Are Restriction Enzymes Recognition Sites Underrepresented in the Organisms That Host Them?

Presented by

Dina Gamaleldin Mansour Aly

In the fulfilment of

Master's degree in Bioinformatics

Supervised by

Björn Canbäck

Coordinator Master's Programme in Bioinformatics

Lund University

2016

Contents

Approval Sheet

Are Restriction Enzymes Recognition Sites Underrepresented in the Organisms that Host them?

Presented by

Dina Gamaleldin Mansour Aly

In the fulfilment of

Master degree of Bioinformatics

1. **Björn Canbäck.** Coordinator of master program of bioinformatics.

2. **Torbjorn Säll**. Professor of Molecular cell biology, and course coordinator of genetic analysis I, II.

3. **Dag Ahrén.** Researcher, PhD, Department of biology, Lund University.

Are RestrictionEnzymes Recognition Sites underrepresented in the Organisms that Host them?

Dina G Mansour Aly

Supervisor: Björn Canbäck.

Department of Biology, Lund University, Lund, Sweden.

Abstract

The restriction modification enzyme system is a vital bacterial defense system against invading phages. Restriction modification system consists of a restriction enzyme and a methyltransferase enzyme that work in a complementary fashion to cut foreign DNA and at the same time methylates and protects host DNA. Recognition site sequence is usually specific for each restriction enzyme. Restriction Recognition Site Representation (RRSR) principally calculates the observed frequency of occurrence of each restriction enzyme recognition site and its expected frequency of occurrence based on the abundance of the four nucleotides in the host's whole genome sequence. A catalogue of restriction enzymes, their properties and annotation was created. The result from RRSR was statistically tested in R using Chi square goodness of fit test and the results showed that many restriction enzymes had their recognition sites underrepresented in their host genome. Adaptive bacterial ada regulon system, a system that protects bacterial genome against exogenous DNA methylation by environmental mutagens interacted with the restriction modification system and may have accounted for the bacterial restriction recognition sites underrepresentation. The dimerization state of the enzyme subunits and their gene expression, as well as the net cellular concentration for methyltransferase enzyme and cofactors; S-adenosyl-methionine, ATP and magnesium ions are vital to determine to achieve a conclusive decision about the efficiency of the restriction modification system in the host. RSRR in phage genome supported the fact that phages are bacterial specific paving a way for using phages as targeted bactericidal agents. RRSR confirmed the co-evolution of phages and bacteria.

Introduction

Restriction Modification System

The bacterial restriction modification system is a self control mechanism against phages first identified by Salvador Luria in 1950's (1). The first known restriction bacterial enzyme was identified in *Escherichia coli* and the restricted phage was lambda. The basic concept of the restriction modification system is that restriction enzymes work in a harmonic fashion with methyltransferases, identifying methylated bacterial DNA while cutting the sugar phosphate backbone of the foreign double helix DNA (Figure 1). Restriction modification system is guided by recognition site sequences. N^6- Adenosyl-methyltransferases methylates the host's DNA by S-adenosyl methionine (SAM) creating a sterically hindered substrate for

the binding of the restriction enzymes or subunits and thus protecting host's DNA from restriction. Restriction enzymes or enzymes' subunits recognize the same restriction recognition sequence that is unmethylated and cuts the DNA.

Nomenclature of the restriction enzymes is derived from the genus of the host bacteria (first initial, upper case letter), the species (second two initials, lower case letters), the bacterial strain (fourth initial, upper case letter) and finally the order of identification (fifth initial, digit, roman digit). For example, EcoRI; is the first restriction enzyme isolated from *Escherichia coli* bacteria strain RY13. Restriction enzymes are sequence specific DNA cutters but the DNA sequence sometimes is not enzyme specific rendering a certain group of restriction enzymes that are ischizomers; restriction enzymes that belong to different or same class of the restriction enzymes and recognize and cleave at the same restriction site. Depending on the type of the restriction modification cleavage mechanism, restriction enzymes are classified according to their cleavage mechanism to five main types.

Type I restriction modification system is the first identified system of restriction enzymes (2,3). The system depends only on one enzyme which consists of three subunits; HsdR is responsible for restriction, HsdM responsible for the methylation and HsdS responsible for specificity of the recognition and cleavage sites. Their dual functional methylation and restriction activity is cofactor dependent; ATP, S-adenosyl-methionine (SAM) and magnesium ions. The cleavage site is spaced from the recognition site by random length spacer; the recognition site is usually asymmetric.

The enzyme's HsdS subunit binds to the specific recognition site sequence which is unmethylated, translocation of the HsdR (Figure S1) over a random length of nucleotide bases on the double strand DNA until it binds to the restriction site (cleavage site) and cleaves it. The HsdM (Figure S2) main function is the methylation of the host DNA using S-adenosyl methionine (SAM).

The efficiency of the restriction modification system I depends on the successful enzyme substrate interaction that is governed by structure activity relationship. The methylation of the host's DNA by the HsdM subunit, first depends on the conformational flipping of the DNA in order for a perfect DNA fit in the SAM pocket and second on the net cellular concentration active methionine S-adenosyl methionine (SAM). The recognition of the specific sequence depends on both the conformation of the HdsS subunit's active catalytic binding site and the spatial orientation of the substrate (DNA). The cleavage of the DNA by HsdR subunit depends on the successful ATP dependent translocation (4) of HsdR subunit and the presence of Mg^{2+} ion in the catalytic active site; successful binding of DNA negatively charged sugar phosphate backbone with the positive catalytic site of HsdR subunit.

Type II restriction modification system is the most abundant system of restriction enzymes; magnesium dependent DNA cutters. This system depends on two separate enzyme groups; methyltransferases and restriction enzymes. These two

enzymes are sometimes fused to make a two enzyme restriction modification system. The recognition cleavage site is the same for both enzymes and is usually palindromic.

The efficiency of restriction modification system II depends on the efficiency of both enzymes, the effective recognition of the host restriction site sequence and the foreign restriction site sequence by the methyltransferase enzyme and restriction enzymes, respectively. The net cellular concentration of the active S-adenosyl methionine is important factor that strongly influence the efficiency of the host's DNA methylation. The recognition and cleavage activity of the restriction enzyme is highly dependent on the Mg^{2+} concentration. This system is not ATP dependent as no translocation of the enzymes is required. There is a number of subtype systems: (i) Type IIS restriction modification system works in the same fashion non-ATP dependent methylation cleavage mode of type II having a multiple protein catalytic domains in the restriction enzyme; a recognition domain and a cleavage domain joint together by small protein chain rendering the recognition site slightly separated from the cleavage site. (ii) Type IIC restriction modification system has one large three domain complex of two enzymes; SAM dependent methylation, recognition and cleavage domains. Their restriction recognition site sequence is separated from cleavage site. (iii) Type IIT restriction modification system has a dual cleavage site enzyme system. This dual cleavage property increases the efficiency of the system against foreign DNA.

Type III restriction modification system represents a very low number of the existing restriction enzymes. It depends on enzymes that have dual functional proteins; ATP dependent restriction protein and SAM dependent methylation protein (5). The RES subunit, restriction subunit recognizes and cleaves two unmethylated non-palindromic sequences asymmetric sequences that are inversely oriented, of 5-8 bps and cleave downstream by 25-27 bps to leave a single strand DNA 5'protrusions. The Mod-subunit (modification subunit), methyltransferase-subunit that methylates the N^6-residue of adenosyl -residues of one strand of DNA. Type III restriction enzymes belong to beta family of the N^6 – Adenosyl-methyltransferases, having multiple motifs that are important for the enzyme catalytic activity, N^6-motif, AdoMet-binding pocket and catalytic motif. Successful enzyme substrate interaction that is governed by structure activity relationship factors that affects the binding of the enzyme system to DNA determines the efficiency of the system. The net cellular concentration of active S-adenosyl methionine (SAM) and ATP are very important for methylation and translocation as well as restriction respectively.

Type IV restriction modification system is a methylation dependent restriction system that recognizes and cleave methylated DNA (6), N6-methylated adenine residues, 5-methylated cytosine residues and 5-hemimethylated cytosine residues. It cleaves away from recognition site.

Other endonucleases are RNA guided nucleases that recognize and cut DNA complementary to their guide RNA (like CRISPR-Cas9) and zinc finger

endonucleases that consist of a DNA binding domain and a restriction subunit usually from Type IIS. These are designed to recognize and cleave large target DNA. They are not covered in the scope of this master's thesis (7).

DNA methylation in bacteria

DNA methylation in bacterial cells consists of two mechanisms, endogenous DNA methylation and exogenous DNA methylation. Endogenous DNA methylation is a vital step of the phage targeted restriction modification machinery that methylates the bacteria DNA to sterically hinder restriction by restriction enzymes. N6-adenosine methyltransferases (Dam), HsdM subunit of restriction modification enzyme and C5-cytosine methyltransferases (Dcm) methylase enzyme of restriction modification system type II, are involved in endogenous DNA methylation (1). Exogenous DNA methylation in bacteria is caused by external mutagens and alkylating agents. Exogenous DNA methylation forms mainly N3-methyladenine and 6O-methylguanine residues in the DNA (8).

Bacterial cells have adaptive response, the Ada regulon system (9,10) that is a group of genes, ada genes (*ada, alkA, alkB, aidB*). The Ada protein is the activator of transcription of the four genes. Ada protein binds to methylated DNA, transfers methyl group to each of its two active sites and then activates the transcription of its own gene *ada* gene and the other three genes. *alkA* gene product is a glycosylase enzyme that catalysis the hydrolysis of methylated bases, N7-methyladenine, N3-methyladenine, O2-methylcytosine from the sugar phosphate backbone. *alkB* gene product is AlkB protein, α-ketoglutarate-Fe(II) dependant protein that by chemical oxidation removes alkyl lesions from DNA. *aidB* gene product (11), AidB protein binds to methylated DNA and degrades DNA endogenous and exogenous methylation.

Restriction proteins that are encoded by *mrr, mrcA, mrcB, mrcC* genes (11-13) are classified as restriction modification system type IV and bind to N6-methyladenine and 5-methylcytosine DNA residues and demethylate DNA.

Restriction enzymes applications

Restriction enzymes have diversity of applications in the field of science. Gene cloning where the gene of interest (from a large genome) and the vector (plasmid) are cut with the same restriction enzyme and followed by ligation and then cloning in a bacterial cell to produce more copies of the required gene or the protein product. Gene cloning and protein product production are valuable techniques that enable the design of gene therapy and replacement therapy for many gene-associated diseases. DNA fingerprinting, Single Nucleotide Polymorphism (SNP) and Restriction Fragment Length Polymorphism (RFLP) detection are techniques that depend on the restriction enzyme specific recognition site sequence. These techniques help to diagnose mutations in certain allelic position without a need of gene sequencing, much more affordable and less time consuming. Restriction site associated DNA mapping (RAD) (14,15), this is useful genotyping tool for associative mapping based on the DNA sequences of restriction sites of restriction

enzymes in the genome of interest; then any mutation in these RAD tags will be in the form of SNPs that can be associated with certain phenotypes. Next Generation Sequencing (Figures S3, S4), restriction enzymes are used to fragment DNA like large chromosomes to get DNA parts of desired length which are able to be used for library construction. Gene Editors; artificial restriction enzymes that are DNA guided, are able to safely and precisely cleave a certain DNA sequence. Anti-viral agents, restriction enzymes act as natural antiviral agents. On the other hand, phages are bacterial specific viruses, bacteriophages that consist of a DNA or RNA enclosed in a protein coat. A bacteriophage needs host bacteria in order to survive. They are bacterial specific because they bind to a specific receptor on the bacterial cell wall. Once stabilized in the cell wall, phages insert their genetic material into bacterial cell. Phages have two cycles, lytic cycle and lysogenic cycle. In the lytic cycle, the phage makes use of the bacterial replication system and bacterial metabolic components in order to replicate phage genetic material and phage protein formation. After a while bacterial cell wall lyses and releases the phages to infect more bacteria. On the other hand, lysogenic phages incorporate their genetic material into the bacterial genome forming a prophage that replicates along with the bacterial genome without bacterial cell lysis. Bactericidal Phage therapy is an application of phages in the treatment of bacterial infections based on the concept the bacteriophages are bacterial specific and bactericidal. The arising bacterial resistance to traditionally used antibacterial and the emergence of life threatening bacterial infections lead to an increased need for a new approach for attacking bacterial infections.

Materials and Methods

Restriction enzymes data

Restriction enzymes data was downloaded from REBASE (16). The included bairoch.txt file was manipulated by a Perl program to output a Restriction Enzyme file with the organism name and restriction enzyme name, restriction enzyme type and restriction enzyme recognition site. Another program was written to retrieve all possible combinations of recognition site sequences in those cases the sequence contained degenerated nucleotides.

Sequence Data

Whole genome sequence data for bacteria and their plasmids were downloaded from National Center for Biotechnology Information (NCBI) (17) using the “wget” linux command. Phages’ sequence data were downloaded from NCBI by wget.

A Perl program was designed to form a catalogue of the restriction enzymes and their host bacteria. The catalogue gives information about the host bacteria name, accession number, taxid (taxonomic identifier) and the restriction enzyme as well as the restriction enzyme type and recognition site.

The main aim of the project is to determine the representation of the restriction enzymes recognition site in the organisms that host them. A Perl program;

"Frequency Calculator" calculates the observed frequency of occurrence of the recognition site based on matching the sequence of the recognition site and the whole genome sequence of the organism. The expected frequency of occurrence of the recognition site was calculated based on the product of the probability of occurrence of each nucleotide base in the sequence multiplied by the whole genome length (bps). The "Frequency Calculator" is designed to read the restriction site recognition sequence from the restriction enzyme organism catalogue and gets its reverse complement for each restriction enzyme recognition site. Then the two strings (recognition site and the reverse complement) are used to calculate the observed frequency of occurrence of the recognition site as well as the expected frequency. Figure 2 illustrates the flow chart of the "Frequency Calculator".

The Perl program outputs two files for each fasta sequence file, first general statistics file (Table S1) that gives the information about the name of the host organism, the whole genome length in base pairs, the probability of occurrence of each of the four nucleotide bases and a complete list of the restriction enzymes, restriction enzyme type and their recognition site sequence and second frequency statistical file (Table S2) that gives information about the host restriction enzymes, recognition site sequence, observed frequency of occurrence and expected frequency of occurrence. The Perl program is run on all the whole genome fasta sequence files in the directory and produces two statistical files for each genome. The frequency calculator program was used to calculate the RRSR in plasmids and phages genomes in the same way as bacterial genomes.

Scripts are available at 130.235.46.13/~dina/Programs.

Statistical analysis of bacteria sequence data

The frequency statistical files for all bacterial genomes are saved in one file. The statistical analysis was performed in R studio (R version 3.2.3) (18) (Figure S5). A data frame is constructed from the observed frequency values and expected frequency values for each restriction enzyme in the frequency statistical file. Chi-square test of goodness of fit was applied to the data frame using "lapply" command. The cut off probability value of significance was set to 0.0001. Bonferroni correction for multiple comparisons gave a cut off false discovery rate of approximately 1.20e-05. The frequency difference between observed frequency of occurrence and the expected frequency of occurrence of each restriction enzyme in the host bacteria was calculated in R. The relative representation was calculated by the ratio of the observed frequency and expected frequency values of each restriction enzyme in host bacteria. Based on both the frequency difference and the relative representation of the restriction enzymes recognition sites in their host bacteria, the restriction enzymes recognition sites were subset into two files; underrepresented and overrepresented. The frequency statistical files for plasmids and phages were statistically analyzed in R in the same manner as the bacterial statistical files. R Scripts are available at 130.235.46.13/~dina/Programs.

