SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
PCR-BASED TECHNIQUES
DNA SEQUENCING

Received
28 January 2014

Accepted
6 May 2014

Published
23 May 2014

Correspondence and requests for materials should be addressed to B.Ü. (ulker@uni-bonn.de)

* Current address: Institute for Microbiology, Department of Biology, Heinrich-Heine University Düsseldorf, 40225 Düsseldorf, Germany.

PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications

Carl Maximilian Hommelsheim, Lamprinos Frantzeskakis*, Mengmeng Huang & Bekir Ülker

Plant Molecular Engineering Group, IZMB (Institute of Cellular and Molecular Botany), University of Bonn, Kirschallee 1, 53115 Bonn.

Designer transcription-activator like effectors (TALEs) is a promising technology and made it possible to edit genomes with higher specificity. Such specific engineering and gene regulation technologies are also being developed using RNA-binding proteins like PUFs and PPRs. The main feature of TALEs, PUFs and PPRs is their repetitive DNA/RNA-binding domains which have single nucleotide binding specificity. Available kits today allow researchers to assemble these repetitive domains in any combination they desire when generating TALEs for gene targeting and editing. However, PCR amplifications of such repetitive DNAs are highly problematic as these mostly fail, generating undesired artifact products or deletions. Here we describe the molecular mechanisms leading to these artifacts. We tested our models also in plasmid templates containing one copy versus two copies of GFP-coding sequence arranged as either direct or inverted repeats. Some limited solutions in amplifying repetitive DNA regions are described.

Breaking the code of DNA binding by TAL (transcription activator-like) effectors led to a new, robust technology for genome engineering[1]. Similarly, the code of RNA recognition by PUF (Pumilio and FBF homology repeat) proteins has been discovered[2,3]. It is also expected that the complete code of RNA recognition by the PPR (pentatricopeptide repeat) proteins is to be discovered in the near future[4–6]. All of these proteins contain repeats of variable nucleic acid-binding domains. Engineering and defined organization of these repeats leads to predictable and highly specific nucleic acid binding. Such modular recognition of nucleic acids by PUF, TALE and PPR proteins offers numerous possibilities for gene editing and genome engineering in prokaryotes and eukaryotes[7,8].

Natural TALE DNA-binding repeats are in the size of approximately 100 nt (coding for 33 to 34 amino acids). The variation in the amino acid sequence of the TALE repeats occurs mainly at positions 12 and 13[9] and is designated as RVDs (Repeat Variable Diresidues). Subsequently, it was determined that each RVD recognizes a single base in the target sequence. The binding preferences of individual RVDs were established experimentally and computationally[1,10]. Following the initial discovery of TALE DNA-binding code, plasmid kits containing all the necessary parts to generate desired DNA-binding TALEs[11,12] made it easy for many labs around the world to use this technology without having to diverge from their budgets or acquire any specialized training. As one of these labs, we have incorporated TALEs in our research. We used the Golden Gate TALE assembly kit[11] and successfully generated various TALE DNA-binding domains with 12 to 18 bp nucleotide binding specificity in the pTAL2 vector supplied with the kit. Here we report difficulties of PCR amplification of repetitive DNA sequences from pure plasmid templates. Such difficulties were also observed by others including the developers of the Golden Gate Kit[11]. Since PCRs are only performed to identifiy succesfully assembled TALe repeats and all the assembly and downstream cloning procedures were only designed for restriction based or Gateway cloning, the designers of the kit probably did not have to solve the PCR amplification problems. They even see PCR generated laddering as an indicator of succesful assembly. It was speculated that the repetitive nature of the amplified fragment is likely the reason. However, in order to use the TALe/TALeN technology to its full potential, it needs to be compatible with the powerful PCR-based cloning methods. Until now, an effective solution for this problem has not been found and the molecular mechanism which leads to the generation of these artifacts remains elusive. We suggest models for why PCR amplifications across repetitive DNAs fail. We also describe how such difficulties could be reduced using certain polymerases, conditions, primers and compounds.

Results

In order to generate a fusion construct for plant transformation and *in-planta* expression, we needed to sub-clone the assembled TALE DNA-binding domains in the pTAL2 vector (Golden Gate Kit, Addgene) targeting 12 to 18 bp DNA region in GFP-coding sequence for specific binding (Supplementary Fig. 2) to a binary T-DNA plant transformation vector (pBAtS1, T. Berson and B. Ülker, unpublished, Supplementary Fig. 3). Due to the absence of convenient restriction sites flanking this fragment in the pTAL2 and desired in-frame fusions to our test proteins, we decided to PCR amplify these fragments using primers 534 and 535 which flank the repeats (See Supplementary Table 1 for primers). However, amplification of the correct sized fragments failed. Instead, PCR amplification products ranging from 350 bp to several thousand base pairs were typically observed (Fig. 1). Even the use of a non-proofreading but robust DNA polymerase, *Taq* (NE Biolabs) failed to amplify only the desired fragment (Fig. 1). A laddering of amplification products below the expected size and also a heavy smear above it were observed. Interestingly, the smallest fragment for the laddering always started from 350 bp and appeared to be in the increments of 100 bp, corresponding to the approximate individual repeat size (80–96 bp).

To solve this issue, we tried to optimize the PCR mix. We followed the manufacturers' troubleshooting suggestions for each polymerase separately, which included the addition of DMSO, the optimization of $MgCl_2$ content or using special buffers for GC-rich templates. These compounds aim to either reduce the formation of secondary structures of the DNA template or increase the activity of the enzyme. None of these alterations significantly improved the amount of specific product (data not shown). Similarly, changes in cycling conditions, including high primer annealing temperatures to decrease unspecific primer-binding were not helpful. We could exclude that the pTAL2 vector backbone is part of the reason because the cut and gel-isolated fragment containing the TALE DNA-binding repeats also caused a similar laddering effect (Supplementary Fig. 1C). Likewise, we have observed such a laddering effect using other vector backbones containing similar TALE DNA-binding repeats (Supplementary Fig. 1A and B). We could also exclude that primers are the cause of these artifacts (Supplementary Fig. 1A, B and C).

Sequencing of artifact fragments were informative for generating hypotheses. In order to better understand why such laddering is generated, we isolated several such distinct fragments from agarose gels after electrophoretic size separation and cloned them into pTOPO® vectors (Fig. 2).

To our surprise, sequencing of these fragments in the TOPO vector was problem-free and the results were informative. Sequencing of two independent clones of the shortest fragment (350 bp, clones 1 and 2) resulted in two highly similar sequences which varied only by a few base pairs (see supplementary sequence files). Alignment of these sequences to the amplified TALE repeat sequence in the pTAL2 vector indicated that these fragments contain the N- and C-terminal domain of the TALE factor with only one DNA-binding repeat. Interestingly, this repeat was a hybrid containing the first repeat RVD (HD1) and the last RVD (LR-HD) (Fig. 2B). Furthermore, the sequence of the second hybrid repeat was somewhat different but was also a hybrid of the first repeat and the last RVD. Therefore, the shortest fragment (350 bp) that we commonly see in our agarose gels after running the samples of TALE PCRs are the result of N+C-terminal end (250 bp long) separated by one TALE DNA-binding repeat of ~100 bp. These results indicate that during polymerase chain reaction, DNA polymerase skipped the 11 other RVDs in the fragment located between the N- and C- terminal ends. Since these repeats are organized as direct repeats and not inverted ones, it is hard to imagine how the middle repeats are excluded from PCR amplification. Similarly, we did not find any complementary sequences flanking of any of these repeats which might lead to hairpin structures and deletions of hairpins by DNA polymerase template slipping[13,14].

Further analysis of four independent clones (3, 4, 5 and 6) containing a 450 bp PCR fragment demonstrated that these clones contained only two TALE DNA-binding repeats between the N-and

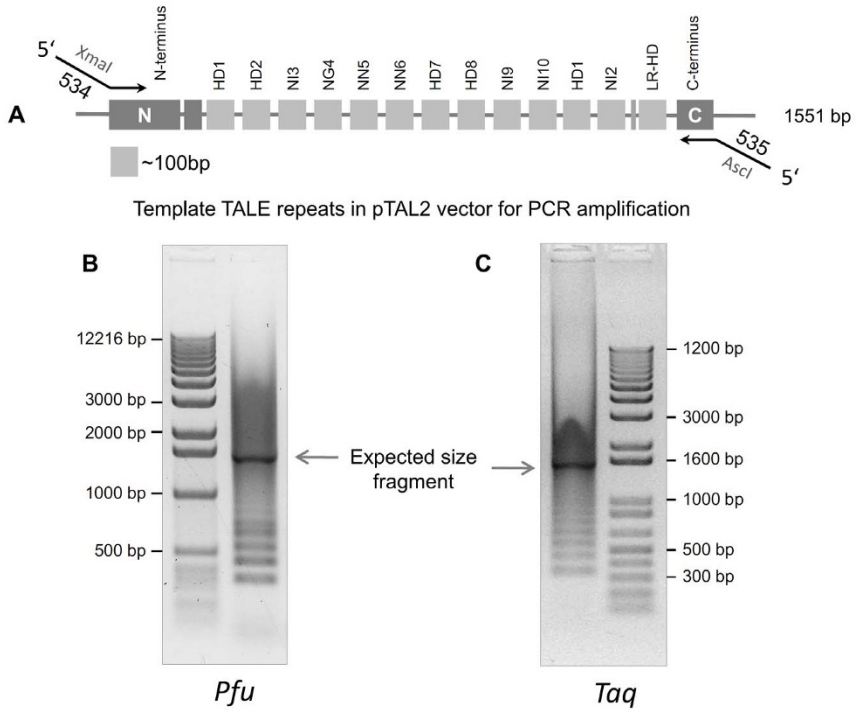

Figure 1 | PCR fragments generated upon a typical PCR amplification from the pTAL2 vector with 12.5 TALE DNA-binding repeats. Plasmid map is shown in Supplementary Fig. 3. Proofreading *Pfu* polymerase (Bioline) and normal *Taq* polymerase (NE Biolabs) were used in PCR amplification. PCR conditions are described in the supplementary material.

