DNAdrive

Imagine the number of times your device has informed you that you ran out of storage space. It is not surprising that in a few years this will be the case for thousands of companies specialized in data storage such as Google, Amazon or Microsoft. It is estimated that because of the exponential accumulation of human-generated data, the amount of information produced annually will increase from 16.3 ZB (zettabytes = 1021 bytes) to 163 ZB by 2025. Unfortunately, the imminent gap between the demand for storage space and the production of storage capacity is not a problem that can be easily solved. Indeed, it is much more difficult to build capacity than to generate data. Building capacity to overcome this stratospheric demand would imply investments worth billions of euros, which is not a cost-effective or even realistic option. In addition, a single solar flare could damage all electronic circuits in the blink of an eye. However, nature has offered us a very effective alternative to storing data: DNA. Hard disk of the living, this incredibly stable and resistant molecule could allow us to store the entirety of Internet in a test tube.

Does the encapsulation of modified DNA in an organism present an effective and sustainable alternative to conventional data storage?

The goal of our project is to show how handling a bacterium capable of integrating foreign DNA into its genome can be used for data storage purposes.

CONTENTS:

1. The CRISPR-Cas system

1.1. History

In 1987, geneticists noticed a strange pattern in some bacterial genomes. A DNA sequence would be repeated several times, with unique sequences between the repeats They called this strange configuration "clustered regularly interspaced short palindromic repeats", or CRISPR. By observing that the spacers present in the CRISPR locus largely correspond to fragments of extrachromosomal elements (such as bacteriophage or plasmid genetic material), CRISPR was finally recognized as a prokaryotic adaptive immune system 19 years after its discovery (Bolotin et al. 2005). In other words, CRISPR corresponds to a specialized region of DNA in the genome of a prokaryotic organism (bacterium or archaea) which has two distinct characteristics: the presence of repetitive sequences from 23 to 55 base pairs and between them, generally unique spacers of 21 to 72 base pairs (Grissa et al. 2007). Bacteria acquire the memory of viral invaders by incorporating invasive DNA extracts into the CRISPR locus of their genome, thereby generating a new spacer between two repetitive DNA sequences. RNA harboring the spacer sequence assists Cas proteins in recognizing and cutting DNA from similar viruses during subsequent invasions (Bolotin et al. 2005; Barrangou et al. 2007; Brouns et al. 2008; Barrangou and Marraffini 2014).

1.2. Locus structure

The three main components of the CRISPR locus are the *cas* genes, the leader sequence, and the spacer-repeat array (Figure 1).

1.2.1. *cas* genes

The CRISPR-Cas immune system requires the presence of a set of CRISPR-associated genes (cas), found only in CRISPR-bearing genomes and located near the spacer-repeat array. These encode the Cas proteins, essential to the immune response of the bacteria. Collectively, 93 cas genes are grouped into 35 families according to the sequence similarity of the encoded proteins. 11 of 35 families form the cas nucleus, which includes the Cas1 to Cas9 families of proteins. A complete CRISPR-Cas locus has at least one gene belonging to the cas nucleus (Grissa et al. 2007; Makarova et al. 2015). Cas1 and Cas2 are the only two Cas proteins universally conserved in all CRISPR-Cas systems, required for the immune system adaptation step (Makarova et al., 2011).

1.2.2. Leader sequence

A leader sequence rich in AT (adenine and thymine) of approximately 60 bp located upstream of the first repeat is essential for the acquisition of spacers (Yosef et al. 2012) and promotes transcription of the spacer-repeat sequences (Pougach et al. 2010). In addition, it contains the binding site of the IHF protein (5 '- WATCAANNNNTTR - 3', where W is A or T, R is A or G and N is any nucleotide) which is essential to the adaptation step - without it, there would be no integration of spacers in the array (Nuñez et al. 2016).

1.2.3. Spacer-repeat array

The leader sequence is immediately followed by a sequence of repetitive sequences separated by unique spacers (Figure 2). Each spacer corresponds to a single exogenous DNA fragment (e.g. that inserted by a virus) and serves as a strategic marker for the bacterium to help destroy similar sequences in future invasions. Some repeats have palindromic properties (when a strand of DNA also has the opposite of its complementary sequence), which implies the formation of a secondary structure such as a stem-loop (‘hairpin’) in RNA (Brouns et al. 2008), while other repeats have no specific structure.

1.3. Mechanism of the bacterial adaptive immune system

Figure 3: The CRISPR-Cas immune system is divided into three stages (van der Oost et al. 2014). In the first step, the adaptation, the Cas1 and Cas2 endonucleases store the invaders' memory in the CRISPR locus in the form of short intermediate sequences, called spacers (Barrangou et al. 2007; Fineran and Charpentier 2012). In the second step, the expression, the spacer-repeat sequence is transcribed and then processed into small CRISPR (crRNA) RNA sequences, each consisting of a spacer and one or two portions of a repeat. crRNAs are linked

to the Cascade multiprotein complex (Brouns et al. 2008). During the last step, the interference, Cascade looks for DNA chains complementary to the crRNA. When a match is found, the target DNA is cleaved several times by the recruited Cas3 nuclease and the attack is stopped, thus immunizing the bacteria.

The study presented in this project was carried out with the type I-E CRISPR-Cas system of Escherichia coli, one of the best studied CRISPR-Cas systems. The following information therefore describes the specific mechanism of this system, unless otherwise noted.

The type I-E CRISPR-Cas system consists of the Cas3 nuclease (which is responsible for the degradation of the target sequence), the Cascade multiprotein effector complex, the Cas1 and Cas2 proteins and the repetition-spacer array of the associated CRISPR locus. The array contains palindromic repeats of 29 base pairs (bp) and intermediate spacers of 32 bp.

The locus must contain at least one repetition, since the repetition preceded by the leader sequence is used as a template for the synthesis of a new repetition (Yosef et al. 2012). Mutations in the leader sequence-repeat boundary permanently hinder integration (Nuñez et al. 2016) and the secondary DNA hairpin structures potentially formed in the repeats can also play a role in Cas1-Cas2 recognition (J.K. Nuñez et al. 2015).

The formation of the complex as well as the nuclease activity of Cas1 are necessary for adaptation. Cas2 nuclease activity is dispensable, indicating a rather structural role for it (Nuñez et al. 2014).

1.3.1. Adaptation – building an immune repertoire

The adaptation of the CRISPR-Cas system is a complex multi-step process in which a protospacer must be extracted from invading foreign DNA and then stored in the CRISPR locus as a spacer. First, the foreign DNA must be recognized as a target for the acquisition of spacers. Second, a sequence of a specific size (typically 30-40 bp, depending on the CRISPR-Cas subtype) must be acquired from the foreign DNA. Finally, the acquired sequence must be integrated as a new spacer in the CRISPR locus and the adjacent repeat sequence must be duplicated. Although the components and prerequisites of the spacer acquisition machinery vary among CRISPR-Cas subtypes and organisms, several components appear to be universally conserved and are essential among all subtypes of the CRISPR-Cas system. These components are the Cas1 and Cas2 proteins, the leader sequence and the first repeat of the CRISPR array.

Adaptation in the type I-E CRISPR-Cas system is the best understood CRISPR-Cas adaptation process. Despite extensive study, the adaptation remains the least understood step of the CRISPR-Cas immune system with respect to expression and interference steps. For example, it remains unclear how viral DNA degradation products (or protospacers) are produced, how the Cas1-Cas2 complex captures the protospacer, whether it is single or double-stranded, if it has overhangs (one strand over flanking the other at one end) and how it is cut by the Cas1-Cas2 complex before it is integrated as a spacer in the CRISPR array (Wang et al. 2015; Künne et al. 2016).

Figure 4: The spacer fragments, generated by Cas1 and Cas2 or other immune systems, are captured by the Cas1-Cas2 complex. The spacer is inserted into the boundary of the leader sequence followed by the repetition-spacer array of the locus through two nucleophilic attacks (1 and 2 in the figure). The PAM (here CTT) helps orientate the new spacer in the array and the last nucleotide of the repetition originates from the integrated DNA of the invader. It is this spacing sequence integrated into the CRISPR array that forms the immune memory. The new repetition is synthesized and the integration is complete.

The acquired spacer is preferentially integrated at the end of the leader sequence followed by the spacer-repeat array. The array is then extended by a spacer-repeat unit (28 + 33 = 61 bp) (Yosef et al. 2012). The integration host factor (the IHF protein) and its binding site in the leader sequence are also necessary for adaptation (Nuñez et al. 2016).

1.3.1.1. PAM

In natural environments, the accidental acquisition of spacers from ‘personal’ DNA - the genome of the cell - rather than from invading DNA is generally harmful because it leads to the degradation of genomic DNA by the interference mechanisms of CRISPR-Cas. Such self-targeting leads to CRISPR-Cas autoimmunity (Stern et al. 2010) and it has been shown that the escape to this autoimmunity usually involves mutational inactivation of Cas genes, mutations in repeats next to the self-derived spacer or escape mutations in PAM (Stern et al. 2010). Therefore, it is necessary for CRISPR-Cas systems to avoid integrating spacers from the genome to minimize these harmful effects. Indeed, the first observations of the E. coli type I-E CRISPR-Cas system showed a strong preference for the acquisition of spacers from foreign DNA and the avoidance of ‘personal’ DNA (Yosef et al. 2012; Díez-Villaseñor et al. 2013; Levy et al. 2015).

The protospacer adjacent motif (PAM), a 2 to 4 nucleotide sequence, is critically important for the recognition and selection of the protospacer during adaptation and interference in Escherichia coli. This is an essential targeting element (absent in the bacterial CRISPR locus) that distinguishes the bacterial genome from foreign DNA, thus preventing the CRISPR locus from being targeted and destroyed by the interference machinery (Deveau et al. 2008; Westra et al. 2013). It has been found that a protospacer could be incorporated into the CRISPR locus as a spacer if and only if it is flanked by the appropriate PAM (Mojica et al. 2009). Four PAM sequences guarantee CRISPR immunity: CAT, CTT, CTC and CTC (Westra et al. 2012). A more general PAM has been suggested, 5' – CHH – 3' (H = A, C or T) (Hayes et al. 2016), which includes all previous PAMs. In Escherichia coli, the most common PAM is 5' – CTT – 3'.

