The discovery of CRISPR in archaea and bacteria

Francisco J.M. Mojica[1,2] and Francisco Rodriguez-Valera[3]

1 Departamento de Fisiología, Genética y Microbiología. Universidad de Alicante. 03690-Alicante. Spain.

2 Instituto Multidisciplinar para el Estudio del Medio "Ramón Margalef". Universidad de Alicante. 03690-Alicante. Spain.

3 Evolutionary Genomics Group, Universidad Miguel Hernandez, Campus de San Juan, 03540 San Juan, Alicante, Spain.

Correspondence

F.J.M. Mojica, Departamento de Fisiología, Genética y Microbiología, Campus de San Vicente del Raspeig, Universidad de Alicante, Ctra. San Vicente-Alicante s/n, 03690-Alicante, Spain

Tel: +34 965 909 761

E-mail: fmojica@ua.es

Article type : Discovery-in-Context Review

Running title

The discovery of CRISPR

Abbreviations

CRISPR, clustered regularly interspaced short palindromic repeats; Cas, CRISPR-associated; DR, direct repeat; TREPs, tandem repeats; SRSR, short regularly spaced repeats; SPIDR, spacers interspersed direct repeats; PAM, protospacer adjacent motif.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/febs.13766

Keywords

CRISPR, Cas, genomics, adaptive immunity, interference, repeat, prokaryote.

Abstract

CRISPR-Cas are self/non-self discriminating systems found in prokaryotic cells. They represent a remarkable example of molecular memory that is hereditarily transmitted. Their discovery can be considered as one of the first fruits of the systematic exploration of prokaryotic genomes. Although this genomic feature was serendipitously discovered in molecular biology studies, it was the availability of multiple complete genomes that shed light about their role as a genetic immune system. Here we tell the story of how this discovery originated and was slowly and painstakingly advanced to the point of understating the biological role of what initially was just an odd genomic feature.

Main text

The exploration of prokaryotic genomes is barely starting to bear fruit. Hidden in this rich ore remain surely many secrets about these ancient cells that have occupied our planet for close to 4 billion years. There is nothing new in this assertion, but microbes are often far from the limelight in the media driven science of today. And still, prokaryotes have provided humans with many tools that we use to study ourselves, the plants and animals that we eat, see and enjoy, and the biological world at large. The field of prokaryotic genomics is often criticized as being merely exploratory (or discovery driven as opposite to hypothesis driven). However, exploratory research has been instrumental in the history of biology providing the main substrate on which to build hypothesis or just opening new unsuspected fields. One wonderful example of the value of such exploratory research has been the discovery of the most outstanding example of heredity of acquired characteristics in biology, the CRISPR-Cas systems.

Clustered regularly interspaced short palindromic repeats, CRISPR for short, form arrays of up to several hundred units in the genome of many bacteria and archaea. Together with the CRISPR-associated (Cas) proteins, they are the main constituents of the CRISPR-Cas systems that provide prokaryotes with acquired immunity and contribute other unrelated functions such as gene regulation. CRISPR were reported for the first time almost 30 year ago. Since then, our knowledge on these peculiar sequences has greatly improved, but the process has been long and, at times, also gruelling.

In 1989, immediately after finishing his compulsory military service, one of us (FJM Mojica) joined the Microbiology group at the University of Alicante, Spain, supported by a part-time contract to analyse the quality of the sea water at the main beaches of the Alicante's region. In parallel, he started his PhD thesis work supervised by F. Rodríguez-Valera and Guadalupe Juez on the regulatory mechanisms that allow extremely halophilic archaea (haloarchaea) to adapt to environmental changes in salinity. The microorganism under study was *Haloferax mediterranei* R-4, a strain some time ago isolated by FR-V and co-workers from ponds of a solar saltern located in the village of Santa Pola, at the Mediterranean Sea coast near Alicante [1]. The synthesis of gas vesicles is a remarkable feature of this strain, prominently affected by salinity. Therefore, the initial thesis project contemplated to unveil the salt-dependent expression of genes encoding the protein components of these buoyancy devices. However, by the time we were analysing data from the Northern blots obtained after hybridizing with vacuole gene probes, salt and growth-phase dependent expression of the major gas vesicle protein of this microorganism was reported by another research group [2]. While still retaining the main objective of the thesis, this publication led to a substantial shift in the particular traits to be studied, focusing then on genetically uncharacterized regions of the *H. mediterranei* genome that seemed to be subjected to some sort of salt-associated DNA modification [3]. Such modification was supported by the observation that some sequences showed different susceptibility to cleavage by restriction enzymes depending on the salinity of the growth medium. It