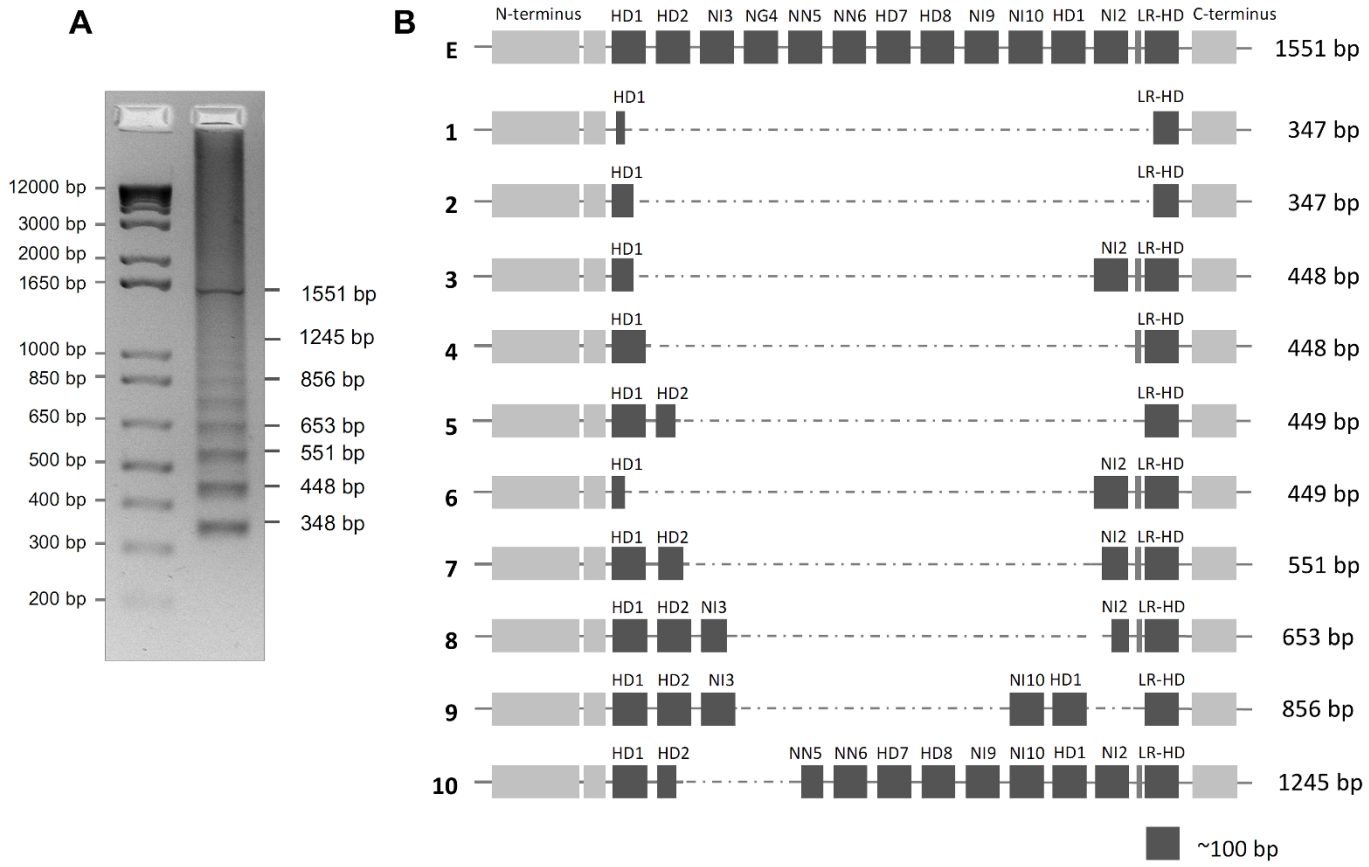

Figure 2 | Cloning and sequencing of PCR artifacts. (A) The TALE repeats were amplified from the pTAL2 vector using primers 534 and 535. The indicated size of PCR amplification fragments were gel-isolated and cloned into pTOPO® vectors from Life Technologies. These clones were sequenced using M13 reverse and forward primers. (B) The alignments of sequencing results to the expected TALE repeat organization are shown graphically. E: expected residue organization and truncated versions. The sizes of the sequence fragments are indicated on the graphical display. Corresponsing sequences are given in the supplementary material.

C-terminal ends. Similar to the 350 bp fragment, these repeats also contained hybrid repeats. In clone 3, a hybrid of HD1 and NI2 and the normal LR-HD RVDs were present between the N- and C-terminal ends (Fig. 2). Clone 4 had a similar organization but the hybrid RVD containing HD1 and NI2 was different. This hybrid contained mostly the sequence from the HD1 and less of the NI2 (Fig. 2). In clone 6, the proportions of the HD1 and NI2 were opposite and the hybrid RVD contained mostly the sequence of the NI2 but less of the HD1 sequence. Clone 5 contained yet another combination of TALE repeats with hybrid RVDs. In this clone, beside the normal LR-HD, a hybrid of repeat of HD1 and HD2 RVDs were present. In all four clones, therefore, only two TALE repeats were present and the rest of 10 repeats were skipped. The sequence of clone 8 containing a 650 bp PCR fragment showed that this clone contained between the normal repeats HD1, HD2 and LR-HD RVDs a hybrid repeat containing parts of the NI3 and NI2. Therefore in this fragment, there were only three TALE repeats and the remaining nine repeats were skipped. These results again demonstrated that the polymerase only amplified two to three extreme TALE repeat domains or their hybrids and skipped 9–10 TALE repeats between them. Generation of hybrid TALE DNA-binding repeats with different RVDs strongly suggests that sequence homology is part of the mechanism of these artifacts. However it is not clear how the polymerase jumps between these repeats.

Analysis of clone 9 containing a 856 bp PCR fragment indicated that deletions can be rather complex and not limited to the TALE repeats in the middle section (Fig. 2). This clone contained the TALE repeats with NI10 and HD1 RVDs but lacked the flanking middle repeats. Again, generation of hybrid repeats were apparent. Similarly, sequencing the 1245 bp DNA fragment in clone 10 showed that it was missing NI3 and NG4 containing repeats and had a hybrid repeat containing HD2 and NN5. Such rearrangements generated by PCR artifacts raise doubts of whether the correct size PCR fragments are reliable as they can contain any mixture of repeats in the range of expected fragment size. Furthermore, the isolation of a desired fragment from an agarose gel containing such a laddering effect is very difficult and requires extreme precision.

The proposed model. These results and the sequence data obtained from the PCR fragments allowed us to draw models explaining how these artifacts are likely generated (Fig. 3 and 4). We speculate that the polymerase does not jump between these repeats but instead the polymerization function of the DNA polymerases is hindered due to complex annealing of DNA fragments containing such repetitive DNAs during repeated denaturing and re-annealing cycles. It is proposed that polymerases can dissociate from template when they encounter hairpins ahead of their polymerization direction[14]. In the system we study, the direct TALE DNA-binding repeats cannot form hairpins and there are no detectable hairpin-causing sequences flanking the repeats, however, out of register annealing of template and later artifact products could generate such paired DNA strands which later might cease polymerization reactions and lead to disengagement of DNA polymerase from template. We believe that fragments resulting from incomplete amplification that arise from such polymerase disengagements are acting as mega primers and anneal out of register but sequence specifically to various positions

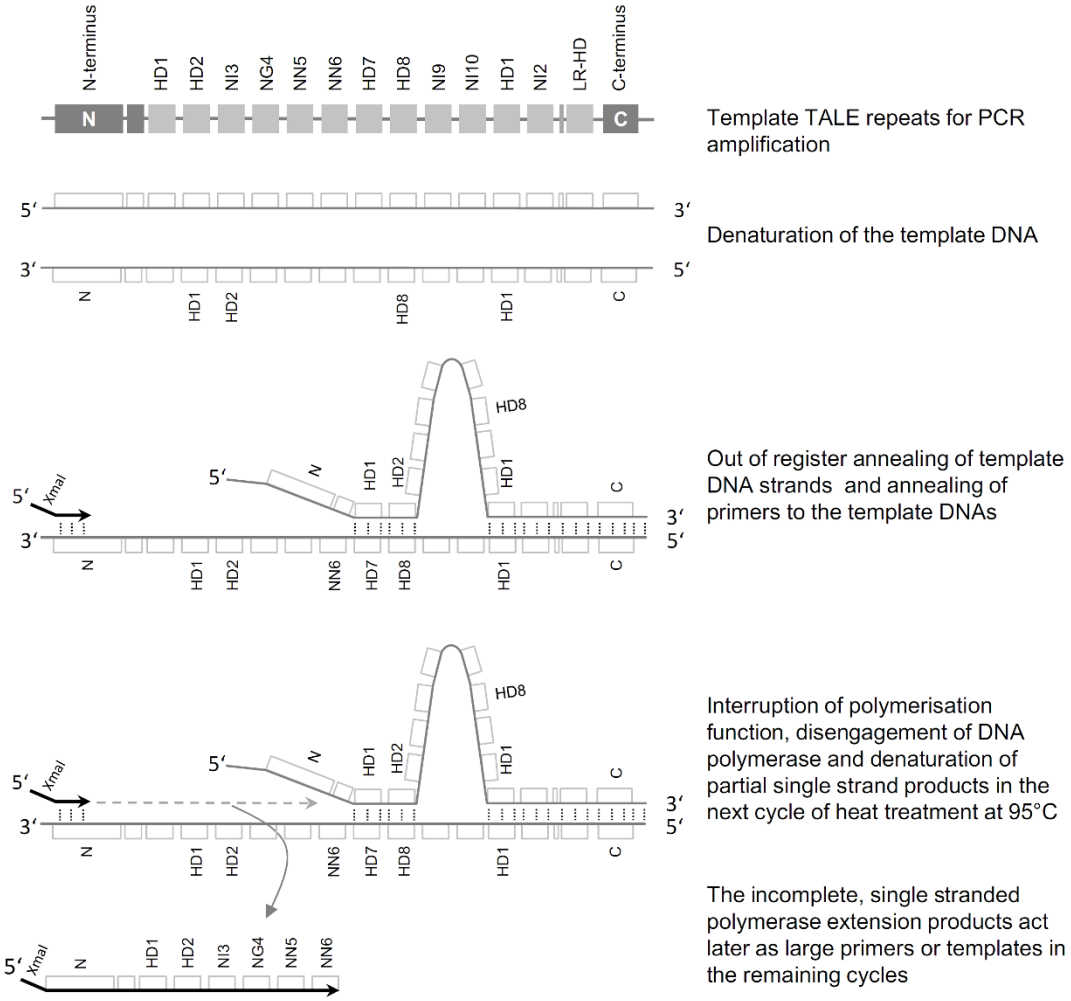

Figure 3 | Proposed model explaining how the PCR artifacts are generated. Polymerization function is hindered by out of register template strand or PCR product alignments ahead of the polymerase. These partially amplified fragments act as mega primers in the following cycles and could bind to the template or other impartially generated fragments in or out of register, leading to the observed laddering effect.

in the repetitive regions in subsequent cycles (Fig. 3 and 4). Such annealed fragments are later filled in by the DNA polymerase, generating various artifact products and new templates for amplification. Since amplification of shorter fragments is more effective during the polymerization reactions, distinct sizes of shorter fragments accumulate along with the correct size of the fragment.

