**MAJOR PROJECT REPORT - II**

**Study of Evolutionary patterns in Whole Genome Sequences: A Case Study of Bacillus Genomes**

Submitted

In partial fulfillment for the degree of

**MASTER OF TECHNOLOGY**

**IN**

**BIOINFORMATICS**



**DEPARTMENT OF BIOTECHNOLOGY**

**DELHI TECHNOLOGICAL UNIVERSITY**

**(*formerly* DELHI COLLEGE OF ENGINEERING)**

Supervised by:

**Dr. A Krishnamachari**

**Assistant Professor, SCIS, JNU**

&

**Dr. Yasha Hasija**

**Assistant Professor, DTU**

By

**RUCHIKA SAHAJPAL**

**14/BINF/2010**

**CERTIFICATE**

This is to certify that the dissertation entitled “**Study of Evolutionary patterns in Whole Genome Sequences: A Case Study of Bacillus Genomes’** hasbeen carried out by **Ms. Ruchika Sahajpal** at the Department of Biotechnology, Delhi Technological University, (formerly Delhi College of Engineering), Delhi as Major Project–II in partial fulfillment for the degree of **Master of technology in Bioinformatics from Delhi Technological University, (formerly Delhi College of Engineering),** Delhi. This is a record of Major Project-I work done under my supervision during the period from January 2012 to June 2012.

University Supervisor:

**Dr. Yasha Hasija**

Associate Head & Assistant Professor

Department of Biotechnology,

Delhi Technological University

**DECLARATION**

I hereby declare that the dissertation **‘Study of Evolutionary patterns in Whole Genome Sequences: A Case Study of Bacillus Genomes’** submitted in the partial fulfillment of the Master of Technology Degree in Bioinformatics is a record work done by me, during the period January, 2012- June, 2012 and has not formed the basis for the award of any Degree or other similar titles under my University in India or Abroad.

**RUCHIKA SAHAJPAL**

**(14/BINF/2010)**

**ACKNOWLEDGEMET**

First I “Praise to the god Almighty” for abundantly pouring his blessings and grace to finish this project successfully.

I want to express deep gratitude to acknowledge the immense help of all those who contributed with valuable suggestions and timely assistance to complete this work.

I am very grateful to my guide **Dr. A. Krishnamachari**, Assistant Professor, SCIS, JNU, for his valuable guidance throughout my project. It was mainly due to his effort that the current project made possible. I also express my indebtedness towards him for constantly inspiring me to complete my project.

Special thanks also go to School of Computational and Integrative Sciences (SCIS), Jawaharlal Nehru University for providing me all the facilities and support to carry out my project work and to **Dr. Naidu Subbarao**, for giving me an opportunity to work as a project trainee at SCIS, JNU.

The support, patience and encouragement rendered by family members are countless that shall be long remembered.

I am also very much thankful to all my friends for their timely support during my dissertation.

RUCHIKA SAHAJPAL

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| S. No | TOPIC | PAGE NO |
| 1 | Introduction | 1 |
|  | 1.1 Background | 2 |
|  | 1.2 Whole genome data analysis | 3 |
|  | 1.3 Computational Biology/ Bioinformatics | 5 |
|  | 1.4 Comparative Genomics | 6 |
|  | 1.5 Evolutionary Studies | 7 |
|  | 1.6 Phylogeny reconstruction | 9 |
|  | 1.7 Tree of life in the light of HGT | 10 |
| 2 | Objective | 13 |
| 3 | Materials & Methods | 14 |
|  | 3.1 Comparative Genome Wide Study | 15 |
|  | 3.2 Genome Size Comparison: | 16 |
|  | 3.3 GC Distribution | 16 |
|  | 3.4 Whole Genome Pair Wise Sequence Alignment: MUMmer 3.0 | 17 |
|  | 3.5 Whole Genome Multiple Sequence Alignment: MAUVE | 18 |
|  | 3.6 Genome Signature – HGT Detection | 18 |
|  | 3.7 Single Gene Tree Construction | 19 |
| 4 | Results & Discussion | 23 |
|  | 4.1 Genome Size Comparison: | 24 |
|  | 4.2 GC Distribution | 25 |
|  | 4.3 Whole Genome Pair Wise Sequence Alignment: MUMmer 3.0 | 29 |
|  | 4.4 Whole Genome Multiple Sequence Alignment: MAUVE | 30 |
|  | 4.5 Genome Signature – HGT Detection | 33 |
|  | 4.6Single Gene Tree Construction | 37 |
| 6 | Conclusion &Future Perspective | 42 |
| 7 | References | 45 |
| 8 | Appendix | 47 |
|  | Appendix I | 50 |
|  | Appendix II | 51 |