Interesting note: in Escherichia coli, the last nucleotide of the new repeat is derived from the first nucleotide of the acquired spacer. This nucleotide is indeed the last nucleotide of the PAM sequence (Datsenko et al. 2012). Since it is almost invariably a C, it is generally considered the last base of the repetition once integrated (thus the model of the 29 bp (28 + 1) repetition and the 32 bp spacer (33 - 1)) (Datsenko et al. 2012; Swarts et al. 2012).

1.3.1.2. Protospacer constraints

Figure 5: The structural requirements of the protospacer for a complete and certain integration are poorly determined. However, the structural and biochemical data of the Cas1-Cas2 complex indicate that, when the DNA protospacer is captured by the complex, it adopts a double-forked shape with a double-stranded central stem (duplex) flanked by single-stranded overhangs derived from the two strands separated by the tyrosine residues (Tyr22) conserved in each Cas1 monomer. Cas2 recognizes the double-stranded region while Cas1 binds to the single-stranded overhangs at the 3' ends. The PAM sequence in the 3' overhang is recognized by the catalytic subunits of Cas1a in a base-specific manner and the subsequent cleavage 5 nt after the duplex boundary generates an intermediate DNA of 33 nt (including the 23 bp dsDNA duplex and two 5 nt ssDNA overhangs with 3'-OH groups) that is incorporated into the CRISPR array via a cut-and-paste mechanism (J. Nuñez et al. 2015; Wang et al. 2015; Amitai and Sorek 2016). It is suggested that this intermediate forms the substrate for integrating the spacer into the CRISPR network. The properties of the Cas1-Cas2 complex show that the preferred in vitro substrate of the protospacer is double-stranded DNA flanked by 7 nt single-stranded overhangs at the 3' ends, one of which contains a PAM 5 nt away from the duplex (J. Nuñez et al. 2015; Wang et al. 2015).

Cas1-Cas2 works as a molecular ruler, the distance between the two active sites of Cas1 specifying the length of the spacer, and probably cuts the protospacer before integration (Wang et al. 2015)*. The interactions between the spacer and the Cas1-Cas2 complex are independent of its sequence (that of the duplex), which allows the integration of any spacer sequence* (J. Nuñez et al. 2015)*.*

1.3.1.3. AAM

Another sequence motif, the adaptation affecting motif (AAM), a 2 to 4 nucleotide sequence (most often dinucleotide), present in the 5' end of the spacer (duplex) at an interval of 30 bp downstream of the 5' - AAG - 3' motif, has been proposed as an important motif for adaptation rather than for interference (Yosef et al. 2013). However, the importance of the AAM needs to be further demonstrated by other studies.

1.3.1.4. Spacer integration

Figure 6: The Cas1-Cas2 complex cleaves the overhangs of the 3' ends of the protospacer 5 nt from the duplex boundary, thus generating a spacer of 33 nt. Both 3' ends of the incoming protospacer are involved in the nucleophilic attack on the CRISPR array (as indicated by the red and blue dashed arrows, respectively). The PAM helps orient the new spacer so that all spacers are in the same orientation with respect to the leader sequence (Mojica et al. 2009; Shmakov et al. 2014). Finally, the gap duplex is repaired by the replication machinery of the host's DNA. The GC base pair from the PAM sequence is highlighted by a green background. It has been suggested that the Cas1-Cas2 complex requires, as part of its binding to the locus, a palindromic sequence potentially capable of forming a cruciform DNA structure / ‘hairpin’, which is a typical requirement of various integrases (Coté and Lewis 2008).

Figure 7 (Wang et al. 2015): The new spacer is integrated in the locus thanks to an integrase mechanism coordinated by Cas1-Cas2 (J.K. Nuñez et al. 2015). The binding of the IHF protein to the binding site present in the leader sequence most likely causes flexion of the array DNA, allowing access of the Cas1-Cas2 complex to the integration site. Thus, the specificity of the end of the leader sequence near the repetition-spacer sequence of the locus can be explained as an integration site of the spacers (Wang et al. 2015).

2. Our project

2.1. The idea

In 2017, scientists took the first steps towards a quantum Internet (Ren et al. 2017), created metallic hydrogen (Silvera and Cole 2010), a treatment for cancer that uses the patient's T cells (Smith et al. 2016), an artificial uterus that could save preterm babies (Partridge et al. 2017), and genetically modified a human embryo for the first time to eliminate a hereditary heart disease (Ma et al. 2017). These examples are only part of a larger group of important breakthroughs that were only done this year. All these discoveries have one essential point in common: in order to make use of the advantages they offer, one needs to store the research, that is, information.

Thanks to the digital dimension of our time, modern society has allowed itself to accelerate the use, creation and sharing of information, which has led to the development of all areas of human existence and to arrive at the highest life quality level the world has ever recorded. Unfortunately, digital storage of data is becoming increasingly expensive and difficult. A new report from Berkeley's Lawrence National Energy Laboratory states that US data centers use an abundance of energy: 70 billion kWh a year. This represents 1.8% of total US electricity consumption (Shehabi et al. 2016). At an average cost of 10 cents per kWh, the annual cost of this consumption is around $ 7 billion. For comparison purposes, 1 kWh is enough to hold ten 100-watt light bulbs on for an hour or for a smartphone to remain charged for an entire year. Producing 70 billion kWh a year would require power plants with a base capacity of 8,000 megawatts, equivalent to about 8 large nuclear reactors, 16 coal-fired power plants (500 MW) or twice as much as the power produced by all of the solar panels in the country (EIA 2017). There are more than 130 data centers in France (containing more than 100,000 computer servers), thousands on the planet, and it is estimated that a large data center can consume more than a French city of 100,000 inhabitants. “If you aggregated the electricity use by data centers and the networks that connect to our devices, it would rank sixth among all countries,” says Gary Cook, Greenpeace’s international IT analyst and the lead author on its report. “It’s not necessarily bad, but it’s significant, and it will grow”.

Humanity has a data storage problem: more data have been produced in the last two years than in the rest of history. The IDC (International Data Company) predicts that by 2025, the global data sphere will reach 163 zettabytes (ZB, or one billion gigabytes) (Reinsel et al. 2017). This is ten times more than the 16.1 ZB of data generated in 2016. This storm of information may soon exceed the capacity of hard drives. To cope with the "Internet of Things", one must separate long-term information from short-term information. A solution to this problem must be found to avoid the increase of an already important pollution, the waste of money and the worst - the loss of these data (for example due to a cyberattack or even a solar flare).

However, we already have at our disposal a solution that has existed for hundreds of millions of years, one of the most capable, small, stable and resistant storage methods of information known to man: DNA. Why this organic molecule? As digital information continues to accumulate, higher density and longer term storage solutions are going to be needed (Gantz and Reinsel 2011). DNA answers these requirements (Bancroft et al. 2001). First of all, it is very dense. At the theoretical maximum, DNA can encode two bits per nucleotide (nt) or 455 exabytes per gram on single-stranded DNA (see calculations). Secondly, unlike most digital storage media, DNA is not limited to a planar layer and is often readable despite degradation under non-ideal conditions over millennia (Pääbo et al. 2004; Bonnet et al. 2010). Moreover, as long as human societies read and write DNA, they will be able to decode it. "DNA won't degrade over time like cassette tapes and CDs, and it won't become obsolete", says Yaniv Erlich, computer scientist at Columbia University. Unlike other high-density approaches, such as manipulating individual atoms on a surface, new technologies can write and read large amounts of DNA at a time, thus making it a scalable data storage approach. Finally, the essential biological role of DNA provides access to the natural enzymes used for reading and writing, and ensures that the DNA remains a readable standard in the foreseeable future.

In this project, we will not only study the possibilities of storing information in DNA, but also exploit the type I-E CRISPR immune system of Escherichia coli in order to store information within its genome. Why bacteria? The reason is simple: not only is the bacterial genome circular, improving the long-term stability of the DNA (since it protects it from exonucleases), but the bacterium also has defense and preservation mechanisms, protecting it against extrachromosomal elements. Also, it can replicate exponentially, spreading information over offspring. When these data are stored on a hard drive that could theoretically increase in size and thus storage capacity over time, scientists could read them by examining the genome of any bacterium in the colony. Moreover, all modern computers are vulnerable to power outages, data theft or even solar flares that could destroy all electronics. As for bacteria, they are immune to cyberattacks and their DNA is resistant to natural disasters and weather-induced degradation. In addition, the CRISPR-Cas system is a very efficient way of storing data thanks to the sequential way in which the data are introduced, allowing the retrieval of stored information much more quickly (via genetic analysis called genotyping). Thanks to new technologies, this information storage system can be extended to other types of organisms as a module: it can be integrated into a plasmid containing the required Cas genes and inserted into any cell. Thus, one could create a 'hard drive' from an extremophile organism such as the hyperthermophilic archaea *Pyrococcus furiosus* and store information in a medium that would otherwise destroy the DNA or current storage devices. We could then have all the filmography of Charlie Chaplin on the surface of volcanic rocks at 120 ° C!

The concept of using DNA for storing digital data is not new, as it appeared a few decades ago. The first person to convert binary data into DNA is the artist Joe Davis, in collaboration with Harvard researchers, in 1988. The DNA sequence, which they inserted into E. coli, encoded only 35 bits. Once organized into a 5 × 7 matrix, with binary 1s corresponding to the dark pixels and 0s corresponding to the bright pixels, they formed the image of an ancient German rune representing life and female Earth ("Microvenus"). A recent study by the European Bioinformatics Institute and Harvard showed that advances in modern DNA manipulation methods could today make the storage of digital data in DNA both practical and feasible. Several research groups, including the University of Illinois at Urbana-Champaign, ETH Zurich and Columbia University are working on this problem. Technology giants have also shown interest in the biological storage of information. Microsoft, for example, plans to add DNA data storage to its cloud services.