DNA methylation genes analysis

The gff files of *Escherichia coli* genome, *Staphylococcus aureus* genome, *Haemophilus influenza* genome, *Salmonella enteric* genome, bacteria with underrepresented restriction sites were downloaded from NCBI genome directory. The gff file provides complete information about the genes in the bacterial genome and their gene products. A complete list of the genes that code for DNA methyltransferases and restriction endonucleases was extracted from the gff file using linux commands. Sequence fasta file for each gene was downloaded from NCBI gene database. A Perl program was designed to report the presence of the genes in the whole sequence fasta genome file of each bacterium.

Cross species RRSR analysis

The whole genome sequence of *Escherichia coli K-12 MG1655, Staphylococcus aureus NCTC 8325, Salmonella enteric subsp. enterica server Typhimurium str. LT2, Haemophilus influenza Rd KW20, Mycobacterium tuberculosis H37RV, Neisseria gonorrhoeae FA 1090, Pseudomonas aeroginosa PAO1, Acinetobacter baumannii ATCC 17978* were downloaded from the genome database of NCBI. The frequency calculator was modified to determine the representation of restriction enzymes recognition sites of one bacterial species in other bacterial species, the representation of *Esherichia coli* restriction enzymes in *Staphylococcus aureus.* This analysis was done for all bacterial species in the restriction enzyme catalogue I.

Database and web programming

A Restriction Enzyme Database is generated which can be accessed through a web interface. A database (Figure 3) with the relevant data is created in SQL (SQLite version 3.8.2). The “bacteria” table is created using restriction enzyme catalogue as an import text file. The relative representation is found in the table “Bacterial RRSR representation table” was created using the “RE_representation.txt” table as import text file. Genes and the diseases they are associated with were downloaded from the GeneCards (19) main webpage. This data was used to create two tables “Allgenes.txt” and “Phages table”. The web interaface (Figure 4) is programmed in HTML using a cascading styling sheet CSS and a common gateway interface (CGI) python program.

Results

Restriction Enzyme Catalogue I

The Restriction Enzyme Catalogue I is formed of only bacteria with whole genome sequence and their restriction enzymes. Tables 1, 2 and 3 demonstrate Restriction Enzyme Catalogue I, a list of bacteria that host the restriction enzymes that are members of the restriction modification system type I, type II and type III respectively. The catalogue provides information about the host bacteria, name,

accession number in the NCBI database and taxid as well as information about the restriction enzyme the bacteria hosts, name, restriction modification system type and the nucleotide sequence of the restriction recognition site.

Restriction Recognition Site Representation (RRSR) in Bacterial Genome

RRSR of restriction modification system type I restriction enzymes in bacterial genomes

Restriction modification system type I restriction enzymes consisted only of 2.94 % (Table 8) of the 680 restriction enzymes studied. Restriction enzymes SauBMKI* *(Staphylococcus aureus)**, Hpy87AI *(Helicobacter pylori),* NgoAV *(Neisseria gonorrhoeae),* LlaG2I *(Lactococcus lactis),* EcoBI *(Escherichia coli B strain),* SpnD39IIIC *(Streptococcus pneumonia D strain),* EcoKI *(Escherichia coli strain K-12 MG1655),* EcoprrI *(Escherichia coli strain K-12 DH108)* and HindI *(Haemophilus influenza),* are restriction enzymes that have their restriction sites spaced from their recognition sites by number of nucleotides ranging from one nucleotide up to nine nucleotides. The former enzymes were the top five enzymes having a relative representation that range from 0 to 0.1 of their restriction recognition sites in their host bacteria (Table 4). *(*Restriction enzyme name **Latin name of host bacteria).*

RRSR of restriction modification system type II restriction enzymes in bacterial genomes

Restriction modification system type II restriction enzymes contributed to about 96% (Table 8) of the 680 restriction enzymes studied. Over than 60% (Table 8) of restriction modification system type II restriction enzymes had their restriction site underrepresented in their host bacteria. NgoAVII *(Neisseria gonorrhoeae),* SgrI *(Streptococcus griseus),* Hpy99II *((Helicobacter pylori),* DsaVI *(Dactylococcopsis salina),* PluTI (*Photorhabdus luminescens*), AbaI (Arthrobacter aurescens), CauIII (*Chloroflexus aurantiacus*), BbrAI (*Bordetella pertussis*), HindIII (*Haemophilus influenza)* and NgoCI *(Neisseria gonorrhoeae)* are restriction enzymes that have their restriction site is the same as the restriction site and are of short palindromic nucleotide sequences 4-6 bases. These top ten restriction enzymes of restriction modification system type II had significance relative representation range of 0 to 0.5 (Table 5). Less than 40% (Table 8) of restriction modification system type II restriction enzymes had their restriction site overrepresented in their host bacteria. BcrI (*Bacillus cersus*), SonI (*Shewanella oneidensis*), BhaII (*Bacillus halodurans*), CgII (*Corynebacterium glutamicum*), BfrAI (*Bacteroides fragilis*), EcoCKI *(Escherishia coli CFT073),* TdeII (*Treponema denticola*), BanAI (*Bacillus anthracis*) and Pae2KI (*Pseudomonas aeruginosa*) are restriction enzymes with palindromic recognition sites and showed significant relative representation range 1.5 to 4.5 (Table 6).

RRSR of restriction modification system type III restriction enzymes in bacterial genomes

Restriction modification system type III represents only 1% (Table 8) of the hosting bacterial restriction modification systems. RRSR had significant overrepresentation of type III. Table 7 has a summary of the RRSR of type III.

DNA methylation genes analysis in bacteria

Escherichia coli genome has *hsdM, mrr, mrcA, mrcB, mrcC, ralR, alkB, aidB* genes intact in its genome. *hsdM* gene that encodes the HsdM subunit in restriction modification system type I (20) is species-specific, the hsdM gene sequence of *Escherichia coli* was not found in other species that had *hsdM* gene of restriction modification system I, *Staphylococcus aureus*, *Salmonella enteric. mrr* gene encodes N6-methyladenine and C5-methylcytosine nuclease (restriction modification type IV) is also species-specific, the nucleotide sequence of the *mrr* gene differ from one bacteria to another. mrr gene is also present in *Klebsiella pneumonia, Salmonella enteric, lactobacillus plantanum. mcrA, mcrB, mrcC* genes, encodes C5-methylcytosine gene restriction nuclease subunits abundant in phage genomes, *Klebsiella* phage, *Escherichia coli* phage, *Streptococcus* phage. *mcrA* gene is present in some bacterial species *Shewanella oneidensis, Lactobacillus salivaris. mcrB* gene is present in *Esherichia coli K-12, Xanthomonas campestris, Deinococcus radiodurans, Salmonella enteric.* ralR prophage gene codes for an antirestriction protein that alleviates the demethylation and favors restriction modification especially type I. *ada* gene, ada regulon transcription activator gene that is an important gene in bacterial adaptive system against exogenous alkylating agents. *ada* gene encodes Ada protein that activates the *alkA* gene, *alkB* gene *and aidB* gene. The later genes are species specific. alkB gene is present in *Escherichia coli, Salmonella enteric, Pseudomonas syringae, Bordetella pertussis, Shigella flexneri and Shigella dysenteriae*. *aidB* gene codes for AidB, DNA demethylating protein is intiated by ada-dependant mechanism and ada-independent mechanism as anaerobosis and pH changes. AidB is present in *Shigella dysentrie, Salmonella enteric and Yersisi pestis.*

Overall relative representation of restriction modification systems

RRSR analysis in bacteria showed that the majority of the bacterial restriction modification systems are of type II. Most of the hosting bacteria for restriction modification system type II had significant underrepresentation of the restriction recognition sites. Table 8 represents the proportions of the restriction modification system in bacteria. Figure 5 is a bar chart illustrating the restriction modification system abundance in hosting bacteria.

Cross species RRSR analysis in bacteria

The cross species RRSR analysis showed varying representation of recognition sites of non-host restriction enzymes. Table 9 summarizes the top restriction enzymes that have their recognition sites underrepresented in non-host bacteria.

AjoI (*Acintobacter johnsonii)* recognition sites, were underrepresented in *Mycobacterium tuberculosis.* M.EacI (*Enterobacter cloacae)* recognition sites, were underrepresented in *Pseudomonas aeroginosa*. Eco71KI (*Escherichia coli*) recognition sites were underrepresented in *Haemophilus influenza* but overrepresented in *Staphylococcus aureus, Mycobacterium tuberculosis, Pseudomonas aeriogenosa* and *Salmonella enterica* (Table 9a).

Restriction Recognition Site Representation (RRSR) in Plasmids

2029 plasmid genomes were screened for the representation of restriction recognition sites. SecII *(Salmonella enteric),* RtrI *(Rhizobium tropic),* SexI *(Salmonella enteric),* Psp29I *(Pseudomonas syringae),* SfaAI *(Shigella flexneri),* LmnI *(Leisingera methylohalidivorans),* NgoJI *(Neisseria gonorrhoeae),* NmeBI (*Nitrosomonas eutropha*), Kpn121 (*Klebsiella pneumoniae*) and Mli40I (*Mesorhizobium loti*) are restriction enzymes of different restriction modification systems that were significantly underrepresented in the plasmid genomes.

NtaI *(Natronococcus occultus),* Chil *(Corynebacterium halotolerans),* AbaI *(Acinetobacter baumannii),* CmaLM2II *(Chamaesiphon minutes),* BtsCI *(Bacillus toyonesis)* are restriction enzymes of different restriction modification systems that were significantly overrepresented in the plasmid genomes (Table 10).

Restriction Recognition Site Representation (RRSR) in Phages

Phages are bacterial viruses that attack the bacteria and kill them. It is only for those bacterial enemies that the restriction modification system is in the host bacteria. Phages that had no relative representation of a specific recognition site of specific restriction enzyme were selected by Perl program that searches the Restriction Enzyme Catalogue I for the susceptible bacteria that has the recognition site for the restriction enzyme that is not represented in the phage. The Perl program outputs "Phage_Therapy.txt".

The overall relative representation of the restriction recognition sites demonstrated a significant relative underrepresentation of bacterial restriction recognition sites in the phage's genome specific to the host bacteria. Hpy99IV* (*Helicobacter pylori J99*)***, NgoAIII (*Neisseria gonorrheae FA1090*), NlaII (*Neisseria lactamica 020-06*), Pdi8503III (*Parabacteroides distasonis ATCC8503*), RpaI (*Rhodopseudomonas palustris CGA009*) were the top of the relative underrepresented restriction enzymes in phages (Table 11). (* restriction enzyme name, *** phage name). Table 12 gives some examples of phage therapy.

Discussion

Restriction modification systems function in different ways to protect host's DNA and cut foreign DNA. Depending on the nature of the restriction modification enzyme, single trimer as EcoKI enzyme in restriction modification system type I, separate enzymes that work collaboratively as EcoRI and M.EcoRI that belong to restriction modification system type II, the relative representation of the restriction recognition site can be interpreted. Each bacterium has one or more restriction

modification systems. Restriction modification system type I and type II restriction recognition sites were significantly underrepresented in their host genome. The significance relative underrepresentation of restriction modification system type II recognition sites was reported in about 60% of the total restriction enzymes that belong to this group of restriction enzymes. An understanding of the DNA methylation of the host bacterial genome offers an explanation for the underrepresentation. DNA methylation in bacteria has two mechanisms. Endogenous DNA methylation occurs when the HsdM subunit of restriction modification system type I and methyltransferases of restriction modification system type II methylates the DNA N6-adenine or 5C-cytosine residues, respectively. Exogenous DNA methylation in bacteria takes place when DNA by alkylating agents (environmental mutagens) methylates the N3-adenine and O6-guanine of the bacterial DNA. Adaptive DNA demethylation mechanisms, ada regulon a bacterial DNA repair system that responds non-selectively to DNA methylation either endogenous or exogenous and demethylates the bacterial DNA. Ada regulon (11) DNA repair system has four main proteins, Ada protein (transcription initiator of the Ada regulon system), AlkA protein (glycosylase enzyme that repairs DNA by hydrolysis of the glycosidic bond of the sugar phosphate backbone of the methylated base), AlkB protein (repairs DNA by chemical oxidation of the methylated base and AidB protein (21) (a restriction enzyme that binds to methylated DNA and cuts it). Another category of enzymes present in some bacteria are restriction enzymes that recognize and cuts methylated DNA, these are classified as restriction modification system type IV and are coded by mrr, *mcrA, mcrB, mcrC* genes in some bacteria species like, *Escherichia coli*. The bacteria harbour two systems, the restriction modification system and the DNA repair system, therefore the interaction between these two systems accounts for the underrepresentation of the recognition sites in the host bacterial genome (22). Underrepresentation of *Eco*KI and *Eco*RI recognition sites in their host bacteria genome can be attributed to the DNA demethylation of the recognition sites by the N6-methyladenine restriction endonucleases and C5-methylcytosine restriction endonucleases of restriction modification type IV and Ada regulon genes, *alkB* and *aidB* genes. *mcrA, mcrB* and *mcrC* genes encoding C5-methylcytosine binding nucleases are abundant in phage genome are phage antibacterial tools against bacterial modification system. The presence of these genes in some bacteria highlights the fact that the lysogenic bacteriophage has phage genome embedded in bacteria genome and proves that the phage genome co-evolve with the bacterial genome.

Other factors that can be considered to affect DNA methylation, are HsdM and HsdS functionality and the gene expression of the methyltransferases enzymes and methylase subunits. The HsdM functionality is affected by the proper conformation of the S-adenosyl methionine (SAM) pocket in its catalytic active site and the net cellular AdoMet (SAM) (8). Unsuccessful binding of the HsdS subunit for recognition and HsdM subunit cause significant hypomethylation of the host DNA. The gene expression and the ribosomal function in the host's cell will greatly affect the net cellular concentration of restriction enzyme subunits. The complexity of the

recognition site sequence can have a profound effect on the relative representation of the host restriction recognition sites. Since, the restriction sites of restriction modification system type II are short and palindromic with low complexity then the abundance of the four nucleotides will affect the relative representation of the host's restriction recognition sites. The more complex the sequence, the more likely will be underrepresented the host restriction recognition sites. The relative cellular concentration of methyltransferase enzymes and restriction enzymes is a very important additional parameter to effectively evaluate restriction modification type II system. Extracellular intracellular environmental interaction; the temperature and hydrogen ion concentration can greatly affect the conformation of the enzyme catalytic site (methyltransferase active site and restriction active site) and the DNA substrate conformation. The net cellular concentration of both SAM and Mg^{2+} ion play an important role in methylation and restriction respectively; decrease in SAM concentration will decrease methylation leading to diminished relative representation of the host's restriction recognition sites.

Cross species RRSR analysis, revealed that many restriction enzymes had their recognition sites underrepresented also in non-host genomes. AjoI (*Acintobacter johnsonii)* recognition site, was underrepresented in *Mycobacterium tuberculosis.* M.EacI (*Enterobacter cloacae)* recognition site, were underrepresented in *Pseudomonas aeroginosa*. The restriction recognition sites of Eco71KI were underrepresented in the non-host *Haemophilus influenza* genome and overrepresented in *Salmonella enteric, Pseudomonas aeriogenosa, staphylococcus aureus* and *Neisseria gonorrhoea.* The underrepresentation of the recognition sites can be explained by the fact that many bacterial species coexist within the same environment and that the restriction enzymes are able to diffuse out of one species bacterial cell and enter other species bacterial cell and cut the recognition sites in the non-host genome.

