, G.A. 1992. Cloning of six new genes with zinc finger motifs mapping to short and long arms of human acrocentric chromosome 22p and q11.2. Genomics. 13:641-648.

Bardeesy, N. and Pelletier, J. 1998. Nucleic Acids Res. 26:1784-1792.

Bellefroid, E.I., Marine, J.C., Ried, T., Lecorq, P.J., Riviere, M., Amemiya, C., Poncelet, D.A., Coulie, P.G., deJong, P., Spiere, C., Ward, D.C. and Martial, J.A. 1993. Clustered organization of homologous KRABzinc finger genes with enhanced expression in human T lymphoid cells. EMBO. J. 12:1363-1374.

Boulay, J. L., Dennefeld, C. and Alberga, A. 1987. The *Drosophila* developmental gene snail encodes a protein with nucleic acid binding fingers. Nature 330:395-398.

Bruening , W., Moffet, P., Chia, S., Heinrich, G. and Pelletier, J. 1996. FEBS Lett. 393:41-47
Brush, D., Dodgson, J.B., Choi, O.R., Stevens, P.W. and Engel, J.D. 1985. Replacement varinat histone genes contain intervening sequences. Mol.Cell.Biol. 5:1307-1317.

Calabro, V., Pengue, G., Bartoli, P.C., Pagliuca, A., Featherstone, T. and Lania, L. 1995. Positional cloning of cDNAs from the human chromosome 3p21-22 region identifies a clustered organization of zinc-finger genes. Hum.Genet. 95:18-21.

Call, K.M., Glaser, T., Ito, C.Y., Buckler, A.J., Pelletier, J., Haber, D.A., Rose, E.A., Dral, A., Yeger, H., Lewis, W.H., Jones, C. and Housman, D.E. 1990. Isolation and characterization of a zinc finger polypeptide gene at human chromosome 11 Wilm's tumor locus. Cell 60:509-520.

Choo, Y. and Isalan, M. 2000. Advances in zinc finger engineering. Curr. Opin. Struct. Biol. 10:411-416.

Clarke, N. D., and Berg, J.M. 1998. Zinc fingers in Caenorhabditis elegans: Finding families and probing pathways. Science282:2018-2022.

Constantinous-Deltas, C.D., Gilbert, J., Barlett, R.J., Herbstrith, M., Roses, A.D. and Lee, J.E. 1992. The identification and characterization of KRAB-domain-containing zinc finger proteins. Genomics. 12:581-589.

Darby, M.K. and Joho, K.E. 1992. Differential binding of zinc fingers from *Xenopus* TFIIIA and p43 to 5SRNA gene. Mol.Cell. Biol. 12:3155-3164.

de Celis, J.F., Barrio, R., Kafatos, F.C. 1996. A gene complex acting downstream of dpp in Drosophila wing morphogenesis. Nature. 381(6581):421-424.

Derry, J.M., Jess, U. and Francke, Y. 1995. Cloning and characterization of a novel zinc finger gene in Xp11.2. Genomics. 30:361-365.

Eltrod-Erickson, M., Benson, T.E. and Pabo, C.O. 1998. High-resolution structures of variant Zif268-DNA complexes: inplications for understanding zinc finger-DNA recognition. Structure 6:451-464.

Engels, W. R., Johnson-Schlitz,D.M., Eggleston, W.B. and Sved, J. 1990 High-frequency *P* element loss in Drosophila is homolog dependent. Cell **62**:515-525.

Engels, W.R., 1997. Invasion of P elements. Genetics 14511-14515.

Friesen, W.J. and Darby, M.K. 1997. Phage display of RNA binding zinc fingers from transcription factor IIIA. J. Biol. Chem. 272:10994-10997.

Friesen, W.J. and Darby, M.K. 1998. Specific RNA binding proteins constructed from zinc fingers. Nat. Struct. Biol. 5:543-546.

Gavel, Y. and von Heijne, G. 1990. Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. Protein Eng. 3:433-442.

Georgopoulos, K., Winandy, S. and Avitahl, N. 1997. The role of the Ikaros gene in lymphocyte development and homeostasis. Annu. Rev. Immunol. 15 :155-176.

Geyer, P. K., Richardson, K.L. Corces, K.L. and Green, M.M. 1988 Genetic instability in *Drosophila melanogaster*: *P*-element mutagenesis by gene conversion. Proc. Natl. Acad. Sci. USA **85**:6455-6459.

Glass, D.B. Smith, S.B. 1983. Phosphorylation by cyclic GMP-dependent protein kinase of a synthetic peptide corresponding to the autophosphorylation site in the enzyme. J. Biol. Chem. 258:14797-14803.

Glass, D.B., El-Maghrabi, M.R., Pilkis, S.J. 1986. Synthetic peptides corresponding to the site phosphorylated in 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase as substrates of cyclic nucleotide-dependent protein kinases. J. Biol. Chem. 261:2987-2993.

GLOOR, G. B., N. A. NASSIF, D. M. JOHNSON-SCHLITZ, C. R. PRESTON, and W. R. ENGELS, 1991 Targeted gene replacement in Drosophila via *P* element-induced gap repair. Science **253**:1110-1117.

Grand, R.J.A. 1989. Acylation of viral and eukaryotic proteins. Biochem. J. 258:625-638.

Hart, M.C., Wang, L. and Coulter. D.E. 1996. Comparison of the structure and expression of *odd-skipped* and two related genes that code a new family of zinc finger proteins in *Drosophila.* Genetics 144:171-182.

Heyting, C., Dettmers, R.J., Dietrich, A.J.J., Redeker, E.J.W. and Vink, A.C.G. 1988. Two major components of synaptonemal complexes are specific for meiotic prophase nuclei. Chromosoma 96:325-332.

Heyting, C., Dietrich, A.J.J., Moens, P.B., Dettmers, R.J., Offenberg, H.H., Redeker, E.J.W. and Vink, A.C.G. 1989. Synaptonemal complex proteins. Genome 31:81-87.

Heyting, C., Dietrich, A.J.J., Redeker, E.J.W. and Vink, A.C.G. 1985. Structure and composition of synaptonemal complexes, isolated from rat spermatocytes. Eur. J. Cell Biol. 36:307-314.

Heyting, C., Moens, P.B., van Raamsdonk, W., Dietrich, A.J.J., Vink, A.C.G. and Redeker, E.J.W. 1987. Identification of two major components of the lateral elements of synaptonemal complexes of the rat. Eur. J. Cell Biol. 43:148-154.

Hoheisel, J.D., Lennon, G.G., Zehetner, G. and Lehrach, H. 1991. Use of high coverage reference libraries of Drosophila melanogaster for relational data analysis. J. Mol. Biol. 220:902-914.

Hoovers, J.M.N., Mannens, M., John, R., Bliek, J., van Heyningen, V., Porteus, D. J., Leschot, N., J., Westerveld, A. and Little, P.F.R. 1992. High resolution localization of 69 potential human zinc finger protein genes: a number are clustered. Genomics 12:254-263.

Huebner,K., Druck, T., Croce, C.M. and Thiesen, H.J. 1991. Twenty-seven nonoverlapping zinc finger cDNAs from human T cells map to nine different chromosomes with apparent clustering. Am. J. Hum. Genet. 48:726-740.

Johnson-Schlitz, D. M. and Engels, W.R. 1993 *P*-element-induced interallelic gene conversion of insertions and deletions in *Drosophila melanogaster.*. Mol. Cell. Biol. **13**:7006-7018.

Kapitonov V.V., Jurka J 1999 direct submission to the database of repeatitive sequences http://www.girinst.org/

Katti, M.V., Ranjekar, P.K. and Gupta, V.S. 2001 Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol. Biol. Evol. 18(7):1161-1167.

Kaufman, P. D. and Rio, D.C. 1992 *P* element transposition in vitro proceeds by a cut-and-paste mechanism and uses GTP as a cofactor. Cell **69**:27-39.

Kishimoto, Y., Nishiyama, K., Nakanishi, H., Uratsuji, Y., Nomura, H., Takeyama, Y. And Nishizuka, Y. 1985. Studies on the phosphorylation of myelin basic protein by protein kinase C and adenosine 3':5'-monophosphate-dependent protein kinase. J. Biol. Chem. 260:12492-12499.

Koetsier PA, Schorr J, Doerfler W. 1993. A rapid optimized protocol for downward alkaline Southern blotting of DNA. Biotechniques 15(2):260-262.

Köster, M., Kühn, U., Bouwmeester, T., Nietfeld, W., El-Baradi, T., Knöchel, W. and Pieler, T. 1991. Structure, expression and in vitro functional characterization of a novel RNA binding zinc finger proteinfrom *Xenopus* EMBO J. 10:3087-3093.

Lammers, J.H.M., Offenberg, H.H., van Aaleren, M., Vink, A.C.G., Dietrich, A.J.J. and Heyting, C. 1994. The gene encoding a major component of the lateral elements of synaptonemal complexes of the rat is related to X-linked lymphocyte-regulated genes. Mol. Cell. Biol. 14:1137-1146.

Larsson, S.H., Charlieu, J.P., Miyagawa, K., Engelkamp, D., Rassoulzadegan, M., Ross, A., Cuzin, F., can Heyningen, V.and Hastie, N.D. 1995. Cell 81:391-401.

Leon, O. and Roth, M. 2000. Zinc fingers: DNA binding and protein-protein interactions. Biol. Res. 33:21-30.

Li, W. and Graur, D. 1991. Fundamentals of molecular evolution. Sinauer Associates, Inc. Sutherland, Massachusetts.

Lichter, P.,Bray, P., Ried, T., Dawid, I.B. and Ward, D.C. 1992. Clustering of C2H2 zinc finger motif sequences within telomeric and fragile site regions of human chromosomes. Genomics. 13:999-1007.

Lindsley, D.L, and Zimm, G.G. 1992. The genome of *Drosophila melanogaster.* Academic Press, Sandiego, CA.

Mak, C.H., Li, Z., Allen, C.E., Liu, Y., Wu, L. 1998. KRC transcripts: identification of an unusual alternative splicing event. Immunogenetics 48(1):32-39.

Mathews, F.S. 1985. The structure, function and evolution of cytochromes. Prog. Biophys. Mol. Biol. 45:1-56.

Michael, S.F., Kilfoil, V.J., Schmidt, M.H., Amnn, B.T. and Berg, J.M. 1992. Metal binding and folding properties of a minimalist Cys_2His_2 zinc finger peptide. Proc.Natl.Acad.Sci.USA. 89:4798-4800.

Miletich, J.P. and Broze, G.J. Jr. 1990. Beta protein C is not glycosylated at asparagine 329. The rate of translation may influence the frequency of usage at asparagine-X-cysteine sites. J.Biol.Chem. 265:11397-11404.

Misra, S., Hecht, P., Maeda, R. and Anderson, K.V. 1998. Positive and negative regulation of Easter, a member of the serine protease family that controls dorsal-ventral patterning in the *Drosophila* embryo. Development. 125(7):1261-1267.

Morgan, B., Sun, L., Avitahl, N., Andrikopoulos, K., Ikeda, T., Gonzales, E., Wu, P., Neben, S. and Georgopoulos K. 1997. Aiolos, a lymphoid restricted transcription factor that interacts with Ikaros to regulate lymphocyte differentiation. EMBO J. 16:2004-2013.

Moses, M.J. 1968. Structure and function of the synaptonemal complex. Genetics. 61(1):Suppl:41-51.

Nassif, N., Penny, J., Pal, S., Engels, W.R. and Gloor, G.B. 1994 Efficient copying of nonhomologous sequences from ectopic sites via *P*-element-induced gap repair. Trends Biochem. Sci. **14**:1613-1625.

O'Kane, C.J. and Gehring, M.J. 1987 Detection in situ of genomic regulatory elements in Drosophila. Proc Natl Acad Sci U S A 84:9123-9127.

Ogura, K., Takechi, S., Nakayama, T. and Yamamoto, M.T. 1996. Molecular structure of the transposable element ninja in Drosophila simulans. Genes Genet Syst 71(1):1-8.

O'HARE, K. and G. M. RUBIN, 1983 Structures of *P* transposable elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. Cell **34**:25-35.

Ohno, S. 1970. Evolution by gene duplication. (Springer-Verlag, Berlin, New York).

Osley, M.A. 1991. The regulation of histone synthesis in the cell cycle. Annu. Rev. Biochem. 60:827-861.

Pavletich, N.P. and Pabo, C.O. 1991. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science 252:809-817.

Pinna, L.A. 1990. Casein kinase 2: an 'eminence grise' in cellular regulation? Biochim. Biophys.Acta 1054:267-284.

Rasmussen, S.W. 1973. Ultrastructural studies of spermatogenesis in *Drosophila* m*elanogaster.* Meigen. Z. Zellforsch. Mikrosk. Anat. 140:125-144.

Rasmussen, S.W. 1974. Studies on the development and ultrastructure of the synaptonemal complex in *Drosophila melanogaster.* C.R. Trav. Lab. Carlsberg. 39:443-468.
Rawlings, N.D. and Barret, A.J. 1994. Families of serine peptidases. Methods Enzymol. 244:19-61.

Reuter, D., Kühnlein, R.P., Frommer, G., Barrio, R., Kafatos, F.C., Jäckle, H. and Schuh, R. 1996. Regulation, function and potential origin of the *Drosophila* gene *spalt adjacent*, which encodes a secreted protein expressed in the early embryo. Chromosoma. 104:445-454.

Robertson, H.M., Preston, C.R., Phillis, R.W., Johnson-Schlitz, D.M.,, Benz, W.K. and engels, W.R. 1988. A stable genomic source of P element transposase in Drosophila melanogaster. Genetics, 118:461-470.

Rosati, M., Franz, A., Matarazzo, M.R., Grimaldi, G. 1999. Coding region intron/exon organization, alternative splicing, and X-chromosome inactivation of the KRAB/FPB-domain-containing human zinc finger gene ZNF41. Cytogenet Cell Genet. 85(3-4):291-296.

Rousseau-Merck, M.F., Hilton, J., Jonveaux, P., Couillin, P., Seite, P., Thiesen, H.J. and Berger, R. 1993. Chromosomal localization of 9 KOX zinc finger genes: physical linkages suggest clustering of KOX genes on chromosomes 12,16 and 19. Hum. Gene. 92:583-587.

Ruberte, E., Marty, T., Nellen, D., Affolter, M. and Basler, K. 1995 An absolute requirement for both the type II and type I receptors, punt and thick veins, for dpp signaling in vivo. Cell 80(6):889-897.

Saleh, M., Selleri, L., Little, P.F.R. and Evans, G.A. 1992. Isolation and expression of linked zinc finger gene clusters of human chromosome 11q. Genomics. 14:970-978.

Sambrook, J., Fritsch. E.F. and Maniatis, T. 1989. Molecular cloning:a laboratory manual. Second edition. Cold Spring Harbor Laboratory Press.

Schuh, R., Aider, W., Gaul, U., Cote, S., Preiss, A., Maier, D., Seifert, E., Nauber, U., Schröder, C., Kemier, R. and Jäckle, H. 1986. A conserved family of nuclear proteins containing structural elements of the finger protein encoded by Krüppel, a *Drosophila* segmentation gene. Cell 47:1025-1032.

Schümperli, D. 1986 Cell-cycle regulation of histone gene expression. Cell 45: 471-472.
Schümperli, D. 1988 Multilevel regulation of replication-dependent hisstone genes. Trends genet. 4:187-191.

Shannon, M., Ashworth, L.K., Mucenski, M., Lamerdin, J.E., Branscomb, E. and Stubbs, L. 1996. Comparative analysis of a conserved zinc finger gene cluster on human chromosome 19q and mouse chromosome 7. Genomics. 33:112-120.

Shannon, M., Kim, J., Ashworth, L., Branscomb, E. and Stubbs, L. 1998. Tandem zinc-finger gene families in mammals: Insights and unanswered questions. DNA Seq. 8:303-315
Shi, Y. and Berg, J.M. 1995. A direct comparison of the properties of natural and designed zinc-finger proteins. Chem.Biol. 2:83-89.

Smith, A. and Benavente, R. 1992. Identification of a structuralprotein component of rat synaptonemal complexes. Exp. Cell Res. 198:291-297.

Sun, L., Liu, A. and Georgopoulos, K. 1996. Zinc finger-mediated protein interactions modulate Ikaros activity, a molecular control of lymphocyte development. EMBO J. 16:2004-2016.

Sun, X. 1994. Expression of two *Drosophila* genes. Ph.D. Thesis.

Supp, D.M., Witte, D.P., Brandford, W.W., Smith, E.P. and Potter, S.S. 1996. Sp4, a member of the Sp1-family of zinc finger transcription factors, is required for normal murine growth, viability, and male fertility. Dev. Biol. 176:284-299.

Thomas, A. and Skolnick, M.H. 1994. A probabilistic model for detecting coding regions in DNA sequences. IMA J Math Appl Med Biol. 11: 149-160.

Tommerup, N. and Vissing, H. 1995. Isolation and fine mapping of 16 novel human zinc finger encoding cDNAs identify putative candidate genes for developmental and malignant disorders. Genomics. 27:259-264.

Towler, D.A., Gordon, J.I., Adams, S.P. and Glaser, L. 1988 The biology and enzymology of eukaryotic protein acylation. Annu. Rev. Biochem. 57:69-99.

Valentin-Hansen, P., Larsen, J.E., Hojrup, P., Short, s.A. and Barbier, C.S. 1986. Nucleotide sequence of the CytR regulatory gene of E.coli K-12. Nucleic Acid Res. 14(5): 2215-2218.

Wells, D. 1986. Compilation analysis of histones and histone genes. Nucleic Acid Res. 14 Suppl, r119-r149.

Wells, d. and McBride, C. 1989. A comprehensive compilation and alignment of histones and histone genes. Nucleic Acid Res. 17 Suppl. 311-346.

Wieschaus, E., Nüsslein-Volhard, C. and Kluding, H. 1984. Krüppel, a gene whose activity is required early in the zygotic genome for normal embryonic segmentation. Dev. Biol. 104:172-186.

