Appl Microbiol Biotechnol (2010) 87:2157–2167
DOI 10.1007/s00253-010-2671-7

Analysis of DNA repeats in bacterial plasmids reveals the potential for recurrent instability events

Pedro H. Oliveira · Kristala Jones Prather · Duarte M. F. Prazeres · Gabriel A. Monteiro

Received: 5 April 2010 / Revised: 5 April 2010 / Accepted: 5 May 2010 / Published online: 23 May 2010

Abstract Structural instability has been frequently observed in natural plasmids and vectors used for protein expression or DNA vaccine development. However, there is a lack of information concerning hotspot mapping, namely, DNA repeats or sequences identical to the host genome. This led us to evaluate the abundance and distribution of direct, inverted, and tandem repeats with high recombination potential in 36 natural plasmids from ten bacterial genera, as well as in several widely used bacterial and mammalian expression vectors. In natural plasmids, we observed an overrepresentation of close direct repeats in comparison to inverted ones and a preferential location of repeats with high recombination potential in intergenic regions, suggesting a highly plastic and dynamic behavior. In plasmid vectors, we found a high density of repeats within eukaryotic promoters and non-coding sequences. As a result of this in silico analysis, we detected a spontaneous recombination between two 21-bp direct repeats present in the human cytomegalovirus early enhancer/promoter (huCMV EEP) of the pCIneo plasmid. This finding is of particular importance, as the huCMV EEP is one of the most frequently used regulatory elements in plasmid vectors. Because pDNA integration into host gDNA can have adverse consequences in terms of plasmid processing and host safety, we also mapped several regions with high probability to mediate integration into the *Escherichia coli* or human genomes. Like repeated regions, some of these were located in non-coding regions of the plasmids, thus being preferential targets to be removed.

Keywords Genetic instability · Deletion formation · Genome integration · DNA vaccine

P. H. Oliveira (✉) · D. M. F. Prazeres · G. A. Monteiro
IBB—Institute for Biotechnology and Bioengineering,
Centre for Biological and Chemical Engineering,
Instituto Superior Técnico,
Av. Rovisco Pais,
1049-001 Lisbon, Portugal
e-mail: pcoliveira@ist.utl.pt

K. J. Prather
Department of Chemical Engineering,
Massachusetts Institute of Technology,
Room 66-458,
Cambridge, MA 02139, USA

Introduction

DNA repeats are considered to be catalysts of molecular evolution by promoting genetic instability and mutational events. These recombination events can give rise to deletions, duplications, translocations, and inversions (Peeters et al. 1988; Mazin et al. 1991; Lovett et al. 1993; Lovett et al. 1994; Bi and Liu 1994; Bi and Liu 1996; Morag et al. 1999; Ribeiro et al. 2008), potentially affecting critical genetic functions. Repeats in bacteria can be classified into two distinct groups: a first one comprising small low-complexity mono- to pentanucleotide DNA tracts usually arranged in head-to-tail configuration such as microsatellites, and a second group comprising larger tandem repeats, transposable elements, and spaced repeats. Both homologous and illegitimate (also termed *recA*-independent) recombination pathways are known to catalyze genetic rearrangements involving repeats and show high dependence on repeat and spacer length as well on the degree of similarity between

homologues (Mazin et al. 1991; Bi and Liu 1994; Bi and Liu 1996; Dianov et al. 1991; Chédin et al. 1994). In other words, close repeats (herein classified as having a spacer sequence smaller than 1 kb) are more prone to recombine than distant repeats, and may eventually proceed in a RecA-independent fashion. In fact, Achaz et al. (2002) propose that newly created close repeats experience strong rates of conversion and deletion, which will ultimately lead to a bias favoring the elimination of close repeats in comparison to more distant ones.

Looking at the particular case of bacterial plasmids, a handful of examples describing structural instability involving repeated regions have been reported. One example is the one of pJS37, a natural hybrid plasmid composed of the streptococcal plasmid pLS1 and the staphylococcal plasmid pC194, in which deletions were observed during rolling-circle replication as a result of pairing between homologous regions from both plasmids (Ballester et al. 1989). Other examples include inverted repeat-mediated multimerization accompanied by a decrease in copy number in derivatives of the IncFI plasmid ColV2-K94 (Weber and Palchaudhuri 1986) and ability of *Cupriavidus necator* (former *Ralstonia eutropha*) to grow on chloroaromatic compounds after homologous recombination in the catabolic plasmid pJP4 (Larraín-Linton et al. 2006).

On the other hand, the frequent use of plasmid vectors has resulted in the detection of repeat-mediated instability phenomena. Some examples include recombination in the pGEX series of plasmids, leading to decreased levels of glutathione S-transferase (GST) expression (Borgi and Gargouri 2008), deletions in pHZ1358, a vector for targeted gene disruption and replacement experiments in *Streptomyces* hosts (Sun et al. 2009), and deletions in *lac*-controlled expression vectors hosted in *Escherichia coli* strains (Kawe et al. 2009). Our group has also identified a direct repeat-mediated deletion formation, taking place in pCIneo, a mammalian expression vector also used for DNA vaccine development (Ribeiro et al. 2008). This recombination was caused by the presence of two 28-bp direct repeats located in non-coding regions of the plasmid and resulted in the deletion of several elements such as the multiple cloning site and SV40 early enhancer/promoter (SV40 EEP) regions. Despite the accumulating evidence for plasmid structural instability, no consistent mapping of repeated regions has been performed to date in natural bacterial plasmids or common cloning and expression vectors. Therefore, in this work, we have focused our attention on the abundance and distribution of direct and inverted repeats in plasmids. While this analysis performed in natural plasmids aims to provide insight about their own plasticity and evolution, the same analysis conducted in cloning and expression vectors is justified by the world-wide growing demand for pDNA (e.g., for transgene expression or biopharmaceutical applications) (see recent reviews by Bower and Prather 2009; Oliveira et al. 2009a). As a result of this hotspot mapping, we were able to detect a novel recombination event occurring within the human cytomegalovirus early enhancer/promoter (huCMV EEP) region of a plasmid vector, a fact that is of concern as this is one of the most commonly used regulatory elements in plasmids.

Materials and methods

Plasmid DNA sequence data

We analyzed the complete nucleotide sequence of 36 natural plasmids from ten different bacterial genera (Table 1), 16 cloning vectors, and 17 mammalian expression vectors (Table 2). Plasmid nucleotide sequences were obtained from the Genbank, or alternatively, from the supplier website. Criteria for plasmid vector selection involved searching those highly used for gene cloning, protein expression, or DNA vaccine development with heterogeneous size dispersion. Although the majority of these vectors share common structural features, we discarded those pertaining to the same category sharing minimal backbone changes (e.g., point mutations, inverted genes). The chosen vectors also harbor the large majority of structural elements typically found in plasmid vectors (such as origin of replication, different antibiotic resistance genes, regulatory sequences, etc).