Incomplete single stranded fragments can also be generated due to other reasons such as polymerase exhaustion and insufficient time to complete the extension of a fragment due to repeated heating cycles. In other words, at any given time, a polymerase could be in the middle of its extension of a certain fragment and an increase of temperature to 95°C would disengage it from the template and from the incompletely extended fragment. Depending on the end of the sequence and the location of binding sites along the template, these fragments could re-anneal to the template out of register in the following cycles and generate these artifacts.

Why are fragments that larger than the expected full length PCR fragments also generated? By extending the model in Fig. 3 to various constellations of amplification, denaturation and out of register annealing of the complete and incomplete amplification products as well as the template DNA strands, we can clearly predict how such larger fragments are generated and why they are so abundant (Fig. 4).

Multiple incomplete strand displacements due to polymerase disengagements are also likely to happen even when using DNA templates with non-repetitive sequences. However, in these cases, since the displaced and incomplete fragments can only bind in register to one location along the template, they cannot generate artifacts.

Searching for polymerases that perform better on repetitive DNA templates. The artifacts observed containing repetitive DNAs uncovered many unnoticed polymerization intermediates, defects and byproducts. The repetitive DNA sequences hence offer good test material for characterizing and selecting DNA polymerases with unrecognized and superior abilities. Therefore we either extended our analyses of DNA polymerases beyond the ones we have already tested (Fig. 1) or followed recommended conditions and used special buffers to improve their accuracy in the amplification of such repetitive DNA sequences.

We have selected some of the most robust DNA polymerases available in the market including *Q5* (NE Biolabs), *Phusion* (NE Biolabs), *Phusion HotStartFlex* (NE Biolabs), *PrimeSTAR HS* (TAKARA), *PrimeSTAR GXL* (TAKARA), *PrimeSTAR Max* (TAKARA), *AccuPrime Pfx* (Life Technologies). These DNA polymerases are advertised as being able to amplify difficult templates. We tested them under the supplier's recommended conditions in the amplification of TALE DNA-binding repeats from the pTAL2 vector. None of these polymerases eliminated the artifacts but *PrimeSTAR* and *AccuPrime Pfx* performed better since the concentration of correct-sized fragment was higher compared to the other polymerases (Fig. 5). These results are reliable as we repeatedly saw these polymerases performing

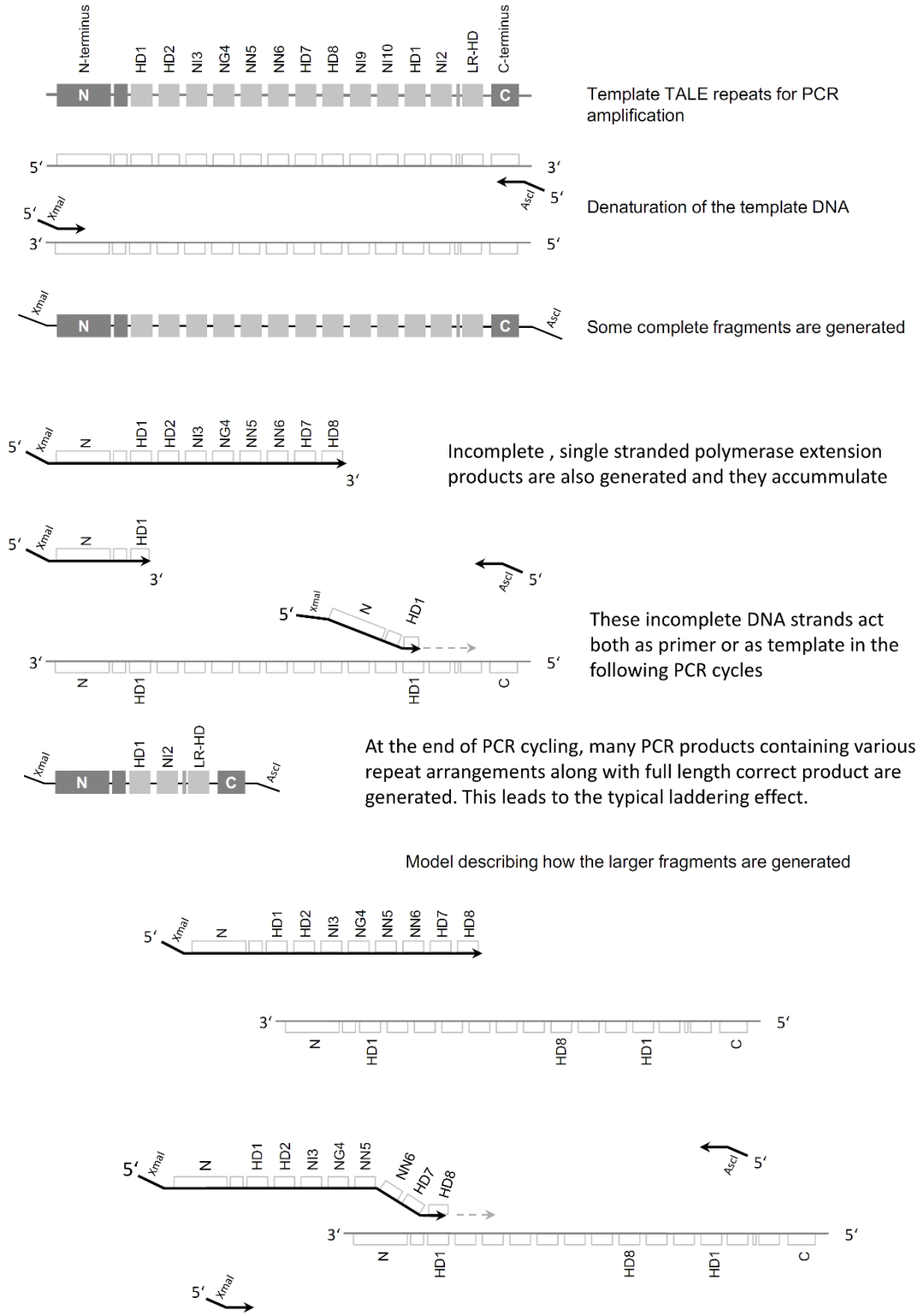

Figure 4 | Extension of proposed model explaining how the PCR fragments of discrete sizes, smaller and larger than the expected fragments, are generated.

somewhat better (data not shown). Cloning of the gel isolated 1551 bp fragment and subsequent sequencing showed that there was no error in the amplification.

Testing DNA polymerases with strand displacement activity. Although these results were encouraging and we could go on with our planned project, we were still interested in understanding why these polymerases were unable to amplify such a difficult template and whether we could eliminate the laddering effect. From our model, we predicted that DNA polymerases with strand displacement abilities should not generate partial amplification products that might act as mega primers in following cycles since they should be able to displace any paired DNA strands ahead of the polymerization direction. The specifications of the DNA polymerases indicate that the *Phusion*, *Q5*, and *Taq* polymerases from NE Biolabs do not have this ability (https://www.neb.com/tools-and-resources/selection-

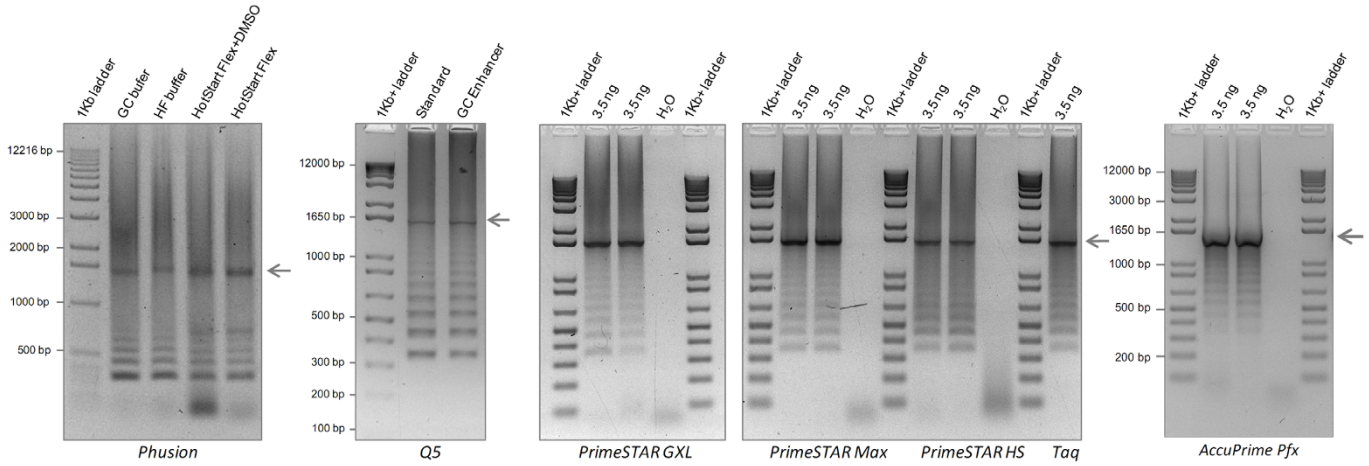

Figure 5 | Testing of various polymerases under supplier's recommended conditions in the amplification of TALE DNA-binding repeats from the pTAL2 vector. The arrows indicate the expected 1551 bp size of the amplification product. PCR conditions are given in the supplementary material.

charts/dna-polymerase-selection-chart) but this feature of the *PrimeSTAR* family TAKARA enzymes and the *AccuPrime Pfx* (Life Technologies) are not known. To further investigate the role of strand displacement ability of polymerases in eliminating these artifacts, we searched for only those DNA polymerases with such ability.

