Premature termination codons signaled targeted repair of frameshift mutation by endogenous RNA-directed gene editing

Xiaolong Wang[*1], *Haibo Peng, Chunyan Li, Xuxiang Wang, Yalei Wang, Gang Chen, Jianye Zhang*

College of Life Sciences, Ocean University of China, Qingdao, 266003, P. R. China

Abstract

It has been presumed that most frameshifted protein coding genes yield truncated, non-functional, and may cause genetic diseases. At present, the mechanism for the identification and repair of frameshift mutation is unknown at the molecular level. Here we report that in *E. coli* a frameshift mutation was repaired spontaneously in a precisely targeted manner. If the premature termination codons (PTCs) were changed all into sense codons, the frameshift gene was expressed without repairing, proved that in the nonsense mRNAs it is the PTCs that signaled the identification and repair of the frameshift mutation. Genome resequencing showed that the revertants' genome is not more variable than the wild-type genome, while transcriptome analysis showed that genes, proteins and pathways for RNA editing, mismatch repair and homologous recombination were upregulated. We proposed a new model for autonomous gene repair, *RNA-directed Gene Editing* (RdGE): by mRNA surveillance the nonsense mRNAs were recognized through PTC signaling, edited and used to direct the repair of the faulty coding DNA sequence. The RdGE pathway is highly conserved among species, which may revise our view on central dogma, and change our perspective on molecular evolution, genetics and gene engineering.

[1] To whom correspondence should be addressed: *Xiaolong Wang, Ph.D., Department of Biotechnology, Ocean University of China, No. 5 Yushan Road, Qingdao, 266003, Shandong, P. R. China*, Tel: *0086-139-6969-3150*, E-mail: *Xiaolong@ouc.edu.cn*.

1. Introduction

It has been generally presumed that most frameshift mutations in protein-coding genes yield truncated, dysfunctional and potentially cytotoxic products [1], and cause death or genetic diseases [2-5]. A variety of molecular mechanisms have been found that involved in the repair of DNA damages [6], such as Nucleotide Excision Repair, Base Excision Repair, Mismatch Repair, Homologous Recombination and Non-Homologous End Joining. However, by far the mechanism for the repair of frameshift mutation is unknown at the molecular level. The back/reverse mutations phenomenon was discovered as early in the 1960s [7], but has been commonly explained: in the DNA replication process random mutation occurred naturally, a frameshift mutation happened restored by *random mutagenesis*.

Here we report that in *E. coli* a frameshifted gene was repaired autonomously, targetedly and initiatively, resulting in diversified variations. Here we propose a new model for the autonomous repair of frameshift mutations, *RNA-directed Gene Editing* (RdGE): in mRNA surveillance the nonsense mRNAs were recognized through PTC signaling, edited and used as a template to direct the repair of the faulty coding DNA sequence.

2. Materials and Methods

2.1 Frameshift mutagenesis and back mutations

A defective plasmid, *pBR322*-(*bla*-), containing a frameshifted β-lactamase gene (*bla*) was constructed by site-directed mutagenesis (Fig 1): in plasmid *pBR322*, base pair 136 (G/C) of the wild-type *bla* gene (*bla*+) was deleted by overlapping extension polymerase chain reaction (OE-PCR). The competent cells of *E. coli* strain *DH5α* was transformed with *pBR322*-(*bla*-), propagated in tetracycline-containing broth (TCB), plated on tetracycline-containing plates (TCPs) to count the total number of bacteria, and the transformed bacteria were plated on ampicillin-containing plates (ACPs) to screen for ampicillin resistant colonies (revertants) whose *bla* gene was repaired. The revertants (*bla**) screened were cultured in ampicillin-containing broth (ACB) to test

their survival rate, growth rate and ampicillin resistance ability, their plasmid DNA were extracted, and their *bla* gene were sequenced by Sanger sequencing.

2.2 Construction and expression of a PTC-free frameshifted bla

A *PTC-free frameshifted bla* (*bla#*) was constructed by a technique called *in vitro readthrough*: (1) the coding sequence of *bla#* was derived from *bla-* by replacing all nonsense codons each with a sense codon, according to the *readthrough rules* (Table 1), and adding a stop codon in the end; (2) the *bla#* gene was chemically synthesized by *Sangon Biotech, Co. Ltd* (*Shanghai*), inserted into the expression vector pET28a, and transformed into *E. coli BL21* competent cells. The transformed bacteria were grown on a kanamycin-containing plate (KCP), propagated in a kanamycin-containing broth (KCB), and plated on ACPs to screen for ampicillin resistant colonies, and then their plasmid DNA were extracted and Sanger sequenced; (3) The frameshifted BLA was expressed by IPTG induction, the total protein of *bla#* bacteria was extracted and analyzed by SDS-PAGE.

2.3 Genomes resequencing and variation analysis of the E. coli strains

The genomic DNA samples were extracted from wild-type (*bla+*) and revertant (*bla**) *E. coli* strains. Library preparation, genome sequencing and data analysis were conducted by a commercial service provided by *Novogene Co. Ltd*. The library was sequenced on an Illumina HiSeq 250PE platform and paired-end reads were obtained. Raw reads in fastq format were processed by removing adapter sequence and low quality reads. The clean reads with high quality were aligned to the reference genome of *E. coli* K12 MG1655 (NC_000913.3) to identify Single Nucleotide Polymorphisms (SNPs), Insertions/Deletions (InDels) and Structure Variations (SVs) of each sample. GATK2 software was used to perform SNP calling. Samtools v0.1.18 was used to sort the reads and reorder the bam alignments. Circos was used to display the coverage of the reads and the distribution of SNPs and InDels on the ring diagram. BreakDancer was used to detect structure variations.

2.4 Transcriptome analysis of the E. coli strains

Total RNA samples were extracted from *E. coli* strains, including wildtype (*bla+*), frameshift (*bla-*) and revertants (*bla**). Library preparation, RNA sequencing and data

analysis were conducted by *Novogene Co. Ltd*. After library preparation and cluster generation, the library was sequenced on an Illumina HiSeq platform. The paired-end raw reads were processed by removing adapter sequence and low quality reads. The clean reads with high quality were aligned to the reference genome, *E. coli* K12 MG1655 (NC_000913.3), and analyzed to identify differentially expressed genes and enriched GO terms and KEGG pathways.

(1) *Quantification of gene expression level*

Bowtie (v2-2.2.3) was used for index building and aligning the clean reads to the reference genome. HTSeq (v0.6.1) was used to count the numbers of reads mapped to each gene. And then FPKM (expected number of Fragments PerKilobase of transcript sequence per Millions base pairs sequenced) of each gene was calculated based on their lengths and reads counts.

(2) *Differential expression analysis*

Prior to differential expression gene analysis, for each strain the read counts were adjusted by the edgeR program through one scaling normalized factor. Differential expression analysis of two conditions was performed using the DEGSeq R package (1.20.0). The P values were adjusted using the Benjamini & Hochberg method. Corrected P-value of 0.005 and log2 (Fold change) of 1 were set as the threshold for significantly differential expression.

(3) *GO and KEGG enrichment analysis of differentially expressed genes*

Gene Ontology (GO) enrichment analysis of differential expression genes was implemented by the GOseq R package, in which gene length bias was corrected. GO terms with corrected P-value less than 0.05 were considered significantly enriched by differentially expressed genes (DEGs). KOBAS (KEGG Orthology Based Annotation System) 2.0 was used to test the statistical enrichment of DEGs in KEGG pathways (http://www.genome.jp/kegg/).

3. Results and Analysis

3.1 The growth of the frameshift mutant and the revertants

When a plasmid *pBR322* containing a wild-type *bla* gene was transformed into *E. coli* (strain DH5*α*, JM109 or BL21), the resulting wild-type (*bla*+) bacteria grow well on an ACP (Fig 2A, left). When one G/C base-pair was deleted in the upstream of the *bla* gene, seventeen nonsense codons emerged in the frameshifted coding sequence, and the active sites of BLA, including the substrate binding site, locate all in the downstream of the deletion. This frameshift mutation is a loss-of-function (*bla*-), and the transformed bacteria were expected cannot grow on ampicillin-containing media (ACM). However, repeatedly a few ampicillin-resistant colonies (revertants) were observed in the ACPs (Fig 2A, middle). The possibility of cross-contamination was excluded as no growth was observed in the blank controls (Fig 2A, right). In addition, the revertants (*bla**) are obviously different from the *bla*+ strain: although they can grow both in TCB and ACB, their growth rates are very much slower when compared with the *bla*+ strain: with 200 rpm shaking, it took up to 36 to 48 hours of culturing at 37°C to reach late log phase.

Hitherto, it seems that there is nothing unusual, as this is a back/reverse mutation, a well-known phenomenon which was discovered as early in the 1960s [7], and has been commonly explained by *random mutagenesis* (Fig 3A): in the DNA replication process random mutation occurred naturally, a few *lucky* cells survive when their *bla* gene happened restored, while most of the bacteria die without a back mutation.

3.2 A reverse mutation is not a random mutagenesis but a targeted gene repair

The traditional random mutagenesis model for reverse mutation sounds faultless. However, in a thought experiment we noticed that the model is somewhat inconsistent: (1) in the model, a cell survived if its defective *bla* gene happened restored in random mutagenesis. While a reverse mutation must occur in a living cell (during the process of DNA replication or repair), but the *bla*- cell itself cannot live in ACM. Therefore, the reverse mutation must have occurred in a cell before it was cultured in ACM. (2) However, for a given coding gene, in *E. coli* the probability of a forward mutation is as low as 10^{-8}. The probability of a back mutation that happened repaired a frameshift mutation should be lower in magnitudes than that of a forward mutation. (3) In addition, the bacterial genome consists millions base pairs and thousands of genes. If

a cell did not "*know exactly*" which gene was damaged, definitely it had little chance to repair it "blindly" by purely random mutagenesis. Therefore, the revertants should be very difficult to be detected if they really rely on random mutagenesis. However, it seems that reverse mutations occur much more frequently than expected, and thus the revertants were detected easily by routine screening.

Therefore, it is rather suspectable that in the history of evolution life has relied solely on such a simple but uncertain random mutagenesis to address this life-or-death issue. It is more likely that a specific mechanism have been developed for the repair of frameshift mutations. In a word, a reverse mutation is not a random mutagenesis, but a targeted gene repair: a cell must first identify the frameshifted gene, and then repairs it targetedly and specifically. However, how a frameshift gene was identified and repaired? By far it is still unknown at the molecular level.

3.3 Sanger sequencing of the bla gene of the revertants

The growth rates of the revertants varied from slow, moderate to normal (or fast). In order to investigate how the frameshifted *bla* was repaired, the revertants' colonies on ACP were picked and propagated in ACB. Their plasmid DNA was extracted and sequenced by the Sanger method. Unexpectedly, in the slow-growing revertants, it seems that their *bla* sequence is neither a wildtype nor a back mutation, but is still "*frameshifted"* (Fig 4A, top). For the moderate-growing revertants, the sequencing diagrams often contain two sets of superposed peaks (Fig 4A, middle), suggesting that they consist two types of *bla*, one was repaired and the other was still frameshifted. In the fast-growing revertants, the majority of their *bla* were repaired (Fig 4A, bottom). In fact, it is likely that the slow-growing revertants also contain repaired *bla* genes, but the proportion of the repaired sequence is too low to be detected by Sanger sequencing (Fig 4A, top). Because pBR322 is a multiple-copy plasmid, even when only a small proportion of its *bla* gene copies were repaired, it might be sufficient to support the survival of the bacteria in ACM.

In addition, in the revertants the repaired *bla* sequences read from the sequencing diagram are not the same as the wild-type but contain a variety of insertions, deletions and substitutions (Fig 4B, top), and their encoded protein sequences are also different

greatly from that of the wild-type BLA (Fig 4B, bottom). Especially, the repaired *bla* coding sequences still contain stop codons. In other words, in the revertants their *bla* coding genes were not restored fully and quickly, but repaired partially and gradually, resulted in diversified variants, called *frameshift homologs*.