Identification and classification of repeats

We used RepSeek (Achaz et al. 2007) to search for direct and inverted repeats with a minimal seed length given by the Karlin and Ost extreme statistics (Karlin and Ost 1985) at a P value<0.001. The output of Repseek was further clustered into the following repeat categories: close direct repeats (CDR), close inverted repeats (CIR), distant direct repeats (DDR), and distant inverted repeats (DIR). The distinction between “close” and “distant” repeats was based on a spacer length respectively shorter or larger than 1 kb. Tandem repeats (TR) were also searched using information provided by Tandem Repeats Finder (TRF) (Benson 1999) with all default parameters set, yet allowing a maximum period size of 2,000 bp. TR are arranged in a contiguous head-to-tail fashion and typically contain a minimal repetitive motif of larger than five nucleotides. When several overlapping tandems were identified in the same region, only the longest one was considered. Tandem motifs of one to five nucleotides (usually known as simple sequence repeats (SSR)) were not considered in the analysis.

Table 1 List of bacterial natural plasmids analyzed

Class	Organism	Plasmid (Genbank #)	Size (kb)	D_N (CDR)	D_N (CIR)	D_N (DDR)	D_N (DIR)	D_N (TR)	Total RRP
Beta Proteobacteria	*Burkholderia cenocepacia* HI2424	1 (CP000461)	164.857	0.024	0.006	0.036	0.018	0.036	0
	Burkholderia cenocepacia J2315	pBCJ2315 (AM747723)	92.661	0.043	0.032	0.032	0.022	0.022	0.002
	Burkholderia vietnamiensis G4	pBVIE01 (CP000617)	397.868	0	0	0.038	0.035	0	0
	Burkholderia vietnamiensis G4	pBVIE05 (CP000621)	88.096	0.011	0	0.011	0.011	0.011	0.005
Gamma Proteobacteria	*Escherichia coli* O157:H7 EDL933	pO157 (AF074613)	92.077	0.022	0.011	0.022	0.065	0.033	0
	Escherichia coli DU1040	NR1 (DQ364638)	94.289	0.032	0.011	0.042	0.032	0	0
	Escherichia coli EH41	pO113 (AY258503)	165.548	0.048	0.006	0.145	0.054	0.006	0.119
	Escherichia coli	pMAR7 (DQ388534)	101.558	0.010	0.030	0.020	0.049	0.010	0
	Escherichia coli	pC15-1a (AY458016)	92.353	0.022	0.022	0.130	0.065	0	0.864
	Escherichia coli K12	Clodf13 (X04466)	9.957	0	0	0	0	0	0
	Salmonella enterica	pAM04528 (FJ621587)	158.213	0.006	0.006	0.303	0.234	0.006	0
	Salmonella choleraesuis	pMAK1 (AB366440)	208.409	0.086	0	0.149	0.058	0.034	0
	Salmonella typhi	pHCM1 (AL513383)	218.160	0.078	0.009	0.083	0.087	0.037	0
	Salmonella typhimurium	R64 (AP005147)	120.826	0.017	0.041	0.033	0.017	0.008	0.268
	Salmonella dublin	pOU1115 (DQ115388)	74.589	0.054	0.013	0	0.027	0.040	0
	Pseudomonas aeruginosa	pMATVIM7 (AM778842)	24.179	0.041	0	0	0.083	0.041	0
	Pseudomonas aeruginosa	RMS149 (AJ877225)	57.121	0.053	0.035	0.053	0.053	0.053	0
	Pseudomonas aeruginosa	pBS228 (AM261760)	89.147	0.022	0	0.034	0.056	0.022	0
	Klebsiella pneumoniae	pCTXM360 (EU938349)	68.018	0.132	0	0.456	0.044	0.015	0
	Klebsiella pneumoniae	pKPN3 (CP000648)	175.879	0.040	0	0.108	0.045	0.011	0.083
Epsilon Proteobacteria	*Helicobacter pylori* P29	pHEl5 (AF469113)	18.291	0.109	0	0.055	0.055	0.164	0
	Helicobacter pylori	pAL202 (AY584531)	12.120	0.083	0	0.165	0.083	0.248	0.324
	Helicobacter pylori	pHP489 (AF027303)	1.222	0	0	0	0	0.818	0.441
	Helicobacter pylori	pHP51 (AY267368)	3.955	0.759	0	0	0	0.506	2.226
	Helicobacter pylori	pHP666 (DQ198799)	8.108	0.123	0	0	0	0.123	0.367
	Helicobacter pylori	pHP69 (DQ915941)	9.153	0.109	0	0	0	0.109	0.320
	Helicobacter pylori	pHEl4 (AF469112)	10.970	0.091	0	0.182	0	0.091	0.314
	Helicobacter pylori HPM8	pHPM8 (AF275307)	7.817	0.128	0	0	0.128	0.128	0.343
Actinobacteria	*Streptomyces coelicolor* A3(2)	SCP1 (AL589148)	356.023	0.003	0.014	0.031	0.034	0.008	0.701
Bacilli	*Bacillus cereus* E33L	pE33L466 (CP000040)	466.370	0.051	0.011	0.268	0.356	0.019	0
	Bacillus anthracis	pX01 (CP001216)	181.773	0.022	0.028	0.033	0.072	0.011	0
	Bacillus subtilis	p1414 (AF091592)	7.949	0.126	0.126	0	0.126	0	0
	Bacillus thuringiensis	pBMB67 (DQ363750)	67.159	0.074	0.030	0.089	0.030	0.015	0
Deinococci	*Deinococcus radiodurans* R1	MP1 (AE001826)	177.466	0.051	0	0.220	0.248	0.006	0.237
	Deinococcus radiodurans R1	CP1 (AE001827)	45.704	0	0	0.219	0.263	0	0
	Thermus thermophilus	pTF62 (DQ058601)	10.402	0.096	0.096	0.192	0.096	0	0

Shown are the values of repeat density (D_N) for close direct repeats, close inverted repeats, distant direct repeats, distant inverted repeats, tandem repeats, and relative recombination potential. The latter was calculated on the basis of those repeats showing the highest recombination potential (strictly identical close and tandem repeats only)

CDR close direct repeats, *CIR* close inverted repeats, *DDR* distant direct repeats, *DIR* distant inverted repeats, *TR* tandem repeats, *RRP* relative recombination potential

Density of repeats and recombination potential

In order to analyze the abundance of repeats, we used the density in number (D_N) (Achaz et al. 2002), defined as:

$$D_N = \frac{\text{Number of repeat copies}}{\text{Plasmid size (kb)}} \quad (1)$$

Because a higher density of repeats does not necessarily imply higher levels of structural instability, we have estimated the relative recombination potential (RRP) for each plasmid in a similar way as previously performed for bacterial genomes (Rocha 2003). We have used the correlation proposed by Oliveira and co-workers (Oliveira

Table 2 List of bacterial and mammalian expression vectors analyzed in this study