Most of the restriction recognition sites had significant low relative representation in the plasmids of the hosting bacteria. Plasmids are circular DNA in the bacteria that is separated from the chromosomal DNA. Plasmids can replicate independently. They usually carry genes that codes for proteins that benefit the bacterial survival; for example, antibiotic resistance. Underrepresentation of restriction enzyme recognition sites in the plasmids can have two possible explanations. First, the restriction-modification system doesn't work on the plasmid DNA due to the conformational structure as circular form that hinders the access of the restriction modification subunits. Second, the plasmid doesn't harbour the recognition site for the enzyme because of the important features that the plasmid gives to the host bacteria.

Phages are bacterial specific, bacteria have specific phages that attacks their genomes. Many bacteria have more than one modification system, *Escherichia coli* has EcoKI (restriction modification type I) in combination of EcoRI (restriction modification type II). This multiple modification system acts a perfect protection against foreign DNA phages. The significant relative underrepresentation of the restriction recognition sites in phages suggests the presence of a evolutionary

relationship between the sequence of the bacterial restriction modification system and the phage genome composition. That is as the bacteria develop more protection by upgrading its restriction modification system, phages also change their DNA composition (mutation and adaptation) in order to maintain their pathogenic power against bacteria. This coevolution provides a reasonable explanation for the significant relative underrepresentation or null representation of the restriction recognition site in the phage genome. This coevolving feature of phages enabled their use as phage therapy in the treatment of life threatening disease caused by bacterial strains that are highly resistant to traditional antimicrobial therapy.

Restriction Enzyme Database Web Interface

Restriction enzyme database was created in SQL (Figure 3). The database can be accessed by web interface. The Restriction Enzyme Catalogue I can be accessed by typing the genus of the bacteria of interest. Restriction recognition site representation for host bacteria of interest can be accesses by the host name which outputs a full list of the restriction enzymes and their relative abundance along with the restriction enzyme type and the recognition site sequence. An access for UniPRotKB (23); protein database to get information about the restriction enzyme classification, structure, function, protein sequence and a complete list of accession number to other databases. For 3D structure view of the restriction enzyme subunits, the catalytic site and the ligands, a link to RSCB PDB (23) workshop can be searched by the PDB id.

For gene cloning, the user can select a disease of interest from a list of diseases and an output table with a list of genes including the gene name and its position and chromosome number. Gene symbols are links to the National Center of Biotechnology Institute (NCBI) gene database to facilitate the access to more information about the gene; gene FASTA sequence, gene products, gene features, publications concerning the gene and disease of interest, gene orthologues, gene transcripts, as well as other possible tools of sequence alignment BLAST. For a detailed mechanism of cloning of a gene, steps and types of gene cloning a link to the textbook “Introduction to Genetic Analysis 7th edition” (24); textbook from the NCBI bookshelf and a summary to gene cloning link. Restriction enzymes play important roles in molecular cloning and protein production.

In order to get information about the organism that host a specific restriction site the user can enter the restriction site of interest in a user-input that will access the Restriction Enzyme Catalogue I and outputs hosting bacteria name, accession number and taxid (taxonomic identifier) as well as the restriction enzyme name and its type. Plasmids and the restriction enzyme recognition sites can be accessed by the restriction site of interest as input and this will execute a table that has a full list of the plasmids that have recognition site and the restriction enzymes targeting the site.

Regarding phage applications, the user can select a pathogenic bacteria of interest that will access the phage therapy table and will give out the pathogenic bacteria list and the suitable bactericidal phage. Advantages and disadvantages for phage

therapy are highlighted in the article linked to the phage therapy (25). Phage therapy research center in Georgia can be accessed by the user to learn more about phage therapy techniques, acute and chronic bacterial infections and on-going research in the field of phage therapy.

Restriction Enzyme Database web interface is user friendly and offers a diversity of applications of restriction enzymes. It is an ideal tool for molecular biology researcher, medical researchers and pharmaceutical researchers.

Conclusion

Restriction Recognition Sites Representation (RRSR) revealed that about 3% of the restriction enzymes included in the study are classified as restriction modification system type I and 96% of the restriction enzymes are classified as restriction modification system type II and 1% of the of restriction enzymes are classified as restriction modification system type III. Overall relative representation of restriction enzymes recognition sites in their host bacteria proved that more than 60% of the restriction enzymes had their recognition sites significantly underrepresented in their host genomes and less than 40% of the restriction enzymes had their recognition sites overrepresented in their host genome. The significance level is set to 0.0001 due to the high false negative on the adjustment of the p-value using Bonferroni correction for multiple testing.

Oligomeric restriction enzymes for example, SauBMKI* *(Staphylococcus aureus)***, Hpy87AI *(Helicobacter pylori),* NgoAV *(Neisseria gonorrhoeae),* LlaG2I *(Lactococcus lactis),* EcoBI *(Escherichia coli B strain),* SpnD39IIIC *(Streptococcus pneumonia D strain),* EcoKI *(Escherichia coli strain K-12 MG1655),* EcoprrI *(Escherichia coli strain K-12 DH108)* and HindI *(Haemophilus influenza),* their restriction modification activity depend on single enzyme composed of three subunits, HsdS, specificity subunit that recognizes the recognition site sequence, HsdM, the methylase subunit that has a S-adenosyl methionine (SAM) pocket to methylate host DNA and HsdR, restriction subunit that is magnesium ion-ATP-dependant and is transloacted few bases away from the recognition site and cleaves the DNA. Significant relative underrepresentation of these restriction modification type I enzymes in their host genome can be supported by the fact that the efficiency of the system is dependent on single complex three-subunits enzyme which is highly influenced by the gene-expression of the enzyme subunits that directly affects the net cellular concentration of the functional enzymes, the complexity of the restriction-recognition mechanism that depends mainly on successful methylation of the host's DNA through the effective binding with the SAM-HsdM subunit. Any mutations or changes in the HsdM active site will cause hypomethylation of the host's DNA.

Over than 60% (Table 8) of restriction modification system type II restriction enzymes had their restriction site underrepresented in their host bacteria. NgoAVII *(Neisseria gonorrhoeae),* SgrI *(Streptococcus griseus),* Hpy99II *((Helicobacter pylori),* DsaVI *(Dactylococcopsis salina),* PluTI (*Photorhabdus luminescens*), AbaI

(Arthrobacter aurescens), CauIII (*Chloroflexus aurantiacus*), BbrAI (*Bordetella pertussis*), HindIII (*Haemophilus influenza)* and NgoCI *(Neisseria gonorrhoeae)* had their recognition sites underrepresented in their host genome.

Underrepresentation of restriction recognition sites of *Eco*KI and *Eco*RI is attributed to the interaction between restriction modification system and DNA repair system ada regulon in *Escherichia coli* (28). Restriction modification system type IV genes *mrr, mcrA, mcrB, mcrC* codes for DNA demethylating proteins that interact with the restriction modification type I, II, III DNA methylation leading to underrepresentation of recognition sites in the host genome. *mcrA, mcrB, mcrC genes* are of phage origin and explains how phages survive the bacterial restriction modification system. Other factors that can affect the efficiency of the restriction modification system are the intracellular balance between the two enzyme groups is strongly controlled by the relative gene expression of the genes encoding these enzymes, cofactors; SAM, Mg^{2+} affected the successful methylation and restriction of the restriction modification system type II system respectively.

Cross species RRSR analysis among different bacterial species showed varying representation of the restriction modification sites that can be explained by environmental coexistence of bacterial species and the abundance of restriction enzymes in the environment.

RRSR in plasmids offered a prompt and precise tool for plasmid selection in gene cloning. The restriction recognition site specificity of restriction modification type II restriction enzymes (R2 sites) enabled the use of plasmids that have R2 sites as vectors for gene cloning in genetic engineering.

RRSR in phages showed significant underrepresentation of restriction recognition sites. Hpy99IV* (*Helicobacter pylori J99*)***, NgoAIII (*Neisseria gonorrheae FA1090*), NlaII (*Neisseria lactamica 020-06*), Pdi8503III (*Parabacteroides distasonis ATCC8503*), RpaI (*Rhodopseudomonas palustris CGA009*) were the top of the relative underrepresented restriction enzymes in phages (Table 10). Phages that had no relative abundance of restriction recognition sites are good bactericidal agents (Table 11). This gave the opportunity to focus on the beneficial applications of phages in control of pathogenic infections. (* restriction enzyme name, *** phage name).

Acknowledgements

Bioinformatics is a very powerful field that enables the manipulation, analysis and interpretation of sequence raw data. Designing of software programs and utilization of available bioinformatics tools are important requirements for successful application of bioinformatics (Fig S4-S7). Master program in Bioinformatics at Lund University is conducted by a professional team of respectable and helpful Professors. My hearted thanks for the following professors:

Björn Canbäck: Coodinator of Master Bioinformatics Program.

Lotta Persmark: Study advisor of Master Bioinformatics Program.

Claes Von Wachenfeldt: Senior Lecturer and course conductor of Bioinformatics and sequence analysis.

Mattias Ohlsson: Senior Lecturer and course conductor of Introduction to programming.

Christof Winter: Postdoc in Translational Medicine at Medicon Village.

Jessica Abott: Researcher (PhD) and course coordinator of Bioinformatics: Processing and Analysis.

Per-Erik Isberg (Dept. of Statistics), and lecturer of Bioinformatics: Processing and Analysis.

Torbjorn Säll: Professor of Molecular cell biology, and course coordinator of genetic analysis I, II.

A warm hope for my colleagues for a successful future***; Amelie Barozet, Gad Hatem, Koithan Thalea, Maria Marin, Maryem Salim, Robert Hafþórsson, Stephen Burleigh and Jelena čalyševa.***

References

1. Loenen WAM. Dryden DTF, Raleigh EA, Wilson GG, Murray NE. Highlights of the DNA cutters: a short history of the restriction enzymes. Nucleic Acids Research. 2014; 42:3-19. doi: 10.1o93/nar/gkt990.
2. Loenen WA, Dryden DT, Raleigh EA, Wilson GG. Type I restriction enzymes and their relatives. Nucleic Acids Res. 2014; 42. doi: 10.1093/nar/gkt847
3. Smith RM, Marshall JJ, Jacklin AJ, Retter SE, Halford SE, Sobott F. Organization of the BcgI restriction-modification protein for the cleavage of eight phosphodiester bonds in DNA. Nucleic acids research. 2013; 41, 391-404.
4. Lyumkis D, Talley H, Stewar A, Shah S, Park CK, Tama F, Potter CS, Carragher B. Allosteric Regulation of DNA Cleavage and Sequence-Specificity through Run-On Oligomerization. Structure. N.C. 2013; 21, 1848-1858. doi:10.1016/j.str.2013.08.012.
5. Bollins JTJ, Szczelkun MD. Re-evaluating the kinetics of ATP hydrolysis during initiation of DNA sliding by Type III restriction enzymes. Nucleic Acids Research. 2015; 43: 10870-10881. doi: 10.1093/nar/gkv1154.

6. Xu S, Corvaglia AR, Chan SH, Zheng Y, Linder P. A type IV modification dependant restriction enzyme SauUSI from Staphylococcus aureus subsp. Aureus USA300. Nucleic Acid Research. 2011; 39: 5597-5610. doi:10.1093/nar/gkr098
7. Chandrasegaran S, Carroll D. Origins of programmable nucleases for genome engineering. J Mol Biol. 2016: 428; 963-989. doi:10.10016/j.jmb.2015.10.014.
8. Casadesus J, Low D. Epigenetic gene regulation in bacterial world. MMBR, 2006; 70: 830856. doi:10.1128/MMBR.00016-06.
9. Teo I, Sedgwick B, Demple B, Li B, Lindahl T. Induction of resistance to alkylating agents in E. coli: the ada+ gene product serves both as a regulatory protein and as an enzyme for repair of mutagenic damage. The EMBO Journal. 1984. 3:2151-2157.
10. Hakura A, Morimoto K, Sofuni T, Nohmi T. Cloning and characterization of the Salmonella typhimurium ada gene, which encodes O6-methylguanine-DNA methyltransferase. Journal of Bacteriology. 1991.173:3663-3672.
11. Mielecki D, Grzesiuk E. Ada response – a strategy for repair of alkylated DNA in bacteria. *FEMS Microbiology Letters*. 2014. 355:1-11. doi:10.1111/1574-6968.12462.
12. Hamill MJ, Jost M, Wong C, Elliott SJ, Drennan CL. Flavin-Induced Oligomerization in *Escherichia coli* Adaptive Response Protein AidB. Biochemistry. 2011. 50:10159-10169. doi:10.1021/bi201340t.
13. Kennaway CK, Obarska-Kosinska A, White JH, Tuszynska I, Cooper LP, Bujnicki JM, Trinick J, Dryden DTF. The Structure of M.Ecoki Type I DNA Methyltransferase with a DNA Mimic Antirestriction Protein. Nucleic Acids Res. 2009; 37: 762. doi:10.1093/nar/gkn988
14. Bonatelli IAS, Carstens BC, Moraes EM. Using Next Generation Sequencing RAD sequencing to isolate multispecies microsatellites for pilosocereus (Cactaceae). PloS ONE. 2015; 10:114. doi: 10.1371/journal.pone.0142602.
15. Wang S, Lv J, Zhang L, Dou J, Sun Y, Li X, et al. Methyl RAD: a simple and scalable method for genome-wide DNA methylation profiling using methylation-dependant restriction enzymes. Open Biol. 2015; 5: 150130. dio:10.1098/rsob.150130.
16. Roberts, Richard J, Vincze T, Posfai J, Macelis D. REBASE-a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 2010; 38: D234–6. doi:10.1093/nar/gkp874.
17. Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I. RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Res. 2014 ;42:D553-559 (PubMed)
18. R Development Core Team 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
19. Rappaport N, Twik M, Nativ N, Stelzer G, Bahir I, Stein T I, Safran M, Lancet D. MalaCards: A Comprehensive Automatically-Mined Database of Human Diseases. Curr. Protoc. Bioinform. 47:1.24:1.24.11.24.19. (URL: www.genecards.org).
20. Fang G, Munera D, Friedman DI, Mandlik A, Chao MC, Baerjee O, et al. Genome-wide mapping of methylated adenine residues in pathogenic

Escherichia coli using single-molecule real-time sequencing. Nature Biotechnology. 2012; 30, 1232-1239. doi: 10.1038/nbt.2432.
21. Landini P, Hajec LI, Volkert MR. Structure and transcriptional regulation of the Escherichia coli adaptive response gene aidB. Journal of Bacteriology. 1994. 176(21):6583-6589.
22. Tesfazgi Mebrhatu M, Wywial E, Ghosh A, et al. Evidence for an evolutionary antagonism between Mrr and Type III modification systems. Nucleic Acids Research. 2011. 39:5991-6001. doi:10.1093/nar/gkr219.
23. The UniProt Consortium**.** UniProt: a hub for protein information. Nucleic Acids Res. 2015; 43: D204-D212.
24. Griffiths AJF, Miller JH, Suzuki DT.An Introduction to Genetic Analysis. 7th edition.New York: W. H. Freeman; 2000.
25. Carrillo CL, AbedonST.Pros and cons of phage therapy. Bacteriophage, 2011; 111–114.doi: 10.4161/bact.1.2.14590
26. Morozova O, Marco A. Marra. Applications of next-generation sequencing technologies in functional genomics. Genomics 2008; 92: 255-264. doi:10.1016/j.ygeno.2008.07.001.
27. Shao k, Ding W, Wang F, Li H, Ma D, Wang H. Emulsion PCR: A High Efficient Way of PCR Amplification of Random DNA Libraries in Aptamer Selection. PLoS One. 2011; 6: e24910.doi: 10.1371/journal.pone.0024910.
28. Shendure J, Porreca GJ, Reppas NB, Lin X, Pe McCutcheon J, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. Science. 2005; 309: 1728–1732. doi:10.1126/science.1117389.
29. García LR, Molineux IJ. Translocation and specific cleavage of bacteriophage T7 DNA *in vivo* by *Eco*KI. Proc Natl Acad Sci U S A. 1999; 96: 12430–12435.
30. Doughty B, Kazer SW, Eisenthal KB Binding and cleavage of DNA with the restriction enzyme EcoR1 using time-resolved second harmonic generation. PNAS. 2011;108: 19979-19984. doi : 10.1073/pnas.1115498108.

