Bachelor thesis

The CRISPR/Cas system in prokaryotes provides resistance against bacteriophages

Ronald Plantinga
20 augustus 2010
Supervisor: Jan Kok

Abstract

Bacteriophages are the biggest threat to bacteria, therefore they have developed several defense systems against bacteriophages that interact with one of the phases of bacteriophage replication. Recently a new mechanism was discovered: the CRISPR/Cas system. This system consists of a number of repeats on the chromosome, called clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated genes (*cas* genes). The CRISPR consist of repeats, interspersed with spacers of nearly the same length which are derived from phage genomes. It has been shown that this system provides resistance against bacteriophages. However, the exact mechanism is still not fully elucidated. The transcription of CRISPR RNAs (crRNAs) and their processing by enzyme complexes encoded by the *cas* genes are possibly important for providing resistance against bacteriophages. In this thesis the scientific progress on the CRISPR/Cas system is described and finally the possible mechanisms will be discussed.

Contents

Introduction

Viruses are the most abundant form of life (although it is discussable that viruses are a form of life) on earth. The total number of viruses on earth is estimated to be 10^{31} (Breitbart and Rohwer, 2005). Viruses are found in different environments such as deserts, hot springs and the human gut (Mc Grath et al, 2007). In 1989 it was estimated by using transmission electron microscopy, that in 1 mL of sea water, +/- 10 million virus like particles are present (Bergh et al., 1989). More recently, metagonomic studies gave additional data on the abundance of viruses in the marine environment. Using the shotgun sequencing method, huge amounts of DNA were sequenced, assembled to whole genomes and analyzed. It was shown, that in 100 L of sea water more than 5,000 viral genotypes are present and that in 1kg of marine sediment more than 1 million viral species are present (Rohwer et al., 2009). Samples of the human feacal contained ~1000 viral genotypes in one sample (Edwards and Rohwer, 2005). Interestingly, the number of viruses in the samples was nearly the same.

Viruses themselves are very diverse in composition, mechanism of replication and host range. Despite their diversity, they have some features in common. All viruses consist of genetic information, which is surrounded by a protein coat, called the capsid. The form of the genetic information is diverse: it can consist of double stranded or single stranded DNA or RNA. Also the composition of the capsid differs between viruses. It can consist of one type of protein or of different proteins with different chemical properties.

Replication is essential for a virus to reproduce itself. Viruses replicate by infecting host cells with their genetic content (ssDNA, dsDNA, ssRNA or dsRNA). This genetic content is used by the host cell to produce the viral components. Viral replication can be divided in five phases (see also Figure 1):

1. *Attachment* or absorbtion of the virus to the cell wall of the host cell.
2. *Penetration*; injection of phage DNA or RNA into the host cell.
3. *Synthesis*; production of the viral component by the host cell.
4. *Assembly and packaging* of the viruses.
5. *Release* of the viruses by lysis or excretion.

Each phase can be performed differently, depending on the virus and the host cell. The first phase, attachment of the virus to the cell wall of the host cell, is mediated by recognition of receptors which are abundant at the cell surface of the host cell.

Another division of viruses is the division between virulent, lytic viruses and temperate viruses. Virulent or lytic viruses kill their host by lysing the host cell and releasing the cell content and the produced viruses. Temperate viruses by definition do not kill the host cell. The viral DNA can be integrated in the DNA, without being expressed and can be passed to the next generation after cell division. Under certain circumstances temperate viruses can switch to the lytic pathway and become deadly for the host cell.

This article is focuses on the interplay between bacteria and viruses which infect bacteria, called bacteriophages (or simply 'phages'). Compared to bacteriophages, bacteria are less abundant in nature. In the marine environment, the number of bacteria can vary between the 10^4 and the 10^7 bacteria ml^{-1}, whereas the number of viruses lie nearly in the same range, between 10^4 and 10^8 viruses ml^{-1} (Wommack and Colwell, 2000). However, in all environments, the number of viruses seems to outnumber the bacteria. For a some of the

environments, the virus-to-bacteria ratio (VBR) was calculated. Their ratio lies mostly between 3 and 10 (Wommack and Colwell, 2000). The VBR values seem to be higher in rich environments, where bacteria grow fast and giving rise to higher amounts of bacteria.

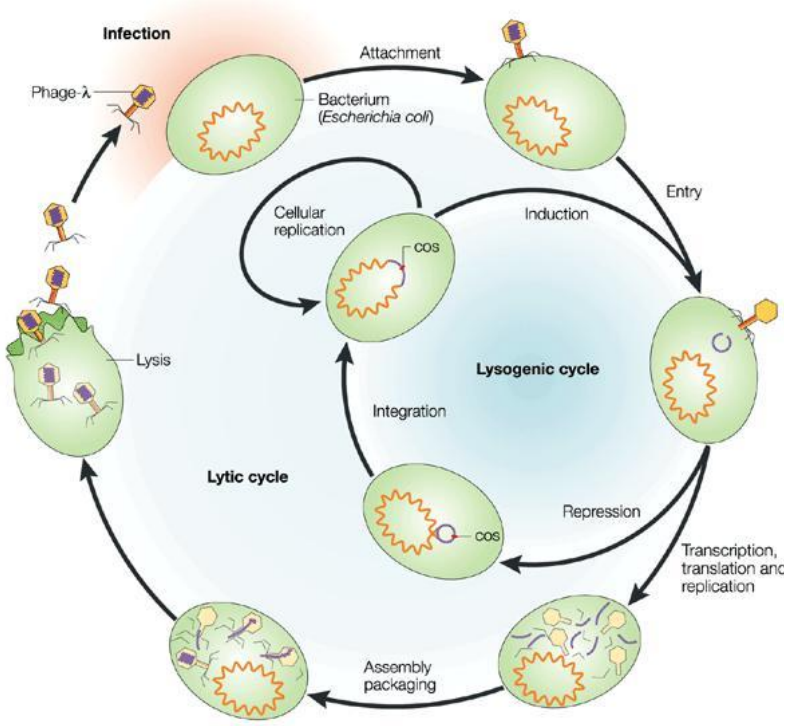


Figure 1: The replication cycle of bacteriophages is shown in this figure. It shows both the lysogenic infection and the lytic infection, in the figure refered to as Lysogenic cycle and the lytic cycle, respectively. The phage can leave the lysogenic cycle when the phage genome that is integrated in the host cell chromosome is transcribed. The big circle show the five phases of phage replication: Attachment, Entry (*Penetration* in text), Transcription, trans-lation and replication (*Synthesis* in text), Assembly packaging and finally lysis. After lysis of the host cell, the phages are released and can infect other bacteria. (Fig. from: Campbell, 2003)

It is shown that phages kill between 4% and 50% of the bacteria produced every day in a marine environment (Breitbart and Rohwer,2005). It is remarkable that despite the big threat of bacteriophages to bacteria, the latter are still very abundant on earth. This is partly due to the presence in bacteria of defense mechanisms against bacteriophages.

