

Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a web-based resource.

France Denoeud, Gilles Vergnaud

▶ To cite this version:

France Denoeud, Gilles Vergnaud. Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains: a web-based resource.. BMC Bioinformatics, 2004, 5, pp.4. 10.1186/1471-2105-5-4 . hal-01158321

HAL Id: hal-01158321 https://ensta-paris.hal.science/hal-01158321

Submitted on 31 May 2015 $\,$

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Database

Open Access

Identification of polymorphic tandem repeats by direct comparison of genome sequence from different bacterial strains : a web-based resource

France Denœud^{*1} and Gilles Vergnaud^{1,2}

Address: ¹Laboratoire GPMS, Institut de Génétique et Microbiologie, Bat 400, Université Paris-Sud, 91405 Orsay cedex, France and ²Centre d'Etudes du Bouchet, BP3, 91710 Vert le Petit, France

Email: France Denœud* - France.Denoeud@igmors.u-psud.fr; Gilles Vergnaud - Gilles.Vergnaud@igmors.u-psud.fr * Corresponding author

Published: 12 January 2004

BMC Bioinformatics 2004, 5:4

This article is available from: http://www.biomedcentral.com/1471-2105/5/4

© 2004 Denœud and Vergnaud; licensee BioMed Central Ltd. This is an Open Access article: verbatim copying and redistribution of this article are permitted in all media for any purpose, provided this notice is preserved along with the article's original URL.

Received: 24 September 2003 Accepted: 12 January 2004

Abstract

Background: Polymorphic tandem repeat typing is a new generic technology which has been proved to be very efficient for bacterial pathogens such as *B. anthracis*, *M. tuberculosis*, *P. aeruginosa*, *L. pneumophila*, Y. *pestis*. The previously developed tandem repeats database takes advantage of the release of genome sequence data for a growing number of bacteria to facilitate the identification of tandem repeats. The development of an assay then requires the evaluation of tandem repeat polymorphism on well-selected sets of isolates. In the case of major human pathogens, such as S. aureus, more than one strain is being sequenced, so that tandem repeats most likely to be polymorphic can now be selected *in silico* based on genome sequence comparison.

Results: In addition to the previously described general Tandem Repeats Database, we have developed a tool to automatically identify tandem repeats of a different length in the genome sequence of two (or more) closely related bacterial strains. Genome comparisons are precomputed. The results of the comparisons are parsed in a database, which can be conveniently queried over the internet according to criteria of practical value, including repeat unit length, predicted size difference, etc. Comparisons are available for 16 bacterial species, and the orthopox viruses, including the variola virus and three of its close neighbors.

Conclusions: We are presenting an internet-based resource to help develop and perform tandem repeats based bacterial strain typing. The tools accessible at http://minisatellites.u-psud.fr now comprise four parts. The Tandem Repeats Database enables the identification of tandem repeats across entire genomes. The Strain Comparison Page identifies tandem repeats differing between different genome sequences from the same species. The "Blast in the Tandem Repeats Database" facilitates the search for a known tandem repeat and the prediction of amplification product sizes. The "Bacterial Genotyping Page" is a service for strain identification at the subspecies level.

Background

Molecular epidemiology, the integration of molecular typing and conventional epidemiological studies, is likely to add significant value to analyses of infections caused by pathogenic bacteria (see [1] for review). Multilocus Sequence Typing (MLST) for instance is now a major reference method for the molecular epidemiology of *Neisseria meningitidis* and other human pathogens [2]. In this

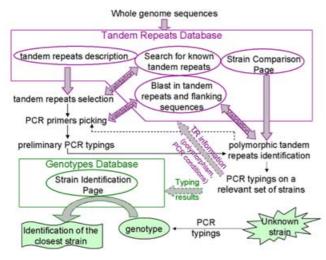


Figure I

The procedure to find polymorphic tandem repeats for use in strain typing. The steps leading from the release of a complete (or incomplete) genome sequence to the validation of new polymorphic markers are described. The purpose of the web-based tools developed is to facilitate the bioinformatics data-management steps.

kind of assay, a set of typically 7 genes is partially sequenced, and the resulting data is converted into sequence types, which can be easily stored in databases, and compared to others. However a number of significant pathogens, including M. tuberculosis [3], B. anthracis and Y. pestis [4] are not amenable to this approach, because of the recent emergence of these pathogens and the resulting rarity of sequence variations. In these pathogens, tandem repeats (TRs) are a source of very informative markers for strain genotyping [5-10]. Tandem repeats in pathogenic bacteria were initially identified within genes associated with bacterial virulence [11,12]. In other instances, the contribution of tandem repeats to genome polymorphism was established after extensive searches based for instance on AFLP (amplified fragment length polymorphism) profiling. This is well illustrated by B. anthracis, in which polymorphic bands in AFLP patterns [13] were subsequently demonstrated by sequencing to be due to tandem repeat variations [14]. Eventually, some of these tandem repeats have been shown to directly contribute to phenotypic variations of the B. anthracis exosporium which makes the outer layer of the spores [15]. The frequent observation that tandem repeat-containing genes are often associated with outer membrane proteins suggests that such genes help bacteria adapt to their environment, and may be to some extent mutation hotspots as a result of positive selection.

Comparison	%matches between flanking sequences	% of flanking seqs not rearranged	% of poly- morphic tandem repeats	positions of matching tandem repeats
Agrobacterium tumefaciens C58 Cereon/UWSC	99.99	99.95	0	
Salmonella enterica typhi CT18/Ty2	99.96	98.83	2.04	$\mathbf{\mathbf{N}}$
Yersinia pestis CO-92/KIM5 P12	99.96	98.51	8.47	1
Staphylococcus aureus Mu50/N315	99.95	98.05	1.38	/
Chlamydia pneumoniae CWL029/AR39	99.94	99.86	1.41	
Chlamydia pneumoniae CWL029/J138	99.92	99.86	2.30	
Mycobacterium tuberculosis H37Rv/CDC1551	99.90	99.50	4.06	
Escherichia coli O157:H7 Sakaï/EDL933	99.87	99.49	2.13	
Brucella suis 1330/ melitensis 16M	99.56	99.08	7.85	
Streptococcus agalactiae NEM316/2603	99.13	85.44	2.76	
Streptococcus pneumoniae TIGR4/R6	99.83	90.01	16.50	
Staphylococcus aureus Mu50/NCTC8325	98.75	92.35	8.12	$\mathbf{\mathbf{X}}$
Streptococcus pyogenes M1GAS/M3GAS	98.71	86.30	6.38	
Staphylococcus aureus Mu50/MW2	98.68	91.04	8.24	1.
Staphylococcus aureus Mu50/MRSA252	98.68	90.17	8.81	1:1
Streptococcus pyogenes M1GAS/M18GAS	98.55	89.19	5.82	
Salmonella typhimurium LT2 / typhi CT18	98.29	87.65	5.44	1
Escherichia coli O157:H7 Sakaï/ K12	97.96	77.23	8.92	
Neisseria meningitidis MC58/Z2491	97.54	92.66	18.88	X
Helicobacter pylori 26695/J99	94.10	87.23	18.91	1
Listeria mono- cytogenes EGDe / innocua Clip	90.19	74.13	3.99	1
Rickettsia prowazekii Madrid E / conorii malish7	89.23	61.05	8.23	
Salmonella typhimurium / Shigella flexneri	86.06	47.16	6.23	1.

Figure 2

Comparison of strains using different indexes. The four columns correspond to (from left to right): (1) mean %identity provided by BLAST when the match occurred on more than half the length of the 500 bp of submitted flanking sequence; (2) proportion (%) of flanking sequences that matched on more than half their length between the two strains; (3) proportion (%) of tandem repeats of a different size in the two strains; and (4) plot of the positions of homologous tandem repeat loci in the two genomes which indirectly reflects large scale genome rearrangements. Species are listed according to the first index (mean %identity)

