## Phage and other MGEs ontologies

Ariane Toussaint, Gipsi Lima-Mendez, Raphaël Leplae SCMBB, Université Libre de Bruxelles

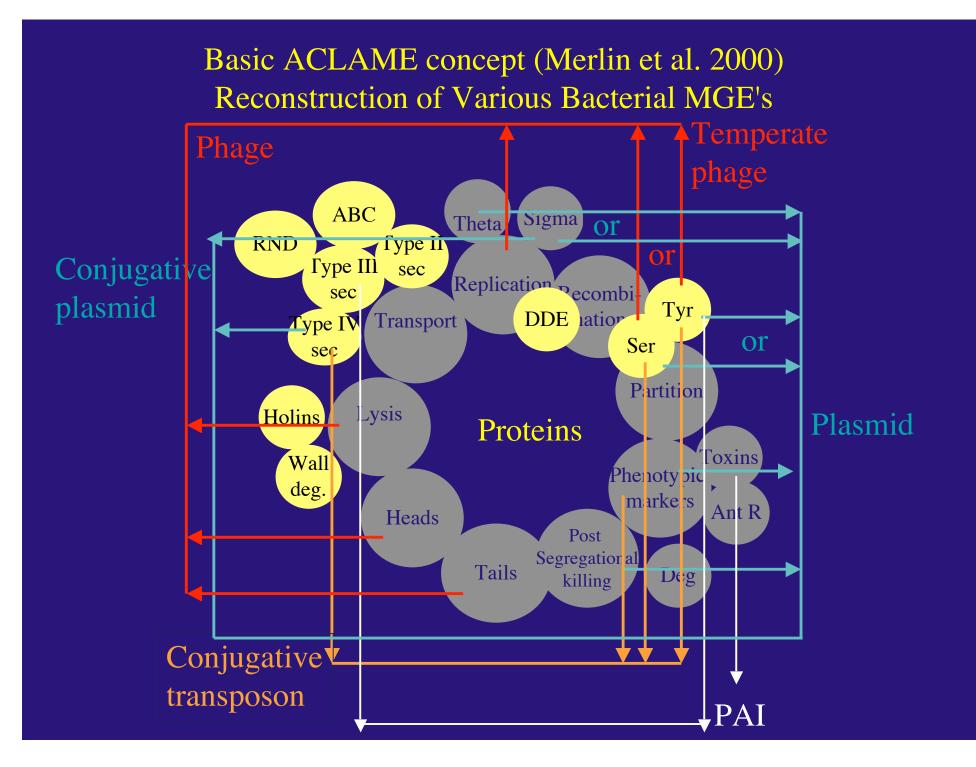


ULB **SCMBB** 

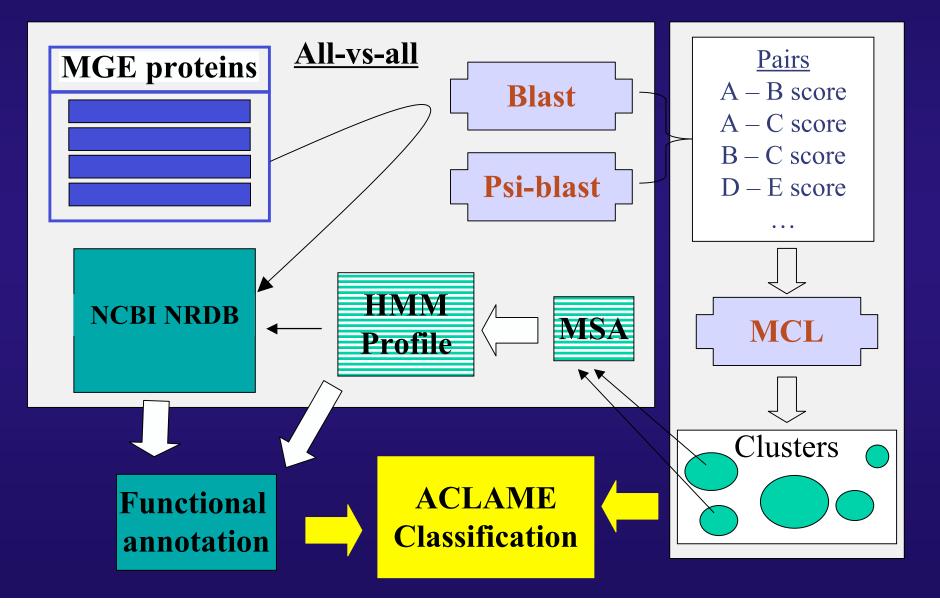
Cesa &CK · CEN







# Generating protein families Proteins clustering



	State of the second sec	ACLAME: Clusters vie	w – Mozilla {Build ID: 2004	4021810}		
Back - Dorward - Reloa	ad Stop	lb.ac.be/perl/Aclame/show_clu	ıster.cgi?mode=list&class=fam	iily&cat=vir	•	💉 Search 📑 👻 🎹
🖞 🚮 Home 🛛 😻 Bookmarks 🥠	mozilla.org 🥠 Latest Builds 🥠	Dictionnaire Françai 🥠 The 🤅	Golden Pages O 🥠 Ministère	e de la rech 🥠 P	lasmid Biology 2004	
ACLAME			ACLAME: Clusters vi lame/show_cluster.cgi?mode=clust	t&id=cluster:vir:2&clas	ss=family&vers 📀 ^ 📿 Google	
	🛱 Rendezvous 🔻 Databases 🔻	Annotation    Biblio    Societes	Personal Toolbar Folder▼ ULB▼	Unclassified <b>v</b>		
Home Clust					Login: P	assword: Go
Classification	_	Classification		Browse	Description	Proposal
Browse	0 Home Cluster					
Description	Classification	sview				
Forums Downle	r Browse Categories	(?): Viruses 🗘 Filter Jun	np (?) : Genomes 🗘 Go	Version 0.2 Chang	ge to (?) : 0.2 🛟 Co	
Ontology	Proposal Clustered p	roteins in families filtered bas	ed on Viruses:			
Resources	Forums Previous	Next Cluster ID Go				
People clus	S Ontology					
Community	Resources Cluster ID	cluster:vir:2	Nb of proteins 83			
Annotation servers clus	S People Function	function:147: DNA recombination	on, site-specific - Tyr recombin	GOUUL	$9037_{3}$	
Ontologies	American comune	HTML View MSA locally View NR	DB matches View SwissProt matche			
Genomes clu	S Ontologies					
Financial supports clus	S Genomes ID V	[sw] ⊽ [nrdb] ⊽ [sco	p] ∇ Description ∇	Length 🗸		Genome and host(s) ⊽
Publications clu		0032 [+] [81] [100] [4]	hypothetical protein	-	Stx1 converting bacteriophage	