The currently known anti-bacteriophage mechanisms of bacteria interact each with one phase of the phage replication. Four different defense mechanisms are known to date. The first mechanism interferes with the attachment- or absorption phase of the bacteriophage replication. Bacteriophages bind to bacteria by recognizing receptors on the cell surface. In order to disturb this interaction bacteria can mask or even remove the receptors on the cell surface (Forde and Fitzgerald, 1999). The second mechanism is associated with the penetration phase of the phage replication. The host cell prevents the phage injecting its DNA. It is supposed that the host cell changes the membrane fluidity in order to hinder DNA or RNA injection (Forde and Fitzgerald, 1999). The third mechanism is active when the phage DNA is present in the host cell. This mechanism is based on restriction and modification of DNA and consists of two components: 1) restriction of the foreign DNA by a site specific restriction endonuclease and 2) protection of the host-cell DNA by methylation (Forde and Fitzgerald, 1999). Phage DNA which is not digested becomes methylated and is capable of infecting the bacterium (Forde and Fitzgerald, 1999). The fourth mechanism is called the abortive infection mechanism and is active when the infection has already taken place. This mechanism consists of a variety of mechanisms and is associated with processes such as

genome replication, transcription, translation, packing or assembly of the phages (Forde and Fitzgerald, 1999).

Recently a new defense system against bacteriophages was discovered, the CRISPR/Cas system. This system consist of clustered regularly interspaced short palindromic repeats (which stands for CRISPR) and CRISPR-associated genes, called *cas* genes. The CRISPR locus consists of region of short sequences which are derived from foreign sequences (Karginov and Hannon, 2010). It is thought that this mechanism can provide resistance against phages by transcribing the incorporated sequences into small RNAs which program an enzymatic complex recognizing the phages (Karginov and Hannon, 2010). A remarkable feature of this system is that it provides inheritable resistance against bacteriophages.

In this article the research on CRISPR will be described. Finally, the question as to what the possible working mechanisms are will be answered, based on the described studies.

The structure of the CRISPR/Cas locus

Short sequence repeats (SSRs) are common in the genomes of prokaryotes. SSRs can be divided into two classes: contiguous repeats and interspersed repeats. Contiguous repeats consist of units adjacent to each other, interspersed repeats are separated by other sequences which are different from the repeats (Jansen et al., 2002-I). In 1987 another class of interspersed SSRs was found, these were the first CRISPR-like repeats described. The researchers found an unusual repeat sequence near the *iap* gene in *E. coli*, consisting of five homologues repeat sequences of 29 base pairs interspersed with spacers of the same length (Ishino et al., 1987). Later, similar repeats were found in other species such as *Haloferax mediterranei, Streptococcus pyogenes, Anabaena* sp. PCC 7120 and *Mycobacterium tuberculosis* (Jansen et al., 2002-II).

In 2000, the earlier found CRISPR-like repeats were described as a new family of prokaryotic repeats by Mojica et al. The general characteristic of these repeats is that they are regularly spaced by intervening sequences of constant length (Mojica et al., 2000). Mojica et al. (2000) refer to them as Short Regularly Spaced Repeats (SRSRs). Using a computer program, they sought in complete genomes and found SRSRs in 9 archaeal- and 10 bacterial species (Mojica et al., 2000). The species which contained SRSRs represent microorganisms of different phylogenetic groups, so the SRSRs are widespread among different archaeal- and bacterial species.

The complete CRISPR locus was identified by Jansen et al. in 2002. Since that time the acronym SRSR (and the acronym SPIDR (Jansen et al., 2002-I) is replaced by the acronym CRISPR, which stands for clustered regularly interspaced short palindromic repeats (Jansen et al., 2002-II). CRISPR-loci seems to be very abundant in archaeal and bacterial species. Approximately 40% of the bacterial genomes contain a CRISPR locus (Kunin et al., 2007). The number of CRISPR-loci in one organism varies between 1 and 20 (Jansen et al., 2002-II).

The CRISPR locus consists of a region with repeats, interspersed with spacers, a leader sequence and genes which encode CRISPR-associated proteins (*cas* genes). Below,

these different components of the CRISPR locus, as defined by Jansen et al., will be discussed (see also Figure 2):

Repeats. The length of the repeats was found to be between 21 and 37 bp and they are almost identical in one CRISPR locus, although most repeats on the end of the locus contained mutations (Jansen et al., 2002-II). Repeats are also similar between related species but started to be more different between distantly related species (Jansen et al., 2002-II). However, some distantly related species still share repeat sequences (Kunin et al., 2007). Part of the repeats contain a palindromic sequence and hence have a predicted secondary structure (Kunin et al., 2007; Horvath et al., 2008). Using the RNAfold program, a folding score was calculated. Of the 167 analyzed organisms, in 66 of them the repeat sequences had a high folding score, 41 had an intermediate folding score and 60 had a low folding score (Kunin et al., 2007).

Spacers. Between the repeat sequences lie spacer sequences of similar length as the repeat sequences and the other spacers (Jansen et al., 2002-II). In contrast to the repeats, the spacer sequences differ within an CRISPR locus. Only in one case, spacers had duplicates within a CRISPR locus.

Leader sequence. At one site of the CRISPR locus, a leader sequence of several hundred base pairs is located. The sequence is always located upstream the CRISPR locus and has the same orientation. Between the analyzed organisms, the leader sequence has approximately 80% sequence identity (Jansen et al., 2002-II).

***Cas* genes.** The CRISPR locus is flanked by CRISPR-associated genes (*cas* genes), which are found on either side of the locus. These genes are clearly associated with the CRISPR locus. Organisms which contain a CRISPR locus always possess at least one *cas* gene and organisms without a CRISPR locus never possess *cas* gene (Jansen et al., 2002-II). Moreover, no CRISPR genes were found in eukaryotic genomes and, nor associated *cas* homologues. If an organism possess more than one CRISPR loci, there is only one set of *cas* genes near one of the CRISPR loci.

Comparison between the *cas* genes revealed that there are four different *cas* genes: *cas1* to *cas4* (Jansen et al., 2002-II). In different organisms, the *cas* genes are in different arrangements on the chromosome and the *cas* genes are not always present at the same time (Jansen et al.,2002-II; Haft et al., 2005). Later, the *cas* genes were classified in more detail by Haft at al. and Makarova et al. Apart from *cas1-4* two other *cas* genes were defined: *cas* 5 and *cas* 6 (Haft et al., 2005; Bolotin et al., 2005). These six *cas* genes were called the *cas*-core genes by Haft et al, and are shared between different organisms. In addition to these core genes there are sort-specific *cas* genes, which are named after the organism, like *cse1-4* for the *E. coli* subtype and *csy1-4* and for *Y. Pestis* (Haft et al., 2005). Finally, there is a group of Repeat-associated-mysterious-proteins (RAMPs) genes (Haft et al., 2005; Makarova et al., 2006). These genes have a very loose conservation, and only share the RAMP signature and a G-rich loop at the C-terminus (Makarova et al., 2006). All organisms possess at least one

RAMP gene and they lie between the other *cas* genes, or on the adjacent to the *cas* genes on the chromosome (Makarova et al., 2006).