LIST OF FIGURES

|  |  |  |
| --- | --- | --- |
| FIGURE NO | TITLE | PAGE NO |
| 1 | Flowchart for single gene tree construction | 20 |
| 2 | Flowchart for whole genome phylogenic analysis | 21 |
| 3 | Graph between Genome size (y axis) and the 10 genome sequences of interest (x axis) | 24 |
| 4 | GC composition of 10 bacillus genomes | 26 |
| 5 | Comparison of the traditional 16s rRNA based phylogenetic tree (above) and pairwise sequence whole genome alignment results from MUMmer (below). | 30 |
| 6 | Whole genome alognment output from MAUVE | 31 |
| 7 | Phylogram obtained from whole genome alignment from MAUVE | 32 |
| 8 | Cladogram obtained from whole genome alignment from MAUVE | 32 |
| 9 | Workflow result for the identification of probable HGTs from GOHTAM. | 34 |
| 10 | Phylogenetic tree based on signature sequence (tetranucleotide/codon usage) of 10 genomes from GOHTAM. | 37 |
| 11 | 16s rRNA based phylogenetic tree | 38 |
| 12 | 23s rRNA based phylogenetic tree obtained from PHYLIP | 39 |
| 13 | Workflow result from phylogenomic analysis of bacillus pumilus genome. | 40 |

LIST OF TABLES

|  |  |  |
| --- | --- | --- |
| TABLE NO | TITLE | PAGE NO |
| 1 | Genome Sequences of interest and their corresponding NCBI Accession IDs | 15 |
|  | Clusters of bacillus species on the basis of genome size. | 25 |
| 2 | Algorithms implemented in GOHTAM and predicted Horizontally transferred genes in Bacillus pumilus genome. | 33 |
| 3 | Atypical regions in Bacillus pumilus obtained via GOHTAM. | 34 |
| 4 | Probable horizontally transferred genes in Bacillus pumilus obtained via GOHTAM | 36 |

**INTRODUCTION**

**1.1 BACKGROUND**

With rapid advancement in data storage and processing technologies, it is not possible to analyze huge amount of data and make it meaningful by sophisticated visualization tools. The IT revolution that is taking place now has a tremendous impact in research and determining activities in a big way[1]. It is not possible to undertake large scale problems and get solution within a timeframe. Needless to mention next generation sequencing technologies produce large amount of raw genomic data that needs to be analyzed. Thus IT tools have become an important part of biological research. It is found to be very useful in deciphering enormous information in the length of genomic sequence.

With the rapid advancements in data capturing technologies with advanced experimental and computational tools, it is possible now to obtain the whole genome data of any organism within a short period. Sequencing era started with the method proposed by Sanger. This approach made a great impact in biological science to study disease related proteins in particular. The race was on thereafter to come up with better and faster ways of sequencing the DNA. [3]

The raw one- dimensional genomic sequences, which are nothing but long character strings, can be processed or analyzed using existing powerful algorithms from the field of computational sciences. Bioinformatics/ Computational Biology, a field which deeply focuses on application of IT tools and techniques to sequence structure expression and other types of data is increasing getting importance in biomedical field. Generally the following types of analysis are carried out routinely using bioinformatics tools: Compositional Analysis, pattern searching, alignment between two sequences and group of sequences, alignment between a sequence and database, evolutionary analysis, protein structure prediction (secondary and tertiary), screening potentially large molecules for drugs, comparing x-ray structures of proteins, carrying out large scale simulations of biomolecules and many more.

**1.2 Whole Genome Data Analysis**

Encouraged by the success and development of DNA sequencing techniques researchers dreamt of whole genome sequencing. Two strategies were adopted i.e BAC sequencing [3] and shotgun method [4]. In BAC sequencing, genomic DNA is cut into fragment of about 150 Mb and is inserted into BAC vectors and later transformed in E.coli for further replication. The BAC inserts are then isolated and mapped to determine the order of each cloned fragment. Shotgun sequencing shears genomic DNA into small pieces randomly and then cloned the pieces into plasmids and sequence on both strands, thus eliminating the BAC step from the former approach. Nevertheless, both strategies make use of Sanger method of sequencing. Sanger method was costly, time consuming and labor intensive [5, 23]. So, there was high demand for cheaper and faster method of sequencing. This demand was fulfilled by the development of high throughput sequencing technologies, called Next Generation Sequencing.