Type of plasmid	Plasmid	Size (kb)	Vendor/Genbank accession no	D_N (CDR)	D_N (CIR)	D_N (DDR)	D_N (DIR)	D_N (TR)	Total RRP
Bacterial expression and cloning vectors	pUC19	2.686	M77789	0	0	0	0	0	0
	pBR327	3.274	L08856	0	0	0	0		0
	pBR322	4.361	J01749	0	0	0	0		0
	pET3a	4.64	New England Biolabs	0	0	0	0		0
	PinPoint XA-1	3.331	Promega/U47626	0	0	0	0		0
	pEXP4-DEST	4.415	Invitrogen	0	0	0	0.227		0
	pSP72	2.462	Promega/X65332	0	0	0	0		0
	pET SUMO	5.643	Invitrogen	0	0	0	0		0
	pCR 2.1-TOPO	3.931	Invitrogen	0	0.254	0	0		0
	pBAD TOPO	4.126	Invitrogen	0.242	0.242	0	0		0
	pGEX-4T2	4.97	GE Healthcare/U13854	0	0	0.201	0		0
	pQE-30	3.461	Qiagen	0.289	0.289	0	0		0.016
	pACYC177	3.941	X06402	0	0.254	0	0		0
	pET161-GW/CAT	6.518	Invitrogen	0	0	0	0		0
	pBluescript II KS (+)	2.961	Stratagene/X52327	0	0	0	0		0
	pTrc His A	4.414	Invitrogen	0.227	0.227	0	0.227		0
Mammalian expression vectors	pCIneo[a]	5.472	Promega/U47120	1.096	0.731	0.183	0	0.365	0.554
	pFN10A (ACT) Flexi(R)	5.867	Promega/DQ487211	0.682	0.511	0.341	0	0.341	0.553
	pVAX1[a]	2.999	Invitrogen	2.001	1.334	0	0	0	0.003
	pACT	5.566	Promega/AF264723	0.898	0.539	0.180	0	0.359	0.553
	Gateway pDEST26	7.481	Invitrogen	0.802	0.401	0	0.401	0.267	0.553
	pAdvantage[a]	4.392	Promega/U47294	0	0	0	0	0	0
	pBIND	6.360	Promega/AF264722	0.629	0.472	0.157	0.157	0.314	0.552
	pG5luc	4.955	Promega/AF264724	0.404	0	0	0	0.202	0.003
	pTNT	2.871	Promega/AF479322	0.348	0	0	0	0	0.019
	pTarget	5.670	Promega/AY540613	0.882	0.705	0.176	0.000	0.353	0.553
	pReg neo	6.802	Promega/EF030522	0.441	0	0.441	0.147	0.735	1.879
	pCat3-Basic	4.027	Promega/U57024	0	0.248	0	0	0	0
	pCat3-Control	4.465	Promega/U57025	0	0.224	0.000	0.896	0.672	0.667
	pSI	3.632	Promega/U47121	0	0	0	0	0.551	0.552
	pcDNA 6.2/cLumio-DEST	6.809	Invitrogen	0.734	0.441	0.000	0.294	0.294	0.553
	gWiz	5.063	Genlantis, Inc	3.753	0.790	0	0	0	0.502
	pCMV•SPORT-βgal	7.854	Invitrogen	0.637	0.509	0.127	0.382	0	0.001

Criteria for pDNA selection were based on their frequent use for protein expression or therapeutic applications, as well as on heterogeneous size distribution. Furthermore, these vectors harbor the largest majority of structural elements typically found in each family (bacterial and mammalian). In a similar way to what was used for natural plasmids, RRP was calculated on the basis of those repeats showing the highest recombination potential (strictly identical close and tandem repeats only)

[a] Vectors used as backbones for DNA vaccine development

et al. 2008) for estimation of recombination frequencies (F_R) in plasmids harboring direct repeats, and defined RRP of a given repeat as

$$\mathrm{RRP} = \begin{cases} 0 & \text{if} \quad L_R < 14 \ \wedge \ L_S \geq 1000 \\ \frac{F_R(L_R, L_S)}{F_R(856, 0)} & \text{if} \quad (14 \leq L_R \leq 856) \ \wedge \ (0 \leq L_S < 1000) \end{cases} \tag{2}$$

where L_R and L_S are, respectively, the length of repeat and spacer sequence in base pairs. F_R (856, 0) is a normalization factor that corresponds to the highest recombination frequency obtained in the meta-analysis developed by Oliveira et al. (2008). F_R also becomes negligible for low values of L_R (<14 bp) and large values of L_S (≥ 1,000 bp) (see Oliveira et al. 2008); thus, RRP was considered to be 0 in these regions. The boundary conditions chosen for L_R and L_S result from the available data on plasmid recombination frequency and also restrict the analysis to repeats showing the highest RRP values (L_S < 1,000 bp). The minimal repeat length considered in

Eq. 2 is also in agreement with the range of values of seed length given by the Karlin and Ost extreme statistics (14–23 bp for natural plasmids and 14–15 bp for plasmid vectors). Unfortunately, there was not enough data available in the literature regarding recombination mediated by inverted repeats; therefore, we were not able to compute the associated RRP. The total relative recombination potential (RRP_T) of each plasmid was computed by adding the contributions of the partial recombination potentials associated with CDR and TR:

$$RRP_T = \sum_i (RRP_i)_{CDR} + \sum_i (RRP_i)_{TR} \quad (3)$$

In the case of TR, the RRP associated with n repeat copies was estimated by taking into account the frequency of the most probable and frequent recombination events given by $(n-1)\cdot F_{R_i}(L_R, 0)$, among all the $n!/[2\cdot(n-2)!]$ possible combinations of two-copy repeats. Yet, three important observations should be drawn:

(a) The assumption of adding several RRPs shown in Eq. 3 does not take into account changes in DNA structure resulting from recombination. In other words, recombination between two repeats A_1 and A_2 may change the RRP of two other repeats B_1 and B_2 by altering, for example, their initial spacer distance. This is actually a complex problem to solve because it would be necessary to predict not only the outcome of each rearrangement (deletion, duplication) but also their specific rate. Therefore, RRP_T taken from Eq. 3 can be seen as a zero or initial recombination potential.

(b) In this work, estimation of F_R in plasmids was performed using correlations developed for $RecA^+$ or $RecA^-$ bacterial strains (Oliveira et al. 2008). These strains are often used for plasmid propagation in order to minimize recombination events. Because strains carrying other mutations, particularly affecting the RecBCD functions, are known to alter recombination frequency, appropriate correlations should be used in those cases.

(c) The correlations for F_R estimation were initially developed on the basis of pDNA harboring strictly identical repeats (Oliveira et al. 2008). However, it is known that F_R decreases abruptly with the introduction of mismatches in the repeats (e.g., a single mismatch introduced in repeats found in *E. coli* plasmids leads to a drop of roughly 100-fold in F_R (Bi and Liu 1994)). We have thus used Eqs. 2 and 3 only for those cases involving perfectly identical repeats, which results in a slight underestimation of RRP.