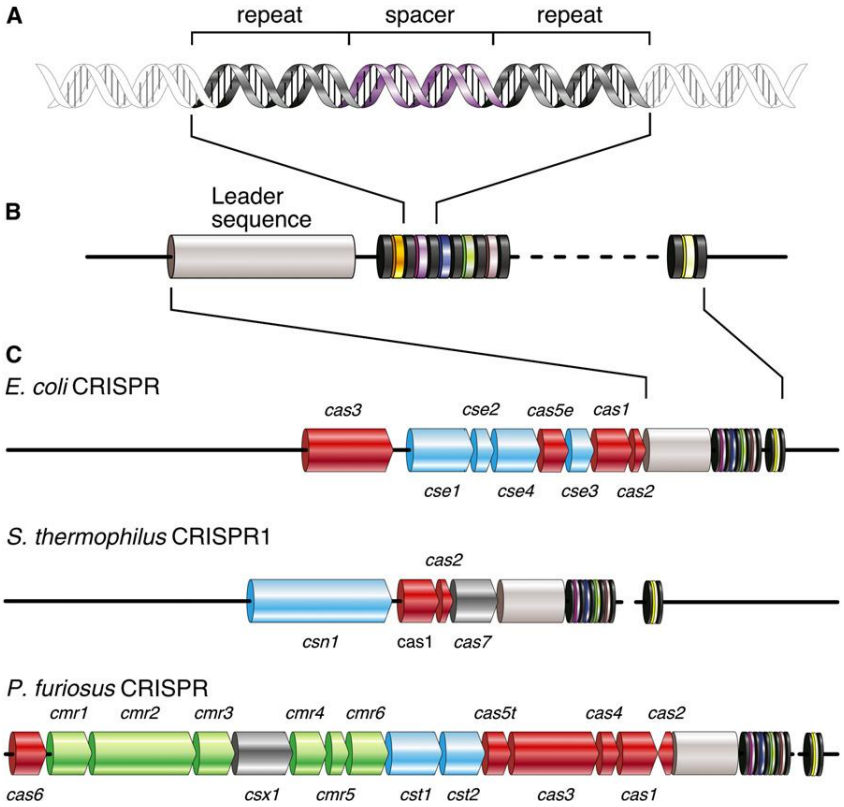


Figure 2: The structure of the CRISPR locus. (A) The repeat-spacer region, both has a length between 21 and 37 bp. (B) A CRISPR locus contains between 2 and 120 several repeats and spacers. Upstream the repeat-spacer region lies the leader se-quences, a non coding part of the DNA of several hundreds of base pairs. (C) The spacer-repeat region and the leader se-quence are accompanied with several *Cas* genes which are differently orienta-ted o the chromosome, as shown for *E. coli*, *S. thermophilus* and *P. furiosus*. (Fig. from Karginov and Hannon, 2010)

The composition of the CRISPR locus is varies between organisms. Each CRISPR locus consists of a repeat-spacer region and associated genes (*cas* genes), but the number and composition varies between every CRISPR loci. The repeat-spacer region can contain between two and 120 repeats interspersed by spacers (Jansen et al., 2002-II). Repeat sequences are most conserved between species and the spacer sequences vary between strains and even between individual cells.

The function of the CRISPR/Cas system

Spacers are derived from phage genomes

Comparing the spacer sequences of CRISPR loci with sequences of genome data-bases revealed that some spacers have homology with bacteriophage sequences, plasmid sequences and chromosomal sequences (Bolotin et al., 2005; Díez-Villaseñor et al., 2009; Horvath et al., 2008; Mojica et al., 2005; Mojica et al., 2008; Pourcel et al., 2004). In 2005 Mojica et al. analyzed the CRISPR loci of 67 strains. In their study, they amplified the CRISPR loci by using PCR and sequenced the amplified products. They found 4500 CRISPR spacers in 67 strains. 88 of these spacers showed homology with known sequences; 47 of the matching spacers matched bacteriophage sequences, 10 matched plasmid DNA and 31 matched chromosomal DNA of other species (Mojica et al., 2005). Other studies found similar results. Of the 349 found spacers, in *Streptococcus thermophilus* and *Streptococcus vestibularis*, 124

had significant homology (E<0.001) with sequences from the NCBI database (Bolotin et al., 2005). Most of the homologues sequences (75%) were homologues to phage genomes of *Streptococci* (Bolotin et al., 2005). A minor part of the matching sequences (20%) had homology to plasmids of *S. thermophilus* and *S. vestibularis* (Bolotin et al., 2005).

CRISPRs of different related organisms were compared to give a clue about the function of the spacer sequences. In a study the CRISPR spacer sequences of 98 different *Y. pestis* strains were compared (Pourcel et al., 2004). The researchers identified and named the spacers. At the start of the repeat-spacer region (so near the leading sequence), 6 spacer sequences (called a-f) were conserved in 43 of 98 strains. The 55 other strains contained at least 2 of this group of conserved spacers (see Fig. 3). In contrast, at the end of the repeat-spacer region the spacers were much less conserved. Two spacers at the end of the repeat-spacer region were conserved in 27 of the 98 strains and the rest of the strains contained different spacers in a different composition (Pourcel et al., 2004). Comparisons between

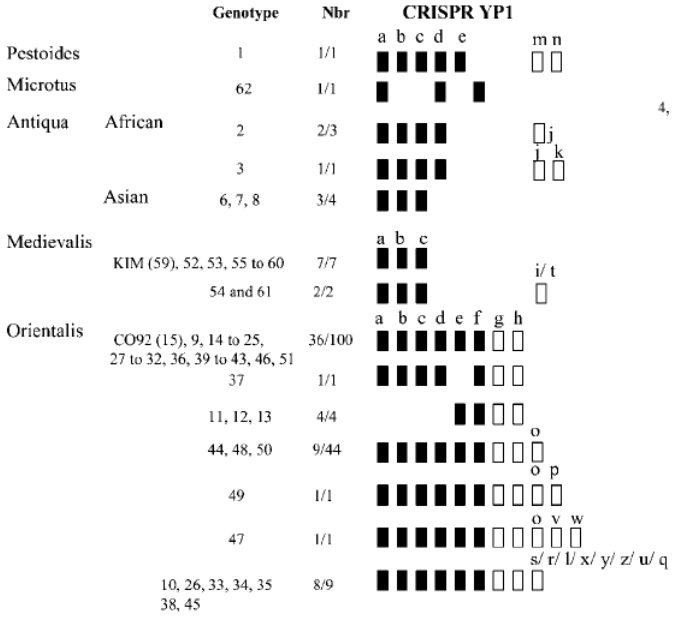


Figure 3: The composition of the spacers on the CRISPR YP1 locus of 60 *Y. pestis* strains is shown in this figure. On the left are the used strains indicated by biovar and Genotype. 'Nbr' indicates the number of alleles sequences of the total number of genotypes. Identified spacers are indicated by boxes and named with letters (a-q). The black boxes are the most conserved spacers, in total 6, which are abundant in all species, the white boxes indicate the spacers which are more unique. (Fig. from: Pourcel et al, 2005)

different strains *E. coli* and *S. thermophilus* showed similar results. In *S. thermophilus* a total of 952 spacers was analyzed(Horvath et al., 2008). The spacers at the beginning of the CRISPR repeat-spacer region showed to be the most conserved between the different strains. Spacers at the end of the repeat-spacer region were less conserved. In different strains of *E. coli* the same was observed (Díez-Villaseñor et al., 2009). At the beginning of the repeat-spacer region a number of spacers are conserved between a part of the 58 analyzed strains, but they are more divergent compared to the start of the repeat-spacer region of *Y. pestis* and *S. thermophilus*. Remarkably, the differences between the members of one group are mostly due to the deletion of spacers.