Conventional storage devices such as DVDs, USB drives, and hard drives store digital data by changing the optical, electrical, or magnetic properties of the material to store them as 0s and 1s. With DNA, the idea is the same, but a different process must be used. DNA molecules include long sequences of units called nucleotides. These nucleotides are called adenine, thymine, guanine, and cytosine, generally referred to as A, T, G, and C. Thus, the information must be stored in the DNA as a nucleotide sequence, instead of sequences of 0s and 1s as in the electronic media. Several methods have been used to achieve this conversion, but most of them fail to overcome the requirements imposed by the DNA's error-prone synthesis or use a non-universal coding scheme, which is useful only to that precise experiment.

In this project, we will suggest several coding schemes that universalize the conversion of a binary sequence into a nucleotide sequence, while respecting the constraints of DNA synthesis and the host bacterium. Then, we will analyze the effectiveness of our algorithm in an in vivo experiment where we will activate and exploit the CRISPR-Cas immune system of a colony of Escherichia coli BL21(DE3) bacteria in order to introduce into their genome two fragments of DNA which encode the text "hello.".

2.2. Compared to other technologies

Theoretically, 1 gram of DNA could store $4.5\,x\,10^{20}$ bytes (1 byte = 8 bits), that is 450 exabytes (EB), or why not 450 000 000 hard disks of 1 terabyte (TB) (Supplementary Info 1). Obviously, the practical maximums would be lower by several orders of magnitude depending on the coding method used, but this method would nonetheless easily surpass any technology currently in use (Table 1).

2.3. The algorithm

Since DNA has 4 possible nucleotides (A, T, G or C), this is a base 4 numeral system. Binary is a base 2 numeral system. Knowing this, the simplest conversion algorithm would be a straightforward conversion from base 2 to base 4: A would correspond to 00, T to 01, G to 10, and C to 11. This algorithm could reach the theoretical limit of information density storable on 1 nucleotide, that of 2 bits per base. However, DNA synthesis is error-prone when the sequence contains long homopolymers (mononucleotide repeats) or extreme AT or GC levels.

Example: ‘E=mc²’ is:

1. « 0100010100111101011011010110001100110010 » in binary,
2. « 10110331123112030302 » in base 4,
3. « TA**TT**A**CCTT**GC**TT**GACACAG » in DNA (homopolymers are bold).

In this example, the homopolymers are negligible. However, if this algorithm was used to convert a computer file that may easily have very long repetitions of 0s or 1s, the resulting homopolymers would be $\frac{l}{2}$ in length, where l is the length of the 0 or 1 repetitive sequence. This could mean hundreds of repetitions.

These homopolymers are not only problematic for DNA synthesis but also interfere with our goal of turning a bacterial colony into a hard disk. This led us to design our own binary to DNA conversion algorithm that meets the requirements of both the bacterium and the error-prone DNA synthesis. Since we are going to exploit the adaptation stage of the Escherichia coli type I-E CRISPR-Cas immune system to insert our own information, the DNA sequences must therefore have the properties of a protospacer.

The following mechanism was designed to store information in Escherichia coli BL21(DE3) which has a type I-E CRISPR-Cas immune system. This mechanism can be adapted to any other CRISPR-Cas system thanks to the versatility of the algorithm whose code can be found on GitHub (link at the end of the project). That said, all the variables that will be mentioned are strictly used for this type of CRISPR-Cas system. This system imposes three requirements (Shipman et al. 2017):

1. **The PAM must not appear within the spacer sequence**

The PAM is like a signal for Cas1-Cas2 to cleave the protospacer, so it should only appear at its beginning. The sequence of the content must therefore not include the 5' – AAG – 3' sequence or its reverse complementary 5' – CTT – 3'.

2. **The protospacer should not have long nucleotide repeats**

The protospacer sequence should not include homopolymers longer than or equal to 4 (e.g. AAAA) since these can cause DNA synthesis errors.

3. **The spacer GC level must be greater than or equal to that of AT**

Because AT-rich sequences are less thermodynamically stable than those rich in GC and often perform special functions in Escherichia coli, (e.g. they correspond to origins of replication (Rajewska et al. 2012)), the GC level of the protospacer must be greater than that of AT.

The information that one wishes to store in the genome of the bacterium corresponds to 33 nt spacers separated by repetitive sequences within the array of the CRISPR 1 locus. Every spacer stores 24 bits of information (the information format will be detailed a little later). First of all, the binary sequence that corresponds to the digital information that we want to store (a fairly easy step) must be divided into 24-bit sub-sequences.

Example: ‘Newton’

[010011100110010101110111][011101000110111101101110]

Then, the appearances of 0s and 1s of each subsequence are counted. The binary digit that appears most will be associated with nucleotides G and C, the other one will be associated with nucleotides A and T. This step ensures that the GC level of the final sequence will be greater than or equal to that of AT, thus complying to rule 3.

[010011100110010101110111] → 10 zeroes, 14 ones → 0 corresponds to A and T, 1 corresponds to G and C

In the last step, the DNA sequence is built. Each binary digit is replaced by the corresponding nucleotide different from the one previously used.

[010011100110010101110111] → [AGTAGCGTACGTACTGACGCTGCG]

The mononucleotide repetitions are then impossible; rule 2 is thus respected. Therefore, rule 3 is also respected, since the canonical PAM contains repetitions (5’ – AAG – 3’).

After having built the DNA sequence, we add a ‘marker bit’ which defines the binary digit to which the nucleotides G and C correspond to. If G and C correspond to 0, we add an A or a T at the end of the sequence. Otherwise, we add a G or a C. The nucleotide alternation rule must be respected all throughout the protospacer sequence.

G et C correspond to 1 → [AGCTGCGATCGATCAGTCGCAGCG][C]

This is the essence of the algorithm. Obviously, as this sequence only encodes part of the content, we must add an index, a number which represents the position of the information in the bit stream. In this project, we chose a length of 4 nucleotides for the index, but it is variable as well. We can count on 4 bits $2^4 = 16$ sequences, which is more than enough for our experiment. As we are going to insert only 2 protospacers, we could have even used a single nucleotide to represent the index, but we will carry on with 4 for demonstration purposes.

In addition to the index, we will include a simple error detection method: a parity bit represented by a nucleotide. If during decoding it equals 0, the number of nucleotides G and C of the index and content must be even. Otherwise, it must be odd. In case the parity of the number of appearances of G and C is not in consensus with the parity bit, it means that the spacer has been mutated and the information is corrupted. The set consisting of the parity bit and the index is called the ‘header’. The set consisting of the header and the content is called a ‘packet’.

Finally, as this sequence is part of a protospacer, one must add the PAM and the AAM, respectively at the beginning and at the end of the constructed sequence. In practice, we add the PAM at the beginning of the encoding process in order to respect the rule of the alternation of nucleotides and the AAM at the end of the encoding process. The PAM, placed before the content, has the 5' – AAG – 3' sequence. The AAM has the dinucleotide sequence 5' – GA – 3' and is placed after the content. Although its importance is not yet established, we chose to include it as well in order to maximize the probability of success of the project.

The steps are then as follows:

1. Add the PAM.
2. Count the number of appearances of each binary digit (0 and 1) in the index and the content.
3. Associate each binary digit to a pair of nucleotides with respect to the number of appearances of each digit in the index and the content.
4. Add the parity bit.
5. Add the index converted to nucleotides.
6. Add the content converted to nucleotides.
7. Add the marker bit.
8. Add the AAM.

See Figure 8 for a visualization of the protospacer format.

Example: ‘Newton’

Protospacer #1:

5’ – [AAG][T][ATAT][ACTAGCGTACGTACTGACGCTGCG][C][GA] – 3’

Protospacer #2:

5’ – [AAG][T][ATAC][TGCGACTATGCAGCGCTGCAGCGT][G][GA] – 3’

In case the length of the binary sequence is not a multiple of 24, 'filler' nucleotides must be added and a way to detect them as well. As one of the requirements of the protospacer is a high GC level, the redundancy added will consist of GC repeats. In order to detect the redundancy (i.e. the end of the valid content), we will use a method that avoids any PAM type sequence (5 '- AAG - 3' or 5 '- CTT - 3') within the protospacer. We then conceived the following rule:

1. If the last two nucleotides of the content are complementary, repeat the last nucleotide different from these two.
2. Otherwise, repeat the penultimate nucleotide.

Example:

5' – [(…)AGT**C**A][**C**][GCGC(…)] – 3' → The last two nucleotides are not complementary (CA), then the penultimate nucleotide of the content (**C**) is repeated.

5' – [(…)A**G**TAT][**G**][CGCG(…)] – 3' → The last two nucleotides are complementary (AT), then the last appearance of a different nucleotide (**G**) is repeated.

Extreme case: If the configuration of the protospacer results in the use of a single pair of bases all throughout the protospacer, the redundancy is introduced by repeating the last nucleotide (e.g. 5 '- [(...) GCGCG] [G] [CGCGC (...)] - 3 ').

While decoding, our algorithm detects that the alternating nucleotide rule is broken and thus detects the beginning of redundancy. All these rules ensure the detection of redundancy and the avoidance of a PAM within the protospacer.

Note: Although the functionality of the algorithm can be extended to other CRISPR systems, i.e. the configuration of the different sectors of the protospacer, in case the PAM differs from 5' – AAG – 3' or 5' – CTT – 3', the guarantee that it will not appear in the protospacer no longer holds. We assume no responsibility for any damage caused by improper use of the algorithm's functionality. On the other hand, we guarantee a proper functionality of the algorithm in type I-E CRISPR-Cas systems like the one found in Escherichia coli.