Tools	Publications vir:protein:	0198 [+] [81] [100] [4]	hypothetical protein	444	Stx2 converting bacteriophage	II Host(s): Escherichia coli
Annotators clu	S Tools vir:protein:	042 [+] [83] [100] [5]	integrase	377	Bacteriophage bIL309 Host(s)	Lactococcus lactis
Blast	8	098 [+] [81] [100] [4]	integrase	394	Bacteriophage bIL310 Host(s)	Lactococcus lactis
clu	Blast vir:protein:	127 [+] [82] [100] [4]	integrase		Bacteriophage blL311 Host(s)	
clu	vii.protein.	195         [+]         [81]         [100]         [4]           106         [+]         [82]         [100]         [4]	integrase		Bacteriophage phi CTX Host(s	-
		196       [+]       [83]       [100]       [4]         29       [+]       [83]       [100]       [3]	similar to phage phi 11 inte integrase	-	Bacteriophage HK620 Host(s)	): Staphylococcus aureus; Meth
clu		399 [+] [82] [100] [5]	putative integrase		Bacteriophage phi3626 Host(s)	
clus	vir:protein:	423 [+] [82] [100] [4]	putative recombinase		Bacteriophage phi3626 Host(s	
clus	st vir:protein:	461 [+] [82] [100] [5]	putative site-specific integr		Bacteriophage phiE125 Host(s	
clus	1	565 [+] [82] [100] [5]	integrase	391	Bacteriophage phig1e Host(s)	Lactobacillus
clus	st vir:protein:	608 [+] [83] [100] [5]	Integrase	374	Bacteriophage r1t Host(s): Lac	ctococcus lactis
ahaa		854 [+] [83] [100] [5]	Int		Enterobacteria phage 186 Hos	
clus	Thip: etc.ini	909 [+] [82] [100] [5]	integrase		Enterobacteria phage HK022	
clus		043 [+] [82] [100] [3]	gpInt, integrase		Enterobacteria phage P2 Host Enterobacteria phage P22 Host	
clus	st vir:protein:2	1059     [+]     [82]     [100]     [5]       10     [+]     [82]     [100]     [6]	integrase integrase		Bacteriophage HK97 Host(s):	17 71
clus	vii protoini.		Int		Bacteriophage K139 Host(s):	7
		f-capsid 23	31 1	404	Otenhuleneenee	
	ster-vir-10 DNA ren		26 1			
🐝 🕮 🎸 🔝 🗹 Do						