So, at the start of the repeat-spacer region, more similarity is found between the spacers of different strains than at the end of the repeat-spacer region. The differences in spacer composition at the start of the repeat-spacer region are due to the deletion of spacers. From this we can conclude that spacers are added at the end of the repeat-spacer region and are sometimes deleted at the start of the repeat-spacer region. This also shows that different

strains are derived from a common ancestor and have acquired new spacers into their genome while living in different environments.

Spacer acquisition after phage challenge

The research described above had mainly a bioinformatic approach, but to show the relation between phage infection and spacer acquisition a more experimental approach was applied. Deveau et al. showed in 2008 the direct link between phage exposure and spacer acquisition. In their study, they exposed a wild type strain (*S. thermophilus* DSCC7710) separately to two different phages (*pac*-type phages 2972 and 858). Subsequently,

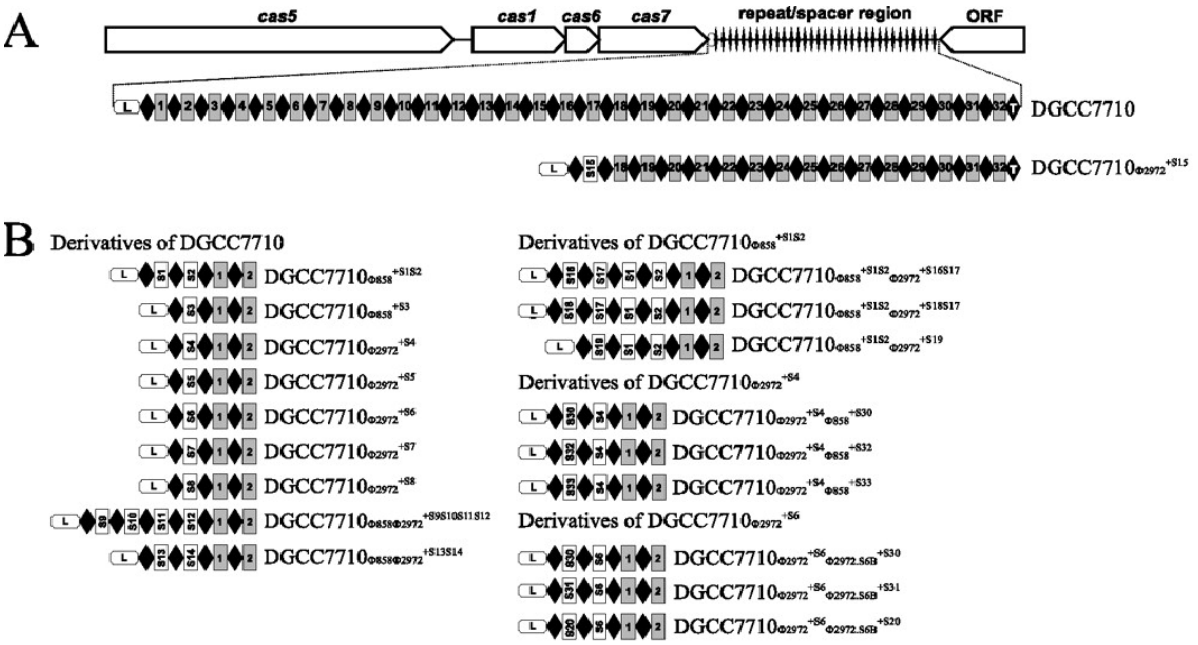
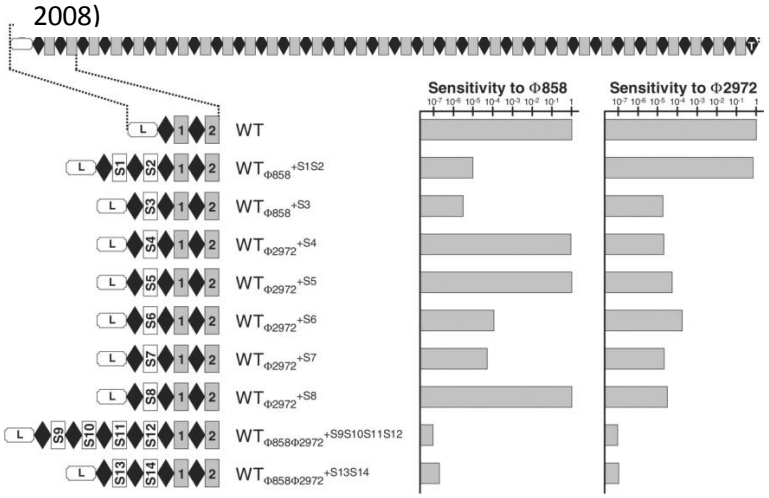


Figure 4: In this figure the spacer acquisition of the different BIMs obtained after phage challenge are shown. (A) The composition of the CRISPR locus in *S. thermophilus* DGCC7710 and *S. thermophilus* DGCC7710 ϕ2972^{+S15}. Repeats are indicated by black squares, spacers are indicated by gray boxes, numbered 1 to 32. (B) Derivatives of different strains exposed to respectively *pac*-type phage 872 and 2972. The spacers which are taken up by the BIMs are indicated by white boxes numbered S1-23. The list of BIMS on the left, derived of DGCC7710, are BIMs from the first generation. The list of BIMs on the left are BIMs of the second

[] 2008)



Figuur 5: In this figure the phage sensitivity of the different *S. thermophilus* BIMS obtained after phage exposure is shown. The upper part of the figure shows the composition of the CRISPR locus. The lower part of the figure shows the sensitivity of different BIMS to *pac*-phages 858 and 2972 in EOP. (Fig. from: Barrangou et al, 2007)

bacteriophage-insensitive-mutants (BIMs) were isolated and the CRISPR loci were sequenced. These were BIMs of the first generation. In total three BIMs were isolated after the exposure to phage 848, four BIMs were isolated after the 2972 exposure and two BIMs were isolated after the exposure to both phages (Deveau et al., 2008). Each BIM had acquired different spacers into their CRISPR locus (see Fig. 4). Subsequently, the different BIMs of the first generation were exposed to another phage to obtain six second generation BIMs. These strains also acquired new spacers. All obtained BIMs (of the first- and of the second generation) were resistant for both phages they were exposed to. Moreover, the efficiency of plaquing (EOP) for the phages that the BIMs were exposed to, was reduced for all the BIMs in the order of 10^4 to 10^5 (see Fig. 5) (Barrangou et al., 2007; Deveau et al., 2008). The EOP is an indication of the efficiency of phage infection. It is based on the number of plaques on a confluently grown agar plate.

The newly acquired spacers of a total of 30 BIMs were analyzed. Of the 30 BIMs, 21 had acquired one new spacer, 7 had acquired two new spacers, one had acquired three new spacers and one had acquired four new spacers into its CRISPR locus (Deveau et al., 2008). In a great part of the BIMS (27 out of 30 BIMs), the original 31 spacers of the wild type strain (*S. thermophilus* DSCC7710) were still abundant. One BIM lost 17 of the original wild type spacers, two BIMs lost 7 of the 31 original wild type spacers (Deveau et al., 2008). These results confirm earlier observations, namely that the spacers at the start of the repeat-spacer region are conserved and that the newly acquired spacers are added at end of the repeat-spacer region. Also the observations that spacers are deleted is confirmed by these results. The difference is that the changes in the CRISPR loci are now directly linked to bacteriophage infection.