2.4. The experiment

2.4.1. Prior information

To transform bacteria into molecular hard disks, we will exploit the adaptation step of the type I-E CRISPR-Cas system of a colony of Escherichia coli BL21 (DE3). This consists of inserting protospacers into the cell that encode the text "hello.", converted from binary to nucleotides thanks to our coding algorithm. The bacteria will treat these protospacers as viral DNA fragments and will trigger the adaptation mechanism of their CRISPR-Cas immune system. The Cas1-Cas2 complex, the adaptation machinery, will integrate these protospacers as spacers into the CRISPR array of the CRISPR 1 locus and thus store our data.

a) The CRISPR-Cas system is naturally dormant

The CRISPR-Cas system is inactive in most Escherichia coli bacteria because of the H-NS repression factor (encoded by the *hns* gene), a protein that binds to hundreds of promoters among which are those of the *cas* genes and blocks their transcription. Therefore, the system does not acquire new spacers naturally and even if it were, these spacers would not protect the bacterium against phages because the important interference proteins are not expressed. The resistance against phages in Escherichia coli will thus be much easier to achieve by other means, in particular by adsorption mutations (for example in LamB). Thus, so far there is no wild Escherichia coli that can be 'trained' to protect against phages using the CRISPR-Cas system. Using a *Δhns* mutant (without the *hns* gene), one could, within two weeks, get protection against plasmids and potentially against persistent phages such as M13. However, the probability that such experiments will be carried out for such a long time in the specific mutant without prior knowledge of CRISPR-Cas is small. On the other hand, in other bacteria, such as in *Streptococcus thermophilus*, the system is in fact a dominant defense system (with almost all phage resistance due to this system).

b) Activation of the immune system

Our experience would not be successful without the activation of the CRISPR-Cas system which is naturally dormant. In addition, the genome of the strain of bacteria that we have at our disposal does not carry any *cas* genes - it only has two CRISPR loci. Thus, we had two possible choices: either activate the system by inserting a plasmid containing the *cas1* and *cas2* genes that code for the Cas1-Cas2 adaptation complex, or use a strain of K12-Δhns bacteria that carries all of the *cas* genes and does not carry the *hns* gene that codes for the H-NS suppression factor. We did not choose the mutant for two reasons: firstly, the suppression of the hns gene does not have a local impact and reduces the competence of the bacteria (Solomon 2012) and secondly, Escherichia coli K12 also carries the genes that code for the Cascade interference machinery that could destroy our protospacers.

To activate the CRISPR-Cas system, we chose the plasmid pCas1+2 deposited by Udi Qimron (*Addgene plasmid # 72676*) which contains the *cas1* and *cas2* genes extracted from a strain of Escherichia coli K12. Since our bacteria will only integrate the spacers and not destroy them because of the absence of genes coding for the Cascade interference machinery, the probability of integration of our DNA is maximized.

The *cas1* and *cas2* genes are placed under a lac operator and a T7 promoter. Thus, their transcription is permitted only in the presence of T7 RNA polymerase and allolactose or IPTG. In order to avoid a ‘leaky’ expression (in case the bacterium does not have the *lacI* gene or does not express it enough - which may result in an involuntary expression of the protein of interest), the plasmid also carries the *lacI* gene which codes for the repressor of the lac operon. The plasmid has CloDF13 as ORI (origin of replication) resulting in 20-40 copies of the plasmid per cell. Finally, it also carries a streptomycin and spectinomycin (SmR) resistance gene to help select cells that have received the plasmid.

As the *cas1* and *cas2* genes are placed under a T7 promoter, our bacteria must synthesize T7 RNA polymerase since otherwise our genes of interest would not be transcribed. T7 RNA polymerase is derived from bacteriophage T7, which means that it is not naturally present inside a bacterium, except in the case of a transduction. Our strain of bacteria Escherichia coli BL21(DE3) contains the λDE3 prophage which houses the lacUV5 (similar to the lac promoter) promoter-regulated T7 RNA polymerase gene. Thus, one can induce the expression of T7 RNA polymerase and the transcription of our genes with IPTG.

These genes are regulated by a lac operator whose repression is not very strict, that is to say that they are transcribed at a basal level even in the absence of IPTG. This would be a problem if the proteins of interest were toxic, but as Cas1 and Cas2 are not, the weak expression of these is not problematic for our project even if the Cas1-Cas2 complex begins to integrate some protospacers present inside the cell (which is very unlikely except in the case of contamination by extrachromosomal DNA).

c) Protospacer design

For this proof of concept, we chose to encode the text "hello.” in two protospacers whose sequences have been established by our conversion algorithm.

Protospacer #1:

5' – [AAG][T][CGCG][CATGACGCGTACGTCAGTACTAGC][A][GA] – 3’

Protospacer #2:

5' – [AAG][C][TATG][ACGTCGATACGTCGCGATCAGCGT][G][GA] – 3’

Since DNA can only be synthesized in the 5' to 3' direction, it is impossible to synthesize double stranded DNA because of the antiparallel property of DNA. We then had two possible choices: either to synthesize two complementary oligonucleotides for each protospacer (which would have annealed), or to synthesize a long self-complementary oligonucleotide that would have formed a 'hairpin' (that is to say that the second half of it would be complementary to the first half in order to anneal). We found that we could use 'hairpin' protospacers consisting of 23 bp duplex DNA at the core of the protospacer (J. Nuñez et al. 2015; Wang et al. 2015) with 7 bases not completed at the 5' end of the lower strand (which includes the PAM) and 5 bases not completed at the 5' end of the upper strand forming a loop. *Reminder: The Cas1a part of the Cas1-Cas2 complex recognizes the 5' – CTT – 3' PAM of the 3' overhang of 7 nt of the protospacer.* Also, the hairpin avoids the dissociation of the complementary strands and their segregation in different cells during heat shock or electroporation. Finally, the structural data of the Cas1-Cas2 complex and the experiments of Seth L. Shipman et al. (Shipman et al. 2017) indicate that 5' – AAG – 3' PAM is dispensable in the upper strand while its complementary 5' – CTT – 3' in the lower strand is indispensable (see the adaptation part of CRISPR-Cas system, section 1.1 of the project). The final design of our protospacers is based on that of Shipman et al. (Shipman et al. 2017).

To obtain the hairpin, one must add to the sequence resulting from the algorithm its reverse-complement whose first 5 nucleotides are deleted (those of the loop). Then, to get the 3' overhang, one must delete the first 7 nucleotides of the final sequence (that is to say the 5' end).

Example:

Sequence:

5' – AAGTCGCGCATGACGCGTACGTCAGTACTAGCAGA – 3'

Reverse-complement:

5' – ~~TCTGC~~TAGTACTGACGTACGCGTCATGCGCGACTT – 3'

Sequence:

5' – ~~AAGTCGC~~GCATGACGCGTACGTCAGTACTAGCAGA
TAGTACTGACGTACGCGTCATGCGCGACTT – 3'

Sequence:

5' – GCATGACGCGTACGTCAGTACTAGCAGATAGTACTGACGTACGCGTCATGCGCGACTT – 3'

The 5' – 3' sequences of the 'hairpin' protospacers encoding "hello." are then as follows (Figure 9):

Protospacer #1:

[GCATGACGCGTACGTCAGTACTA][GCAGA][TAGTACTGACGTACGCGTCATGC][GCGACTT]

Protospacer #2:

[GACGTCGATACGTCGCGATCAGC][GTGGA][GCTGATCGCGACGTATCGACGTC][ATAGCTT]

2.4.2. Protocol

First, we inserted the pCas1+2 plasmid (*Addgene plasmid # 72676*, a plasmid carrying the *cas1* and *cas2* genes) via heat shock into chemically competent Escherichia coli BL21(DE3) bacteria to activate the adaptation process of the CRISPR-Cas system. Then, we induced the expression of the T7 polymerase and the Cas1 and Cas2 proteins in a 1 mM IPTG LB solution. We made the bacteria electrocompetent and we electroporated our protospacers encoding "hello." (at a 5 µM concentration). To verify the integration of the protospacers, we performed several PCR reactions on the genomes of cells surviving the electroporation, followed by gel electrophoresis on the products until we identified an expansion of the CRISPR array. The positive PCR products (~3%) were purified and digested with restriction enzymes (MluI and NruI) whose recognition sites were unique in each of the protospacers and did not appear in the CRISPR locus. Samples that were successfully digested were sent to sequencing.

2.5. Conclusion

Following our experiments, we can thus conclude on our project: the successful integration and recovery of the message shows that the storage of data in the genome of an organism is possible.

The storage of information in DNA is sustainable thanks to the stability of the molecule. Indeed, it can resist tens, even hundreds of thousands of years under appropriate conditions. For example, the oldest sequenced DNA could be dated between -450,000 and -800,000 years after its discovery in Greenland (Willerslev et al. 2007) and we were able to sequence the entire mammoth genome (Palkopoulou et al. 2015). Theoretically, DNA could resist 6.8 million years since its half-life is 521 years, but according to scientists, it would no longer be readable after 1.5 million years.

Moreover, this method is sustainable thanks to the defense mechanisms of the host cell. Not only can DNA survive for thousands of years, but it is also protected from exogenous organisms within the bacteria. For example, it is especially protected from exonucleases by the plasmid form of the genome. In addition, even if the bacterium dies for example because of lack of food, the genomic DNA is not damaged (prokaryotic organisms do not contain lysosomes) and can easily be recovered. Therefore, we would not need any infrastructure to protect the DNA.

Then, storing data in DNA is very effective thanks to its compactness. As 1 gram of DNA can theoretically store 450 exabytes of information, a few grams of DNA could store all the data produced by humanity throughout history. A more concrete example would be the storage of the 100 petabytes of a data center in just 220 μg of DNA. Thus, we would no longer need to use hundreds of hectares to build data centers since we could archive all the information in DNA.

Moreover, the almost null amount of energy required by bacteria shows the effectiveness of this method. Although bacteria actually require nutrients to subsist, the amount of culture medium needed to reach a large enough colony is tiny. To avoid too many mutations in the stored data, at a given moment, the replication of the bacteria must be stopped either by freezing them or by starvation. As for DNA, it does not require any source of energy for its existence as a salt. As a result, carbon emissions from energy consumption by existing data storage devices would be significantly reduced.

Finally, this method is effective thanks to the exponential replication of bacteria. In fact, by replicating, the bacteria make several copies of the same genomic or plasmid DNA. Having multiple copies of the same DNA can be used to correct mutations that DNA can undergo without the need to synthesize other oligonucleotides. Thus, we could have a method of error correction resulting from the natural replication of the bacteria, which would mean again a reduction of data storage costs.

Moving on to the process of integrating DNA into the bacterial genome, although it may seem long, complex, and exhausting, the sustainability benefits that it presents must not be ignored in a world where electronics are so fragile in the face of natural disasters, which, moreover, are unpredictable.