3.4 The genome of the revertant is not more variable than the wild-type

A mutator strain, *e.g.*, *E. coli mutD5*, has a high level of spontaneous mutagenesis over the whole genome, owing to its defective proofreading and mismatch correction ability during DNA replication [8]. To investigate whether the above *bla* variants is a result of random mutagenesis, the genome DNA of *bla*+ and *bla*- were surveyed by NGS resequencing. As shown in Fig 4C-4D, Table 2 and 3, the level of SNP/InDel of the revertant is not higher, but lower than that of the wild-type, suggesting that the revertants had adopted a more stringent proofreading and mismatch correction during the replication of their genomic DNA. It suggests that the genome of the revertants is not more variable, but more stable than the wild-type, and therefore in the revertants the high level of variation observed in the *bla* gene is not a random mutagenesis at the whole-genome level, but a targeted gene repair specific for the frameshifted *bla* gene.

In addition, the genome sequencing report issued by the service provider showed that there are many *structure variations* (SVs) in both of the two bacterial genomes tested (Fig 4D). However, most of them length in 100~200 bp, equals approximately to the length of the reads, therefore, they were considered as falsely mapped reads rather than true structure variations. It has been well known that the limitations of short reads and amplification biases of the second generation platforms cause serious difficulties in sequence assembly, mapping and the analysis of genome structure [9]. However, the SNP/InDel data is reliable because of the extremely high sequencing depth and coverage of the Illumina HiSeq 2500 platforms.

3.5 PTCs signaled the spontaneous repair of the frameshift mutation

We speculate that it is the PTCs that signaled the repair of the frameshift mutation in the nonsense mRNA. In order to prove this, a PTC-free frameshifted *bla* (*bla#*) was constructed by replacing the PTCs each with an appropriate sense codon according to the *readthrough rules* (Table 1). The *bla#* gene was synthesized chemically, cloned in

the expression vector *pET-28a*, and expressed in *E. coli* BL21. As shown in Fig 5, by SDS-PAGE a 34-kDa protein band representing the frameshifted BLA was detected. The transformed bacteria were plated on ACPs, while no revertant was observed. The *bla* gene of the transformants was Sanger sequenced, but no variation was found. In a word, the PTC-free frameshifted *bla* gene was not repaired but expressed normally. The only difference between *bla-* and *bla#* is: *bla-* contains PTCs while *bla#* does not. Therefore, in *bla-* the targeted gene repair must be signaled by the PTCs.

3.6 Identifying genes/proteins involved in the repair of frameshift mutations

We expected that the repair of the frameshift mutation requires the upregulation of relevant genes, proteins and pathways. By analyzing the transcriptome of wild-type (*W*), frameshift (*FS*) and revertant (*RE*) strains, as shown in Fig 6, hundreds of genes were identified as upregulated in the frameshift and revertant when compared with the wild-type (Supplementary dataset 1-2). And dozens of pathways were identified as upregulated in a KEGG pathway enrichment analysis (Supplementary dataset 3-4). As shown in Table 4-5, genes and pathways related to transportation, localization, RNA surveillance, RNA degradation, DNA replication, RNA editing/DNA mismatch repair and homologous recombination were significantly upregulated in the frameshift and revertant strains when compared with those in the wild-type. We are further validating the regulation and function of these genes in the repair of frameshift mutation.

3.7 A new model for DNA Repair: RNA-directed Gene Editing

At present, the mechanism for the identification and repair of frameshift mutation is unknown at the molecular level. *However, the transcripts (nonsense mRNAs) must be involved, because obviously the PTCs cannot function in the DNA double helix, but only in the transcripts*. Based on the above experiments, and an in-depth survey of the literatures, we proposed a new model for the repair of frameshift mutations (Fig 3B), *RNA-directed Gene Editing* (RdGE), which consist four main steps:

(1). ***Nonsense mRNA recognition/processing***: When a frameshift mutation occurred in the upstream of a coding gene, in the downstream of the frameshift it results in a substantial change of the encoded amino acid sequence and often the emerging of a number of nonsense codons, called *hidden stop codons* (HSCs) [1]. When a

frameshifted coding gene was expressed, the transcripts will be recognized as nonsense mRNAs by the mRNA surveillance system [10-16]. In the nonsense mRNAs, the stop codons are called *premature termination codons* (PTCs). When the nonsense mRNAs are translated, the ribosome complex will be blocked when it encounters a PTC [17]. The nonsense mRNA was then recognized by the mRNA surveillance system, using the presence of a second signal downstream of a termination codon to distinguish a PTC from a true stop codon [18]. Once identified, the nonsense mRNAs are subject to one of the following pathways: *nonsense-mediated mRNA decay (NMD), translational repression, alternative splicing, or transcriptional silencing* [19-21]. However, it has been observed that sometimes a frameshifted coding gene can be still active and functional through some special mechanisms, *e.g., translational readthrough* or *translational frameshifting* [22-25].

(2). ***RNA editing***: base on the above data, we hypothesized that nonsense mRNAs can be repaired through RNA editing [26-30]. By PTC signaling, a cell is capable of identifying a nonsense mRNA, but it has no way to locate the missing/inserting base that caused the frameshift exactly, and therefore it has to insert, delete and substitute nucleotides within or nearby a PTC, producing a variety of homologous mRNA variants.

(3). ***Translation*****:** The edited mRNAs were then translated. The remaining PTCs, if exist, were readthrough by translating them into appropriate amino acids [31-35]. A cell survived if and only if the RNA editing produced a functional protein. The chance of producing a functional protein by editing the nonsense mRNA is still fairly small, but is greater than that of random mutagenesis in genomic DNA.

(4). ***RNA-directed Gene Editing***: if a cell survived, a functionalized transcript were transported back into the nucleus/plasmid DNA to localize their own coding gene, and to direct the repair of their faulty coding sequence, through mismatch repair and/or homologous recombination.

Comparing with the random mutagenesis, this RdGE model for frameshift repair better explains the slow growth rates of the frameshift mutant and the revertants, and the high frequency variations observed only in the *bla* genes, but not in the genome:

(1). The frameshift strain grows slowly not only in ACB, but also in TCB, because the targeted gene repair is initiative and has already happened even before the adding of ampicillin, *i.e.*, ampicillin is not the cause of reverse mutation, but a selection pressure to screen and display the revertants, which were the result of reverse mutation that had already happened before the adding of ampicillin.

(2). RdGE is signaled by the PTCs in the nonsense mRNAs, and the aim of RdGE is to clear away PTCs. In the frameshift and the initial slow-growing revertants, there were quite many PTCs in their *bla*, the chance of producing a functional mRNA by RNA editing is very small, and most cells died before their *bla* was repaired, so their survival rate is very low and their growth rate is very slow.

(3). In the fast-growing revertants, with more and more PTCs were replaced, less and less PTCs were left in the *bla* gene, gene repair become easier and easier, so their survival and growth rate become higher and higher.

(4). Finally, when all PTCs were eliminated, the gene repair process was complete. The *bla* coding sequence is restored, not necessarily same to the wild-type, but may contain a variety of variants.

Interestingly, as shown in Fig 2, the wild-type colonies are equal sized, but the sizes of the revertant colonies varied, showing that their growth rates differed greatly, suggesting that the activities of their BLA enzyme are also diversified. Previously, we have found that frameshift homologs are widespread within and across species [36]. In the above, we demonstrated that the PTCs signaled the repair of the nonsense mRNAs and their coding gene. Therefore, when a frameshift mutation happens in a coding gene it is repaired by RdGE, and resulted in a variety of frameshift homologs. Therefore, the widely-exist frameshift homologs are indeed the residual clues for this RdGE model preserved in nature. If so, RdGE is a novel mechanism not only for the targeted gene repair, but also for the molecular evolution of protein coding genes.

4. Discussion

4.1 A mechanism for the repair of frameshift mutations

The molecular mechanisms involved in the repair of DNA damage has been well studied [6], however, no mechanism is designed specific for the repair of frameshift mutations. In cell division and multiplication, DNA replication happens daily in cells. Because of the imperfect nature of the DNA replication, InDels can be caused by a spontaneous mutation (replication errors or slipped strand mispairing) or an induced mutation (radiation, pollutants or toxins), which may lead to frameshift mutations. Therefore, in a cell, maintaining the reading frame of every protein coding gene is a fundamentally important task. As a target-specific and well-controlled mechanism, this RdGE model not only well explains our observations, but also is consistent with numerous relevant previous studies.

Since the original transcripts were nonsense mRNA, they must be functionalized through RNA editing prior to direct the repair of their own coding gene. It is pretty surprising that cells do not edit the coding DNA itself directly, but the transcripts are first edited and then used to direct the DNA repair. However, RdGE is indeed more reasonable and feasible than directly editing the coding DNA sequence: the gene must be first transcribed and then repaired, because in a cell a frameshift mutation cannot be identified in the DNA level. It would be more efficient to repair the transcript first, make it be functional, and then the repaired mRNA is used to direct the gene repair. Otherwise, if cells edit the coding DNA directly, the edited gene must be transcribed again for a second round to make it be functional. Conceivably, the RdGE strategy has much fewer difficulties and more opportunities to succeed than editing DNA directly.

4.2 Evidences connecting DNA Repair and RNA-editing

All types of RNAs are subject to processing or degradation, and various cellular mechanisms are involved [37]. During the last decades, some studies have established a link between DNA repair and RNA surveillance. Several proteins that respond to DNA damage, *e.g.* base excision repair enzymes, SMUG1, APE1 and PARP1, have been shown to participate in RNA surveillance, turnover or processing [38].

In the above, we demonstrated that deaminase and mismatch repair proteins were both upregulated in the frameshift and the revertant. Deamination of DNA bases generates deoxyinosine from deoxyadenosine. Several studies provided some evidences that deamination is related to DNA repair and RNA editing. For example, endonuclease V, which is highly conserved from *E. coli* to human [39], is involved in alternative excision repair that removes deoxyinosine from DNA. It is reported that human endonuclease V, localizes to the cytoplasm, is also a ribonuclease specific for inosine-containing RNA, hydrolyses the second phosphodiester bond located 3' to the inosine in unpaired inosine-containing ssRNA regions in dsRNA, and controls the fate of inosine-containing RNA in humans [39]. For another example, editing of the pre-mRNA for the DNA repair enzyme NEIL1 causes a lysine to arginine change in the lesion recognition loop of the protein, and the recoding site is a preferred editing site for the RNA editing adenosine deaminase ADAR1, suggested a link between the regulatory mechanism for DNA repair and RNA editing [40].

4.3 Evidences for RNA-directed DNA Repair

As is well known, RNAs are transcribed in the nucleus and transported from the nucleus to the cytoplasm [41]. Recently, increasingly more evidences suggested that RNAs can be transported back to the DNA to repair double strand breaks (DSB), a process referred to as *RNA-directed DNA Repair* (RdDR). Not only a synthetic RNA carried by a DNA oligonucleotide, but also RNA-only oligonucleotides can precisely repair a DSB in homologous DNA, serving as direct templates for DNA synthesis at the chromosomal level [42-45]. The capacity of short RNA patches to directly modify DNA was reported in *E. coli* [44, 46], and precisely repaired a DSB in yeast [42, 47] and human embryonic kidney (HEK-293) cells [44].

RdGE is an extension of RdDR, as the transcripts were edited prior to directing DNA repair. As described above, in the literature there are many evidences support the link between RNA editing and gene repair, which also support this RdGE model indirectly, while further systematic investigations are needed to clarify the molecular process in details and validate this mechanism in various species.

4.4 RNA-directed genome editing and targeted gene repair

Gene therapy can potentially be used for the repair of disease-causing frameshift and point mutations. CRISPR/Cas9 has been widely used for RNA-guided genome editing [48-50], and for the introduction or correction of specific mutations [51, 52]. However, RNA-guided genome editing requires the transfection of an exogenous gene in the host cell. On the other side, deoxyoligonucleotides-directed targeted gene repair (ODN-TGR) is capable of targeting a single-base pair mutation in a highly specific manner [53, 54]. Being non-transgenic, ODN-TGR could be more favorable in gene therapy, genome engineering and molecular breeding. Unfortunately, however, they have often suffered from low efficiency, because they rely on the induction of the endogenous DNA repair system, while they are suppressed by the DNA mismatch repair mechanisms [55].

Moreover, synthetic DNA oligonucleotides and DNA/RNA complexes have been widely used for targeting the gene repair [53-55]. And synthetic RNA or endogenous transcripts have also been reported to direct the repair of the double-strand breaks of DNA [42, 43, 47, 56-58], but it has never been reported that endogenous RNA can be edited and direct the repair of a gene. Since the basic components of RdGE are highly conserved from bacteria to human, this endogenous pathway is potentially useful for gene editing or gene repair using only RNA molecules, and without introducing any exogenous gene or protein.