Finally Deveau et al. compared the newly acquired spacers with the genomes of the infecting phages (*pac*-type phages 2972 and 858). This showed that 37 out of 39 analyzed spacers had 100% identity with one of the genomes of the infecting phages (Deveau et al., 2008). The two spacers which did not match had one mismatch with the phage genomes. Subsequently the region of the matching sequence of the phage genome was analyzed. The matching sequence in the phage genome is called proto-spacer. Deveau et al. (2008) identified a sequence which is always present two nucleotides downstream of the proto-spacers, called the proto-spacer-associated motif (PAM). They showed that 34 of the 39 analyzed proto-spacers had an AGAAW (W stands for either a Adenine or Thymine on that position) motif at the 3' end. Other studies also identified conserved sequences near the proto-spacers. These sequences were shown to be NGG (Mojica et al., 2008) or NGGNG, both in *S. thermophilus* infecting phages (Horvath et al., 2008). The differences in PAM sequences are probably specific to the analyzed bacteriophages and bacteria. If there is a conserved PAM, this sequence possibly plays a role in recognizing phage DNA by certain factors.

It was also shown that bacteria with CRISPRs containing spacers matching the template DNA of the phages are less sensitive to phages than CRISPR loci containing spacers matching the coding DNA (Brouns et al., 2008). In the experiment, artificial CRISPRs against phage lambda were designed and subsequently the EOP was derived, 8 CRISPRs containing spacers derived from non-coding phage DNA and coding phage DNA corresponding to

essential phage genes. The sensitivity to phage lambda varied between an EOP of 1 (no resistance) for most of the coding-sequence derived spacer and 10^{-4} for a non-coding - sequence derived spacer (Brouns et al., 2008). Thus the non-coding DNA provides higher resistance to phages.

Spacer acquisition at the population level

Deveau et al. (2008) showed under laboratory conditions the relation between phage exposure and spacer acquisition. It was shown that this process also takes place under environmental conditions (Andersson and Banfield, 2008; Tyson and Banfield, 2008). In metagenomic studies, large amounts of DNA from two different spots were sequenced and complete genome sequences were reconstructed. A group of the *Leptospirillum* species was identified (Tyson et al., 2004). Subsequently CRISPR loci from the two different spots were identified and yielded a number of reconstructed CRISPR strains. The CRISPR loci of the strains from two different samples were compared to each other (see Fig. 6 for the outline of this experiment). There are some similarities between all CRISPRs found. Conforming to the earlier results, the spacers at the beginning of the repeat-spacer region of the CRISPR are shared among most strains. The spacers in the middle of the repeat-spacer region are only conserved in the strains originating from the same spot and the spacers at the end were almost all strain-specific (Tyson and Banfield, 2008).

In another study, Andersson and Banfield analyzed the CRISPR loci of two samples and compared them with the sequences of bacteriophages in the same two samples. They found 6044 spacer sequences in 37 CRISPRs and among them were 2348 unique spacers (Andersson and Banfield, 2008). These sequences were compared to Spacer-containing non-CRISPR (SNP) contigs, these are sequences with homology to spacer sequences which are not part of any CRISPR region. Subsequently contigs were derived from the sequences of two different samples from the soil. The contigs were grouped based on tetranucleotide frequencies (GC-content) and sub grouped based on mate-pairs. The subgroups (AMDV1-5) were seen as different bacteriophages. Of all the analyzed spacers, up to 40% matched (100% identity) to the found contigs. Interestingly, the two samples were taken six months after each other and already showed signs of new spacer acquisition and spacer loss. These results seemed questionable since the samples were not taken on exactly the same location. But still, these results show the dynamics of spacer acquisition and spacers loss at the population level. The spacer matches between the bacteriophage genomes are much higher, compared to the NCBI database searches (40% against 2-35%). This difference in percentage of matching spacers is clearly a symptom of the incompleteness of the NCBI database and a hallmark of viral diversity.

It is clear that the CRISPR locus takes up DNA sequences of phages by adding new spacers to their genome. This is show by matching the spacers to already known phage sequences, by the observation of newly acquired spacers after infection of known phages under laboratory conditions, and also by metagenomic studies on the population level. This suggests that the CRISPR loci are involved in providing phage resistance. In the next chapter the mechanisms underlying the supposed phage resistance are explained.

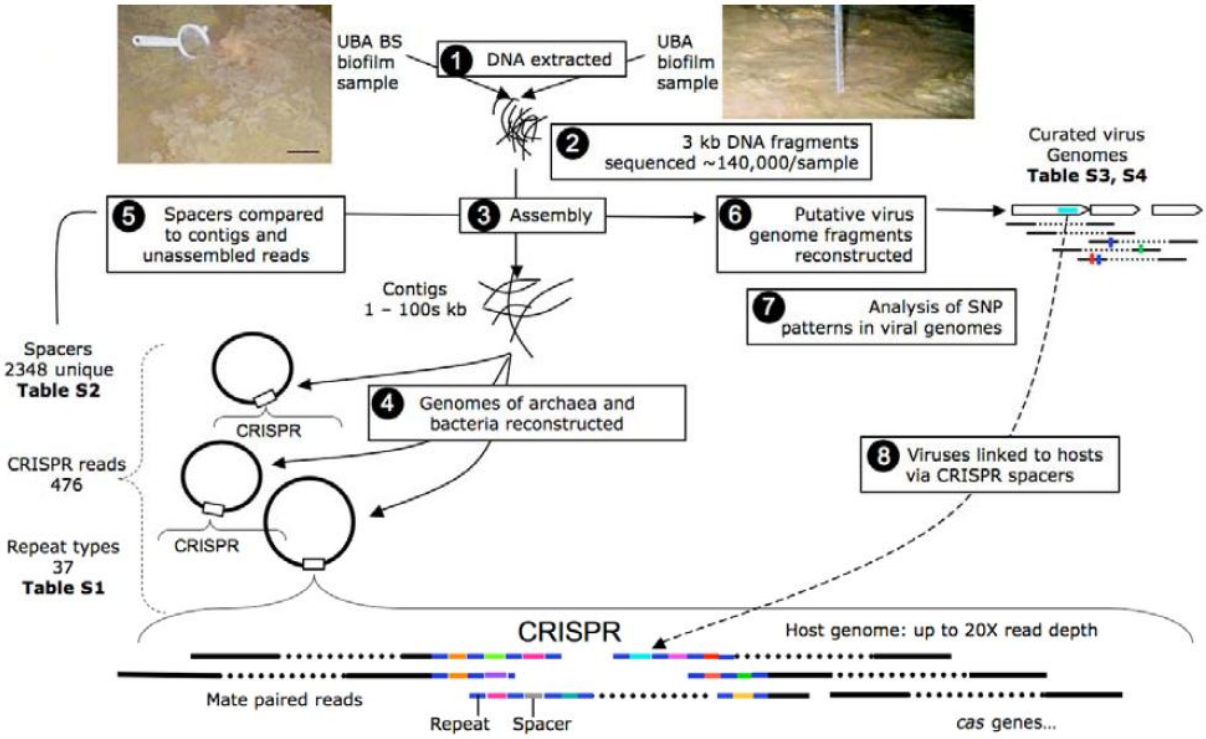


Figure 6: The experimental set-up of the metagonomic experiment of Anderson and Banfield. In this experiment first, the DNA of two different samples was extracted (1), sequenced (2)and assembled to whole sequences of archaea, bacteria (4) and viruses (6). The spacers of the CRISPR loci of the reconstructed archaean and bacterial genomes were compared with contigs of unassembled reads (5) and viral genomes (8). (Fig. from: supplemental data of Anderson and Banfield, 2008)

The mechanisms of the CRISPR/Cas system

The functions of the Cas proteins

We have seen that the CRISPR/Cas system provides resistance to bacteriophages by acquisition of spacers sequences. Beside the spacer sequences, *cas* genes are found near CRISPR-loci in all organisms containing them (Jansen et al., 2002-II). The *cas* genes are divided into different classes, core-*cas* genes of which one (*cas1i)* is shared among all CRISPR-containing organisms, species specific *cas* genes and RAMP-genes (Haft et al., 2005; Makarova et al., 2006). In silico analyses of Makarova et al. (2006) gave some clues about the functions of the Cas proteins. Using the COG (Clusters of Orthologues groups) database, the Cas proteins were divided into different families with different functions. The Clusters of Orthologues groups are conserved domains of proteins, based on the comparison of 138458 proteins (Tatusov et al., 2003).