Bacterial strains and plasmid DNA

Escherichia coli DH5α (F^- *ϕ*80d*lacZΔ*M15 *Δ*(*lacZYA-argF*) U169 *recA1 endA1 hsdR17*(r_k^-, m_k^+) *phoA supE44* λ^- *thi-1 gyrA96 relA1*) (Invitrogen) was used for pDNA propagation. The plasmid used in this work was pCIneo, a 5,472-bp mammalian expression vector (Promega).

Detection of recombination within the 21-bp direct repeat cluster of the huCMV promoter

Recombination within the 21-bp direct repeat cluster of the huCMV EEP was detected through PCR amplification of an internal 362-bp huCMV fragment encompassing the three copies of 21-bp direct repeats (Fig. 3b). PCR amplification was performed with primers CMV-For21 (5′ AATATGACCGC CATGTTGG 3′) and CMV-Rev21 (5′ GCCAAGTAG GAAAGTCCC 3′). All PCR reactions were performed in 25-μl final volume containing 2.5 mM $MgCl_2$, 0.1 μM each primer, 0.4 mM dNTPs and 2.0 μl of sample. The following amplification program was used: 10 min at 95 °C followed by 35 cycles of 30 s at 94 °C, 20 s at 53 °C and 30 s at 72 °C.

DNA sequence analysis

PCR fragments were gel-purified using the QIAquick gel extraction kit (Qiagen). Nucleotide sequence of the fragments was determined by STAB Vida (Portugal) by using primers CMV-For21 and CMV-Rev21.

Results

Abundance and distribution of repeats in bacterial natural plasmids

We found 1,084 total repeats among the 36 natural plasmids analyzed (Table 1). These plasmids were isolated from several well-known organisms representative of proteobacteria, actinobacteria, deinococcus-thermus, and firmicutes. Among all plasmids analyzed, the ones from *Helicobacter pylori* caught our attention. Although these plasmids are typically small, they show particularly high values of CDR density and RRP (Table 1). In fact, one of these plasmids, pHP51, was found to have the highest values of CDR density and RRP among all plasmids analyzed, mainly due to the presence of a perfectly similar TR having three copies of 228 bp each (previously termed R3 by Song et al. 2003), located in a non-coding region. Interestingly, a nucleotide BLAST search using this TR as query revealed identity with plasmid pHPO100 from *H. pylori* in only one repeat copy (data not shown). pHPO100 is actually very similar to pHP51 in terms of structure and functionality (hence, it was excluded from our analysis), but a close analysis of its sequence revealed a deletion of approximately two copies of the TR. This observation suggests that pHPO100 might have originated from pHP51,

for example, as a result of *recA*-independent slipped misalignment between adjacent repeats. Moreover, the generalized high density of CDR found in *H. pylori* plasmids (Table 1) points to a high rate of repeats being created and, therefore, to a higher plasticity and potential for adaptation. These findings also agree with a previous suggested role for repeats from *H. pylori* plasmids as hot spots for recombination and site-specific integration events (Hofreuter and Haas 2002).

We next focused our attention on those repeats typically showing the highest RRP values (close repeats and TR) and, therefore, being more prone to recombine. CDR were found to be more abundant (25.0% among close+distant repeats) than CIR (9.6 %) (Fig. 1a, b). This tendency is better seen in Fig. 1c, where the overrepresentation of direct repeats is obvious compared with inverted repeats for spacer lengths shorter than 10 kb, while the opposite is observed for larger spacers. In fact, similar behavior involving the avoidance of sequences capable of producing inversions has been observed in prokaryotic and eukaryotic genomes (Achaz et al. 2001; Achaz et al. 2002), suggesting that selection might have acted towards the minimization of inversions, for example, by keeping existing inverted repeats further apart.

Both close repeats and TR mapped preferentially within intergenic regions (48–70%) and, to a minor extent, within regions involved in replication and mobilization, mobile elements, or other regions with no function assigned (Fig. 1d). Because intergenic regions are less exposed to selective pressure, repeats located in these regions are usually too unstable to persist (particularly if they are long) and tend to be gradually deleted. This bias favoring deletion of close repeats and TR located in poorly selected regions has been proposed by Achaz et al. (2002) for bacterial genomes, and is suggested to be involved in reductive evolution of some genomes. In our case, unless the intergenic repeats are moved farther apart by other genetic rearrangements (such as translocations or inversions), they will tend to mediate a gradual process of plasmid minimization.