Deep-VentR DNA polymerase (NE Biolabs) was advertised as the polymerase with one of the highest strand displacement activity and is suitable for thermal cycling. This DNA polymerase was also able to amplify the expected fragment from such difficult templates containing TALE repeats despite a much lower yield (data not shown). However, the results were not very reproducible because sometimes the artifacts were mostly eliminated but other times they were similar to the results obtained with other tested polymerases. To understand why, several parameters were tested. We found that the amount of enzyme in the reaction is highly sensitive because two to three-fold changes in the amount of enzyme used resulted in completely different results ranging from no amplification, some but specific amplification to production of commonly observed artifacts with fragments smaller and larger than the expected fragment size (Fig. 6).

Could single strand DNA-binding proteins improve PCR amplification results? The use of single strand DNA-binding proteins (SSBPs) was shown to improve PCRs with difficult templates[14,15]. We reasoned that such proteins could reduce reannealing of DNA templates to each other by binding the template DNA and mega primers, thereby preventing out of register annealing of template and artifact strands. Primers are expected to be only marginally affected by SSBPs because they are less likely to be bound by SSBPs due to their short size and their excess amounts. Therefore we tested ET SSB (NE Biolabs, Cat.#: M2401S), a heat resistant single strand DNA-binding protein isolated from thermophilic bacteria along with *Deep-VentR* polymerase. We also used other proteins as controls including RNAseA, a heat stable RNA-binding protein and BSA which is not heat stable at temperatures above 60°C. All of these proteins are known to have a general positive charge, therefore they could interact with negatively charged single strand DNA backbones. Additionally, we tested whether the addition of random hexamers could prevent out of register annealing of incomplete PCR products to the template DNA strands.

Addition of ET SSB and to a minor degree RNAseA reduced some of the artifacts but BSA and random hexamers did not improve PCR amplification (Fig. 7). In the case of ET SSB, the reduction of artifact bands lead to the amplification of a higher amount of the correctly sized fragment. We then tested ET SSB along with two other polymerases that do not have strand displacement activity. As shown in

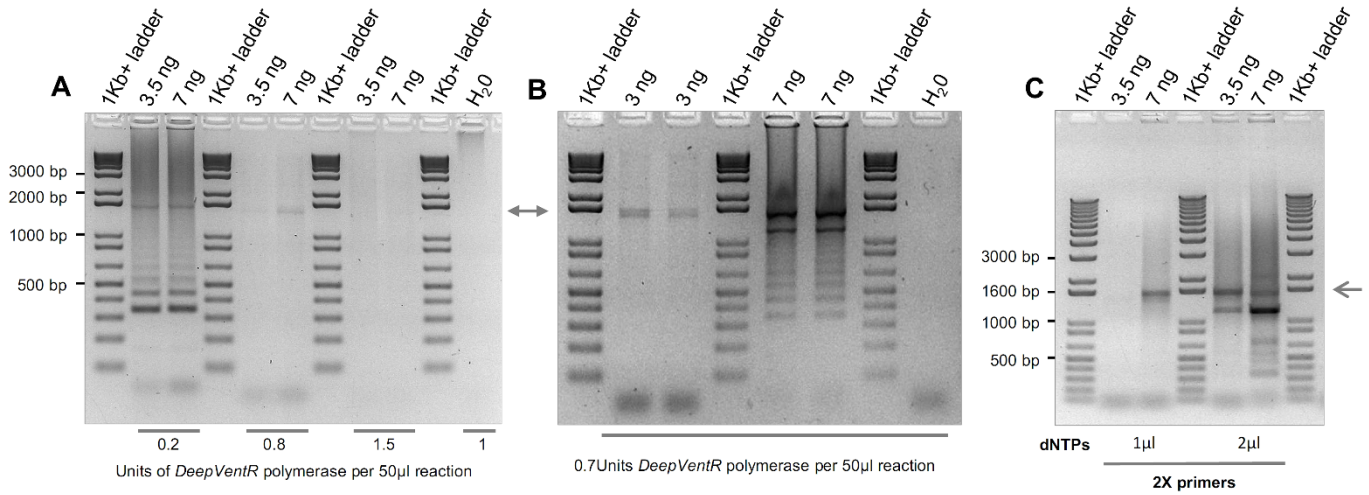

Figure 6 | Testing a DNA polymerase with high strand displacement activity in the amplification of the 12 TALE DNA-binding repeats in pTAL2 vector. Various parameters influencing the activity and the outcome of the PCR amplification of the *Deep-VentR* DNA polymerase are shown. The arrows indicate the expected size of the amplification products. PCR conditions are given in the supplementary material.

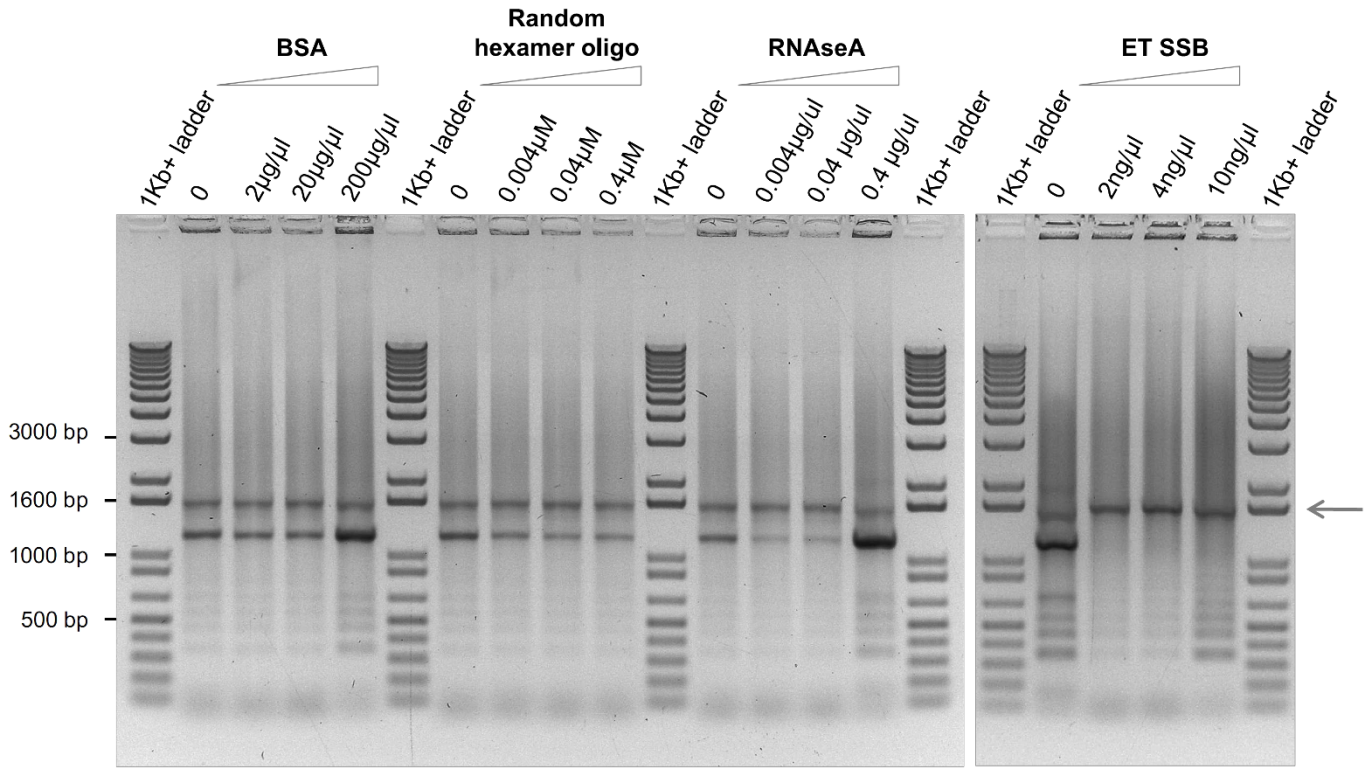

Figure 7 | ET SSB slightly improves the performance of *Deep-VentR* DNA polymerase in amplifying TALE DNA-binding repeats. The arrow indicates the expected size of amplification products. ET SSB, a heat resistant single strand DNA-binding protein isolated from thermophilic bacteria (NE Biolabs). PCR conditions are given in the supplementary material.

Figure 8, ET SSB had only a minor effect on PCR amplification but did not reduce the artifact bands.

The distance of primer annealing sites to the repetitive DNA has a role in artifact production. Using the models in Fig. 3 and 4, we speculated that increasing the length of the non-repetitive region in relation to the repetitive region in the amplified DNA should reduce the production of artifact products. This is because longer, non-repetitive DNA sequences would facilitate mostly in register re-annealing of template or unfinished polymerization products, which should eliminate most of the miss-annealed DNA ahead of polymerization. Therefore, we tested primers that anneal to 1544 bp upstream and 1152 bp downstream of the 1312 bp-long repetitive DNA-containing 12.5 TAL DNA-binding domains and compared these with the earlier combinations of primers that are annealing much closer (211 and 24 bp) to the repetitive DNA region (Fig. 9). Indeed primers binding further upstream and downstream from the repetitive DNA region amplified much more of the expected product and had fewer artifacts compared to primers annealing to flanking regions much closer to the repetitive DNA (Fig. 9).