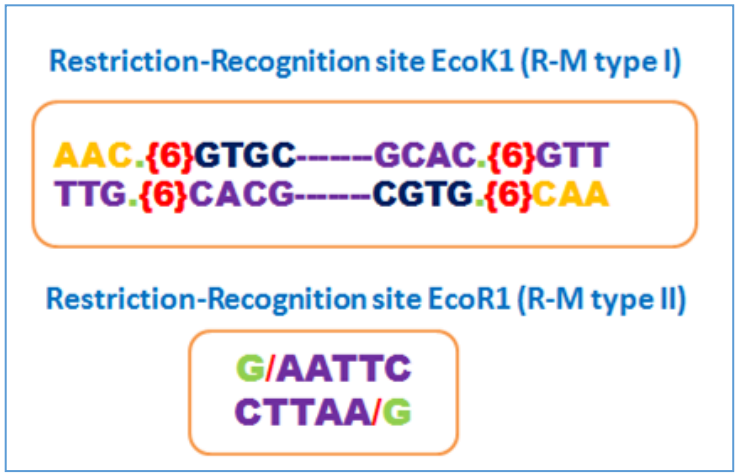


Figure 1. Restriction Modification System. The restriction enzyme EcoK1 (type I) restriction site (GTGC) is spaced from recognition site (AAC) by six nucleotides, restriction enzyme EcoR1 (type II) recognition site is the same as the restriction site and is palindromic. (GAATTC); (20,21).

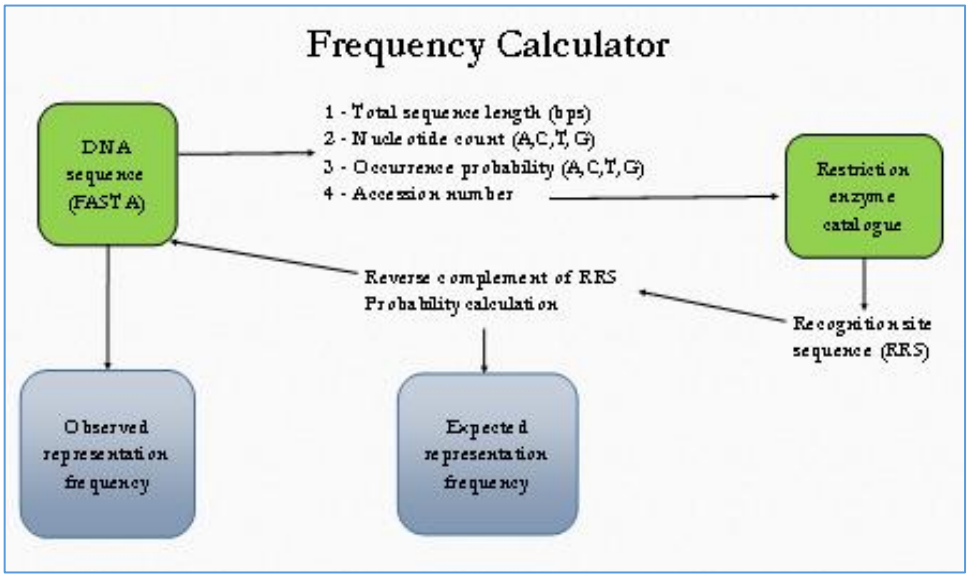


Figure 2. The flow chart of Frequency Calculator Program. Step 1: The frequency calculator calculates total sequence length in base pairs, the probability of each nucleotide for the DNA sequence (input file 1). The accession number of the genome is used to select the host restriction enzymes and their sequences from the Restriction Enzyme Catalogue (input file 2). Step 2: The frequency calculator forms the reverse complement of the restriction-recognition site. Step 3: The frequency calculator calculates the observed frequency of the restriction-recognition site and its reverse complement. Step 4: The frequency calculator calculates the expected frequency of the restriction-recognition site and its reverse complement based on the probability of the nucleotides present in the sequence and its reverse complement.

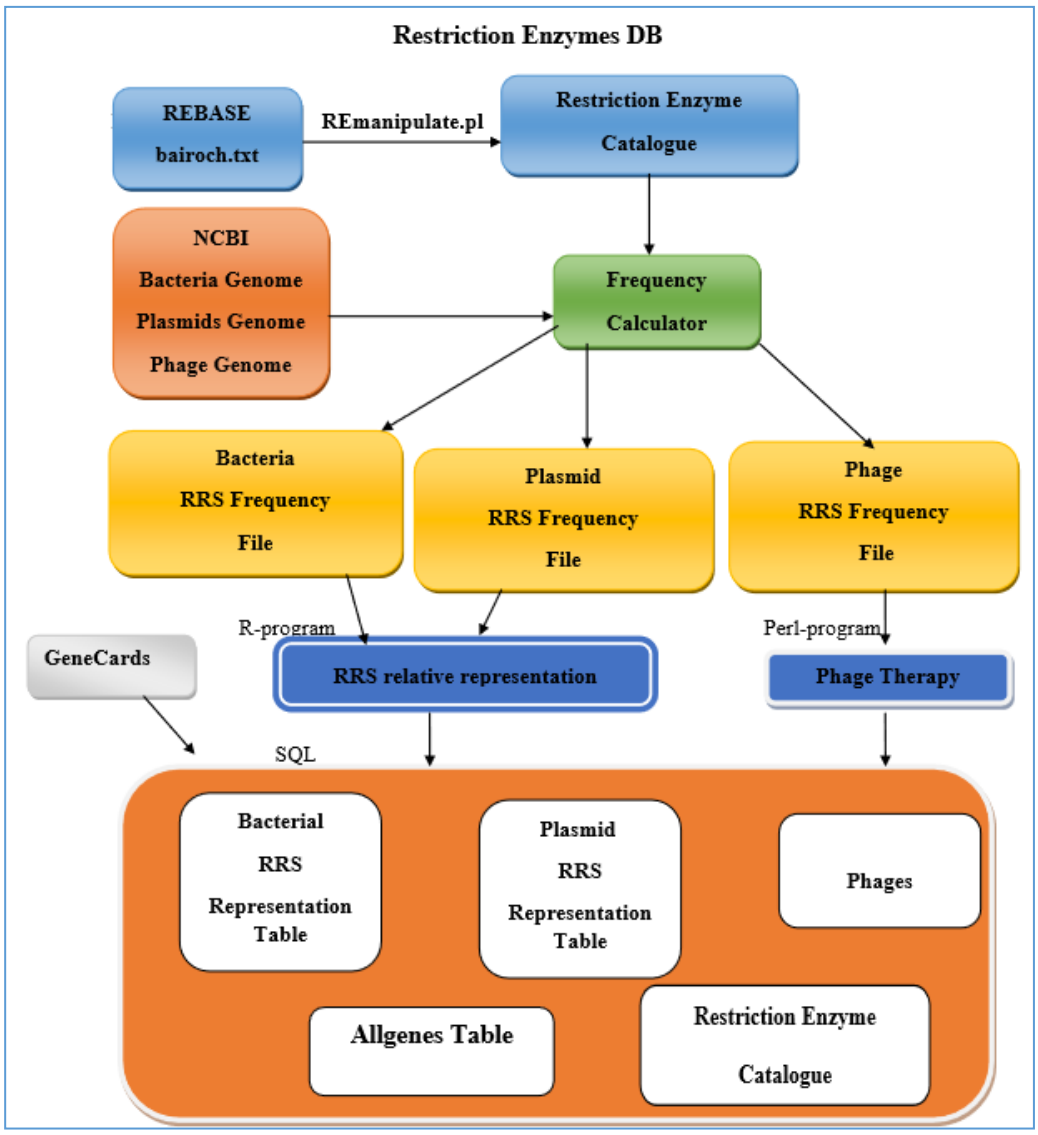


Figure3. Restriction Enzyme DB in SQL. REBASE bairoch.txt file is manipulated using a Perl program to create Restriction Enzyme Catalogue (Table1) which was the input table in SQL. NCBI download of bacteria and their plamids fasta sequences and phages fasta sequences. The frequency calculator Perl program (Figure 2) is run on all the fasta sequence files of the bacteria sequences and outputs a frequency file for each genome. Restriction-recognition site observed frequency and expected frequency. The frequency files for all bacterial genomes were in one file (bacteria RRS frequency file). In R, chi-square goodness of fit test was used to create RRS relative representation file which is the input file for Bacterial RRS representation table in SQL. Plasmid genomes were done as the same manner as bacterial genome. RRS relative representation was an input file for Plasmid RRS Representation Table. Phage genomes were processed in the same way by the frequency calculator. The phage RRS frequency file was manipulated by a Perl program that produces a phage therapy table (Table 9) which was the input of phages table in SQL. A series of gene associated disease files were downloaded from GeneCards to produce Allgenes Table in SQL.

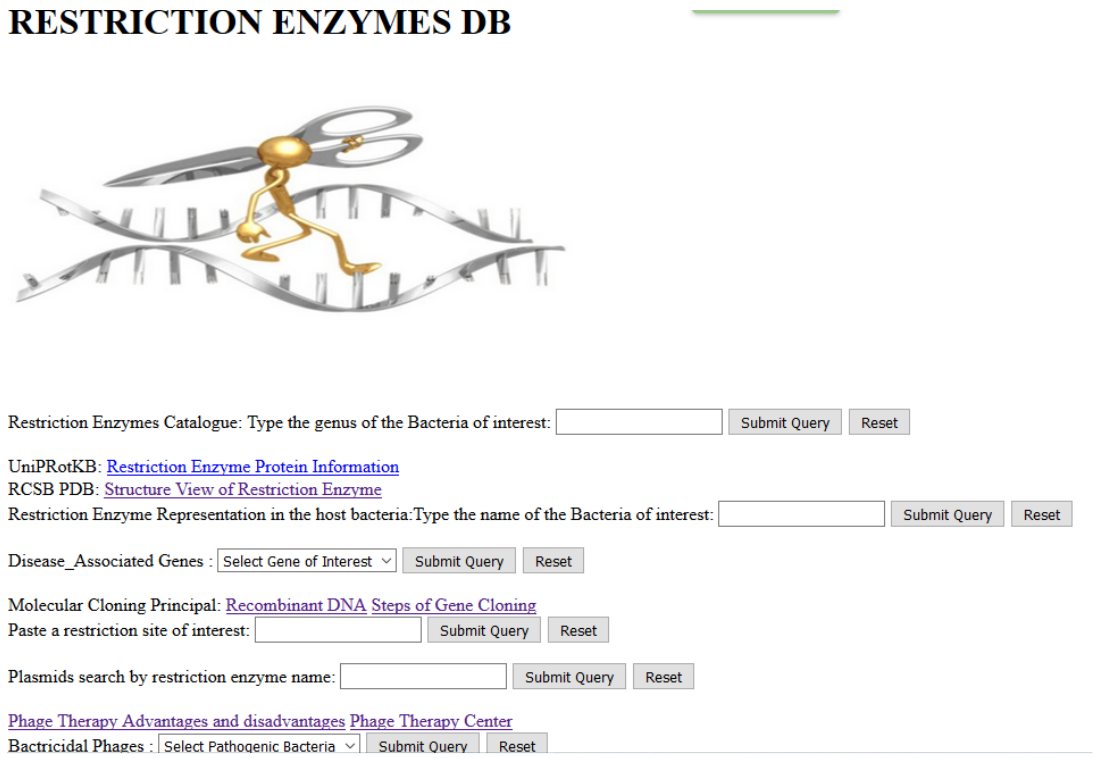


Figure 4. Restriction Enzyme DB web interface. The bacteria name is used to access the Restriction Enzyme catalogue. UniPRotKB link provides an access to UniPRotKB protein database were the user can get information about the restriction enzyme of interest (PDB ID, protein family, protein aminoacid sequences) using the name of the restriction enzyme supplied by the Restriction Enzyme Catalogue. RCSB PDB link enable the user to view the 3D-structure of the restriction enzyme and its subunits of interest using the PDB ID or the name of the enzyme. Bacteria name as a user input is used to access Bacterial RRSR representation Table to get information about the RRSR in bacteria of interest. The user selects from a list of diseases one disease of interest which is used as input to access allgenes table and produces a list of genes associated with the disease of interest, gene name and gene position as well as the genes are linked to NCBI gene database to get more information about each gene, its fasta sequence, differential gene expression and publications related to the gene of interest. For general information about the gene cloning and recombinant DNA, two links for an article and a textbook are provided. To get information about the suitable restriction enzyme to use in gene cloning, the user can input any restriction-recognition site of interest and choose from the list restriction-modification type II restriction enzymes and get information about their host bacteria. Plasmids information can be accessed by the name of the restriction enzyme of interest. Phage therapy summary and research center are added as links to give the user information about the advantages and disadvantages as well as applications and on-going research. Phages therapy table can be accessed by the user to select the suitable pathogenic bacteria of interest.

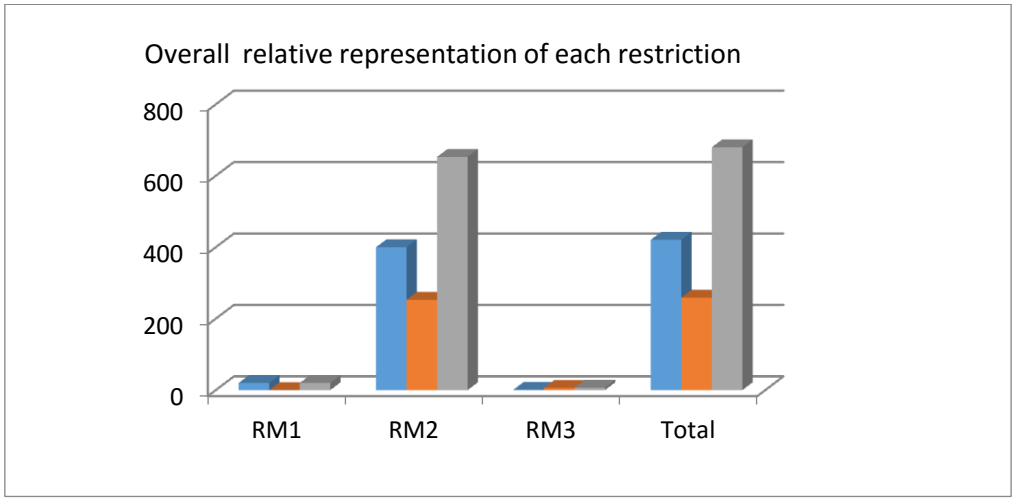


Figure 5. Overall relative representation of restriction modification systems. The figure illustrates the three types of the restriction modification system, RM1 restriction modification type I, RM2 restriction modification type II and RM3 restriction modification type III. About 96.029 % of 680 restriction enzymes were RM2 (653). About 60.24% of RM2 (400) were underrepresented and 39.76% (253) were overrepresented. Blue bar refers to the underrepresented, orange bar refers to overrepresented, grey bar refers to the total.