Strain Comparison Page

	C	Agrobacteriu	m tumei	aciens	C58 (Cereor	/UWSC) [99	.999%]		• Brucella st	iis (133	(0)/ B	rucell	a mel	itensis ()	16M) [99.569	6]		
	c	Helicobacter	pylori	26695	/J99) <mark>(94.10</mark> 9	6]			• Listeria mo	onocyto	genes	(EGI)-e)/	L innocu	a(Clip11262	2) (90.1	9%)	
	ø	Mycobacteri	um tuber	culosi	s (H37Rv/Cl	DC1551) [99.	90%]		· Neisseria	neningi	tidis	MC58	/Z249	1 [97.55	%]			
	Ċ	Rickettsia pr	owazek	ii (Mad	rid E)/ R co	norii (malish	7) (89.23)	×-]	Streptococ	cus aga	lactio	e NEI	M316	/2603 [9	9.13%]			
	C	Streptococcu	s pneun	ioniae	TIGR4/R6)	[98.83%]			• Yersinia p	estis (C	0-92	/KIM	5 P12)) [99.96*	%]			
	C	Chlamydia p	neumoni	ae CW	L029/J138 [99.92%]			Chlamydia	a pneum	onia	CWL	029/4	AR39 [96	9.94%]			
		Chlamydia J	neumo	uae 3	strains comp	arison (CWL	029/313	8/AR	(39) availabl	e at: li	nk							
	C	Escherichia d	oli (O1	57:H7	Sakai/ K12)	[97.96%]			CEscherichi	ia coli (0157	:H7 (S	akaī/l	EDL933) [99.87%]			
		Escherichia	colt 3 s	trains	comparison	(0157:H7 Sa	ikat/015	7:H7	EDL933 / K	12) ava	ilable	r at: 1	nk					
	c	Staphylococo	us aure	us Mu5	0/MW2 [98.	58%]			C Staphyloc	occus a	ureus	Mu50	MRS	A252 [9	8.69%]			
	c	Staphylococo	us aure	us Mu5	0/N315 [99.	6%]			C Staphyloc	occus a	ureus	Mu50	/NCT	C8325	[98.75%]			
						arison (MuSt	/ NB15/	MW2										
	c	Streptococcu													GAS2832) (S	98.55%	1	
						arison (MI		GASS										
	C	Salmonella t							C Salmonell		-				la typhi Tv2	[99.96%	10	
		Salmonella t					%]		C Salmonell								644	
		The second states of the			and the second sec	himurium / 1	a second							0.000		50 - 10 7 105 - 1		
	0	variola/came							C variola/va									
		variola/cowp							⊂ vaccinia/c									
		10.00			All sure seams	tola / vaccin	a lamel											
	Pal	lect a crite			rectances (real				company and	indure .		102						
		Total Lo	ength		Unit Leng		Copy Nur	mber		match	es		min	%GC	_			
		min : 0			min : 0		nin : 0		min :	P	_		min	: 0				
		max : 500	00	1	max : 500	п	ax : 6000		max	100			max	100				
						-		multij					max					
		max : 500	nly tan	dem r		n length dif		multij	ple of unit				max					
		max : 500	nly tan	dem r	epeats witl	n length dif		multij	ple of unit				max					
		max : 500	nly tan	dem r	epeats with	min : 1 max : 1	ference	multi _j Su	ple of unit				max					
oba		max : 500	nly tan	dem r	epeats with	min : 1 max : 1	ference	multi _j Su	ple of unit				max					
	cteriu	max : 500	nly tan loci whe osis (H	dem n re the '	epeats with TR matches:) / Mycoba	min : 1 max : 1 cterium tub	ference i	multij Su s (CD	ple of unit i	length			max					
	cteriu	max : 500	nly tan loci whe osis (H gth dif	dem r re the 137Rv) ference	epeats with TR matches: / Mycoba ce betweer	min : 1 max : 1 cterium tub	ference i	multij Su s (CD	ple of unit i	length 00 bp		s (CD		: 100				
dem	cteriu	max : 500	nly tan loci whe osis (H gth dif	dem r re the 137Rv) ference	epeats with TR matches: / Mycoba ce betweer (H37Rv)	min : 1 max : 1 cterium tub	ference i	s (CD 5 bp	ple of unit i ibmit	length 00 bp		s (CD)	01551	; [100	TR name		Nbr of	
ition	cteriu repe Total length	max : 500	loci whe osis (H gth dif um tuber it Copy th nbr	dem n re the 37Rv) ference culosis	epeats with TR matches: / Mycoba ce betweer (H37Rv) GC Sequenc	min : 1 max : 1 sterium tub the two st	ference i erculosi- rains >= Position 24648-	s (CD 5 bp Total length 160	ple of unit i abmit)C1551) and <= 50 Mycobacteriu	00 bp	culosi Copy nbr	% match	C1551 % SC S	; 100	TR: name H37Rv_0024	diff	Nbr of matches	orien
dem ition 1648- 4825	repe Total length 178	max : 500 T Select or Number of Im tubercul ats with len Mycobacter Contig Un eny MT_H37Ry 18	loci whe osis (H gth dif um tuber it Copy th nor 10	dem r re the 37Rv 137Rv ference % match 70%	epeats with TR matches: / Mycobae ce betweer (H37Rv) (H37Rv) 6 6 6 6 6 6 6 6 6 6 9 9 9 10 9 10 9 10	min : 1 max : 1 terium tub the two st	ercolosi rains >= Position 24648- 24807	s (CD 5 bp Total length 160	ple of unit i abmit CC1551) o and <= 50 Mycobacterite Contig MT_CDC1551	00 bp m tuber Unit length	culos Copy nbr 9	96 match 6796	01551 % S GC S 73% g	; [100) iequence	H37RV 0024	diff 18	matches 1	orien +
ition 1648- 4825 9503- 9562	repe Total length 178	max : 500 r Select or Number of um tubercul ats with len Mycobacteri Contig len MT_H37RV 18 MT_H37RV 9	osis (H gth dif um tube it Copy th nbr	dem n re the 137Rv) ference reulosis % match 70%	rR matches: / Mycoba- ce betweer (H37Rv) % % % Sequenc 72% planmen	min : 1 max : 1 terium tub the two st H37Rv 0024	ercolosi rains >= Position 24848- 24807 74907 79550	s (CC 5 bp Total length 160	ple of unit i abmit C1551) and <= 50 Mycobacterit Contig MT_CDC1551	00 bp m tuber Unit length 18	culosi Copy nbr 9 7	96 match 6796 9696	C1551 % SGC S 73% a 77% a	; [100) iequence ignment	TRC FIBILITY	diff 18 9	matches 1 1	orien + +
ition 1648- 4825 1503- 9562 1881-	repe Total length 178	max : 500 T Select or Number of Im tubercul ats with len Mycobacter Contig Un eny MT_H37Ry 18	osis (H gth dif um tube it Copy th nbr	dem n re the 137Rv) ference reulosis % match 70%	epeats with TR matches: / Mycobae ce betweer (H37Rv) (H37Rv) 6 6 6 6 6 6 6 6 6 6 9 9 9 10 9 10 9 10	min : 1 max : 1 terium tub the two st H37Rv 0024	rercolosi rains >= 24648- 24607 79432- 79550 149870-	s (CC s 5 bp Total length 160	ple of unit i abmit CC1551) o and <= 50 Mycobacterite Contig MT_CDC1551	00 bp m tuber Unit length 18	culosi Copy nbr 9 7	96 match 6796 9696	C1551 % SGC S 73% a 77% a	; [100) iequence	H37RV 0024	diff 18	matches 1	orier + +
ition 1648- 4825 1503- 9562 1881- 0864 1010-	repe Total India 178 60/	max : 500 r Select or Number of um tubercul ats with len Mycobacteri Contig len MT_H37RV 9	osis (H gth dif um tuber it Copy th nor 10 6 113	dem n re the 337Rv) ference culosis ference 70% 53%	rR matches: / Mycobae ce betweer (H37RV) GC Sequenc 72% elignmen 77% elignmen	min : 1 max : 1 terium tub the two st H37Rv 0024	reculosi rains >= Position 24648- 24807 79482- 79550 149870- 151033 424005	multi Su S (CC) S (CC)	ple of unit i abmit C1551) and <= 50 Mycobacterit Contig MT_CDC1551	00 bp m tuber Unit jength 18 19	culosi Copy nbr 9 7 134	96 match 6796 9696 5396	C1551 % SC SC S73% a 80% a) iequence ianment ianment	H37RV 0024	diff 18 9	matches 1 1	orier + +
dem ition i648- 4825 i503- 9562 i881- 0864 i010- 4141	Total Total 178 60 1 984 1	max : 500 r Select or Number of im tubercul ats with len Mycobactor Contig Ur eny MT_H37Rv 9 MT_H37Rv 9 MT_H37Rv 51	loci whe osis (H gth dif 10 10 6 113 3	dem n re the 337R v) ference 100% 53% 100%	epeats with TR matches: / Mycoba- ce betweer (H37Rv) % % % % % % % % % % % % % % % % % % %	n length diff min : 1 max : 1 :terium tub the two st H37Rv 0024 H37Rv 0024 H37Rv 0024	recolosi rains >= Position 24648- 24907 79432- 79550 149870- 151033 424095 424298	Su	ple of unit i ibmit CC1551) and <= 50 Mycobacterit Contig MT_CDC1551 MT_CDC1551 MT_CDC1551	length 00 bp m tuber 18 9 9 51	culosi Copy nbr 9 7 134 5	96 match 6796 9696 5396 10096	C1551 96 SC S 973% a 277% a 80% a 74% a) iequence ignment ignment ignment	H37Rv 0024	diff 18 9 180	natches 1 1 1	orier + +
dem 1000 1648- 4825 1503- 9562 1881- 0864 1010- 4141 1871- 7298	Total Total length 178 984 132 2428	max : 500 r Select or Number of um tubercul ats with len Mycobacter Contog MT_H37Rv 9 MT_H37Rv 9 MT_H37Rv 9 MT_H37Rv 51 MT_H37Rv 15	osis (H gth dif it Copy it nor 10 6 113 3 268	dem n re the 337Rv) ference (culosis ference 53% 53% 53% 53%	rR matches: / Mycoba- ce betweer (H37Rv) % Sequenc 72% plignmen 73% plignmen 50% plignmen 51% plignmen 51% plignmen	n length diff min : 1 max : 1 oterium tub the two st H37Rv_0024 H37Rv_0024 H37Rv_0424	reculosi: rains >= Position 24648- 24807 79482- 79550 149870- 151033 424005 424298 425294	multij Su 5 bp Total length 160 69 1164 234 2413	ple of unit l abmit CC1551) and <= 50 Mycobacteric Contig MT_CDC1551 MT_CDC1551 MT_CDC1551 MT_CDC1551	00 bp um tuber Unit Nength 18 9 9 51	culosi Copy nbr 9 7 134 5 132	96 match 6796 9696 5396 10096 5996	C1551 % S GC S 7396 a 7796 a 8096 a 8196 a 8196 a) iequence ignment ignment ignment ignment	H37Rv 0024	diff 18 9 180 102 15	1 1 1 1 1 1	orier + +
dem ition i648- 4825 i503- 9562 i881- 0864 i010- 4141	Total Total length 178 984 132 2428	max : 500 r Select or Number of im tubercul ats with len Mycobactor Contig Ur eny MT_H37Rv 9 MT_H37Rv 9 MT_H37Rv 51	osis (H gth dif it Copy it nor 10 6 113 3 268	dem n re the 337Rv) ference (culosis ference 53% 53% 53% 53%	epeats with TR matches: / Mycoba- ce betweer (H37Rv) % % % % % % % % % % % % % % % % % % %	n length diff min : 1 max : 1 oterium tub the two st H37Rv_0024 H37Rv_0024 H37Rv_0424	rains >= Position 24648- 24807 7982- 7950 142870- 151033 4240256 425029-	multi Su Su Su Su Su Su Su Su Su Su Su Su Su	ple of unit i ibmit CC1551) and <= 50 Mycobacterit Contig MT_CDC1551 MT_CDC1551 MT_CDC1551	00 bp um tuber Unit Nength 18 9 9 51	culosi Copy nbr 9 7 134 5 132	96 match 6796 9696 5396 10096 5996	C1551 % S GC S 7396 a 7796 a 8096 a 8196 a 8196 a) iequence ignment ignment ignment	H37Rv 0024	diff 18 9 180 102	matches 1 1 1 1	oner + + +
dem 4648- 4825 2503- 9562 2881- 0864 1010- 4141 1871- 7298 7284-	Total Total length 178 984 132 2428	max : 500 r Select or Number of um tubercul ats with len Mycobacter Contog MT_H37Rv 9 MT_H37Rv 9 MT_H37Rv 9 MT_H37Rv 51 MT_H37Rv 15	osis (H gth dif it Copy it nor 10 6 113 3 268	dem n re the 337Rv) ference (culosis ference 53% 53% 53% 53%	rR matches: / Mycoba- ce betweer (H37Rv) % Sequenc 72% plignmen 73% plignmen 50% plignmen 51% plignmen 51% plignmen	n length diff min : 1 max : 1 cterium tub 1	rains >= Position 24648- 24807 79452 79452 79452 42029 42442 578726	multij Su Su Su Su Su Su Su Su Su Su Su Su Su	ple of unit l abmit CC1551) and <= 50 Mycobacteric Contig MT_CDC1551 MT_CDC1551 MT_CDC1551 MT_CDC1551	00 bp m tuber 118 9 9 51 300	culosi Copy nbr 9 7 134 5 132 3	96 match 6796 5396 10096 5996	C1551 96 SC S 7796 a 8096 a 8096 a 8196 a 6196 a) iequence ignment ignment ignment ignment	H37Rv 0024	diff 18 9 180 102 15	1 1 1 1 1 1	orien

Figure 3

960165-960321

Example of a query in the Strain Comparison Page. On the top, the query page shows the 28 comparisons currently available (others will be added as new genome sequences are finished and released). Bottom, the result of a query performed for *Mycobacterium tuberculosis* strains H37Rv and CDC1551 is summarized.