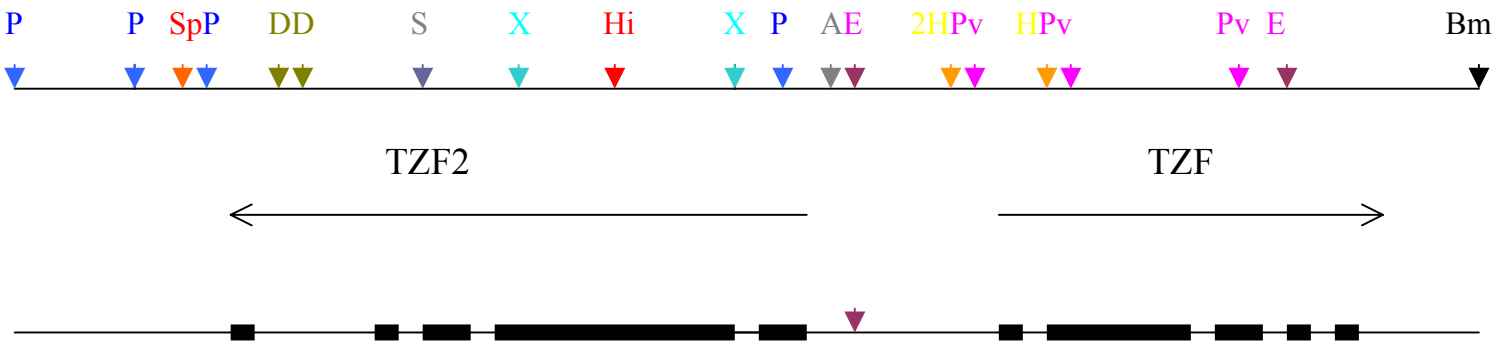
Wolfe, S.A., Nekludova, L. and Pabo, C.O. 2000. DNA recognition by Cys_2His_2 zinc finger proteins. Annu. Rev. Biophys. Biomol. Struct. 29:183-212.

Woodgett, J.R., Gould, K.L. and Hunter, T 1986. Substrate specificity of protein kinase C. Use of synthetic peptides corresponding to physiological sites as probes for substrate recognition requirements. Eur. J. Biochem. 161:177-184.

Xiong, Y., Burke, W.D. and Eickbush, T.H. 1993 Pao, a highly divergent retrotransposable element from Bombyx mori containing long terminal repeats with tandem copies of the putative R region. Nucleic Acids Res., 21: 2117-2123.

Appendix I:

The genomic structure of the genes *Tzf* and *Tzf2*

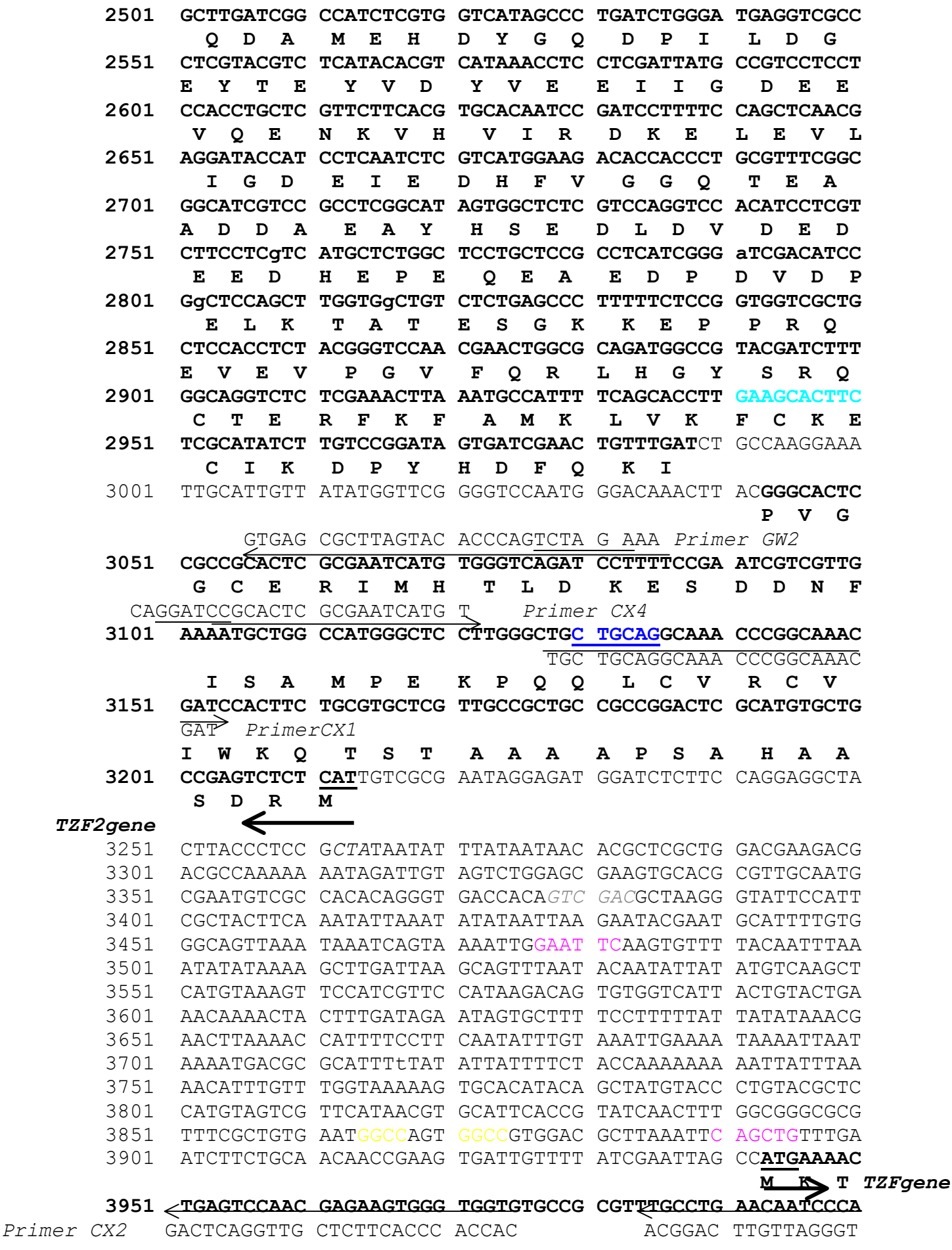


Genomic organization of TZF and TZF2

1. The ATG of TZF and TZF2 are underlined. Bold letters are from the coding exons.
2. The restriction sites used in subcloning are highlighted in the same colour as in the restriction map.
3. Some primers have several altered nucleotides compared to the original sequence to create a restriction site at the 5′ end. The restriction sites are underlined.

```
   1  CTGCAGGCCT CCCACAGCGC GCCCCTGCAC ACGAACGGTG GCCAACAGTT
  51  GGTCAGTCCG GGTGGCAATA ATAACAATGT TAAGATCAAC TCACTGATTA
 101  GCCGGGAGCC GCTGGGGCGC ACCAAGTCAA ATACCATCGA GCAGGTATGT
 151  CAGGGCTTCT CCGCTCAAAT GGATCTTTCA GAGGGTGCCG TGGAATCCAG
 201  CCAGCCCAAC CGCCACAAGA AGCTGCAGGC CTCCCACAGC GCGCCCCTGC
 251  ACACGAACGG TGGCCAACAG TTGGTCAGTC CGGGTGGCAA TAATAACAAT
 301  GTTAAGATCA ACTCACTGAT TAGCCGGGAG CCGCTGGGGC GCACCAAGTC
 351  AAATACCATC GAGCAGGTAT GTCAGGGCTT CTCCGCTCAA ATGGATCTTT
 401  CAGAGGGTGC CGTGGAAGCC AGCCAGCCCA ACCGCCACAA GAAGCTGCAG
 451  CGCCAACAAC CGCCGCCCGC CTACCGCCTG CTGGTGCCCA CCTACAGTGC
 501  TCCCCTCCAG CAACAACAAC ACCAGGCACA GCAGCAGCAT CAACAGTCCA
 551  ACTCCAGCAC CAACTATCAC CACCAGTATC TGACGCGCAC GCCATCCGCC
 601  CCGGTCACGG ATCAGGGACT GGGTCTGCCA CTGCCCGCTC ACAACTTCGC
 651  CCATCTGTCT GCCTCCGATT CGCGCATCAA CGAGGAGCTG CACGCCTCGC
 701  AGCAGCTTCC GCGGGAGGAG CAGCGCCGAT TGCTGCGCTA TCACCTGGGC
 751  AGCCTTTTCC CACCGCACCA GGTGCATGCC GTACTGCAGC TCTATCCGGA
 801  GGAGACCGAC GCCAAGACCA TATGCGCGGC TATTCTTAAT TTATTTCCGC
 851  ATAATTAGGC TAGGCTTAAA TTTCATCAAT ATAATTAGAT GTGTAATTAC
 901  ACGCGTGTTT CTGCTTAAAG TGTAAAGTGT CGCTTTGGGA CGACGTGGAG
 951  TCCTCGCTCG AAATTTACAT ATCCAGGCCA ACGATGTTGG CCGTGTCGTA
          CGCTCGAGAATTTACAT ATCCAGGCCA ACGATG  PrimerGW1
                    *  M   D  L  G  V   I  N  A   T  D  Y
1001  GGACACCAGC CCGGCTACGG ATTCCCGTGC GGACTGGCCG CGCCTCTCGT
      S  V  L  G   A  V  S   E  R  A   S  Q  G  R   R  E  H
```

```
1051  GGATGACCCG GTGGTTACGC AGTTTGTACG CCTGCGGGAA TCCCTTGCCG
        I   V   R   H   N   R   L   K   Y   A   Q   P   F   G   K   G   C
1101  CACACATCGC AGCTAAAGAG AGAGAGCAAA CGTGTTTTAA ATTGATTAGA
        V   D   C   V
1151  aTCCtTGATT AATTATGTCT TCCtTATaAC TTATaATATT TAAACACTAA
1201  TACTTCATTC CCTGTTTATC TAATGTTATT TACGCCTCAG GCACCTAATA
1251  AACAATCAaC TAAACTCAAA ATATTAATAT ACTACGTAAT ATTCGTAAAC
1301  AAtAATCTTT TGAAACAGCA TGGCTCTTTC AAAATAGTTT CCACTGTGAA
1351  TTGTACATTC TTATAGAATA AAAAGTATAT TTTTATAAaC GATTTCCATC
1401  TCTCTATATC AAAAAAAAAA CTCAATCTGG TTCTGAAGCT AGTGTATGCA
1451  TAAACAATCT GTGAAACAGT GAAAGTTCGT GCTAGAGAAT ACGGACTGTA
1501  ACCGCCAGGG GGAAAGCGAA ACTTACACGT GCGGCTTCTC TCCCGTATGA
                                      H   P   K   E   G   T   H   I
1551  ATCATTTTGT GGAACTTCAA GGTGTTGGTG TAAGTGAATG TCTTGTGGCA
        M   K   H   F   K   L   T   N   T   Y   T   F   T   K   H   C
1601  AACATCGCAC TCGTAGGGTC GCTCGTTGGT GTGGATTCTG CAAACGGGTA
      V   D   C   E   Y   P   R   E   N   T   H   I   R

1651  ATTATTCAAT CAAATCGGGT TTGGATTGGA GCTCACCaAC TtACCTgTGG
                                                          H   H
1701  TGCTTgTTAC gCGTGgACAA gTCGGCGAAG gCTgCCGGgC AATAgCTGCA
        K   N   R   T   S   L   D   A   F   A   A   P   C   Y   S   C
1751  TTTGTACGGC CGATTTCCGG TGTGGgTGTT CATGTGGCGT GCCAGCTGCT
      K   Y   P   R   N   G   T   H   T   N   M   H   R   A   L   Q   Q
1801  GAGCCTGGGC GAAGCAATGA CCACAGATCC TGAAAAATGA GATTTTAAGT
        A   Q   A   F   C   H   G   C   I   E
1851  TAATGAAGGC TTATTGAAAA TGATTAATTA CTTCCACACA ATCATTATAC
1901  CCGAATTAGA CGGATATATG AGCTAAAATC GTAATAACAC TCCAAAGAAC
1951  ATAGTGAACA CGCAGTAAAA ACTGTGAAGT TACTTACTCG CACTCGTGCG
                                                  C   E   H   P
2001  GCTTCACGCC GGAGTGGACC TTGATGTGCT CGGTAAGGCG ACTCTGGGAG
        K   V   G   S   H   V   K   I   H   E   T   L   R   S   Q   S   P
2051  GGATACATGT TGCCGCAGAC ATCGCAGATG TACTTGAACT CCTTCTtGTC
        Y   M   N   G   C   V   D   C   I   Y   K   F   E   K   K   D
2101  GcCAAGAAGG aTCTTGTGGc CgCCTtTGGT tTTAaTGcCG GAATGcTTGC
      G   L   L   I   K   H   G   G   K   T   K   I   G   S   H   K   R
2151  GCGCCAGCAC CTCGCCAATC GAGATGTCCT TCTCGGATTT GATGCTCATT
        A   L   V   E   G   I   S   I   D   K   E   S   K   I   S   M   K
2201  TTGCTGCCAG CCGAATCACT GTTACCCCGC CGTACCTTAA GTGGATTGCC
        S   G   A   S   D   S   N   G   R   R   V   K   L   P   N   G
2251  GCGGTCAGTT GTCTTGGAGG TACTCGATTC TACAGCCACC GAGGCCGTCG
      R   D   T   T   K   S   T   S   S   E   V   A   V   S   A   T   S
2301  ATGTAGCTGA CCTGCGGGGA TTCACTCTCC GCTTGGTGGC ATTCCGCGCA
        T   A   S   R   R   P   N   V   R   R   K   T   A   N   R   A   H
2351  TGGATGGACG CGCGAAcGCT TTTTGAGGGA ACGAACTCCT CTTCGGTTGA
        I   S   A   R   V   S   K   S   P   V   F   E   E   E   T   S
2401  GTTCAGATCG ACTTCCGCAT CATCCTCGTG GGCACTTTCG GTCAACTGAT
      N   L   D   V   E   A   D   D   E   H   A   S   E   T   L   Q   D
2451  CATGCTCGAC CTGGTCCAGG TATTCGATTT CGGCAGATAA CTCGGACAGG
        H   E   V   Q   D   L   Y   E   I   E   A   S   L   E   S   L   A
```