Cas1 proved to be the most conserved protein of all the Cas proteins and is conserved between all CRISPR-containing organisms (so in all archaeal species) it has a predicted nuclease/integrase function (Makarova et al., 2006). C*as* genes 2-4, are also abundant in species of different phylogenetic groups but are not abundant in all CRISPR-containing species, like Cas 1 (Makarova et al., 2006). The functions of Cas2-4 are thought to be respectively a putative helicase (Cas2), a DNA helicase (Cas3) and a RecB-like nuclease

(Cas4) (Makarova et al., 2006). *Cas* genes 1-4 are, because of their high abundance, called the *cas* core genes (Haft et al., 2005). Because of the high conservation of c*as1*, the corresponding protein seems to be essential for the function of the CRISPR/Cas system.

In contrast to the widely conserved *cas* core genes, the *RAMP* genes are much less conserved. Whereas the core *cas* genes can be identified with one (*cas1)* or two (cas1-3) COG groups, the RAMP proteins belong to 10 COG-groups (Makarova et al., 2006). The RAMP proteins are shown to have similarity to ferrodoxin fold proteins and are therefore thought to be a putative RNA binding protein (Makarova et al., 2006).

The role of Cas proteins

These results are just an indication about the function of the Cas proteins. Experimental data gave a clearer insight into the possible functions of the different Cas proteins and other components of the CRISPR locus in *S. thermophilus*. In their study, Barrangou et al. (2007) knocked out several components of the CRISPR locus and determined the sensitivity of the *S. thermophilus* strains to *pac*-phages 858 and 2972 (the different synthetic CRISPRs and their effect on phage resistance are summarized in Fig. 7).

After removing the repeat-spacer region from the CRISPR locus phage resistance was lost completely (Barrangou et al., 2007). So, the repeat-spacer region and the *cas* genes of the CRISPR locus provide resistance to phages. To show that the two spacers which were taken up after the phage challenge, are essential to phage resistance, these two spacers were added to the CRISPR locus of a strain that is sensitive to that particular phage. These new strains were resistant to that particular phage (Barrangou et al., 2007).

The resistance of the strain was also affected when a sequence was placed between the repeat-spacer region and the *cas* genes (Barrangou et al., 2007). So the composition of the CRISPR locus proved to be essential for the functioning of the CRISPR system. The researchers also inactivated *cas5* and *cas7* and determined the sensitivity against *pac*-phage 858 and phage 2972. Inactivation of *cas5* resulted in loss of phage resistance while inactivation of *cas7* had no effect on the resistance to phage 858 (Barrangou et al., 2007). Cas5 was predicted to be a nuclease, because it contains a helix-turn-helix domain (Barrangou et al., 2007). It seems that the nuclease function of Cas5 is essential to provide phage resistance and that Cas7 plays another role in the functioning of the CRISPR/Cas system.

The role of RNA

It has been shown that RNA is transcribed from the repeat-spacer sequences (Tang et al., 2002). In this study, total RNA of the organism was isolated and a Northern blot was performed with the sequences of the repeats as probe. The transcribed RNAs showed to be of discrete lengths, namely multiples of 68 (68, 136, 204, 272, 340 and 408 nucleotides) (Tang et al., 2002).

Small RNA molecules are also transcribed from the CRISPR locus (Brouns et al., 2008; Hale et al., 2009; Lillestøl et al., 2006). In the archeon *S. acidocaldarius* the transcribed RNAs of the CRISPR locus also had discrete lengths of respectively 72, 115 and 180 bp. Remarkably, the RNAs extracted from bacteria grown to stationary phase contained an extra RNA molecule of 36 basepairs (Lillestøl et al., 2006). This suggests that during the stationary

phase, a different processing step is done by the (endo)nucleases. It is clear however that the RNAs are processed from a big transcript into shorter fragments. The small RNA products seemed to be derived from a large RNA product of 180 bp which is present in the samples taken from the stationary- and exponential phases. If the smallest RNA fragment is 36 bp, the large RNA fragment of 180 bp is cut at 4 sites to yield 5 RNA fragments of 36 bp (these results are summarized in Fig. 8).

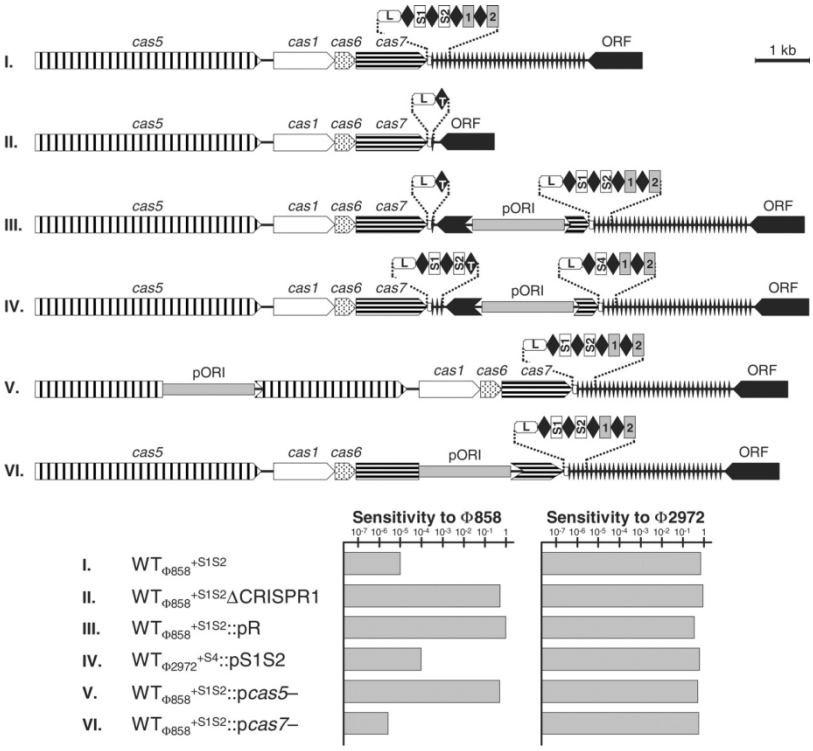


Figure 7: This figure illustrates the experiment of Barrangou et al. Different CRISPRs containing spacers providing resistance against phage 858 were constructed and their phage sensitivity was (EOP) measured. The upper part of the figure shows the composition of the CRISPRs (I-VI). The lower part of the figure shows the sensitivity of the different strains to phage 858 and 2972 as a control. (Fig. from: Barrangou et al., 2007)

Hale et al. (2009) analyzed the RNA products of the CRISPR loci in *Pyrococcus furiocus*. They isolated 20 short RNA's (from now on called crRNA; CRISPR RNA) of 45 nucleotides and 31 crRNAs of 39 nt. and sequenced them. It appeared that all crRNAs contained the same part of the repeat sequence of 8 nt. long and a part the spacer sequences. The length of the spacer sequence which is part of the crRNA, is shorter or longer, depending on whether it is part of the 39 nt. crRNA or of the 45 nt. crRNA. The part of the repeat sequence is conserved between organisms and can function as a tag for RNA cutting enzymes.