was hypothesized that epigenetic changes were the basis of the cleavage resistance. This DNA modification could be playing a physiological role through affecting expression of genes that might be involved in the adaptation to growth at different salinities. Under this premise, two such regions were selected for sequencing and gene expression analysis. These sequencing reactions in 1992 were among the very first ever performed at the University of Alicante. In one of the few readable sequencing films we got in these initial attempts (Fig. 1), an unexpected pattern was seen: DNA segments, 30 bp long, were repeated at regular distances (Fig. 2A). This peculiar arrangement was initially deemed to be a sequencing artefact. However, additional sequencing experiments confirmed the presence in this region of at least 14 almost perfectly conserved (one occasional mismatch) repeats flanked at one end by a highly degenerated copy. Each repeat included short inverted repeats, thus being partially palindromic. They were located in a, seemingly, noncoding area next to open reading frames (ORFs) that did not show homology to any known protein [4]. In contrast to the single RNA molecules detected for each ORF, Northern blots hybridized with probes of the repeat region revealed multiple transcripts, suggesting that highly processed RNAs derived from the repeat arrays (Fig. 3).

Whereas regularly spaced repeats had not been seen before in archaea, similarly arranged structures had been reported in the chromosome of Gram-negative (*Escherichia coli*) [5] and Gram-positive (*Mycobacterium spp.*) [6] bacteria. Two DNA segments with 14 and 7 copies, respectively, of a 29-bp sequence were identified in *E. coli* K12 (Fig. 2B). As in the case of *H. mediterranei*, the *E. coli* repeats were highly conserved within each array (with the exception of the terminal unit), contained a dyad symmetry and were probably located in noncoding regions. Hybridization assays suggested that similar repeats were present in other *E. coli* strains as well as in isolates of the closely related *Salmonella enterica* and *Shigella dysenteriae* species [7]. An array with 49 copies of a 36 bp direct repeat (DR), separated by nonrepetitive spacer DNA of 35-41 bp in length (Fig. 2C), was found in *Mycobacterium bovis* BCG strains [6]. Southern blot hybridization suggested that these repeats are a

general feature in the *Mycobacterium tuberculosis* complex, where a high polymorphism of the DR-containing region among strains was inferred. Taking advantage of such variability, the DR loci were soon harnessed for typing purposes [8].

Putative biological functions such as chromosomal rearrangement or regulation of neighbouring genes were proposed for the bacterial repeats [6,7]. However, functional studies aiming to discover their biological role were only addressed in haloarchaea [9]. Even though salt-dependent transcription of genes surrounding the haloarchaeal repeats (by then called TREPs after Tandem REPeats) was detected, their implication in exclusively transcription regulation of a few genes was questionable due to the large size of the repeat loci. TREP sequences were found in the chromosome and a large plasmid of both *H. mediterranei* and *Haloferax volcanii*, and hybridization assays disclosed the presence of similar repeats in a second haloarchaeal genus, *Haloarcula*. Taking advantage of a genetic manipulation system developed for *H. volcanii* [10], the effect of the presence in *H. volcanii* cells of recombinant plasmids carrying fragments with TREPs was investigated. Firstly, the possibility that TREPs could serve as frequent recombination sites, and hence recombination as the main biological role of TREPs, was dismissed. Then, a reduction in cell viability detected in these cultures, concomitant with the appearance of cells with reduced chromosomal content, was seen as a sign of incompatibility between replicons containing the repeats, hinting at a role for TREPs in replicon partitioning [9]. Unarguably, a different explanation to the perceived incompatibility could have been envisaged if data on the origin of spacers, unveiled a decade later [11], would have been known at the time. Indeed, it might be possible that the extra set of repeat arrays provided additional insertion sites for spacers, increasing the likelihood of spacer acquisition. In principle, being chromosomal DNA the most abundant in the cell, the vast majority of newly acquired spacers would derive from it. Hence, the chromosome would become preferentially targeted and degraded by interference guided by the novel spacers. However, at the time we could not guess that we were dealing with an adaptive immune system.