2501 GCTTGATCGG CCATCTCGTG GTCATAGCCC TGATCTGGGA TGAGGTCGCC
Q D A M E H D Y G Q D P I L D G
2551 CTCGTACGTC TCATACACGT CATAAACCTC CTCGATTATG CCGTCCTCCT
E Y T E Y V D Y V E E I I G D E E
2601 CCACCTGCTC GTTCTTCACG TGCACAATCC GATCCTTTTC CAGCTCAACG
V Q E N K V H V I R D K E L E V L
2651 AGGATACCAT CCTCAATCTC GTCATGGAAG ACACCACCCT GCGTTTCGGC
I G D E I E D H F V G G Q T E A
2701 GGCATCGTCC GCCTCGGCAT AGTGGCTCTC GTCCAGGTCC ACATCCTCGT
A D D A E A Y H S E D L D V D E D
2751 CTTCCTCgTC ATGCTCTGGC TCCTGCTCCG CCTCATCGGG aTCGACATCC
E E D H E P E Q E A E D P D V D P
2801 GgCTCCAGCT TGGTGgCTGT CTCTGAGCCC TTTTTCTCCG GTGGTCGCTG
E L K T A T E S G K K E P P R Q
2851 CTCCACCTCT ACGGGTCCAA CGAACTGGCG CAGATGGCCG TACGATCTTT
E V E V P G V F Q R L H G Y S R Q
2901 GGCAGGTCTC TCGAAACTTA AATGCCATTT TCAGCACCTT GAAGCACTTC
C T E R F K F A M K L V K F C K E
2951 TCGCATATCT TGTCCGGATA GTGATCGAAC TGTTTGATCT GCCAAGGAAA
C I K D P Y H D F Q K I
3001 TTGCATTGTT ATATGGTTCG GGGTCCAATG GGACAAACTT ACGGGCACTC
P V G
GTGAG CGCTTAGTAC ACCCAGTCTA G AAA Primer GW2
3051 CGCCGCACTC GCGAATCATG TGGGTCAGAT CCTTTTCCGA ATCGTCGTTG
G C E R I M H T L D K E S D D N F
CAGGATCCGCACTC GCGAATCATG T Primer CX4
3101 AAAATGCTGG CCATGGGCTC CTTGGGCTGC TGCAGGCAAA CCCGGCAAAC
TGC TGCAGGCAAA CCCGGCAAAC
I S A M P E K P Q Q L C V R C V
3151 GATCCACTTC TGCGTGCTCG TTGCCGCTGC CGCCGGACTC GCATGTGCTG
GAT PrimerCX1
I W K Q T S T A A A A P S A H A A
3201 CCGAGTCTCT CATTGTCGCG AATAGGAGAT GGATCTCTTC CAGGAGGCTA
S D R M
TZF2gene
3251 CTTACCCTCC GCTATAATAT TTATAATAAC ACGCTCGCTG GACGAAGACG
3301 ACGCCAAAAA AATAGATTGT AGTCTGGAGC GAAGTGCACG CGTTGCAATG
3351 CGAATGTCGC CACACAGGGT GACCACAGTC GACGCTAAGG GTATTCCATT
3401 CGCTACTTCA AATATTAAAT ATATAATTAA GAATACGAAT GCATTTTGTG
3451 GGCAGTTAAA TAAATCAGTA AAATTGGAAT TCAAGTGTTT TACAATTTAA
3501 ATATATAAAA GCTTGATTAA GCAGTTTAAT ACAATATTAT ATGTCAAGCT
3551 CATGTAAAGT TCCATCGTTC CATAAGACAG TGTGGTCATT ACTGTACTGA
3601 AACAAAACTA CTTTGATAGA ATAGTGCTTT TCCTTTTTAT TATATAAACG
3651 AACTTAAAAC CATTTTCCTT CAATATTTGT AAATTGAAAA TAAAATTAAT
3701 AAAATGACGC GCATTTTTAT ATTATTTTCT ACCAAAAAAA AATTATTTAA
3751 AACATTTGTT TGGTAAAAAG TGCACATACA GCTATGTACC CTGTACGCTC
3801 CATGTAGTCG TTCATAACGT GCATTCACCG TATCAACTTT GGCGGGCGCG
3851 TTTCGCTGTG AATGGCCAGT GGCCGTGGAC GCTTAAATTC AGCTGTTTGA
3901 ATCTTCTGCA ACAACCGAAG TGATTGTTTT ATCGAATTAG CCATGAAAAC
M K T TZFgene
3951 TGAGTCCAAC GAGAAGTGGG TGGTGTGCCG CGTTTGCCTG AACAATCCCA
Primer CX2 GACTCAGGTTG CTCTTCACCC ACCAC ACGGAC TTGTTAGGGT

```
       E  S  N  E  K  W  V  V  C  R  V  C  L  N  N  P  S

4001   GCGAGGGCGA GGAGCTGCTC CACGACATAT TCAGCGAAAC GGCAAGCACG
       CGCTCCTAGGCA  Primer DM5
         E  G  E  E  L  L  H  D  I  F  S  E  T  A  S  T
4051   CGACTGGACC AAATGCTGCA CATTTGCGCA GGCATTCCAG TAAGTGTGAA
       R  L  D  Q  M  L  H  I  C  A  G  I  P
4101   TCCTGCTTGG TTTGcCCATT TTAAATTATT CCAACCCAAA CAGGTCAGCC
                                                  V  S  L
4151   TAGATGACAA CTTCCCGGAC AAGATGTGCA GCAAGTGCGT GCGCTGCCTG
         D  D  N  F  P  D  K  M  C  S  K  C  V  R  C  L
4201   CGGCTCTGCT ACAAGTTCCG TCTGACATGC CAGCGATCCC ATCAGCACAT
       R  L  C  Y  K  F  R  L  T  C  Q  R  S  H  Q  H  I
4251   TATGGACATG CTGGACCGGG AGGCCAGCAA TGCTAACGCC GCCGGCGAAG
         M  D  M  L  D  R  E  A  S  N  A  N  A  A  G  E  G
4301   GGGATTTGCT TAGCATCGCG GAGGACCTTT CGGTGGAGAG CGTACTGAAG
         D  L  L  S  I  A  E  D  L  S  V  E  S  V  L  K
4351   TCGTGGGAGG ACTACGCCAG TCAGCTGGAT GGCGGCATGA AAGTGGAGGG
       S  W  E  D  Y  A  S  Q  L  D  G  G  M  K  V  E  G
4401   CGAGGAGGAT CAGCAACATC AGGTTATCAC CTATGTCGTG GAGGATGGTG
         E  E  D  Q  Q  H  Q  V  I  T  Y  V  V  E  D  G  D
4451   ATACTGATGA TACCAATATG TTCGATGTGC ACGATCCCAC GCAGCCGGTG
         T  D  D  T  N  M  F  D  V  H  D  P  T  Q  P  V
4501   CCAAATGAGA TCGAGGAGGC TGAAACCTAT GCTGAATACG AGGAATACGA
       P  N  E  I  E  E  A  E  T  Y  A  E  Y  E  E  Y  E
4551   ACTGCTCACC AACGAAAACT CGCCGGAAAT CGCACAGGAA AAAGGCTCCA
         L  L  T  N  E  N  S  P  E  I  A  Q  E  K  G  S  T
4601   CCGGCACAGA TGTTGCCACA GAGGAGCCGC CCGAAGAAGA AATTGCTGAA
         G  T  D  V  A  T  E  E  P  P  E  E  E  I  A  E
4651   GACATACTCG ACTCTGACGA AGATTATGAC CCAACTCATG CTAAGCCGGA
       D  I  L  D  S  D  E  D  Y  D  P  T  H  A  K  P  E
4701   AAAATGCGAC CGATCGGGCA GGAAGCCAGT TGcATACCAC AAGAaTAGCC
         K  C  D  R  S  G  R  K  P  V  A  Y  H  K  N  S  P
4751   CaAAAGTGGA AACCTtCAAA AAGAAGGTGG GCAGAAAGCC GcgCAACAAg
         K  V  E  T  F  K  K  K  V  G  R  K  P  R  N  K
4801   CTGAGcAcat AcAtCtGCGa TgTgTGCGGA AAtATCTATC CGACTCAGGC
       L  S  T  Y  I  C  D  V  C  G  N  I  Y  P  T  Q  A
4851   gCgTCTCACC GAgCACATGA aATTCCATTC TGgTGTTAAa CCACACGAGT
         R  L  T  E  H  M  K  F  H  S  G  V  K  P  H  E  C
4901   GCGAGTAGGT TTAAAAGTTG AAAGTAAATA GTAACTATGg CTAATAGGAT
         E
4951   GTTATTATAG GATCTGCGGA AGAGGCTTTG TGCAGAATCA ACAGCTGGTA
                  I  C  G  R  G  F  V  Q  N  Q  Q  L  V
5001   CGGCACATGA ACACTCACAC GGGGAACCGA CCATACAAGT GCAACTACTG
       R  H  M  N  T  H  T  G  N  R  P  Y  K  C  N  Y  C
5051   TCCAGCTGCC TTCGCCGATC GATCCACGAA AACCAAACAT CATAGGTGAG
         P  A  A  F  A  D  R  S  T  K  T  K  H  H  R
5101   AAATATTCGA AACAGTTGAA AGTAAATTAA ACTGACATTA TTTCTCTCCA
5151   GAATTCACAC TAAGGAGCGT CCCTACGTGT GCGACGTTTG CTCCAGAACG
         I  H  T  K  E  R  P  Y  V  C  D  V  C  S  R  T
5201   TTTACCTACT CGGACAACCT GAAGTTCCAC AAGATGATTC ACACGGGGGA
       F  T  Y  S  D  N  L  K  F  H  K  M  I  H  T  G  E
```

```
5251  GAAGCCGCAT GTGTAAGCAT CAACATATTT TACTACTTCA TTATCCTGCA
        K   P   H   V
5301  ATAATAGCTA TTCTCTTTCA GCTGTGATCT TTGTGGCAAA GGATTTGTGA
                            C   D   L   C   G   K   G   F   V   K
5351  AGGCCTACAA ATTGCGTTTG CATCGGGAAA CGCATAATAG ACGTATCACC
        A   Y   K   L   R   L   H   R   E   T   H   N   R   R   I   T
5401  TGGAGAAATG ACGCAGAAGA GAGCACCAAA GCAGAAGATG TCAAGGGGGA
      W   R   N   D   A   E   E   S   T   K   A   E   D   V   K   G   E
5451  AaCGCCGGAG TTTCTCAATG AACTCCCCAA AGAGTGACAT GTTCTTTTTA
        T   P   E   F   L   N   E   L   P   K   E   *
5501  GTTCTTATGC AAAGTTTAGT CTAAGTATTT AGTAAGCCGT TGTTTAAGTT
5551  CTTCCATTAA GCAAATAAAT GTACACAGGA TATAtTTTTT TGTACAATTT
5601  GTTTTTTATT TCTTATAAAA AATTAAATAA GGAAAATGAC AAATTTTAAT
5651  GGCGGTGCCC TATGGCTTAG AACTTATCCA TTAATATATT GTACAATTTT
5701  CAACGAGAAA CTTATTTAAT TTGCACCTCA ATGTTTCCTC TTAGTCACTG
5751  CAATCATTTC ACCTCTTTCG TCCTTTTGTG TGACTTTTCG TGGGCGTTTT
5801  GTGGTGCAGC TGGGAGAAGG ACTTCTGGCA GTACTGGCAG CTAAATGGTC
5851  GTTCACCGGT ATGGGTTTTC AGATGAACCC GGAGGACGTT GGAGTAGCCA
5901  AAGGACCTGC TGCAGAACTC GCATTTGTAC GGTCGTTCGT TGGTGTGGAT
5951  CC
```

Enzyme Abbreviations:

Bm: BamHI
P : PstI
Sp : SphI
D : DraI
S :SacI
X : XmnI
Pv: PvuII
Hi : HincII
E : EcoRI
A: SalI
H: HaeIII

Appendix II:

The complete sequence of the cosmid 19G11

```
   1  GGATCAAGTT GTGGAAGTTC GAGTTCAATT GGGTAGGGAG AGATACTGTG
  51  GTGAATCACC CACCTGGACG TCGTTGCCCC GCACTCCGCA CTGATTATCA
 101  CGGAGCAGGT TGCGATCGCT CTGGCTGAGA TTGCCGGATA GCAGGATTCT
 151  CATGAAGTAG TCACACTCCC GGATCGAAAT GCAGTGACCG GTGACACGGC
 201  CACTCGGAAT TTTTGCAACA CAATTGATGG GCGGGAGTTG AGCTGGAATA
 251  TCGGAGATCG GGGATCGTCC CACGTTAACT TAAGTTCACA GCCAGACTTG
 301  TTTACGATCC CGTATTCCGT ACACACTCAC CATTCGCGCC CAATGCCAGC
 351  GAGCCGACAA CAACAAGTAG AGCTGGAAAA CTTCCCATGC CGGACCGGAT
 401  CGATCAGATA CGATACAGCT GAGCGGAACG AATGAGACTG CGATCGAAGC
 451  TGAGTCTGAG ATTAACATTA AGTTCGGCCG GTGCCTGTGA ATTACCAGAG
 501  AGTCGCTTTC GGACCGGGTT CAGATAACAT AATCGTAGTT GTGCAACTGG
 551  GGGGCTTTGT AGAGCTTTTA ATGCCCTGCA CTACGGCATA TTCACTTACA
 601  ATGAGTTGTC GAGATTTCAG CGTGTTACGC TTAGATGAGA TCTAGTAATT
 651  AATATTTTAT TTGTGGAATA AGCTCACAAG CTAATACTGA TATACTGATC
 701  GTGATTTTAA AAGTCTCCCA TTTGGAGACG TTAGAAATTC GAATGCCGAA
 751  CATCTTAAGT GGAGCCTCTC ATATGTTCGG CATCTCGGCG TTTAGTTGTG
                WG209
 801  GTGTGAATTA ACCATGACCG GGTTGAAAAT TTGTCGGAAC TCGCTTGCAG
 851  GCTGATGAAG GATGTTTGAA GTCGGAAATT GAATGACTAC ACGATGACCG
 901  CCAACTAGTG TGCCCAGGTG GACGAGATCG CAGAGTTGCC TTGGTCAGCT
 951  AGTATGACCA GCTGTGCAGC GTTCGATAAC GTGATATCGG CGCGAATTGA
1001  ATTTTGGCGC GGAGGTTTTT CCTCTATGAC TCTTCTTTAA ACCTACTTTT
1051  GTTGTTCTTT ATGTCGTTAT TTAATGGTAT TTGTCGGGTA TATGATCTTT
1101  GGCTTGCCTT TACTCGAGAA TCCATCTGGG TATCAGCCGC GAGAGAAATC
1151  TTTCAGTGCA ACCCACACGG GTATCATTCG CACACTCCCC AGGAGGAGTC
                                      V4
1201  TATGGGCTCA CAACTTCAAA TTCAGCAGCA GCAGACATTG TCCGGACCCG
1251  GATCCATCAC ATTAACAGCA TCGCCACAAC TGCCTATCTC TATCAGTAGA
1301  ATGCGACAGC ATCAGCGGCT ACTCCAACAT AGTTCATCTG GCTCTGGGAG
1351  GCCTTTTGCG GCCGATCATT TTATCCTGCC GCCAAGAAAT ATTGCGCAAT
1401  AGTGGTTCCA AAGCTAAACC ACAGTAGAGA TACAAATCAA ACTGCAGAGC
1451  CGACCAATCA GACCGCCGGT GCTCTGATTG GCCGCCCATT GGAGACTGCG
```

```
1501  CGTTATGGCT AAGGTAACCA GTTTCAAGAT GGCAACGCGC ACGTACCAAT

1551  CAGAATCAAG TTAGAACGCT ATATTCGAGT TCCTCGACTA TCAGATTCCT

1601  ACTACCTACT ATACGTACCT ACTACGTTAA TAGGAGTGCG AACTAGAACA
                                                <-------- B8M2
1651  TTTACATATT TTTGGCGTAC CGATACTAAT CAGCAAGACA AATAAAATGT

1701  AAAGAAATTG TTCAAAAGTG GGAGTTTTGG GCGGTTAGTG AGCGTTAAAG

1751  TGGACGTGGT AGTCCTTACT CTCTCGCCTT TATAGTTCCT AATATCTCTA

1801  CATTCATTCG GAAGCAATCG GCTATAGTAT ACCCTTTAAC TGTACGAGTA

1851  AGGGGTATAA AAAACGTAAA GGTGGCGGTG GCTGTAGCTT ACGACACTCC
                                      WG218
1901  TTGACACTGG GTCCCAAAAT AGCAGTTCAG TGTGTTTGGT TTTATTAATA
      ->

1951  TAATATGGTT GGTTCTTCGC TGAATGTATA ATATTATTAT TAACTGTTGT

2001  TACTTAATGG TTATCAATCC TGAAGCATAA ATTTAGCACT TGCTTATTTT

2051  GTTTGTATAA ATAGCAGCAA AATAGTGATA TTCGATAGAT CATGTGAATA

2101  AATATTTATT TTTCTTCAAA AATATCGATA TATTGAAATG TAGATAAATT

2151  TGTCACATCC CTAGTTCGCC CGAATGGCTG CGTGTGTGTC TGTGCGCGCG

2201  CCTGTATTGT CCGCCATCTT GTCAGCCCGA CTTCATCGAA AATTACAAAT

2251  TTAATCGTTT AACGCGTTTT ATGCCCACTT AACACACCAG AAAGTGCTGC

2301  AGTACACATT TTCCCACAAA AAGGATATCG TCTTGAGTAC GCTGCGCTCA
                                <-------------- WG219
2351  GCAAAGGGGG ATAAAATTGC ATTTGAAAGT GGAATTGTTG GTGCGGAGAA
                          WG192 <--------
2401  AAAATTGTGC AGCAAAAAAT TCCCAGGTCT GTGCTGTATG TGTGTGTGAG

2451  AGGCAGGCCA GGGTTGCCGA TCCTCCTATT TATATGTGCA TAAAATAGAT

2501  TTTTAACAGT AAATCCCATT AACTTGTGAT CAATTGAGTG CGAATCGCTG

2551  TAAATAATTA TTTGGCGGGC TGTTATAATT TTTATTTCCC AAAAAACCGC
                                      WG272 <----------
2601  ATTATTCTGC CACTTTACTA ATTCCCAAAA ACATGACAAC CCTGTCCGTG
      ->

2651  GGAAGGGTAC TCAAACAACA AAAACGAAGT ACGTTGCGCG GCGCCACTTT

2701  GACGTTTGAA TCGTCATCAG ACGGTTGGGT AAATTTTCTC GCTCTCGTTG
            ->
            WG194
2751  CTCTCGGTTT CGGCCCCACT GTATGCGTGT GTTGTATGTG TGTGAGCCTT
                                                <----------
2801  TATGTGAGTG TGTGCGGTCC GTGACGCATT CTTGTTTTGC ATCGAAAATC
      <------
      WG271
2851  GCCAAAGCAA AGAAAAGCAG AAAAAGCTGT TAAAGCTAAA AATGATTAAA

2901  TGCTGAAAAT CAAAGCACAG TTAAAAATCA AAAGACCCTG CGTGTGCAAC

2951  TCTAAACGCA AGTGTATCTG TGAGTGTGTG GGTTTGGTGG GGGGCTGTGT

3001  GGCGGAAGGA AGCGAAAGAG CCGCTAAAGA GAAGCAGAAA AAAGGGCGAG
```