Testing the generality of the model and simplifying the test system. In order to determine whether our model of how PCR artifacts with repetitive DNAs are generated is applicable to sequences other than the TAL DNA-binding repeats of 100 bp, we generated constructs carrying one or two copies of the complete GFP (717 bp) coding sequence in another vector backbone (pBasicS1, T. Berson and B. Ülker, unpublished) (Fig. 10A and supplementary Fig. 4). *GFP* clones containing two copies of *GFP* were arranged either as sense-sense (*GFPs* + *GFPs*) or as sense-antisense (*GFPs* + *GFPa*) orientations and were separated by a filler DNA-containing *NOS* terminator and *lac* promoter (Fig. 10A). We used primers 572 and 390, annealing 129 bp and 139 bp away from the repeated GFP-coding sequences, respectively. In the case of the single GFP-coding sequence, these primers should allow amplification of a 1658 bp fragment and in the case of two copies of the GFP-coding sequences, a fragment size of 2381 bp is expected (Fig. 10A). However, if incomplete polymerization products are generated and can reanneal the template DNA out of register, it is expected that constructs carrying two *GFP* sequences arranged as direct and inverted repeats should produce amplification artifacts. Indeed the double *GFP*-containing vectors but not the single copy *GFP*-containing vector produced artifact products shorter than the

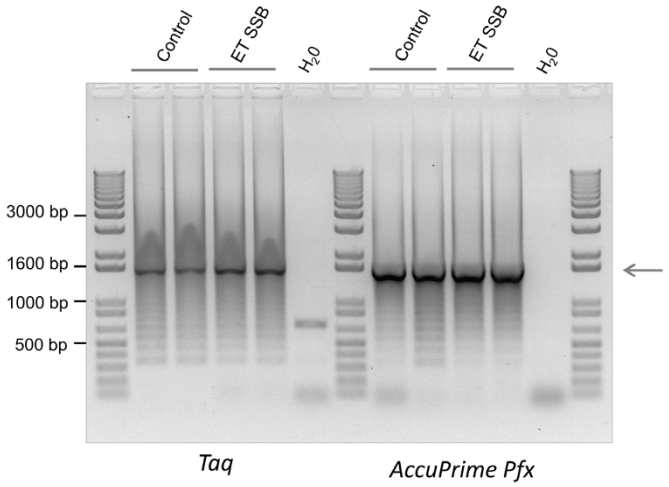

Figure 8 | Single strand binding protein, ET SSB, only has a minor effect on the reduction of artifacts. *Taq* (NE Biolabs) and *AccuPrime Pfx* (Life Technologies) DNA polymerases were used in amplification of TALE DNA repeats. The arrows indicate the expected size of the amplification products. PCR conditions are given in the supplementary material.

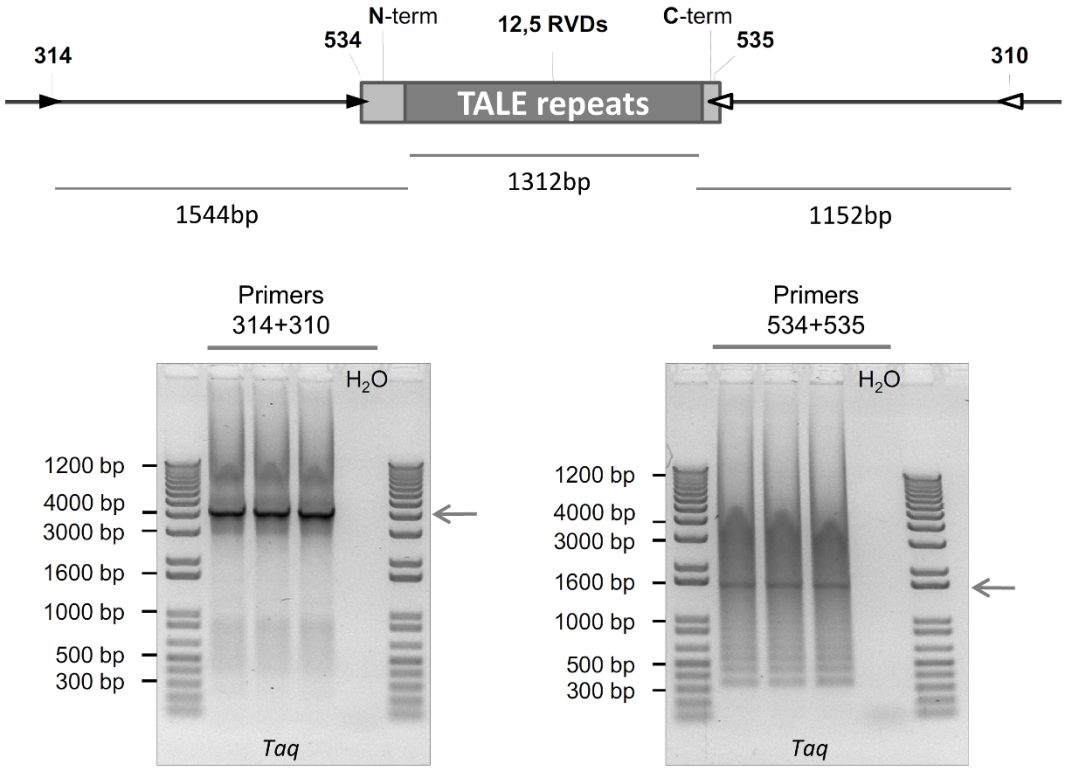

Figure 9 | Primers that anneal far away from the repetitive DNA perform much better in amplifying the desired product. *Taq* DNA polymerase (NE Biolabs) was used in the PCR amplification of the indicated region of the pdTALE12 plasmid. PCR conditions are given in the supplementary material.

expected size using the primers 572 and 390 with an extension time of 2 minutes and 40 seconds (Fig. 10B). We describe how these *GFP* PCR artifacts are likely generated with a model (Fig. 11).

Consistent with our models (Fig. 3, 4 and 11), PCR amplification of *GFP* copies arranged as direct repeats generated the expected 985 bp artifact fragment containing only one *GFP* flanked by binding sites for primers 572 and 390 (Fig. 11). Sequence analysis of this fragment also confirmed this predicted organization and sequence (See the supplementary sequence material). These data indicate that fragments between duplicated sequences as well as one of the duplicated sequences can be deleted during PCR amplification using common DNA polymerases.

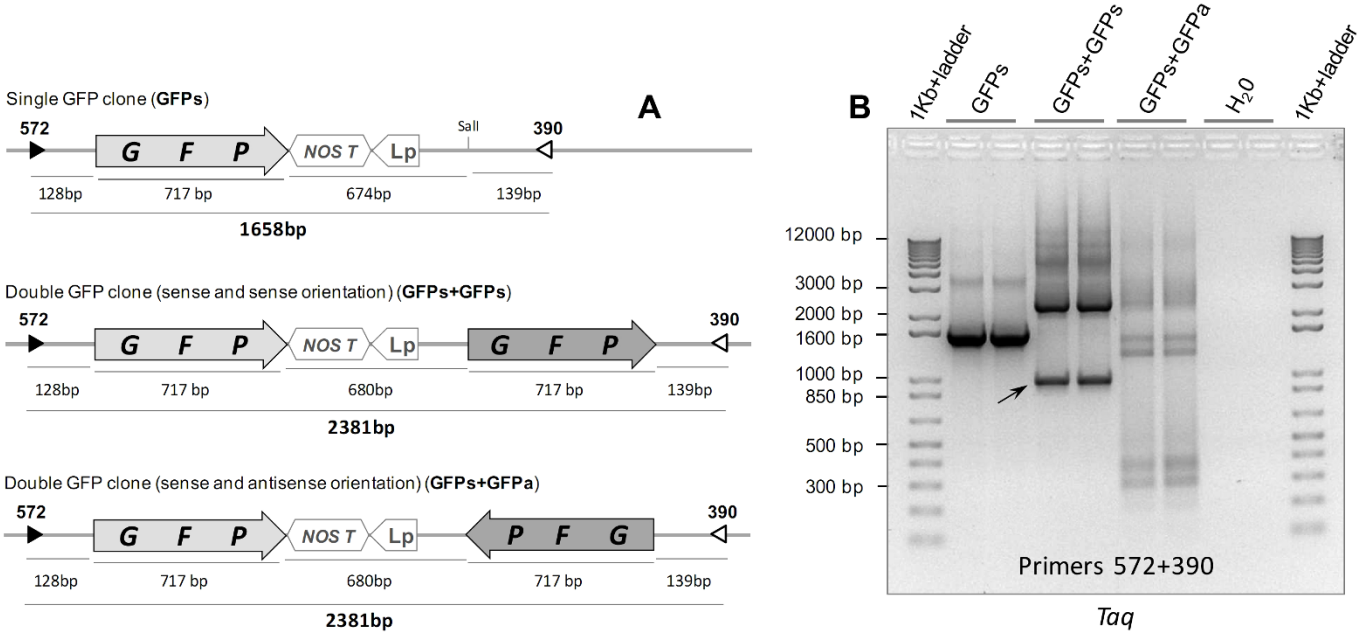

Figure 10 | Testing the generality of the model to other template DNAs with repetitive sequences. (A). GFP-coding sequences were cloned into the pBasicS1 vector and their integrity were checked by sequencing and restriction enzyme digestions. (B). PCR results obtained with primers 390 and 570. *Taq* DNA polymerase (NE Biolabs) was used in PCR amplification. The arrow indicates the sequenced artifact product which contained only one copy of the *GFP* lacking the filler sequence. See the supplementary material for PCR conditions.