European Molecular Biology Laboratory (EMBL)[10] initiated the first full-fledged DNA sequence database in 1982, and shortly Gen Bank[11] was created by the National Institute of Health. National Center for Biotechnology Information (NCBI) is a part of National Library of Medicine at the National Institute of Health (NIH) USA, while EMBL was then based in Heidelberg, Germany. Both database compile data and information that is published in print journals that published in print journals to an electronic format and make it available for the research community. Another famous database joined the data-collecting collaboration a few years later was DDBJ [12] that is The DNA data bank of Japan. All these three made collaboration in 1988 referred as International Nucleotide Sequence database collaboration with an agreement to use a common format for data elements for each record and they update their data base when the records were directly submitting to it.

It is well known that the prokaryotic genome organization is comparatively simpler than eukaryotes. An intrinsic order and correlation exists in the long chain of nucleotides and they also exhibit many statistical regularities. Many biological features which are embedded in the whole genome sequences can be identified using appropriate algorithms and strategies. A typical genome sequence data may have genes or coding segnments, splica sites, untranscribed regions like non-coding segments, many protein and transcriptional binding sites, repeats etc. it is not possible to carry out experimental procedures to identify these features in the genomes as it may take time, energy and cost. Hence, biologists rely on computational methods to try and identify these features with the help of known information from the model organisms. *E-coli, Saccharomyces cerevase, Bacillus subtilis, Drosophila melanogaster, Arabdopsis* etc are examples of typical model organisms and which are living proofs to identify such biologically important features. Once these features are identified, their use for drug development and disease prevention may be taken up by experimentalist. St it is evidently clear that the wealth of information a genome sequence can provide can be exploited using bioinformatics tools.

**1.3 Computational Biology/ Bioinformatics**

Sequence databases led to a new branch of science called Bioinformatics which can be defined as the application of information technology to the management of biological data. Bioinformatics is a highly inter-disciplinary subject area and the Standard definition provided by NIH (USA) is following:

“Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.”[13]

The complete annotations of the genomes are carried out by *in silico* methods. This approach has replaced the prevailing cumbersome, costly and time consuming wet lab experiments to a large extent. Bioinformatic approaches which employ sophisticated mathematical, statistical and computational tools captures the biological signals or features that is hidden in the gene or genome sequences. Algorithms which are well established in other fields are being applied or used on sequence and structure data to infer biological functions. Bioinformatics tools also make use of the existing information and enable us to build prediction models.

New fields are emerging focusing on a specific area with large scale data analysis. They include genomics, metabolomics, transcriptomics, glycomics etc. With the powerful computational methods it is now possible to get a macroscopic picture of the biological process in question. Bioinformatics approaches have a major role in important field called Functional genomics and Comparative genomics.

Functional genomics: It aims to determine the role of the sequences of the living cell either as transcribed and translated unit or as regulator motif. In a very simple way if we view DNA sequence as a message, then functional genomics is concerned with the meaning of the message. This has led to the experimental analysis of RNA transcripts ( transcriptome ) and the repertoire of expressed protein ( the proteome), each of which presents fresh informatics challenges . The main goal is to make link or connections between biological process or key molecules and go beyond what is encoded in the nucleotide sequences.

**1.4 Comparative Genomics**

This part of bioinformatics has an objective to trace the ancestries and make correlation with disease or even elucidating function by comparing sequences from different organisms or even from different individuals. Comparative genomics make use of the capabilities of both alignment and genome comparisons tools. Every alignment is based on a collinear arrangement of sequence similarities and all alignment tools Wassume this collinearity [2]. Alignment helps to study coding and non –coding regions from different species.

Several comparative genome analysis uses alignments from the corresponding non -coding and coding regions across different species [2,14,15]. Identification coding and non-coding region of a species through cross -species comparative genomics greatly depends on the evolutionary distances of the species. A synteny map i.e. gene order arrangement between two or more genome sequences can give clue to the possible evolutionary events.