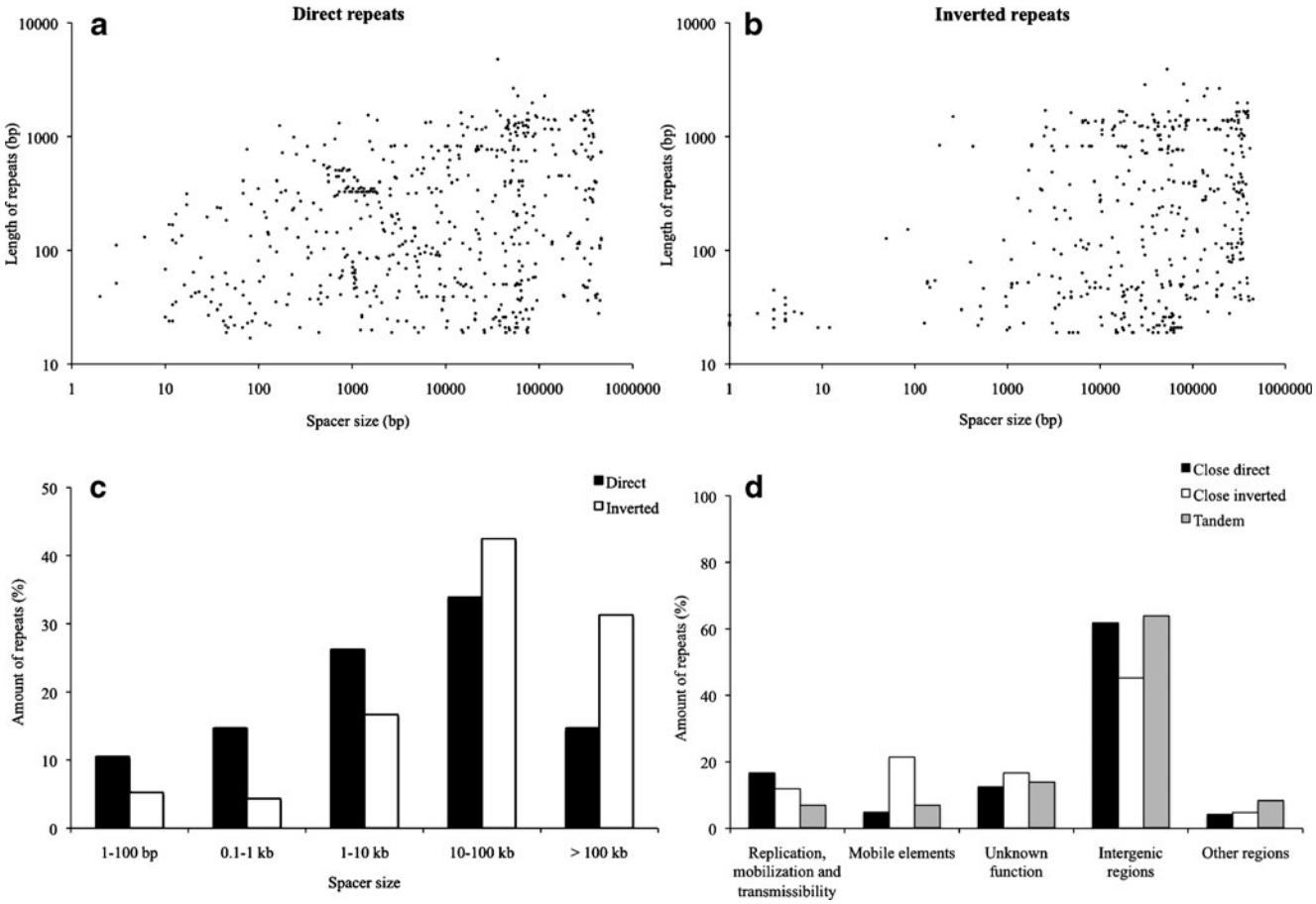

Fig. 1 Graphical representation of repeat vs spacer length for natural plasmids reveals a tendency for overrepresentation of direct repeats in comparison with inverted repeats (**a**, **b**). This overrepresentation was found to be valid for spacer lengths smaller than 10 kb, while an opposite tendency was seen for larger spacers (**c**). Repeats highly prone to recombine (CDR, CIR, and TR) were found to preferentially locate within intergenic regions, thus being good candidate regions to suffer genetic rearrangements (**d**)

Abundance and distribution of repeats in cloning and expression vectors

We have found an unexpectedly high density of repeats, particularly in mammalian expression vectors (Table 2). The large majority of mammalian vectors exhibits a higher density of direct repeats in comparison with inverted ($P < 0.05$, binomial test), while similar analysis was not statistically relevant in bacterial expression vectors. It is worth noticing that the plasmid showing the highest density of CDR was gWiz, although pReg neo was the one having the highest RRP (due to the presence of several tandems within the SV40 promoter and lambda operator sites) (Table 2).

From the standpoint of plasmid design (particularly for protein expression purposes or development of biopharmaceuticals), there is significant interest in identifying exactly the type of repeats present and their location in the vector. This will allow the manufacturer to avoid, minimize, or redesign the regions harboring such sequences. Therefore, we looked at the location and preferential length of repeats showing the highest RRP values (close repeats and TR). We found that TR were absent from bacterial vectors, while CDR and CIR were predominantly located within non-coding regions and regulatory sequences (such as transcriptional terminators or binding sites) (Fig. 2a). Concerning mammalian expression vectors, TR were almost exclusively located within the SV40 EEP (Fig. 2b) and mainly comprised the well-known 21- and 72-bp clusters involved in promoting its enhancer activity. CDR and CIR were predominantly found within eukaryotic promoters and, to a lesser extent, within non-coding regions. The large percentage of close direct and inverted repeats lying within the huCMV EEP mainly stems from the presence of 17-, 18-, 19-, and 21-bp arrays of transcriptional regulatory repeated elements (Boshart et al. 1985). In both bacterial and mammalian vectors, the large majority of CDR and CIR found were 14–40 bp long (Fig. 2c, d). TR found in

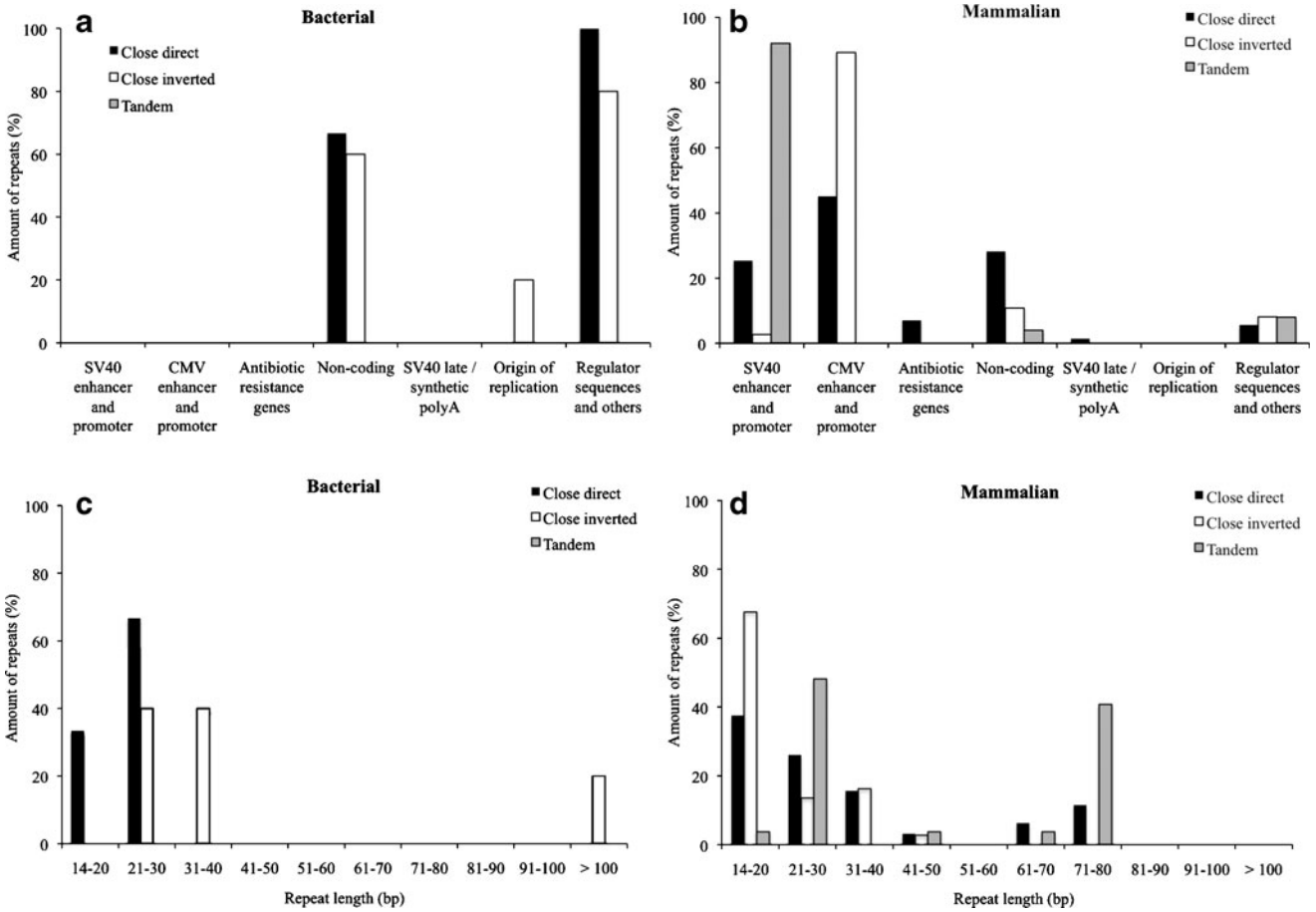

Fig. 2 Repeat mapping in expression and cloning vectors has shown that CDR and CIR are preferentially located within non-coding and regulator sequences (**a**), while in mammalian expression vectors, they essentially map within eukaryotic promoters and non-coding regions (**b**). TR were predominantly found (**b**). Because many of the repeats found encompassed more than one structural element of the plasmid, the sum of the partial percentages is higher than 100%. The majority of CDR and CIR found were 14–40 bp in length (**c**, **d**) while TR had typical periods of 21 and 72 bp (**d**)

mammalian expression vectors were mainly distributed within the 21- and 72-bp clusters of SV40 (Fig. 2d).