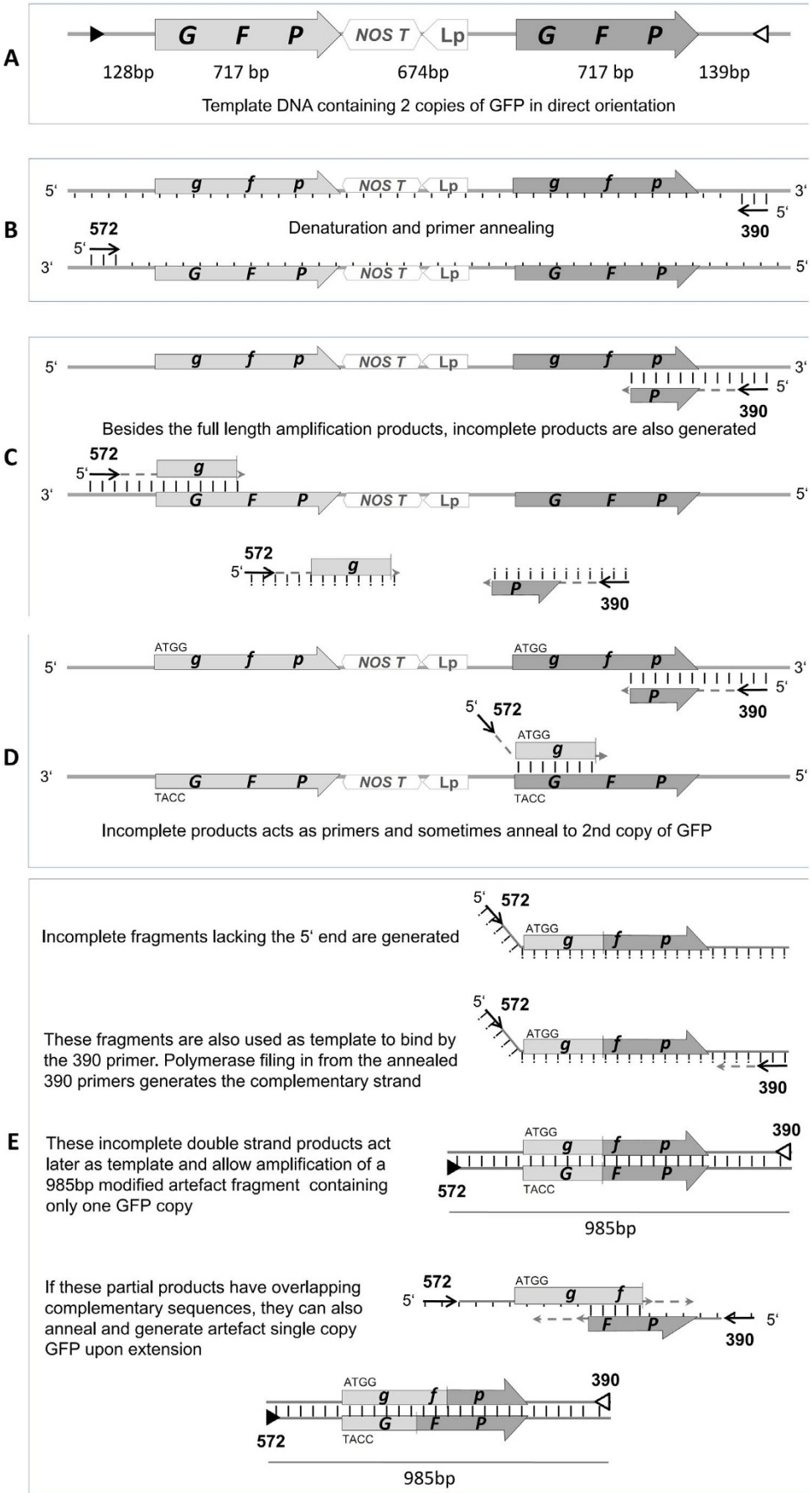

Figure 11 | The model explaining how *GFP* sequences arranged as direct repeats separated by a linker DNA generate artifact single copy *GFP*. See supplementary Fig. 5 for the model describing the generation of artifacts in inverted repeats.

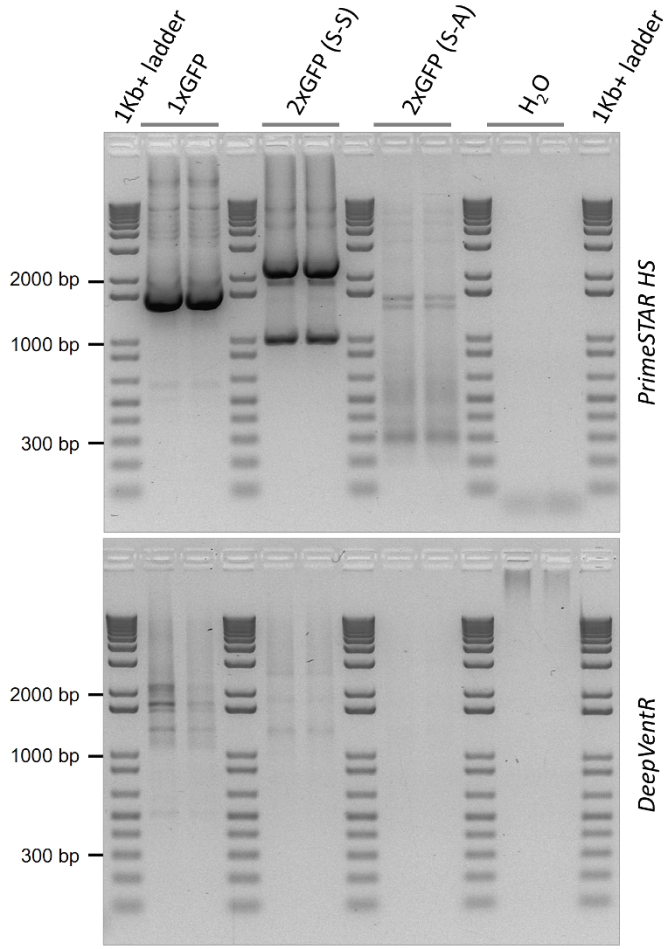

Figure 12 | Testing the generality of the model to other template DNAs with repetitive sequences and better-performing polymerases. See the supplementary material for PCR conditions.

In the arrangement of direct repeats of *GFP*, partial polymerization fragments starting from the primer 572 and ending before the first copy of the *GFP*, ending in the linker region or ending after the second *GFP* copy are all expected to reanneal to the template in register upon repeated cycles of denaturing and reannealing. Therefore these fragments should not generate any artifacts (Fig. 11). However, those incomplete fragments containing parts of the first copy or the second copy of the *GFP* can reanneal out of register to the second or the first copy of the *GFP* (Fig. 11). In the end, these should generate characteristic 985 bp and 3776 bp artifact fragments together with the primer 390. A similar situation is also expected from partial polymerization products starting from the direction of the 390 primer in that, besides the correct expected size fragment of 2381 bp, the artifact products of 985 bp and 3776 bp are expected to be generated.

The 985 bp artifact fragment could also be generated upon the annealing of two partial polymerization fragments, one of which containing some parts of the first copy of the *GFP* and the other one containing overlapping parts of the second copy of the *GFP* to each other instead of annealing to the template strand followed by filling in by polymerase.

Testing better-performing DNA polymerases on templates other than TALE DNA-binding repeats. Polymerases that were found to amplify the TALE repeats to some limited satisfaction failed to perform well on templates containing the *GFP* sequence either in direct or as indirect repeats (Fig. 12). However, as observed with TALE repeats, increased distance of the primer annealing site to the repetitive *GFP* sequence significantly improved the correct PCR product and reduced the artifacts (Fig. 13, primer combinations 239 + 284).

Discussion

Repeated DNA sequences are known to be significant components of genomes and they are highly dynamic[16]. Although most repeats are located in intergenic regions, some are also located in coding sequences or pseudogenes. The presence of a high amount of repetitive DNA in genomes, which can be as high as 80%, adds an enormous difficulty to assembling sequences in genome sequencing projects[17,18]. In many genome sequences available in databases, there are still many regions with no reliable sequence data due to the ambiguity of repetitive DNA. PCR-based amplification and sequencing techniques are used in aiding the assembly of the regions of genomes with no or low sequence information[19].

Repetitive DNA plays various roles in genomes ranging from genome organization, centromere assembly, telomere formation and related aging process, epigenetic modulation of associated loci, rapid genetic variation in times of stress and adaptive immune system in vertebrates and possibly speciation[20,21]. Repetitive DNAs were also involved in human diseases. For example, expansion of intragenic triplet repeats in humans is associated with various diseases, including Huntington chorea, myotonic dystrophy, synpolydactyly and fragile X syndrome[22–27]. Therefore repetitive sequences are of evolutionary, biological, biotechnological and medical significance and cannot be ignored[17].

Working with repetitive DNA has many challenges from cloning to maintaining in bacteria as they are frequently recombined or deleted. Amplifying fragments containing repetitive sequences arranged as tandem or inverted repeats or their combinations is difficult using PCR[28]. Besides, PCR amplification of highly similar sequences could generate PCR-mediated artifacts such as recombination or chimera formation[29,30]. Molecular mechanisms leading to generation of such artifacts are mostly unknown or ill-defined. Repetitive nature of TALE DNA-binding domains as well as the generation of artifacts from such repeats offers a suitable platform to gain insights into the mechanisms of how these PCR artifacts are generated. Similarly, the results obtained from these sequences could be informative on the natural events *in-vivo* leading to copy number variation of repetitive DNAs[31]. Therefore we systematically investigated several aspects of how artifacts are formed and how they can be eliminated.