Table 1. Restriction modification system type I in Restriction Enzyme Catalogue I: Bacteria is the host bacteria**, Accession** is the accession number of the bacteria in the NCBI database; **taxid** is the taxonomy identification number in NCBI database; **RE**name is the restriction enzyme name; **RE**seq is the restriction-recognition site sequence (A: Adenine, C: Cytosine, G: Guanine, T: Thyamine).

Bacteria	Accession	taxid	Rename	REseq
Acetobacter_pasteurianus_386B	NC_021991	1266844	ApaLI	ACC.{5}CTTC
Acetobacter_pasteurianus_386B	NC_021991	1266844	ApaLI	ATC.{5}CTTC
Acinetobacter_baumannii_1656-2	NC_017162	696749	AbaBGI	TTCA.{6}TCC
Bacillus_cellulosilyticus_DSM_2522	NC_014829	649639	BceRI	CCC.{5}CTC
Desulfurobacterium_thermolithotrophum	NC_015185	868864	DvuI	CACC
Escherichia_coli_042	NC_017626	216592	EcoprrI	CCA.{7}ATGC
Escherichia_coli_042	NC_017626	216592	EcoprrI	CCA.{7}GTGC
Escherichia_coli_042	NC_017626	216592	EcoprrI	GCAC.{7}TGG
Escherichia_coli_042	NC_017626	216592	EcoprrI	GCAT.{7}TGG
Escherichia_coli_BL21(DE3)	NC_012892	469008	EcoBI	AGCA.{8}TCA
Escherichia_coli_BL21(DE3)	NC_012892	469008	EcoBI	GGA.{8}ATGC
Escherichia_coli_BL21(DE3)	NC_012892	469008	EcoBI	TGA.{8}TGCT
Escherichia_coli_BL21(DE3)	NC_012892	469008	EcoprrI	CCA.{7}ATGC
Escherichia_coli_BL21(DE3)	NC_012892	469008	EcoprrI	CCA.{7}GTGC
Escherichia_coli_B_str._REL606	NC_012967	413997	EcoprrI	GCAT.{7}TGG
Escherichia_coli_BW2952	NC_012759	595496	EcoBI	AGCA.{8}TCA
Haemophilus_influenzae_Rd_KW20	NC_000907	71421	HindI	GTCAAC
Haemophilus_influenzae_Rd_KW20	NC_000907	71421	HindI	GTCGAC
Haemophilus_influenzae_Rd_KW20	NC_000907	71421	HindI	GTTAAC
Haemophilus_influenzae_Rd_KW20	NC_000907	71421	HindI	GTTGAC
Haemophilus_influenzae_Rd_KW20	NC_000907	71421	HindNI	GATC
Helicobacter_pylori_26695	NC_018939	85962	Hpy87AI	CT.{2}TCC
Helicobacter_pylori_26695	NC_018939	85962	Hpy87AI	GGA.{2}AG
Helicobacter_pylori_26695	NC_018939	85962	HpyAXVIII	CT.{2}TCC
Helicobacter_pylori_26695	NC_018939	85962	HpyAXVIII	GGA.{2}AG
Helicobacter_pylori_J99	NC_000921	85963	Hpy99XVI	AAG.{6}TAAAG
Helicobacter_pylori_J99	NC_000921	85963	Hpy99XVI	ATAC.{5}ATAC
Helicobacter_pylori_J99	NC_000921	85963	Hpy99XVI	ATAC.{5}ATAT
Helicobacter_pylori_J99	NC_000921	85963	Hpy99XVI	ATAT.{5}ATAT
Helicobacter_pylori_J99	NC_000921	85963	Hpy99XVI	ATAT.{5}GTAC
Helicobacter_pylori_J99	NC_000921	85963	Hpy99XVI	ATAT.{5}GTAT
Helicobacter_pylori_J99	NC_000921	85963	Hpy99XVI	GTAC.{5}ATAC

Table 2. Restriction modification system type II in Restriction Enzyme Catalogue I: Bacteria is the host bacteria**, Accession** is the accession number of the bacteria in the NCBI database; **taxid** is the taxonomy identification number in NCBI database; **RE**name is the restriction enzyme name; **RE**seq is the restriction-recognition site sequence (A: Adenine, C: Cytosine, G: Guanine, T: Thyamine).

Bacteria	Accession	taxid	Rename	Reseq
Bacillus_cereus_ATCC_10987	NC_003909	222523	BceSII	CGAAG
Bacillus_cereus_ATCC_10987	NC_003909	222523	BceSII	GGACC
Bacillus_coagulans_2-6	NC_015634	941639	Bco10278I	GAATC
Bacillus_coagulans_2-6	NC_015634	941639	Bco10278I	GATTC
Bacillus_halodurans_C-125	NC_002570	272558	BhaI	GATGC
Bacillus_halodurans_C-125	NC_002570	272558	BhaI	GCATC
Bacillus_megaterium_DSM_319	NC_014103	592022	BmgAI	GGGCCC
Bacillus_megaterium_DSM_319	NC_014103	592022	BmgAI	GTGCAC
Bacillus_megaterium_DSM_319	NC_014103	592022	BmgAI	GTGCCC
Bacillus_subtilis_PY79	NC_022898	1415167	Bsu1192II	CCGG
Bacillus_subtilis_PY79	NC_022898	1415167	Bsu1192II	CGCG
Bacteroides_fragilis_638R	NC_016776	862962	BfrAI	ATCGAT
Bifidobacterium_breve_UCC2003	NC_020517	326426	BbrUII	GGCGCC
Bifidobacterium_breve_UCC2003	NC_020517	326426	BbrUII	GTCGAC
Bordetella_bronchiseptica_253	NC_019382	568707	BbrAI	AAGCTT
Caldicellulosiruptor_bescii_DSM_6725	NC_012034	521460	CbeI	CACGAG
Caldicellulosiruptor_bescii_DSM_6725	NC_012034	521460	CbeI	CTCGTG
Caulobacter_crescentus_NA1000	NC_011916	565050	CcuI	CGCCGGCA
Caulobacter_crescentus_NA1000	NC_011916	565050	CcuI	CGCCGGCG
Cellulophaga_lytica_DSM_7489	NC_015167	867900	Cma23826I	AAAAGAG
Cellulophaga_lytica_DSM_7489	NC_015167	867900	Cma23826I	AAAAGGG
Clostridium_difficile_BI1	NC_017179	699034	CdiAI	GC.{1}GC
Clostridium_perfringens_SM101	NC_008262	289380	CpfI	GATC
Clostridium_perfringens_SM101	NC_008262	289380	CpfI	GGCC
Corynebacterium_diphtheriae_241	NC_016782	698966	Cdp1	CAAAAA
Corynebacterium_diphtheriae_241	NC_016782	698966	Cdi11397I	CTGCGC

Table 3. Restriction modification system type III in Restriction Enzyme Catalogue I: Bacteria is the host bacteria**, Accession** is the accession number of the bacteria in the NCBI database; **taxid** is the taxonomy identification number in NCBI database; **RE**name is the restriction enzyme name; **RE**seq is the restriction-recognition site sequence (A: Adenine, C: Cytosine, G: Guanine, T: Thyamine).

Bacteria	Accession	taxid	REname	REseq
Bacillus_cereus_ATCC_10987	NC_003909	222523	BceSI	CGCG
Bibersteinia_trehalosi_USDA-ARS-USMARC-192	NC_020515	1171377	Btr192II	ACATC
Bibersteinia_trehalosi_USDA-ARS-USMARC-192	NC_020515	1171377	Btr192II	GATC
Bibersteinia_trehalosi_USDA-ARS-USMARC-192	NC_020515	1171377	Btr192II	GATGT
Geobacter_metallireducens_GS-15	NC_007517	269799	GmeII	CCTGGA
Geobacter_metallireducens_GS-15	NC_007517	269799	GmeII	GGATC
Geobacter_metallireducens_GS-15	NC_007517	269799	GmeII	TCCAGG
Neisseria_gonorrhoeae_FA_1090	NC_002946	242231	NgoAX	CCACC
Neisseria_gonorrhoeae_FA_1090	NC_002946	242231	NgoAX	GGTGG
Staphylococcus_aureus_subsp._aureus_USA300_FPR:	NC_007793	451515	SauBI	GATC
Staphylococcus_aureus_subsp._aureus_USA300_FPR:	NC_007793	451515	SauBI	GG.{1}CC
Staphylococcus_aureus_subsp._aureus_USA300_FPR:	NC_007793	451515	SauBMKI	GG.{1}CC

Table 4. Top underrepresented restriction enzymes recognition sites of restriction modification system type I in Bacteria: Index refers to the index and bacteria species in Table 4a **, REname** is the restriction enzyme name; **REseq : Reversecomplement** is the restriction-recognition site sequence and its reverse complement (.{} is any nucleotide A,C,T or G with the number of nucleotides is indicated in the curly brackets); **RR** is the relative representation of the restriction-recognition site, the ratio of the observed frequency of occurrence to the expected frequency of occurrence, RR is less than 1 that indicates underrepresentation; **Difference** is the difference between the observed and expected frequencies, **p-value** (significance level = 0.0001).

Index	REname	Rseq:Reversecomplement	RR	Difference	p-value
1	SauBMKI	GG.{1}CC:GG.{1}CC	0	-1901	0
2	Hpy87AI	CT.{2}TCC:GGA.{2}AG	0.00083682	-1.09E+003	9.03E-261
3	Bsu6633I	GG.{1}CC:GG.{1}CC	0.56	-4.11E+003	2.61E-250
4	NgoAV	GCA.{8}TGC:GCA.{8}TGC	0.005172414	-5.78E+002	8.77E-126
5	NgoBI	GAG.{5}TAC:GTA.{5}CTC	0	-3.20E+002	3.13E-114
6	LlaG2I	CT.{1}GATG:CATC.{1}AG	0	-4.77E+002	7.81E-107
7	HcuI	GAA.{6}TCGG:CCGA.{6}TTC	0	-4.63E+002	2.90E-102
8	BcrAI	GG.{2}CC:GG.{2}CC	0.65	-1.58E+003	5.05E-084
9	EcoKI	AAC.{6}GTGC:GCAC.{6}GTT	0	-2.87E+002	6.07E-064
10	EcoprrI	GCAT.{7}TGG:CCA.{7}ATGC	0.003460208	-2.88E+002	9.82E-064
11	Spn23FI	CAC.{7}CTG:CAG.{7}GTG	0	-2.86E+002	1.00E-063
12	EcoBI	TGA.{8}TGCT:AGCA.{8}TCA	0	-2.77E+002	9.16E-062
13	HindI	GTTGAC:GTCAAC	0.110215054	-3.31E+002	2.30E-059
14	SrfI	GCA.{6}TTAA:TTAA.{6}TGC	0	-1.99E+002	9.34E-045
15	PaeAI	AGG.{5}TTCA:TGAA.{5}CCT	0	-1.90E+002	2.28E-042
16	SpnD39IIIC	GAA.{9}TTTG:CAAA.{9}TTC	0.104478	-1.31E+002	2.57E-033

Table 4a. The index of table 4 and the host bacteria full name.

Index	Bacteria
1	Staphylococcus_aureus_subsp._aureus_MSHR1132
2	Helicobacter_pylori_26695
3	Bacillus_subtilis_subsp._subtilis_str._168
4	Neisseria_gonorrhoeae_FA_1090
5	Neisseria_gonorrhoeae_FA_1090
6	Lactococcus_lactis_subsp._lactis_KLDS_4.0325
7	Hahella_chejuensis_KCTC_2396
8	Bacillus_cereus_ATCC_10987
9	Escherichia_coli_str._K-12_substr._MG1655
10	Escherichia_coli_str._K-12_substr._DH10B
11	Streptococcus_pneumoniae_D39
12	Escherichia_coli_B_str._REL606
13	Haemophilus_influenzae_Rd_KW20
14	Streptococcus_pyogenes_MGAS10750
15	Pseudomonas_aeruginosa_PA7
16	Streptococcus_pneumoniae_D39

Table 5. Top underrepresented restriction enzymes recognition sites of restriction modification system type II in Bacteria: Index refers to the index and bacteria species in the Table 5a, REname is the restriction enzyme name; **REseq : Reversecomplement** is the restriction-recognition site sequence and its reverse complement, Difference is the difference between the observed and expected values, **RR** is the relative representation of the restriction-recognition site, the ratio of the observed frequency of occurrence to the expected frequency of occurrence, RR is less than 1 that indicates underrepresentation; **Difference** is the difference between the observed and expected frequencies, **p-value** (significance level = 0.0001).

Index	REname	Reseq : Reversecomplement	RR	Difference	p-value
1	NgoAVII	GATC:GATC	0.52	-3.99E+003	2.29E-272
2	SgrI	CTCGAG:CTCGAG	0.28	-1.99E+003	2.23E-241
3	Hpy99II	GTGAC:GTCAC	0.09	-1.03E+003	1.04E-188
4	DsaVI	GTCTAC:GTAGAC	0.012806	-8.48E+002	2.35E-181
5	PluTI	AGGCCT:AGGCCT	0.18	-7.93E+002	1.32E-119
6	AbaI	TTAA:TTAA	0.61	-2.22E+003	9.78E-117
7	CauIII	CTGCAG:CTGCAG	0.36	-1.00E+003	1.79E-102
8	BbrAI	AAGCTT:AAGCTT	0.09	-3.61E+002	6.11E-066
9	HindIII	GTTGAC:GTCAAC	0.11	-3.31E+002	2.30E-059
10	BpeI	AAGCTT:AAGCTT	0.11	-2.79E+002	2.42E-049
11	XcaI	TCTAGA:TCTAGA	0.29	-3.23E+002	1.60E-045
12	NgoCI	GGCGCT:AGCGCC	0.4	-3.85E+002	5.91E-037

Table 5a. The index of table 5 and the host bacteria full name.

Index	Bacteria
1	Neisseria_gonorrhoeae_FA_1090
2	Streptomyces_griseus_subsp._griseus_NBRC_13350
3	Helicobacter_pylori_J99
4	Dactylococcopsis_salina_PCC_8305
5	Photorhabdus_luminescens_subsp._laumondii_TTO1
6	Arthrobacter_aurescens_TC1
7	Chloroflexus_aurantiacus_J-10-fl
8	Bordetella_bronchiseptica_RB50
9	Haemophilus_influenzae_Rd_KW20
10	Bordetella_pertussis_Tohama_I
11	Xanthomonas_campestris_pv._vesicatoria_str._85-10
12	Neisseria_gonorrhoeae_FA_1090

Table 6. Top overrepresented restriction enzymes recognition sites of restriction modification system type II in bacteria: Index refers to the index and bacteria species in Table 6a**, REname** is the restriction enzyme name; **REseq : Reversecomplement** is the restriction-recognition site sequence and its reverse complement (.{} is any nucleotide A,C,T or G with the number of nucleotides is indicated in the curly brackets); **Difference** is the difference between the observed and expected frequencies, **RR** is the relative representation of the restriction recognition site, the ratio of the observed frequency of occurrence to the expected frequency of occurrence, RR is greater than 1 that indicates overrepresentation; **p-value** (significance level = 0.0001).