The popular technique known as phylogenetic footprinting which uses alignment based comparison to annotate conserved functional elements [16]. Comparative analysis can also be done on phylogenetically diverse organism as well as on very close species. If analysis is done on phylogenetically distant genomes, it may give clues about selective pressures governing genes / operon clustering and horizontal gene transfer (HGT), and mechanism of evolution. We may also understand the common mechanisms of virulence antibiotic resistance and host range determination.

Genome comparisons of closely related species can be used to establish the genetic basis for phenotypic variation and may provide some insights. Specific regions which can be regarded as signature of genome can be used for species identification. Genome comparison analysis of closely related species is currently used for precise strain determination.

Comparative study has been done among various species especially in prokaryotes and our study will focus specifically on bacilli family; *Firmicutess* Division, *Bacilli* Class, *Bacillales* Order, *Bacillaceae* Famliy, *Bacillus* Genus. Bacillus genus is an extensive heterogeneous group comprising 83 validly described species to date (http://www.bacterio.cict.fr/b/bacillus.html) [17]. Many species in this taxon are of major clinical importance e.g Bacillus cereus group. Bacillus anthracis cause anthrax in human and bacillus thuringiensis whose genes are used in BT cotton to produce toxin against insects. Few species in this genus are also used for industrial production of various enzymes, antibiotics, etc. Thus, this group is of economic importance. Currently genome data of ten species of bacillus and total of twenty five strains are available in public domain (i.e http:// www.ncbi.nlm.nih).

**1.5 Evolutionary Studies**

The importance of knowing the phylogenetic history of all living organisms has become increasingly widely acknowledged. Phylogenies are now behind applied to diverse fields, including subjects like immunology, epidemiology and conservation. Also, many new areas such as evolutionary development and evolutionary ecology have started to make extensive use of this study. Identifying and interpreting phylogenies is therefore now an important component of all of the biological sciences. As explained by Lecointre and Le Guyader in 1993 [1]

*‘Phylogenies form the framework within which we can best arrange our knowledge of all aspects of biology*’

The most significant application of phylogeny is explaining and predicting organismal features. Phylogenetic classification scheme organizes our knowledge in a way that maximizes information content by being both descriptive and predictive. There are many other important uses of phylogenies including the study of co-phylogeny of hosts and pathogens - understanding the role of hosts in pathogen evolution) and biogeography - understanding the origins and spread of species. For example, the phylogenetic tree can be compared to the geographic locations of the samples in order to investigate the spread of a disease; or, so-called “molecular clocks” can be applied to estimate the age of important events in the origin and spread of new pathogens. [18]

Darwin was the first to recognize that living species are not independently created, but have been generated through descent with modification from ancestral species [3]. Darwin further envisioned that the propinquity of descent among species could be depicted in the form of a Tree of Life (TOL). Almost 150 years later, one of the major goals in biological research is to convert Darwin’s monumental vision into reality by assembling the complete TOL.[4] TOL promises to deepen our understanding of the history of life and sheds light on the evolution of molecules, phenotypes, and developmental mechanisms [19], as well as directly impact the research on key areas such as human health, agriculture, and biodiversity [5].

Phylogenetic analysis attempts to group organisms on the basis of their common ancestry. This is founded on a widely held view of the mode of the evolutionary process: *species are lineages undergoing divergent evolution with modification of their intrinsic attributes, the attributes being transformed through time from ancestral to derived states*. [6] Phylogeneticists usually interpret both the individual tree and the species tree as proceeding from the past towards the present. The species tree is then interpreted as proceeding from a single common ancestor, through time via parents to children, with branching due to speciation events.

Advances in three disciplines - statistical phylogenetics, information technology, and molecular biology and genomics, have led to enormous effort in the prospect of assembling the TOL [6, 20]. The advances in genomics are most important. Since the sequencing of the first prokaryote [7] and eukaryote [8] genomes, more than 300 prokaryote and 40 eukaryote genomes have been decoded, whereas several more scheduled for completion [9]. Thus enormous data is generated to be analyzed. The impact of this humongous data flow has been large, for example, several recent studies have featured data matrices consisting of tens to hundreds of genes [10-13].