The RNA is processed by an enzyme complex consisting of different Cas-proteins (Brouns et al., 2007; Hale et al., 2009). In *E. coli* an enzyme complex consisting of the proteins CasA, CasB, CasC, CasD and CasE was detected and was called the Cascade complex (Brouns et al., 2007). This complex cuts large RNA fragments (from now on called pre-crRNA; pre CRISPR-RNA) into small RNA fragments (crRNA) of around 57 nt. (Brouns et al., 2007). To elucidate the function of the individual Cascade components, knockout strains for the different components were made. It appeared that the 57 nt. crRNA fragments were present in the knock-out strains for CasA, CasB and CasC, but absent in the knock out strains for CasD and CasE (Brouns et al., 2007). The Northern blot of the CasE knockout showed a long RNA fragment of 150 nt. When the complex was expressed in a different *E.*

coli strains lacking other *Cas* genes, the pre-crRNA was fully cut. In the knockout strains lacking the *CasE* gene no crRNA products were detected. So it seems that casE is an essential component of the Cascade complex and is needed to process the pre-crRN. These results are confirmed by an similar study in *P. furiocus*, in which another protein complex consisting of seven proteins and crRNA was discovered (Hale et al., 2009). The complex consists of 5 RAMP proteins having putative RNA binding activity (Cmr3, Crm4, Cmr6, Cmr1-2 and Cmr1-1), one polymerase/nuclease (Cmr2) and a protein without a derived function (Cmr5). To elucidate the function of the different components of the protein complex, its activity was measured by excluding one of the 7 proteins at a time. This study revealed that all proteins affect the RNA cleaving ability, except Cmr5 (Hale et al., 2009).

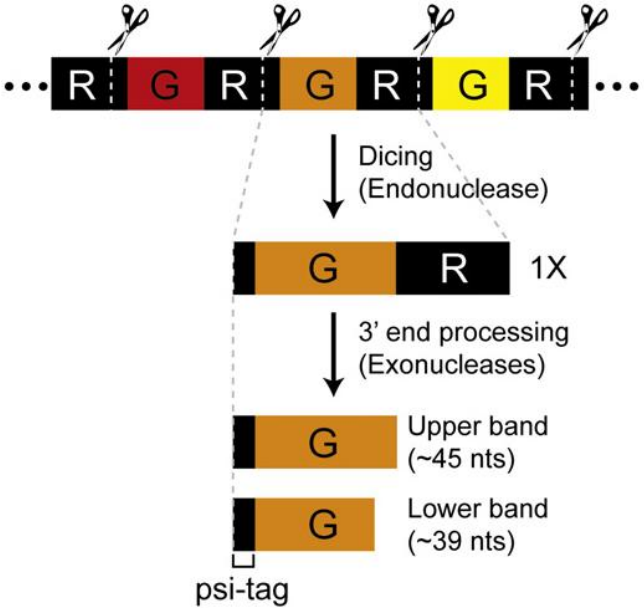


Figure 8: Possible mechanism of processing of the CRISPR RNA (crRNA). The upper part of the figure shows the repeat-spacer region of the CRISPR locus. The R stands for repeats and the G stands for Guide-sequence, which is the same as the repeat sequence. The other parts of the figure show the different products of the processing steps. The intermediate consisting of G and R (second from above) is cut out of the long RNA product from the repeat-spacer region of the CRISPR locus. This product is cut in two ways, shown in the figure. Every product has an psi-tag, a part of the repeat sequence and a part of the spacer sequence (G). (Fig. from: Hale et al., 2009)

Thus, it seems likely that protein cleavage is done by large protein complexes, but the exact mechanism is still difficult to elucidate. In the protein complex of *E.* coli, just one protein proved to be essential for RNA cleavage. In contrast, in the protein complex of *P. furiocus*, at least three components are essential for RNA cleavage.

Moreover, the cascade complex from *E. coli* does not seem to be solely essential for providing phage resistance (Brouns et al., 2007). CRISPRs possessing only the Cascade complex and none of the other *cas* genes such as cas3, cas 1, cas 2 did not provide resistance against phage infection. On the other hand, a CRISPR with only Cas3 also did not specify phage resistance. Interestingly, only the CRISPRs which contained both the genes encoding Cascade complex and Cas 3 possessed phage resistance (Brouns et al., 2007). CRISPRs which contained the *Cascade* genes and *cas1-3*, were equally sensitive to phage attack as the CRISPR containing only the *Cascade* genes and *cas3* (Brouns et al., 2007). (these results are summarized in Fig. 9.)

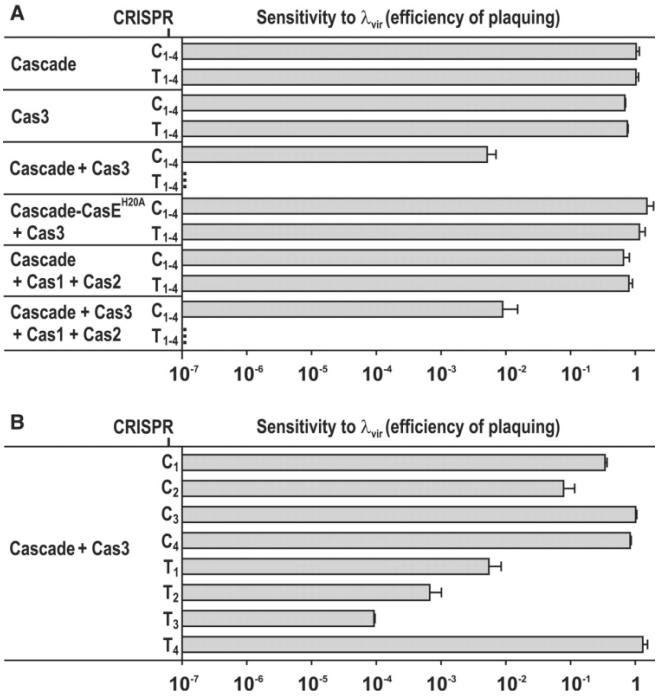


Figure 9: The sensitivity of an *E. coli* strain with different CRISPRs to phage lambda in EOP. (A) On the left of the diagrams, the composition of the different CRISPRs are indicated. C1-4 and T1-4 refers to the composition of the spacer sequences in the CRISPR loci. Spacers indicated with a C, are derived from the coding strain of the phage DNA and the spacers indicated with the letter T are derived from the spacers derived from the template strain of the phage DNA. (B) The sensitivity of the CRISPRs containing the *cas* genes for the Cascade complex, *cas3* and one of the 8 spacer sequences. (Fig. from: Brouns et al., 2008)

Conclusion

The CRISPR locus provides resistance against bacteriophages. It is shown that the acquisition of new spacers, derived from the infecting bacteriophage, is essential for providing resistance against these phages. Beside the spacers, the *cas* genes are essential for the functioning of the CRISPR locus. Some of the Cas proteins encoded by these *cas* genes form protein complexes and cut RNA, that is transcribed from the CRISPR locus.