960053- 263 MT_CDC1551 53 960315

3297804

-3297841 38

336763 alignment /

-3367673 38

2

19

						au di	il ou	amp	arison	aye. I	lore	and	-	su anis							
							-		be compa	red:											
									157:H7 EDI	933/E12/	UPBC	TOTA:	1								
									enterica typi					hi Tyz							
				C Sale	nonella typi	inaia	m / Sala	scorfla	enterica typi	E CTIS/S	lacoell	a enter	ica 179	hi Ty2 / Ship	illa Bezzeri						
				C Stag	hylococcus	areus	Mu50	/NG1	1/MW2/M	RSA252/N	CICSE	25									
									AB GASSIS		52832	MJ S	511								
				C Ord	hopox virus	HC VARI	ola/va	ccinia (canalpox l	compos											
				N	umber of c	listinc	t allele	s four	nd in the se	ected str		in:1	=								
				c 1	Select only	r tande	em rep	eats v	with known	lengths in	all the	strain	ns								
									l	submit											
									eria below i												
					Total Leng			init Le		Copy Nu		T	%m	atches	%	GC					
					iin : 0 ARC : 50000	3	min max	500		min : 0 max : 0000			nin : 0 nax : 1		min:0 max:1	00					
										submit											
				o157:H7																	
train B: train C:		nerichi		0157:H7	(EDL93	3)															
				UPEC C	FT073																
andem	repe	ats w	ith 2	to 4 dist	inct alle	les ar	nong	the 4	4 strains:												
A	A	A Unit	A	A	B	B Total	8 Unit	B	B	с	C Total	C	Copy	С	D	D Total	D	D	D	TR	Noro
Position	length	length	rter	sequence elignment /	Position	length	liength	nbr	Sequence	Posãon	length	length	nbr	Sequence	Position	length	length	nor	Sequence	Name	ateles
46590	53	6	10			80		+0	alignment /	201 225	-	+0		alignment /	201 017		40	2.2	alignment /	-	
				Inear seg		56	8	10	Inear seg	201235	36	12	3	anear seg	201947	747	12	3,2	alignment / Inear.seg	_	4
46724	135	57	2,3	inear.sed alignment / inear.seg	48591- 48725	56 136	6 57	10 2,3	linear seg alignment / linear seg	42183 42311	36 129	12 57	3	inear seg algoment / inear seg	44835 44969	747 135	12 57	3,2 2,3	linear seg alignment / linear.seg		4
	136 189	57 85		inear.sed alignment / inear.sed alignment /	48591- 48725			-	Inear seg alignment / Inear seg alignment /	42183 42311	1000	_	-	alignment / alignment / alignment /	44835 44969			-	linear.seg alignment / linear.seg alignment /		-
71495 71683 133896		-	2,3	inear sea alignment / inear sea alignment / inear sea alignment /	46591- 46725 71495- 71684 133897	136	57	2,3	Inear seg alignment / Inear seg alignment / Inear seg alignment /	42183 42311 66554 66837 129524	129	57	2	inear seg algoment / inear seg algoment / inear seg algoment /	44835 44969 69078 69351 139926	135	57	2,3	inear seg alignment / inear seg alignment / inear seg alignment /		2
71495- 71683 133896- 134273 134629-	189 378	85 96	2,3 2,2 3,9	Inear sed elignment / Inear sed elignment / Inear sed elignment /	46591 46725 71496 71684 133897 134274 134630	135 189	57 85 96	2,3 2,2 3,9	Inear seg alignment / Inear seg alignment / Inear seg alignment /	42183 42311 66564 66837 129524 129901 130257	129 274	57 85 98	2 3	Brear seq alignment / Brear seq alignment / Brear seq alignment / Brear seq	44835 44969 69078 69351 139926 140207 140563	135 274 282	57 85 96	2,3 3,1 2,9	Inear seg alignment / Inear seg alignment / Inear seg alignment / Inear seg alignment /		2 2 2
71495- 71683 33896- 134273 34629- 134668	189 378 40	85 96 21	2,3 2,2 3,9 1,9	Inear sea alignment / Inear sea alignment / Inear sea alignment / Inear sea alignment / Inear sea	48591 48725 71495 71684 133897 134274 134630 134669	136 189 378 40	57 85 96 21	2,3 2,2 3,9 1,9	Inear seg algrment / Inear seg algrment / Inear seg algrment / Inear seg algrment / Inear seg	42183 42311 66564 66837 129524 129901 130257 130296	129 274 378 40	57 85 98 9	2 3 4 7	inear seg algoment / inear.seg algoment / inear.seg algoment / inear.seg algoment / inear.seg	44835 44969 69078 69351 139926 140207 140563 140626	135 274 282 64	57 85 96 24	2.3 3.1 2.9 3,08	Inear seg alignment / Inear seg alignment / Inear seg alignment / Inear seg alignment / Inear seg		2 2 2 2
71495- 71683 133896- 134273 134629- 134668 191255-	189 378	85 96	2,3 2,2 3,9	inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg	46591 46725 71496 71684 133897 134274 134630 134669 391257 391707	135 189 378	57 85 96	2,3 2,2 3,9	Inear seg alignment / Inear seg alignment / Inear seg alignment /	42183- 42311 66564- 66837 129524- 129901 130257- 130296 338982-	129 274 378	57 85 98	2 3 4	inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg	44835- 44969 69078- 69351 139926- 140207 140563- 140626 430253- 430612	135 274 282	57 85 96	2,3 3,1 2,9	Inear seg bigrment / integr.seg bigrment / integr.seg bigrment / integr.seg bigrment / integr.seg bigrment / integr.seg		2 2 2
71495- 71683 133896- 134273 134629- 134688 191255- 391705 111014-	189 378 40	85 96 21	2,3 2,2 3,9 1,9	Inter seg alignment / Inter seg alignment / Inter seg alignment / Inter seg alignment / Inter seg alignment / Inter seg alignment /	46591 46725 71495 71684 133897 134274 134630 134669 391257 391707 411016	136 189 378 40	57 85 96 21	2,3 2,2 3,9 1,9	Inear seg elignment / Inear seg elignment / Inear seg elignment / Inear seg elignment / Inear seg elignment /	42183 42311 66564 66837 129524 129901 130257 130296 338982 339339 353823	129 274 378 40	57 85 98 9	2 3 4 7	inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg	44835 44969 69078 69351 139926 140207 140563 140626 430253 430612 443971	135 274 282 64	57 85 96 24	2.3 3.1 2.9 3,08	inear.seq alignment / inter.seq alignment / inter.seq alignment / inear.seq alignment / inear.seq alignment / inear.seq not detected		2 2 2 2
71495- 71683 33896- 134273 34629- 134688 191255- 391705 111014- 411748 161213-	189 378 40 451 735	85 96 21 93 94	2,3 2,2 3,9 1,9 4,9 8	inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg	48591- 48725 71496- 71684 133897- 134274 134630- 134669 391257- 391707 411016- 411750 862867-	136 189 378 40 451 736	57 85 96 21 93 94	2,3 2,2 3,9 1,9 4,9 8	Inear sea eignment / Inear sea eignment / Inear sea eignment / Inear sea eignment / Inear sea eignment / Inear sea eignment /	42183- 42311 66554- 68837 129524- 129901 130257- 130296 338082- 339339 353823- 353993 775930-	129 274 378 40 358 171	57 85 96 9 93 ?	2 3 4 7 4 ?	inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg alignment / inear.seg not detected by TRF alignment /	44835 44969 69078 69351 139926 140207 140563 140626 430253 430612 443971 444137 799377	135 274 282 64 360 167	57 85 96 24 93 7	2.3 3.1 2.9 3,08 3,9 7	ineer.seg alignment / inter.seg alignment / ineer.seg alignment / ineer.seg alignment / ineer.seg alignment / ineer.seg not detected by TRF		2 2 2 2 2 3
71495- 71683 133896- 134273 134629- 134688 191255- 391705 111014- 411748 861213- 851641	189 378 40 451	85 98 21 93	2,3 2,2 3,9 1,9 4,9	Inear.seg algorment / Inear.teg algorment / Inear.teg algorment / Inear.seg algorment / Inear.seg algorment / Inear.seg algorment / Inear.seg	48591- 46725 71495- 71684 133897- 134630- 134630- 134630- 391257- 391707 411016- 411750 862867- 863295	136 189 378 40 451	57 85 96 21 93	2,3 2,2 3,9 1,9 4,9	Inear sea eigement / Inear sea eigement / Inear seg eigement / Inear seg eigement / Inear seg eigement / Inear seg eigement / Inear seg eigement / Inear seg	42183- 42311 66554- 66837 129501- 130256- 130256- 338982- 353833 353823- 353893 775930- 775930- 775930-	129 274 378 40 358	57 85 96 9 93	2 3 4 7 4	inear seg alument / inter.aca alument / inear.aca alument / inear.seg alument / inear.seg not devected devected funder.seg inear.seg inear.seg inear.seg	44835 44969 69078 69351 139926 140263 440263 430253 430612 443971 444137 799377 799886	135 274 282 64 360	57 85 96 24 93	2.3 3.1 2.9 3,08 3,9 7	Inear.seg algoment / Inear.seg algoment / Inear.seg algoment / Inear.seg algoment / Inear.seg not detected by TRF algoment / Inear.seg		2 2 2 2 2
71495- 71683 133896- 134273 134629- 134688 891255- 391705 411014- 411748 861213- 861641 865089-	189 378 40 451 735	85 96 21 93 94	2,3 2,2 3,9 1,9 4,9 8	Inear.seg alazment / Inear.teg alazment / Inear.seg alazment / Inear.seg alazment / Inear.seg alazment / Inear.seg alazment / Inear.seg alazment /	48591- 46725 71495- 71684 133897- 134630- 134630- 134630- 391257- 391707 411018- 411750 862867- 883295 886743-	136 189 378 40 451 736	57 85 96 21 93 94	2,3 2,2 3,9 1,9 4,9 8	Inear sea eignment / Inear sea eignment / Inear sea eignment / Inear sea eignment / Inear sea eignment / Inear sea eignment /	42183- 42311 66564- 60837 129524- 129901 1302256 338982- 353993 353823- 353993 775930- 776439 779887-	129 274 378 40 358 171	57 85 96 9 93 ?	2 3 4 7 4 ?	inear.seg alumners/ intear.seg alumners/ intear.seg alumners/ intear.seg alumners/ intear.seg alumners/ intear.seg not detected by Thef alumners/ intear.seg not detected detected	44835 44969 69078 69351 139925 140207 140626 430253 4430253 4430512 443971 799377 799885 901115	135 274 282 64 360 167	57 85 96 24 93 7	2.3 3.1 2.9 3,08 3,9 7	Insersed bigment / Insersed bigment / Insersed bigment / Insersed bigment / Insersed by TRF bigment / bigment / bigment /		2 2 2 2 2 3
71495- 71683 133896- 134273 134629- 134668 991255- 391705 411014- 411748 861213- 861641 965089- 865244 905293-	189 378 40 451 735 429 156	85 96 21 93 94 15 78	2.3 2,2 3,9 1,9 4,9 8 29 2	Inear sea eligement / Inear sea eligement /	48591- 46725 71495- 71684 133897- 134274 134630- 134669 391257- 391707 411016- 411750 862867- 863295 865295 866943- 866898 868947-	136 189 378 40 451 735 429 156	57 85 96 21 93 94 15 78	2,3 2,2 3,9 1,9 4,9 8 29 2	Inear sea element / Inear sea	42183- 42311 66554- 66837 129524- 129524- 130257- 130296 338939 353823- 353823- 775930- 776439 7759887- 779887- 779887- 779887- 7798191-	129 274 378 40 358 171 610 51	57 85 96 9 93 7 15 7	2 3 4 7 4 7 34 2	inear.seg algement / inear.seg algement / inear.seg algement / inear.seg algement / inear.seg not detected by TRF algement / inear.seg not detected by TRF algement / inear.seg	44835- 44969 69078- 69351 139926 140207 140583- 140628 430612 430612 4430971- 4430971- 799886 803335- 803490 803539-	135 274 282 64 360 167 510 155	57 85 96 24 93 7 15 78	2.3 3.1 2.9 3.08 3.9 7 34,46 2	Iner sea algoment / Inder sea Bigoment / Inder sea Bigoment / Ineer sea algoment / Ineer sea algoment / Ineer sea by TRF algoment / Ineer sea bigoment / Ineer sea algoment / Ineer sea algoment / Ineer sea algoment /		2 2 2 2 3 3
71495- 71683 33896- 134273 34629- 134629- 134688 91255- 391705 411014- 411748 861213- 851641 865089- 865244 865293- 865449	189 378 40 451 735 429	85 98 21 93 94 15	2,3 2,2 3,9 1,9 4,9 8 29	Inext sea eligement / Inext tog eligement / Inext tog eligement / Inext sea eligement / Inext sea eligement / Inext sea eligement / Inext sea eligement / Inext sea eligement / Inext sea	48591- 48725 71496- 71684 133897- 134274 134630- 134629 391257- 391707 411016- 411750 8628429 863826 868674- 866898 868647- 867103	136 189 378 40 451 735 429	57 85 96 21 93 94 15	2,3 2,2 3,9 1,9 4,9 8 29	Inear sea element / Inear sea	42183- 42311 66564- 60837 129524- 129901 130256 338982- 339339 353823- 353993 775930- 776439 776439 779887- 779887- 779847- 779847- 779845	129 274 378 40 358 171 510	57 85 96 9 93 ? 15	2 3 4 7 4 2 34	inear sea alignment / inear sea alignment / inear sea alignment / inear sea alignment / inear sea alignment / inear sea of delected by TRF alignment / inear sea	44835- 44969 69078- 69351 139925- 140207 140526 430253- 430612 443971- 444137 7993876 803335- 803490 803539- 803597	135 274 282 64 360 167 510	57 85 96 24 93 7 15	2,3 3,1 2,9 3,08 3,9 7 34,46	Inear sea algoment / Inear sea Incer sea Incer sea algoment / Inear sea algoment / Inear sea algoment / Inear sea by The' algoment / Inear sea algoment / Inear sea algoment / Inear sea algoment / Inear sea algoment / Inear sea		2 2 2 2 2 3 3 2 2 2 2 2
71495- 71683 33896- 134273 34629- 134688 91255- 391705 411014- 411748 851641 851641 855089- 865244 865293- 865449 065508-	189 378 40 451 736 429 156 167	85 96 21 93 94 15 78	2.3 2,2 3,9 1,9 4,9 8 29 2	Inegr. seg alignment / Inegr. seg	48591- 48725 71486- 71884 13897- 134274 134630- 134274 134680 391257- 391707 411018- 411750 863295 8652657- 865295 865743- 866898 888743- 866983 1154731- 1154806	136 189 378 40 451 735 429 156 157	57 85 96 21 93 94 15 78	2,3 2,2 3,9 1,9 4,9 8 29 2	Inear sea element / Inear sea	42183- 42311 66554- 68837 129524- 129901 130255- 130296 3383892- 353893 353893 775830- 776439 776439 776887- 779887- 779887- 779887- 779887- 779885- 779885- 7980591- 780591- 80591- 80591- 80591-	129 274 378 40 358 171 610 51	57 85 96 9 93 7 15 7	2 3 4 7 4 7 34 2	inear.seg algement / inear.seg algement / inear.seg algement / inear.seg algement / inear.seg not detected by TRF algement / inear.seg not detected by TRF algement / inear.seg	44835- 44969 69078- 69351 139925- 140207 140583- 140683- 430253- 430253- 430253- 430612 443971- 444137 799885 803335- 803395- 803597 988559-	135 274 282 64 360 167 510 155	57 85 96 24 93 7 15 78	2.3 3.1 2.9 3.08 3.9 7 34,46 2	Iner sea algoment / Inder sea Bigoment / Inder sea Bigoment / Ineer sea algoment / Ineer sea algoment / Ineer sea by TRF algoment / Ineer sea bigoment / Ineer sea algoment / Ineer sea algoment / Ineer sea algoment /		2 2 2 2 2 3 3 2 2 2 2 2