In 1995 FJMM moved to the University of Oxford for a postdoctoral stay to work on regulation of gene expression in *E. coli.* Subsequently, back in Alicante, he attempted to start his own research group and resume TREP research in *Haloferax*. However, applications for grants did not get through: halophilic archaea were not regarded as a good model to investigate this issue. Without specific funding and with very limited infrastructures in terms of lab space and facilities, the partitioning hypothesis was challenged now in *E. coli*. When experiments equivalent to those previously performed in *H. volcanii* were replicated in *E. coli*, no clear phenotype was seen. Moreover, additional studies did not provide any evidence of an involvement of the *E. coli* repeats in the partition apparatus. Otherwise, the peculiar arrangement provided by palindromic, regularly spaced repeats, suggested that they might serve as exceptional landmarks for binding to a cellular structure (i.e., the cytoplasmic membrane) or to soluble proteins in a cooperative way. Moreover, they could also influence the structure of the carrier DNA molecule. However, repeat binding proteins could not be recovered from cell extracts and topology of recombinant plasmids was not affected by the presence of repeat stretches.

Meanwhile, a new opportunity to tackle the study of the repeats arose. From 1987 to 1995, the amount of publicly available sequence data was very limited. As a result, regularly spaced repeats had only been seen in the few above mentioned microorganisms. Moreover, its presence was in most cases just inferred after hybridization experiments instead of by direct sequencing, arguably due to the technical hitches. This panorama progressively changed thanks to the improvement of sequencing techniques that led to the publication of the first complete genome of a free-living organism, a bacterium, in 1995. This milestone achievement accounted for a radical transformation of the way biological questions will be addressed henceforth, and a new period in biology began. Sequencing was greatly facilitated and data of additional prokaryotic genomes were released during the late 1990s. The first comprehensive report of regularly spaced repeats in a complete genome was published in 1996 [12] and additional descriptions joined it in the following years. By the end of

the century, similarly arranged repeats had been discovered in 12 species of archaea and bacteria, and open access to 26 completed microbial genomes was granted. However, there was a lack of computer tools for the accurate detection of such peculiar sequences. Therefore, taking into account the length range of known interspersed repeats and spacers, César Díez-Villaseñor (University of Alicante) implemented a specific computer programme to detect, with a wide tolerance, analogous regions in complete genomes. Even though such permissiveness frequently retrieved false-positives, this allowed for the identification of novel repeat units and complete arrays in prokaryotes where their presence had already been reported, as well as in additional completed or close to completion genomes. This comprehensive study demonstrated the wide distribution of this sort of redundant sequences across a wide microbial diversity, hinting at an ancestral origin and high biological relevance [13]. Furthermore, it led to the recognition and definition of a family of prokaryotic repeats termed SRSR after **S**hort **R**egularly **S**paced **R**epeats, an acronym that also mirrored the **S**pacer-**R**epeat-**S**pacer-**R**epeat organization of the clusters. This initial description was followed by a second one, authored by Jansen and collaborators, where the designation **SP**acers **I**nterspersed **D**irect **R**epeats (SPIDR) was proposed [14]. The potential naming conflict was solved after mutual agreement of the two research groups to use CRISPR (pronounced krisper) after "Clustered Regularly Interspaced Short Palindromic Repeats". Jansen immediately accepted the new definition and acronym rather than the other, less descriptive or not so distinctive alternatives, that were proposed [15].