However, the exact mechanism of action of the CRISPR system is not clear. The property of the Cas protein complexes to restrict RNAs could be used to restrict phage RNAs and hence make them incapable of reproducing. Still the question remains why the spacers are incorporated. The spacers possibly play a role in the restriction reaction. The Cas protein complex in *P. furiosus* contained transcribed RNAs from the CRISPR locus (Hale et al., 2009).

Another hypothesis is that the CRISPR/Cas system works by RNA interference. This mechanisms silences phage RNA by binding to the complementary single strand of the phage RNA. In this mechanism the Cas protein complexes can execute the spacer acquisition. This can also explain the presence of the crRNA in this complex that is used as a sort transporter of the spacers to the chromosome.

These hypotheses are highly speculative, as the amount of research yet done is insufficient to describe a complete working mechanism of the CRISPR/Cas system. But still, a few conclusions can be drawn. It is clear that the CRISPR/Cas system is a defense mechanism against bacteriophages completely different from the mechanisms already known. The CRISPR/Cas system provides an inheritable resistance against bacteriophages and hence is important for adaptation to the environment. Bacteria adapt themselves to the environment by uptake, among others, of phage DNA. On the other hand, bacteriophages mutate very fast,

so different spacer sequences have to be taken up by bacteria continuously to provide resistance against the infecting phage.

The complete mechanism of the CRISPR/Cas system is not yet elucidated. The research done to date focused on a number of organisms, while the CRISPR loci differed between the species, so the working mechanisms could differ between organisms. This is also shown by the different effects of knocking out components of the enzyme complexes in *E. coli* and *P. furiosus*. In *E. coli* just one protein of the Cascade complex proved to be essential for RNA cleavage, while in *P. furiosus* all but one of the components proved to be essential to DNA cleavage. Because of the differences in CRISPR loci, more research should be performed to fully elucidate the working mechanisms of the CRISPR/Cas system.

References

Books

Clark DP, Dunlap PV, Martinko JM, Madigan MT. (2009) *Brock biology of microorganisms* (12ed) San Francisco: Pearson education.

Articles

Anderson F, Banfield F. (2008) Virus population dynamics and acquired Resistance in Natural Microbial Communities *Science* 320, 1047-1050

Barrangou R, Fremauc C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. (2007) CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* 315, 1709-1712

Bergh, Ø. et al. (1989) High abundance of viruses found in aquatic environments. *Nature* 340, 467–468

Breitbart M, Rohwer F. (2005) Here a virus, there a virus, everywhere the same virus? *TRENDS in Microbiology* 13(6), 278-284

Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Ambrosius P, Snijders L, Dickman MJ, Makarova KS, Koonin EV, van der Oost J. (2008) Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science* 321, 960-963

Bolotin A, Quinquis B, Sorokin A, Ehrlich SD. (2005) Clustered regularly interspaces short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* 151, 2552-2561

Campbell A. (2003) 'The future of bacteriophage biology' Nature Reviews Genetics 4, 471-477

Deveau H, Barrangou R, Garneau JE, Labonté J, Fremaux C, Boyaval P, Romero DA, Horvath P, Moineau S. (2008) Phage response to CRISPR-encoded Resistance in *Streptococcus thermophilus*. *Journal of Bacteriology*. 190, 1390-1400

Díez-Villaseñor C, Almendros C, García-Martínez J, Mojica FJM. (2010) Diversity of CRISPR loci in *Escherichia Coli*. *Microbiology* 156, 1351-1361

Edward RA, Rohwer F. (2005) Viral Metagenomics. *Nature Reviews Microbiology* 3,504-510

Forde A, Fitzgerald GF. (1999) Bacteriophage defense systems in lactic acid bacteria. *Antonie van Leeuwenhoek* 76, 89-113

Hale CR, Zhao P, Olson S, Duff MO, Graveley BR, Wells L, Terns RM, Terns MP. (2009) RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. *Cell* 139, 945-956

Haft, D.H., Selengut, J., Mongodin, E.F., and Nelson, K.E. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol* 1, e60

Horvath P, Romero DA, Coûte-Monvoisin A. (2008) Diversity, Activity, and Evolution of CRISPR Loci in Streptococcus thermophilus *Journal of Bacteriology* 190,1401-1412

Ishino Y, Shinagawa H, Makino K, Amemura M, Nakata A. (1987) Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli , and identification of the gene product. *Journal of Bacteriology* 169, 5429-5433

Jansen R, van Embden JD, Gaastra, W, Schouls, LM. (2002-I) Identification of a novel family of sequence repeat among prokaryotes. *OMICS* 6, 23-33

Jansen R, van Embden JDA, Gaastra Wm Schouls LM. (2002-II) Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology* 43, 1565-1575

Karginov FV, Hannon GJ. (2010) The CRISPR system: small RNA-guided Defence in Bacteria and archaea. *Molecular Cell Review* 37, 7-19

Kunin V, Sorek R, Hugenholz P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biology* 8, R61

Lillestøl RK, Redder P, Garrett RA, Brugger K. (2006) A putative viral defence mechanism in archaeal cels. *Archaea* 2, 59-72

Makarova KS, Grishin NV, Shabalina SA, Wolf YI, EV Koonin. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct* 1, 7

Mc Grath S, Fitzgerald GF, van Sinderen D. (2005) Bacteriophages in dairy products:Pros and cons. *Biotechnology Journal* 2,450-455

Mojica FJ, Ferrer C, Juez G, Rodriguez-Valera F. (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Molecular Microbiology* 36, 244-246

Mojica FJ, Díez-Villaseñor C, Garcia-Martinez J, and Soria E. (2005 .Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol* 60, 174–182.
Mojica FJ, Díez-Villaseñor C, García-Martínez J, Almendros C. (2008) Short motif determine the targets of the prokaryotic CRISPR defence system. *Microbiology* 155,733-740

Mojica FJM, C. Díez-Villaseñor C, García-Martínez J, Almendros C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR system. *Microbiology* 155, 733-740

Pourcel C, Salvignol G, Vergnaud G. (2005) CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies *Microbiology* 151, 653-663

Rohwer F, Thurber RV. (2009) Viruses manipulate the marine environment. *Nature* 459,207-212

Tang TH, Bachellerie JP, Rozhdestvensky T, Bortolin ML, Huver H, Drungowski M, Elge T, Brosius J, Huttenhofer A. (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeo-glocus filgidus. *Proc. Natl. Acad. Sci. USA* 99, 7536-7541

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. (2003) The COG database an updated version included prokaryotes. *BMC Bioinformatics*. 11, 41

Tyson W, Chapman J, Hugenholz P, Allen EE, Ram RJ, Richardson PM et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*. 428, 37-43

Tyson W, Banfield F. (2008) Rapidly evolving CRISPRs implicated acquired resistance of microorganisms to viruses *Environmental Microbiology* 10, 200-207

Wommack KE, Colwell RR. (2000) Viroplankton: Viruses in Aquatic Ecosystems. *Microbiology and Molecular Biology Reviews* 64(1), 69-114