In addition, our algorithm can also be used to encode information in isolated DNA in salt form. Indeed, the properties of it are favorable to an error-free synthesis of DNA. For example, the algorithm constructs sequences that do not contain homopolymers and whose dominant base pair can be controlled. The algorithm also has the advantage of being able to decode the message using either the strand on which the spacer is introduced or its complementary in the 5' - 3' direction. If one wishes to have a balanced ratio of AT and GC, one can simply encode the binary data using a Huffman code to balance the number of binary 0s and 1s.

Then, the information storage system can not only be used in multiple bacterial colonies, but can also be extended to other strains: thanks to new technologies, it is possible to integrate into a plasmid an artificial CRISPR locus accompanied by the genes coding for the proteins necessary for CRISPR-Cas adaptation. By inserting this plasmid, it would now be possible to store data for example in extremophile organisms to ensure the survival of information in hostile environments. The versatility of our algorithm adds to the efficiency of this information storage system, making it modular and accessible to all organisms (even eukaryotes).

However, although this method is indeed sustainable, its effectiveness can still be debated because of the price of DNA and the speed of its synthesis and sequencing. Today, it is impossible to replace a hard disk of an ordinary computer with a DNA disk or bacteria since it could not meet the real-time requirements of the user. The synthesis of a nucleotide costs about $0.50, which is too expensive to be able to store data such as digital images domestically. Therefore, the current practical use is in data archiving and not in short-term storage.

Nevertheless, this process can certainly be improved in the future to address inefficiency issues. This proof of concept is intended to pave a new way of storing information to meet the archiving needs of permanent data such as scientific research that is essential to the intellectual survival of our species. Moreover, the current statistics show an exponential increase in the speed of synthesis and sequencing of DNA, accompanied by an exponential decrease in their costs.

Thus, we have just shown that the encapsulation of a modified DNA in a bacterial organism is feasible. Although today it is only an efficient and sustainable alternative to archiving data and not storing daily information, this method of storing digital information in DNA still shows its potential to extend and be mass distributed (scalable), and to be made available to the public, depending on investments and the advancement of technologies that seem to be favorable (Graph 1).

Acknowledgements:

We thank Dr. Andrei-José Petrescu, PhD – Head of the Department, Dr. Cătălin Lazăr, PhD – Viral Glycoproteins Department, Dr. Norica Nichita, PhD – Head of Viral Glycoprotein Department, and their colleagues from the Institute of Biochemistry of the Romanian Academy that have helped us both materially and intellectually throughout the project, Dr. Seth L. Shipman, PhD, postdoctoral researcher of the Department of Genetics of Harvard Medical School for the advice and help offered, and our science teachers for the help, encouragement and advice they have given us.

Useful links:

The DNAdrive algorithm: https://github.com/alexbrt/DNAdrive

Supplementary Material:

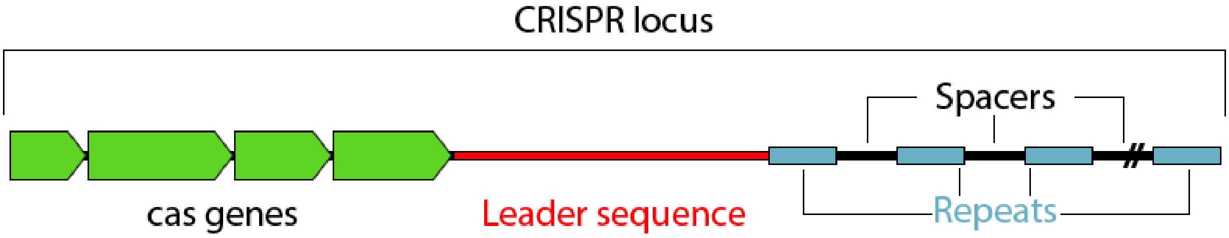


Figure 1 – CRISPR locus structure

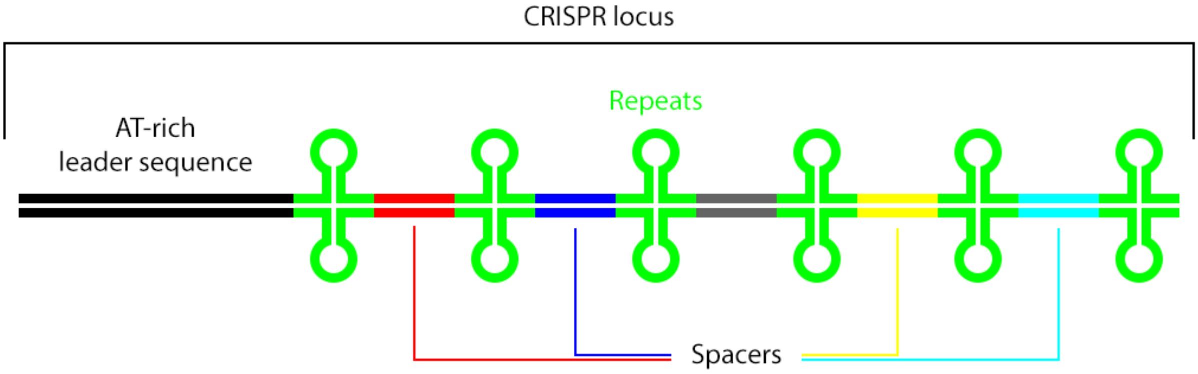


Figure 2 – CRISPR spacer-repeat array

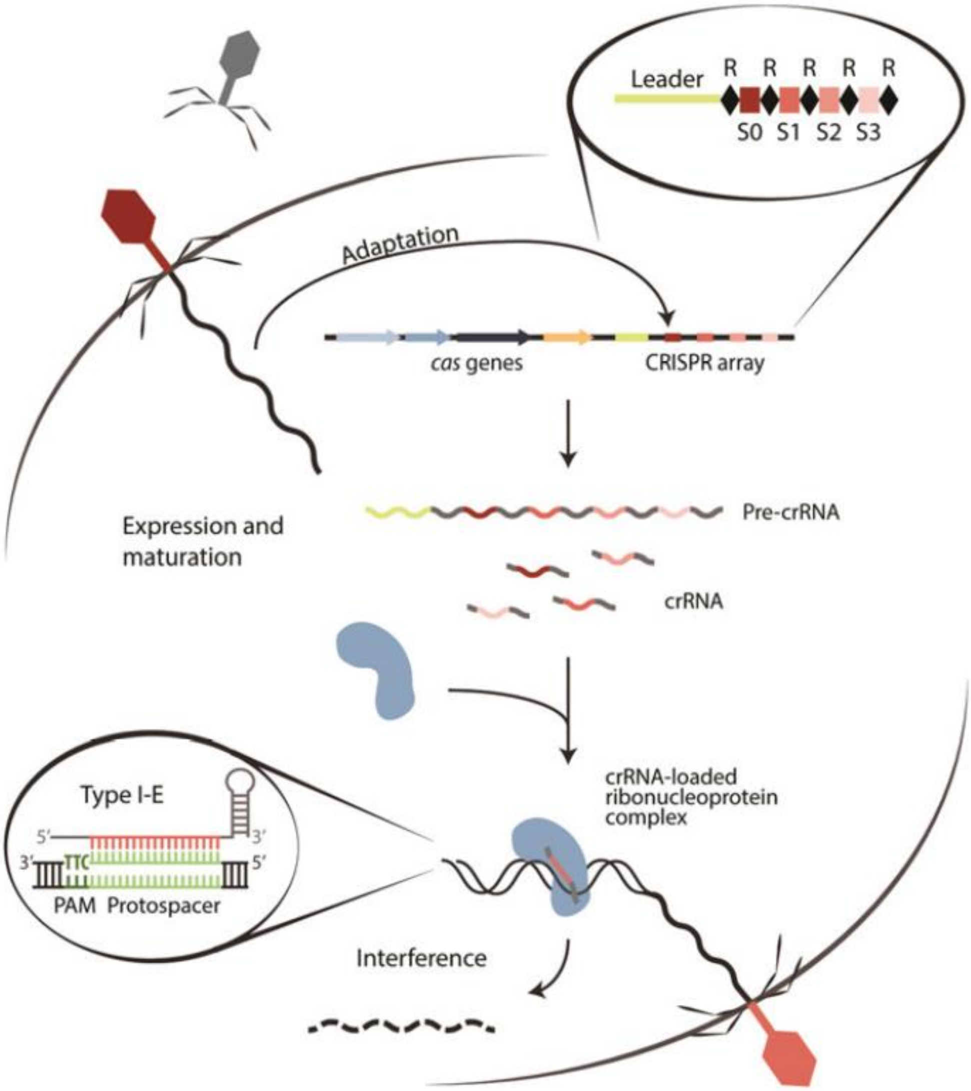


Figure 3 – CRISPR-Cas immune system mechanism stages (image by Amlinger, Lina)