3051 GGCGACTGCT GGTCGAAACA AAAAAAAAAG AGAAAAAAAC GGATGGGGGA

3101 AACAGAGAAA TCATAAAAAC GCCGGCAACA GCGATTCTCG AGTGCCCTTC

3151 CCGCCACTCC CCAACAGTTG CCATCCCACC CCTCGCAGCG GCGGCACACC

3201 CACAAACCAC AAAGCGAATG ACAAACAAAA CTAAACAAAA AGCTAAAAGA

3251 AATTCAAGAC GCGCTCCCCA CTGAAAGTTC GTTTACACGG AATTGGTTAT

3301 TTGTTTTGTT TAGTTTTGGA AATTTGTTTA TTGGCGACAT ATGTGTGGTG

3351 ATGAACCTCC CCCTCCGCCC GCTCTTAGTT ATCGTGTGCA TGTGTGTGTT

WG274 →

3401 TATATTGCTG AAATTCTTCA CAAGAGTTGA GTAAAGTCCG TGAGAAACCG

3451 AGATTATCGT TAACTTGGTT TTAACGGAAC AGATGTGTGT GTTTTTTTTT

WG251

3501 TCATACCAAC AACAGCTGCC GCTCACTACC TACAACAACA ACAAAAAGAG

3551 CAACATAATA AAACCTCGAA TAACGAAAAT TTCGTACACT TTTATAAAAA

3601 AACAAAATAG TGATCGTAAG CCTTGGATAT TTTATTGTGT GAATAATTCT

← WG273

3651 GTTTTAACTT AATTCGCCTT ATAATTATGA TACTGTTATA AAATAATACT

3701 TATTGCACCA AGAAATCACA AAACAGCTAA GGCCAAAGGA AGTGTTTTAT

← WG192

3751 ATCGGAGATT TACTTTCATG TTAAAATGGA AAAGTAAAGA AAATTACCAC

3801 ATTAATTGGA TCTATCATGC GGATGAATTG TTTAAACAAC TTAAGTGTCT

← WG193

3851 TTTAAAAATT ATACATGTTT ATCTTTCCAT TTAATATGGA TGCACAATCA

3901 ATTTGGATTA TAAAAATATA TTATTTTGCA TTCGAAAACA ATACCAAACG

3951 GCTTATTTAG GAAACCTTTC AAAAGCTTTA AATATTCCTA CCAGCGTTAT

4001 TAAATTTCCA AATTGAAATA AATATTTATT TGAATAATTC AGTGAGTCAT

4051 TTTATGGAAC CCATTAATAT TTATTGTCTT TTGACCGCGA AAAAGCAAAG

4101 AAAAATGCAA AGCAAAATCT TTGACTGTGA AGCCCATAAT TATCTCGTAT

4151 GACTCACGAT TTTGGAAGTT TGTTTGCAGA CAAAACGAGC AAGATATATG

4201 CTCCATAAGT TATGCCAACA TTGTTGGCAG CACTCCCAAG GCTTCCCCGG

4251 GTTGCTTCCA GCCAGTGTCC ACTGTACCAC TGCTCTGGAG CCATGAGACC

4301 ACTCTGTGGG ATGTGGACGG GCCTACGTGC AATGCAAAAG TGGCCAAGTG

4351 GAGCTGCAGA TGGAGTTCGT TTGGCTTCGC TGGCTGGCTG AATAAACTGA

4401 ACCGCATGTC TGGCTCTTGG CTTTTACCCC GCATCCCGCA CCGTCCTCCC

4451 CACCGCATCA AAAATCCATC TCCAGCTGCC TTTGCGGTGG TTTCCAGCCC

4501 CGCCCTTCTT GCTTTCCCCC ACCCGGTCGT CATTATATTA CGTTTTTTGC

4551 ACATTTCTTG CTTTGCACAA TTTAGCCTTC AGTTTCCGGT TATGAAGGAA

WG195 →

4601 GCAGCAAACG GGTTCGGACT CAGTATCTCC CTCTCTCTCT CCCTCTGCCC

4651 GTTTTCCTCA CTTACGGGCA ATGTAATGTG CATTTAAATG CAATTTCTGC

4701 TGCTGTGCGT CAATGTCGAT GAGTTTGCAA ATATCATGAA AGCGCGCCGA

4751 AAATATGGGA ATCATTTCTC ACCCTTATCG TAAATCTGAA ACACAAGTCA

4801 ATCATCACTA TTTTTAAACA TTGCTTATTT ACTCCAGGAA CAATCACTAT

4851 CTTATCATGT TCATATTAAC AATAAAGGAA CCATTATATG CAGATAGTTT

4901 GATGGCATAG ATTTGATTAA TATATCAGTT TCTGGTTCTT GATAAAACGT

4951 ACTAATATGA CAGCAGTTAT AAAATGTAGT TCATGTCTGG TTATAAACTA

5001 ATTTTAAAAG CTTTTTATCG CTGCTATTTT AACAGTTGTC GGAATTATTG

5051 TAAAACGCCC GACTAATTAG CTGTTGAAAT ATTGAGCAAC ATTTCTGAGA

5101 TACAACTAAA ATGTGAAGAA ACAAGGGTGT TAGGGCCCCC GACACTTACA

5151 AATGTTTTGC AAATGAATGC CTAGTGTTTT TCCTACCAGC ATGTTGAAAA

5201 CCATTCAACT GGTAGTCGTC GTACTCGTAT ATTTTCCAGT TCTGGTGTTT

5251 TTATGCCGCC GTAGCCCAGT GAGCGTAACG CCTGGAAACG CTTTGCTCTC

5301 GTAATTAACC TCTCGACCGA CCGGCTTTCT GCCAGCCTAT CGGTTTACCT

5351 GGTCGCTTGC TACCTGGTCC AAGCAAAGCA ACAATCCATC TGTATCTGTG

5401 TGTAGCAGCT ATCCACAGCT CGTTTAACCA GCAGACCTTT GAGCACGGTT

WG275 →

5451 TCTTGGGCCA TCACAAGCCG AAAACCTTCG TGTTCCACGG CGTGCGTTTT

5501 CCCCGGGGGT TTTCTTCGTT CTTCGCTAAT TCTGCCGAAA GGTGCCTCAA

5551 AATTTTCCAT TAATGTTAGA GTAGCAGCAG ACTAATCGCC TTGAAAACAC

5601 ATGAACATGA TCTATTATTC CATTGCTTGT GCGTTTCACA TCGAGTATTC

← WG276

5651 AATCAGTAGT CAGTCAAGTC GTTTTAAAAT TAAAGAATCT AAACTTGTCG

5701 ATGACTCAGG GGGATATTTA AATTATTCTT TGTTTTGAAC ATAATCGTTC

5751 CACTTTAAAC CCTTTAGCTT AAAAAAAATA ATAAAAATGG TAATAAGGCA

5801 GAGTTGTAAT TTAATACATA TTTTTCAATT TCTTTTTAAG CCACTCGTCC

5851 AATTTGTAAT TGGATTTTTA TTTCATTGTT GAATATGAGC AAATTGAAGT

5901 TGATGTGCTG CGCATTTTTC ATGAAAAACC ATTAACACAA AAGCACTCGT

5951 TTGTATAATA CAGCCATTCC AAATATAGAA AATGGATAGA CATTTGCTGG

← WG196

6001 TATTGCAATA AAATTTTATT TGTTTATCTT AAATGTTTCA CGCTTTAAAG

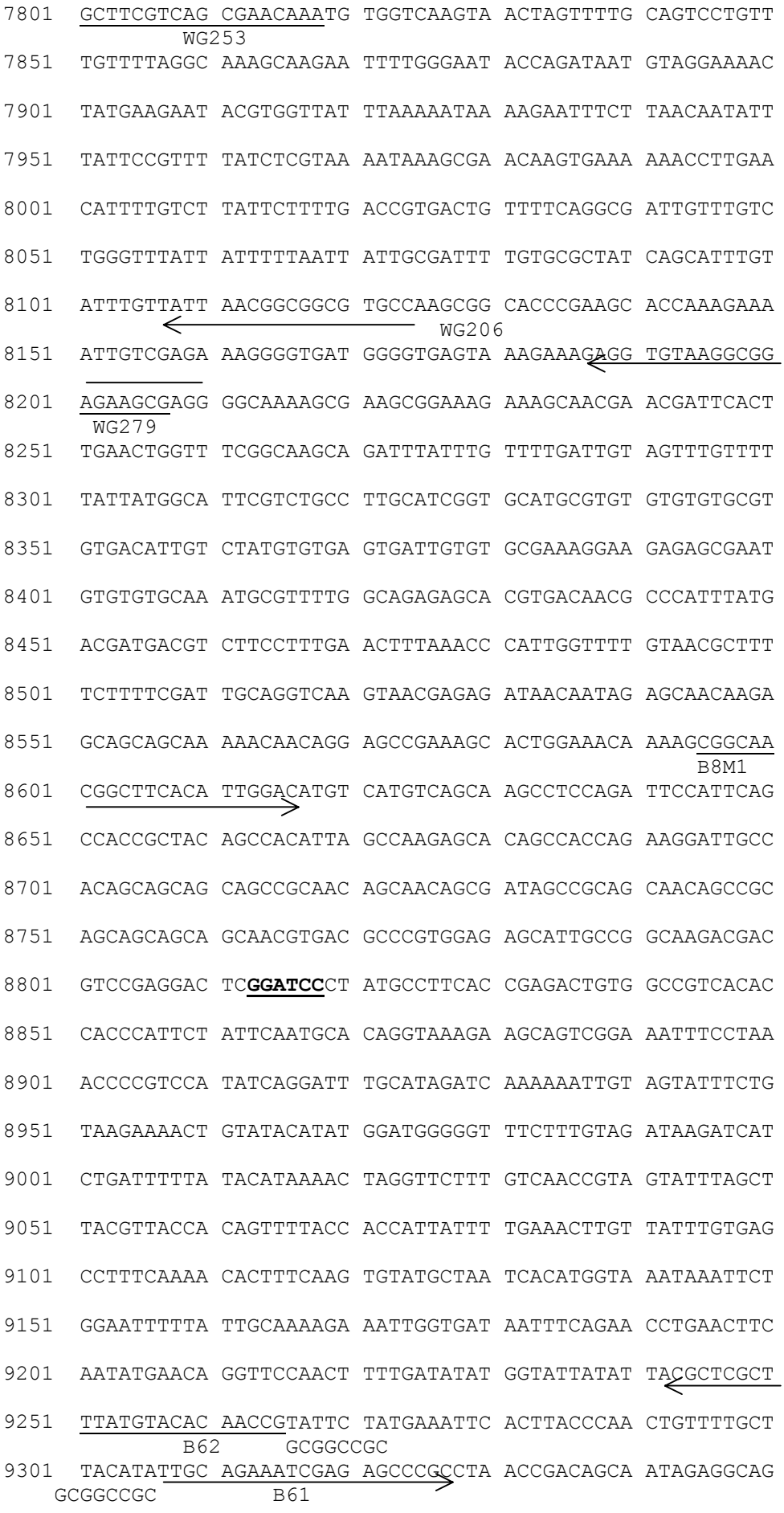
6051 TTAAAACTCC CATCGAAAAG TCACTCAATT GTTTAGGTTA AGTATTCTTG

6101 GAATGGCTAA ATTGGAATGG CATTAAAGTT GCAGTAGAAG GCATTTACAA

6151 TTTTGAAAGA CATTTGATGT CATGGTTCTA TTAGTCTGAC GCCTATTATC

WG223

```
6201  TCGTTTCCCA TCCACTCTAC TTTTATTATG AAGCTTTGTG ACTAAGAATT
6251  AAGCATCTAA AGATCACTGA TCTCGGGCTT TTTATGCCAC CATCCAGCGG
6301  CTATTTGTAA GGTTTCTACC CGCTTGTCTT GTATTTTCTG TTCAGGTTCG
6351  ATGGCTTGTC TCTGCAGCTT TGACTTCTTT TGTCTCGGCG TTCTTTGGCC
6401  TTTCTTGTTT TCTCGTTTCT TCTTCTTCTC GTTTCTCGCT TCTTATCATT
6451  CGAGCGTGTT TCCAAGTTGG ACCAAAAAAA CGAGTTAACA TTTAAGACCT
6501  AAAATGGCCA AAAAGCAGCG ACTTGCTAGC AGATCGCTCG AAAGAACAAC
6551  AAAAGCAAAC AATAACAAAT TCCCAGCATC GTCATTATTT GCAATTATCG
6601  ACAGTTTGCT TTTCTGCGGT TTCCTCCATC TCTTCCTCTC TCTTTCTTTT
6651  TGTTTGCACG CTAAAATCTT GGACTTCGTA TTCGTCTGCG GTCTTGAGAT
6701  CAAAAACCTT TGCTTGGTTG TTATTGAGGT ACATTTATCG AAAGGAAAAT
6751  ACCGATATGA TGAATGTCTT GTAGGTTATT TCATATAATG TTTAAATCGT
6801  ATTCTTAGGT TTTCTTTTAT AGGCAAGGAA ATGATCAGAT GTTTCGCGTG
6851  ATATATATGT ATGTATGTAT ATATATGCGA CTATATAGCC TCTATAAAGA
6901  TTTCAATATG CGACATCCAC AACTGGAGTT GTTTGATTTA TGATCGGGGT
6951  GTTTTCGCCT AGGCCTTAGC CTTTAAGATA TCTTTCAATG TCTGCCTGCC
                WG205
7001  GTTCCGGATG AATAAGTAAA AATAAGAAGC ATAAACAAAG GTGTCACGTA
7051  CACTTTTGCA CTCACACGTA CTGCAAAGGA GTCGGATAGA ATGAGATAGT
7101  GGGTGCGGAA AAGGGGGATA TTCGGATTTG GATTGGGATC TAGTATGTGT
7151  GCTGCTGTGG AGTCGGGCGA AGTAAAGTGA AGGGAACTGC TGGCAAATGC
                  V4                          WG252
7201  TTTCAAAGTC ATTTCCTTCT ACTCGGAAAA CGTAAAGAAA GCATAAACAA
7251  ACAAAGAAAC ATTCCACTCG CAGCTACATA CATATGTATG TATGTGTGTA
7301  CGTGTAAGTC GGATAACTTG TTCCTCCCAA GCACATATTG TAATCATAAA
7351  AATACTTGCC TTCTCGGTCG GAATGGAAAT CAATATGTGC GAGACTAATC
7401  AAAAAGCTGT AAGGCGAATT GAAACAAAAC AACTGAGCGG AAATGTAAAA
7451  TATTTCCCCA TCTCATATTT TTTTTATTTA TATGGTTTGT GATGGTGTTG
7501  TCTGTAATTG CACCGAATCA TTCTCGTATT TCTCCATTTC CTCCCCGATT
7551  CGCTGTAAAA AGTCCAGCAA TTCGCCCCCG ATTTGTGCTT ATCTTGTTAC
                          WG278
7601  TCTTCTCGCA TTTTTTTCTT TCACCGCCTC TGTTTTCTTT CTCTGCCCCA
7651  ATAATTGCTG TTTCGCTACG TTCAATAAGA TTATAGTATT TGTATTCCTG
7701  CGCAGTTGTT TTTGTTTCGT CTGCCATACA TATATTTGTG CACACAAACG
7751  TGCTTCTTTG ATTTGCTTAG CATGCAAATT TGCGATTTTT TCTTTGACGC
```

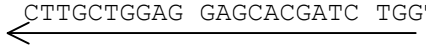

7801 GCTTCGTCAG CGAACAAATG TGGTCAAGTA ACTAGTTTTG CAGTCCTGTT