Author Contributions

Xiaolong Wang conceived the study, designed the experiments, analyzed the data and wrote the paper; Haibo Peng, Chunyan Li, Yalei Wang and Xuxiang Wang performed the experiments; Gang Chen and Jianye Zhang discussed on the paper.

Acknowledgements:

This study was supported by The *National Natural Science Foundation of China* (Grant No. 81072567).

Figure legends

Fig 1. The introduction of a frameshift mutation in the upstream of the *bla* gene in the plasmid pBR322. Sanger sequencing result of: (A) the wild type (*bla*+); (B) the frameshift mutant (*bla*-). (C) Alignment of the nucleotide sequence of the wild type and that of the frameshift mutant.

Fig 2. Growth of *E.coli* on ACPs and TCPs: *bla*+: Wild-type; *bla*-: Frameshift; *bla**: revertants; *blank*: Blank control

Fig 3. Different models for the repair of frameshifted protein-coding genes:

(A) The traditional "*random mutagenesis*" model: **Left**: in a frameshifted coding gene, a number of HSCs (*red bars*) emerged in the CDS, and in mRNA the PTCs (*red bars*) caused *translational termination*, resulting in truncated products; **Right**: when the coding DNA sequence is happened repaired by a *back mutation*, the reading frame is restored, the stop codons are hid (*green bars*), and the translation proceed;

(B) This "*initiative repair*" model: **Left**: in a frameshifted gene, a number of HSCs (*yellow bars*) emerged in the CDS. Signaled by the PTCs (*yellow bars*), the nonsense mRNA was identified by *mRNA surveillance*, and processed by *translational termination* and *mRNA decaying*. **Middle**: The nonsense mRNA can also be repaired or revised by *RNA editing*. **Right**: the coding sequence is repaired by RdGE, most stop codons are hid (*green bars*), the reading frame restored partially, and the translation proceed by *translational stop codon readthrough*.

Fig 4. Sequencing and sequence analysis: (A) The Sanger sequencing diagram of the revertants, shows two set of overlapped peaks, one is frameshifted, the other is repaired; (B) The repaired *bla* coding DNA sequence (top) and the translated proteins sequence (bottom) read from the Sanger sequencing diagram of the revertants. *bla*: wild-type; *bla*-: frameshift; *A*-*E*: different revertants; *red box*: the base deleted in the mutagenesis; *blue boxes*: bases inserted or deleted in the revertants; (C) The distribution of SNPs/InDels on the genome of wild-type and revertants (linear view); (D) The distribution of SNPs/InDels on the genome of wild-type and revertants (circular view); (E) The falsely reported "structure variations (SVs)" in the genome of wild-type and revertants;

Fig 5. SDS-PAGE of the frameshifted BLA by expression of pET28a-*bla #* in *E. coli* BL21.

Fig 6. The transcriptome analysis of different *E. coli* strains: *W_1*: wildtype; *FS_2*: frameshift; *R_3*: initial revertant; *RE_4*: subculture of a revertant; (A) The number of differential expression genes (DGEs) and the heat map; (B) the most enriched GO terms by comparing FS_2 with W_1; (C) the most enriched GO terms by comparing R_3 with FS_2; (D) DNA mismatch repair pathway (DpoIII gene was upregulated); (E) DNA homologous recombination pathway (4 genes were upregulated). (F) RNA degradation pathway (6 genes were upregulated);

Table 1. The natural suppressor tRNA for nonsense mutations (the *in vivo readthrough rules*).

Site	tRNA (AA)	Wild type		Correction	
		Code	**Anti-code**	**Code**	**Anti-code**
supD	Ser (S)	→ UCG	CGA←	→ UAG	CUA←
supE	Gln (Q)	→ CAG	CUG←	→ UAG	CUA←
supF	Tyr (Y)	→ UAC	GUA←	→ UAG	CUA←
supG	Lys (K)	→ AAA	UUU←	→ UAA	UUA←
supU	Trp (W)	→ UGG	CCA←	→ UGA	UCA←

Table 2. The summary of InDel in the *bla*+/- genome

Sample	Insertion	Deletion	Het	Hom	HetRate(%)	Total
bla+	2	4	0	6	0	6
bla-	2	3	0	5	0	5

Table 3. The summary of SNP in the *bla*+/ - genome

Sample	ts	tv	ts/tv	Het	Hom	HetRate(%)	Total	Density(SNP/Kb)
bla+	43	27	1.59	4	66	0	70	0.02
bla-	42	27	1.56	2	67	0	69	0.01

Table 4. Possible RdGE-relevant genes and pathways that were upregulated in the frameshift

Pathway Term	Database	ID	Input number	Background number	P-Value	Corrected P-Value
ABC transporters	KEGG	eco02010	32	171	0.0001	0.0090
RNA degradation	KEGG	eco03018	1	15	0.7419	0.9762
DNA replication	KEGG	eco03030	1	17	0.7822	0.9762
Mismatch repair	KEGG	eco03430	1	22	0.8575	0.9796
Homologous recombination	KEGG	eco03440	1	27	0.9069	0.9796

Table 6. Possible RdGE-relevant genes and pathways that were upregulated in the revertant

Pathway Term	Database	ID	Input number	Background number	P-Value	Corrected P-Value
ABC transporters	KEGG	eco02010	32	171	3.53E-05	0.0018
Homologous recombination	KEGG	eco03440	2	27	0.6539	0.9519
RNA degradation	KEGG	eco03018	3	15	0.1017	0.5470
DNA replication	KEGG	eco03030	1	17	0.7564	0.9519
Mismatch repair	KEGG	eco03430	1	22	0.8357	0.9519

References

1. Seligmann, H. and D.D. Pollock, *The ambush hypothesis: hidden stop codons prevent off-frame gene reading.* DNA Cell Biol, 2004. **23**(10): p. 701-5.
2. Owens, M., et al., *SOS1 frameshift mutations cause pure mucosal neuroma syndrome, a clinical phenotype distinct from multiple endocrine neoplasia type 2B.* Clin Endocrinol (Oxf), 2016. **84**(5): p. 715-9.
3. Yan, S.E., et al., *In-depth analysis of hyaline fibromatosis syndrome frameshift mutations at the same site reveal the necessity of personalized therapy.* Hum Mutat, 2013. **34**(7): p. 1005-17.
4. Amiel, J., et al., *Polyalanine expansion and frameshift mutations of the paired-like homeobox gene PHOX2B in congenital central hypoventilation syndrome.* Nat Genet, 2003. **33**(4): p. 459-61.
5. Kang, S., et al., *GLI3 frameshift mutations cause autosomal dominant Pallister-Hall syndrome.* Nat Genet, 1997. **15**(3): p. 266-8.
6. Jensen, N.M., et al., *An update on targeted gene repair in mammalian cells: methods and mechanisms.* J Biomed Sci, 2011. **18**: p. 10.
7. Streisinger, G., et al., *Frameshift mutations and the genetic code. This paper is dedicated to Professor Theodosius Dobzhansky on the occasion of his 66th birthday.* Cold Spring Harb Symp Quant Biol, 1966. **31**: p. 77-84.
8. Schaaper, R.M., *Mechanisms of mutagenesis in the Escherichia coli mutator mutD5: role of DNA mismatch repair.* Proc Natl Acad Sci U S A, 1988. **85**(21): p. 8126-30.
9. Roberts, R.J., M.O. Carneiro, and M.C. Schatz, *The advantages of SMRT sequencing.* Genome Biol, 2013. **14**(7): p. 405.
10. Ge, Z., et al., *Polypyrimidine tract binding protein 1 protects mRNAs from recognition by the nonsense-mediated mRNA decay pathway.* Elife, 2016. **5**.
11. Schweingruber, C., et al., *Nonsense-mediated mRNA decay - mechanisms of substrate mRNA recognition and degradation in mammalian cells.* Biochim Biophys Acta, 2013. **1829**(6-7): p. 612-23.
12. Niu, D.K. and J.L. Cao, *Nucleosome deposition and DNA methylation may participate in the recognition of premature termination codon in nonsense-mediated mRNA decay.* FEBS Lett, 2010. **584**(16): p. 3509-12.
13. Muhlemann, O., et al., *Recognition and elimination of nonsense mRNA.* Biochim Biophys Acta, 2008. **1779**(9): p. 538-49.
14. Muhlemann, O., *Recognition of nonsense mRNA: towards a unified model.* Biochem Soc Trans, 2008. **36**(Pt 3): p. 497-501.
15. Muhlrad, D. and R. Parker, *Recognition of yeast mRNAs as "nonsense containing" leads to both inhibition of mRNA translation and mRNA degradation: implications for the control of mRNA decapping.* Mol Biol Cell, 1999. **10**(11): p. 3971-8.
16. Zhang, J. and L.E. Maquat, *Evidence that the decay of nucleus-associated nonsense mRNA for human triosephosphate isomerase involves nonsense codon recognition after splicing.* RNA, 1996. **2**(3): p. 235-43.
17. Hwang, J. and Y.K. Kim, *When a ribosome encounters a premature termination codon.* BMB Rep, 2013. **46**(1): p. 9-16.

18. Scofield, D.G., X. Hong, and M. Lynch, *Position of the final intron in full-length transcripts: determined by NMD?* Mol Biol Evol, 2007. **24**(4): p. 896-9.

19. Le Hir, H., et al., *The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay.* EMBO J, 2001. **20**(17): p. 4987-97.

20. Gonzalez, C.I., W. Wang, and S.W. Peltz, *Nonsense-mediated mRNA decay in Saccharomyces cerevisiae: a quality control mechanism that degrades transcripts harboring premature termination codons.* Cold Spring Harb Symp Quant Biol, 2001. **66**: p. 321-8.

21. Zhang, J. and L.E. Maquat, *Evidence that translation reinitiation abrogates nonsense-mediated mRNA decay in mammalian cells.* EMBO J, 1997. **16**(4): p. 826-33.

22. Fox, T.D., *Five TGA "stop" codons occur within the translated sequence of the yeast mitochondrial gene for cytochrome c oxidase subunit II.* Proc Natl Acad Sci U S A, 1979. **76**(12): p. 6534-8.

23. Pai, H.V., et al., *A frameshift mutation and alternate splicing in human brain generate a functional form of the pseudogene cytochrome P4502D7 that demethylates codeine to morphine.* Journal of Biological Chemistry, 2004. **279**(26): p. 27383-27389.

24. Xu, J., R.W. Hendrix, and R.L. Duda, *Conserved translational frameshift in dsDNA bacteriophage tail assembly genes.* Molecular Cell, 2004. **16**(1): p. 11-21.

25. Baykal, U., A.L. Moyne, and S. Tuzun, *A frameshift in the coding region of a novel tomato class I basic chitinase gene makes it a pseudogene with a functional wound-responsive promoter.* Gene, 2006. **376**(1): p. 37-46.

26. Kugita, M., et al., *RNA editing in hornwort chloroplasts makes more than half the genes functional.* Nucleic Acids Res, 2003. **31**(9): p. 2417-23.

27. Bock, R., *Sense from nonsense: how the genetic information of chloroplasts is altered by RNA editing*. Biochimie, 2000. **82**(6-7): p. 549-57.

28. Horvath, A., E.A. Berry, and D.A. Maslov, *Translation of the edited mRNA for cytochrome b in trypanosome mitochondria.* Science, 2000. **287**(5458): p. 1639-40.

29. Yoshinaga, K., et al., *Extensive RNA editing and possible double-stranded structures determining editing sites in the atpB transcripts of hornwort chloroplasts.* Nucleic Acids Res, 1997. **25**(23): p. 4830-4.

30. Sloof, P., et al., *RNA editing in mitochondria of cultured trypanosomatids: translatable mRNAs for NADH-dehydrogenase subunits are missing.* J Bioenerg Biomembr, 1994. **26**(2): p. 193-203.

31. Kopelowitz, J., et al., *Influence of codon context on UGA suppression and readthrough.* J Mol Biol, 1992. **225**(2): p. 261-9.

32. Weiss, R.B., *Ribosomal frameshifting, jumping and readthrough.* Curr Opin Cell Biol, 1991. **3**(6): p. 1051-5.

33. Engelberg-Kulka, H., L. Dekel, and M. Israeli-Reches, *Regulation of the escherichia coli tryptophan operon by readthrough of UGA termination codons.* Biochem Biophys Res Commun, 1981. **98**(4): p. 1008-15.