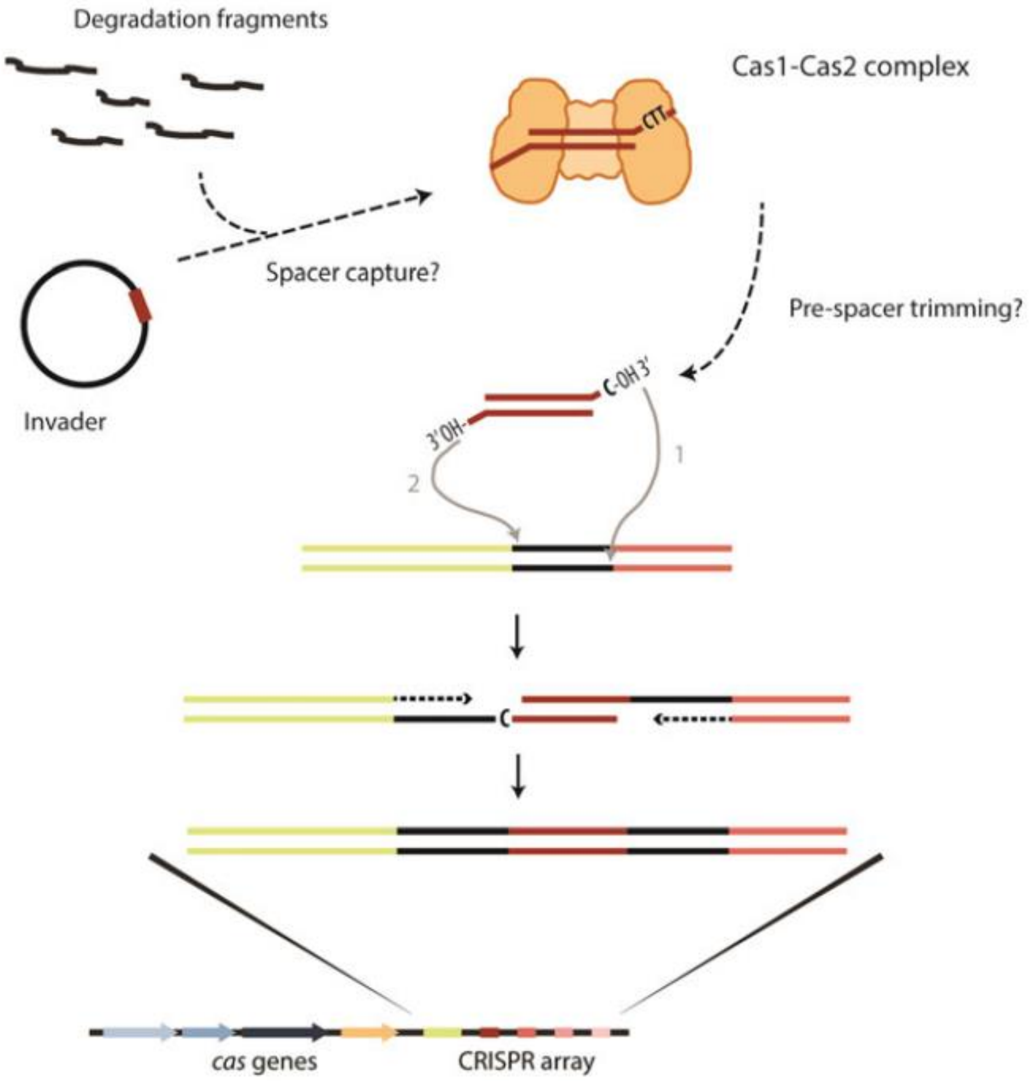


Figure 4 – CRISPR adaptation (image by Amlinger, Lina)

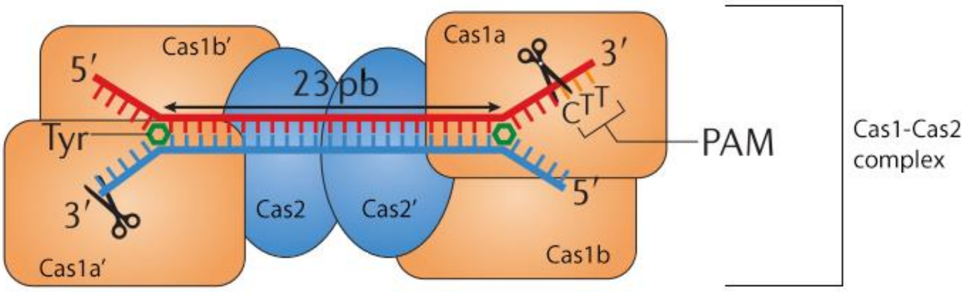


Figure 5 – Cas1-Cas2 complex bound to a double-forked protospacer (image by Amitai, Gil, and Sorek, Rotem)

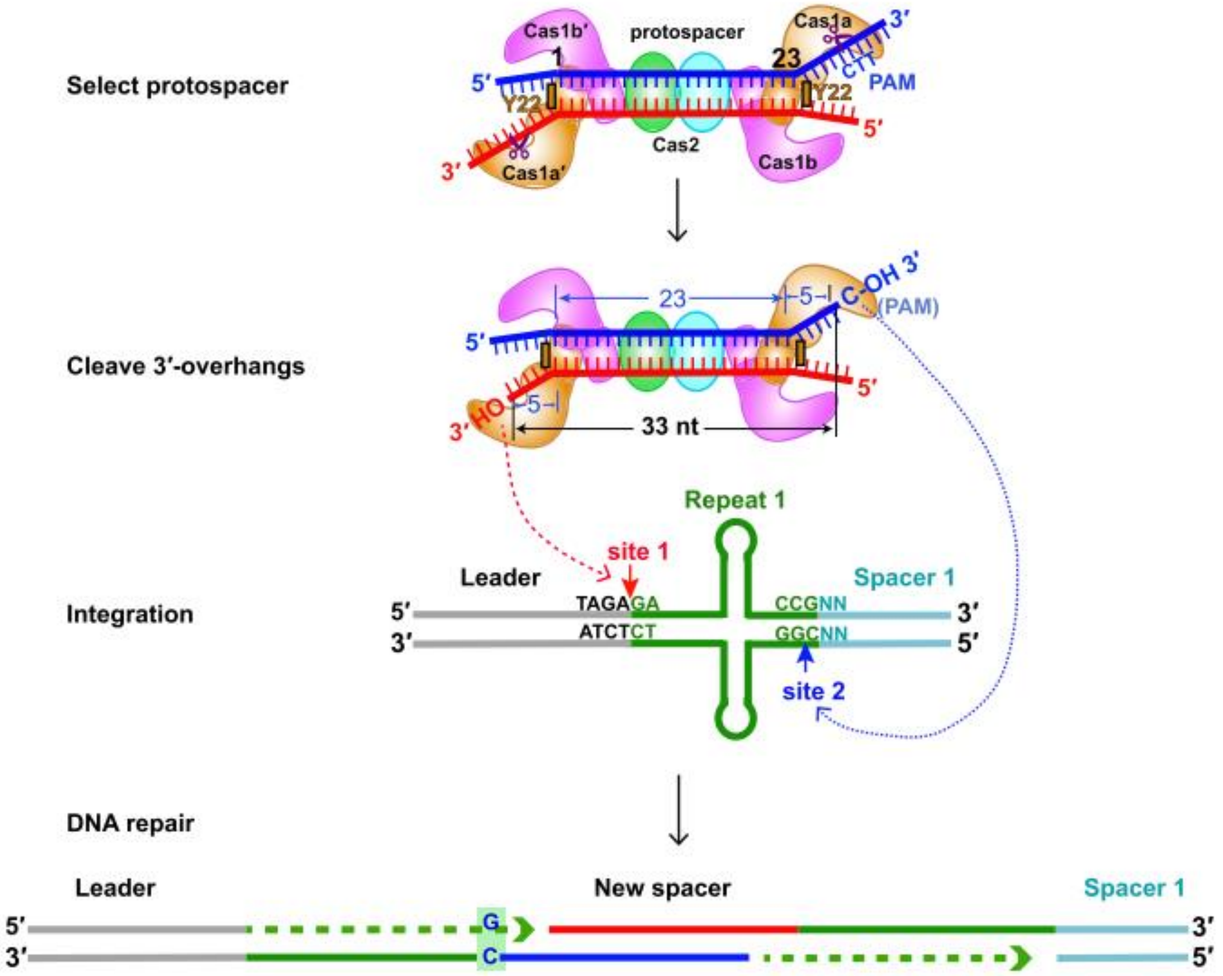


Figure 6 – Spacer integration overview (image by Wang et al.)

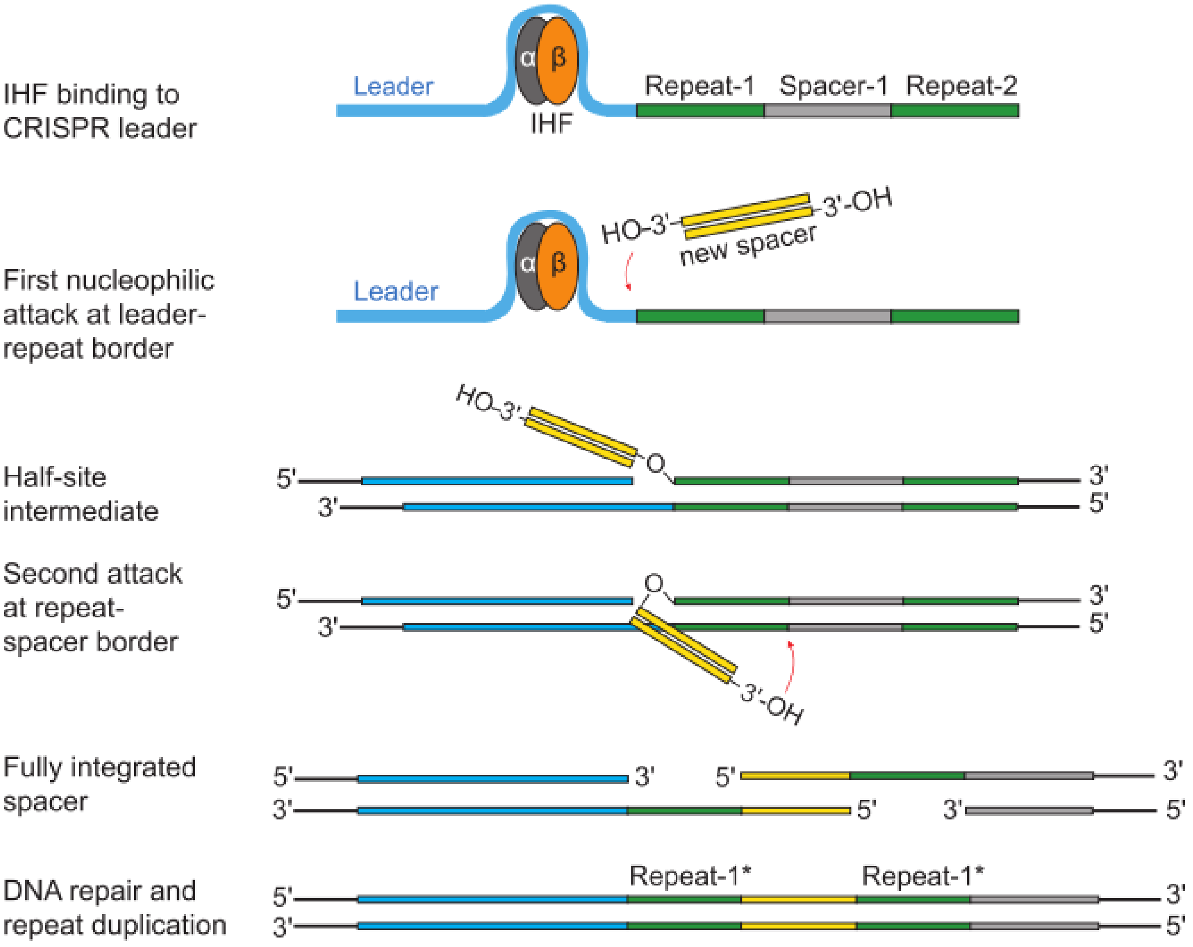

IHF binding to CRISPR leader
Leader
α
β
IHF
Repeat-1
Spacer-1
Repeat-2
First nucleophilic attack at leader-repeat border
HO-3'
3'-OH
new spacer
Leader
Half-site intermediate
HO-3'
O
5'
3'
3'
5'
Second attack at repeat-spacer border
O
3'-OH
Fully integrated spacer
DNA repair and repeat duplication
Repeat-1*
Repeat-1*

Figure 7 – Spacer integration steps (image by Nuñez et al.)