WG253
7851 TGTTTTAGGC AAAGCAAGAA TTTTGGGAAT ACCAGATAAT GTAGGAAAAC
7901 TATGAAGAAT ACGTGGTTAT TTAAAAATAA AAGAATTTCT TAACAATATT
7951 TATTCCGTTT TATCTCGTAA AATAAAGCGA ACAAGTGAAA AAACCTTGAA
8001 CATTTTGTCT TATTCTTTTG ACCGTGACTG TTTTCAGGCG ATTGTTTGTC
8051 TGGGTTTATT ATTTTTAATT ATTGCGATTT TGTGCGCTAT CAGCATTTGT
8101 ATTTGTTATT AACGGCGGCG TGCCAAGCGG CACCCGAAGC ACCAAAGAAA
WG206
8151 ATTGTCGAGA AAGGGGTGAT GGGGTGAGTA AAGAAAGAGG TGTAAGGCGG
8201 AGAAGCGAGG GGCAAAAGCG AAGCGGAAAG AAAGCAACGA ACGATTCACT
WG279
8251 TGAACTGGTT TCGGCAAGCA GATTTATTTG TTTTGATTGT AGTTTGTTTT
8301 TATTATGGCA TTCGTCTGCC TTGCATCGGT GCATGCGTGT GTGTGTGCGT
8351 GTGACATTGT CTATGTGTGA GTGATTGTGT GCGAAAGGAA GAGAGCGAAT
8401 GTGTGTGCAA ATGCGTTTTG GCAGAGAGCA CGTGACAACG CCCATTTATG
8451 ACGATGACGT CTTCCTTTGA ACTTTAAACC CATTGGTTTT GTAACGCTTT
8501 TCTTTTCGAT TGCAGGTCAA GTAACGAGAG ATAACAATAG AGCAACAAGA
8551 GCAGCAGCAA AAACAACAGG AGCCGAAAGC ACTGGAAACA AAAGCGGCAA
B8M1
8601 CGGCTTCACA TTGGACATGT CATGTCAGCA AGCCTCCAGA TTCCATTCAG
8651 CCACCGCTAC AGCCACATTA GCCAAGAGCA CAGCCACCAG AAGGATTGCC
8701 ACAGCAGCAG CAGCCGCAAC AGCAACAGCG ATAGCCGCAG CAACAGCCGC
8751 AGCAGCAGCA GCAACGTGAC GCCCGTGGAG AGCATTGCCG GCAAGACGAC
8801 GTCCGAGGAC TCGGATCCCT ATGCCTTCAC CGAGACTGTG GCCGTCACAC
8851 CACCCATTCT ATTCAATGCA CAGGTAAAGA AGCAGTCGGA AATTTCCTAA
8901 ACCCCGTCCA TATCAGGATT TGCATAGATC AAAAAATTGT AGTATTTCTG
8951 TAAGAAAACT GTATACATAT GGATGGGGGT TTCTTTGTAG ATAAGATCAT
9001 CTGATTTTTA TACATAAAAC TAGGTTCTTT GTCAACCGTA GTATTTAGCT
9051 TACGTTACCA CAGTTTTACC ACCATTATTT TGAAACTTGT TATTTGTGAG
9101 CCTTTCAAAA CACTTTCAAG TGTATGCTAA TCACATGGTA AATAAATTCT
9151 GGAATTTTTA TTGCAAAAGA AATTGGTGAT AATTTCAGAA CCTGAACTTC
9201 AATATGAACA GGTTCCAACT TTTGATATAT GGTATTATAT TACGCTCGCT
9251 TTATGTACAC AACCGTATTC TATGAAATTC ACTTACCCAA CTGTTTTGCT
B62 GCGGCCGC
9301 TACATATTGC AGAAATCGAG AGCCCGCCTA ACCGACAGCA ATAGAGGCAG
GCGGCCGC B61

```
 9351  CAACAAGAGG CAGACGGCAG CAACAGCTGC GGCCAACAGA AAGGCGAACC

 9401  TGGTGGCCCA ACTGAGTGTC ACAGAGGCAG CAAAGGCGCA GGCGTCTTTG

 9451  GCAAGCAACA ACACAACGAA TTTCCATCAT GTCACGCAAT CTCAGAGACA

 9501  GTCGACGGCG CTGCAGTTGC AATTGCCACT GCAATCCCAG TCACAGTCGC

 9551  AGGCCTCGCC GAAGCGGGCC ACCAACGTGT GCATAGTCCG CCCGCAGCAA

 9601  CAGCAGCTGG AGAAGATAGC CACCTCGGAG TCCTGCCAGT CGCCGGCAGC

 9651  ACCACCACCG CTTTACGCCC ACACTCCATC GCTGTGGCAG ACGCCGCTGC

 9701  TCATAGACAA TGGGCAAAAG CAACAGCTCC TCCAGCAGCA GCATCAGCAA
                                      <-------- B63

 9751  CCGCAACAGC AACAGTCCGT TGCTATTGCG TTGGTCAGTC CGCCCACATC
       ------- CGCCGGCG

 9801  GCCCGCCTCA TTACCTTCGC CCACTCTGCC GCCTGCCACC GCTGCAAGTG

 9851  ACCGCCATGG TGGCACCGAT TTCCGTATCG CCCAAGGGTG GATTACCTTT

 9901  GCCGCCATCG AAGTTCCATC ACACCACACC TGGCGCAACA TCTGCAGAAG


 9951  GTGGAGTGCT TGAAAAAAAA GAAATCCTTG CCACTGGCCT GCCAAAACAA

10001  CAACAATAAC AGCAATTTGC CGAATAACAA CAATGTGGAG TCGCTTAAGA

10051  AACCGGTGGT GCAGGGAACG AGCTACAATC AGACTCATCC GCCACCGCTG

10101  ATGGTTTTCA ATACGGGAAC AGTTGCAGTT CCCGCGCAGA GTCCGCAGAC

10151  TGCTGCTCCA CAGAAACATT CCACCGGCAA CAGCGTAGAT GACAGCGATC
              <-------- 6int

10201  TCAACGAGAT ACCCGTCAAT GTTATCTTCA GAAAGCCGCA AGAGGCAGGC

10251  GGACGGCGAA AACAGGTGGA CCGGGAGGAT TAAGTGCACC TGTTTCGGGA

10301  ACGCCGCAAA CTCGTCCAGC TGAAGTGAAA ATGGTGACTC CTCTCACGCC

10351  GCCCACTCCA CCAGAGATGA GCGCACCGCC CCCTGTAGCG CAAATGCAAC

10401  CCCCGCAGAT ACCCACGTCT TGTGTTCCAG CTTTAGCTCC CAGCTTCAAA

10451  GTGTCGTCAC CAGCAGTTCT CAGCCCGAAG GTGATCTCAC CAGCTCCTGC

10501  AAGCCCAAAG CTCTTGTGTC CGACAGCACC CGCTTCAACG AACTCACTGC
              B64            GCGGCCGCC-------------------->

10551  AAATTGCGCC AAAAGTGTTC CAGCCACTAC AACCTCATCT GCACCAGCAT

10601  TAGCCACGAA ATCAGAACAA ATGTCTTCCA AAGTGGCCAA TTTAAACGCC

10651  TTCAACCGTC AAGCACCCAT AGCTTCAAAG GGCGTTAGAA ATGGCATCAC

10701  TAACAACAAT AACAACAGAA ACAGCAATGT CGTTGCGAAG AAATCGTCAC

10751  CAACTTCGAT GCCGCCTCCA AAGGATCCGA TTGCACCGAT TGCAGCAAAC

10801  GAGTTGACGG ATTCCGAACA TCGTCAGCGA CGCCGCAAGC CCGCCTCCGC

10851  CTGCAGACGC AGCTTGGACA CTCTGAGTGA GAACGAGTCC TTTTCAGTGG
```

```
10901  ATTCGCCCTA CTGCCTACAA CAGCATTGGC TGCACTCCGG CTTTAATAAT

10951  AAGTCCCACG ACGCGCCACT ATCGCAAAGT AACCGGCGTG AGGATCGAAT

11001  TGCTGTTCGT AAAGGAGCCC TGAGGCGACA AGCGCTGCAG CTACTGTCTA

11051  CGCGCTCCCT GCAGGAGCTT CCCATGCGAG CGGCCAAACA GCGGCTGCAG

11101  TGCGTCCAGA ACATGCTGAT CAAGTACCAA GATCACGCTG GCCAGCAGCA
                                                    B8s2
11151  GGGAATGTAA GTTTAATGTA GTGTAATCAT TACCCAATTT ATACTAATTG

11201  TACTGTTTTC CGTTAAAGTG GAATTGGCAA CCACTGTCTG GTGGCCAGCT


11251  GCAAGCAGCC CACGCTTAGT ATGGCAGCGC ACTGCGAGCG TCACATCGTG

11301  AACAACAGCA CCCAGCAGCT TTTCCAGCCT TGCGTCGCCT GGCGGATGGA

11351  TGGAACCGCT TGCCAGGCTC CTGTCTTCGA TGTGCTGCAC ACATTGGCGC

11401  TCTGCAAGGT GCACAGCCAC CTGCGTTCTG GTATGGATGG AGCCCGTCCA

11451  GCATCGAAAC AGCCGCCCGT TAGTTTACCA GTCGCTGGAG TGGCGACTCT

11501  TTATGTGCCT GTGAAGCAGC AGCGAAAGCG CAAGGCTAAT ACGAATGCCG

11551  TGGCTCGTCC TCAGAAACGT GGTAGGAAGC CGGCGAATGA ACCGATTGCT

11601  AACCAGATCA GCAGCCAGAT AAACAAACTG ATACCTGGGG CAGGAATGCA

11651  GCGTAAAAGC AGCACCACAT CGCTTGAGTC CATTGCCAGT AATTCGSAAT
                                                  WG215
11701  CGTCGGCCAC CTCACACTCA CAGCCGCCTT ACAAGCCCGG AAATGTGTTG

11751  GCCGCTGTAC CTGCAGCACA GAACTTGTCC CAACGGTCCA TTCCTCCAGC

11801  GCTGGCTCCG CTCAGCAGTG ATTTTCAACC CAACCATCAG CAGCAGGAAC

11851  CGCTTTTGGT TCCCAAGTTG GAGGTGGATT CATTGTTTAA ATTTGATGCC

11901  GATCAACAGC AAAATCAGCA GCAGCAATCG ACTTTGGACC CCAGCTTACT
                          WG214
11951  TTCTTTGGAC ATTGCGAATA TCAAGGCCGA GGAGATTAGC CAAATTGTAG

12001  CACAATTGGC GGCGGCCGGT GGTGATTTAC AAAATCCCAA CAATAATAAT

12051  AACAACAACA TAGGAATCCA TAATAACAAT AGCGTGCACT TCAACAACAA

12101  CAACAACAAT AGTATGAACT ACAGCAACAA TAACAATAAC AGTTTCCCCT

12151  CATTCAACAC GGCCTTTGGC AACGCCATGG GCCAGCCAAG CAACACAATT

12201  ACGCACAATG GCCACGCTGT TTGGCCAATA CGGCCGATCT TCTTGGCCAG
                                                    L6
12251  GACATGTTTG GTATTTGCGA GAACAGCTCG GCATACGCCA GTTCCGAGGA

12301  TACCGGATTG GGCGGTCTCA GCGAGTCGGA GCTGATAGGC ACTAACGATG

12351  CTGGTAAGTT TCAAAACGGT GAAACTTTTA TGACTTCCAT ATAATAGCTA

12401  TGTATATATA GAAGGAAGCT GGACAAATTC CCTAAGATTT TTTCAGTATT
```