We knew that the CRISPR arrays were transcribed, at least those of *Haloferax*, and the fact that these transcripts were playing a role, either on their own or in collaboration with other components of the cell, could be anticipated. In 2002, four genes encoding CRISPR-associated (Cas) proteins were identified nearby the repeat loci [16]. Apparently, they were exclusively present in genomes carrying CRISPR. Based on sequence similarity, some Cas were tentatively linked to activities related to interaction with nucleic-acids. Therefore, Cas proteins were candidates to assist the CRISPR RNAs

whatever they were doing. The key to unravel the function of the two-component CRISPR-Cas systems came from the repeat-intervening spacers. The origin of these sequences had remained enigmatic. The name itself (spacer) hints at an irrelevant role in the repeats just separating the palindromes. They were unique in the carrier-genome, except for some occasional duplication in the CRISPR locus. Perhaps they were synthesized de novo upon the generation of new repeats [14]. Maybe the yet uncharacterized Cas proteins were responsible for their synthesis. In the early 2000s, the research group of FJMM was working on a project to assess the polymorphism of CRISPR loci in *E. coli.* To that end, a collection of phylogenetically diverse isolates of the species was chosen for a representative analysis. Initially, the arrays of a few strains could be successfully PCR amplified and eventually sequenced. As in many other occasions before, spacers were systematically probed against public nucleotide databases. However, this time one of the queries retrieved a matching sequence. To our surprise, the spacer homolog was located in the genome of a coliphage. Then, three additional spacers were found to be similar to sequences in either that bacteriophage or in a conjugative plasmid of *E. coli*. To corroborate this finding, the spacers of 61 additional strains, including all available complete genomes, were subjected to the same sort of analysis, turning out that about a 2% of them matched sequences in non-CRISPR loci, invariably in genetic elements of the corresponding spacer-carrier species [11]. The CRISPR meaning suddenly clicked into place; these arrays are crisper-like compartments for storing DNA chunks of invaders, to keep a fresh memory of past infections. The next question was then why the cell dedicated part of its limited genome space to this “trunk of souvenirs”. Four microorganisms were selected, as representatives of their respective prokaryotic groups, for further analysis: *E. coli* (Gram-negative bacteria), *Streptococcus pyogenes* (Gram-positive bacteria), *Sulfolobus* (crenarchaea) and *Methanothermobacter thermoautotrophicum* (euryarchaea). They greatly differ in their physiology, ecology and phylogeny. Hence, common observations related to their CRISPR loci would provide conclusions applying to the prokaryotes in general. Notably, spacer-homologs were found in mobile genetic elements that, according to published studies, failed to efficiently infect the corresponding spacer-carrier strain,

even though they proficiently disseminated in populations of closely related strains lacking the spacer. Moreover, plasmids or viruses containing spacer-matching sequences (named protospacers) were usually absent in the spacer-carrier. In the exceptional cases where the cell harboured one of these protospacer-carriers, the similarity between the spacer and its homolog segment was markedly diminished with respect to the protospacer consensus sequence. Taken together, these findings strongly supported a connection between CRISPR and immunity against foreign DNA, most likely guided by CRISPR-RNA molecules. Nonetheless, given that protospacers were also detected in chromosomal regions, a regulatory role for CRISPR was inferred, resembling the eukaryotic RNA interference (RNAi) mechanism [11].

After one decade (1993-2003) of unsuccessful attempts, the mystery behind the TREPs function seemed to be solved. Now, the challenge reached the next level: the role inferred had to be proven. Immediately after this discovery, the immunity hypothesis was investigated by FJMM's team in *E. coli* strains carrying resident spacers that matched sequences in bacteriophages or engineered plasmids. However, highly variable results were obtained upon deletion of phage-homologous spacers on susceptibility to infection, as well as when analysing transformation efficiency of recombinant plasmids containing spacer-matching sequences (unpublished results). It was later reported that the CRISPR system of *E. coli,* at least that of K12-derivative strains, is usually silenced due to repression executed by the histone-like protein H-NS. Furthermore, this inhibition can be relieved by the activator protein LeuO. This on-off swift of CRISPR immunity [17] could account for the results we obtained: interference was sporadically triggered by fortuitous de-repression. It is a remarkable coincidence that, H-NS [18] and a promoter regulated by LeuO [19], had been the subject of FJMM's project for the period of its allegedly parenthesis in CRISPR research during his postdoctoral stay in Oxford.