#### Results for query sequence gi|56692569|ref|YP\_164037.1| Blast search results in Viruses

Hit vir:protein:1974 [+]
Description Original annotation: Mor; ACLAME function(s): transcriptional activator Clusters:
cluster:vir:997
Bits| 34

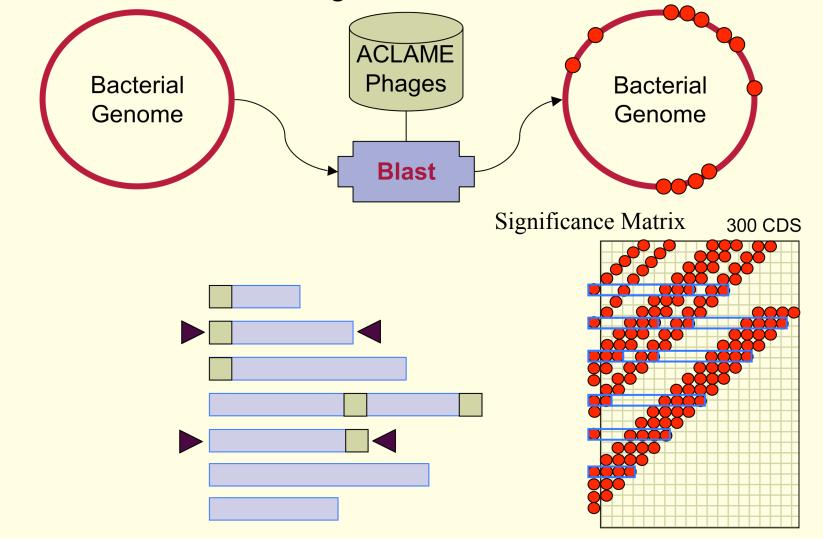
**Eval** | 3.e-03 **%ID** | 26| | 24| 88| 4| 42| 45| 4| 4

vir:protein:1974 [+] (Length: 129) - Original annotation: Mor ACLAME function(s): transcriptional activator Clusters: cluster:vir:997 Bits: 34 Eval: 3.e-03 %ID: 26 Query range: 15-106 Hit range: 31-118 gi|56692569|ref|YP\_164037.1| 15 MLSELAELIVEAFVRNGLSREKAVSESEELVFQLHRRWAGITFVFPVKDELARKRLELHILQRYDGSNADKLVREFGVTE vir:protein:1974 15 MLSELAELIVEAFVRNGLSREKAVSESEELVFQLHRRWAGITFVFPVKDELARKRLELHILQRYDGSNADKLVREFGVTE 31 LLAELNDLLRGELSRLGVDP----AHSLEIVVAICKHLGGGQVYIPRGQALDSLIRDLRIWNDFNGRNVSELTTRYGVTF vir:protein:1974 107 NTVYKAIRRMRR

Results for query sequence gi|56692571|ref|YP\_164039.1| Blast search results in Viruses No hits found!

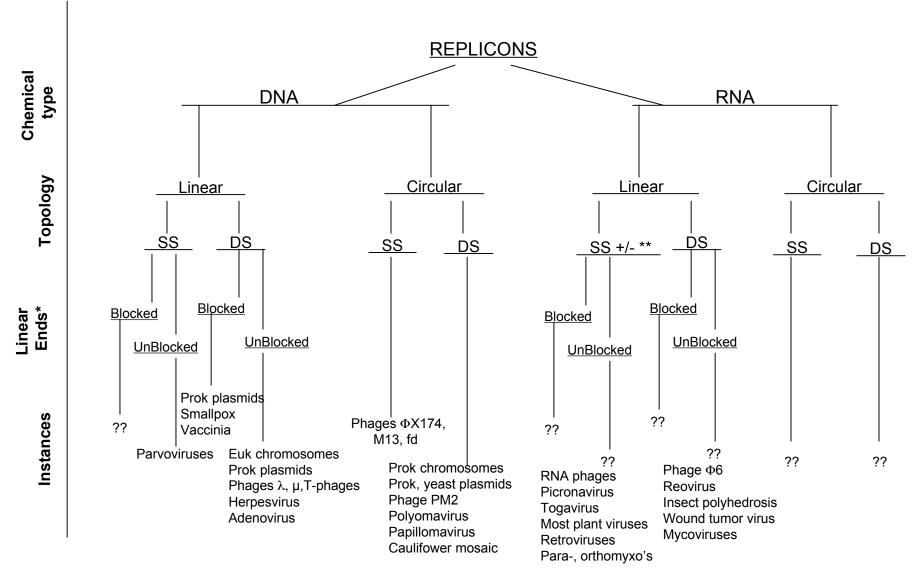
## Prophinder

Using the information in ACLAME to detect other MGEs in bacterial genomes.