12451 TTGACACGTT ATTTCTCTCT TCCAGATGAT ATAGCATTGA ATGGCGCCCA

12501 CTTGCTGGAG GAGCACGATC TGGTAAATGT GTTCGACACG CTGTCAGACG



12551 ATGCCTTCAA CGAGCTGTTC CAATCCGGTG TGTATTTAAC CACAATTTTA

12601 CTTGTTCAGC TTTTCATTGT ACTAATCCAT GTGTGCCTTC GCTGCCAACG

12651 CCACCGCCAT CGCCTTGGAC CTGTGCCTAT GCAGTGCAAC AAGCCGAGTG

12701 CGAGGCTATG GACCGGGCTT TGGACCGGGC CTTACAGCAG ACAATGGGCG

12751 GCTCGGCGGA CAGCGCCTTT CTTAACGATT TCCTGGACGT CGGCGACGAT

12801 CTGCTGGCCG ATGCTGTGAT GCACTCACCA AACACGTCCG GCATCGATGC

12851 TCCTCCCCTC TTTGGGGACA GCAGCAGCGG CGGTGGCAAT AGCAGCAGCA

12901 ACGGCGCCTC CGACATCCGG GGCTTGGTGC AGACCTAATT CCGGCATCAG

12951 GTCGGATTTA TGGCCTACAG AATTTACAGA TTCATTTAGA GAGACAGAGA

13001 GAGAGATTTT CAAGCTTAAT TTCCCATTCA CTTTTAGGCA GTTTTCGCTT

13051 AGAATTTCGG TATTTTCTTT TTGGCCATTT CTTACCTGCG ATTCTAGTTT

13101 GGCACAATGT TTCTATATGC AGCTTCAATG TTATGCATGC ATTCACCAGC

13151 AGATATGCAT GAATAAAATA GCATTCAAAA ACCATATAGT TATGCGTTAA

13201 TGTGAAAAAC AAAAAAAAAA AACACACACA AAAAGTAAAA AATTTGAAAA

13251 AATGTCCAAA AATATTAAAT TCAACTTTAA ATCAACATTA AGAATGAAAT

13301 GTATATGTAT CCGTACGCAA AAACATAATA TTGTAAATTC GCGGAGAGGT

13351 GTAGCTTTAC GTACATCTTA ACGTAACTTA ATTTGTTAGT GAACCAACAA

13401 AAGGATTGTA GAGATCAGCT TTATTATACA TATAAACGAG AAATTATAAA

13451 ACAATACCAG AGCAAACAGA TAAAATGAAA CCGAAATTGT AGGAAATCTA

13501 GCAACAAATT TTGAATTCTG AATTAAAGTT ACGAACGCGA TACATTGGAA

13551 AGTTCGTTAA CATAAACAAG CCTAAAATTT GTTTAAAAAT TTTGTGAACG

13601 CATTTACAAC AACAACCACA ACACCATACA CTCAAATGAG AATTGCAAGA

13651 ACTACGAAAA CAAAACAGGT GAAGCTGAAA TTGTTTAGTT TGATTTTAGT

13701 TCTTGGTTTG CACCTAATAC ACATACACAT ACACACCCAC ACCTAAACAC

13751 CAACACACAT ACATACAACG TACAAGGGAC TAGTTTGAAA CCGCGAATCG

13801 TATTACTAAA TATATATACA ACTATATATA CTTATATATA TAAAATATAG

13851 CTATATATAC ATAATACCAC TTAATATATC TGTACACATA CTTTTAAGTG

13901 TAAGCTTTAA TTTAAACCAA CCCACAAACA AAGAAAAGAA AACCCTATAA

13951 CAAAAAAAAA TTAACAGATT GTAAAAGTGT ATACTTGATT TATTTTTAAG

14001 CTCTTTTATT TAGCCATTGC TCGATGGCCA CAAACTTTGT TCGATTTTAA

14051 TTAAGCCTAA CAAATTTTAA AAGCAAAATT AGGCTGACCG ACAATTTTTG

14101 AGTTTTAGCA TGCGTTTGTC TATTCTAGAT TATGTAAATT AACAAATATT

14151 ATATAGAACA TAAATGAATT GTAATTTTCT TTAGTCTAGT GACAAGCAGC

14201 AACTTATAAA TCGAGATATA TTTATACAAC AATACACAAA ACAAGCAAGA

14251 AAACTGTTAC GCTTTAAAAT ACGATATATA CCGCTCACTG GAGACCGATC

14301 CCCAAATCAC AGCCCTCAAG AAAACAAAAC AAAAAAATTC ACACCCAAAC

14351 CATTCCATCT CCATCCATTG TAAGTTTTCT CATTTCTCAT CTGTCACGCG

14401 GAGTCTAAGC AAGTATTAAA TGAATGAAAA ACGTTTAATT ACACTAGATA

14451 ACTATGTATT TGTATATGTA TCACGCATCG AATTTTCATT ACGTTAATGC

14501 ACCACCCATG GCACACTCAA ACACTTGCAT CTTTCATAGG AACTTTCGTT

14551 TGGGCATAGC GAAACGATAG AGCAGAAGGA ACTACTTTAA AGATAATGGA

14601 ACTCTGTAAA ATATGATATA TAATAATTCC CCCACACTCT CACTCTCTAT

14651 CTCTAAAACC CTATATAGAT ATATTATATA AACGTATGAA TTCGTTCTTG

14701 AGTTAGTTGA TTACCGATTA CCGGGTTCAC TTGATTACCG ATTGAGTTGG

14751 CCCGCACTTT ACTCTGATCT ACTTCGCTGT GCAAGTGACG CGAACAAAAT

14801 TAATGTCCTT TTTGTAGCGA AACTACAAAA CCATGTATTT GATTATATGC

14851 TTTAAACCCT ATTTTTATAA CACTTTAATT TGTAAGAACT GTAGCGATGA

14901 ACCTACAACC CCGCATCCCT CCAGTAGCCG CCTCTTTTTC ACCATTTGAA

WG212

14951 GTTTTGTCTG TTTTCGGTAT GAGGTACAAT TTCATTTTGT AATTTACGTG

15001 CATTTTTGTG ATCTAATTTG TGTAACTATT ATAAACGAAA AAAAAAAGAA

15051 ATACAAAATA CTTTAAGAAA ACGAAGCAAA AAAAAAAAAA AATAAAACAA

15101 ATCAAAGCAA ATGAAAAAAA GAATTGATTG ATTTTATGGC AACAGAATAT

15151 TTTACCCTGA AAATATAATA CAGTATTTTT AGTAAAAAAA AAAAACCAAA

15201 GCATCCTGGC TTGTCATTGA TCAATGGGAA ATCGGTATGG GGAACTGATG

15251 GCTGGGTATC GGGAATGGGA ATCTCGGCAG GCGGTGACCA GATTTTCATA

L4

15301 GACTCCCATT CGTTTAACAA CATAATTTTC TCCTAGGAAA CAGTTAGTTT

15351 GCCTTGACAC ATACTCCGAG TAACTCCCGA AAACAAAATA CAAAAAACAT

15401 CAGCAGAACC GCCGAATTAA GAAACCCGCT AATCCTTCCC AGGATTCTTT

15451 CAGTGCGTTT CGTCGAGTTG TGGTCGAGGA TTCAAATTCA AAGGTTATAT

15501 TGAAAATTAT TATTTTCTAT TTTGTTTTCC TTGCTCGACC ACCAACCCAA

15551 TCGCATCTAA TCGCAAGGAG CATTCAGTCC AGTGCAAAAG AGACAAAAAC

15601 TGACCAGATC TGGTCCGGAT TATCCCCGTT TTGCTAAGTT GATTGAGTGC

L3

15651 CATGTGGCGG CATTTCCTAC TCTGGGTTGT CCGTATGCAG CGATTATTAC

15701 ATCATATCCT AGCAGGCGTG TCGTTTGCTG TCGGTTGTAA TGCATACAAA

15751 TGCCTATCTG AGCTGCCGTC TCCCCGATTA GTCACTTTTT TTGTACGTTT

15801 GTAAGCTGCC GCCAGTTTTC AGAGTGGCGC CACGGGGATA CGTGGAATAG

15851 CGTGGTAAGT GGGCGCCACA TGCTCCACCA TCGACCCCAC TAACCGACTC

15901 CTCACCACCA GTTGTGCTCT ACGTATATTT TTATATCATC ATTGCGGAAC

15951 CACAAAGCTC TCGACTACTT TCTAACTGAG GAACTGAATC AAAGGTGAGT

16001 TCAATTCAGC ATATTCTGTA TATTTGCGCT ATGACTCTTT CCCGCATAAA

WG211

16051 TTGATTTAAA TAATATGTAG TTATCACCGA CAATTATTAC TTTATTTGCT

16101 AATACAATTT CCTTAATTTG TACGAATGCG TATGTCGTAG GACATAGCAG

16151 GACTAAAACT GTGTTTTAAG CTTCGATATA TTTACATGGT CCCTACAAGC

16201 TGAACTTAAG CTATCCTTAA TCCAAGTGCC ATAAATAAGA GGGGGGGGGG

16251 ACTGTTTGGT GCTATAACTC CGTGGTCTCC AGTTTTCGGG CCAGGATCTG

16301 AGCATTGTTG GTCATTTTGT CGATTTGGTT GTATAGATTA AAGTGCCGAT

16351 TAAGATCATT GCTGTTGTTG AATAGTTTGG CAGTTCCCTT GAGCGTTTCG

16401 TGCATATCCT GTGTGGCCAC GGCCAGAAAG TTATCATCCA AAGAGTCATG

WG210

16451 GCCACCAATA CTATGGAGAA ATAAGTGTTA TATATTTAAG TAAGGGAAGA

16501 TTTAGTGATT TCCTCTATTA TCACACCTGC CACTTTCTTC TGATTTGAAC

16551 CACTCGTTGT CGTAACCCTT TAGCCAAATA GGATTTGATC CTTGGACACT

16601 GTGCTTGTTG GTATTTGGCA CTCGAGTGGC CGGCTCCTGA GCGTTTTGCA

16651 GGGACATCGA GGAGTGAGTA CCTAAATGGA ATTCGTGTAT TAAATACATT

16701 AATTTCTCTT GTATCCTAAT CGTTACTTAC GGAATATATT GACTTTAGCG

L2

16751 GCTCTCAGCT GCCTGCGGTA TCCATTCCGT TGAGAGGCGC AAAGTGCTAT

16801 AGTGACTACC AGTAGAGTGG CCAGAAAGAG GTTGGTGAAA ATCAGCCACA

16851 CGAATAAAGG ACCTCGGCTG GGTCCTGCGG TCAGGAGCTG AGCCTCTGCC

16901 GCCTGCGTGT CCAGCACGTT GAGCTCTTTA AATAGCCCAT TCAGGCTTTC

16951 GATGTGCGAG TCAATCAGTT TGAGCACCTC GGATACCTCG TAGATTGAGT

17001 TATCCTCCCG CTCGACTAGA TGCATGTACA GATCCGTCTT TGTCTTGTCC

17051 ACCGAACCGT CCTTGTTCTC ATGAACCTTA ATCTCATCAA TGTTTACAAT

L1

17101 CGATTCTGTT ATGTTGCTCA GGGTGCTAAG GTATGGAAAT AACCATTTGA

17151 AGGGGGGTTA GACTTTAAAA TCAACTCAAT ATTTACTCTC TGAAGGAGTC

17201 CACTCGAGAT CGCAGTTCGT CCGGCTGAAG TCTAAACACG AATCGCACCT

17251 TTTGGTCCTC TCTTAACAGG TAGATAAACA CATGGGCCAC ATCCTTCATT

17301 CCTGAGGTGT CGTTGGCCAA AACGACGAAA TCGAAATAGC CCTTCATTCC

17351 CTTTTGTGGG TCAAAGTTTA GCTGCACTTC TCCGGTTTCC TGATCGAGAA

17401 GGAAGGGAGC TTTGCGTACG TTCTCCAATC CCTCGGAAAG AGTCTGTCGG

B8s1 →

17451 ATTTCACCCA CTTGGTAGTA TCCAATCCTT CCATTATCTC CTTCATCTGC

17501 ATCCGTAGCC TCCACTCGCA TGAATTTTAA ACCGAAATCA GCATTGGTCG

17551 TTATACCACC AGTAAAGATT TTAGAACGGA ATCTCGGTGG ATTATCGTTG

17601 ATGTCCAGCA CGCGCACCTT TACTTTTATT ACAGTGCTAT CTAGTTCTGG

17651 ATGGGTTGTT GCCAATCCTC CGGTT**GGATC C**TCCTGACTA CTGTCATAGG

17701 CTTCATACGA TAACAGCTCT GCATCTTCTG GTCCGTATTC ATTACTCTCT

17751 GTCGATCTCA ATTGTCGAGA GTGTTTGAAA CGATCATATC CAAACTGTTG

17801 GTTATGCACA ATGCCTCCCA ATCTATTGTT AATCTCGATT CCCATACGAC

17851 GACGTTTTTC GCCACCAGAT ATAGAATCAT TTCTACAATT CTCAGTGGCA

B12 GCGGCCGC ←

17901 CGAACATAGA GCGTAAAATT GGCAATCACC TCGCGATCCA GTTCACGGTC

CGCCGGCG B11

→

17951 AACAGTCAAT ATATGGGTTT CGGGATCGAG ACGAAAGTAT CCCGCTTCGT

18001 TTCCATTCAC TATAAAGTAG CACACCTGGC TGGGCGTATC ATTCGGATCG

18051 TCCAGTTCCA ACTGATCCTT ATCTATGGTG TCCGGTAGCT TTATTCTTTC

18101 **GGATCC**T**GGA TCC**GAGTGCT CCGTGAAGTT TACGGATATT TCGTTAACCA

18151 TAAACTGAGG CTCATAATCG TTAACATTCC GAACATAAAT CGTTAGATCC

18201 AAGTCCGAAC TCAGTGAGGT TGGCAATCCC TGATCATAGG CTTCTATACG

18251 TATCTGTAAT GAAAATTATA TCATATAGTA TATTCTTTTG TGTGTTTATA

18301 CTTTAAAATT TGTAGGTATA AGGAGATTTG GCATTTACCT CGTGAATCTT

18351 CTGCTTTTCG CGGTTAAGTT GTTCCTTCAG AAAGAGCTCC CCTGTGCTGC

18401 CGTCTACTTC AAACATTTTG TAGCTGCCCA ATGGATCGGG GTTAGTCTGT

18451 ATCTCACCAA AGCATTCTGC CCCATATCCC CATCACTGGC ATAAGCCTGT

CGCCGGCG B52

18501 AGTATCAGCG TTCCCACAGT CGCATTCTCC GGAATGCGAA CAGTGGTGTT

18551 ATGAAGTGGC CTCACAAAAA CCGGAGCGTG ATCATTGAAA TCCGATACGC

18601 AAATATCCCA AGGTGTTGTG CACAATATTG GCTGGGAAGG CCCATATCTC

18651 GCGTAGTTAT AGTAAGGCTA TAGTTTCCGT ATCTATTGCG CCAACTTTTG

18701 CTAGCGAGAT TTGGGCATTC TAGCGTCTAT CTGGCGCCTT CCAAACATAC

18751 CTTTAGTACA GATACATTTA GTGTTTGAAA GACGACTTAT TCTCCTTTAA

18801 ATCCTTACCA TCTTTGTTAC CCCGAGTGAT GGCAAAATCG AGTCTACCAT

18851 TTGGCGTAGT GGGGTCATCT GCATCCTTGC CCACTATGCT TGCCACTCGT

18901 TCATGCAACT CGTGATACTC AGTGATTAAA ACACAACTCA TGGGCAAGTC

18951 AACCTCTGGT GGATTATCGT TCTCATCGAG TATGGAAATG CTGAAAGAGC

19001 TAAATGAAAT AAATTTGTCA AATGCTTAAA TTTTTATAAC TTACAAAACC

19051 TGTTTGAAGG CATTGCGACT TTCTCCGGCC AGGAATCCGA ATTGATAGTT

19101 ATCCCAGGCC TCGATAACCA GGTTGTAGGA ATCCTTAGAC TCCCTGTCCA

19151 AACGATCGGC CACCGTTAGC ACACCAGTGT CTGCATCTAT GGTGAACTTT

19201 CCTTGCGAAC TGATGCGATC CATCAGGAAG GTGATCTTGC CAAAATCTCC

19251 CGAATCCGCA TCTCCAGCCT GGAGAGTGGC TACTCTGGTG CCTGGAGCTG

B51 GCGGCCGC ⟶

19301 CATTTTCACT AATCGTATAG TTCTTGCTGC CACCCACAAA GTAGGGATTG

19351 TTATCATTCT CATCCAGCAC TGTTATATAC ACATCCACCA GTGAGGATCT

19401 TGCTGGACTT CCAGAATCCA TGGCCCTCAC ACTGAAATTT AGCCACTGAT

19451 GCTGCTCATG GTCAATCTTG CTGGCCACCA CAATTTCCCC TGTTTGC**GGA**

19501 **TCC**AGATGCA TTAGTGATCG GTACGTGGGA TTTCCTTCCA GGGCATAGGT

19551 TATGGTGCGA TTCTTGTCCA CATCTGAGGC CAGCACATTC ACAATCATTG

19601 CCCCGTTTAT GCTGTTCTCC GTGATCGATT GACGGTAAAA GGGAAGTCGA

19651 AACTTTGGGT TATTATCATT TTCATCCAGC ACTTGAATGC TCAGGAAACC

19701 CTCAGCGATT TGTCTGCCCT TGGCCGCAGC CAGATCCTCA ACAGTAATGG

⟵ B42 CGCCGGCG

19751 CTAGCTTAAT ATGCTCGACT CTTTCGCGAT CCAACAGCTT AACCACCTTG

19801 AGGGTTCCCT CGATGGAGTC CACCTCGAAG GCGCCCAAGA AGTCGTATTC

19851 GCTACTCTTC ACCAGCGCCC CCTCTTCAGT TCGACCTTCG CAGTGTTCTG

19901 GGTTCAGTTT GTACCGCAGA ATGGCCTTGT GGTCCAGGTC AGTTGCCTGA

19951 ATGCGATACA CCAGTGTGCC CACAGGCGTG TTCTCTAAGA TCTGCAATGC

20001 TGGCATTTCC TTTAGCACAG GCGGCTTATT ATTCACATCC TGAATGCTAA

20051 TGTTCACCGT ACATGTGGTC ATCAACTGGG AATTGCCCAA TCCACCATCC

20101 AAAGCAATCT GTGGTAAACA ATTACGAAAG GGTTTTAATA TTCTACAAAA

20151 AGGAATCTCT TTTGAACGTA CCACTGACAG AGTATAGAGT GATCTCTTGC

20201 TCTCTGTGAG ATCTGGATCT AGATTGGCTC CGTGTGCCAC CGATATAACT

20251 CCCGTTTCCG AGTTGATGAT AAACTTATCA CCGGCACCCG TCTGTATGCG

```
20301  ATACACCACC ACATTGTTGG GTGCTGATCC ATCTCGATCT ATGGCCCGTT
             B41   GCGGCCGC------------------------>
20351  ACCTGCAGCA CCGAGCTCCC ACCTGGCAGA TCCTCGGGAA CCGTCTTTGC

20401  ATAGAAGCTT CGCTGGAATA TCGGCGCATT GTCGTTAACA TCCTGCACGT

20451  AAATAAGAAC AGGAACCACA GTGGAGAGCA TTGGAATGCC CGAGTCTCGT

20501  GCTCGAACCA GCAGATCGAT CTCTCGAATG CTAAAGGATC CACCCGTGTA

20551  AGGATCACTT CTGCGACTGC TCCCATCTAC CAGTTCCTCG AAATCAAAAC

20601  TGTGCACCGG ACGCAGAAGG CCACTCTGGG GATCTATAGT GAAGTTGGAG

20651  CGGTACAGAC CCTCGACAAT CTCGTAGGTC ACCTGGCTGT TCTCCGTGCC

20701  ATTGAGATCC GCATCTCTGG CCTCCAATTG CAGAGGGGTT TCGAACTCAG
                                                  <------------
20751  CCTGATTCTC CAGCAACTTG GTCTCGTATT GCCGCTGAGG AAATGTGGGT
              B32      CGCCGGCG
20801  GCGTTATCGT TCACATCCAG GATGTCTACG ATTATTTGGG CCGTATTCCT

20851  GTTGCCCTGA CCAGCGTTGT CGATGGCTTC CACGGTGAGA TAGTGTCGTG

20901  AAATGATTTC ACGATCGAAG GCAGTTCCAC CTGCCTGTTT GATGGTTATC

20951  ACTCCCGTAA TGGGGTTGAG ATTGAGTCTT AAAGGGGAAA GGTATGTTAG

21001  ACCCGTATCA TAGTTTTAAT TTTAAAACTT ACAAATTAGC GATTCCTCCT

21051  CTGAGGTTGG TGTAACGTAT ACCCATAGTT CCATAATCCC CAGAATCCAC

21101  ATCAACCGCC TGGACATGGG TGATGATTGT ATCCTGCTCG CTGTTTTCCA

21151  AAACGCTCGC ATTATAAATG GTTTGACTAA ACTCTGGGAA GTTGTCGTTT

21201  TGGTCTCGTA TAAAAATCTG AACATGAGCA GAACTATATA AGAGAAGATA

21251  TATTTTAGAC ATTACTCTAT ATGCATTTTC CAGCAGCCAT AAGCTTACCT
                                        B31   GCGGCCGC
21301  CCATCGACTG GGCTCGTCGA CCTCCCTGGC GAATATCGTA AAATTGACCT
       ---------->
21351  CGGTGAATTG TTCAAAATCG AGAGACTTTG AGTTCTTCAC CCGCAGCATA

21401  AAGTTGGCCT CATTAACAGC CAACTCGGGT ACGATCTCGA ACAGATCGTT

21451  GGGCGGATCC AGAAACAAGC GGAAGGTGCC ATTGTTGCCC TCATCGTGGT

21501  CGAAGACCAC GTTTTGCACC TCCTCGTCGA TGAAATTAAG CGGGGTATTG

21551  GTTTGGGCAT TCTCGTTGAC CTCACAACGA TAGACTGTTT CTCCAAATGT

21601  GGGTATTTCG TCATTTACAT CGCTTACAAT GACCGTAACT TCGGTGCGTA

21651  CAGTTGTGGG CGCCATTTGG GTATTCGACT TGGACAACTC GGTGGCCGAA
                           <------------------------------ B8B1
21701  ATACGCAGGA TATGAGCTCC ATTCACCTGA TCACTCTGCT CCTCGCGATC

21751  CAGTTTGGTT AGCGTGTGAA CTATGCCTGT GTGCGGATTA ATGTCGAACA

21801  GATCGTTGGC CTCCAGGGAA TAGGCTATGG GATTATTGAT CCCACGATCT

21851  CCATCTATAG CTCGAACTCG CAGAACTTTT GTGCCCACTG GAGCATCTTC
```

```
21901  GGCAATTCGG GCCACTGCCT GTACTTCTAC GAACTCCGGC GGCTGATCCT

21951  CCAGATCCTT TACCTTCACC AATATGGCAG CAGTTCCCGT ATTGATCGGT

22001  CCCTGATTGG CACGATCGAT GGCCAGGACT CTCAGTTGAT ACAGACTCTT
         WG221 ------------------------>
22051  TCTCTCATAG TCCAGCTCCT TTTGCAACCT CAATATGCCC TTGCCTTGGT

22101  GCGTAGCTAT GGAGAACACA TCATTGTCGC CATCCAATTC TTGCAGATAG

22151  TAAACCACCT GGCCGTAGGC CCCTTCATCC GCATCCGTGG CCTCCAGAGT

22201  GCTGACCACT CCCGGTGCAC TGCCCTCCGG TATTTCAATG GCATTTTGAT

22251  AGGGCAGAAA TGTGGGAACA TTATCGTTTA TATCCTCAAC CAGGAGCAAA

22301  AAACTCTGGG TTACATAGTT GTGATCGCTG TAGTGACTGT CAGTCAGCGT

22351  TAGCACTATG GCATACTCGT CCTGCAGCTC CCGATCCAGT TCCTTGGCCA

22401  AAAAGATCTT GGCCTCGTTG CCACCCGTGT TCTCAATCCG AATGATCTCG

22451  CTATCATGCG AATTGCGCTT GCCAAAGGTC AGCGGATCGT TGTCCGGATC
                          <------------------------
                                          WG220
22501  GTAACCCTTT AGCGTGTATA TCAGGGTTCC TTACGAGGAA AAAAAAGTGC

22551  AAAAATTCAA TCAAAAATTA TTTACTATCT ATCTTTTGCA GATTATTTAT

22601  AGGAAATAAA TGTACAGCCA TTCAGCAAAA GATGATTATA CGGATAAGCA

22651  AACCATTTTT AAACACGACA CATTGTCATT GTCTGGTCGA ATTTATATTA

22701  CGATTTTATT TTGCACAACT TCAAGATTAA GTATATTTAG CTTAAAATTG


22751  TTGAGACATG ACACACTACC ATTCTTTGTA TAGATAATAA GTTTGCTATC

22801  TGATATTTCA CATTTCTGCT TAGCTTTTGA AATAATTTTC TTTTTTTGTA

22851  ATATATTAAC TCTAGACTGG ACTTACCCAC TTTCGTTTCA GGACTCTCCT

22901  TCAGACGCAG TACAATCTCC GACTGGCCGT CGATAGCGAA ACGCGGCGGT

22951  CGATTTGAGG CAACCAAGCT GATTAGTGTG GTGAGTATTA TGACTTGGAT

23001  GGATATGGAG CTTATATAAC TAGAGAAACT TCGCCGCCTT TGTCCTTTTA

23051  AGCGCACCTT CTTGTAGTCC ATGCTGTCGA TTCCATTACC ATTCGCCGGC

23101  GGCTGTGATT CGTTGTGCAT CGTGTGTGGC GTTTGGGAAC GACACCTTTG

23151  GGGTTACTGG TTATAATAAT AATAGTCACT ACCACATGAC AGGTAGAATC

23201  GAGCGGCGAC ACCTAGAAAA TAGACGAAGT GTTACGACAG AAGCCAAGTG

23251  GAGTGTTAAC CCATTTTCCC CACATGTCGC ACGTTTTATG GGCCGTGATC

23301  TATGACCCGC TGGCAGTTTC GCTCGAGAAG CTAACTTAAA GTGGAGGGAA

23351  GAATCCACCA GAGAAGTTCT GTGCTCGTTT GCCCAAAAAA AAAAAAAAAC

23401  AGAAACAAGA AGTGAATTGG GTCGTGCATT AAGAAATCCG AATGGAACTG
```