In spite of the novelty and interest of the revelation, that prokaryotes harbour an adaptive interference system, the lack of solid experimental evidence obstructed the publication of the manuscript. Initially submitted in 2003, after being rejected by four different journals, a substantially lightened version was eventually accepted in October 2004 thanks to very positive comments of two anonymous reviewers and constructive suggestions made by the reviewing editor of Journal of Molecular Evolution [11].

Another paper reporting the origin of the spacers in *Yersinia* species [15,20], suffered a similar ordeal. Subsequently, a third paper along the same lines appeared [21], showing a positive correlation between spacer number and resistance to infection. They also discovered a relevant aspect of the spacer homologs: the protospacers of a CRISPR system in *Streptococcus thermophilus* were juxtaposed to a short conserved sequence. These signatures, today known as Protospacer Adjacent Motifs (PAMs), were afterwards shown to be a common feature in diverse CRISPR systems [22] where PAM recognition is a requisite for both spacer uptake and target interference [23].

Arguably, the advent of three independent reports pointing in the same direction helped to overcome the almost general reluctance, as inferred from most journal editors and referees comments, to admit the existence in prokaryotes of a widespread, adaptive immune system. The CRISPR field experienced a golden age that strongly contrasted with the very little attention attracted previously. Diverse bioinformatic tools and resources were soon implemented [24] for the analysis of CRISPR loci and their dynamics [25] and the diversity of the CRISPR-Cas systems became evident [26]. Distinct CRISPR sequences could be grouped after their sequence, and a substantial number of canonical CRISPR were not palindromic at all [26]. However, attempts to replace "palindromic" with "prokaryotic" in the definition of this family of repeats did not reach an ample agreement among the CRISPR community.

The empirical demonstration that CRISPR provides acquired resistance against genetic elements, came from researchers of a Danish bio-based company working in collaboration with the group of Sylvain Moineau (Université Laval, Canada). Thus, Barrangou *et al*. [27] found that *S. thermophilus* becomes resistant to infection by bacteriophages after the insertion of new spacers matching the virus genome. The identity of the spacer was a key determinant of the specificity of the resistance phenotype. Notably, they proved that the Cas proteins were indeed a functional component of the CRISPR systems. Soon after this ground-breaking work, CRISPR interference against plasmid transfer through DNA targeting was shown [28], CRISPR-RNA molecules were identified as the guides of this immune system [29] and target cleavage was unveiled as the mechanism of interference [30]. Many Cas proteins were biochemically characterized and their activity established [31].

We know now that CRISPR can acquire spacers from, and eventually interfere with, RNA [32] and/or DNA molecules, depending on the particular system. Besides, target cleavage is not the only way CRISPR may act and roles beyond immunity, notably through regulation of gene expression [33], are played by specific systems. This is the case of some complete CRISPR-Cas systems, but also of CRISPR that function without the assistance of Cas proteins, and vice versa. There are orphan CRISPR arrays in prokaryotic genomes and viruses [34,35], as well as CRISPR-like stretches in eukaryotic viruses [36] and mitochondria [13], most of which are enigmatic regarding their activity or function. The discovery of canonical CRISPR almost three decades ago opened an avenue with many side paths to unknown destinations. The role in genome editing that has brought up Cas9 to notoriety might be played by other unrelated proteins as well [37].

But CRISPR-Cas are much more than a prokaryotic immune system. Their existence proves that prokaryotic genomes are environment tuned. They can adapt by Lamarckian inheritance keeping track of genomic encounters, what probably helps in fine tuning the delicate equilibrium between conservation (maintaining the status quo) and variation providing novel genomic features and

phenotypic properties in bacterial populations. There might be many more unknown mechanisms among the roughly 30% function-unknown proteins that appear in each prokaryotic genome regardless of how common the microbe is, not to mention the average 10% non-coding DNA [38] present in all these genomes.