Supplementary Info 1 – DNA information density calculation

The theoretical DNA information density was calculated using 2 bits per ssDNA nucleotide.

DNA cannot exist as an isolated molecule outside the cell of a living being because the repulsion of the two strands, caused by the negative charges of the phosphodiester bonds, would cause its denaturation. In addition to a neutral pH (7.0) and a low temperature, Na$^+$ ions must be added to neutralize the molecule. Whenever two nucleotides are bound together, a molecule of water is eliminated (or 'lost') by a process called dehydration synthesis. That being said, in order to obtain an accurate estimate of the molecular weight of a DNA strand, the loss of water and the addition of sodium must be considered.

$$\begin{aligned} Molar\ mass\ of\ a\ DNA\ strand = {} & nA * (mA - mH - mO + mNa) \\ & + nT * (mT - mH - mO + mNa) \\ & + nG * (mG - mH - mO + mNa) \\ & + nC * (mA - mH - mO + mNa) \end{aligned}$$

Where *nA* is the number of adenine bases, *mA* is the molar mass of deoxyadenosine monophosphate, *nT* is the number of thymine bases, *mT* is the molar mass of deoxythymidine monophosphate, *nG* is the number of guanine bases, *mG* is the molar mass of deoxyguanosine monophosphate, *nC* is the number of cytosine bases, *mC* is the molar mass of deoxycytidine monophosphate, *mH* is the molar mass of hydrogen, *mO* is the molar mass of oxygen, and *mNa* is the molar mass of sodium. This means that:

$$Molar\ mass\ of\ a\ DNA\ strand = nA * 335.2 + nT * 326.2 + nG * 351.2 + nC * 311.2$$

For a DNA strand with balanced GTCA levels and with sodium associated to it, the average molar mass of a nucleotide would be:

$$\frac{335.2 + 326.2 + 351.2 + 311.2}{4} = 330.95\ g.mol^{-1}$$

The number of mol in 1 g of DNA is $\frac{1}{330.95}\ mol$

The number of nucleotides in 1 g of DNA is equal to:

$$Avogadro's\ constant * number\ of\ mol = \frac{6.022 * 10^{23}}{330.95}\text{nucleotides}$$

The mass of 1 nucleotide is $\frac{1}{\frac{6.022*10^{23}}{330.95}}\,g$ so the 'mass' of 1 bit is:

$$\frac{1}{2 * \frac{6.022 * 10^{23}}{330.95}} = 2.75 * 10^{-22}\ g$$

So, theoretically, 1 gram of DNA could store $4.5\ x\ 10^{20}$ bytes (1 byte = 8 bits), or 450 exabytes (EB) or why not 450 000 000 hard disks of 1 terabyte (TB). Obviously, the practical maximums would be lower by several orders of magnitude depending on the coding method used.

Table 1 – Other technologies

Label	Date	Density	Info
Magnetic tape	1928	256 bits / in^2	The first magnetic tape recorder, the Uniservo Univac, recorded at a density of 128 bits / in on a half-inch magnetic tape.
Hard disk	1956	2 000 bits / in^2	The IBM 350 disk storage unit, the first disk drive, stores 3.75 MB on 50 24-inch diameter magnetic disks containing 50,000 sectors, each of which holds 100 alphanumeric characters, for a capacity of 5 million characters.
CD	1982	~0.50 Gbit / in^2	CD-ROM of 120 mm diameter and 700 MB capacity (actually 847 MB).
DVD (SL)	1996	~3.27 Gbit / in^2	DVD-SS-SL of 120 mm diameter and 4.7 GB capacity (actually 5.5 GB).
Blu-ray (SL)	2002	~14.73 Gbit / in^2	Blu-ray-SL 120 mm diameter and 25 GB capacity.
Modern hard disk	2012	1 000 Gbit / in^2	Seagate 1 Tb (terabit) / in^2.
Atom-based magnetic memory	2012	~71 684 Gbit / in^2	12 iron atoms / bit, 9 nm^2 / bit, low-temperature nonvolatile memory.
Atom-position-based memory	1992	645 160 Gbit / in^2	"IBM" written with xenon atoms spaced 1 nm across a 14 * 5 nm^2 lattice, 1 bit / nm^2.
Quantum holography	2008	~890 320 Gbit / in^2	35-bit image pairs, 17 * 17 nm^2 aerial atoms and reading space of ((4 * 5) / (17 * 17)) * 20 bits / nm^2 = 1.38 bits / nm^2.
DNA	2017	~1 897 529 Gbit / in^2 for ssDNA or ~3 795 058 Gbit / in^2 (but rather 3.6e12 Gbit / g)	Ordered DNA, 2 nm helix diameter, nucleotide length of 0.34 nm, 10 base pairs per complete rotation.

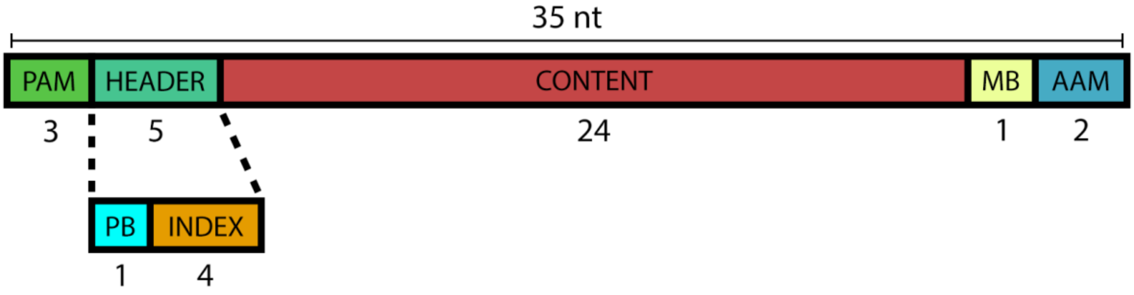
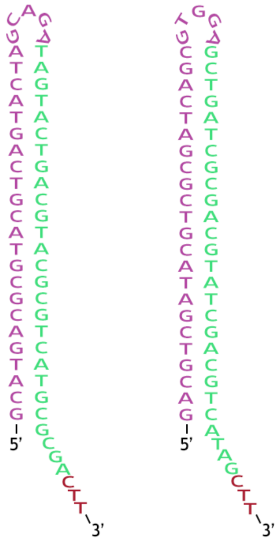
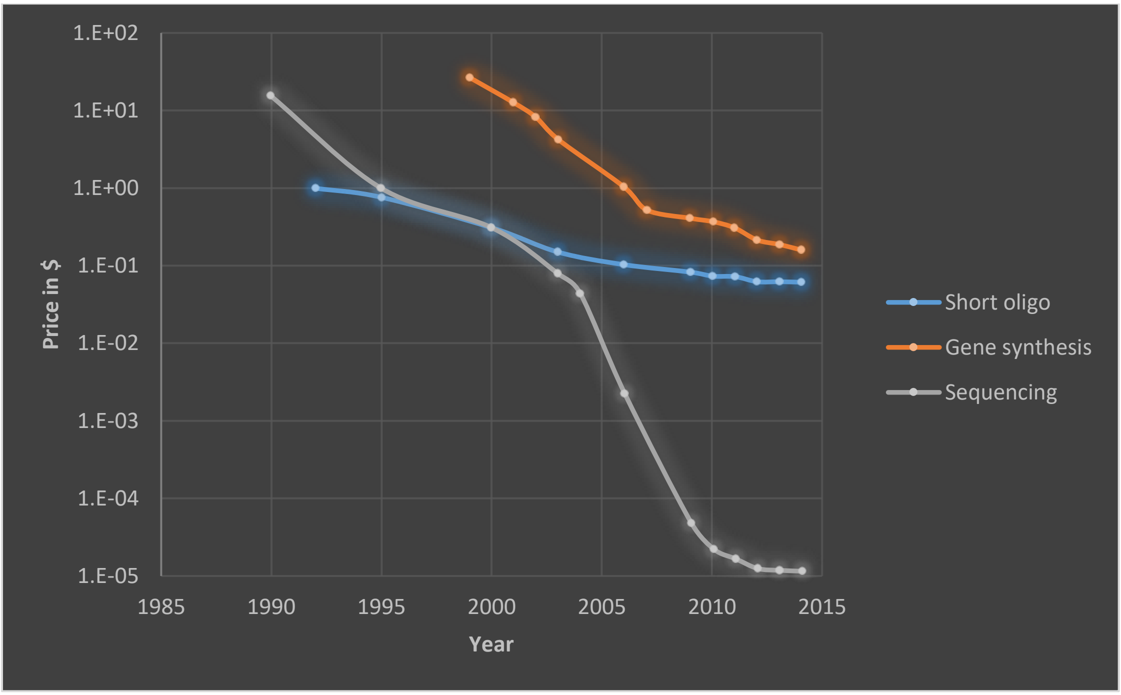


Figure 8 – Protospacer structure: PB is the parity bit, MB is the marker BIT

Figure 9 – Hairpins encoding “hello.”


5'
GCATGACGCGTACGTCAGTCATAGCAGATAGTACTGACGTACGCGTCATGCGCGACTT
3'
5'
GACGTCGATACGTCGCGATCAGCGTGGAGCTGATCGCGACGTATCGACGTCATAGCTT
3'



Graph 1 – Price evolution

References and bibliography:

Amitai G, Sorek R. 2016. CRISPR-Cas adaptation: Insights into the mechanism of action. Nat. Rev. Microbiol. 14:67–76. doi:10.1038/nrmicro.2015.14.