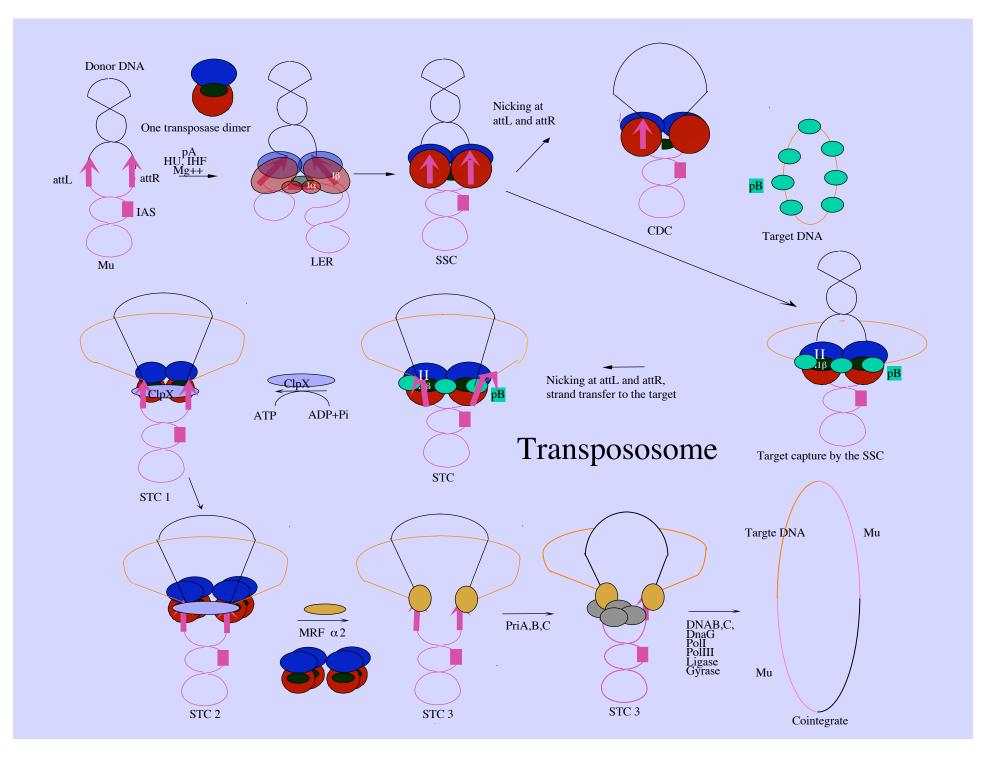
000		ACLAME: Prophinder							
	S C + Antr://aclame.s	.scmbb.ulb.ac.be/nerl/Aclame/Pronhages/view_nronhage.cgi?tvne=genome&bactGen=NC_0043GQGoogle ACLAME: Prophinder							
🛱 Ren	+ Shttp://aclame.scmbb.ulb.ac.	c.be/perl/Aclame/Prophages/view_prophage.cgi?type=prophage&bactGen=NC_004347&proph=NC. S ~ Q- Google							
Browse	Databases  Annotation  Biblio	Societes  Personal Toolbar Folder ULB  Unclassified		1					
Descriptic			6						
Proposal	Prophage ID NC_004347:3	:300:2 Size 26.87 kb (26873 bp) Return to genome view	1	h					
Foru	Bacterial genome NC_004347 - Shewanella oneidensis MR-1, complete genome. Size 4969.80 kb - (4969803bp) circular								
Ontology	Nb CDS in prophone 20 With ACLAME bits 25 Number of DBs 0 Info								
Resor	▲ ► 🏠 😋 + Chitp://aclame.scmbb.ulb.ac.be/perl/Aclame/show_cluster.cgi?mode=clust&id=cluster:vir:997&class=family&vers= ^ Q- Google								
People	NC_004347 0•0mb □ Rendezvous ▼	NC_004347 0.000 CR Rendezvous v Databases v Annotation v Biblio v Societes v Personal Toolbar Folder v ULB v Unclassified v							
Communi	300 CDS ACLAME	Provide Provid		l					
Annotatio		Login: Password:	Go	l					
Ontologie	NC_004347:300:2	Classification Browse Description	Proposal						
Genomes	CDS Home								
Financial	Classification	Clusters view							
Publicatio	Browse	Categories (?) : Viruses () (Filter) Jump (?) : Genomes () Go Version 0.2 Change to (?) : 0.2 () Go	L L						
Το	Control panel			l					
Annotator	Proposal Forums	Clustered proteins in families filtered based on Viruses:		J					
Blast	Ontology	Previous Next Cluster ID Go							
1	Resources	Cluster ID cluster:vir:997 Nb of proteins 2 Cluster level Families							
	Details (?)	Function function:182: transcriptional activator (	h						
	Prophage NC_0 Annotation servers	GOO016563	•						
	List of CDS	660010202							
	GI number C Genomes	$\begin{tabular}{ c c c c c } \hline ID & & & [sw] & $\nabla$ [nrdb] & $\nabla$ [scop] & $\nabla$ & Description & $\nabla$ & Length & $\nabla$ & Genome and host(s) & $\nabla$ & $$							
	24374207 2796 Financial supports								
	24374208 2796 Publications	vir:protein:1978 [+] [5] [33] gpC, late transcription activator 140 Enterobacteria phage Mu Host(s): Escherichia							
	24374209 2797 Annotators	Previous Next Cluster ID Go							
	24374210 2798 Blast								
	24374211 2798 24374212 2798Generated on 12 Septe	ntember 2006 Copyright (C) 2003-2006 The ACLAME team.	Raphael Leplae						
	24374213 2799	Terms of use							
	24374214 2799								
	24374215 2800 24374216 2800								
	24374217 2800								
	24374218 2801								
	24374219 2801								
	24374220 2801			L					
	24374221 2803 24374222 2805			4					
C	24274222 2000			1					
			11	11					