Detection of a mutation within the huCMV EEP of pCIneo

We have shown in a previous work that recombination between two 28-bp direct repeats in the mammalian expression vector and DNA vaccine backbone pCIneo gives rise to monomeric and heterodimeric products (Ribeiro et al. 2008). After submitting the nucleotide sequence of pCIneo to Repseek, we actually found a total of six possible combinations of direct repeat-mediated mutations (repeat length ≥14 bp) (Fig. 3a). Among these, five combinations of CDR mapped within the huCMV EEP repeat cluster (Fig. 3a). Although it would be expected that these repeats could undergo spontaneous recombination, we only found a single report describing genetic instability in this eukaryotic element. It involved the generation of aberrant replication intermediates in *E. coli* during amplification of pVAX1 and pL33A as a result of replication-fork stalling at the direct repeat cluster of the huCMV EEP (Levy 2003). Yet, no deletions were reported. This fact led us to search pCIneo for such mutations. We chose to focus on the 21-bp region because it contains the only two (out of three) perfectly similar repeats among all the CMV repeats (Fig. 3b). As a result, recombination between the latter repeats would, in theory, occur at a higher frequency, thus being detected more easily.

PCR amplification of the region encompassing the 21-bp region gave rise to an extra 169-bp fragment besides the expected 362-bp fragment (Fig. 3c). Sequencing of this fragment led us to conclude that the two perfectly similar 21-bp direct repeats were recombining, resulting in the deletion of the spacer sequence plus one direct repeat (Fig. 3d). No evidence was found for recombination between any of these repeats with the third non-perfectly similar sequence.

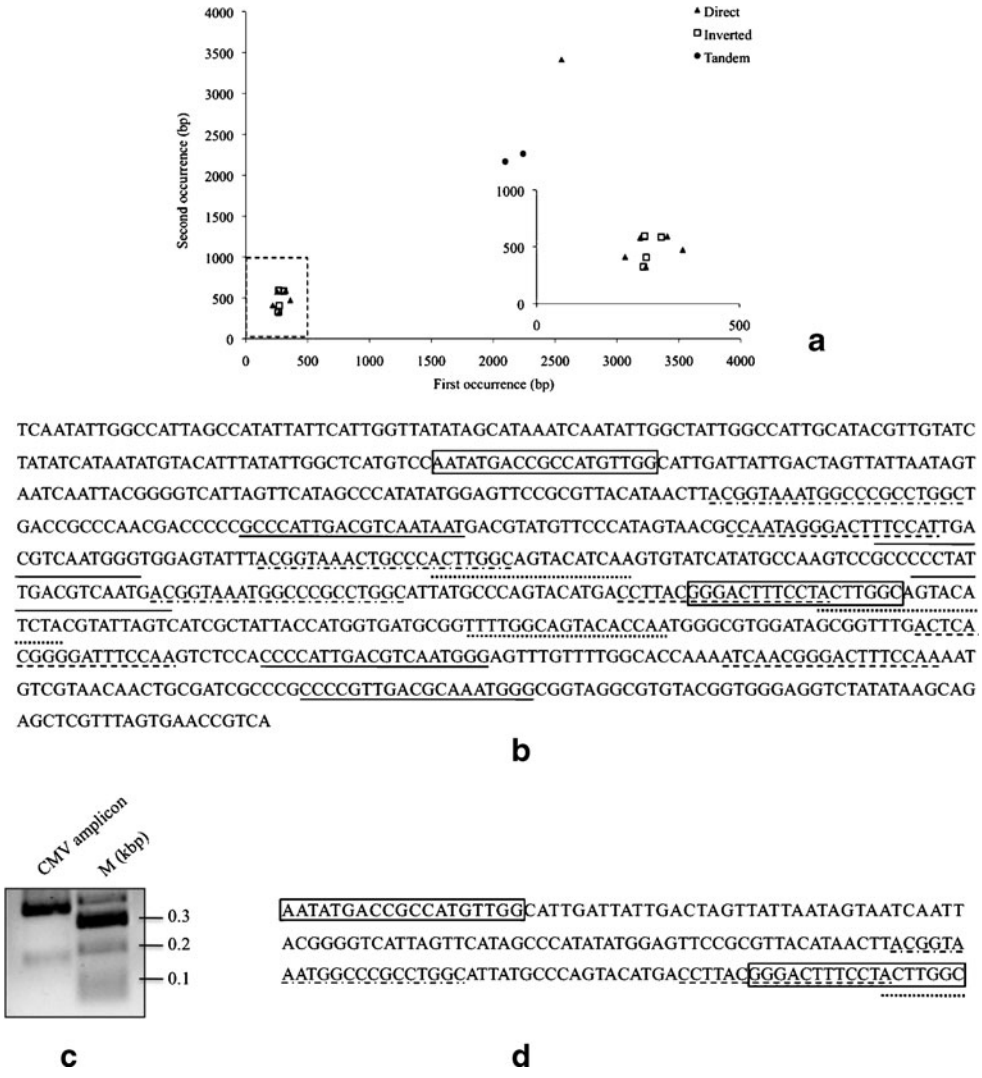

Fig. 3 **a** Repeat distribution in pCIneo indicating the first and second occurrence of CDR, CIR, and TR. The magnified region represents the distribution of repeats found within the huCMV EEP. **b** Major EEP region of the huCMV gene. The four classes of imperfect direct repeats are underlined as follows: *dotted lines* 17 bp; *dashed lines* 18 bp; *solid lines* 19 bp, and *dash dotted lines* 21 bp. *Boxed regions* indicate the CMVFOR21 and CMVREV21 primers. **c** PCR amplification of pure pCIneo using primers CMVFOR21 and CMVREV21 gave rise to a 362-bp fragment, as well as a least abundant 193-bp fragment. **d** Sequencing of the 193-bp amplicon led us to conclude that recombination occurred between two (out of the three) 21-bp direct repeats

Evaluation of the insertion potential of pDNA into the *E. coli* and human genome

Apart from intramolecular recombination, there is the possibility that pDNA might integrate into host gDNA if similar regions happen to be present. Cases of spontaneous pDNA integration have been documented (Richardson and Park 1997), a fact that gains even more relevance if the molecules are to be used for therapeutic purposes (Wang et al. 2004). In order to address this issue, we started by performing a similarity search between the vectors shown in Table 1 and the *E. coli* genome. We found several highly similar regions (length higher than 50 nucleotides and percentage of similarity higher than 90%) (Table 3). Although plasmid sequences such as the ones pertaining to the *lac* operon and *rrn*B T1 and T2 transcription terminators were already expected to reveal similarity with the *E. coli* genome, non-coding plasmid sequences also contained partially deleted regions from *lac* genes (Table 3). Because pVAX1, pCIneo, and pAdvantage are also used as backbones for DNA vaccine development, we also searched for similarity between these vectors and the human genome. No significant similarity was obtained for pVAX1, whereas high similarity was obtained between the chimeric intron of pCIneo and the beta/delta globin human genes and between a non-coding region of pAdvantage and an intergenic region within chromosome 5.