For many years, the molecules of greatest interest were ribosomal RNA (rRNA). This is because they perform the same function in every cell, and so changes in sequences are not biased by differing functional selection in different taxa. Also, rRNA has domains that evolve at very different speeds. This occurs because the structure of rRNA is maintained by base pairing. The base paired regions evolve extremely slowly because a single base change will disrupt base pairing, and this disrupts the function of the molecule. Therefore only very rare simultaneous double mutations that do not disrupt base pairing are likely to be fixed by evolution. Conversely, the regions of rRNA molecules that are not base-paired evolve much more rapidly. Therefore, a comparison of rRNA sequences can reveal evolutionary relationships over greatly different time scales.

Until now the phylogenetic analysis was based on the available genes from an organism. However the presence of reticulate events such as horizontal gene transfer, hybridization, duplication, and speciation and recombination largely affect gene based phylogenetic analysis. However genome-scale phylogenetics has the potential power to overcome this incongruence observed in gene-scale phylogenetics. The availability of such larger amounts of data has enabled researchers to increase the analytical power of their studies, shedding light on key clades of the TOL. Genome-scale phylogenetics has also enriched the gene-scale perspective of evolution at the molecular level. Notable examples include several experimental demonstrations of the occurrence and impact of lineage sorting in phylogenies of closely related species [14], including our own evolutionary branch, the primates [15]; the discovery of the major impact of horizontal gene transfer in prokaryotic evolution [16]; the presence of incongruence in gene-scale phylogenetics and its severity [17]; and the demonstration that, given genome-scale data, the lack of phylogenetic resolution may be the signature of closely spaced series of cladogenetic events [10].

One may infer that the problems encountered in gene-scale phylogenetics—such as incongruence and poor resolution—may be overcome simply by scaling up the amount of data utilized. However, the acquisition of genome-scale data—even if analyzed by ‘state-of-the-art methodology' does not guarantee that the resulting phylogenetic inference is correct for three reasons: (i) the existence of short stems in the TOL, (ii) the substitutional saturation of phylogenetic signal from very ancient clades and (iii) the effect of systematic errors.

**1.6 Phylogeny reconstruction**

Reconstructing a tree-like phylogenetic history is conceptually straightforward. The objective is to infer the ancestors of the contemporary organisms, and the ancestors of those ancestors, etc., all the way back to the ‘most recent common ancestor’ of the group of organisms being studied. Ancestors can be determined because the organisms share unique characteristics. That is, they have features that they hold in common and that are not possessed by any other organisms, only inherited from the ancestor to the organism. The ancestor acquired a set of inheritable, genetically controlled characteristics, and passed those characteristics on to its offspring. We observe the offspring, note their shared characteristics, and thus infer the existence of the unobserved ancestor. There are basically three ways to deal with conflicting evidence when reconstructing an evolutionary history:

1. Select the largest set of characters that are consistent with some tree (traditional approach);

2. Select the largest set of taxa for which the character data are compatible;

3. Construct a network rather than a tree, thus accommodating the incompatibilities.

The purpose of phylogenetics is to create a diagram representing character differences among a set of taxa, in a way that reveals the evolutionary history. Indeed, the phylogenetic inference is now being widespread based on datasets of multiple unlinked gene sequences [18]. Phylogenetic analysis of molecular sequences usually consists of three distinct procedures: (i) sequence alignment; (ii) character coding; and (iii) tree building. These steps are performed in this order, and all three of them need to be fully described for a phylogenetic analysis to be reproducible. However there are many known artifacts potentially associated with each of these procedures, thus they need to be performed cautiously.

Alignment is the process of establishing the possible homology relationships among the sequence residues [19,20]. Homology refers to the relationships of features that are shared among taxa due to common ancestry i.e. they all inherited the feature from their most recent common ancestor. Character coding [21,22] is often overlooked as an important step in sequence analyses. The parts of the sequence alignment involving length variation such as gaps are sometimes considered to be uncertainly aligned, and are treated as missing data by some computer programs. These regions are often excluded from the tree-building analysis. Thus the phylogenetic information is lost. This issue is dealt by coding the length-variable regions as a set of independent characters, which are then included in the tree-building analysis. Tree building is the third step of a phylogenetic analysis, and it displays the information obtained from the sequence alignment and coding steps in the form of a branching diagram [23]. That is, conceptually all it changes the tabular data, the alignment, in the form of a tree.