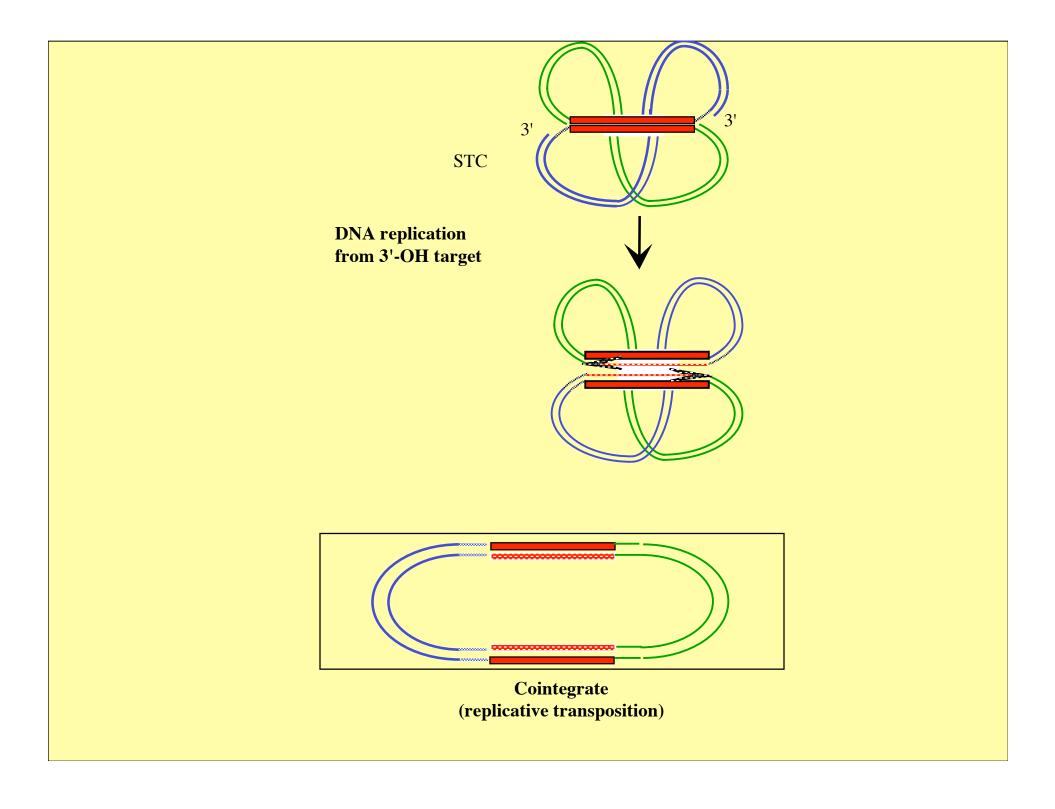


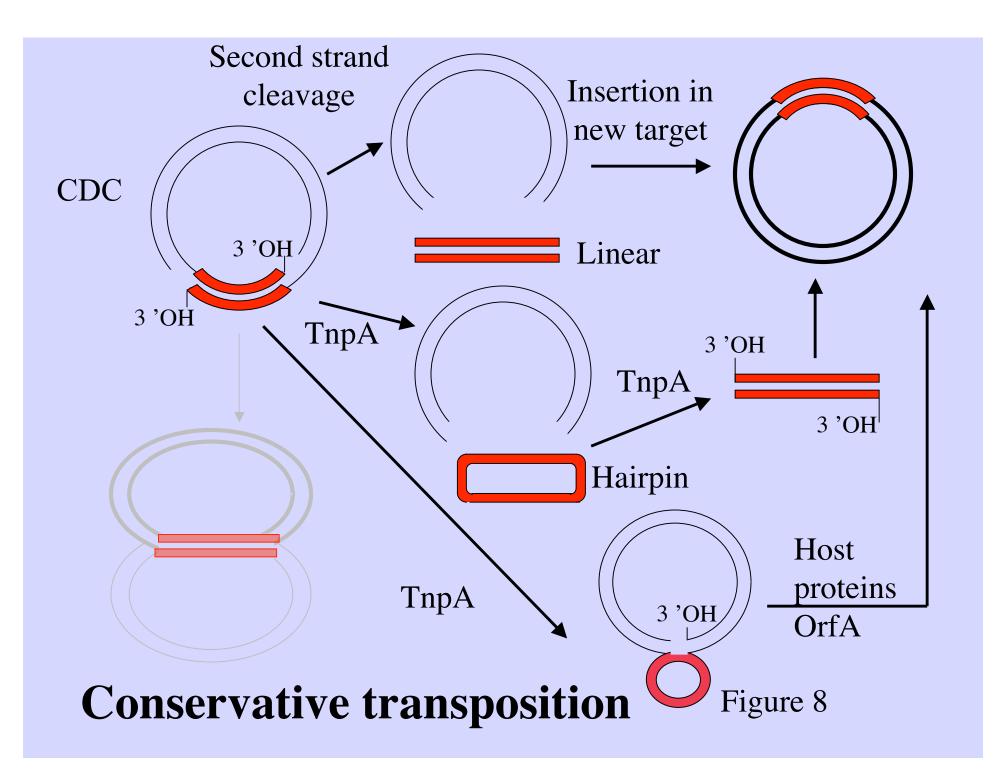


\*Ends of linear replicons can either be free (unblocked) 5'phosphates and 3' hydroxyls or blocked in short hairpins (in SS) or covalently crosslinked (DS). \*\* SS RNA replicons can either be directly translatable mRNA's (+) or complementary to mRNA (-), needing to be copied before used as mRNA.

## **Replicative transposition**







#### attL attR are TnpA binding sites (transpososome) IAS (internal activating sequence of transposition) IHF (integration host factor protein) binding site (specific) HU (histone-like protein) binding site (non specific usually) TnpA (transposase) nicking sites at the ends of attR and attL (concerted) DNA replication priming site? (primosome assembly site)

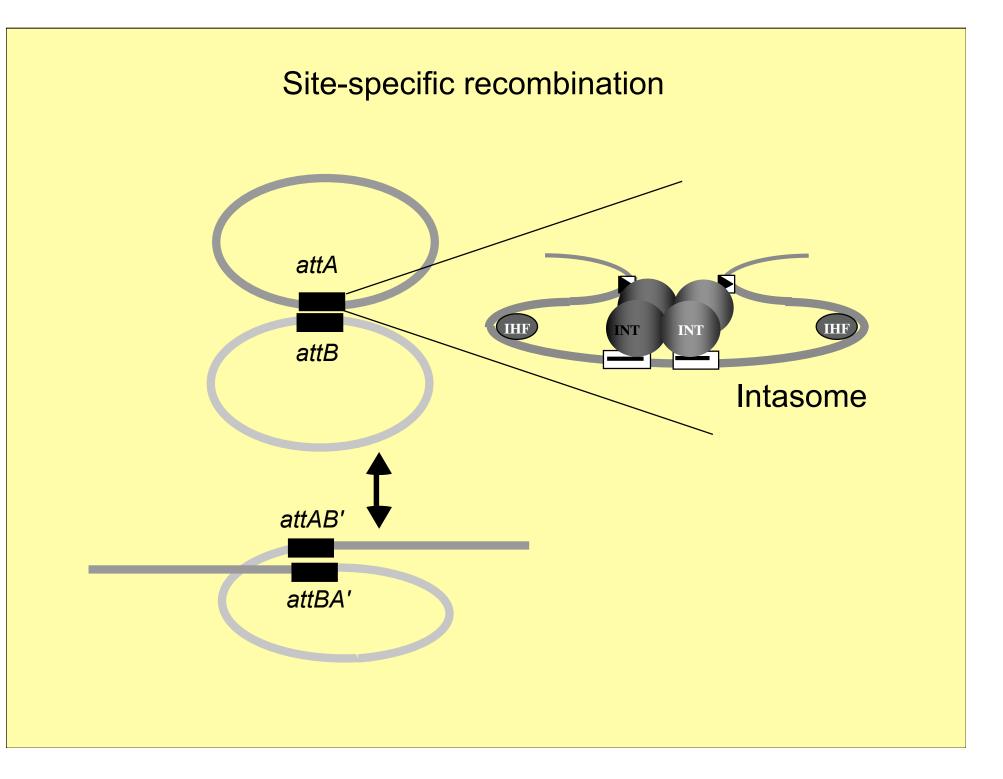
Except for *IAS* and, in many cases, priming assembly site, these are also valid for IS and Tn's, which use DDE transposases. Are usually called *IRL* and *IRR* and in composite transposons distinguish outside and inside *IRs* 