34. Seidman, J.S., B.D. Janssen, and C.S. Hayes, *Alternative fates of paused ribosomes during translation termination.* J Biol Chem, 2011. **286**(36): p. 31105-12.

35. Su, D., Y. Li, and V.N. Gladyshev, *Selenocysteine insertion directed by the 3'-UTR SECIS element in Escherichia coli.* Nucleic Acids Res, 2005. **33**(8): p. 2486-92.

36. Wang, X.W., X.; Chen, G.; Zhang, J.; Liu, Y.; Yang C. , *The shiftability of protein coding genes: the genetic code was optimized for frameshift tolerating.* PeerJ PrePrints 2015. **3** p. e806v1.

37. Isken, O. and L.E. Maquat, *Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function.* Genes Dev, 2007. **21**(15): p. 1833-56.

38. Vohhodina, J., D.P. Harkin, and K.I. Savage, *Dual roles of DNA repair enzymes in RNA biology/post-transcriptional control.* Wiley Interdiscip Rev RNA, 2016.

39. Kuraoka, I., *Diversity of Endonuclease V: From DNA Repair to RNA Editing.* Biomolecules, 2015. **5**(4): p. 2194-206.

40. Yeo, J., et al., *RNA editing changes the lesion specificity for the DNA repair enzyme NEIL1.* Proc Natl Acad Sci U S A, 2010. **107**(48): p. 20715-9.

41. Nakielny, S. and G. Dreyfuss, *Transport of proteins and RNAs in and out of the nucleus.* Cell, 1999. **99**(7): p. 677-90.

42. Storici, F., et al., *RNA-templated DNA repair.* Nature, 2007. **447**(7142): p. 338-41.

43. Keskin, H., et al., *Transcript-RNA-templated DNA recombination and repair.* Nature, 2014. **515**(7527): p. 436-9.

44. Shen, Y., et al., *RNA-driven genetic changes in bacteria and in human cells.* Mutat Res, 2011. **717**(1-2): p. 91-8.

45. Shen, Y. and F. Storici, *Detection of RNA-templated double-strand break repair in yeast.* Methods Mol Biol, 2011. **745**: p. 193-204.

46. Thaler, D.S., G. Tombline, and K. Zahn, *Short-patch reverse transcription in Escherichia coli.* Genetics, 1995. **140**(3): p. 909-15.

47. Storici, F., *RNA-mediated DNA modifications and RNA-templated DNA repair.* Curr Opin Mol Ther, 2008. **10**(3): p. 224-30.

48. Shalem, O., et al., *Genome-scale CRISPR-Cas9 knockout screening in human cells.* Science, 2014. **343**(6166): p. 84-7.

49. Ran, F.A., et al., *Genome engineering using the CRISPR-Cas9 system.* Nat Protoc, 2013. **8**(11): p. 2281-308.

50. Ran, F.A., et al., *Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity.* Cell, 2013. **154**(6): p. 1380-9.

51. Paquet, D., et al., *Efficient introduction of specific homozygous and heterozygous mutations using CRISPR/Cas9.* Nature, 2016. **533**(7601): p. 125-9.

52. Zimmer, C.T., et al., *A CRISPR/Cas9 mediated point mutation in the alpha 6 subunit of the nicotinic acetylcholine receptor confers resistance to spinosad in Drosophila melanogaster.* Insect Biochem Mol Biol, 2016. **73**: p. 62-9.

53. Parekh-Olmedo, H., et al., *Gene therapy progress and prospects: targeted gene repair.* Gene Ther, 2005. **12**(8): p. 639-46.

54. de Semir, D. and J.M. Aran, *Targeted gene repair: the ups and downs of a promising gene therapy approach.* Curr Gene Ther, 2006. **6**(4): p. 481-504.

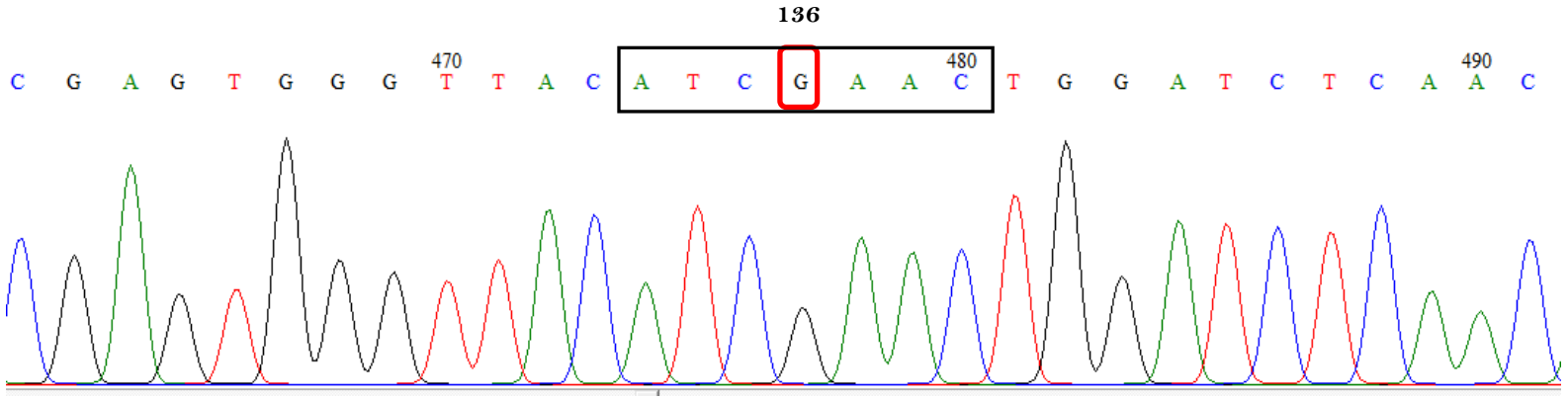
55. Dekker, M., C. Brouwers, and H. te Riele, *Targeted gene modification in mismatch-repair-deficient embryonic stem cells by single-stranded DNA oligonucleotides.* Nucleic Acids Res, 2003. **31**(6): p. e27.

56. Trott, D.A. and A.C. Porter, *Hypothesis: transcript-templated repair of DNA double-strand breaks.* Bioessays, 2006. **28**(1): p. 78-83.

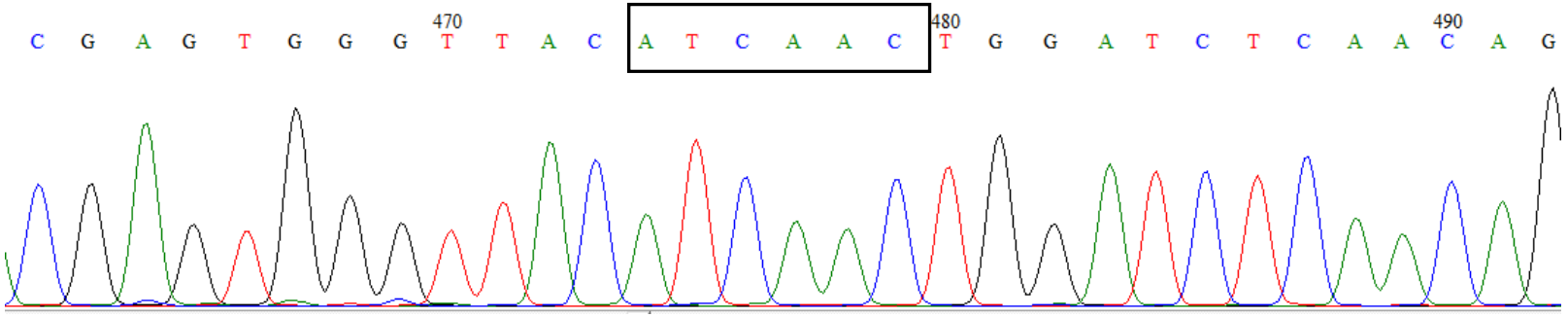
57. Keskin, H., C. Meers, and F. Storici, *Transcript RNA supports precise repair of its own DNA gene*. RNA Biol, 2016. **13**(2): p. 157-65.

58. Wei, L., et al., *DNA damage during the G0/G1 phase triggers RNA-templated, Cockayne syndrome B-dependent homologous recombination.* Proc Natl Acad Sci U S A, 2015. **112**(27): p. E3495-504.