Index	REname	Reseq : Reversecomplement	Difference	RR	p-value
1	BcrI	GAAGAG:CTCTTC	3469	4.49	0
2	BctI	CTCTTC:GAAGAG	3424	4.58	0
3	SonI	ATCGAT:ATCGAT	3282	3.34	0
4	BhaII	GGCC:GGCC	5691	1.59	3.02E-285
5	CglI	GCGGC:GCCGC	3457	1.74	9.79E-206
6	M.CglI	GCGGC:GCCGC	3457	1.74	9.79E-206
7	BfrAI	ATCGAT:ATCGAT	2046	2.28	1.33E-177
8	BfrBI	ATCGAT:ATCGAT	2060	2.29	8.11E-177
9	EcoCKI	CACAG:CTGTG	2809	1.54	3.04E-132
10	TdeII	GAAGAG:CTCTTC	1098	3.02	3.82E-126
11	BanAI	ATCGAT:ATCGAT	1518	1.85	1.42E-100
12	Pae2kI	CTCGAG:CTCGAG	1447	1.67	2.83E-081
13	CglAI	GCATGC:GCATGC	937	1.97	2.93E-066
14	CpfI	GGCC:GGCC	1029	1.8	1.48E-065
16	EcoHK31I	TGGCCG:CGGCCA	914	1.75	1.34E-056
17	LlaG2I	CT.{1}GATG:CATC.{1}AG	613	2.38	2.18E-056
18	Xgl3216I	TTCGAA:TTCGAA	178	1.31	9.62E-007
19	EciI	CCGCGG:CCGCGG	1496	1.05	0.2126306

Table 6a. The index of table 6 and the host bacteria full name.

Index	Bacteria
1	Bacillus_cereus_ATCC_10987
2	Bacillus_cereus_ATCC_10987
3	Shewanella_oneidensis_MR-1
4	Bacillus_halodurans_C-125
5	Corynebacterium_glutamicum_ATCC_13032
6	Corynebacterium_glutamicum_ATCC_13032
7	Bacteroides_fragilis_YCH46_DNA
8	Bacteroides_fragilis_NCTC_9343
9	Escherichia_coli_CFT073
10	Treponema_denticola_ATCC_35405
11	Bacillus_anthracis_str._Ames
12	Pseudomonas_aeruginosa_PAO1
13	Corynebacterium_glutamicum_ATCC_13032
14	Clostridium_perfringens_str._13
16	Escherichia_coli_str._K-12_substr._MG1655
17	Lactococcus_lactis_subsp._lactis_Il1403
18	Xylella_fastidiosa_Temecula1
19	Enterobacter_cloacae_EcWSU1

Table 7. Relative representation of restriction enzymes recognition sites of restriction modification system type III in Bacteria: Index refers to the index and bacteria species in Table 7a**, REname** is the restriction enzyme name, **REseq:Reversecomplement** is the restriction recognition site sequence and the Reversecomplement, **Difference** is the difference between observed and expected frequencies, **RR** is the relative representation of the restriction recognition sites, the ratio of the observed frequency of occurrence to the expected frequency of occurrence, RR is greater than 1 that indicates overrepresentation, **p-value** (significance level = 0.0001).

Index	REname	Reseq: Reversecomplement	Difference	RR	p-value
1	GmeII	TCCAGG:CCTGGA	4537	4.53	0
2	BceSI	CGCG:CGCG	4106	1.78	1.89E-253
3	Btr192II	GATGT:ACATC	1384	1.53	4.08E-065
4	NgoAX	CCACC:GGTGG	814	1.33	3.08E-027
5	NgoAX	GGTGG:CCACC	762	1.3	8.05E-024

Table 7a. The index of table 7 and the host bacteria full name.

Index	Bacteria
1	Geobacter_metallireducens_GS-15
2	Bacillus_cereus_ATCC_10987
3	Bibersteinia_trehalosi_USDA-ARS-USMARC-192
4	Neisseria_gonorrhoeae_FA_1090
5	Neisseria_gonorrhoeae_FA_1090

Table 8. Total number of restriction enzymes in each restriction modification system and their restriction recognition sites overall relative representation. Underrepresented and Overrepresented are the numbers of restriction enzymes with their recognition sites underrepresented and overrepresented in each restriction modification system respectively.

	Underrepresented	Overrepresented	Total
RM1	20	0	20 (2.94) %
RM2	400 (61.255%)	253 (35.9%)	653(96.029%)
RM3	1	6	7(1.029%)
Total	421	259	680

Table 9. Cross species RRSR in Bacteria. Relative representation of restriction enzymes recognition sites of restriction modification system type I, II, III in non-host bacteria: Index refers to the bacteria species and is explained in the Table 9a, **REname** is the restriction enzyme name, **REseq:Reversecomplement** is the restriction recognition site sequence and the Reversecomplement; **RR** is the relative representation of the restriction recognition sites, the ratio of the observed frequency of occurrence to the expected frequency of occurrence, **RR** is greater than 1 that indicates overrepresentation; **p-value** (significance level = 0.0001).

Index	REname	Rtype	Rseq:Reversecomplement	Difference	RR	p-value
1	AjoI	R2	CCTGG:CCAGG	-242	0.705238	1.03E-010
2	M.EacI	R2	GGATC:GATCC	-682	0.039437	1.04E-138
3	M.Tsp32I	R2	TCGA:TCGA	-731	0.104167	1.04E-130
3	M.TspPM5I	R2	TCGA:TCGA	-731	0.104167	1.04E-130
3	M.TthHB8I	R2	TCGA:TCGA	-731	0.104167	1.04E-130
4	BscUI	R2	GCATC:GATGC	-189	0.670157	1.04E-009
4	BscWI	R2	GCATC:GATGC	-189	0.670157	1.04E-009
5	HsuI	R2	CTAG:CTAG	-195	0.170213	1.05E-031
6	TaaI	R2	AC.{1}GT:AC.{1}GT	-800	0.002494	1.06E-174
7	XmiI	R2	GTCTAC:GTAGAC	-798	0.001252	1.08E-174
3	M.EcoDR2	R1	TCA.{12}GTCG:CGAC.{21}TGA	-775	0.001289	1.08E-169
8	CalI	R2	GAT.{8}TCGT:ACGA.{8}ATC	-146	0.633166	1.08E-008
6	PctI	R1	TCG.{7}CGT:ACG.{7}CGA	-683	0.001462	1.09E-149
3	M.Csp205I	R2	TTCGAA:TTCGAA	-247	0.711111	1.09E-010
3	M.Csp68KII	R2	TTCGAA:TTCGAA	-247	0.711111	1.09E-010
3	TscAI	R2	ACGT:ACGT	-632	0.22549	1.10E-088
2	HsoII	R2	GCGC:GCGC	-604	0.220645	1.11E-085
2	HspAI	R2	GCGC:GCGC	-604	0.220645	1.11E-085
5	PdmI	R1	CA.{7}TAAAG:CTTTA.{7}TG	-375	0.00266	1.11E-082

Table 9a. The index of table 9 and the non-host bacteria full name.

Index	Bacteria
1	Mycobacterium tuberclosis H37RV
2	Pseudomons aeroginosa PAO1
3	Staphylococcus aureus subsp. Aureus NCTC 8325
5	Esherichia coli str. K-12 substr. MG1655
6	Neisseria gonorrhoeae FA 1090
7	Acintobacter baumanii ATCC 17978
8	Salmonella enterica subsp. Enterica serovar Typhimurium str. LT2
9	Haemophilus influenza Rd KW20

Table 9b. Eco71KI recognition sites representation in different bacterial species. Index refers to the bacteria species and is explained in the Table 9a, **REname** is the restriction enzyme name, Rtype is the restriction modification system type, **REseq:Reversecomplement** is the restriction recognition site sequence and the Reversecomplement; **RR** is the relative representation of the restriction recognition sites, the ratio of the observed frequency of occurrence to the expected frequency of occurrence, **RR** is greater than 1 that indicates overrepresentation; **p-value** (significance level = 0.0001).

Index	REname	Rtype	Rseq:Reversecomplement	Difference	RR	p-value
7	Eco71KI	R2	ATGCAT:ATGCAT	1769	16.11965812	0
7	Eco71KI	R2	GGTCTC:GAGACC	2712	5.178736518	0
5	Eco71KI	R2	ATGCAT:ATGCAT	379	14.06896552	3.91E-073
5	Eco71KI	R2	GGTCTC:GAGACC	1979	37.64814815	0
9	Eco71KI	R2	ATGCAT:ATGCAT	2937	5.231988473	0
9	Eco71KI	R2	GGTCTC:GAGACC	-165	0.48757764	5.34E-014
1	Eco71KI	R2	ATGCAT:ATGCAT	3116	7.050485437	0
1	Eco71KI	R2	GGTCTC:GAGACC	988	6.710982659	6.48E-161
6	Eco71KI	R2	ATGCAT:ATGCAT	1113	2.899317406	8.22E-120
6	Eco71KI	R2	GGTCTC:GAGACC	186	1.271929825	2.41E-006
2	Eco71KI	R2	ATGCAT:ATGCAT	3685	6.625954198	0
2	Eco71KI	R2	GGTCTC:GAGACC	2382	9.270833333	0
8	Eco71KI	R2	ATGCAT:ATGCAT	4	1.133333333	0.622460656
8	Eco71KI	R2	GGTCTC:GAGACC	5851	52.77876106	0
3	Eco71KI	R2	ATGCAT:ATGCAT	1158	2.354385965	1.28E-103
3	Eco71KI	R2	GGTCTC:GAGACC	1238	4.762917933	1.26E-177

Table 10. Relative representation of restriction enzymes recognition sites in Plasmids: REname is the restriction enzyme name; **Represenatation:** is dependent on the relative representation (RR) of the restriction recognition site which is the ratio of observed frequency of occurrence and expected frequency of occurrence of the recognition site; **p-value** (significance level= 0.0001).

Rename	Plasmid_name	Representation	p-value
SecII	Salmonella_enterica	underrepresented	3.38E-57
RtrI	Rhizobium_tropici	underrepresented	1.01E-13
SexI	Salmonella_enterica	underrepresented	1.57E-11
Psp29I	Pseudomonas_syringae	underrepresented	7.78E-11
SfaAI	Shigella_flexneri	underrepresented	1.34E-09
LmnI	Leisingera_methylohalidivorans	underrepresented	9.17E-09
NgoJI	Neisseria_gonorrhoeae	underrepresented	1.99E-08
NmeBI	Nitrosomonas_eutropha	underrepresented	4.12E-06
Kpn12I	Klebsiella_pneumoniae	underrepresented	4.30E-06
Mlu40I	Mesorhizobium_loti	underrepresented	1.75E-05
NtaI	Natronococcus_occultus	Overrepresented	2.77E-279
ChiI	Corynebacterium_halotolerans	Overrepresented	6.54E-70
AbaI	Acinetobacter_baumannii	Overrepresented	8.14E-68
CmaLM2II	Chamaesiphon_minutus	Overrepresented	4.38E-40
BtsCI	Bacillus_toyonensis	Overrepresented	3.46E-13
I-CreI	Candidatus_Rickettsia	Overrepresented	1.14E-05
LpnI	Legionella_pneumophila	Overrepresented	3.86E-05
NgbI	Neisseria_gonorrhoeae	Overrepresented	4.39E-05
M.Lph145II	Legionella_pneumophila	Overrepresented	5.85E-05

Table 11. Underrepresentation restriction enzymes recognition-sites in Phages: REname is the restriction enzyme name; **RR** is the relative representation of the restriction-recognition site, the ratio of the observed frequency of occurrence to the expected frequency of occurrence, RR is less than 1 that indicates that underrepresentation; **p-value** (significance level= 0.0001).

Rename	Phage_name	RR	p-value
Hpy99IV	Helicobacter_pylori_J99	0.43	3.29E-020
NgoAIII	Neisseria_gonorrhoeae_FA_1090	0.64	3.29E-012
NlaII	Neisseria_lactamica_020-06	0.64	3.29E-012
Pdi8503III	Parabacteroides_distasonis_ATCC_8503	0	3.29E-024
RpaI	Rhodopseudomonas_palustris_CGA009	0	3.29E-024
Saf8902III	Spirochaeta_africana_DSM_8902	0	3.29E-024
SauBI	Staphylococcus_aureus_04-02981	0	3.29E-024
MfeI	Mycoplasma_fermentans_M64	0	3.29E-024

Table 12. Phage Therapy. The pathogenic bacteria and its restriction enzymes that is susceptible to the suitable bactericidal phage.

Pathogenic Bacteria	RE_name	Bactericidal Phage
Acetobacter_aceti	AauI	Arthrobacter_aurescens_TC1
Bacillus_anthracis	BanAI	Bacillus_anthracis_str._A0248
Bacillus_subtilis	Bsu1192I	Bacillus_subtilis_QB928
Caseobacter_polymorphus	CpoI	Chlorobium_phaeobacteroides_BS1
Caulobacter_crescentus	CcuI	Caulobacter_crescentus_NA1000
Caulobacter_species	CatHI	Croceibacter_atlanticus_HTCC2559
Enterobacter_cloacae	EciI	Escherichia_coli_NA114
Escherichia_coli	EciI	Enterobacter_cloacae_EcWSU1
Escherichia_coli	EciI	Enterobacter_cloacae_SCF1
Escherichia_coli	EciI	Escherichia_coli_042
Haemophilus_influenzae	HindIII	Haemophilus_influenzae_Rd_KW20
Helicobacter_pylori	Hpy303I	Helicobacter_pylori_UM299
Klebsiella_oxytoca	KoxII	Klebsiella_oxytoca_E718
Klebsiella_oxytoca	KoxII	Klebsiella_oxytoca_KCTC_1686

Supplement

Table S1. The Frequency Calculator Output for Escherichia_coli_str._K-12_substr._MDS42_DNA genome: REname is the restriction enzyme name; REtype is the Restriction-Modification system type; REseq: Reversecomplement is the restriction-recognition site sequence and its reverse complement (A: Adenine, G: Guanine, C: Cytosine, T: Thyamine.{} is any nucleotide A,C,T or G with the number of nucleotides is in the curly brackets) ; Obs. Freq: Observed Occurrnce Frequency; Exp. Freq: Expected occurrence frequency.

REname	REtype	REseq:Reversecomplement	Obs.Freq	Exp. Freq
EcoprrI	R1	GCAT.{7}TGG:CCA.{7}ATGC	656	246
M.EcoprrI	M1	GCAT.{7}TGG:CCA.{7}ATGC	656	246
Eco53kI	M2	CCGCGG:CCGCGG	1130	1110
Eco53kI	R2	CCGCGG:CCGCGG	1130	1110
EcoHAI	R2	TGGCCG:CGGCCA	1774	1058
EcoHK31I	R2	TGGCCG:CGGCCA	1774	1058
EcoprrI	R1	CCA.{7}GTGC:GCAC.{7}TGG	509	260
M.Eco12581I	M1	AGCA.{6}TGA:TCA.{6}TGCT	833	236
M.Eco29kI	M2	CCGCGG:CCGCGG	1130	1110
M.Eco9001Dcm	M2	CCTGG:CCAGG	10469	4155
M.Eco9001I	M1	CCA.{8}TGAA:TTCA.{8}TGG	678	237
M.Eco9387Dcm	M2	CCTGG:CCAGG	10410	4155
M.EcoEc67Dam	M	GATC:GATC	33540	15575
M.EcoGI	M	GATC:GATC	33540	15575
M.EcoGI	M2	GATC:GATC	33540	15575
M.EcoNI	M2	CCT.{5}AGG:CCT.{5}AGG	437	1014
M.EcoNIH1I	M1	CCT.{5}AGG:CCT.{5}AGG	0	1014
M.EcoprrI	M1	CCA.{7}GTGC:GCAC.{7}TGG	15	260
M.EfaBMDam	M1	CCA.{7}GTGC:GCAC.{7}TGG	0	260

Table S2.The Restriction enzymes list for host Escherichia_coli_str._K-12_substr._MDS42_DNA genome: REname is the restriction enzyme name; REtype is the Restriction-Modification system type; REseq : Reversecomplement is the restriction-recognition site sequence and its reverse complement (A: Adenine, G: Guanine, C: Cytosine, T: Thyamine, .{} is any nucleotide A,C,T or G with the number of nucleotides is in the curly brackets).