Frameshifting sequence (IS3 family at least) Q: isn't it a process? Signals for frameshift are in RNA but process occurs during translation! Frameshift signals more OK for SO?

X\_XXN ou X\_XXZ\_ZZN and SD 9-15 bp upstream of previous motif or hairpin (IR GC rich) 5-9 bp long after the previous motif.

**Looking at VDJ definitions we will need every gene on the planet** 



For temperate phages:

attP (and also attB for bacterial chromosomes and, attBP', attPB' when inserted prophage).
Int binding sites (distinction to be made between Int N-term and C-term binding site)
Int nicking site (Int stands for Tyr or Ser recombinases here!)
When Ser or Tyr recombinases are resolvases: res site to be subdivided into Res binding sites and Res cleavage site.

These binding and cleavage sites are also valid for other islands, which use Tyr or Ser recombinases to integate and excise, *attI* from integrons and *attC* from <u>gene cassettes</u> - Do they need to be distinguished and if yes how? *Sensu...*  For Tn7-like make sure to also consider TnsA and C (?) binding (cleavage) sites and preferential chromosomal *att* site *attTn* and its TnsD binding site.

And ???

**PROCESSES** Under GO:0006310: DNA recombination. site-specific recombination: integration resolution (transposon-sensu) resolution (plasmid-sensu) transpositional recombination: replicative transpositional recombination conservative transpositional recombination hairpin intermediate figure 8 intermediate linear intermediate illegitimate recombination ?? more??

#### ACTIVITIES In GO:

integrase? tyrosine-based site-specific recombinase activity serine-based site-specific recombinase activity relaxase **PROPOSAL:** site-specific recombinases integrase (sensu retrovirus): see DDE transposase integrase (sensu phage-prophage, genomic island and conjugative transposon) tyrosine-based site-specific recombinase activity serine-based site-specific recombinase activity excisionase resolvase? DDE transposase relaxase

Also need to deal with

- Tn916 family transposases that are Tyr-Rec integrases
- IS607 transposases which may be Ser recombinases
- transposases that are 'relaxases'
  - -IS605-608, Y transposases
  - -YII IS91 RC replication
  - -...MIMIVIRUS

And : <u>IS91 is a helitron</u> NO rather helitrons are similar to IS91 family (see IS-Finder DB) now called IS*CR* elements (Toleman et al. MMBR 70, 296 (2006). PNAS ref in OBO only refers to fly.

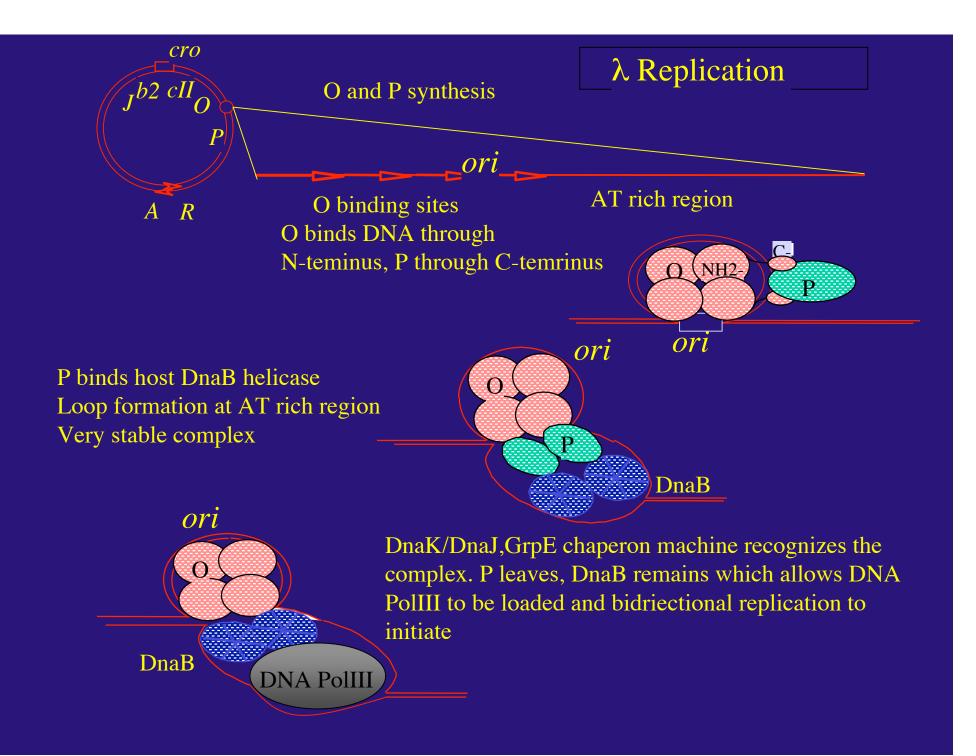
COMPONENTS (primosome and replisome exist in GO) intasome