A



B



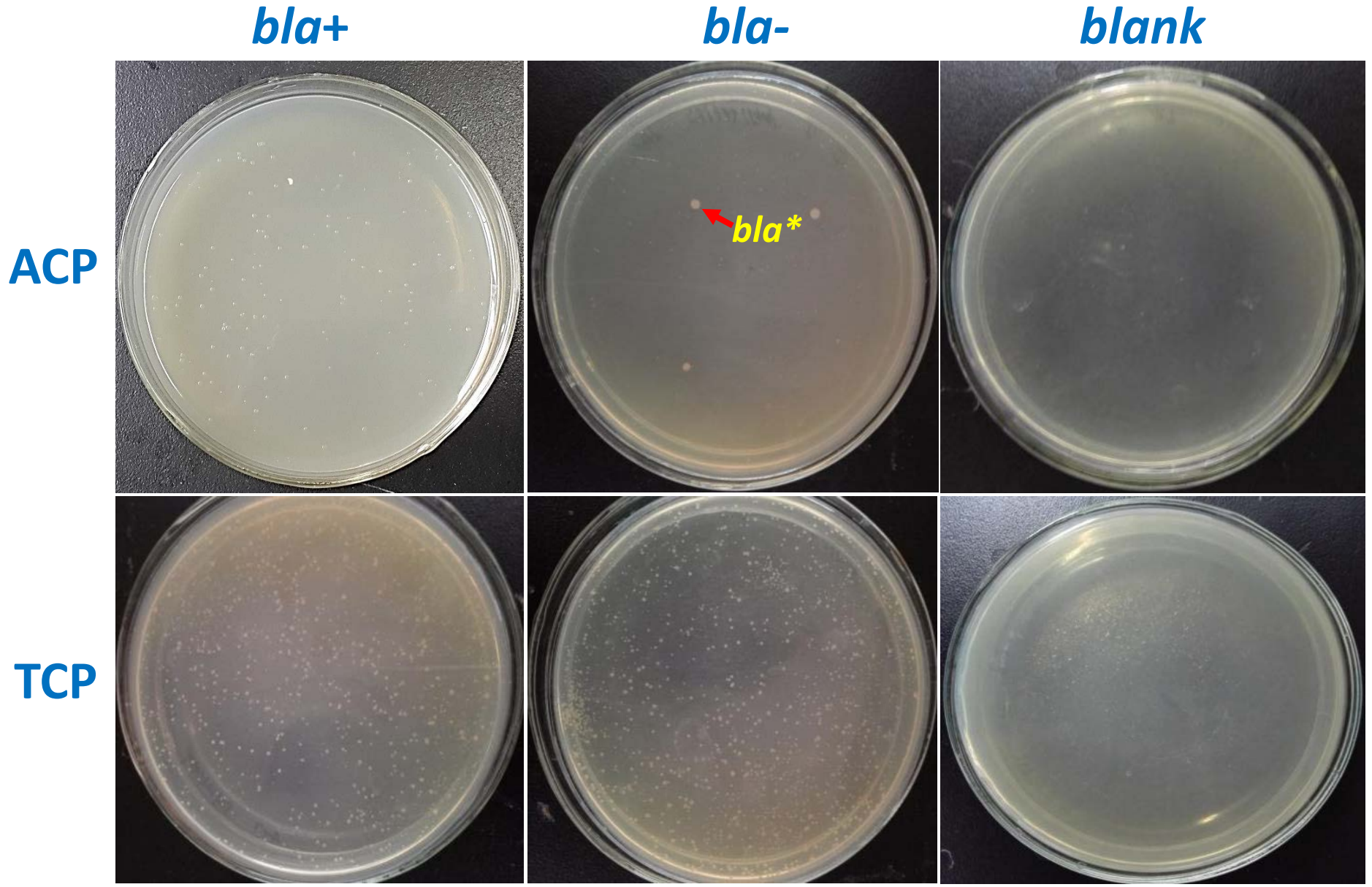
C

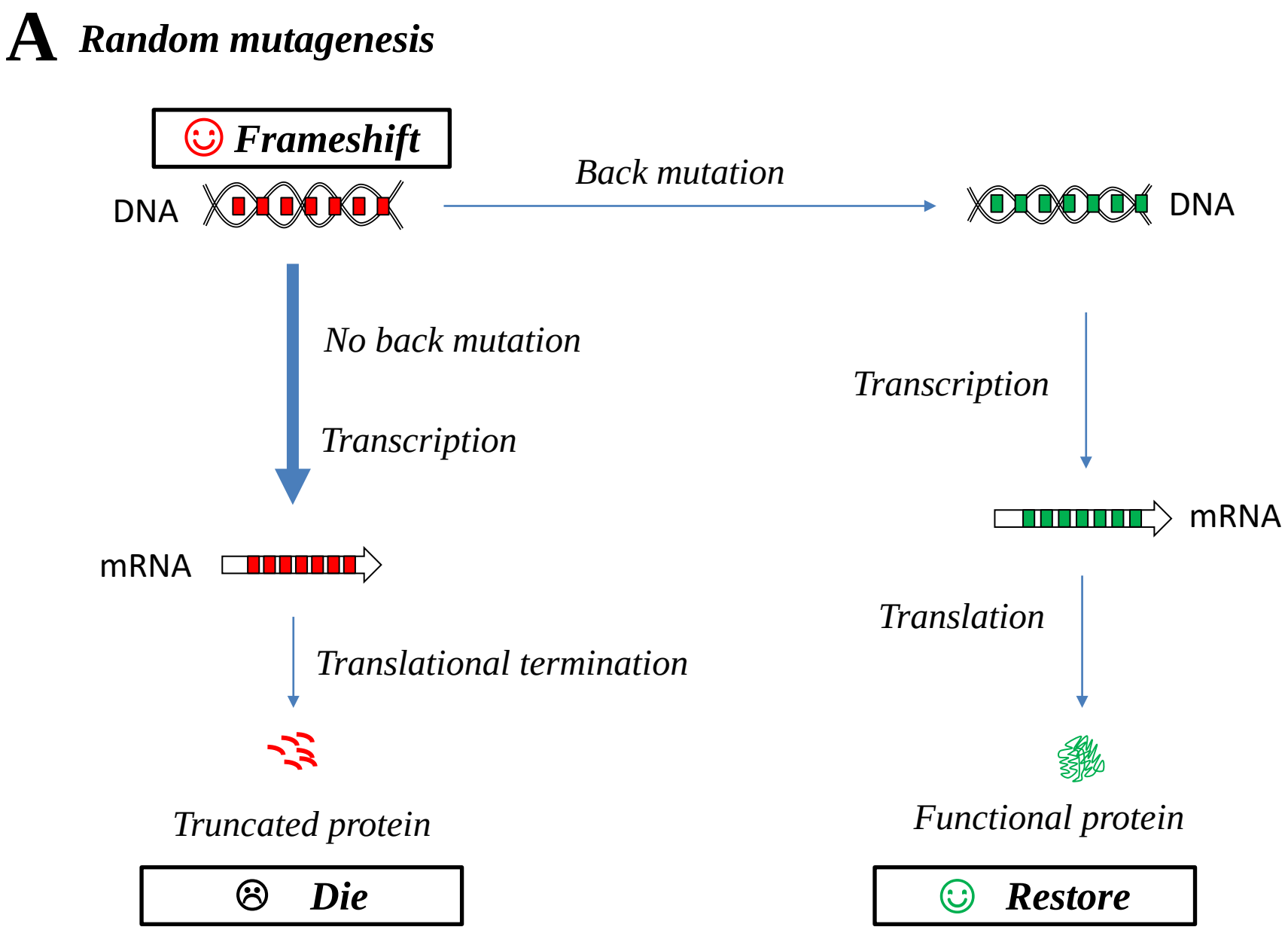
136

Wild-type (*bla* +): …A T C G A A C…

Frameshift (*bla* -): …A T C - A A C…

Fig 2. Growth of different *E.coli* strains on ACPs and TCPs




A
Random mutagenesis
Frameshift
DNA
Back mutation
DNA
No back mutation
Transcription
Transcription
mRNA
mRNA
Translational termination
Translation
Truncated protein
Functional protein
Die
Restore

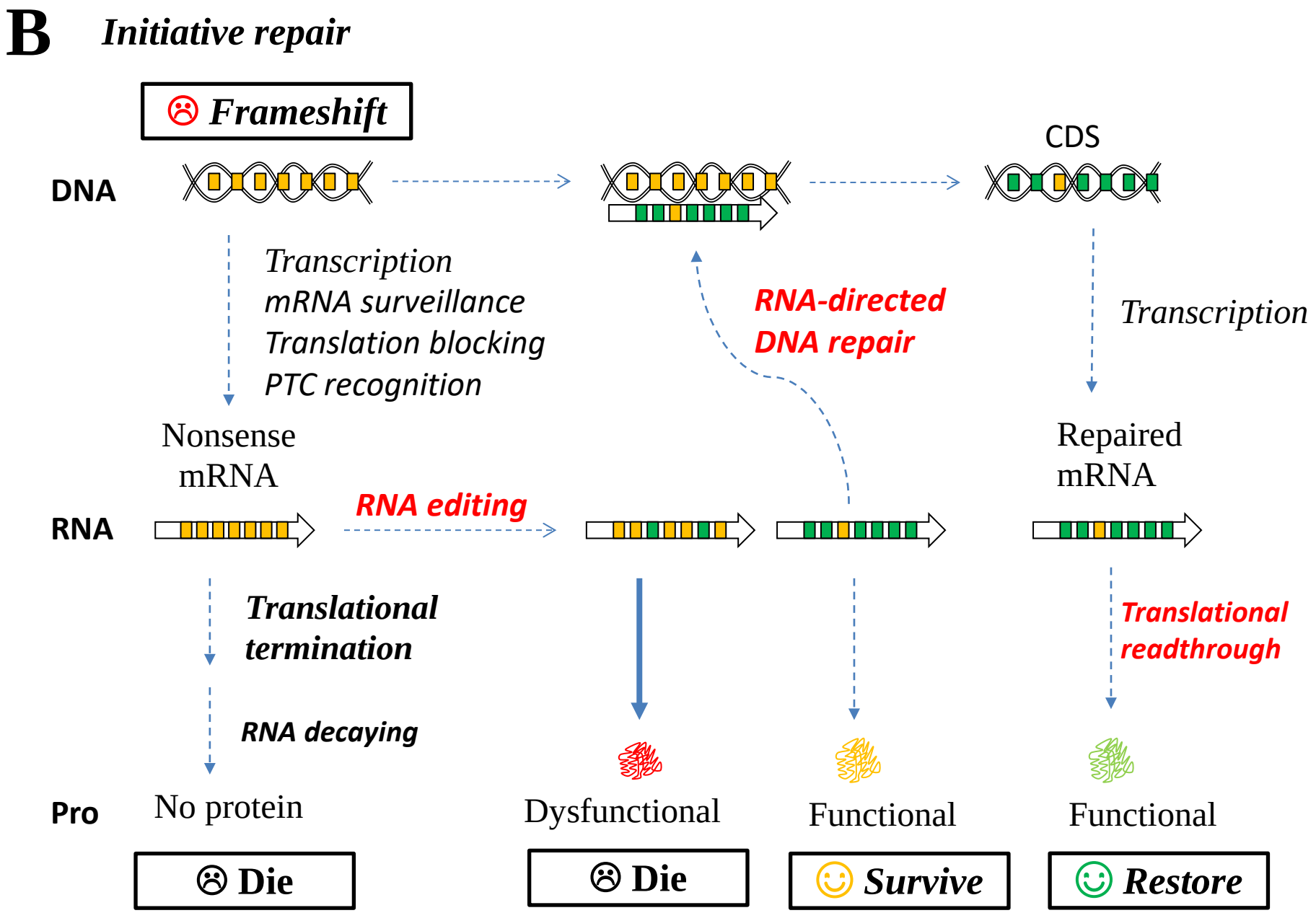

B
Initiative repair
☹ Frameshift
DNA
CDS
Transcription
mRNA surveillance
Translation blocking
PTC recognition
RNA-directed
DNA repair
Transcription
Nonsense
mRNA
Repaired
mRNA
RNA
RNA editing
Translational
termination
Translational
readthrough
RNA decaying
Pro
No protein
Dysfunctional
Functional
Functional
☹ Die
☹ Die
☺ Survive
☺ Restore