Despite exhausting several possibilities, we could not find an ideal solution that allows error free amplification of repetitive DNA regions. Nevertheless, we describe many minor improvements when amplifying repetitive DNAs including the choice of suitable polymerases. Although direct repeats are somewhat amplifiable, inverted repeats appeared to be far more challenging. We show that there is a major need for better DNA polymerases with strand displacement activity. Perhaps the use of enzyme mixtures besides DNA polymerase that are involved in DNA replication *in-vivo* in thermophilic organisms could be tested. For example, we observed minor improvement with SSBPs which are also known to play a role in DNA replication *in-vivo* and improve PCRs[15]. Use of other DNA polymerase-associated proteins could possibly help to solve the bottleneck. Better performance of *AccuPrime Pfx* DNA polymerase compared to other DNA polymerases on repetitive DNA templates is also likely due to its mixture of other proprietary thermostable accessory proteins in this polymerase mix (http://tools.lifetechnologies.com/content/sfs/manuals/accuprimepfx_man.pdf)[15]. Similarly, reaction conditions where out of register annealing is prevented could be developed to improve correct PCR amplification. Furthermore, our results obtained from primers that bind far away from the repetitive DNA suggest that artifact products could be reduced by forcing the incomplete PCR fragments to anneal in register. PCR amplification

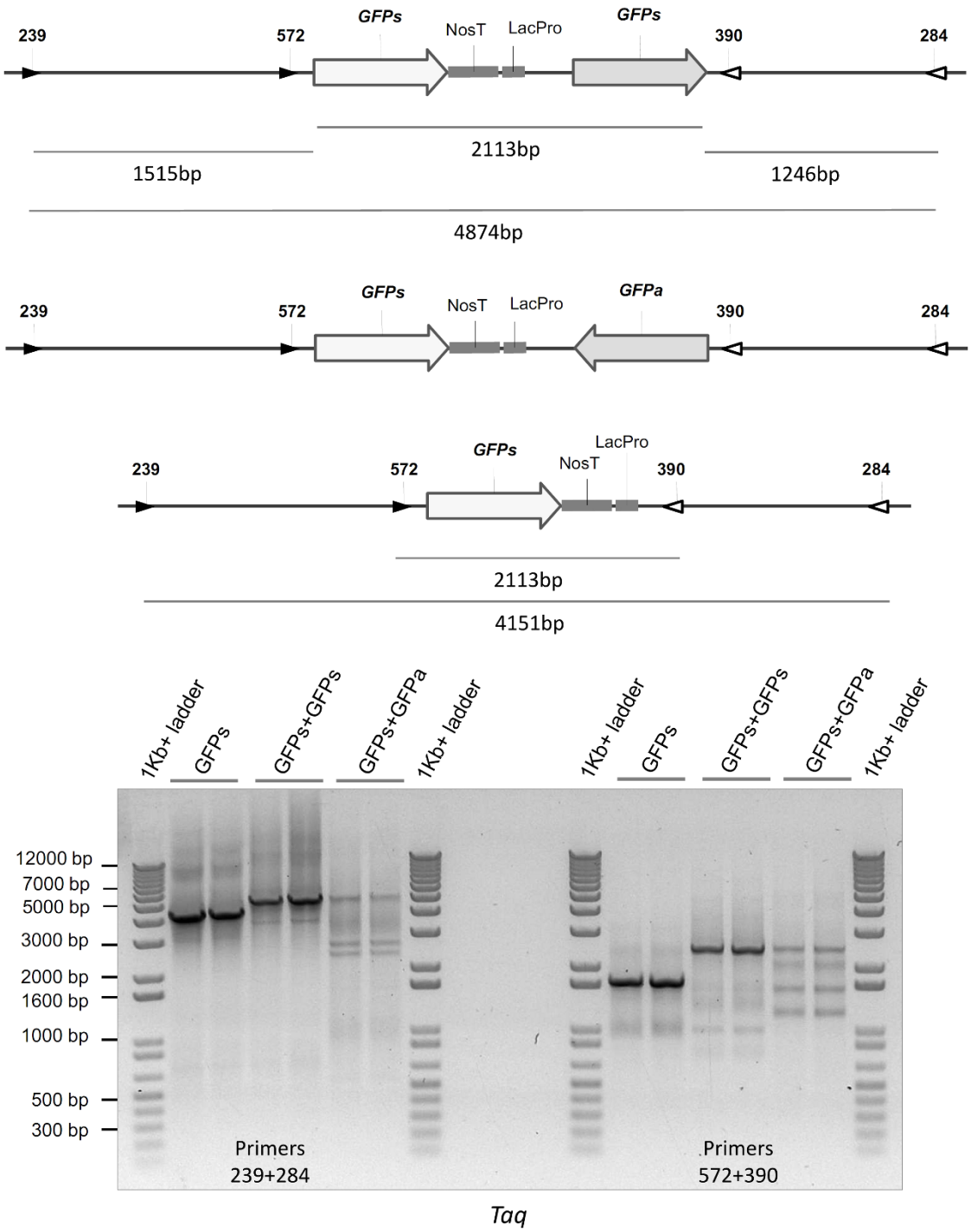

Figure 13 | Increasing the distance of the primer annealing site to the repetitive *GFP* reduces artifacts. Notice that even for the primer combinations 572 + 390, the amount of artifact products were reduced and the amount of the correctly sized fragment was increased when the extension time increased from 2 minutes to 5 minutes (compared to Fig. 10). See supplementary Fig. 4 for the plasmid maps used in PCR amplification.

by primers annealing to locations far away from the repeats was also reported to alleviate the PCR amplification problems from constructs containing TALE DNA-binding repeats[32]. However, the authors did not offer any explanation to why this leads to better amplification. We explain the improvement by the following. As the artifact product length increases, the time needed for annealing at 60°C also increases. Therefore it is expected that shorter fragments generated from primers binding closer to the repetitive region anneal more frequently and flexibly to different repeats. Since annealing starts not necessarily from the 3′ end but could happen throughout the entire length of the strand, increasing the length of non-repetitive region in PCR fragments would force correct and in-register annealing to the non-repetitive regions. This would result in reduced artifacts. Nevertheless, such a strategy was not an option for our cloning due to strict requirements for proper fusion construct.

Commercial companies specialized in synthesis of DNA sequences from scratch are also facing difficulties in synthesizing repetitive DNA regions and hence both the price and time to get the desired sequences are increasing. Codon alteration is an alternative route that can be taken to reduce difficulties in synthesis and amplification of repetitive DNAs. Although it is not experimented here, the use of the highly processive and by far the highest strand

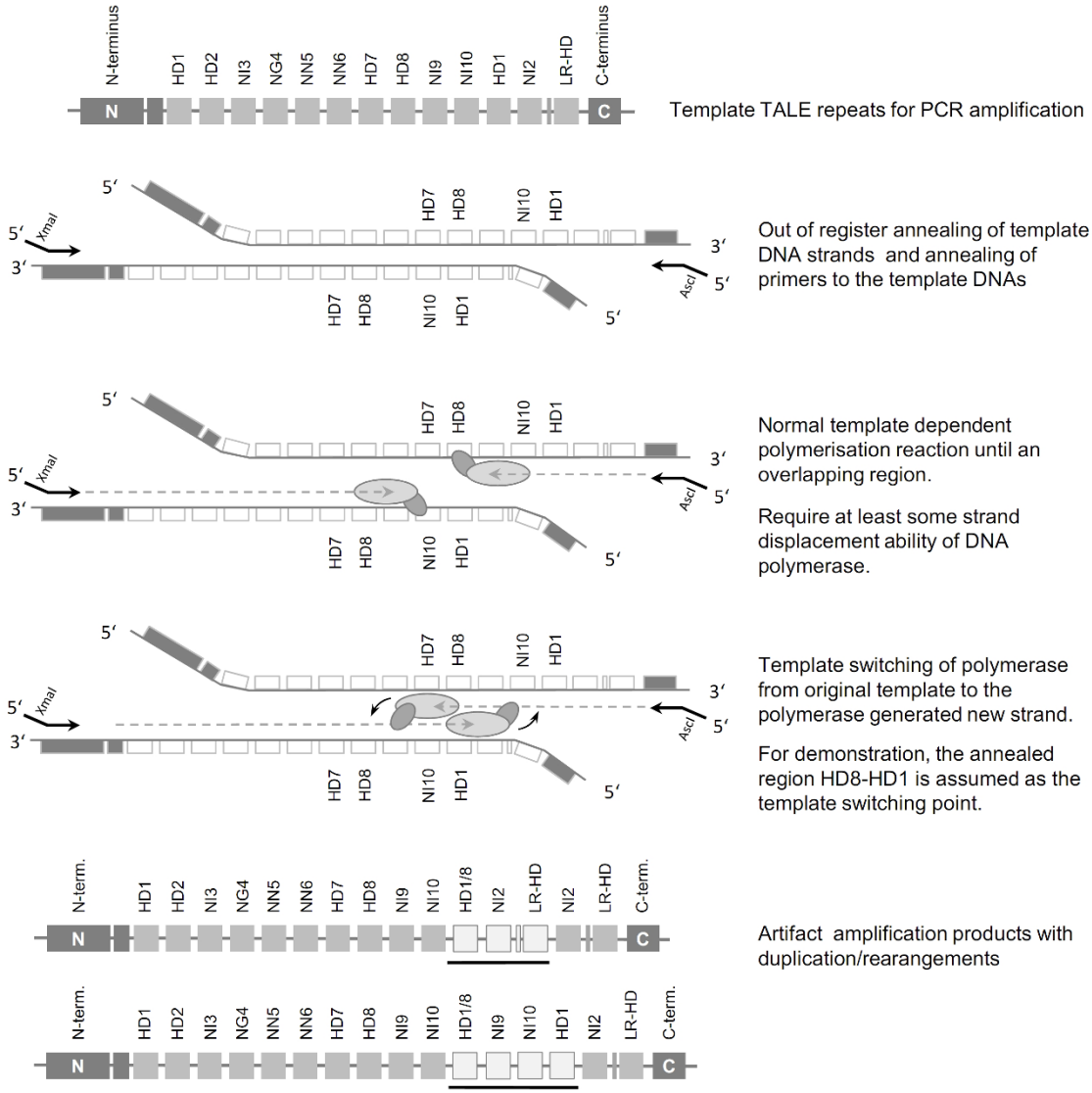

Figure 14 | Template switching by DNA polymerase might also be involved in the generation of PCR artifacts.

displacing enzyme Phi29 DNA polymerase might perform much better on such difficult templates[33]. Phi29 DNA polymerase is isolated from *Bacillus subtilis* phage Phi29, therefore is not suitable for thermal cycling but is a great tool in isothermal reactions. It replicates very long stretches of duplex DNA in the absence of any helicase because it possesses a strong strand displacement activity[34]. Its activity can also be modified by engineering its DNA-binding domain[35].