71495- 71683 33896- 134273 34629- 134668 91265- 391705 111014- 411748 651641 65089- 865244 865244 865293- 8655449 065508- 1065583 066515-	189 378 40 451 735 429 156 157 76	85 96 21 93 94 15 78 79 39	2,3 2,2 3,9 1,9 4,9 8 29 2 2 2 1,9	Inegr. seg alignment / Inegr. seg	48591- 46725 71486- 71884 13897- 134274 134630- 134274 134680 391257- 391707 411018- 411750 863295 8652657- 865295 865743- 866898 888743- 866983 1154731- 1154806	136 189 378 40 451 735 429 166 167 76	57 85 96 21 93 94 15 78 79 39	2,3 2,2 3,9 1,9 4,9 8 29 2 2 2 1,9	Inear sea eigement / Inear sea eigement /	42183- 42311 66564- 66837 129601 130257- 130256 338939 353823- 353993 775930- 776439 776439 779887- 779987 779987- 779987- 933654 933654 933654 933654	129 274 378 40 358 171 610 51 255 76	57 85 96 9 93 ? 15 ? 79 39	2 3 4 7 4 7 34 7 2 2 2	inear.seq alument/ inear.seq alument/ inear.seq alument/ inear.seq alument/ inear.seq alument/ inear.seq detected by Tief alument/ inear.seq alument/ inear.seq alument/ inear.seq alument/ inear.seq alument/ inear.seq alument/ inear.seq alument/ inear.seq	44835- 44969 69078- 69351 139926- 140207 140563- 430253- 430612 443971- 799886 803547 980359- 803597 988559- 988673 989605-	135 274 282 64 360 167 510 156 69 115	57 85 96 24 93 7 15 78 79 39	2.3 3.1 2.9 3.08 3.9 7 34,46 2 2 3.1	Inser sea algement / Inser sea Bornest / Inser sea algement / Inser sea algement / Inser sea not detected by The algement / Inser sea by The algement / Inser sea by The algement / Inser sea algement / Inser sea algement / Inser sea algement / Inser sea algement / Inser sea algement / Inser sea		2 2 2 2 2 3 3 2 2 3 3 2 2 3 2
71495- 71683 33896- 134273 33629- 134688 91265- 391705 11014- 411748 81641 851641 851641 851641 85293- 885449 065508- 1065583 066515- 1066790	189 378 40 451 735 429 156 157 76	85 96 21 93 94 15 78 79	2.3 2,2 3,9 1,9 4,9 8 29 2 2 2	Inext.seg elegement / Inext.seg elegement / Inext.seg	49591- 48725 71496- 71684 133897- 134630 134630 391257- 391707 411016- 411750 862687- 863295 868743- 866898 86847- 867103 1154731- -1154006 1155736- -1156013	136 189 378 40 451 735 429 166 167 76	57 85 98 21 93 93 94 15 78 79	2,3 2,2 3,9 1,9 4,9 8 29 2 2 2	Inear sea eigement / Inear sea alerment / Inear sea	42183- 42311 66554- 68837 129901 130027- 1300296 338082- 339339 353823- 353893 775930- 775930- 775930- 775930- 775930- 775930- 775937 7764597 780191- 780191- 780454- 933654- 933654- 933654- 933729 934661- 934897	129 274 378 40 358 171 610 51 255	57 85 96 9 93 ? 15 ? 79	2 3 4 7 4 7 34 7 34 7 2	inear.seq algament / inear.seq algament / inear.seq algament / inear.seq algament / inear.seq algament / inear.seq algament / inear.seq algament / inear.seq algament / inear.seq	44835- 44969 69078- 69351 139926- 140583- 140583- 140525- 4430612 4430612 4430612 4430612 443071- 444137 799886 803335- 803490 803539- 803597 9888673 9888673 9889605- 989856	135 274 282 64 360 167 510 155 59	57 85 96 24 93 7 15 78 79	2.3 3.1 2.9 3.08 3.9 7 34,46 2 2 2	Ineer sea Internet / Internet /		2 2 2 2 2 3 2 2 2 3 2 2 3 3
71495- 71683 33896- 134273 34629- 134688 91255- 391705 111014- 411748 61213- 861641 65089- 885449 665508- 1065583 0665563 066555 1066790 520733-	189 378 40 451 735 429 156 167 76 276	85 96 21 93 94 15 78 79 39	2,3 2,2 3,9 1,9 4,9 8 29 2 2 2 1,9	Inear sea discriment / Inear sea	49591- 48725 71496- 71684 133897- 134630 134630 391257- 391707 411016- 411750 862687- 863295 868743- 866898 86847- 867103 1154731- -1154006 1155736- -1156013	136 189 378 40 451 735 429 156 167 76 276	57 85 96 21 93 94 15 78 79 39	2,3 2,2 3,9 1,9 4,9 8 29 2 2 2 1,9	Ineer ses elgement / Ineer ses elgement /	42183- 42311 66554- 68837 129901 130027- 1300296 338082- 339339 353823- 353893 775930- 775930- 775930- 775930- 775930- 775930- 775937 7764597 780191- 780191- 780454- 933654- 933654- 933654- 933729 934661- 934897	129 274 378 40 359 171 610 51 255 76 237	57 85 96 9 93 ? 15 ? 79 39	2 3 4 7 4 7 34 7 2 2 2	inear.seq Blarmert / Inear.seq alignmert / Inear.seq alignmert / Inear.seq alignmert / Inear.seq alignmert / Inear.seq not delected by TRF elignmert / inear.seq alignmert / alignmert / inear.seq alignmert / inear.seq alignmert / alignmert / a	44835- 44969 69078- 69351 139926- 140207 140563- 430253- 430612 443971- 799886 803547 980359- 803597 988559- 988673 989605-	135 274 282 64 360 167 510 155 69 115 252	57 85 96 24 93 7 15 78 79 39	2.3 3.1 2.9 3.08 3.9 7 34,46 2 2 3.1	Inser sea diarment / Indef sea		2 2 2 2 2 3 3 2 2 3 3 2 2 3 2
71495- 71683 33896- 33896- 134688 338273 334629- 391725- 391705 11014- 411748 881213- 881241 881293- 885244 885293- 885244 885293- 885524 885293- 885524 8855293- 085552- 0085529 0085	169 378 40 451 735 429 156 167 76 276 226	85 96 21 93 94 15 78 79 39 39 6	2,3 2,2 3,9 1,9 4,9 8 29 2 2 1,9 7 4,3	Inser, sea discriment / Inser, sea	48591 48725 71496- 71684 133897- 134630- 134630- 134630- 134630- 134630- 134630- 134630- 134630- 134630- 391257- 391707 411018- 411750 862867- 8628647 866898 868847 867103 1154731- -1155013 1155738- -1155013	136 189 378 40 451 736 429 166 167 76 276 33	57 85 96 21 93 94 15 78 79 39 39 39 6	2,3 2,2 3,9 1,9 4,9 8 29 2 2 2 1,9 7	Inear sea elgament / Inear sea	42183- 42311 66554- 129901 130257- 130256 338982- 338982- 353893 775930- 775930- 775930- 775930- 775930- 775930- 775930- 779937 780191- 933654- 933654- 933729 933654- 933729 933654- 1160874- 1150874- 1150874-	129 274 378 40 358 171 610 51 255 76 237 14	57 85 98 9 93 ? 15 7 79 39 39 29 ?	2 3 4 7 4 ? 34 ? 2 2 8 ? 2	ineer sea internet / internet / i	44835- 44899 69078- 69351 139925- 140523 4300512 440626 430253- 430612 444137 7998377- 799886 803335- 803490 803539- 803490 803539- 803597 989655- 9898573 989655- 989856 1305139- - 1305139- 1474673-	135 274 282 64 360 167 510 156 59 115 252 14	57 85 96 24 93 7 15 78 79 39 3 9 3 7	2,3 3,1 2,9 3,08 3,9 7 34,46 2 2 3,1 84 2 3,1	Inser sea learner / index.sea learner / index.sea learner / index.sea learner / inser.sea learner / inser.sea		2 2 2 2 2 3 3 2 2 3 3 2 2 3 3 3
17495- 77683 33896- 33896- 134688 91225- 991255- 991705 11014- 81213- 861641 85293- 86544 85293- 86544 85293- 86549- 0005583 00005583 00005583 520733- 1520758 878733-	169 378 40 451 735 429 156 167 76 276 226	85 98 21 93 94 15 78 79 39	2,3 2,2 3,9 1,9 4,9 8 29 2 2 2 1,9 7	Inear sea discriment / Inear sea discriment /	48591 48725 71496- 71684 133897- 134630- 134630- 134630- 134630- 134630- 134630- 134630- 134630- 134630- 391257- 391707 411018- 411750 862867- 8628647 866898 868847 867103 1154731- -1155013 1155738- -1155013	136 189 378 40 451 736 429 166 167 76 276 33	57 85 96 21 93 94 15 78 79 39 39	2,3 2,2 3,9 1,9 4,9 8 29 2 2 2 1,9 7	Ineer ses elgement / Ineer ses elgement /	42183- 42311 66564- 129601 130057- 130026 338862- 339839 353893 353893 353893 775930- 775930- 775930- 775930- 775930- 775930- 775937 780191- 780497 933654- 933654- 933654- 933654- 933654- 933654- 934897 1100861- -1160874	129 274 378 40 358 171 610 51 255 76 237 14	57 85 96 9 93 7 15 7 79 39	2 3 4 7 4 7 34 7 2 2 2 6	ineer seq intervent / intervent / intervent / intervent / intervent / intervent / intervent / intervent in	44835- 44969 69075- 69351 139926- 140207 140563- 140026 430253- 430612 443027- 430612 443027- 799886 803397- 903539- 803539- 803539- 803597 988559- 988673 988655- 988655- 1305139- -1305152	135 274 282 64 360 167 510 156 59 115 252 14	57 85 96 24 93 7 15 78 79 39 3	2.3 3.1 2.9 3.08 3.9 7 34,46 2 3,1 84	Inser sea ingrament / inser sea ingrament /		2 2 2 2 2 3 3 2 2 3 2 2 3 3 2 3
71405- 71683 33806- 134273 334629- 34629- 391705 11014- 861641 861213- 865449 865508- 865548 865549 0665508- 10865508- 10865508 10865508 10865508 10865508 10865750 520733- 1678638 1678733- 1678608	169 378 40 451 735 429 156 167 76 276 255 76	85 96 21 93 94 15 78 79 39 39 6	2,3 2,2 3,9 1,9 4,9 8 29 2 2 1,9 7 4,3	Inser, sea discreter / Inser, sea	46591- 46725 71496- 71894 13397- 19827 411018- 411018- 411018- 80287- 411018- 802847- 802847- 802847- 115573- 115573- 115573- 1155013 1055852 177013- 1155013	135 189 378 40 451 735 429 156 167 76 276 33 76	57 85 96 21 93 94 15 78 79 39 39 39 6	2,3 2,2 3,9 1,9 4,9 8 29 2 2 1,9 7 6	Inear sea elarment / Inear sea	42183- 42311 129624- 69656- 129624- 129624- 130207- 130226- 338382 353993 776339 776339 776339 77639 776439 776937 778045- 933654- 933654- 933654- 934897 1100874- 1100874 -1100874	129 274 378 40 358 171 610 51 265 76 237 14 43	57 85 98 9 93 ? 15 7 79 39 39 29 ?	2 3 4 7 4 ? 34 ? 2 2 8 ? 2	ineer sea internet / internet / i	44835- 44899 69078- 69351 139925- 140523 4300512 440626 430253- 430612 444137 7998377- 799886 803335- 803490 803539- 803490 803539- 803597 989655- 9898573 989655- 989856 1305139- - 1305139- 1474673-	135 274 282 64 360 167 510 156 59 115 252 14 43	57 85 98 24 93 7 15 78 79 39 3 9 3 7	2,3 3,1 2,9 3,08 3,9 7 34,46 2 2 3,1 84 2 3,1	Iner sea learner / indef, sea learner / indef, sea learner / indef, sea learner / iner sea not detected by Thef iner sea learner / iner sea		2 2 2 2 3 2 2 3 2 2 3 2 3 3 3 3
71425- 71683 33806- 134273 34629- 134273 34629- 134688 917255- 391705 391705 391705 391705 391705 391705 411748 865244 45523- 665508- 1005508 005508- 1005508 005508- 10055080	169 378 40 451 735 429 156 167 76 276 255 76	85 96 21 93 94 15 78 79 39 39 6 33	2,3 2,2 3,9 1,9 4,9 8 29 2 2 1,9 7 4,3 2,3	Inser, sea discreter, i Inser, sea	46531- 46725 71496- 71694 133897- 71694 134274 134679 391257- 411016- 411750 865295 411016- 665295 865295 865295 865295 1154731- 1154005 1155738- 11560325 115635 115655 115655 115655 1156555 1156555 1156555 1156555 1156555 11565555 11565555 11565555 115655555555	135 189 378 40 451 735 429 156 167 76 276 33 76	57 85 96 21 93 94 15 78 79 39 39 6 33	2,3 2,2 3,9 1,9 4,9 8 29 2 2 2 2 1,9 7 6 2,3	Inear sea elarment / Inear sea	42183- 42311 66564- 06837- 129604 130257- 33082- 338839 353823- 353833 353839 353823- 353833 353839 353823- 353833 353839 353823- 3538	129 274 378 40 358 171 610 51 265 76 237 14 43	57 85 96 9 33 7 15 7 79 39 39 2 7	2 3 4 7 4 ? 3 4 ? 3 4 ? 2 2 8 ? ? 2 ? ? ? ?	inear sea internet / internet / i	44835- 44989 69078- 69078- 69078- 140027 140583- 440327- 4443371- 4443371- 4443371- 4443371- 4443371- 803359- 803359- 989856 989856 1305139- 1305132 1474873- 11075152 2712440- 2712440	135 274 282 64 360 167 510 156 59 115 252 14 43	57 85 96 24 93 7 15 78 79 39 39 3 2 7 7 7	2.3 3.1 2.9 3.08 3.9 7 34,46 2 2 3.1 84 7 7 7	Inser sea insert sea insert of insert sea insert of inser sea insert sea inser sea ins		2 2 2 2 2 3 2 2 3 2 2 3 3 2 3 3 3 2 2 3 3 2
71495- 71683 38886- 34273 34628- 91255- 91255- 917055 917055 917055 917055 917055 917055 917055 917055 917055 915008 915008 915008 915008 9157084- 915708- 915700	189 378 40 451 735 429 156 157 76 276 255 76 284	85 96 21 93 94 15 78 79 39 39 6 33	2,3 2,2 3,9 1,9 4,9 8 29 2 2 1,9 7 4,3 2,3	Instruction Instru	46531- 46725 71496- 71684 13397- 71684 134274 134274 134274 134274 141150 682087- 682087- 682087- 682087- 867103 867103 867103 1155728- 11	136 189 378 40 451 735 429 156 167 76 276 333 76 284	57 85 96 21 93 94 15 78 79 39 39 6 33	2,3 2,2 3,9 1,9 4,9 8 29 2 2 2 2 1,9 7 6 2,3	Inear sea elarment / Inear sea	42183- 42311 19654- 06554- 199001 130257- 130282 338823- 338839 35823- 353893 35823- 353893 35823- 353893 35823- 353893 35823- 353893 776803- 776803- 77690- 77600-	129 274 378 40 358 171 610 51 255 76 237 14 43 102	57 85 96 9 33 7 15 7 79 39 39 2 7	2 3 4 7 4 ? 3 4 ? 3 4 ? 2 2 8 ? ? 2 ? ? ? ?	ineer sea internet / internet / i	44835- 44980 69078- 09351 140267 140267 1402683- 140207 1402683- 140267 1402683- 140267 1402683- 140267 1402683- 140268 03359- 08359- 998866 03359- 998859 03359- 998859 1305139- 1305132 1474673- 147471 1305132 2112540 2211550 2211540 2211550 2211540 2211550 2211540 2211550 2211540 2211550 2211550 2211540 2211550 2211540 2211550 2211540 2211550 2211540 2211550 2211540 2211550 221550 22115	135 274 282 64 360 167 510 155 59 115 252 14 43 375	57 85 96 24 93 7 15 78 79 39 39 3 2 7 7 7	2.3 3.1 2.9 3.08 3.9 7 34,46 2 2 3.1 84 7 7 7	Inser, sea Barment / Inser, sea		2 2 2 2 2 3 2 2 3 2 2 3 3 2 3 3 3 2 2 3 3 2
71495- 71683 133896- 134273 134629- 134668 391255-	1899 378 40 451 735 429 156 157 76 276 26 76 284 167	85 96 21 93 94 15 78 79 39 39 6 33 91	2.3 2,2 3,9 1,9 4,9 8 29 2 2 2 2 1,9 7 4,3 7 4,3 2,3 3,1	Inter, sea discreter, i Inter, sea	46291- 46725 71496- 71496- 71496- 71496- 71496- 71496- 71496- 71496- 71496- 71496- 71496- 71496- 800147- 800147- 800047- 800047- 800047- 800047- 800047- 1155738- - 1156033 1155738- - 1156039 1155738- - 1156039 1155738- - 1156039 1155738- - 1156039 1155738- - 1156039 1155738- - 1156039 1155738- - 1156039 1155738- - 1157019 1155738- - 1157019 1155738- - 1157019 1155738- - 1157019 1155738- - 1157019 1155738- - 1157019 1155738- - 1157019 1155738- - 1157019 1155738- - 1157019 1155738- - 1157019 1155738- - 1157019 1155738- - 1157019 1155738- - 1157019 1157739 1157731- 1157019 1155738- - 1157019 1157731- 1157019 1157731- 1157019 1157731- 1157019 1157731- 1157019 1157731- 1157019 1157738- 1157739 11577739 11577739 11577739 115777575 1157775 1157775 1157775 1157775	135 189 378 40 451 735 429 156 167 76 276 33 76 284 155	57 85 96 21 93 94 15 78 79 39 39 6 33 39 6	2,3 2,2 3,9 1,9 4,9 8 29 2 2 2 2 1,9 7 6 2,3 3,1	Inear sea electment / Inter sea	42183- 42311 10654- 129524- 129624- 129601 130257- 130285 338839- 339839 353823- 339839 353823- 339839 353823- 339839 353823- 339839 353823- 776439 776439 776837 77097- 776439 776837 77097- 776439 776837 77097- 776837 77097- 776837 77097- 776837 77097- 776837 77097- 776837 77097- 776837 77097- 776837 77097- 776837 77097- 776837 77097- 776837 77097- 776837 77097- 776837 77097- 776837 77097- 776837 77097- 776837 77097- 77097- 72097- 110087- 110	129 274 378 40 358 171 510 51 255 76 237 14 43 102 155	57 85 96 9 33 7 15 7 39 39 39 7 7 7 7 7	2 3 4 7 4 7 34 7 2 2 2 6 7 7 7 7	ineer sea internet / internet / internet / ineer sea isprment / ineer sea isprment / inter sea isprment / inter sea of one sea of one sea isprment / inter sea of one	44835- 44989 69078- 69078- 69078- 140027 140583- 440327- 4443371- 4443371- 4443371- 4443371- 4443371- 803359- 803359- 989856 989856 1305139- 1305132 1474873- 11075152 2712440- 2712440	135 274 282 64 360 167 510 155 59 115 252 14 43 375	57 85 96 24 93 7 15 78 79 39 3 7 9 7 91	2.3 3.1 2.9 3.08 3.9 7 34,46 2 2 3.1 84 7 7 7 4.1	Inser sea ingreent / inser sea		2 2 2 2 3 2 2 2 3 2 2 3 3 3 3 3 3 3 3 3