This is the Complete genome of Escherichia_coli_str._K-12_substr._MDS42_DNA,_complete_genome.
The whole genome is 3976195 bps.
The probability of Guanine is 0.254998560181279.
The probability of Cytosine is 0.256352367024253.
The probability of Adenine is 0.244299638221968.
The probability of Thyamine is 0.2443494345725.
List of Accession number and the restriction enzyme names and their recognition site sequences:

Organism Name	Rename	REtype	RE seq:Reverse Complement
*Escherichia_coli_str._K-12_substr._MDS42	EcoprrI	R1	GCAT.{7}TGG:CCA.{7}ATGC
*Escherichia_coli_str._K-12_substr._MDS42	M.EcoprrI	M1	GCAT.{7}TGG:CCA.{7}ATGC
Escherichia_coli_str._K-12_substr._MDS42_	Eco53kI	M2	CCGCGG:CCGCGG
Escherichia_coli_str._K-12_substr._MDS42_	Eco53kI	R2	CCGCGG:CCGCGG
Escherichia_coli_str._K-12_substr._MDS42_	EcoHAI	R2	TGGCCG:CGGCCA
Escherichia_coli_str._K-12_substr._MDS42_	EcoHK31I	R2	TGGCCG:CGGCCA
Escherichia_coli_str._K-12_substr._MDS42_	EcoprrI	R1	CCA.{7}GTGC:GCAC.{7}TGG
Escherichia_coli_str._K-12_substr._MDS42_	M.Eco12581I	M1	AGCA.{6}TGA:TCA.{6}TGCT
Escherichia_coli_str._K-12_substr._MDS42_	M.Eco29kI	M2	CCGCGG:CCGCGG
Escherichia_coli_str._K-12_substr._MDS42_	M.Eco9001Dcm	M2	CCTGG:CCAGG
Escherichia_coli_str._K-12_substr._MDS42_	M.Eco9001I	M1	CCA.{8}TGAA:TTCA.{8}TGG
Escherichia_coli_str._K-12_substr._MDS42_	M.Eco9387Dcm	M2	CCTGG:CCAGG
Escherichia_coli_str._K-12_substr._MDS42_	M.EcoEc67Dam	M	GATC:GATC
Escherichia_coli_str._K-12_substr._MDS42_	M.EcoGI	M	GATC:GATC
Escherichia_coli_str._K-12_substr._MDS42_	M.EcoGI	M2	GATC:GATC
Escherichia_coli_str._K-12_substr._MDS42_	M.EcoNIH1I	M1	CCT.{5}AGG:CCT.{5}AGG
Escherichia_coli_str._K-12_substr._MDS42_	M.EcoNIH1I	M2	CCT.{5}AGG:CCT.{5}AGG
Escherichia_coli_str._K-12_substr._MDS42_	M.EcoprrI	M1	CCA.{7}GTGC:GCAC.{7}TGG
Escherichia_coli_str._K-12_substr._MDS42_	M.EcoprrI	R1	CCA.{7}GTGC:GCAC.{7}TGG
Escherichia_coli_str._K-12_substr._MDS42_	M.EfaBMDam	M1	CCA.{7}GTGC:GCAC.{7}TGG

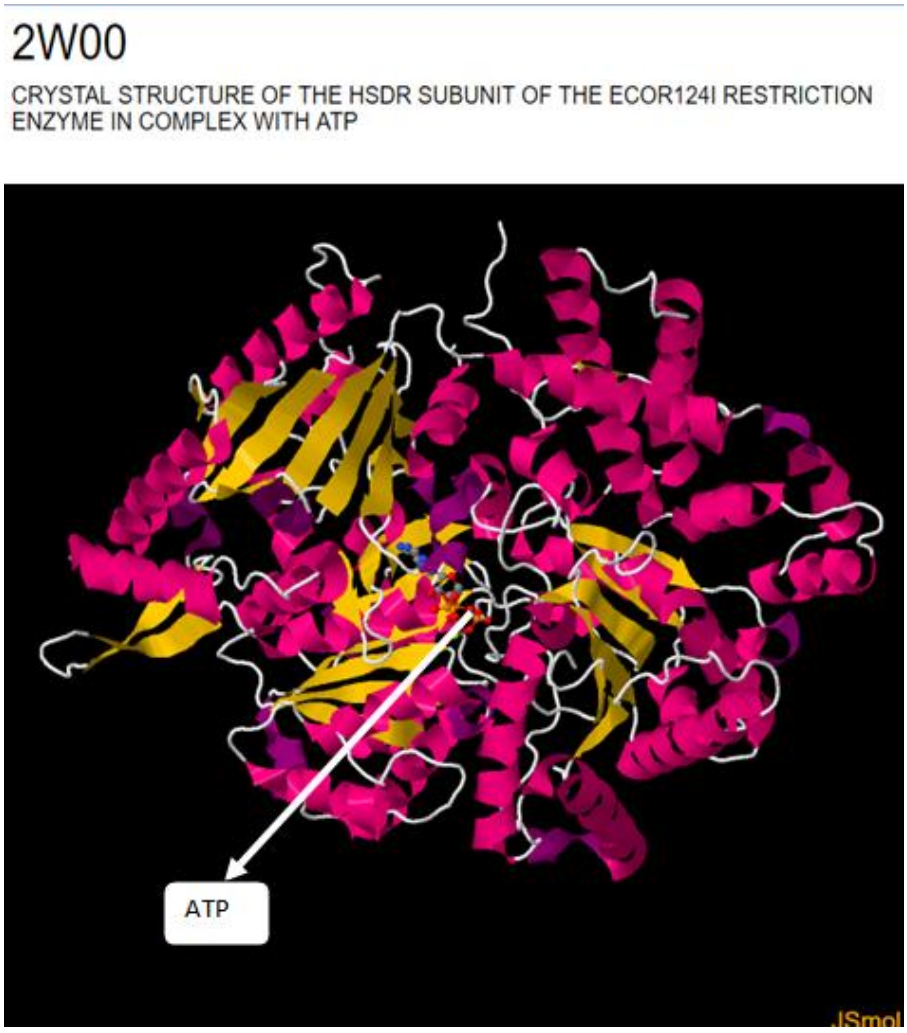


Figure S1. The HsdR subunit of EcoR1 (23,30). EcoR1 is a restriction – modification system type 1 that consists of three subunits; HsdS (specificity subunit that recognizes the restriction-recognition site sequence), HsdR (restriction subunit; ATP-bound subunit that is translocated to the cleavage site), HsdM(methylase subunit, SAM, S-N6 adenosyl methionine pocket). This RCSB PDB illustrates cartoon model of the secondary structure of the restriction subunit with the ATP molecule. The pink ribbons are alpha helix, yellow strips are Beta sheets and purple is the DNA double helix.

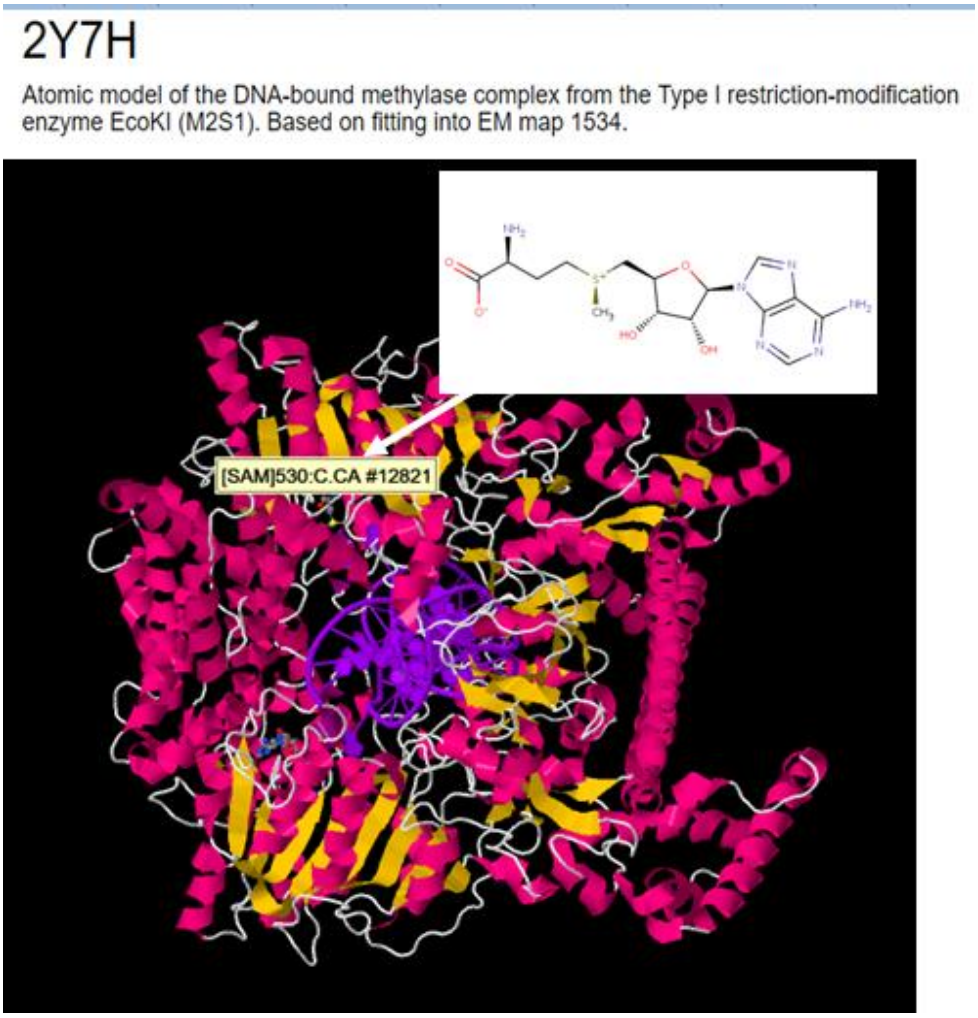


Figure S2. The HsdM subunit of EcoK1 (23, 29). EcoK1 is a restriction – modification system type 1 that consists of three subunits; HsdS (specificity subunit that recognizes the restriction-recognition site sequence), HsdR (restriction subunit; ATP-bound subunit that is translocated to the cleavage site), HsdM (methylase subunit, SAM, S-N6 adenosyl methionine pocket). This RCSB PDB illustrates cartoon model of the secondary structure of the methylase subunit with the SAM molecule necessary for methylation of the restriction-recognition site. The pink ribbons are alpha helix, yellow strips are Beta sheets and purple is the DNA double helix.

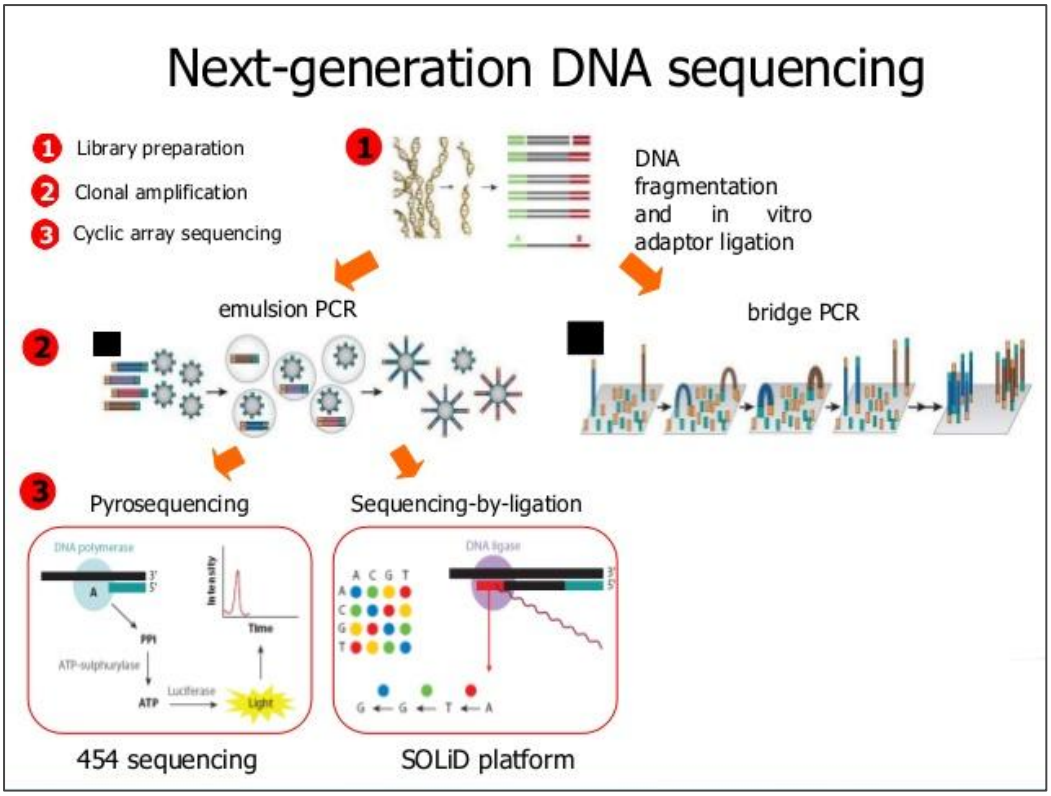


Figure S3: Next-generation DNA sequencing (26). The principal components of next generation sequencing is library preparation, Clonal amplification and massive parallel sequencing. Library preparation from the DNA isolated from the sample is done by fragmentation of DNA using high sound energy (sonication) or by force of DNA through small hole (nebulization). DNA fragments are bound to emulsion beads (adaptors). Denaturing of DNA by heating at 95◦C to form two single stranded DNA (ssDNA) from each double-stranded DNA. Clonal amplification of the fragments of DNA ligated to in vitro adaptors is performed by polymerase chain reaction (PCR) to produce more copies of the insert. Emulsion PCR (27) amplifies ssDNA in streptavidin-coated beads with complementary primers to the adaptors are emulsified in water in oil droplet with DNA polymerase, buffer, dNTPs; deoxynucleotides triphospahtes; denaturation, annealing and elongation. Bridge PCR is used by Illumina for sequencing. Pyrosequencing sequencing is a step wise addition of each dNTP to the emulsion beads and when the dNTP is complementary to the DNA template base a pyrophosphate is released which releases light on formation of ATP. Sequencing by ligation uses DNA ligase enzyme to bind the complematary bases on the different strands (primer-bound template DNA and fluorecent dNTP). DNA ligase will not ligate the bases unless complementary.