Although, we favor the hypothesis that truncated amplification products act as mega-primers and their misalignments on the template lead to artifacts, we cannot exclude other acting mechanisms such as template switching (Figure 14 and Supplementary Figure 6). One compelling piece of evidence indicating that template switching can generate recombinants was obtained from a single round of primer extension reaction in the absence of subsequent heat denaturation using *Thermus aquaticus* DNA polymerase I[36]. It is also possible that both mechanisms contribute to the observed phenomenon. Perhaps this was the reason why Deep-VentR DNA polymerase, which has high strand displacement ability, instead of reducing the production of partial amplification products acting as mega-primers, made only some minor improvements.

In conclusion, through the sequencing of PCR artifact products from two different repetitive DNA templates (12 × 100 bp as direct repeats and 2 × 717 bp as direct or inverted repeats) and the systematic analyses of selected DNA polymerases available in the market along with various other conditions, we were able to model the mechanisms leading to these artifacts. Despite better understanding these mechanisms, we could only make minor improvements, hence it was a great awareness for us how troublesome the repetitive sequences are. Thus, the data described in this paper should alert researchers utilizing PCR/RT-PCR techniques in diagnostics, forensics, fingerprinting, trans-splicing, homologous recombination, cloning, metagenome sequence analyses based on 16S rDNA, and genome sequence analyses making use of PCR that the presence of repetitive DNAs across the amplified regions in the templates can lead to artifacts and false conclusions. Similarly, samples containing mixtures of DNA from related organisms or multi-allelic templates can generate chimeric sequences. This is beautifully exemplified by the finding that artificial 16S rDNA sequences are being accumulated in public databases, suggesting the presence of non-existent organisms[37–39]. Our results might also stimulate development of better polymerases, kits and solutions to reduce or eliminate PCR amplification artifacts using templates with repetitive DNA.

Methods

Methods of generating TALE repeats, sequences obtained from gel isolated PCR artifact fragments and PCR conditions used for each figure in the main manuscript

and supplementary figures are given in full detail in the supplementary information. Likewise, supplementary information contains a table showing the names and sequences of primers used.

1. Moscou, M. J. & Bogdanove, A. J. A simple cipher governs DNA recognition by TAL effectors. *Science* **326**, 1501 (2009).
2. Filipovska, A., Razif, M. F., Nygard, K. K. & Rackham, O. A universal code for RNA recognition by PUF proteins. *Nat Chem Biol* **7**, 425–427 (2011).
3. Lu, G., Dolgner, S. J. & Hall, T. M. Understanding and engineering RNA sequence specificity of PUF proteins. *Curr Opin Struct Biol* **19**, 110–115 (2009).
4. Barkan, A. *et al.* A combinatorial amino Acid code for RNA recognition by pentatricopeptide repeat proteins. *PLoS Genet* **8**, e1002910 (2012).
5. Yagi, Y., Hayashi, S., Kobayashi, K., Hirayama, T. & Nakamura, T. Elucidation of the RNA recognition code for pentatricopeptide repeat proteins involved in organelle RNA editing in plants. *PLoS ONE* **8**, e57286 (2013).
6. Yin, P. *et al.* Structural basis for the modular recognition of single-stranded RNA by PPR proteins. *Nature* **504**, 168–171 (2013).
7. Filipovska, A. & Rackham, O. Modular recognition of nucleic acids by PUF, TALE and PPR proteins. *Mol Biosyst* **8**, 699–708 (2012).
8. Fichtner, F., Castellanos, R. U. & Ülker, B. Precision genetic modifications: a new era in molecular biology and crop improvement. *Planta* **239**, 921 (2014).
9. Hopkins, C. M., White, F. F., Choi, S. H., Guo, A. & Leach, J. E. A family of avirulence genes from Xanthomonas oryzae pv.oryzae. *Mol. Plant-Microbe Interact* **5**, 451–459 (1992).
10. Boch, J. *et al.* Breaking the code of DNA binding specificity of TAL-type III effectors. *Science* **326**, 1509–1512 (2009).
11. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res* **39**, e82 (2011).
12. Miller, J. C. *et al.* A TALE nuclease architecture for efficient genome editing. *Nature Biotechnol* **29**, 143–148 (2011).
13. Viguera, E., Canceill, D. & Ehrlich, S. D. In vitro replication slippage by DNA polymerases from thermophilic organisms. *J Mol Biol* **312**, 323–333 (2001).
14. Canceill, D., Viguera, E. & Ehrlich, S. D. Replication slippage of different DNA polymerases is inversely related to their strand displacement efficiency. *J Biol Chem* **274**, 27481–27490 (1999).
15. Rapley, R. Enhancing PCR amplification and sequencing using DNA-binding proteins. *Mol Biotechnol* **2**, 295–298 (1994).
16. Hartl, D. L. Molecular melodies in high and low C. *Nat Rev Genet* **1**, 145–149 (2000).
17. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet* **13**, 36–46 (2012).
18. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat Rev Genet* **12**, 363–376 (2011).
19. Korbel, J. O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
20. Jurka, J., Kapitonov, V. V., Kohany, O. & Jurka, M. V. Repetitive sequences in complex genomes: structure and evolution. *Annu Rev Genomics Hum Genet* **8**, 241–259 (2007).
21. Shapiro, J. A. & von Sternberg, R. Why repetitive DNA is essential to genome function. *Biol Rev Camb Philos Soc* **80**, 227–250 (2005).
22. Li, Y.-C., Korol, A. B., Fahima, T. & Nevo, E. Microsatellites Within Genes: Structure, Function, and Evolution. *Mol Biol Evol* **21**, 991–1007 (2004).
23. Jin, P., Alisch, R. S. & Warren, S. T. RNA and microRNAs in fragile X mental retardation. *Nat Cell Biol* **6**, 1048–1053 (2004).
24. Muragaki, Y., Mundlos, S., Upton, J. & Olsen, B. R. Altered growth and branching patterns in synpolydactyly caused by mutations in HOXD13. *Science* **272**, 548–551 (1996).
25. Scherzinger, E. *et al.* Huntingtin-encoded polyglutamine expansions form amyloid-like protein aggregates in vitro and in vivo. *Cell* **90**, 549–558 (1997).
26. Lee, J. E. & Cooper, T. A. Pathogenic mechanisms of myotonic dystrophy. *Biochem Soc Trans* **37**, 1281–1286 (2009).
27. Kunkel, T. A. Nucleotide repeats. Slippery DNA and diseases. *Nature* **365**, 207–208 (1993).
28. Sahdev, S., Saini, S., Tiwari, P., Saxena, S. & Singh Saini, K. Amplification of GC-rich genes by following a combination strategy of primer design, enhancers and modified PCR cycle conditions. *Mol Cell Probes* **21**, 303–307 (2007).
29. Brakenhoff, R. H., Schoenmakers, J. G. & Lubsen, N. H. Chimeric cDNA clones: a novel PCR artifact. *Nucleic Acids Res* **19**, 1949 (1991).
30. Meyerhans, A., Vartanian, J. P. & Wain-Hobson, S. DNA recombination during PCR. *Nucleic Acids Res* **18**, 1687–1691 (1990).
31. Dover, G. Slippery DNA runs on and on and on. *Nat Genet* **10**, 254–256 (1995).
32. Briggs, A. W. *et al.* Iterative capped assembly: rapid and scalable synthesis of repeat-module DNA such as TAL effectors from individual monomers. *Nucleic Acids Res* (2012).
33. Dean, F. B., Nelson, J. R., Giesler, T. L. & Lasken, R. S. Rapid amplification of plasmid and phage DNA using Phi 29 DNA polymerase and multiply-primed rolling circle amplification. *Genome Res* **11**, 1095–1099 (2001).
34. Blanco, L. *et al.* Highly efficient DNA synthesis by the phage phi 29 DNA polymerase. Symmetrical mode of DNA replication. *J Biol Chem* **264**, 8935–8940 (1989).
35. de Vega, M., Lazaro, J. M., Mencia, M., Blanco, L. & Salas, M. Improvement of phi29 DNA polymerase amplification performance by fusion of DNA binding motifs. *P Nat Acad Sci USA* **107**, 16506–16511 (2010).
36. Odelberg, S. J., Weiss, R. B., Hata, A. & White, R. Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. *Nucleic Acids Res* **23**, 2049–2057 (1995).
37. Hugenholtz, P. & Huber, T. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int J Syst Evol Microbiol* **53**, 289–293 (2003).
38. Berney, C., Fahrni, J. & Pawlowski, J. How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys. *BMC Biol* **2**, 13 (2004).
39. Wintzingerode, V. F., Göbel, U. B. & Stackebrandt, E. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* **21**, 213–229 (1997).

Acknowledgments

Research in the Ülker lab is supported in part by a grant from the German Research Foundation, DFG (UE 160/1-1) and basic funding from IZMB, University of Bonn. We thank members of the Ülker lab particularly Tobias Berson, Franziska Fichtner and Ahmet Bakirbas for reading the manuscript and helpful discussions. We also thank other members Reynel U. Castellanos and Désirée Waidman for sharing their unpublished data. We appreciate the help of Ms. Kaitlyn Jane Courville, Heinrich-Heine University Düsseldorf, in editing the article.

Author contributions

B.U., L.F. and C.M.H. conceived the study and planned the experiments. C.M.H. contributed most of the data presented in the manuscript, studied the literature and edited the article. L.F. made the initial observations, gel isolated and sequenced the artifact fragments, studied the literature and edited the article. M.H. generated the plasmids containing repetitive *GFP* fragments and edited the article. B.U. wrote the article, studied the literature and prepared the figures. All authors contributed to the discussions, read and approved the final version of the article.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Hommelsheim, C.M., Frantzeskakis, L., Huang, M.M. & Ülker, B. PCR amplification of repetitive DNA: a limitation to genome editing technologies and many other applications. *Sci. Rep.* **4**, 5052; DOI:10.1038/srep05052 (2014).