Figure 4

Example of a query in the Strain Comparison Page for more than two strains. Top, the query page shows the 6 comparisons currently available (others will be added as new genome sequences are finished and released). Bottom, the result of a query performed for Escherichia coli strains O157:H7 Sakaï, O157:H7 EDL933, K12 and UPEC-CFT073 is summarized. In several loci, the size of the repeat is listed differently for the different strains, which is due to different detections by the Tandem Repeats Finder, usually as a result of internal variations within the tandem array. Total length is calculated from positions of matching flanking sequences in the different strains, and does not necessarily correspond to the length of the tandem repeat detected by TRF in the locus. "Number of alleles" refers to the number of predicted sizes differing by at least 5 bp among the strains compared.

-2579737 38

oi grmert

2852483

2

19

elignment i

19 2 2

Polymorphic tandem repeats (VNTRs, for Variable Number of Tandem Repeats), once identified, provide convenient tools requiring ordinary molecular biology equipment and the data can be easily exchanged and compared. The resulting assay, called MLVA (for multiple locus VNTR analysis) can even be automated [16]. We have developed tools which facilitate the bioinformatics step of genome analysis required to start a project. A previously described Tandem Repeats Database enables the identification of tandem repeats across entire genomes [9,10,17-19]. It has been constantly updated, with now more than a hundred bacterial genomes available, compared to 35 at the onset of the database. We present here a new and major development of this resource which takes advantage of the fact that more than two different strains from the same species have now been sequenced at least for a number of major human pathogens. As a result, the tools accessible over the Internet at http://minisatel lites.u-psud.fr now comprise four complementary parts. The newly added resource, the Strain Comparison Page, takes advantage of the availability of genome sequences from more than one strain from a growing number of species to directly identify tandem repeats differing between the sequenced strains. This is of interest because the vast majority of tandem repeats is often not polymorphic [19]. The "Blast in the Tandem Repeats Database" page facilitates the search for a known tandem repeat, the prediction of PCR amplification products size, and the verification of primer specificity. Once an MLVA assay has been set up, and carefully validated by typing collections of isolates, it is relatively easy to construct databases of genotypes to be used locally or which can be queried across the Internet. The "Bacterial Genotyping Page" illustrates a freely accessible, fast and easy to use internet-based service for strain comparisons, in which a user can compare a genotype produced for one of his isolates to the existing data.