It seems appropriate to remember here that the great motor of molecular biology during the XXth century, molecular cloning, was made possible by restriction endonucleases and plasmid vectors, both derived from serendipitous discoveries of prokaryotic cell biology features involved in self/non-self discrimination as well. The lesson here to scientists, science policy makers and mankind at large is that the only way forward is enlarging evenly the sphere of knowledge supporting fundamental research. In words of Louis Pasteur: "There does not exist a category of science to which one can give the name applied science. There are science and the applications of science, bound together as the fruit of the tree which bears it".

Acknowledgements

FJMM is funded by the Spanish Ministerio de Economía y Competitividad (BIO2014-53029P) and the European Commission / Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (291815 Era-Net ANIHWA). FR-V is funded by projects MEDIMAX BFPU2013-48007-P from the Spanish Ministerio de Economía y Competitividad, MaCuMBA Project 311975 of the European Commission FP7 and PROMETEO II/2014/012 project AQUAMET from the Generalitat Valenciana. The authors thank Lluis Montoliu for critical reading of this manuscript.

Author contributions

The authors contributed equally to writing this manuscript.

References

1. Rodríguez-Valera F, Juez G & Kushner DJ (1983) Halobacterium mediterranei spec, nov., a new carbohydrate-utilizing extreme halophile. *Syst Appl Microbiol* **4**, 369-381.

2. Englert C, Horne M & Pfeifer F (1990) Expression of the major gas vesicle protein gene in the halophilic archaebacterium Haloferax mediterranei is modulated by salt. *Mol Gen Genet* **222**, 225-232.

3. Juez G, Rodríguez-Valera F, Herrero N & Mojica FJM (1990) Evidence for salt-associated restriction pattern modifications in the archaeobacterium Haloferax mediterranei. *J Bacteriol* **172**, 7278-7281.

4. Mojica FJM, Juez G & Rodríguez-Valera F (1993) Transcription at different salinities of Haloferax mediterranei sequences adjacent to partially modified PstI sites. *Mol Microbiol* **9**, 613-621.

5. 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. *J Bacteriol* **169**, 5429-5433.

6. Hermans PW, van Soolingen D, Bik EM, de Haas PE, Dale JW & van Embden JD (1991) Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains. *Infect Immun* **59**, 2695-2705.

7. Nakata A, Amemura M & Makino K (1989) Unusual nucleotide arrangement with repeated sequences in the Escherichia coli K-12 chromosome. *J Bacteriol* **171**, 3553-3556.

8. Groenen PM, Bunschoten AE, van Soolingen D & van Embden JD (1993) Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method. *Mol Microbiol* **10**, 1057-1065.

9. Mojica FJM, Ferrer C, Juez G & Rodríguez-Valera F (1995) Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. *Mol Microbiol* **17**, 85-93.

10. Lam WL & Doolittle WF (1989) Shuttle vectors for the archaebacterium Halobacterium volcanii. *Proc Natl Acad Sci U S A* **86**, 5478-5482.

11. Mojica FJM, Díez-Villaseñor C, García-Martínez J & Soria E (2005) Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* **60**, 174-182.

12. Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NS & Venter JC (1996) Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. *Science* **273**, 1058-1073.

13. Mojica FJM, Díez-Villaseñor C, Soria E & Juez G (2000) Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Mol Microbiol* **36**, 244-246.

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

15. Mojica FJM & Garrett RA (2013) Discovery and seminal developments in the CRISPR field. In *CRISPR-Cas Systems: RNA-mediated adaptive immunity in Bacteria and Archaea* (Barrangou R & van der Oost J, eds), pp. 1-31. Springer-Verlag, Berlin Heidelberg.

16. Jansen R, van Embden JD, Gaastra W & Schouls LM (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* **43**, 1565-1575.

17. Mojica FJM & Díez-Villaseñor C (2010) The on-off switch of CRISPR immunity against phages in Escherichia coli. *Mol Microbiol* **77**, 1341-1345.