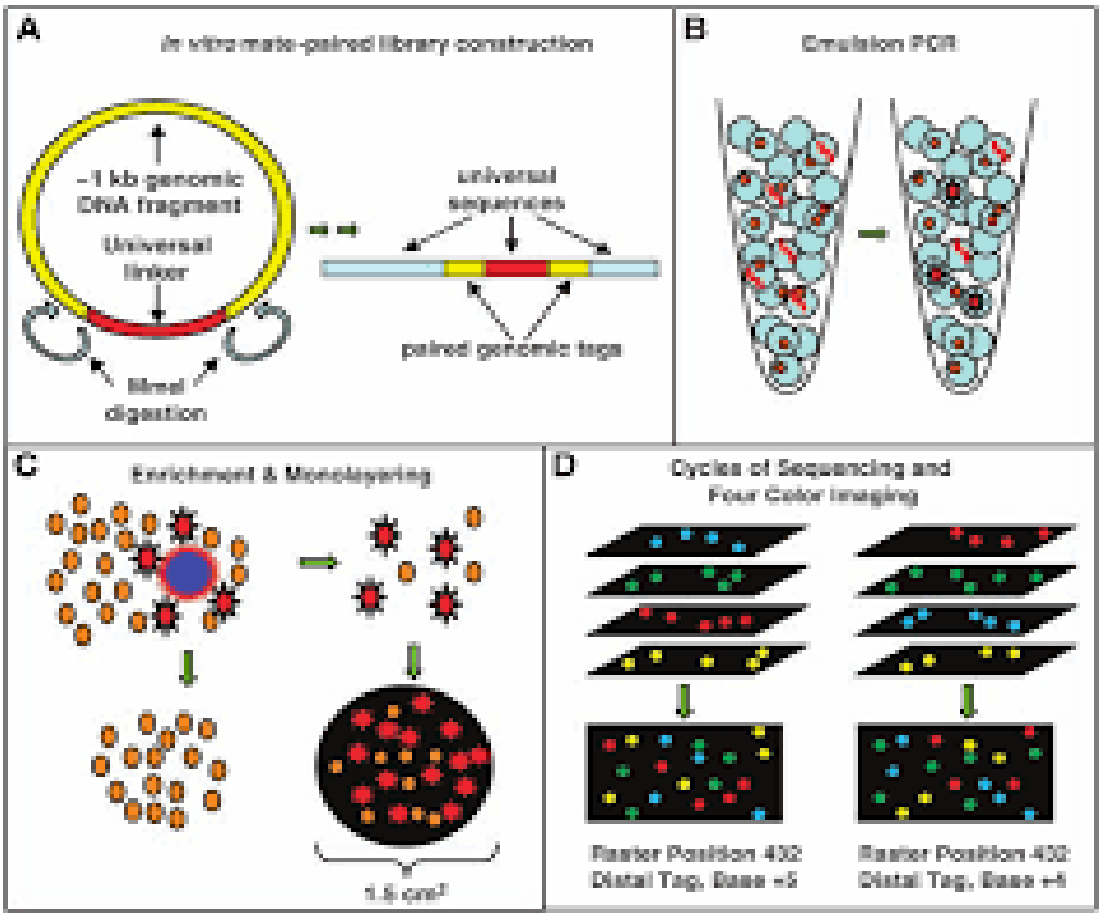


Figure S4: Polony Sequencing (28). 1) Formation of the library: DNA is sheared and repaired. Then circularized using the T-tailed 30 base pairs long synthetic oligonucleotides T30 which has a two restricition-recognition sites of Mme1. The circular DNA is then replicated by Rolling cycle amplification. The circular DNA templates are digested by Mme1 restriction enzyme. The DNA fragments are then repaired and primers added. 2) Clonal elongation by emulsion PCR (18) amplifies ssDNA in streptavidin-coated beads with complementary primers to the adaptors are emulsified in water in oil droplet with DNA polymerase, buffer, dNTPs; deoxynucleotides triphospahtes; denaturation, annealing and elongation. 3) Sequencing by ligation completes the sequencing process.

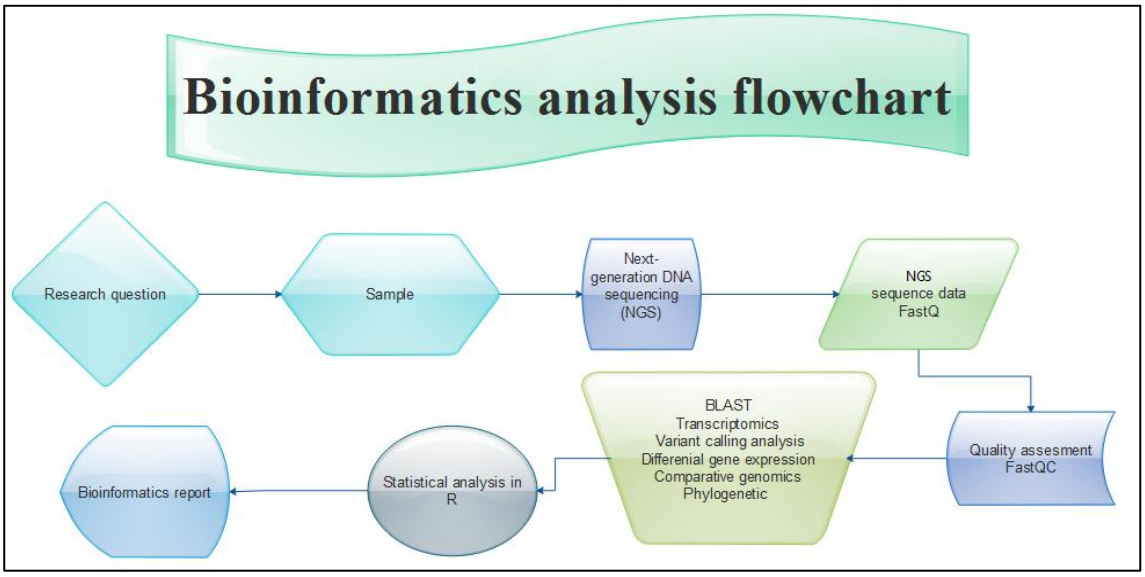


Figure S5: Bioinformatics analysis flowchart. A simple schematic flowchart theat illustrates the main steps in bioinformatics analysis. The research question is any proposed hypothesis that is to be studied by sequencing of any sample (blood, tissue) using next generation sequencing technology. Next generation sequencing platforms output is a raw sequence data file FASTQ that has the precise arrangement of the four nucleotide bases (Adenine A, Guanine G, Cytosine C, Thyamine T). The raw sequence data quality assessment using FASTQC generates a report about the quality of reads. DNA analysis software is a program that analyze the raw sequence data file (sequence alignment to a reference genome, sequence annotation, transcriptomic analysis that includes variant calling analysis tools and differential gene expression analysis, phylogenetics analysis). Statistical analysis in R of the DNA sequence analysis is a vital process for the acceptance or the rejection of the null hypothesis based on the research question. The bioinformatics report documents the overall results of the bioinformatics analysis process.

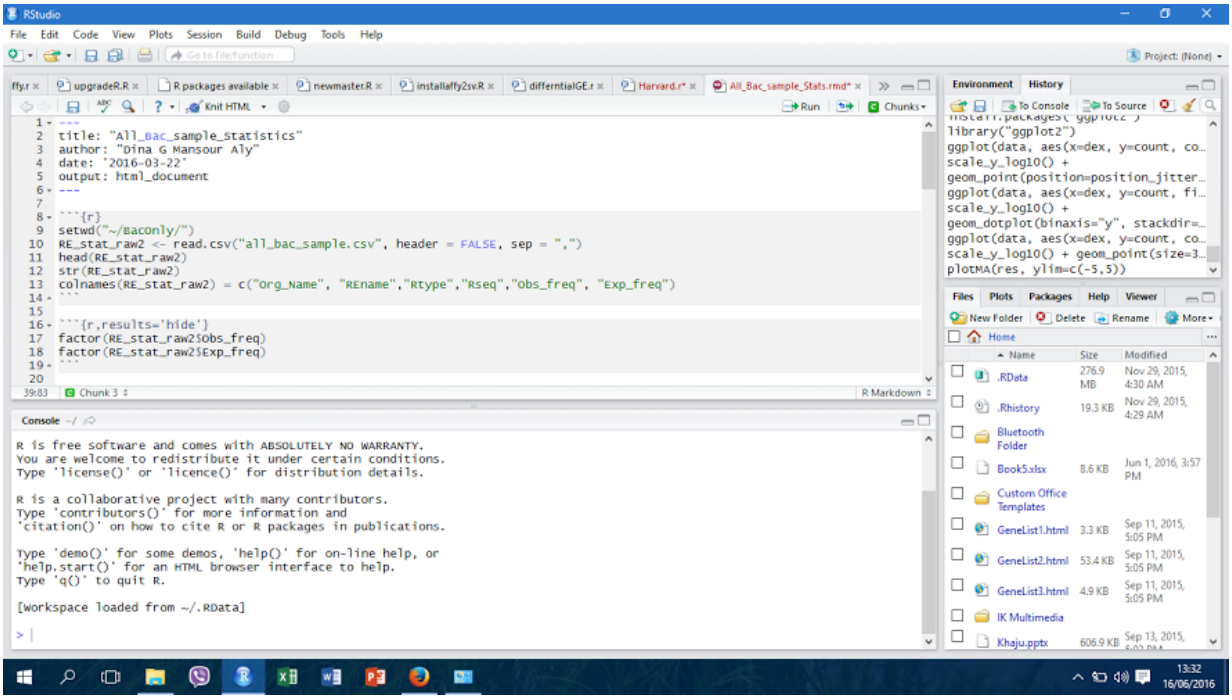
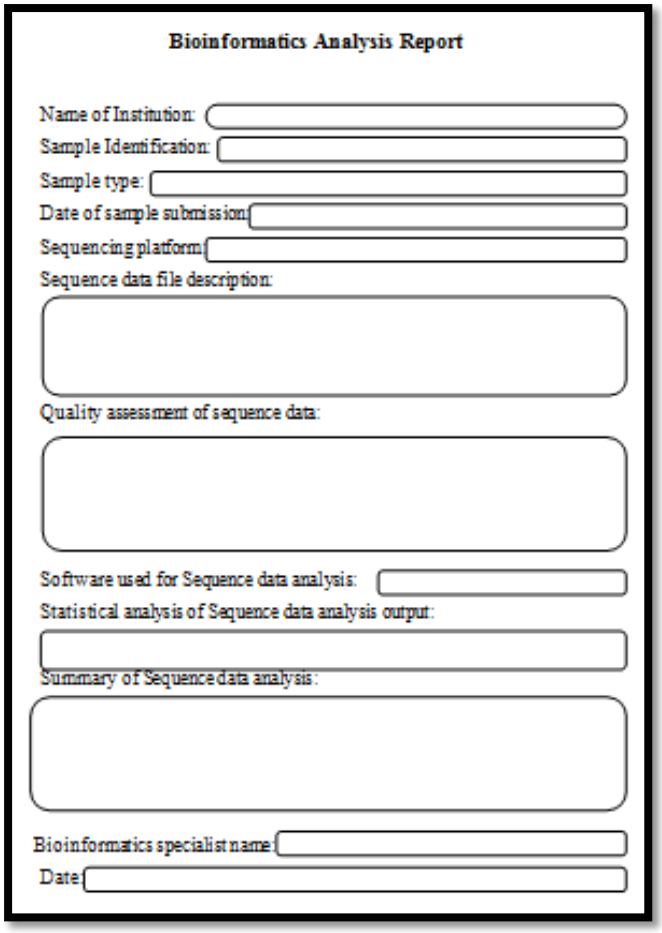


Figure S6: R-studio desktop. Rstudio is a open-source free user interface for R (22). R is a statistical software for statistical analysis. R has many packages that are very useful for sequence data analysis.


Bioinformatics Analysis Report

Name of Institution:
Sample Identification:
Sample type:
Date of sample submission:
Sequencing platform:
Sequence data file description:
Quality assessment of sequence data:
Software used for Sequence data analysis:
Statistical analysis of Sequence data analysis output:
Summary of Sequence data analysis:
Bioinformatics specialist name:
Date:

Figure S7: Bioinformatics report. This is a documentation of the bioinformatics analysis . A full summary of the name of the institution that supplied the sample for sequencing. The sample identification can be a number, or the name of the individual that supplied the sample (patient). This report can be for bioinformatic analysis of one sample or of many samples. The sample type refers to whether the sample submitted for sequencing is blood or tissue and the organism (blood sample of human patients). Date of sample submission is the date at the beginning of bioinformatics analysis. This is an important parameter to measure the overall duration of the analysis. Sequencing platform is the name of the sequencing platform. The description of the raw sequence data file usually has information about the sequence data file format (FASTQ) , the number of reads, the number of samples. Quality assessment of the sequence data is usually a summary of FASTQC report. Software name is the software used to analyse the sequence data; samtools, vcftools, ensemble tools. Statistical analysis package , the package/packages used for the statistical testing in R. Summary of the bioinformatic analysis is complete report of the results and its evidence-based interpretation supported by references from the scientific articles and review. This bioinformatic report has many applications depending on the research question at the beginning of the analysis: 1- Personalized medicine in medical settings. 2- Comparative genomics and evolution studies. 3- Population genetics and gene-association studies. 4- Pharmaceutical sector for novel target gene therapy.

Glossary

Bioinformatics

Interdisciplinary field of science where biochemical and biological data are analyzed using mathematics and computer science for the study of genomes.

Comparative Genomics

Comparing genome features of different related organisms or unrelated organisms in order to understand evolutionary relationships. Phylogenetics software can be used to construct phylogenetic tree that illustrates the evolutionary relationships between organisms.

Genome Assembly

The alignment of the sequenced DNA fragments to reform the whole genome of a certain organism.

Gene Annotation

The identification of genes and coding region in a DNA sequence (genome) and the genes functions.

Hierarchical Shot gun sequencing

Hierarchical Shot gun sequencing is used for sequencing of large DNA strands. First the genome is fragmented randomly and cloned into a suitable vector. A scaffold is formed from the random fragments. Then the vector that have the scaffolds are sheared to sequence length and sequenced

Library

This is the first step in next generation sequencing. DNA of a sample is fragmented and cloned in bacteria host that allow replication of the DNA fragments. This collection of cloned DNA is a library.

Next generation sequencing

The use of sequencing technology to read the precise order of the nucleotide bases , (Adenine, thiamine, cytosine, guanine). The main steps are DNA fragmentation and cloning; library formation, Clonal amplification and massive parallel sequencing.

Polymerase Chain Reaction (PCR)

Method uses heat for making multiple copies of fragments of DNA. PCR reagents include DNA polymerase enzyme, suitable Mg^{2+} buffer, dNTPs (deoxyribonucleotides triphosphate). Process has three steps: Denaturing, annealing and extending.

Sanger Sequencing

Sanger sequencing (Chain termination sequencing), This is the first sequencing method. Single-strand DNA attached to primer is elongated by one fluorochrome-

dideoxynucleotide base by DNA polymerase. The four dideoxynucleotide bases labelled by four different fluorochromes. This method is used to sequence short DNA strands (100-1000bps).

Sequence Alignment

The alignment of certain nucleotide sequence to a reference genome. Global Sequence Alignment (Needleman–Wunsch algorithm) and local sequence alignment.

Whole genome sequencing

Whole genome sequencing is used for sequencing of large DNA strands. First the genome is fragmented randomly to the sequencing size. Then they are sequences and assembled.