Discussion

It is now becoming clearer that hotspots able to engage in recombination are widespread among plasmids. The reason why we do not detect these mutations more often relies on the simple fact that most of them do not carry any selective advantage over other non-mutant plasmid forms already present. Thus, each of these mutants is generally maintained at a low frequency within a plasmid population. In other words, we may have a considerable sum of different mutants in a population, yet not be easily detecting them (e.g., by agarose gel). Yet, the frequency at which these deletions or amplifications are detected in pDNA may eventually rise in the presence of damaging/stressful exogenous factors such as antibiotics, temperature, medium composition, and others (Chowdhury et al. 1996; Sandegren and Andersson 2009; Oliveira et al. 2009b).

In this work, it was shown that CDR are more abundant than CIR, particularly in *H. pylori* plasmids. Because colonization of the primate stomach by *H. pylori* may extend for decades, this bacterium might have developed a highly programmed system of diversification based on extensive nonrandomly distributed repeats in order to subsist in a highly stressful environment (Aras et al. 2003). On the other hand, when looking at plasmid vectors, we found that regions such as the huCMV EEP, SV40 EEP, and non-coding ones harbor a high density of repeats. To overcome the high recombination potential of multiple repetitive elements contained within the huCMV EEP and SV40 EEP, the use of more stable regulatory elements should, in some cases, be considered. To test this possibility, we submitted to Repseek the nucleotide sequences of three alternative promoters: the one from the human *UbC* ubiquitin gene (Genbank accession number D63791, base pairs 3561–4771), from rous sarcoma virus (RSV) (taken from the mammalian expression vector pRc/RSV, Invitrogen) and from human elongation factor 1α promoter (taken from pEF1/V5-HisA, Invitrogen). Interestingly, none of these alternative promoters revealed the presence of any type of

Table 3 Identical regions found between several plasmid vectors and the *E. coli* and human genomes

Regions in pDNA	Regions in *E. coli* gDNA	Examples	Location in vector (bp)/identity (%)
lac operon sequences	*lac* operon sequences	pET SUMO pTarget pET161-GW/CAT pBluescript II KS(+)	4341–5543/99%; 1066–1226/90% 1363–1499/99%; 5028–6230/99%; 460–618/100% 802–1031/100%
rrnB T1 and T2 terminators	Ribosomal RNA 5 S genes	pBAD TOPO pQE 30	496–921/100% 1064–1161/100%
tac promoter	Intergenic sequence and *yci*Vgene involved in tryptophan regulation	PinPoint XA	3223–3277/96%; 3032–3224/100%
Arabinose promoter and regulatory elements	Intergenic	pBAD TOPO	1–299/99%
Non coding sequences	Portions of *lac* elements	pCIneo pSP72 pTNT	1412–1482/100% 140–193/100% 216–286/100%
Regions in pDNA	Regions in human gDNA	Examples	Location in vector/identity (%)
Chimeric intron	Beta globin gene	pCIneo	864–980/94%
Non coding	Intergenic	pAdvantage	1733–1836/94%

Only regions larger than 50 bp and sharing an identity equal or higher than 90% were considered

internal repeated element, thus representing a clear advantage in terms of stability over the multiple repeats of the huCMV EEP and SV40 EEP. Moreover, the former have proven to be valuable alternatives in terms of generating identical (or higher) levels of transgene expression (Gill et al. 2001; reviewed by Garmory et al. 2003). However, if human clinical applications are envisaged, the use of the *UbC* ubiquitin gene promoter or elongation factor 1α promoter in vectors should be avoided due to the inherent risk of genome integration.

Concerning non-coding regions, it was shown in this work that they harbor a considerable amount of repeats in both bacterial and mammalian vectors (Fig. 2a, b). Aiming at the elimination of some potential instability hotspots, few strategies have been proposed that involve deletion of any non-essential and potentially detrimental backbone sequences. Examples of these reduced molecules include the minimalistic immunogenic defined gene expression (MIDGE) vectors (Moreno et al. 2004), minicircles (Darquet et al. 1999; Bigger et al. 2001), plasmid-mediated repressor titration systems (Cranenburgh et al. 2001), and RNA-based selection systems (Mairhofer et al. 2008). Because sequences identical to the *E. coli* and human genome were seen to be present in non-coding regions of pDNA, minimization might also be beneficial in avoiding potential genome integration events. We personally foresee a promising future for minimal plasmids in tackling some structural instability events, although the currently available methods of plasmid minimization suffer from low efficiency and a lack of cost-effective scale-up processes, problems that must be solved to make them suitable for clinical application.

Acknowledgements This work was supported by Fundação para a Ciência e a Tecnologia (POCI/BIO/55799/2004 and Ph.D. grant BD/22320/2005 to Pedro H. Oliveira) and by the MIT-Portugal Program.

References

Achaz G, Netter P, Coissac E (2001) Study of intrachromosomal duplications among the eukaryote genomes. Mol Biol Evol 18:2280–2288

Achaz G, Rocha EPC, Netter P, Coissac E (2002) Origin and fate of repeats in bacteria. Nucleic Acids Res 30:2987–2994

Achaz G, Boyer F, Rocha EPC, Viari A, Coissac E (2007) Repseek, a tool to retrieve approximate repeats from large DNA sequences. Bioinformatics 23:119–121

Aras RA, Kang J, Tschumi AI, Harasaki Y, Blaser MJ (2003) Extensive repetitive DNA facilitates prokaryotic genome plasticity. Proc Natl Acad Sci USA 100:13579–13584

Ballester S, Lopez P, Espinosa M, Alonso JC, Lacks SA (1989) Plasmid structural instability associated with pC194 replication functions. J Bacteriol 171:2271–2277

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580

Bi X, Liu LF (1994) recA-independent and recA-dependent intramolecular plasmid recombination. Differential homology requirement and distance effect J Mol Biol 235:414–423

Bi X, Liu LF (1996) A replicational model for DNA recombination between direct repeats. J Mol Biol 256:849–858

Bigger BW, Tolmachov O, Collombet JM, Fragkos M, Palaszewski I, Coutelle C (2001) An araC-controlled bacterial cre expression system to produce DNA minicircle vectors for nuclear and mitochondrial gene therapy. J Biol Chem 276:23018–23027