Fig 4A. Sanger sequencing diagram of the revertants

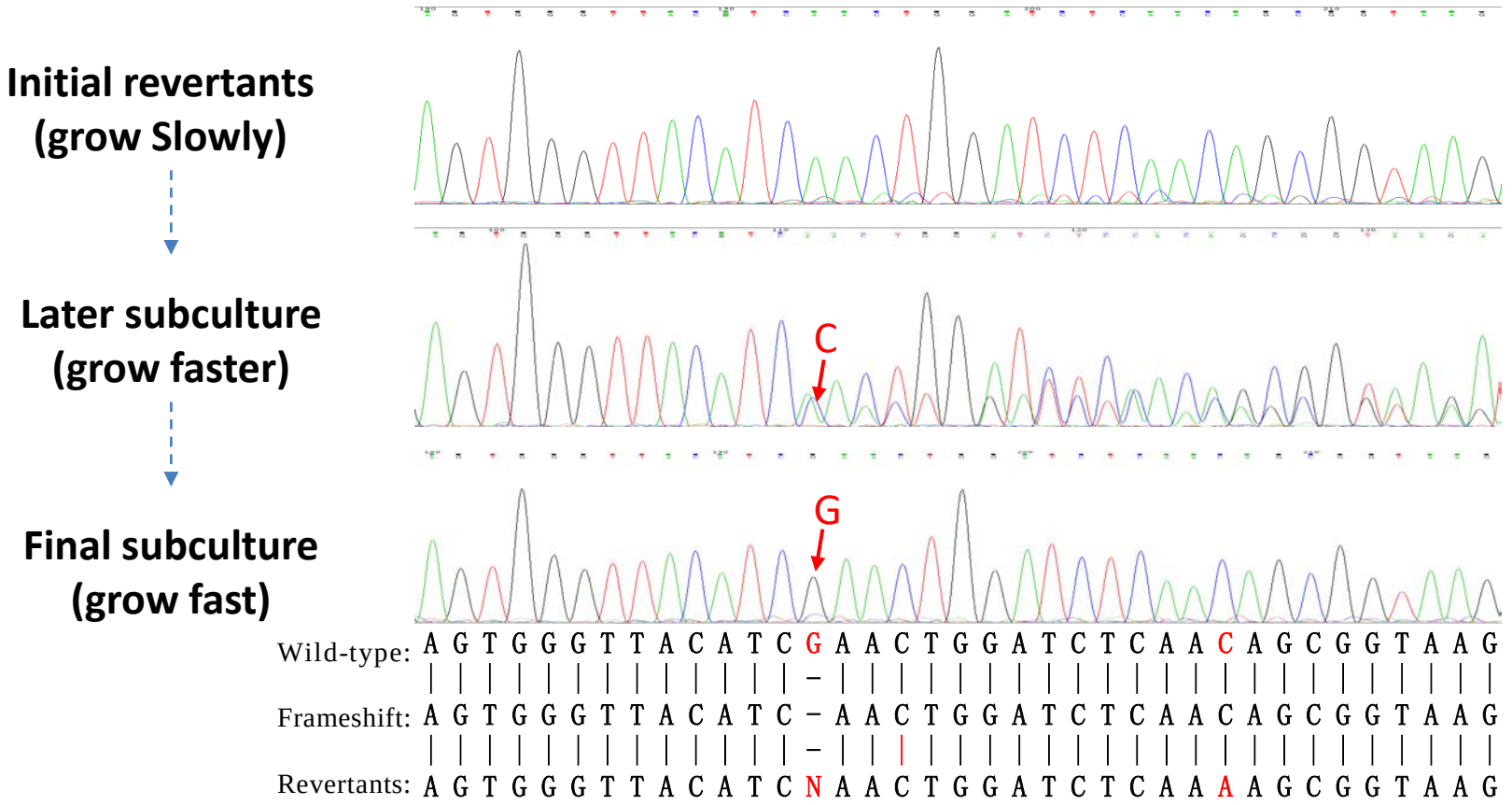


Fig 4B. The *bla* sequences of the wild-type and revertants

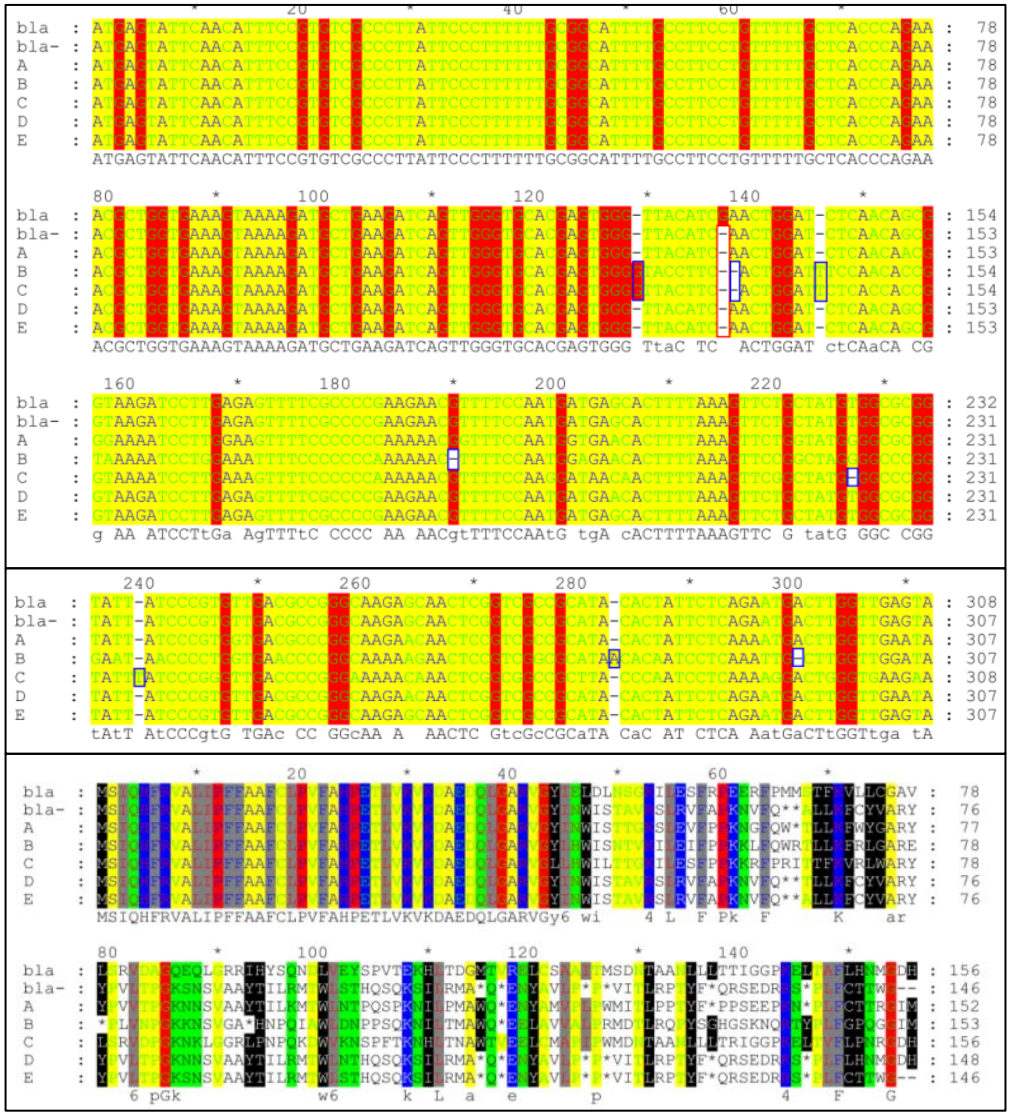


Fig 4C. The SNP/InDel in the genome of the wild-type and revertants

bla+

bla*

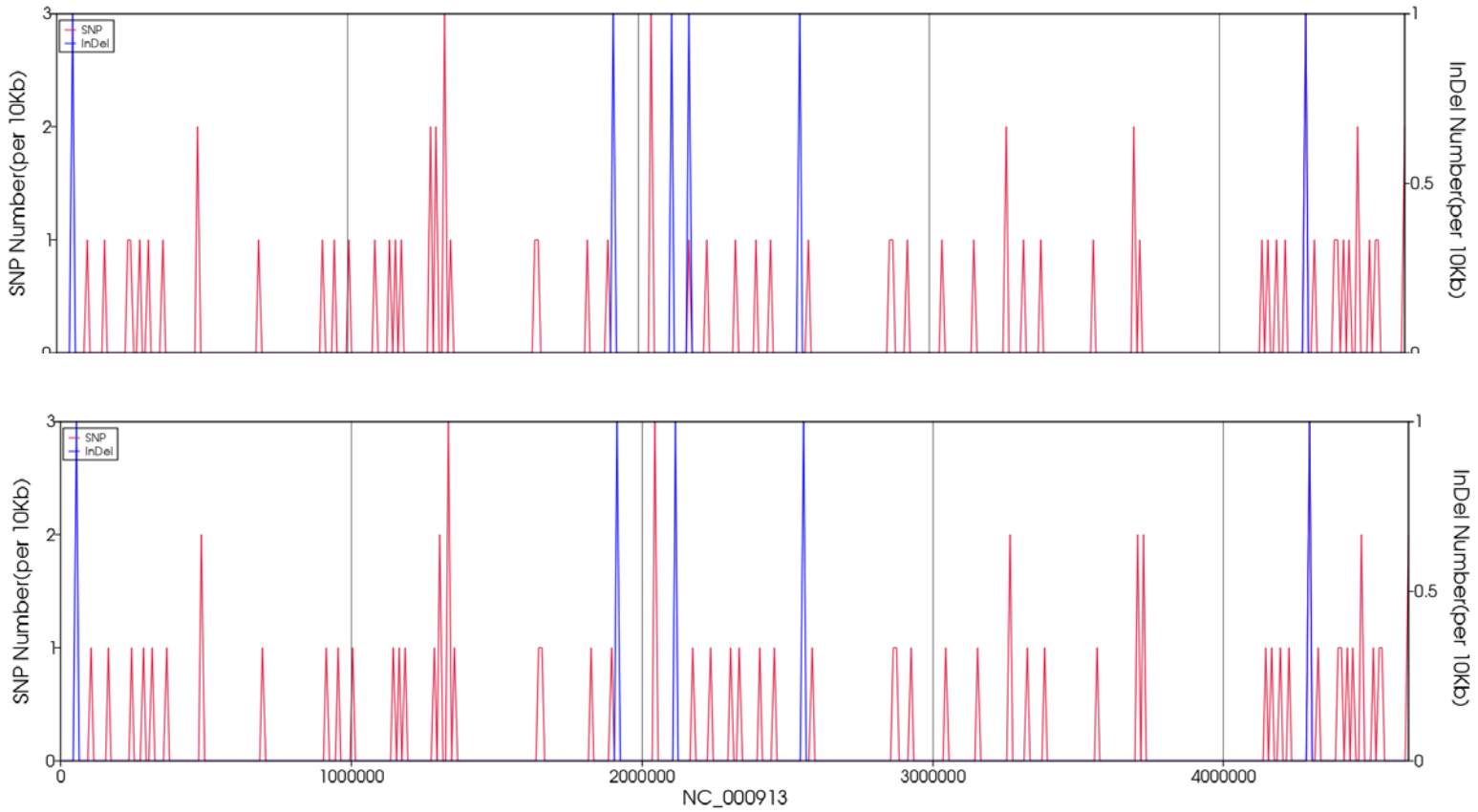


Fig 4D. The SNP/InDel in the genome of the wild-type and revertants