18. Mojica FJM & Higgins CF (1997) In vivo supercoiling of plasmid and chromosomal DNA in an Escherichia coli hns mutant. *J Bacteriol* **179**, 3528-3533.

19. Mojica FJM & Higgins CF (1996) Localized domains of DNA supercoiling: topological coupling between promoters. *Mol Microbiol* **22**, 919-928.

20. 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.

21. Bolotin A, Quinquis B, Sorokin A & Ehrlich SD (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**, 2551-2561.

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

23. Shah SA, Erdmann S, Mojica FJM & Garrett RA (2013) Protospacer recognition motifs: Mixed identities and functional diversity. *RNA Biol* **10,** 891-899.

24. Grissa I, Vergnaud G & Pourcel C (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* **8**, 172.

25. Andersson AF & Banfield JF (2008) Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* **320**, 1047-1050.

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

27. Barrangou R, Fremaux 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.

28. Marraffini LA & Sontheimer EJ (2008) CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* **322**, 1843-1845.

29. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, Snijders AP, Dickman MJ, Makarova KS, Koonin EV & van der Oost J (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960-964.

30. Garneau JE, Dupuis ME, Villion M, Romero DA, Barrangou R, Boyaval P, Fremaux C, Horvath P, Magadán AH & Moineau S (2010) The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**, 67-71.

31. Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, van der Oost J & Koonin EV (2011) Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* **9**, 467-477.

32. Silas S, Mohr G, Sidote DJ, Markham LM, Sánchez-Amat A, Bhaya D, Lambowitz AM & Fire AZ (2016) Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* **351**, aad4234.

33. Westra ER, Buckling A & Fineran PC (2014) CRISPR-Cas systems: beyond adaptive immunity. *Nat Rev Microbiol* **12**, 317-326.

34. García-Heredia I, Martín-Cuadrado AB, Mojica FJM, Santos F, Mira A, Antón J & Rodríguez-Valera F (2012) Reconstructing viral genomes from the environment using fosmid clones: the case of haloviruses. *PLoS One* **7**, e33802.

35. Makarova KS, Wolf YI, Alkhnbashi O, Costa F, Shah S, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft D, Horvath P, Moineau S, Mojica FJM, Terns RM, Terns MA, White MF, Yakunin AF, Garrett RA, van der Oost J, Backofen R & Koonin EV (2015) An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* **13**, 722-736.

36. Levasseur A, Bekliz M, Chabriere E, Pontarotti P, La Scola B & Raoult D (2016) MIMIVIRE is a defence system in mimivirus that confers resistance to virophage. *Nature* **531**, 249-252.

37. Gao F, Shen XZ, Jiang F, Wu Y & Han C (2016) DNA-guided genome editing using the Natronobacterium gregoryi Argonaute. *Nat Biotechnol,* advanced online publication*,* doi:10.1038/nbt.3547.

38. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S & Ussery DW (2015) Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics* **15**, 141-161.

Fig. 1. Autoradiograph of an unpublished Sanger sequencing gel, dated 21/08/1992, where regularly spaced repeats were discovered in *H. mediterranei*. Repeats are marked with side bars.

Fig. 2. Sequences of regularly spaced repeat regions. DNA stretches containing four regularly spaced repeats (highlighted in yellow) are shown as representative examples of the sequences originally reported in *H. mediterranei* (A) [4], *E. coli* (B) [5] and *M. bovis* (C) [6]. Inner inverted repeats are underlined.

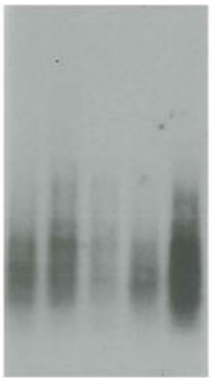

Fig. 3. Transcription of regularly spaced repeat regions. Northern blot hybridization of total RNA samples obtained from *H. mediterranei* cultures at different growth conditions. A DNA fragment from the repeat array described in [4] was used as probe (FJMM, unpublished doctoral thesis).