Construction and content The Tandem Repeats Database main page

Tandem repeats were identified from finished microbial genome sequences (as listed by the Genome OnLine Database [20]) using the tandem repeats finder (TRF) software [21,22] with the following options: alignment parameters, "2,3,5" (these parameters are the less stringent ones), minimum alignment score to report repeat, 50 (this score allows to detect short structures), maximum period size, 500 base-pairs. When the program reported redundant (overlapping) repeats, the redundancy was eliminated as described in [23], before import in the database. The database uses Microsoft Access 2000 and the querying process uses Active Server Pages (ASP, Microsoft) with Perlscripts or VBscripts. Perl was obtained from the ActiveState Programmer Network [24]. The database is hosted on a server running under Windows 2000 server

(Microsoft). The tandem repeats database main page is described in more detail in [9].

The Strain Comparison page

Sequence comparisons used BLAST [25]. The BLAST software was obtained from the NCBI FTP site [26]. The flanking sequences of TRs from one strain were compared to the whole sequence of the other strain (and reciprocally, to avoid missing some tandem repeats that would not appear in the tandem repeats database for one strain because they were not detected by the Tandem Repeats Finder [21] -for instance because there is only one copy of the repeated unit in the considered strain). The resulting list of matching tandem repeats was then imported in the database, where it can be queried. The comparison of more than two strains was made possible through a supplemental step before import in the database: the synthesis of several 2-strains comparisons, of the same "reference" strain against each of the others (matching between TRs of the different strains was deduced from the positions on the reference strain).

The Blast page

The Blast Page allows users to run BLAST [25] in the tandem repeats and flanking sequences from the database via Perlscripts. The Blast outputs are linked to the database, in order to easily obtain the description of identified tandem repeats.

The Bacterial Genotyping page

The web-page site performing identifications was developed using the BNserver application (version 3.0, Applied-Maths, Belgium) and ASP (Microsoft) using Perlscript. The typing results (gel images and resulting data) were managed using the Bionumerics software package as described in [10]. The output of a query is a list of strains and genotypes from the database together with similarity scores.

Utility

The procedure to find polymorphic tandem repeats (TRs) for use in strain typing

Figure 1 shows the steps leading from a genome sequence to the exploitation of polymorphic tandem repeats for bacterial strain genotyping. Although Tandem Repeats are easily identified using the Tandem Repeats Database, TR polymorphism must be evaluated by typing across a set of relevant strains. If the sequences of several strains of the species of interest are available, the Strain Comparison Page can be used to directly identify tandem repeats predicted to be polymorphic in size between the two (or more) sequenced strains. However, it is important to keep in mind that the tandem repeats predicted as being polymorphic will depend on the sequenced strains and wellplanned surveys of isolates will still be necessary. The available tools do not replace this validation step, as the value of each marker must be carefully established on an appropriate set of isolates. The definition of an appropriate set of isolates depends upon the question which is being addressed, i.e. large scale or local epidemiology. The Blast Page has been implemented in the tandem repeats database in order to easily determine the size of the expected PCR amplification products. The database is also manually updated to contain PCR conditions as well as polymorphism index, and links to the original reports [27] (input from users is welcome). Eventually, when an MLVA assay has been fully developed and validated, typing data can be made accessible so that individual queries can be run. The Bacterial Genotyping Page illustrates how this could work. The genotyping data for a strain can be entered and submitted via this page. The output is the description of the closest strains. The data which has been submitted is not incorporated in the database itself, since this would require stringent data validation steps. In the following sections, we are presenting the web-based resources associated with this procedure.

The "Strain Comparison" pages

The strain comparison pages are available via [28]. The comparison of two strains is based on a pre-computed BLAST [25] analysis of the flanking sequences of tandem repeats from one strain against the other, and vice-versa. Figure 2 summarizes the results of this first step for 23 comparisons. Three indexes are scored (see figure legend): (1) the "mean %identity" between the flanking sequences is a measure of single nucleotide polymorphism (SNPs) frequency (not insertions-deletions), (2) the proportion (%) of flanking sequences that matched the flanking sequence of its homologue in the other strain on more than half of the 500 bp assayed here -i.e. that were not rearranged, by insertion of mobile elements for instance -, (3) the proportion (%) of tandem repeats that were found to be of a different length between the two strains being compared. In addition, the positions of matching tandem repeats in the two genomes is plotted to reveal large-scale genome rearrangements. A number of situations are observed: for instance Yersinia pestis orientalis strain CO-92 [29], and medievalis strain KIM5 P12 [30] show a very high "mean %identity" (99.96 %), in agreement with the recent emergence of Yersinia pestis [4]. In spite of this, the two strains differ by a high number of large rearrangements (as seen on the plot), which reflects the high genome plasticity observed in this species [31], together with a relatively high rate of polymorphic tandem repeats (8.47%). In contrast, Listeria monocytogenes strain EGD-e and Listeria innocua strain Clip 11262 have a lower homology (90.19%) and only 3.99% of polymorphic tandem repeats in spite of the evolutionary distance (see Figure 2).

The strain comparison page allows queries in the tandem repeats database according to the tandem repeat length difference between the two strains compared, and also to other tandem repeats characteristics (unit length, copy number, etc...). Figure 3 illustrates a query done for *Mycobacterium tuberculosis* strains H37Rv and CDC1551 [32]: the query "length difference \geq 5 bp" identifies 58 tandem repeats (8 are shown on Figure 3). This prediction has been tested for the 30 loci amenable to PCR analysis and polymorphism has been confirmed in all cases [10].

When more than two strains have been sequenced, a synthesis of the results of several 2-strains comparisons is also available. Figure 4 illustrates a query made for *Escherichia coli* strains O157:H7 Sakaï, O157:H7 EDL933, K12, and UPEC-CFT073 [33-35]: 87 tandem repeats were found with 2 to 4 alleles among the 4 strains (18 of which are listed in Figure 4).

The "Blast in the Tandem Repeats Database" page

To facilitate the identification of already studied tandem repeats, we implemented BLAST [25] against the tandem repeats from the database, *i.e.* the tandem repeats themselves and their flanking sequences. The Blast page is available at [36]. All bacteria can be queried at once, which allows the identification of tandem repeats families, conserved in several bacterial species. Another page is dedicated to the Blast of PCR primers and provides the size of the PCR products in all the species/strains where the primers match. Figure 5 shows the results of searching the PCR primer pair from tandem repeat H37Rv_0024_18 bp [10] in all bacteria: as expected, the PCR primer pair matches *Mycobacterium tuberculosis* strains H37Rv and CDC1551, providing different PCR product lengths.

The Bacterial Genotyping page

The Bacterial Genotyping page [37] provides one illustration on how tandem repeat typing data can be made available via internet to allow external users to query genotyping data (Bacillus anthracis, Yersinia pestis, Mycobacterium tuberculosis, Pseudomonas aeruginosa for the moment) and compare a new strain to existing data as previously described in [10]. For each locus, allele sizes can be selected among a list of possibilities (observed sizes). The results of the query indicate a similarity score and include links to the complete data recorded for each strain listed. This page is just meant as an illustration and prototype. MLVA reference data could also be made available for downloading as tabular data files, or can be copied from published datasets, which can then be complemented by in-house data, and analyzed by the appropriate clustering software.

	G Whole sequences C Only Tandem Repeat	CGACGACCGCTA	¥.
ast in:	Whole sequences Only Tandem Repeat	s (and flanking	
ect a g	enome to blast:		
rchaea:	Aeropyrum pernix K1 Archaeoglobus fulgidus DSM4304 Halobacterium sp RC-1 Methanobacterium thermoautotrophicum delH Methanococcus jannaschii DSM2661 Methanopyrus kandleri AV19 Methanosarcina mazei Go1 (DSMZ 3647)	Bacteria:	Vibrio cholerae El Tor N16961 Wigglesworthia glossinidia brevipalpis Xanthomonas axonopodis pv. citri 306 Xanthomonas campestris ATCC33913 Xylella fastidiosa 9a5c Yersinia pestis CO-92 Yersinia pestis KIM5 P12 All bacteria
	Arabidopsis thaliana chromosome 4 Caenorhabditis elegans chromosome 1 Human chromosome 20 Human chromosome 21	Viruses:	Bovine adenovirus B Bovine adenovirus D Canine adenovirus type 1 Duck adenovirus 1 Fowl adenovirus A Fowl adenovirus D

BLAST of PCR primers in: bacteria

BLASTN 2.2.1 [Apr-13-2001]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

List of primer pairs matches:

```
    Match in strain MT_H37Rv => PCR product: 328 bp
left primer matches at pos 24598: +/+ (exact match)
right primer matches at pos 24925: +/- (exact match)
Search for corresponding tandem repeat
    Match in strain MT_CDC1551 => PCR product: 310 bp
left primer matches at pos 24598: +/+ (exact match)
```

```
left primer matches at pos 24598: +/+ (exact match)
right primer matches at pos 24907: +/- (exact match)
Search for corresponding tandem repeat
```

Figure 5

Example of a query in the "Blast of PCR primers" page, providing the length of the PCR products in the strains/species where the primer pair matches, and links to the corresponding tandem repeats descriptions.

Discussion

Bacterial genomes evolution

As shown by the indexes from Figure 2, there are different ways to represent the divergence/similarity between two strains. They are not correlated, suggesting independent evolution processes. First, the "mean %identity" between two genomes reflects point mutations, and is an indicator of the time passed since the two strains diverged. For instance, *Yersinia pestis* is known to be of recent emergence [4] and shows a high "mean %identity" between strains CO-92 (orientalis) and KIM5 P12 (medievalis). In contrast, and as shown by the dot plot, large genome rearrangements occurred in this genome, which is representative of a high genome plasticity [31]. The index "% of flanking sequences not rearranged" is an indicator of small-scale genome rearrangements, such as the inser-

tions of mobile elements. This index is low for genomes rich in mobile elements, like Streptococcus agalactiae, in which such elements significantly contribute to strain diversity [38]. Finally, the index "% of polymorphic tandem repeats" between two strains represents the tandem repeats evolution rate. For the moment, the mechanisms of bacterial VNTRs mutations have not been precisely investigated, but it seems likely to be independent of the other processes mentioned, as there are no correlations between the indexes. Figure 2 provides clues to assess which typing method(s) will be efficient in the different species. For instance, the two bacterial species Salmonella typhimurium strain LT2 [39] and Shigella flexneri strain 2a301 [40] share only 86.06% of sequence identity, clearly making the identification of matching tandem repeats between the two species difficult and of low significance. MLVA analysis appears to be of highest interest for the subspecies typing of highly monomorphic species including Yersinia pestis, Bacillus anthracis, Mycobacterium tuberculosis and Brucella [9,10,41].

Strain comparison efficiency

The sequencing of more than one strain for some bacterial species allows direct identification of polymorphic tandem repeats, assuming that no sequencing errors occurred. Earlier investigations provide good reasons to believe that tandem repeats in the size range considered here (a few hundred base-pairs) are correctly sequenced, and consequently, that the strain comparison data is reliable. As a negative control, the comparison of two independent sequences from the same strain of Agrobacterium tumefaciens strain (C58), one from Cereon genomics [42] and the other from Washington University [43], shows that no length polymorphism is detected among tandem repeats (Figure 2) between the two independent sequences. As a positive control, the tandem repeats predicted to be polymorphic by genome sequence comparison between the two strains of M. tuberculosis have indeed been proved polymorphic by PCR typing of isolates [10].