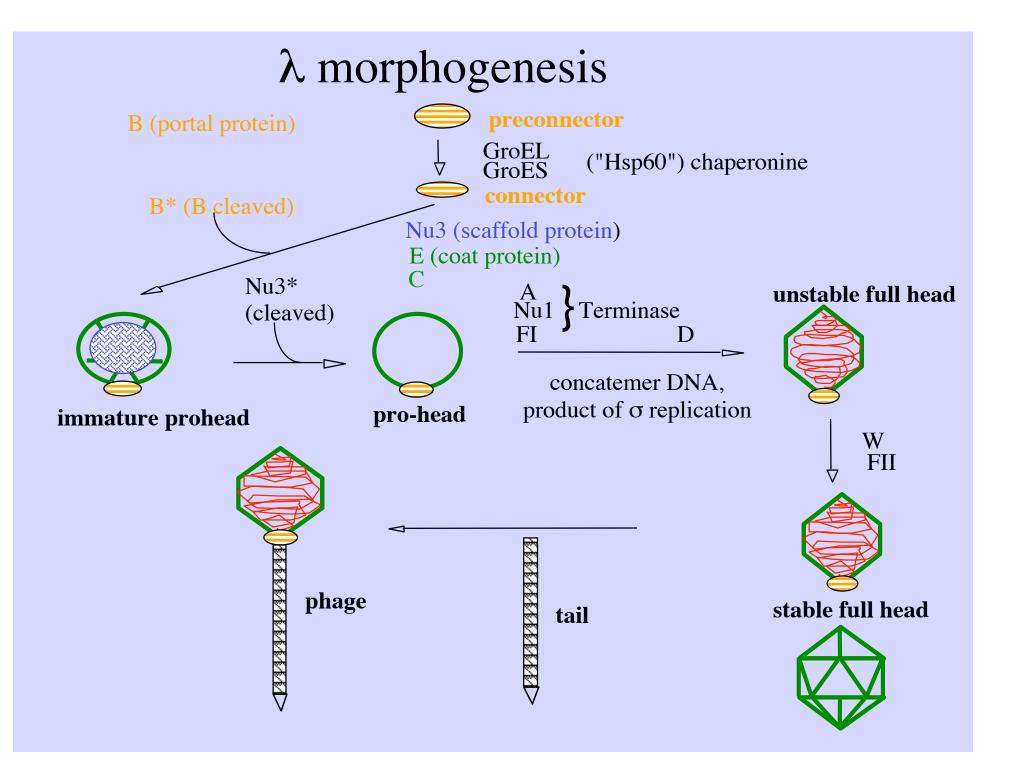
intermediates?? excisome? intermediates?? relaxosome intermediates?? transpososome LER complex stable synaptic complex (or type 0) cleaved donor complex (or type 1) diverse strand transfer complex(es) (including type 2) others?

#### BACK to PHAGES...

Infection and lytic cycle...







That means for all phages but transposable ones: *oriV*, replication proteins binding and nicking sites (primosome, replisome) For all phages:

regulatory sequences (same as for bacteria including transcription antitermination sites)

terminase binding sites

*pac cos* (3' and 5' extended) site, terminase nicking site

*Note: RCR to be ditinguished between phages and e.g. helitrons?* 

Various questions:

Where do we enter the phage lytic cycle, the related subprocesses, phage components and related molecular functions (more GO than SO problem)?

Intra vs inter chromosomal transposition. Translocation?

*SO:0000182 etc.* transposition is via a RNA/DNA intermediate: all have DNA, some have RNA intermediates!

DNA region: gene vs cistron (polycistronic and not polygenic mRNA!) gene array gene cassette



Raphaël: ACLAME, developper, Prophinder on the web



Gipsi: Prophinder, graph analysis