23451 GACTCGCAGT CTGCGTGTCT ACAGGTGTTC TATATGTTGT CCGTCTGTTC

23501 TGCGTGAGAA TTGGCCAGTG GATTGGCCGA GGCAGATCTT TGCGAAAGTT

WG233

23551 TTATGCTCAA ATTAATTAAG CGTCAGCTCT TTATTGGCTT TATACTTCAG

23601 TTGAGTTTAC TTAAATGCTG GCTCAGCGCC AGCCTGGAAT TCCGCTAAAG

23651 AAATTTTATG GGTTTCTTCA CAGCTGCCTT TGGGTTTTGA TTGTTATTTT

23701 GTTCGCGGTT TAACGACCGG AACGTTTTCT AAACTGCCCG ACAATTATTA

WG232

23751 CTATTTTGGC TTATTCATGT GCCAGATTTG TCTCATTGCA AATTTCAAAC

23801 AAATTGCGAA TAAGTTCTCG CTCAAGTGGA GTGAAAATTG AAAGTTGTCT

23851 TCGAAAGATG AAGCTATCAC GAGCTTAATG AAAACCGCAT TTGCGAATAT

23901 TTTTAATGAT GTTGCTACAC TTTTGGAAAA ACTTCCATCG CCATCGGCTT

23951 ATCTTTTCAC AGAGGGGGGT GCGTGTGAAA AATGCATTTG TAGAGCGATG

24001 ACTGCATCTG TCAAATGCAG ATGCATATAT GCATTATTAA TGCCATTGTT

WG227

24051 CGCGAGTTCA ACTCGGTTAG GCTCATTTGC AATCGGAGCT GCGAATCCGC

24101 AAGATTTGCC AACTGTGGTT CATTGATTTC GCAATTGCCG CACGGCAATT

24151 TCTTAATGGG CTGCCATCGT AAATGCCTCC GTGGATGGTA ATTGATTTAC

24201 CTGCCACTGC AGGTTGCATC AACCTCTTTA TCGCCGGGAA AAATCGCGGC

24251 AAGCAGCCAG CAAGGTCGCT AGAGAGTCGC CTTAACTGAT CACGGTTAAG

24301 TGATTCAAAA GTAATCACAA TCTGCGCCTG ATTTCGCTTT CAACACGCGC

24351 CGCATGCAAC ATTTTCAATG GTGCGACTGC CACTAACACA ACTTGTATCT

24401 GTATCTGCCG CATCTTCTGC TTTTGGGCGG CTAATTCTCT CCATCGCCGT

24451 CCGCGGATAA TTATCCATGC ATCCGTGGCC AGCATGTTTG TGTAACGTTG

24501 GCAATTGGGT TGTTTGGTCA GTTGGACGAT GCAGGGATTT TGGATGCTCT

24551 GCTGCCCAGG CAATTTCATG CTGCACTGGC TCAGTCGAAT CGGGGTTATC

24601 ATAGCTGTCG ATTATTTACG ATGGGTAAAT GAAGAGAACA CTTCGCTCGG

24651 TACATAAAAT TGTATGCTTT TGTTCTGCGT ACAGATATTG TTATTCATGT

24701 GTTTTTTACA ACACCCATAC GCCTTTGSAT TTTCGGCGAT CTTATAGCTT

WG225

24751 TCTAATTAGT TCAATTTATA TAAATGAAAT GAAATCAAAA TACGGTTAAG

24801 AAATAGAGGA AATGCGAGGG GTGGAAGTTC CGCCACACTT TTCATTAGCC

24851 TAATGTCTGG GCTATCACTT CCCTTCAAAT AGCCCATTCA AAATCGCATT

24901 TAGTTTAGTT AACTCCATTA GTGGCAAACA CACTCCCACT TTCTTGCTGC

24951 CAATTTTCCA CACAGCTAAA ACCCACTGAA TTAAGGCAAA TTAATAAGTT

```
25001  ACGCCGCCTC GGCAAGATCG GCAGGGCTCT GCTCTGCCAT TTCCCCTCCA

25051  GATTTACGTA AGTGGAAAGT GGCAAAAGTG GCTGAGGCTG CTGAGGAGTC

25101  GCCGAAAGTC AAGCAAAACA CTTGGCCCAG CCACCATGAC GACGGCGCCT

25151  TTCCGCAGTT CAGATATGTA GCTCTGCGAC CTGCAAGATA CACATGTGAT

25201  GTATCTCAAT CCGAAACCGA GCGAATTTCG GTGCTAATGA CTCTGGTGGC

25251  AGCAGGGTTC ATGCTTGGAC AAGTTCTGAC CCCATCCGAG TGGGCCAAAA

25301  AGTGAAATCC TTTGCGAAAA TGTGTCCAAC GTACGCGCGG TAACCAACGT

25351  CGGTTGGCGT CGCCAATCTT GCGGCTTTTT GGGTTAACCG CAAACAAACC

25401  TGGTGGCAGC TCCAAGGTAA TCGCCACCGA AATGGAAATC CCAATAAACA

25451  ATTTTGATTT GTCTGCCGCT GTGACGAGTT TCCCGATCTC TTTCACCGCC
                <-------------------------------------- WG224
25501  CCCCTTTTTC CAGCAACAGT ATAATCGCCG GTTCATGACT TTCCAAGCCA

25551  GCCAAGGGCT TACTTAATTA TATTAAATAT TTTAAGCCCG CCTGAGAGCA

25601  ACAGTCACAT ATGTACACTC GAAAAAATAA GTGCTCTGCT GTGGGTCAAT

25651  ATATGTATGT TTAGCTAATA TCAAAACTAA TCTCAGAAAT CTAAGAAACC

25701  TTATAGATCA GGGAACTATC CAAAAATCAT AGAAATGCAT GACAAAAACA

25751  TAAAATGTTA ATAATACGAT TAATGAGAAG TCGTTGATAT CAACCCTTCT

25801  AAAATATATC AATATTTTTT TCCGAGTGCA GCCATGGAGA TCGAGATTGG

25851  CGGGAATCGT GTTGTGACCA ATGTTGCCTC CAGTTTCCGG CCCGCGAGCA

25901  AACAATCGGT CCCCTGGCGT TCAGGCGTTC CGAGTACTCC AATTTGAACA
                        WG229 ---------------------------->
25951  CTTCCTCCCC CGTATTAGGG CCCTTTTGTC ATGCCCATCG ATTCCGATAT

26001  GTGCGGCACT CCGACAGTCA AGTTCAACGC CAGCAGCCAA CGATTGTTTG

26051  CTGGTGGCTG GCTGGCAAAC AAAAGGAGCT CGAGCTGGAT TGGGCGGTGG

26101  GGCAGGAGAG TTTGGGGAGC TGACTCCGTT GCGGTGGCAT GGTTTTTGAC

26151  TCATAGCCCG GGTTCAAGCG ACGCCCGCAT CCGTCTCACT ATTCGAGCCA

26201  AGCTCTCCGA GACTTGGCTT AGATGTTATA CTATTGTACT ATGCATTCAT

26251  AGCTGGTGTC GTGAGCAATT TGTCCGTTAC ACTCCGGCTC CTTTTCTGGC

26301  TTAAGTTTTT GCACTTTTCT ATTTGCACCT GCTTCGCAAG TATTTCTACA

26351  TTTTGGGGTA ATTACACAAT TTTTATTGGA CTGGCTGGTC TTGGAAGAGC

26401  AGCCAAGTCC TTGACACGAG CTATGGCCTT AGTGGCAGTA TTACAAGAAG

26451  AGCAGAAGTC AGCCAACCCT TCGAAGTTGA ACTTCAGGCT CCAATGCACT

26501  CGACGAGCTT GTAAAGTGCG AATTTACTAC TCGAGAGTTT TGGCAATCAC

26551  TTAACAATAT TTTAATAAGC CAATAGATAT TTCTAGGTTC ATTAATATTT
```

```
26601  AATTCCATTT AATGCTATTA ATAGTTTGAT TATAGGAGAA GATCTTCTTG
26651  GTACAAATAT TTTTACATAC CGTGGTAAAA GGGTATATAA ACTCTGGTAC
26701  TATCTTTGAG CTCTTTTTCT TTCCGTAAAG TGTTTTGTTT ATGTTTTTTC
26751  ACCCCCAACG ACTGCTTGCA AGGAAACCTC CTAACTTCCC ACTTTATCGA
26801  AGGAAAGTCG CATGCCAGGA CAGGAATGCA GACCGGTTTC GCTTACTTAC
                WG230
26851  CCCCGTATAA AAGATAGATA TTCCTATAGT CCGAGTATCC TTTGGCGGTT
26901  TTCTGGTTGG GGTCCTTTTC CCTATACAAT TTTTGAAGCC AGCGAACAGT
26951  GACACACTCG AGTGTCTTTC AGCAAGTGTC TTTCACCAGA GAACTGACAA
27001  ATTTCGTGGC TGACACATAT TTTCGAAAAG GCTTCGACGT CACTGGTAAT
27051  TAGGCATAAG TCAAAGAACC CCGCAAAACA CACATATGGA GTACACGATG
27101  AAGCCGCATA AAATAAACGT GTCGCACGCT TCGAAATTGG GCAATGCACG
27151  AACCGGAACT CGACTTTAAA GAGCGGTGTT TGTGGTAAGG GGCCTCGAAA
27201  CCTGGATGGT GGTCAGATCA CTTGATGACA CTTTAAGCCG ACAACGGAAC
27251  GTGCCGTAGA ACACGCAGAG TTTCGAGTTC TTGGCCAAGC GAATAAATTA
27301  CGTTGGAGGC TATTTTGCAC TTTCGTAAGC CGAAAACCGA AAACCGTTGA
27351  GTGTTTATTT CACGGAAAGC CAATTTTGAA CGAAGTGTTT ACCGCAACAG
27401  CCGCTCGACG ACTGACGTTG AGTTTGGAAA ACAAGCTAAC GAGACTTTGC
27451  AGCTCAACTG GAATCGAGTT ATCGCCGGAA TGGCATGGCA ACCCTGTTGC
                  WG231
27501  GGGAGAGACA GATATAGCCG GGTCGTAACT AAAGAGAGAG AAACGCACCA
27551  AAGGCACGCG AGTAATTAGC CAACCGTAAA TTCGACCGGC AAGTGGGCGG
27601  TGGGTGAGCT CTGTTTGTTT ACAATTTGCG AGCTCGTTTA CTGGATCGCA
27651  AGTGGGTGCT CACTGGAAGG CTGGAATGGT GGTCAGCGGA AACCGGAAGT
27701  GCGTGCGCTT ATGCAATTTT GAAGGTTGAG CTTTCCATCG CAAAAACCAA
                                                  WG234
27751  TTTTAGTTTG CTTAGCATCA GTTGATAATC TTTGAAACCT TGAACTTTTT
27801  TGGATCGGTT TTCTCTTAAC CTAATAATTT CAGTACTGTC ATTCTAATTT
27851  GTCATTAATT TAATCCAGTG ATTTGCCATA TCATATTGAC AGATACGTAA
27901  CGTCATTATT TTTCCTAGAA TAATATGTCC GTGAGTCGAG TGACTATGAT
27951  GCGAAAGGGC CACTCCGGGG AGGTAGCACG CAAGCCCAAC ACTGTGGTGG
28001  TGTCGGTTCC ACCGCTGGTG AAGAAGTCCA GCAAGAGCCG CTCGTTCCAC
28051  TTCCGCTATC TGGAGCTGTG CCGGGCCAAG AATCTGACGC CGGTGCCGGA
28101  AATCCGCAGC AAGTCGAATG CGACCACCAC CTTTCTGGAG CTGTGCGGCG
```

```
28151  ATAAGCTGGC GGTCAGCGAT TGGCAGCTCC TAACCGAGGC GCTCCACTAT

28201  GATCTCGTGC TCCAGCATCT GGTGGTGCGC CTGCGACGCA CATATCCACA

28251  AAGTAGGTGA TCTTTGGTAG TCTGCTACTC TGTATGGAAT GTGAATTATA
                                            (Pseudo BamhI)
28301  CCATTTCGTT TATTCATTTT TTTCGGCAGC CAACATTGATCCCATTGACA

28351  CCGAGAAACG AGCCCGACTT TTTCGCCAGC GGCCAGTGAT CTATACTCGC

28401  TTCATATTCA ACAGTTTGGT CCAGGCGATT GCCAACTGTG TTTCGAGCAA

28451  CAAAAATCTA AGTGTGTTGA AGCTGGAGGG ATTGCCATTG CAGGATGGAT
                                                  (Pseudo BamhI)
28501  ATATCGAGAC CATTGCCAAG GTGCGTTGAA AAGTTTTGCG GTCGGGATCCA

28551  CGTTTCCAAG TACACACATA TATCCAATCC ATTCACAGGC ACTGGCAGAC

28601  AACGAATGCC TCGAAACAGT GAGTTTTCGC AAATCCAACA TTGGCGATAA

28651  GGGCTGCGAG GTGGTGTGCA ACACAGCCAA ATACCTGAAT CGCATCGAAG

28701  TGTTCGACCT CTCTGAATGC GGACTTACTT CCAAGGGAGC CGAGCACGTG

28751  GCCGACATGC TCAAGGTAAT CTCTTAATCC CTGTAGATCT TTTTCTTTCT

28801  TTCTAGAAAT ATATCCTGGC GTTTTCAGTA GTATATGATG AAATTCCATT

28851  GCATATTTTT ATTACCCTTC TTAATGGTCT TCCTATAACT GCGTTTATGT

28901  TTTTTATGTA ATCAGATGCA AAAGATCACC CGTTTCACGG AGGGATGGGA

28951  GAAGTCCCTC CGCTACCGTA GCGTCGATGT GAACACGATT GGCGGTCTGC

29001  GCACGGTTTT GTTGGCTGAC AACCCGGAGA TTGGCGACGT GGGCATCCGG

29051  TGGATAACCG AGGTGTTGAA AGAGGATGCT TGGATAAAAA GTACGTAGAG

29101  TCCGAATAGT GCGATCCATA CAGTTCAAAT ACCCCACAGA AATCGACATG

29151  GAGGGCTGCG GCCTGACGGA TATCGGGGCA AATCTAATTC TCGATTGCCT

29201  GGAGCTGAAC ACGGCCATTA CGGAGTTCAA TGTGCGAAAC AACGAAGGAA

29251  TCAGTAAGTT CCTGCAGCGA AGTATCCATG ATCGTCTTGG CTGTTTACCA

29301  GAGGAGAAAC AGGAGCCAGA GTATGATCTC AGTTGCGTCA ACGGGCTACA

29351  GAGCCTGCCC AAGAACAAGA AGGTCACCGT CTCTCAACTG CTGTCCCACA

29401  CCAAAGCATT GGAGGAGCAG CTCTCCTTCG AGCGAACGTT GCGCAAGAAG

29451  GCCGAGAAGC TGAATGAGAA GCTTAGCCAC CAGCTCATGA GACCCGACTC

29501  CAATCACATG GTTCAAGAGA AGGCCATGGA GGGAGGATCA CAAACAAACA

29551  TTTCGAGGGA ATATGTGGCG CGGAATGATG TTATGCCAGA AGTCATCAAA

29601  AAGTGGGCTA GCTCTCAATT CTCAATTCAA TGCCTATCTG ATTTGAGTAT

29651  TTCTTCACAG TTCCCAAAGC TACCGCCAGT CGCACTTCAA CCGGCTGGTC

29701  AACAGTGCGG CCACCAGTCC CGAAGTCACA CCCCGCAGCG AGATTGTCAC
```

```
29751  ATTGCGCAAG GAGCAGCAGC TGCAACGTCA ACAACCCCCA CCAATGGAGG

29801  TCAAGCATCT TTCCTTGGAG CAGCAAATCC GAAATCTGCG CGACGTGCAG

29851  AAAAAGGTGG ACTTGGACGT GGAGGAAGAG GAGGAGGAGG AGGAACAGCA

29901  GGCGGAGGAA AGTCAATCCG AGTCGGAGCT GCAGAACGAG GAGCAACAGC

29951  ATTACGAACA GCAAATGCAG GTCCAACGCA AACATCTCCA GGTGCGCAAG

30001  GTTCGCAGTG AGATTAAGTA TGTGGAAAAC AATCCCAAGG AGGCAGCCAA

30051  AAAGAATCGC GAGTCCAAGT CGGACCATGA GTTTGCCAAC GAGAGAGATG

30101  TGAGTAGTAT CCACAAGATA AATTAACCAG TACTAGGAAT TTATTCTCTT

30151  GTACGATAAC CCATCTTATC CCCATAGTTC AAGCTTAATC CCTCTGTGCA

30201  ATTCGAGACG GACATTGGCG ACAATTTGAT GGTCAATCCT GGCCACCGAT
           B8B2
30251  ACGAGGGCGG CGGGGGCGAT ACGGGCTATG TCTACAACTA CGAGCATGAG

30301  CAGCAGCAGC AGCCAGTCAA GCGGGGCTAC GAGCACGGCT ATGTGGTGGG

30351  CGTGGGTGAC GGATCCCACA GGAGGCAGAG GCAATCTCAA CTGGTCGAGG

30401  CTTTGGTGCA AAAACGTGTC CCAGGCGCCA GCGATGGACA TGTGGCGCAG

30451  TTCGTTAGCA ATCTGGAACG ACAAGCGAAT GCTGGTAAAA CGGGGAAAAA

30501  GCGCCTTAAA CCTCGGCCTG AGGACGATCT TCAGGTACCA GTCGGCGACA
                                                    B21
30551  TGCACATGGA GTCGTCGTAT ATGTCCCGCT CCGAAGAACT CTCCTCAACG
           CGCCGGCG
30601  GACGTTACGC TGGAGAACTC AGACTACGAG ACGGAGGCGA CGGACTCCAC

30651  GTTACTTAGT AGCTCGAAAT ACTCCTCCAT GCATGTCTTT GTGCGGCGCA

30701  AGCAATCGGA GTCCATGTCA CTCACAGAAG AGGCCGGCGA CGGAGATGCC

30751  GGCGGTGGTG GAGGCTCTGG CGATTTCGGC GACCAAAATG TCATATCCCC
                 B22          GCGGCCGC
30801  GGCCAATGTC TACATGTCCC TGCAGCTCCA GAAGCAGCGG GAGCAGAGCG

30851  CCTAGGGGTC GTGACTTTAG GTTCCTCCAT TCAAAATCCC CGCAGCCATT

30901  CGCAGCCGGA GCAGCGACAC CACAGCTGGC TGATACGCCG AGCAGTAACA

30951  ACAACAACAC CACTACCACC ACCATCACAC CCACAACAAA GCAGCCAACT

31001  CGATTCGATA GGATCCCCAC TGGGAGAGCA AAATTTGCAG TCATACTCCC

31051  GGGCAGTAGC AGTAGAAAAG GGAACCTGCT CCTTTTTGCA CGTCCTGCCA
                                                    B72
31101  TCTGCGAATT CAAGCTGCAC CAAAAATGTA CAAAAATATA CTTTGGTCTT
         CGCCGGCG
31151  ACTAATTTCC ATTAAAGATT TATTATTTGT GTGCTCCCTA AGCGAGTTTG

31201  TTGGACCTTA TTTTCGGATG ATTAACCGCC CCAAGTGAAA ATGCACATAT

31251  GCATGATATT GAGTTCTCTA TGTGATATCT TTTGATTTAT TTAGTCCGTA
```