Selection based on comparison of sequence data from two strains will miss some polymorphic loci. Indeed, the results provided by the approach rely upon the phylogenetic distance between the two strains being compared. If the strains are very closely related, only a few TRs will be found different between them, but these tandem repeats will probably be the most polymorphic ones. Conversely, if the strains are distant in the phylogenetic tree, a larger number of polymorphic TRs will be found, some of them will be only moderately polymorphic. Obviously, when a few well-selected strains have been sequenced, it is likely that very few polymorphic tandem repeats are undetected in the Strain Comparison pages. It is of course still going to be very important to determine the TR allele frequency for isolates carefully selected to be representative of the global diversity of a given pathogen before suggesting the configuration of an MLVA assay to use in subsequent studies. In addition, those TR markers that are highly polymorphic in diverse test panels of isolates may be monomorphic when applied to isolates responsible for local outbreaks. The configuration of TR markers used to make up an assay needs to be determined empirically with representative local isolates and tailored to the study population and study questions.

Polymorphic tandem repeats selection for species with only one sequenced strain

The identification of simple criteria able to predict tandem repeat polymorphism when genome sequence data is available for only one strain would indeed greatly facilitate the development of MLVA assays. It would seem reasonable for instance to expect that the number of copies and the internal homogeneity of tandem arrays are strong predictors [23]. We take advantage here of the many strain comparisons which are made available via the strain comparison pages to evaluate such criteria.

We have analyzed bacteria with at least three sequenced genomes (Staphylococcus aureus: 6 strains, Escherichia coli: 4 strains, Streptococcus pyogenes: 4 strains and Salmonella typhi and typhimurium: 3 strains). We assume that in such cases, only a few polymorphic tandem repeats are missed in the comparisons. We compared the distribution of tandem repeats sequence characteristics among the group of "polymorphic" loci (differing in at least two of the strains compared, excluding length differences between strains that resulted from microdeletions in the flanking sequences) and the others. Comparisons were performed for the following sequence characteristics: unit length, copy number, total length, %GC, GC bias (=|%G-%C|/ (%G+%C)), %matches, and HistoryR (a score derived from tandem repeat history reconstruction algorithm [44] as described in [23]). None of the variables were normally distributed, as tested with Kolmogorov-Smirnov test, so a non-parametric Wilcoxon test was used to compare the distributions, which were judged significantly different at the .05 level of the statistic (2 tailed). Distributions were significantly different for all 4 species studied for %matches, total length and copy number. As shown on Figure 6, polymorphic TRs have a higher internal conservation and total length than monomorphic ones. Copy number, which is correlated with total length, is also higher among polymorphic TRs.

Selecting the longest and most conserved tandem repeats should thus improve polymorphic TRs identification. Table 1 illustrates the query "total length ≥ 80 bp and %matches $\geq 80\%$ " applied to the four species used to find

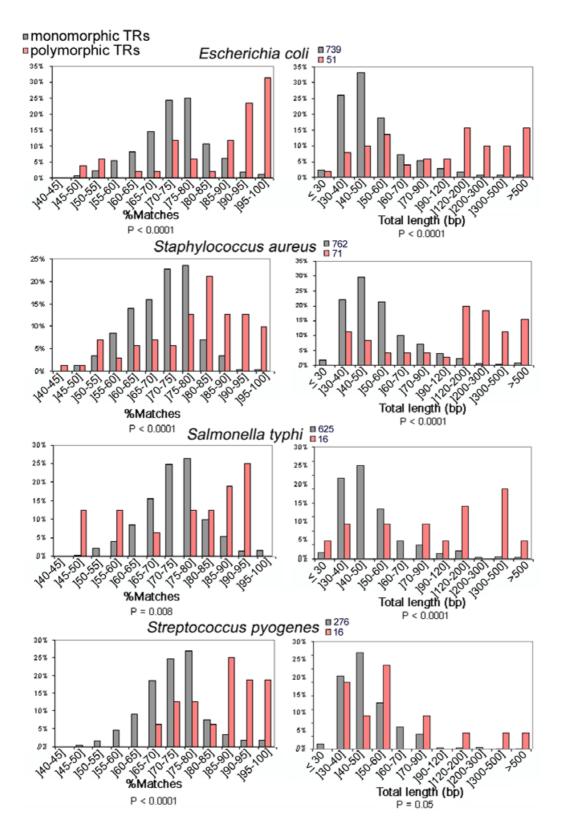


Figure 6

Proportion of predicted polymorphic (pink) and monomorphic (grey) tandem repeats according to different parameters (internal homogeneity of the repeat array (%matches) or total length). P-values obtained for the non-parametric Wilcoxon tests appear below each histogram.

Table 1: Use of the criterion "total length \ge 80 bp and %matches \ge 80%" on 4 species for which 3 strains or more were compared. The number of monomorphic, polymorphic (2 alleles or more) and highly polymorphic (3 alleles or more) TRs in whole set, and positive and negative groups are listed. (a) "criterion" refers to the selection of TRs with L \ge 80 bp and %M \ge 80%

Comparison (total number of TRs)		e set (propo otal numbe		Tandem repeats with L≥80 bp AND %M≥80% (proportion among the set)			Tandem repeats with L<80 bp OR %M<80% (proportion among the set)			% of the polymorphic TRs (2 alleles or more) that were detected by criterion ^a	% of the TRs with 3 alleles or more that were detected by criterion ^a	% of all TRs that fulfil the criterion ^a
	l allele	2 alleles or more	3 alleles or more	l allele	2 alleles	3 alleles or more	l allele	2 alleles or more	3 alleles or more			
S aureus (833 TRs)	762 (91.5%)	71 (8.5%)	38 (4.5%)	5 (13%)	8 (20%)	25 (66%)	757 (95%)	25 (3.5%)	13 (1.5%)	46%	66%	7.23%
E coli (790 TRs)	739 (93.5%)	51 (6.5%)	12 (1.5 %)	12 (38%)	I 3 (40%)	7 (22%)	727 (96%)	26 (3.5%)	5 (0.5%)	39%	58%	4.86%
S typhi / typhimurium (641 TRs)	625 (97.5%)	16 (2.5%)	2 (0.3%)	13 (68%)	4 (22%)	2 (10%)	612 (98%)	10 (2%)	0 (0%)	37.5%	100%	3.27%
S pyogenes (292 TRs)	276 (94.5%)	16 (5.5%)	3 (1%)	4 (67%)	0 (0%)	2 (33%)	272 (95%)	14 (4.7%)	l (0.3%)	12.5%	67%	2.71%

Table 2: Use of the criterion "total length \ge 80 bp and %matches \ge 80%" on 10 species for which 2 strains were compared. The numbers of tandem repeats with equal lengths and different lengths between the two strains in the whole set, and positive and negative groups are listed.

Comparison (total number of TRs loci)	Whole set (proportion)		Criterion + (L≥80 bp, %M≥80%)		Criterion -		Sensitivity (% of the TRs with different lengths that were detected by criterion)	Specificity (% of the TRs predicted by the criterion that have different length)	% of all TRs that fulfil the criterion
	equal length	differen t length	equal length	differen t length	equal length	differen t length			
H pylori 26695/J99 (624 TRs)	506 (81%)	8 (19%)	0	11	506	107	9%	100%	2%
N meningitidis MC58/Z2491 (642 TRs)	528 (82%)	4 (18%)	10	23	518	91	20%	70%	5%
M tuberculosis H37Rv/ CDC1551(1502 TRs)	1441 (96%)	61 (4%)	35	27	1406	34	44%	44%	4%
L monocytogenes EGD-e/L innocua Clip I 1262 (576 TRs)	553 (96%)	23 (4%)	2	3	551	20	13%	60%	1%
S agalactiae NEM316/2603 (398 TRs)	387 (97%)	(3%)	2	I	385	10	9%	33%	1%
S pneumoniae TIGR4/R6 (406 TRs)	339 (83%)	67 (17%)	14	29	325	38	43%	67%	10%
Y pestis CO-92/KIM5 P12 (1499 TRs)	1372 (92%)	127 (8%)	44	19	1328	108	15%	30%	4%
R prowazekii Madrid E/R conorii malish 7 (316 TRs)	290 (92%)	26 (8%)	0	2	290	24	8%	100%	1%
Brucella suis 1330/ Brucella melitensis 16 M (739 TRs)	681 (92%)	58 (8%)	2	4	679	54	7%	67%	1%
X fastidiosa 9a5c/grape Temecula1 (573 TRs)	440 (77%)	133 (23%)	2	28	438	105	21%	93%	5%

predictive criteria. For all four species, the group fulfilling the criterion is, as expected, enriched in polymorphic (at least two alleles) tandem repeats: in *Staphyloccocus aureus*, polymorphic tandem repeats represent only 8.5% of the whole population of tandem repeat loci but are predominant (87%) in the criterion positive group. The enrichment is even greater for highly polymorphic TRs, *i.e.* with 3 alleles or more: for example from 4.5% in the whole set to 66% in the positive group for *Staphylococcus aureus*. However this simple criterion misses more than half of the polymorphic loci. In addition, the efficiency of the criterion is highly variable in the different species: it is relatively satisfying in *Staphylococcus aureus* (54% of polymorphic tandem repeats would be missed) but very inefficient in *Streptococcus pyogenes* (almost 90% are missed). The results for highly polymorphic loci (3 alleles or more) are more consistent (the proportion of TRs with 3 alleles or more detected by the criterion ranges from 58% for *Escherichia coli* to 100% for *Salmonella*).

It is tempting to speculate that these observations are applicable to other species. Subsequently, we applied the criterion to ten of the 2-strains comparisons available on the Strain Comparison Page (Table 2). In all ten instances, the criterion positive group is enriched in TRs with different lengths between the two strains, compared to the whole set. This proportion varies from less than 3% in *Streptococcus agalactiae* to more than 20% in *Xylella fastidiosa* in the whole set. It is increased to 33% and 93% respectively among the set of loci which satisfy the criterion (these percentages correspond to the predictor's specificity), but the vast majority of polymorphic loci will be missed (90% and 80% respectively). Sensitivity, that is % of the TRs with different lengths that were detected by criterion varies from 6.90% for *Brucella* to 44.26% for *Mycobacterium tuberculosis*.

The finding that polymorphic tandem repeats have, on average, a higher internal conservation, total length, and copy number than monomorphic ones is in agreement with previous observations that TR polymorphism is correlated with conservation in Yersinia pestis and with total length in Bacillus anthracis [9]. It is also reminiscent of the behavior of microsatellites (also called short sequence repeats: SSR, see [45] for review), which are stabilized by internal variations [46] and by reduction of the number of repeats [47]. Unfortunately, we show here that such simple prediction criteria may miss a very large proportion of polymorphic tandem repeats, and provide highly variable results in different species. This indicates that, in the absence of sequence data from two strains or more, the systematic testing of tandem repeats polymorphism across a set of relevant strains remains the most appropriate way to develop an MLVA assay. Consequently, the Strain Comparison page is of great use when two strains or more have been sequenced.

Conclusions

Bacterial strain typing at the subspecies level is essential for epidemiological issues in the context of disease control. This can be used to determine if an S. aureus or P. aeruginosa infection for instance has been acquired in an hospital environment or not. On a larger scale, it can be used to trace the emergence of new, more virulent or drug resistant M. tuberculosis strains. It is also of interest in the field of bioterrorism and bioweapons control, as was shown by the investigations following the 2001 B. anthracis attacks. Tandem repeats typing has recently emerged as one way to address this issue. Indeed, in the case of a number of highly monomorphic bacterial species, including B. anthracis and Y. pestis, tandem repeats typing is the method of choice for subspecies typing. In addition to the fact that these loci represent an important fraction of the existing polymorphism, it offers a number of practical advantages, including the ease of typing, and of data exchanges among different countries. It is hoped that the tools which are described here will help evaluate the potential of tandem repeats typing assays for a larger range of pathogens.

Availability

All the tools presented are freely available from <u>http://</u><u>minisatellites.u-psud.fr</u>.

List of abbreviations used

ASP: active server pages

MLVA: multiple locus VNTR analysis

PCR: polymerase chain reaction

TR: tandem repeat

TRF: tandem repeats finder

Authors contributions

FD is the developer of the database and web site, and the curator of the database. GV participated in the development of the initial procedure for the tandem repeat size comparisons between two genomes. The two authors contributed equally to the writing.

Acknowledgments

This work was funded by grants from Délégation Générale de l'Armement (DGA, France) aimed at facilitating the typing of dangerous pathogens.