Borgi I, Gargouri A (2008) A spontaneous direct repeat deletion in the pGEX fusion vector decreases the expression level of recombinant proteins in Escherichia coli. Prot Express Purif 60:15–19

Boshart M, Weber F, Jahn G, Dorsch-Häsler K, Fleckenstein B, Schaffner W (1985) A very strong enhancer is located upstream of an immediate early gene of human cytomegalovirus. Cell 41:521–530

Bower DM, Prather KJ (2009) Engineering of bacterial strains and vectors for the production of plasmid DNA. Appl Microbiol Biotechnol 82:805–813

Chédin F, Dervyn E, Dervyn R, Ehrlich SD, Noirot P (1994) Frequency of deletion formation decreases exponentially with distance between short direct repeats. Mol Microbiol 12:561–569

Chowdhury R, Sahu GK, Das J (1996) Stress response in pathogenic bacteria. J Biosci 21:149–160

Cranenburgh RM, Hanak JAJ, Williams SG, Sherratt DJ (2001) Escherichia coli strains that allow antibiotic-free plasmid selection and maintenance by repressor titration. Nucleic Acids Res 29:E26

Darquet AM, Rangara R, Kreiss P, Schwartz B, Naimi S, Delaère P, Crouzet J, Scherman D (1999) Minicircle: an improved DNA molecule for in vitro and in vivo gene transfer. Gene Ther 6:209–218

Dianov GL, Kuzminov AV, Mazin AV, Salganik RI (1991) Molecular mechanisms of deletion formation in Escherichia coli plasmids I. Deletion formation mediated by long direct repeats. Mol Gen Genet 228:153–159

Garmory HS, Brown KA, Titball RW (2003) DNA vaccines: improving expression of antigens. Genet Vaccines Ther 1:2

Gill DR, Smyth SE, Goddard CA, Pringle IA, Higgins CF, Colledge WH, Hyde SC (2001) Increased persistence of lung gene expression using plasmids containing the ubiquitin C or elongation factor 1α promoter. Gene Ther 8:1539–1546

Hofreuter D, Haas R (2002) Characterization of two cryptic Helicobacter pylori plasmids: a putative source for horizontal gene transfer and gene shuffling. J Bacteriol 184:2755–2766

Karlin S, Ost F (1985) Maximal segmental match length among random sequences from a finite alphabet. In: Cam LML, Olshen RA (eds) Proceedings of the Berkeley Conference in honor of Jerzy Neyman and Jack Kiefer. Wadsworth, Belmont, CA, pp 225–243

Kawe M, Horn U, Plückthun A (2009) Facile promoter deletion in Escherichia coli in response to leaky expression of very robust and benign proteins from common expression vectors. Microb Cell Fact 8:8

Larraín-Linton J, De la Iglesia R, Melo F, González B (2006) Molecular and population analyses of a recombination event in the catabolic plasmid pJP4. J Bacteriol 188:6793–6801

Levy J (2003) Avoidance of undesirable replication intermediates in plasmid propagation. US Patent 7390654

Lovett ST, Drapkin PT, Sutera VA Jr, Gluckman-Peskind TJ (1993) A sister-strand exchange mechanism for recA-independent deletion of repeated DNA sequences in Escherichia coli. Genetics 135:631–642

Lovett ST, Gluckman TJ, Simon PJ, Sutera VA Jr, Drapkin PT (1994) Recombination between repeats in Escherichia coli by a recA-independent, proximity-sensitive mechanism. Mol Gen Genet 245:294–300

Mairhofer J, Pfaffenzeller I, Merz D, Grabherr R (2008) A novel antibiotic free plasmid selection system: advances in safe and efficient DNA therapy. Biotechnol J 3:83–89

Mazin AV, Kuzminov AV, Dianov GL, Salganik RI (1991) Mechanisms of deletion formation in Escherichia coli plasmids II: Deletions mediated by short direct repeats. Mol Gen Genet 228:209–214

Morag AS, Saveson CJ, Lovett ST (1999) Expansion of DNA repeats in Escherichia coli: effects of recombination and replication functions. J Mol Biol 289:21–27

Moreno S, López-Fuertes L, Vila-Coro AJ, Sack F, Smith CA, Konig SA, Wittig B, Schroff M, Juhls C, Junghans C, Timón M (2004) DNA immunisation with minimalistic expression constructs. Vaccine 22:1709–1716

Oliveira PH, Lemos F, Monteiro GA, Prazeres DMF (2008) Recombination frequency in plasmid DNA containing direct repeats – predictive correlation with repeat and intervening sequence length. Plasmid 60:159–165

Oliveira PH, Prather KJ, Prazeres DMF, Monteiro GA (2009a) Structural instability of plasmid biopharmaceuticals: challenges and implications. Trends Biotechnol 27:503–511

Oliveira PH, Prazeres DMF, Monteiro GA (2009b) Deletion formation mutations in plasmid expression vectors are unfavored by runaway amplification conditions and differentially selected under kanamycin stress. J Biotechnol 143:231–238

Peeters BPH, de Boer JH, Bron S, Venema G (1988) Structural plasmid instability in Bacillus subtilis: effect of direct and inverted repeats. Mol Gen Genet 212:450–458

Ribeiro SC, Oliveira PH, Prazeres DMF, Monteiro GA (2008) High frequency plasmid recombination mediated by 28-bp direct repeats. Mol Biotechnol 40:252–260

Richardson PT, Park SF (1997) Integration of heterologous plasmid DNA into multiple sites on the genome of Campylobacter coli following natural transformation. J Bacteriol 179:1809–1812

Rocha EPC (2003) An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. Genome Res 13:1123–1132

Sandegren L, Andersson DI (2009) Bacterial gene amplification: implications for the evolution of antibiotic resistance. Nat Rev Microbiol 7:578–588

Song JY, Choi SH, Byun EY, Lee SG, Park YH, Park SG, Lee SK, Kim KM, Park JU, Kang HL, Baik SC, Lee WK, Cho MJ, Youn HS, Ko GH, Bae DW, Rhee KH (2003) Characterization of a small cryptic plasmid, pHP51, from a Korean isolate of strain 51 of Helicobacter pylori. Plasmid 50:145–151

Sun Y, He X, Liang J, Zhou X, Deng Z (2009) Analysis of functions in plasmid pHZ1358 influencing its genetic and structural stability in Streptomyces lividans 1326. Appl Microbiol Biotechnol 82:303–310

Wang Z, Troilo PJ, Wang X, Griffiths TG, Pacchione SJ, Barnum AB, Harper LB, Pauley CJ, Niu Z, Denisova L, Follmer TT, Rizzuto G, Ciliberto G, Fattori E, Monica NL, Manam S, Ledwith BJ (2004) Detection of integration of plasmid DNA into host genomic DNA following intramuscular injection and electroporation. Gene Ther 11:711–721

Weber PC, Palchaudhuri S (1986) An inverted repeat sequence of the IncFI plasmid ColV2-K94 increases multimerization-mediated plasmid instability. J Gen Microbiol 132:989–995