Bancroft C, Bowler T, Bloom B, Clelland CT. 2001. Long-term storage of information in DNA. Science (80-.). 293:1763–1765.

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 (80-.). 315:1709–1712. doi:10.1126/science.1138140.

Barrangou R, Marraffini LA. 2014. CRISPR-cas systems: Prokaryotes upgrade to adaptive immunity. Mol. Cell 54:234–244. doi:10.1016/j.molcel.2014.03.011.

Bolotin A, Quinquis B, Sorokin A, Dusko Ehrlich S. 2005. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 151:2551–2561. doi:10.1099/mic.0.28048-0.

Bonnet J, Colotte M, Coudy D, Couallier V, Tuffet S. 2010. Chain and conformation stability of solid-state DNA : implications for room temperature storage. 38:1531–1546. doi:10.1093/nar/gkp1060.

Brouns SJJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJH, Snijders APL, Dickman MJ, Makarova KS, Koonin E V., van der Oost J. 2008. Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. 321:960–964.

Coté AG, Lewis SM. 2008. Mus81-Dependent Double-Strand DNA Breaks at In Vivo-Generated Cruciform Structures in S. cerevisiae. Mol. Cell 31:800–812. doi:10.1016/j.molcel.2008.08.025.

Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. 2012. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. Nat. Commun. 3:945–947. doi:10.1038/ncomms1937.

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. J. Bacteriol. 190:1390–1400. doi:10.1128/JB.01412-07.

Díez-Villaseñor C, Guzmán NM, Almendros C, García-Martínez J, Mojica FJM. 2013. CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli. RNA Biol. 10:792–802. doi:10.4161/rna.24023.

EIA. 2017. Electric Power Monthly with Data for November 2016. U.S. Energy Inf. Adm.

Fineran PC, Charpentier E. 2012. Memory of viral infections by CRISPR-Cas adaptive immune systems: Acquisition of new information. Virology 434:202–209. doi:10.1016/j.virol.2012.10.003.

Gantz BJ, Reinsel D. 2011. Extracting Value from Chaos State of the Universe : An Executive Summary. :1–12.

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 8:1–10. doi:10.1186/1471-2105-8-172.

Hayes RP, Xiao Y, Ding F, van Erp PBG, Rajashankar K, Bailey S, Wiedenheft B, Ke A. 2016. Structural basis for promiscuous PAM recognition in Type I-E Cascade from E. coli. Nature 530:499–503. doi:10.1038/nature16995.

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. doi:10.1128/jb.169.12.5429-5433.1987.

Künne T, Kieper SN, Bannenberg JW, Vogel AIM, Miellet WR, Klein M, Depken M, Suarez-Diez M, Brouns SJJ. 2016. Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Mol. Cell 63:852–864. doi:10.1016/j.molcel.2016.07.011.

Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, Edgar R, Qimron U, Sorek R. 2015. CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature 520:505–510. doi:10.1038/nature14302.

Ma H, Marti-Gutierrez N, Park S-W, Wu J, Lee Y, Suzuki K, Koski A, Ji D, Hayama T, Ahmed R, et al. 2017. Correction of a pathogenic gene mutation in human embryos. Nature 548:413–419. doi:10.1038/nature23305.

Makarova KS, Craft DH, Barrangou R, Brouns

SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, et al. 2011. Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. 9:467–477. doi:10.1038/nrmicro2577.Evolution.

Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, et al. 2015. An updated evolutionary classification of CRISPR–Cas systems. Nat. Rev. Microbiol. 13:722–736. doi:10.1038/nrmicro3569.

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. doi:10.1099/mic.0.023960-0.

Nuñez J, Harrington LB, Kranzusch PJ, Engelman AN, Doudna JA. 2015. Foreign DNA capture during CRISPR–Cas adaptive immunity. Nature 527:535–538. doi:10.1038/nature15760.Foreign.

Nuñez JK, Bai L, Harrington LB, Hinder TL, Doudna JA. 2016. CRISPR Immunological Memory Requires a Host Factor for Specificity. Mol. Cell 62:824–833. doi:10.1016/j.molcel.2016.04.027.

Nuñez JK, Kranzusch PJ, Noeske J, Wright A V., Davies CW, Doudna JA. 2014. Cas1 – Cas2 complex formation mediates spacer acquisition during CRISPR – Cas adaptive immunity. Nat. Struct. Mol. Biol. 21:528–534. doi:10.1038/nsmb.2820.

Nuñez JK, Lee ASY, Engelman A, Doudna JA. 2015. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature 519:193–198. doi:10.1038/nature14237.

van der Oost J, Westra ER, Jackson RN, Wiedenheft B. 2014. Unravelling the structural and mechanistic basis of CRISPR–Cas systems. Nat. Rev. Microbiol. 12:479–492. doi:10.1038/nrmicro3279.Unravelling.

Pääbo S, Poinar H, Serre D, Jaenicke-Després V, Hebler J, Rohland N, Kuch M, Krause J, Vigilant L, Hofreiter M. 2004. Genetic Analyses From Ancient DNA. Annu. Rev. Genet. 38:645–679. doi:10.1146/annurev.genet.37.110801.143214.

Palkopoulou E, Mallick S, Skoglund P, Enk J, Rohland N, Li H, Omrak A, Vartanyan S, Poinar H, Götherström A, et al. 2015. Complete Genomes Reveal Signatures of Demographic and Genetic Declines in the Woolly Mammoth. Curr. Biol. 25:1395–1400. doi:10.1016/j.cub.2015.04.007.

Partridge EA, Davey MG, Hornick MA, McGovern PE, Mejaddam AY, Vrecenak JD, Mesas-Burgos C, Olive A, Caskey RC, Weiland TR, et al. 2017. An extra-uterine system to physiologically support the extreme premature lamb. Nat. Commun. 8:15112. doi:10.1038/ncomms15112.

Pougach K, Semenova E, Bogdanova E, Datsenko KA, Djordjevic M, Wanner BL, Severinov K. 2010. Transcription, processing and function of CRISPR cassettes in Escherichia coli. Mol. Microbiol. 77:1367–1379. doi:10.1111/j.1365-2958.2010.07265.x.

Rajewska M, Wegrzyn K, Konieczny I. 2012. AT-rich region and repeated sequences - the essential elements of replication origins of bacterial replicons. FEMS Microbiol. Rev. 36:408–434. doi:10.1111/j.1574-6976.2011.00300.x.

Reinsel D, Gantz J, Rydning J. 2017. Data Age 2025 : Don’t Focus on Big Data; Focus on the Data That’s Big. :1–25.

Ren J, Xu P, Yong H, Zhang L, Liao S. 2017. Ground-to-satellite quantum teleportation. doi:10.1038/nature23675.

Shehabi A, Smith SJ, Sartor DA, Brown RE, Herrlin M, Koomey JG, Masanet ER, Horner N, Azevedo IL, Lintner W. 2016. United States Data Center Energy Usage Report. Lawrence …. doi:LBNL-1005775.

Shipman SL, Nivala J, Macklis JD, Church GM. 2017. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. Nature 547:345–349. doi:10.1038/nature23017.

Shmakov S, Savitskaya E, Semenova E, Logacheva MD, Datsenko KA, Severinov K. 2014. Pervasive generation of oppositely oriented spacers during CRISPR adaptation. Nucleic Acids Res. 42:5907–5916. doi:10.1093/nar/gku226.

Silvera IF, Cole JW. 2010. Metallic hydrogen: The most powerful rocket fuel yet to exist. J. Phys. Conf. Ser. 215. doi:10.1088/1742-6596/215/1/012194.

Smith AJ, Oertle J, Warren D, Prato D. 2016. Chimeric antigen receptor (CAR) T cell therapy for malignant cancers: Summary and perspective. J. Cell. Immunother. 2:59–68. doi:10.1016/j.jocit.2016.08.001.

Solomon M. 2012. Deleting the hns Gene for the H-NS Protein in Escherichia coli Reduces the Transformation Efficiency Following the Heat Shock Transformation Protocol. J. Exp. Microbiol. Immunol. 16:119–122.

Stern A, Keren L, Wurtzel O, Amitai G, Sorek R. 2010. Self-targeting by CRISPR: gene regulation or autoimmunity? Trends Genet. 26:335–340. doi:10.1016/j.tig.2010.05.008.

Swarts DC, Mosterd C, van Passel MWJ, Brouns SJJ. 2012. CRISPR Interference Directs Strand Specific Spacer Acquisition. PLoS One 7:1–7. doi:10.1371/journal.pone.0035888.

Wang J, Li J, Zhao H, Sheng G, Wang M, Yin M, Wang Y. 2015. Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell 163:840–853. doi:10.1016/j.cell.2015.10.008.

Westra ER, van Erp PBG, Künne T, Wong SP, Staals RHJ, Seegers CLC, Bollen S, Jore MM, Semenova E, Severinov K, et al. 2012. CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. Mol. Cell 46:595–605. doi:10.1016/j.neuron.2009.10.017.A.

Westra ER, Semenova E, Datsenko KA, Jackson RN, Wiedenheft B, Severinov K, Brouns SJJ. 2013. Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition. PLoS Genet. 9. doi:10.1371/journal.pgen.1003742.

Willerslev E, Cappellini E, Boomsma W, Nielsen R, Hebsgaard MB, Brand TB, Hofreiter M, Bunce M, Poinar HN, Dahl-Jensen D, et al. 2007. Ancient Biomolecules from Deep Ice Cores Reveal a Forested Southern Greenland. Science (80-.). 317:111–114. doi:10.1126/science.1141758.

Yosef I, Goren MG, Qimron U. 2012. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res. 40:5569–5576. doi:10.1093/nar/gks216.

Yosef I, Shitrit D, Goren MG, Burstein D, Pupko T, Qimron U. 2013. DNA motifs determining the efficiency of adaptation into the Escherichia coli CRISPR array. Proc. Natl. Acad. Sci. 110:14396–14401. doi:10.1073/pnas.1300108110.