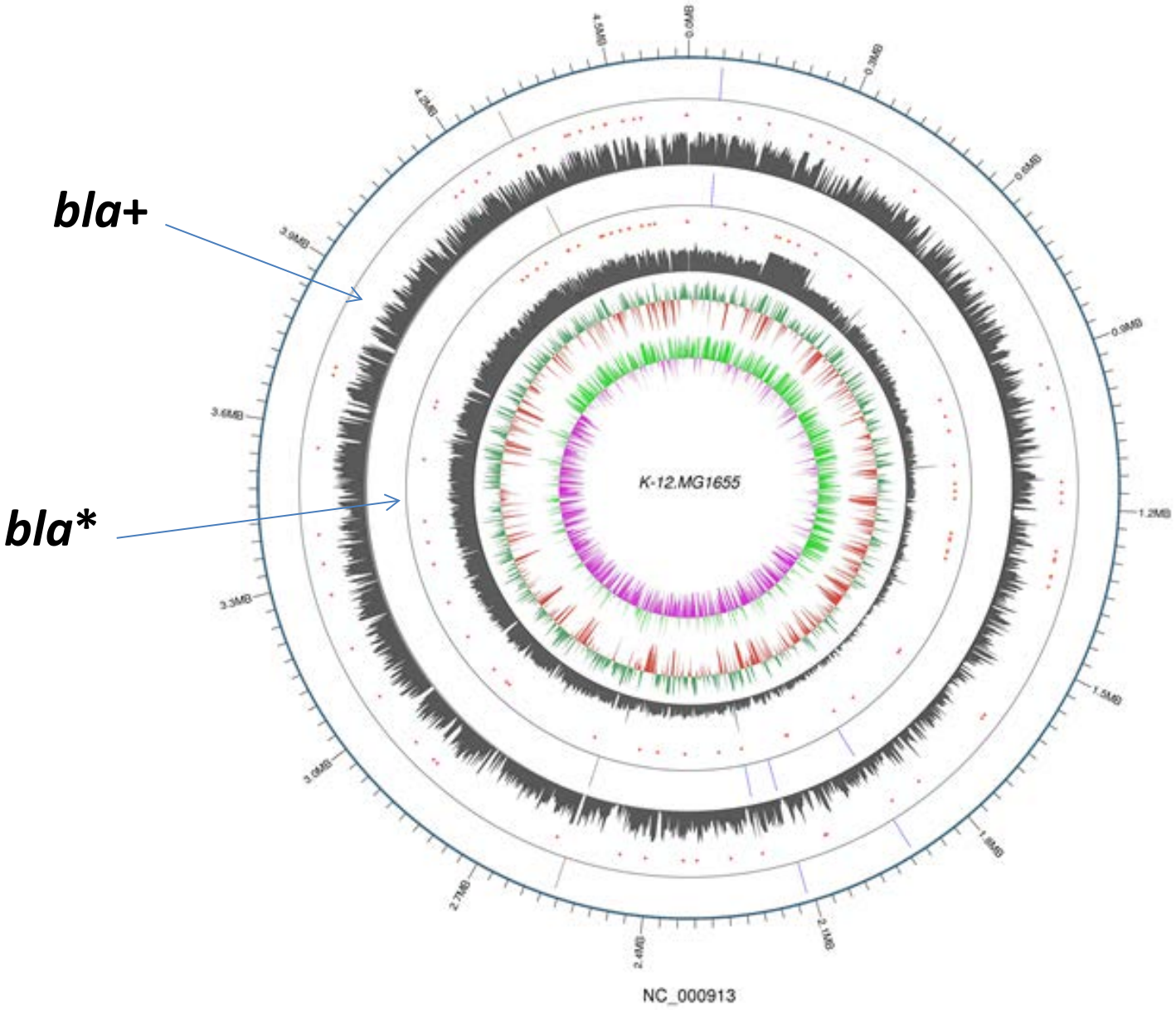


Fig 4E. The structure variation (SV) in the genome of wild-type and revertants

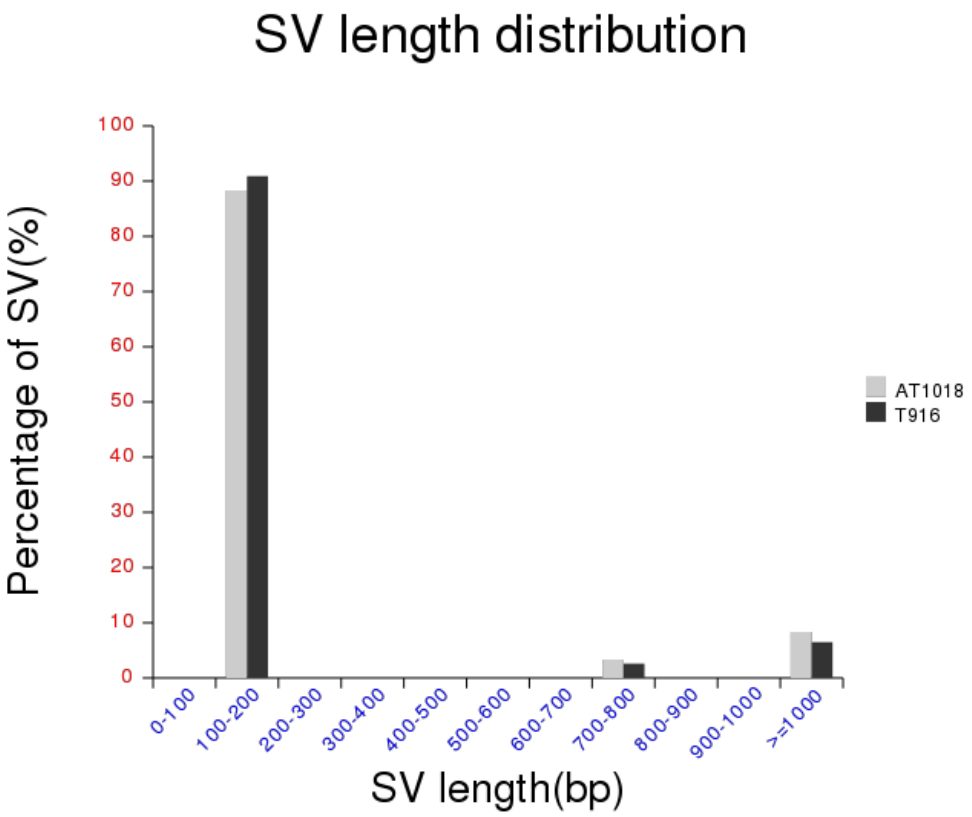


Fig 5. *bla#*

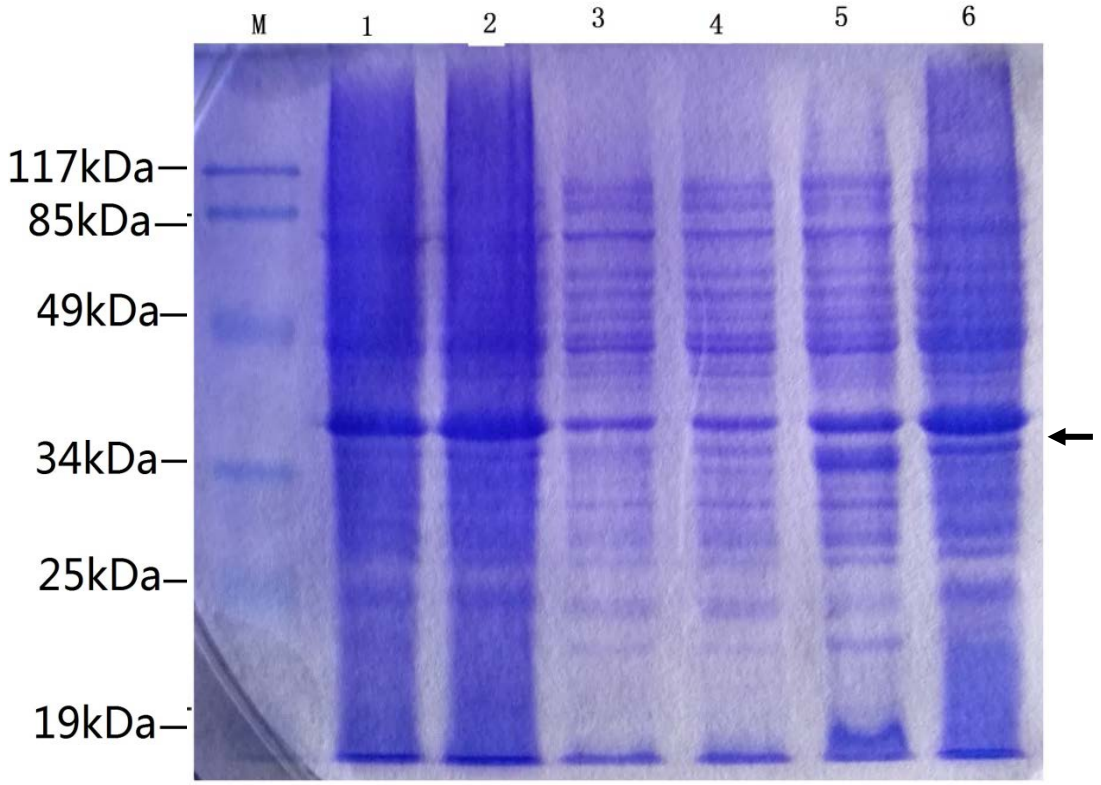
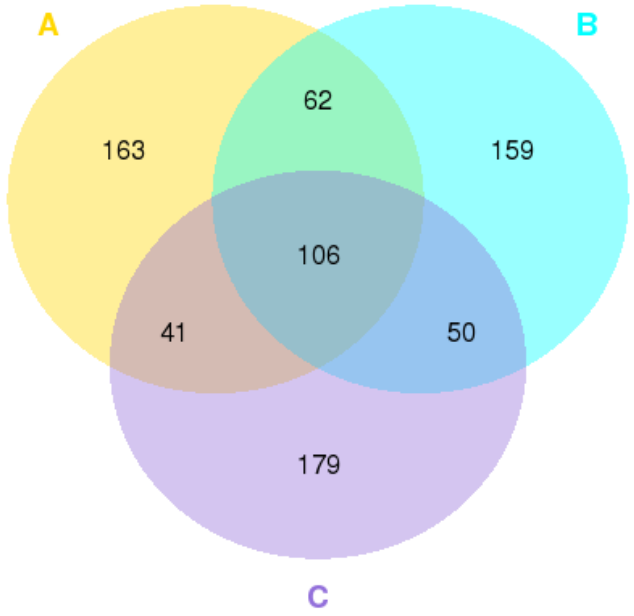
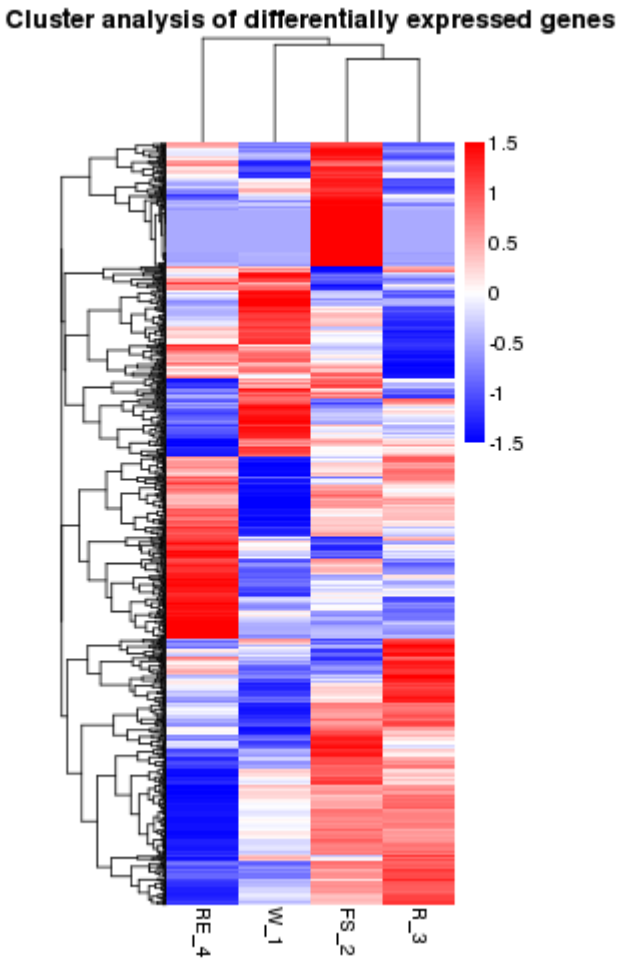


Fig 6. Transcriptome

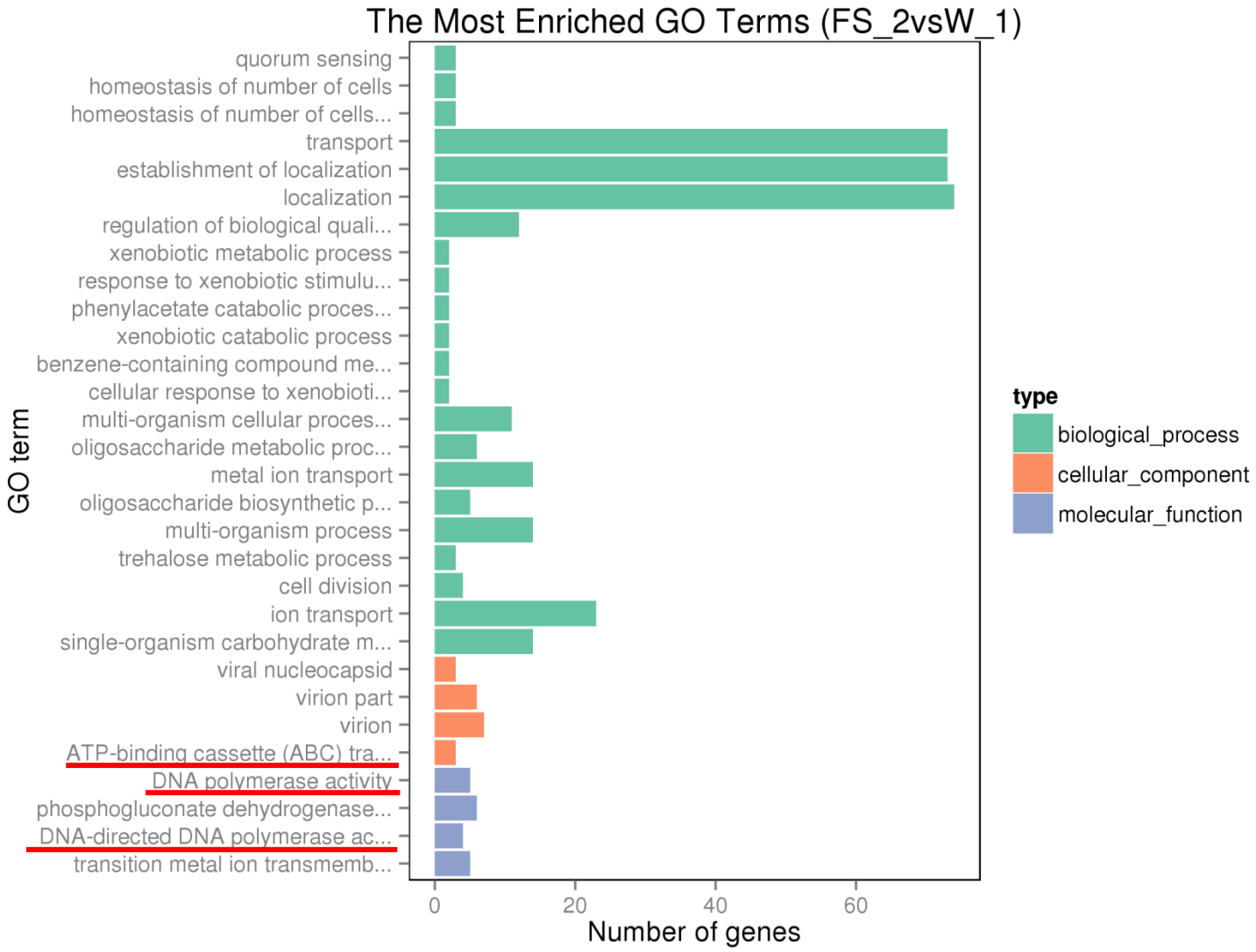
A differential expression genes (DGE)