31301 TATTGTGAAA TTATAAAATA TTTGCGTTCG GTTAAGTTTG CGTTGAATAC

31351 TACTGATTAA GATTCAATTG GATCAAGTAT TTTGAATGTA GTTAAGTTAA

31401 CTCCATGGTT ATTTATCAAG TGGAGGTTAG AGGGCGTGCT TAACCGCCAA

31451 ATCCGTAAAG GGTGCGTCCC TGGCGCTTGA GGGCATAGAC CACGTCCATG

31501 GCGGTCACGG TCTTGCGCTT GGCGTGCTCG GTGTAGGTGA CAGCGTCACG

31551 GATAACGTTC TCAAGGAATA CCTGTGAATT TTCCGAGAAA TTAATCAGTT

31601 TTTCACCACT TTCCCTTCGT TTCGCCAGAA CGTACCTTTA GCACACCGCG

31651 AGTTTCCTCG TAAATCAAGC CAGAGATACG CTTAACACCG CCGCGACGAG

31701 CCAAACGGCG AATAGCAGGC TTGGTGATAC CCTGGATGTT ATCACGAAGC

31751 ACCTTACGAT GACGCTTGGC GCCCCCCTTT CCCAATCCTT TGCCACCCTT

31801 TCCACGACCA GTCATTTCTC AGTTGCTTCG TAAAGTTGGC TGAAAAGAAG

31851 AGAAGACACA ATAAACCGTA AAGCGACGCC ATGTTTGGTG AAAGGTAAAC

31901 GTTCATTTGA CAGGTGAAAA CTGTTGTTAA TAAGTCTGTT TTCATACAAA

31951 ATAATGGTCT CCAATTATTT TTTTTATGTT TATGCACAAA AAAACTTTGT

32001 TGACCGATTC TTCTATTATT TACTTAAATT TCTCAGCAAC AAAACTCACT

32051 TCAAGAAATC TCCAGAAAAT TGCTTGTGAG CGTGACCAGA TCTCGCACAC

32101 CTCAAAATAC CTTTCGCAAA ATGCGCGTAT CTAGTATTTC TATTGTCTCA

32151 CTACCAGCCC TGGTCAGTTA TCGCCTATCG GCCATCGCTA TCGCCGCCCC

32201 GTTGCGACAA AAAAGATTTA ATCGAACCTA AAGACGTGCT CGCACTAAAA

32251 TCGCTTGAAA ACAGGCCCGC AGACCTGCGA AAAACGAAAA AGTGCAGCGC

32301 GCATATACTT TTTCAACTGT GCCCCTCTAG CTTAAAATTA AGTCGCGGCG

32351 AAAAGTCGAG TAAAAACCGC GGAAATGCGC ATGCAAACGG TGTGTGGCCA

32401 GCAAAATCGC TGCCAAGGCA CCGCACACAC ACTCGGCCAC CCACACATAC

32451 ACACTTAGTG CTGTACTCGA AAAGTGCGAA GACAACAGGA ACTCTGTGCC

32501 AAAATAATAT TAACAGTACC CAGTTATTCC ATTCCACTGC ACCTGTCCCC

32551 GAAACATCGA AATATTCGCG TTACGTATAC GCAACGAGTG CTGTAAACAA

32601 GTTTGCACAG GCGGTAACAA TGTCCAAATA CGATCCGCTT TATCTAACGG

32651 CGCAGCTAAC GCTGGGTATG TATTGCACTG CGAGTGTGTG GGTGCATACT

32701 TATTGGCGCC TTAATTCCCA GTCGCTATAT TATGTACATA TGTACATAAT

32751 ATTATATTTT GTCTCTGCCT TCCGATTTCC CCCTGTAATG CAAGCCGACT

32801 TCGATTTATA ATCATTCGCG CGTATGATTT TTGATGATAT GCGATGCGTA

32851 ATGCTATTAC TTTATTAATA TCACATGGCG CTTGAAAAGT ACATATATAG

32901 GGTGTATAGG CTTTGTGAAA GCTTTGTCCT ATTAACCAAC ACATCAGAAC

32951 ATCTTAAATT TTCAGAAATT AGGTATAACG AATGTAAATG CTAGGAATCA

33001 AATCTATTTG TGAAGTGAAT AATGTATACA CTAGATACTC TATTAAAGCA

33051 CCAACAATTG TACTACACAA ATGTAGTAAA ATATCGTAAG AATTGTTGAC

33101 AGTGGGTACT ACGAATTCAT ATGTTGAGCG CCAAACGCAT ACTCGGTATA

33151 AATCGATCGG ATACCTTAAT TCCCAATCTT AATCCATGTG TGTGTGCGCT

33201 GAGCATTTCT TTGAATGGAT TTTGCTGCAT TAGGAACTTG TTGTATGGCA

33251 GGTTCTGGTT TCTCCCACCT TTTGCTCCAC CGGCCTCCCT CTCTTTCGCT

33301 CGCCATAGCC CTCTCTCTTT CTGTGGCGCA GTGCGCTCTC TCCCCTGTTG

33351 CTGCTGCTGT TGTTGTTGTT TCTGCTGCTG CTGTTTCTTT TTTCCGCACA

33401 AGCGCCACGT TAAACCCATC CCTTTCTGTG TGCATAATGT GCCTGCTTAT

33451 GTATCTACTT TACTATAAAG CATGCACGCT TACGAACACA CACATACATG

33501 GCCAGCGATC CGATTACCTA AAGGACAGTG GGACGAAAAG TGTTTCAAAG

33551 GATTGAAAAC AGTTTGCTTA GCCAATGGTC AAGAAATATC AATCAATATA

33601 TCGCAGGCTC GTTGCAATAA ACGTTTTCAA AATTAAATTT TTTGTATTTA

GCGGC → B71

33651 CACATTTAAA AAGTGAATTA TCATTTTAAT TTTTATCAAC AAAAATTTCG

33701 GTATTTGGAT GA**GGATCC**CA GATTAAATCG AATTAATGAA TGTAAATAGA

33751 TTTTCAAGAA ACCTTACTGC AGGCCCACTG TGCACTGCAT ATCCGTGCAT

33801 TGGGGCCATG CCCACTCAAA GGCTGGTGTG TGGCCAGTGG GGCTAATTCC

33851 GTGCCCAAAT GGGGTGGCTT TCAATGGCAG AGGCCCCAGC TTAGGACGCT

33901 ACTCCTGCTA CTTCTGGGCA TGCGATATGT GTACAAAGGA TAGCGCCCAC

33951 AAAGAGCTCG CTGAGCGCCC TCCCTTTCAG TCTTATTCCC CAAATAGGCT

34001 CGACTTTATT TGCCCACCCT TTGAGCACTT CCAACCGATA ATTCCATTAA

34051 CTTTGA*TCCT AAGCTTCCGG TCTCCCTATA GTGAGTCGTA TTAATTTCGA*

vector Lawrist4

34101 TAAGCCACCT CGAGGCGAAT TAGCCCGCCT AATGAGCGGG CTTTTTTTGG

34151 CCGTTTCGGC CGAATTCTCT AGAGATCTTC CATACCTACC AGTTCTCCGC

34201 CTGCAGGGGG GGGGGGGGGG GGGGGGGGGG GGGGACATGA GGTTGCCCCG

34251 TATTCAGTGT CGCTGATTTG TATTGTCTGA AGTTGTTTTT ACGTTAAGTT

34301 GATGCAGATC AATTAATACG ATACCTGCGT CATAATTGAT TATTTGACGT

8V2

34351 GGTTTGATGG CCTCCACGCA CGTTGTGATA TGTAGATGAT AATCATTATC

34401 ACTTTACGGG TCCTTTCCGG TGATCCGACA GGTTACGGGG CGGCGACCTC

34451 GCGGGTTTTC GCTATTTATG AAAATTTTCC GGTTTAAGGC GTTTCCGTTC

34501 TTCTTCGTCA TAACTTAATG TTTTTATTTA AAATACCCTC TGAAAAGAAA

34551 GGAAACGACA GGTGCTGAAA GCGAGGCTTT TTGGCCTCTG TCGTTTCCTT

34601 TCTCTGTTTT TGTCCGTGGA ATGAACAATG GAAGTCCCCC CCCCCCCCCC

34651 CCCCCCCCCC CCCTGCAGCA ATGGCAACAA CGTTGCCCGG ATCGATCGGT

34701 CGCGCGAATT GATCCGACCA AAGCGGCCAT CGTGCCTCCC CACTCCTGCA

34751 GTTCGGGGGC ATGGATGCGC GGATAGCCGC TGCTGGTTTC CTGGATGCCG

34801 ACGGATTTGC ACTGCCGGTA GAACTCCGCG AGGTCGTCCA GCCTCAGGCA

34851 GCAGCTGAAC CAACTCGCGA GGGGATCGAG CCCGGGGTGG GCGAAGAACT

34901 CCAGCATGAG ATCCCCGCGC TGGAGGATCA TCCAGCCGGC GTCCCGGAAA

34951 ACGATTCCGA AGCCCAACCT TTCATAGAAG GCGGCGGTGG AATCGAAATC

35001 TCGTGATGGC AGGTTGGGCG TCGCTTGGTC GGTCATTTCG AACCCCAGAG

35051 TCCCGCTCAG AAGAACTCGT CAAGAAGGCG ATAGAAGGCG ATGCGCTGCG

35101 AATCGGGAGC GGCGATACCG TAAAGCACGA GGAAGCGGTC AGCCCATTCG

35151 CCGCCAAGCT CTTCAGCAAT ATCACGGGTA GCCAACGCTA TGTCCTGATA

35201 GCGGTCCGCC ACACCCAGCC GGCCACAGTC GATGAATCCA GAAAAGCGGC

35251 CATTTTCCAC CATGATATTC GGCAAGCAGG CATCGCCATG GGTCACGACG

35301 AGATCCTCGC CGTCGGGCAT GCGCGCCTTG AGCCTGGCGA ACAGTTCGGC

35351 TGGCGCGAGC CCCTGATGCT CTTCGTCCAG ATCATCCTGA TCGACAAGAC

35401 CGGCTTCCAT CCGAGTACGT GCTCGCTCGA TGCGATGTTT CGCTTGGTGG

35451 TCGAATGGGC AGGTAGCCGG ATCAAGCGTA TGCAGCCGCC GCATTGCATC

35501 AGCCATGATG GATACTTTCT CGGCAGGAGC AAGGTGAGAT GACAGGAGAT

35551 CCTGCCCCGG CACTTCGCCC AATAGCAGCC AGTCCCTTCC CGCTTCAGTG

35601 ACAACGTCGA GCACAGCTGC GCAAGGAACG CCCGTCGTGG CCAGCCACGA

35651 TAGCCGCGCT GCCTCGTCCT GCAGTTCATT CAGGGCACCG GACAGGTCGG

35701 TCTTGACAAA AAGAACCGGG CGCCCCTGCG CTGACAGCCG GAACACGGCG

35751 GCATCAGAGC AGCCGATTGT CTGTTGTGCC CAGTCATAGC CGAATAGCCT

35801 CTCCACCCAA GCGGCCGGAG AACCTGCGTG CAATCCATCT TGTTCAATCA

35851 TGCGAAACGA TCCTCATCCT GTCTCTTGAT CAGATCTTGA TCCCCTGCGC

35901 CATCAGATCC TTGGCGGCAA GAAAGCCATC CAGTTTACTT TGCAGGGCTT

35951 CCCAACCTTA CCAGAGGGCG CCCCAGCTGG CAATTCCGGT TCGCTTGCTG

36001 TCCATAAAAC CGCCCAGTCT AGCTATCGCC ATGTAAGCCC ACTGCAAGCT

36051 ACCTGCTTTC TCTTTGCGCT TGCGTTTTCC CTTGTCCAGA TAGCCCAGTA

36101 GCTGACATTC ATCCGGGGTC AGCACCGTTT CTGCGGACTG GCTTTCTACG

36151 TGTTCCGCTT CCTTTAGCAG CCCTTGCGCC CTGAGTGCTT GCGGCAGCGT

36201 GAAGCTAGCT TGGCTTGGAG CCTGTTGGTG CGGTCATGGA ATTACCTTCA

36251 ACCTCAAGCC AGAATGCAGA ATCACTGGCT TTTTTGGTTG TGCTTACCCA

36301 TCTCTCCGCA TCACCTTTGG TAAAGGTTCT AAGCTCAGGT GAGAACATCC

36351 CTGCCTGAAC ATGAGAAAAA ACAGGGTACT CATACTCACT TCTAAGTGAC

36401 GGCTGCATAC TAACCGCTTC ATACATCTCG TAGATTTCTC TGGCGATTGA

36451 AGGGCTAAAT TCTTCAACGC TAACTTTGAG AATTTTTGTA AGCAATGCGG

36501 CGTTATAAGC ATTTAATGCA TTGATGCCAT TAAATAAAGC ACCAACGCCT

36551 GACTGCCCCA TCCCCATCTT GTCTGCGACA GATTCCTGGG ATAAGCCAAG

36601 TTCATTTTTC TTTTTTTCAT AAATTGCTTT AAGGCGACGT GCGTCCTCAA

36651 GCTGCTCTTG TGTTAATGGT TTCTTTTTTG TGCTCATACG TTAAATCTAT

36701 CACCGCAAGG GATAAATATC TAACACCGTG CGTGTTGACT ATTTTACCTC

36751 TGGCGGTGAT AATGGTTGCA TGTACTAAGG AGGTTGTATG GAACAACGCA

36801 TAACCCTGAA AGATTATGCA ATGCGCTTTG GGCAAACCAA GACAGCTAAA

36851 GATCTCGGCG TATATCAAAG CGCGATCAAC AAGGCCATTC ATGCAGGCCG

36901 AAAGATTTTT TTAACTATAA ACGCTGATGG AAGCGTTTAT GCGGAAGAGG

36951 TAAAGCCCTT CCCGAGTAAC AAAAAAACAA CAGCATAAAT AACCCCGCTC

37001 TTACACATTC CAGCCCTGAA AAAGGGCATC AAATTAAACC ACACCTATGG

37051 TGTATGCATT TATTTGCATA CATTCAATCA ATTGTTATCT AAGGAAATAC

37101 TTACATATGG TTCGTGCAAA CAAACGCAAC GAGGCTCTAC GAATCGAGAG

37151 TGCGTTGCTT AACAAAATCG CAATGCTTGG AACTGAGAAG ACAGCGGAAG

37201 CTGTGGGCGT TGATAAGTCG CAGATCAGCA GGTGGAAGAG GGACTGGATT

37251 CCAAAGTTCT CAATGCTGCT TGCTGTTCTT GAATGGGGGG TCGTTGACGA

37301 CGAGATGGCT CGATTGGCGC GACAAGTTGC TGCGATTCTC ACCAATAAAA

37351 AACGCCCGGC GGCAACCGAG CGTTCTGAAC AAATCCAGAT GGAGTTCTGA

37401 GGTCATTACT GGATCTATCA ACAGGAGTCA TTATGACAAA TACAGCAAAA

37451 ATACTCAACT TCGGCAGAGG TAACTTTGCC GGACAGGAGC GTAATGTGGC

37501 AGATCTCGAT GATGGTTACG CCAGACTATC AAATATGCTG CTTGAGGCTT

37551 ATTCGGGCGC AGATCTGACC AAGCGACAGT TTAAAGTGCT GCTTGCCATT

37601 CTGCGTAAAA CCTATGGGTG GAATAAACCA ATGGACAGAA TCACCGATTC

37651 TCAACTTAGC GAGATTACAA AGTTACCTGT CAAACGGTGC AATGAAGCCA

37701 AGTTAGAACT CGTCAGAATG AATATTATCA AGCAGCAAGG CGGCATGTTT

37751 GGACCAAATA AAAACATCTC AGAATGGTGC ATCCCTCAAA ACGAGGGAAA

37801 ATCCCCTAAA ACGAGGGATA AAACATCCCT CAAATTGGGG GATTGCTATC

37851 CCTCAAAACA GGGGGACACA AAAGACACTA TTACAAAAGA AAAAAGAAAA

37901 GATTATTCGT CAGAGAATTC TGGCGAATCC TCTGACCAGC CAGAAAACGA

37951 CCTTTCTGTG GTGAAACCGG ATGCTGCAAT TCAGAGCGGC AGCAAGTGGG

38001 GGACAGCAGA AGACCTGACC GCCGCAGAGT GGATGTTTGA CATGGTGAAG

38051 ACTATCGCAC CATCAGCCAG AAAACCGAAT TTTGCTGGGT GGGCTAACGA

38101 TATCCGCCTG ATGCGTGAAC GTGACGGACG TAACCACCGC GACATGTGTG

38151 TGCTGTTCCG CTGGGCATGC CAGGACAACT TCTGGTCCGG TAACGTGCTG

38201 AGCCCGGCCA AACTCCGCGA TAAGTGGACC CAACTCGAAA TCAACCGTAA

38251 CAAGCAACAG GCAGGCGTGA CAGCCAGCAA ACCAAAAACTC GACCTGACAA

38301 ACACAGACTG GATTTACGGG GTGGATCTAT GAAAAACATC GCCGCACAGA

38351 TGGTTAACTT TGACCGTGAG CAGATGCGTC GGATCGCCAA CAACATGCCG

38401 GAACAGTACG ACGAAAAGCC GCAGGTACAG CAGGTAGCGC AGATCATCAA

38451 CGGTGTGTTC AGCCAGTTAC TGGCAACTTT CCCGGCGAGC CTGGCTAACC

38501 GTGACCAGAA CGAAGTGAAC GAAATCCGTC GCCAGTGGGT TCTGGCTTTT

38551 CGGGAAAACG GGATCACCAC GATGGAACAG GTTAACGCAG GAATGCGCGT

38601 AGCCCGTCGG CAGAATCGAC CATTTCTGCC ATCACCCGGG CAGTTTGTTG

38651 CATGGTGCCG GGAAGAAGCA TCCGTTACCG CCGGACTGCC AAACGTCAGC

38701 GAGCTGGTTG ATATGGTTTA CGAGTATTGC CGGAAGCGAG GCCTGTATCC

38751 GGATGCGGAG TCTTATCCGT GGAAATCAAA CGCGCACTAC TGGCTGGTTA

38801 CCAACCTGTA TCAGAACATG CGGGCCAATG CGCTTACTGA TGCGGAATTA

38851 CGCCGTAAGG CCGCAGATGA GCTTGTCCAT ATGACTGCGA GAATTAACCG

38901 TGGTGAGGCG ATCCCTGAAC CAGTAAAACA ACTTCCTGTC ATGGGCGGTA

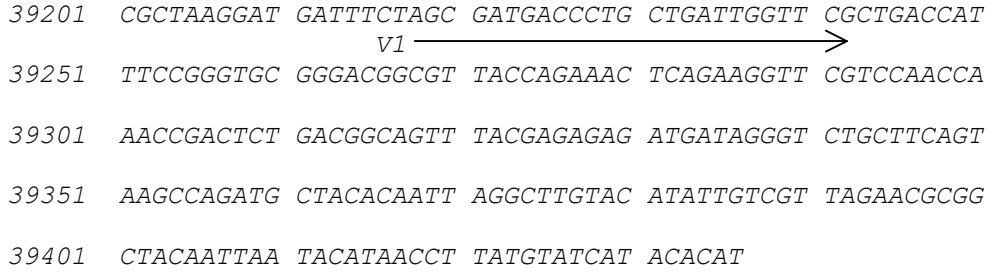
38951 GACCTCTAAA TCGTGCACAG GCTCTGGCGA AGATCGCAGA AATCAAAGCT

39001 AAGTTCGGAC TGAAAGGAGC AAGTGTATGA CGGGCAAAGA GGCAATTATT

39051 CATTACCTGG GGACGCATAA TAGCTTCTGT GCGCCGGACG TTGCCGCGCT

39101 AACAGGCGCA AACGTAACCA GCATAAATCA GGCCGCAGCT CGCCCGGGGA

39151 TCTGGCTAGA ATTCGGCCGG GGCGGCCAGA TAAAAAAAAT CCTTAGCTTT


39201 CGCTAAGGAT GATTTCTAGC GATGACCCTG CTGATTGGTT CGCTGACCAT
V1
39251 TTCCGGGTGC GGGACGGCGT TACCAGAAAC TCAGAAGGTT CGTCCAACCA
39301 AACCGACTCT GACGGCAGTT TACGAGAGAG ATGATAGGGT CTGCTTCAGT
39351 AAGCCAGATG CTACACAATT AGGCTTGTAC ATATTGTCGT TAGAACGCGG
39401 CTACAATTAA TACATAACCT TATGTATCAT ACACAT

ZUSAMMENFASSUNG der Dissertation von Wenli GU

Fachbereich Biologie der Johannes Gutenberg-Universität Mainz

Thema: The study of a novel zinc finger gene cluster *TZF* and a genomic region flanking the histone replacement gene H4r of drosophila melanogaster

Part I : A zinc finger gene *Tzf1* was cloned in the earlier work of the lab by screening a λ-DASH2 cDNA expression library with an anti-Rat SC antibody. A λ-DASH2 genomic DNA library and cosmid lawrist 4 genomic DNA library were screened with the cDNA fragment of *Tzf1* to determine the genomic organization of *Tzf1*. Another putative zinc finger gene *Tzf2* was found about 700 bp upstream of *Tzf1*.

RACE experiment was carried out for both genes to establish the whole length cDNA. The cDNA sequences of *Tzf* and *Tzf2* were used to search the Flybase (Version Nov, 2000). They correspond to two genes found in the Flybase, CG4413 and CG4936. The CG4413 transcript seems to be a splicing variant of *Tzf* transcripts. Another two zinc finger genes *Tzf3* and *Tzf4* were discovered *in silico*. They are located 300 bp away from *Tzf* and *Tzf2*, and a non-tandem cluster was formed by the four genes. All four genes encode proteins with a very similar modular structure, since they all have five C2H2 type zinc fingers at their c-terminal ends. This is the most compact zinc finger protein gene cluster found in *Drosophila melanogaster*.

Part II: 34,056 bp insert of the cosmid 19G11 , containing the Histone H4 replacement gene H4r was completely sequenced by mapping and subcloning. Computer analysis was carried out with this sequence and 9 new putative genes except for the known genes H4r and *punt* were identified.

Part III: P-element mediated excision was carried out trying to obtain *Drosophila melanogaster* strains with mutated H4r gene.About 5,000 F2 candidate flies were sorted according the phenotypes and screened with PCR. No mutant could be identified.

Genemigt vom 1. Gutachter/von der 1. Gutachterin _____