References

- 1. van Belkum A: High-throughput epidemiologic typing in clinical microbiology. *Clin Microbiol Infect* 2003, **9:**86-100.
- Enright MC, Spratt BG: Multilocus sequence typing. Trends Microbiol 1999, 7:482-7.
- Gutacker MM, Smoot JC, Migliaccio CA, Ricklefs SM, Hua S, Cousins DV, Graviss EA, Shashkina E, Kreiswirth BN, Musser JM: Genomewide analysis of synonymous single nucleotide polymorphisms in Mycobacterium tuberculosis complex organisms: resolution of genetic relationships among closely related microbial strains. Genetics 2002, 162:1533-43.
- Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, Carniel E: Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. Proc Natl Acad Sci U S A 1999, 96:14043-8.
- van Belkum A, Scherer S, van Leeuwen W, Willemse D, van Alphen L, Verbrugh H: Variable number of tandem repeats in clinical strains of Haemophilus influenzae. Infect Immun 1997, 65:5017-27.
- Frothingham R, Meeker-O'Connell WA: Genetic diversity in the Mycobacterium tuberculosis complex based on variable numbers of tandem DNA repeats. Microbiology 1998, 144:1189-1196.
- 7. Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C: Variable human minisatellite-like regions in the Mycobacterium tuberculosis genome. *Mol Microbiol* 2000, **36**:762-771.
- Adair DM, Worsham PL, Hill KK, Klevytska AM, Jackson PJ, Friedlander AM, Keim P: Diversity in a variable-number tandem repeat from Yersinia pestis. J Clin Microbiol 2000, 38:1516-9.
- Le Flèche P, Hauck Y, Onteniente L, Prieur A, Denoeud F, Ramisse V, Sylvestre P, Benson G, Ramisse F, Vergnaud G: A tandem repeats database for bacterial genomes: application to the genotyping of Yersinia pestis and Bacillus anthracis. BMC Microbiol 2001, 1:2.
- Le Flèche P, Fabre M, Denoeud F, Koeck JL, Vergnaud G: High resolution, on-line identification of strains from the Mycobacterium tuberculosis complex based on tandem repeat typing. BMC Microbiol 2002, 2:37.
- 11. Spanier JG, Jones SJ, Cleary P: Small DNA deletions creating avirulence in Streptococcus pyogenes. Science 1984, 225:935-8.
- 12. Hollingshead SK, Fischetti VA, Scott JR: Size variation in group A streptococcal M protein is generated by homologous recom-

bination between intragenic repeats. Mol Gen Genet 1987, 207:196-203.

- Keim P, Kalif A, Schupp J, Hill K, Travis SE, Richmond K, Adair DM, Hugh-Jones M, Kuske CR, Jackson P: Molecular evolution and diversity in Bacillus anthracis as detected by amplified fragment length polymorphism markers. J Bacteriol 1997, 179:818-24.
- Keim P, Price LB, Klevytska AM, Smith KL, Schupp JM, Okinaka R, Jackson PJ, Hugh-Jones ME: Multiple-Locus Variable-Number Tandem Repeat Analysis Reveals Genetic Relationships within Bacillus anthracis. J Bacteriol 2000, 182:2928-2936.
- Sylvestre P, Couture-Tosi E, Mock M: Polymorphism in the collagen-like region of the Bacillus anthracis BcIA protein leads to variation in exosporium filament length. J Bacteriol 2003, 185:1555-63.
- Supply P, Lesjean S, Savine E, Kremer K, van Soolingen D, Locht C: Automated high-throughput genotyping for study of global epidemiology of Mycobacterium tuberculosis based on mycobacterial interspersed repetitive units. J Clin Microbiol 2001, 39:3563-3571.
- 17. Vergnaud G, Denoeud F: Minisatellites: Mutability and Genome Architecture. Genome Res 2000, 10:899-907.
- Pourcel C, Vidgop Y, Ramisse F, Vergnaud G, Tram C: Characterization of a Tandem Repeat Polymorphism in Legionella pneumophila and Its Use for Genotyping. J Clin Microbiol 2003, 41:1819-1826.
- Onteniente L, Brisse S, Tassios PT, Vergnaud G: Evaluation of the polymorphisms associated with tandem repeats for Pseudomonas aeruginosa strain typing. J Clin Microbiol 2003, 41:4991-7.
- 20. GOLD Genomes OnLine Database [http://ergo.integratedge nomics.com/GOLD/]
- 21. Benson G: Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 1999, 27:573-580.
- 22. The Tandem Repeats Finder [http://tandem.bu.edu/trf/trf.html]
- Denoeud F, Vergnaud G, Benson G: Predicting Human Minisatellite Polymorphism. Genome Res 2003, 13:856-867.
- 24. The ActiveState Programmer Network (ASPN) ActivePerl download page [http://www.activestate.com/products/ActivePerl]
- download page [http://www.activestate.com/products/ActivePerl]
 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997, 25:3389-402.
- 26. The NCBI BLAST ftp site [ftp://ftp.ncbi.nih.gov/blast/]
- 27. The tandem repeats database [http://minisatellites.u-psud.fr]
- 28. The Strain Comparison Page [http://minisatellites.u-psud.fr/ comparison]
- Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, Sebaihia M, James KD, Churcher C, Mungall KL, Baker S, Basham D, Bentley SD, Brooks K, Cerdeno-Tarraga AM, Chillingworth T, Cronin A, Davies RM, Davis P, Dougan G, Feltwell T, Hamlin N, Holroyd S, Jagels K, Karlyshev AV, Leather S, Moule S, Oyston PC, Quail M, Rutherford K, Simmonds M, Skelton J, Stevens K, Whitehead S, Barrell BG: Genome sequence of Yersinia pestis, the causative agent of plague. Nature 2001, 413:523-7.
- Deng W, Burland V, Plunkett G 3rd, Boutin A, Mayhew GF, Liss P, Perna NT, Rose DJ, Mau B, Zhou S, Schwartz DC, Fetherston JD, Lindler LE, Brubaker RR, Plano GV, Straley SC, McDonough KA, Nilles ML, Matson JS, Blattner FR, Perry RD: Genome sequence of Yersinia pestis KIM. J Bacteriol 2002, 184:4601-11.
- 31. Radnedge L, Agron PG, Worsham PL, Andersen GL: Genome plasticity in Yersinia pestis. Microbiology 2002, 148:1687-98.
- Fleischmann RD, Alland D, Eisen JA, Carpenter L, White O, Peterson J, DeBoy R, Dodson R, Gwinn M, Haft D, Hickey E, Kolonay JF, Nelson WC, Umayam LA, Ermolaeva M, Salzberg SL, Delcher A, Utterback T, Weidman J, Khouri H, Gill J, Mikula A, Bishai W, Jacobs WR Jr, Venter JC, Fraser CM: Whole-Genome Comparison of Mycobacterium tuberculosis Clinical and Laboratory Strains. J Bacteriol 2002, 184:5479-5490.
- 33. Hayashi T, Makino K, Ohnishi M, Kurokawa K, Ishii K, Yokoyama K, Han CG, Ohtsubo E, Nakayama K, Murata T, Tanaka M, Tobe T, Iida T, Takami H, Honda T, Sasakawa C, Ogasawara N, Yasunaga T, Kuhara S, Shiba T, Hattori M, Shinagawa H: Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 2001, 8:11-22.

- Perna NT, Plunkett G 3rd, Burland V, Mau B, Glasner JD, Rose DJ, Mayhew GF, Evans PS, Gregor J, Kirkpatrick HA, Posfai G, Hackett J, Klink S, Boutin A, Shao Y, Miller L, Grotbeck EJ, Davis NW, Lim A, Dimalanta ET, Potamousis KD, Apodaca J, Anantharaman TS, Lin J, Yen G, Schwartz DC, Welch RA, Blattner FR: Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature 2001, 409:529-33.
- 35. Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR: Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A 2002, 99:17020-4.
- 36. The Blast in the tandem repeats database page [http://minis atellites.u-psud.fr/Blast]
- 37. The Bacterial Genotyping Page [http://bacterial-genotyp ing.igmors.u-psud.fr]
- Tettelin H, Masignani V, Cieslewicz MJ, Eisen JA, Peterson S, Wessels MR, Paulsen IT, Nelson KE, Margarit I, Read TD, Madoff LC, Wolf AM, Beanan MJ, Brinkac LM, Daugherty SC, DeBoy RT, Durkin AS, Kolonay JF, Madupu R, Lewis MR, Radune D, Fedorova NB, Scanlan D, Khouri H, Mulligan S, Carty HA, Cline RT, Van Aken SE, Gill J, Scarselli M, Mora M, Iacobini ET, Brettoni C, Galli G, Mariani M, Vegni F, Maione D, Rinaudo D, Rappuoli R, Telford JL, Kasper DL, Grandi G, Fraser CM: Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V Streptococcus agalactiae. Proc Natl Acad Sci U S A 2002, 99:12391-6.
- McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, Hou S, Layman D, Leonard S, Nguyen C, Scott K, Holmes A, Grewal N, Mulvaney E, Ryan E, Sun H, Florea L, Miller W, Stoneking T, Nhan M, Waterston R, Wilson RK: Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. Nature 2001, 413:852-6.
- 40. Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, Wang J, Liu H, Yang J, Yang F, Zhang X, Zhang J, Yang G, Wu H, Qu D, Dong J, Sun L, Xue Y, Zhao A, Gao Y, Zhu J, Kan B, Ding K, Chen S, Cheng H, Yao Z, He B, Chen R, Ma D, Qiang B, Wen Y, Hou Y, Yu J: Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157. Nucleic Acids Res 2002, 30:4432-41.
- 41. Bricker BJ, Ewalt DR, Halling SM: Brucella 'Hoof-Prints': strain typing by multi-locus analysis of variable number tandem repeats (VNTRs). *BMC Microbiol* 2003, 3:15.
- Goodner B, Hinkle G, Gattung S, Miller N, Blanchard M, Qurollo B, Goldman BS, Cao Y, Askenazi M, Halling C, Mullin L, Houmiel K, Gordon J, Vaudin M, lartchouk O, Epp A, Liu F, Wollam C, Allinger M, Doughty D, Scott C, Lappas C, Markelz B, Flanagan C, Crowell C, Gurson J, Lomo C, Sear C, Strub G, Cielo C, Slater S: Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58. Science 2001, 294:2323-8.
- 43. Wood DW, Setubal JC, Kaul R, Monks DE, Kitajima JP, Okura VK, Zhou Y, Chen L, Wood GE, Almeida NF Jr, Woo L, Chen Y, Paulsen IT, Eisen JA, Karp PD, Bovee D Sr, Chapman P, Clendenning J, Deatherage G, Gillet W, Grant C, Kutyavin T, Levy R, Li MJ, McClelland E, Palmieri A, Raymond C, Rouse G, Saenphimmachak C, Wu Z, Romero P, Gordon D, Zhang S, Yoo H, Tao Y, Biddle P, Jung M, Krespan W, Perry M, Gordon-Kamm B, Liao L, Kim S, Hendrick C, Zhao ZY, Dolan M, Chumley F, Tingey SV, Tomb JF, Gordon MP, Olson MV, Nester EW: The genome of the natural genetic engineer Agrobacterium tumefaciens C58. Science 2001, 294:2317-23.
- Benson G, Dong L: Reconstructing the duplication history of a tandem repeat. Proc Int Conf Intell Syst Mol Biol 1999:44-53.
- van Belkum A, Scherer S, van Alphen L, Verbrugh H: Shortsequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev 1998, 62:275-93.
- 46. Schumacher S, Fuchs RP, Bichara M: Two distinct models account for short and long deletions within sequence repeats in Escherichia coli. J Bacteriol 1997, 179:6512-7.
- 47. De Bolle X, Bayliss CD, Field D, van de Ven T, Saunders NJ, Hood DW, Moxon ER: The length of a tetranucleotide repeat tract in Haemophilus influenzae determines the phase variation rate of a gene with homology to type III DNA methyltransferases. Mol Microbiol 2000, 35:211-22.