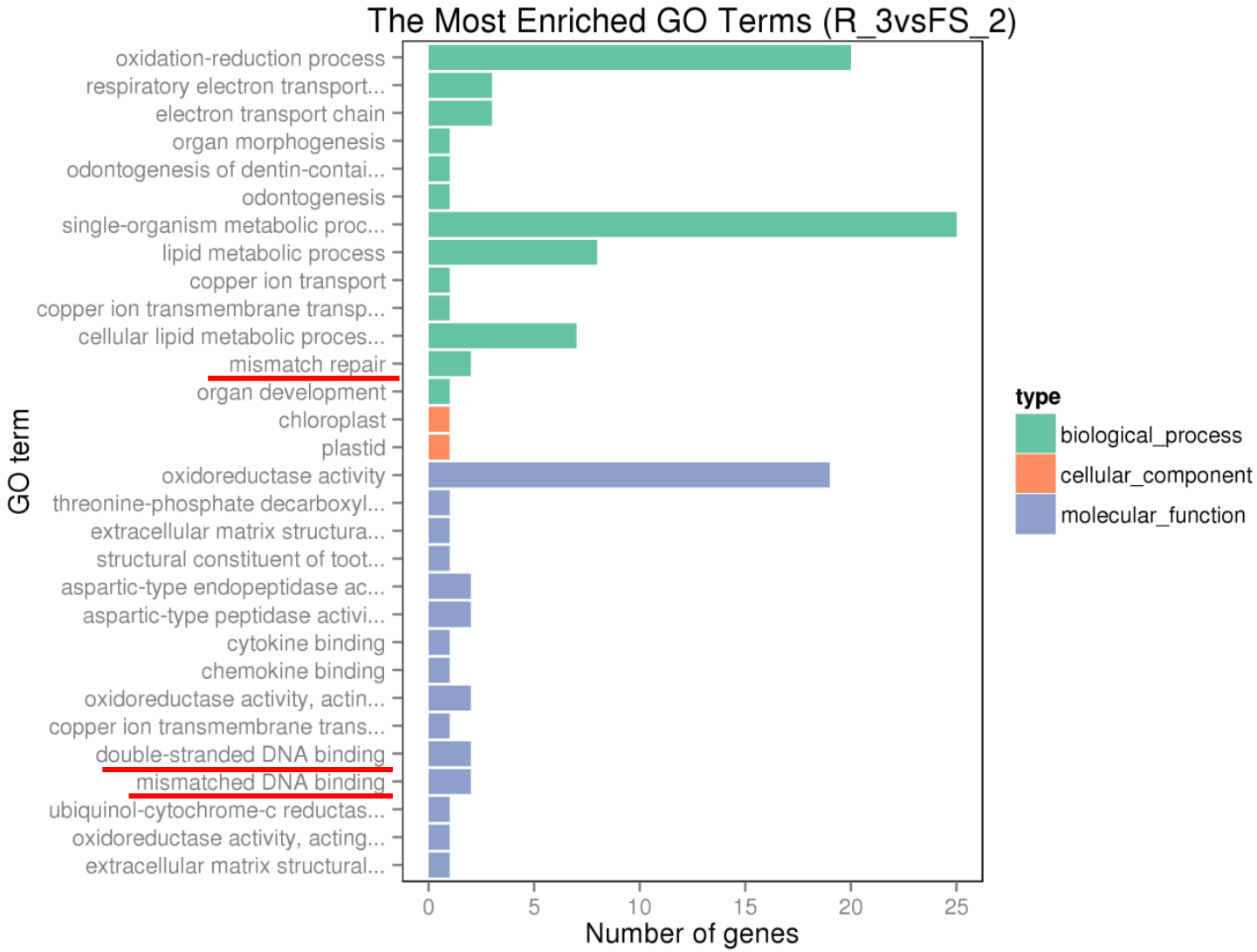
A: FS_2vsW_1
B: R_3vsW_1
C: RE_4vsW_1



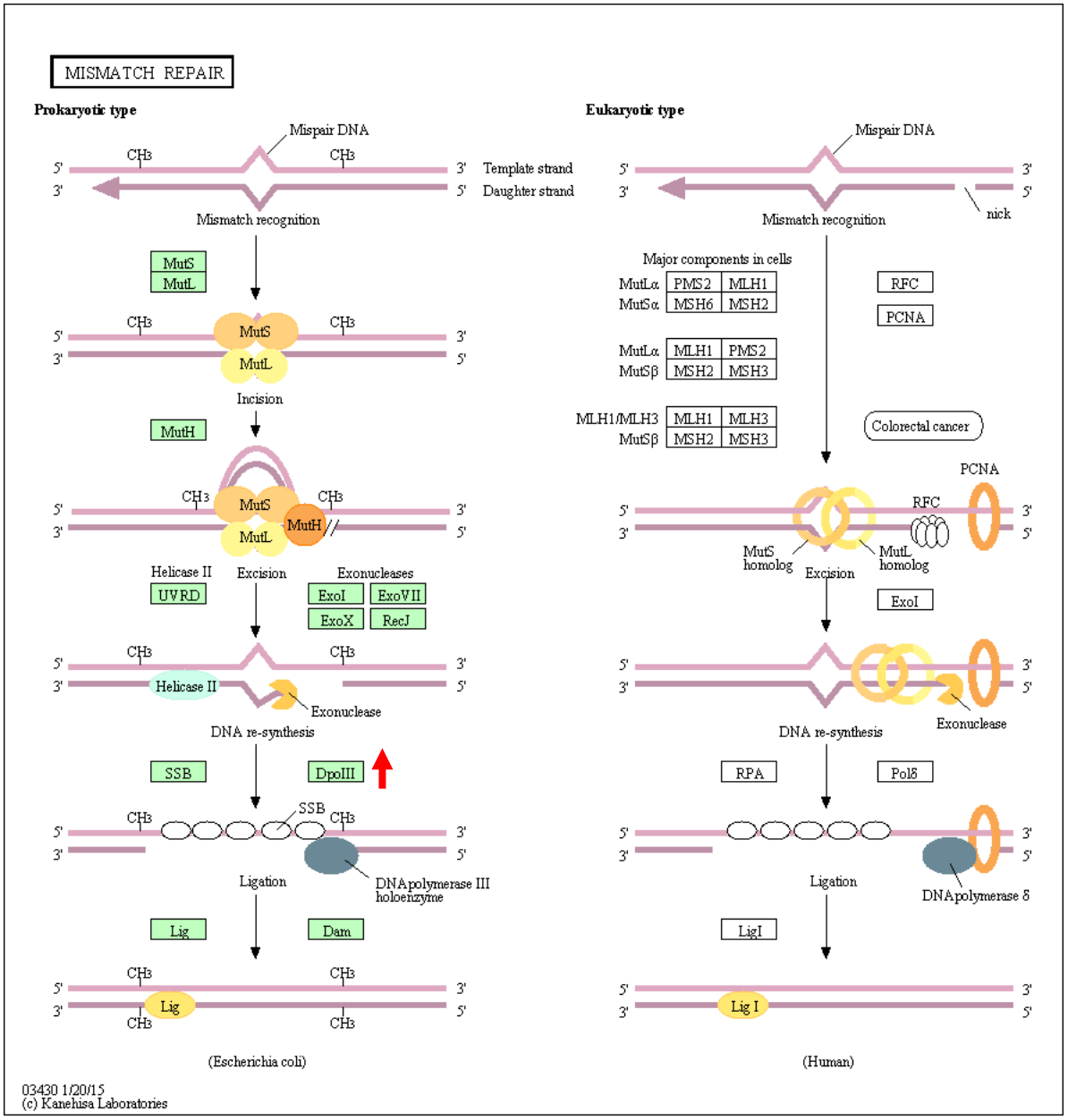
B


The Most Enriched GO Terms (FS_2vsW_1)
GO term
quorum sensing
homeostasis of number of cells
homeostasis of number of cells...
transport
establishment of localization
localization
regulation of biological quali...
xenobiotic metabolic process
response to xenobiotic stimulu...
phenylacetate catabolic proces...
xenobiotic catabolic process
benzene-containing compound me...
cellular response to xenobioti...
multi-organism cellular proces...
oligosaccharide metabolic proc...
metal ion transport
oligosaccharide biosynthetic p...
multi-organism process
trehalose metabolic process
cell division
ion transport
single-organism carbohydrate m...
viral nucleocapsid
virion part
virion
ATP-binding cassette (ABC) tra...
DNA polymerase activity
phosphogluconate dehydrogenase...
DNA-directed DNA polymerase ac...
transition metal ion transmemb...
0
20
40
60
Number of genes
type
biological_process
cellular_component
molecular_function

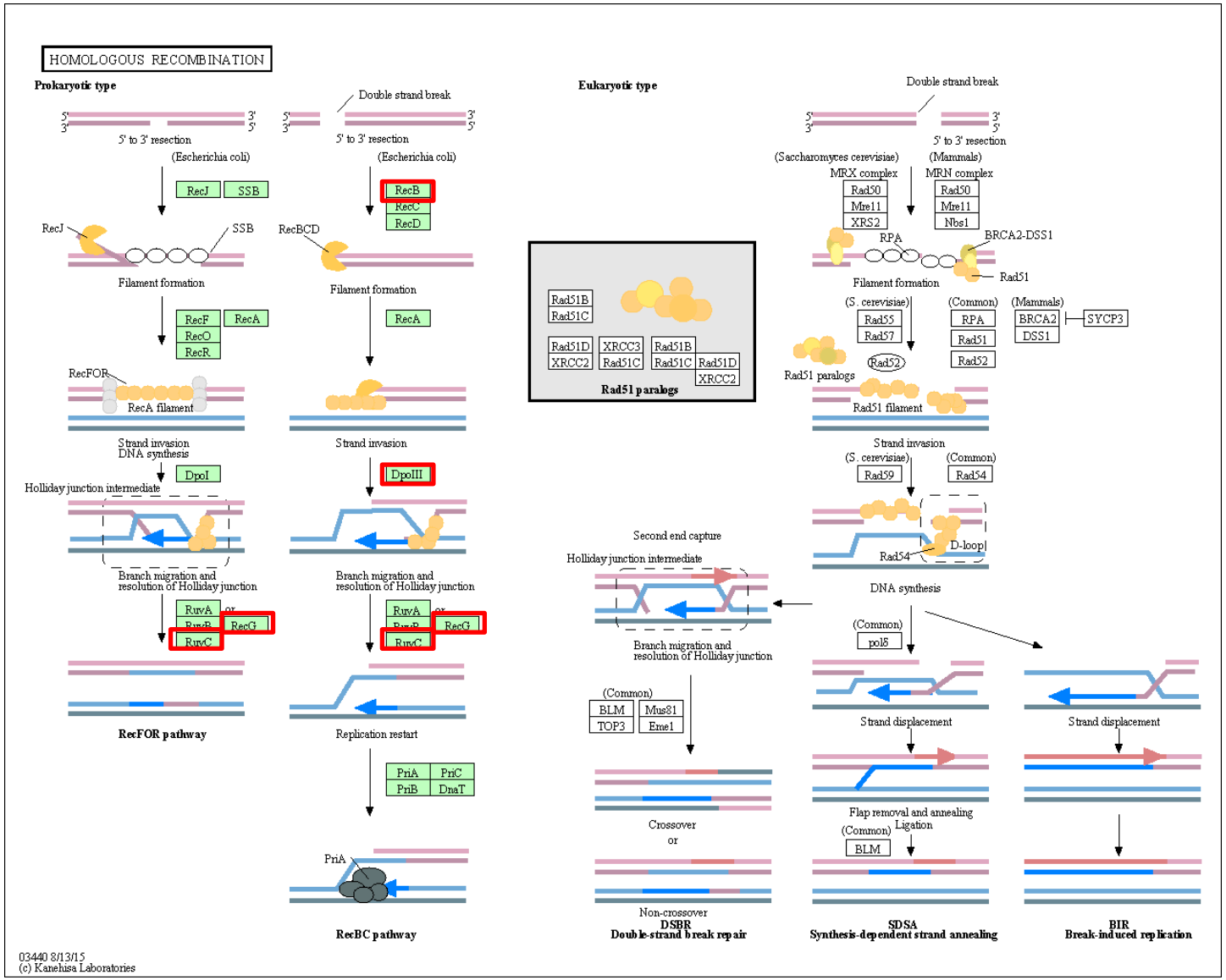
C


The Most Enriched GO Terms (R_3vsFS_2)
oxidation-reduction process
respiratory electron transport...
electron transport chain
organ morphogenesis
odontogenesis of dentin-contai...
odontogenesis
single-organism metabolic proc...
lipid metabolic process
copper ion transport
copper ion transmembrane transp...
cellular lipid metabolic proces...
mismatch repair
organ development
chloroplast
plastid
oxidoreductase activity
threonine-phosphate decarboxyl...
extracellular matrix structura...
structural constituent of toot...
aspartic-type endopeptidase ac...
aspartic-type peptidase activi...
cytokine binding
chemokine binding
oxidoreductase activity, actin...
copper ion transmembrane trans...
double-stranded DNA binding
mismatched DNA binding
ubiquinol-cytochrome-c reductas...
oxidoreductase activity, acting...
extracellular matrix structural...
GO term
0
5
10
15
20
25
Number of genes
type
biological_process
cellular_component
molecular_function

D DNA mismatch repair



E DNA homologous recombination



F RNA Degradation