**1.7 Tree of life in the light of HGT**

Horizontal gene transfer complicates the reconstruction of life’s history. Via phylogenetic analysis we may be able to trace an individual gene, or a highly coadapted set of genes and their products like the ribosome, to a molecular ancestor. However, these genes are likely to have been transferred between lineages at least a few times. While successful transfers occur most frequently between closely related organisms, [24] these transfers are sufficient to cause molecular ancestors to have existed in different lineages and at different times. [25,26]. Gene transfer between different lineages provides challenges and opportunities. It provides a challenge, because the reconstruction of phylogenetic networks from molecular data is even more complex than the reconstruction of phylogenetic trees. It provides opportunities, because genes transferred horizontally between divergent organisms allow the reconstruction of the evolution of metabolic pathways [27,29], allow the correlation of evolutionary events occurring in different domains of life, [30] and can provide a shared derived character for the recipient and its descendants. [31-34]

It is becoming increasingly apparent that many genes within eukaryotes and prokaryotes have been acquired by horizontal transfer, but not all genes are equally likely to be transferred [35-43]. The preferential horizontal transfer of genes in both eukaryotes and prokaryotes is strongly correlated with gene function. Specifically, genes participating in transcription, translation, and related processes (informational genes) are far less likely to be horizontally transferred than genes participating in housekeeping functions (operational genes) [43]. Furthermore, the frequency of horizontal transfer in prokaryotes is not related to evolutionary rates, nucleotide substitution rates, because evolutionary rates for operational and informational genes have not differed significantly since the cyanobacteria and proteobacteria diverged [43].

There are several approaches to detecting HGT, all of which are best used in conjunction with each other. Most HGT computational prediction methods can be roughly grouped into two main categories: compositional methods, which identify anomalous sequence signatures within a prokaryotic genome suggestive of a region of HGT, and phylogenetic methods, which analyze the incongruence of a gene tree versus its associated species tree.

Sequence composition methods: Sequence composition methods depend on different species having differences in their genome signatures. These methods identify HGT by searching for genomic regions that have an abnormal sequence composition (G+C, dinucleotide bias, and so on) compared to the rest of the genome. Atypical base composition of a particular gene, for example a GC content that differs significantly from that of most other genes in the genome, provides another piece of evidence that HGT may have recently taken place and this approach has been used to detect host-derived regions in poxviruses and other small genomes. This signal, however, is expected to fade over time and therefore this approach is only useful for detecting recent HGT. An in-depth study of HGT in the Salmonella lineage indicated that ancient horizontally transferred gene sequences tended to share a greater similarity in sequence composition with their host compared to more recently acquired genes [44], clearly supporting the idea that transferred genes ameliorate to their host genome over time [45]. However, very recently acquired prophage elements tended to have sequence compositions that were more similar to the host genome, not representing amelioration but rather specialization and adaptation to their hosts [44]. Although this study may suggest that more sensitive measures of sequence composition are needed to better predict HGT events, these methods must be carefully designed so that they do not result in an increase in false positives.

Wealth of information generated from our analysis may provide information on a genomic scale which may help us to understand certain biological features/functions. Bacillus genomes are clustered into ten groups[49] based on analysis of 16S RNA ,16S-23S RNA etc. However classification based on partial sequence information is not desirable. It is natural to maximally use the available information to infer classification among organisms. In that direction, we are planning to carry out our full blown analysis using the available full length genomic sequences of these ten genomes and see whether any distinct sub-group or classes emerges from the bacillus family. We are also interested in finding species specific characteristics based on genomic data.

**OBJECTIVE**

Given the whole genome sequences of a family we tried to answer the following questions

1. Whether they cluster as one or do they exhibit sub clusters (or) groups.
2. If so, then carrying full blown genome analysis may reveal the information about possible evolutionary drift or events?
3. Can the genome trees provide a better insight to the evolutionary patterns than gene trees
4. Can we obtain any meaningful information regarding evolutionary events such as horizontal gene transfer which can enlighten possible evolutionary path

**MATERIALS**

**&**

**METHOD**

**STUDY I: COMPARATIVE GENOMICS OF BACILLUS GENOMES**

Comparative study has an objective to trace the ancestries and make correlation with disease or even elucidating function by comparing sequences from different organisms or even from different individuals. Comparative genomics make use of the capabilities of both alignment and genome comparisons tools. Genome comparisons of closely related species can be used to establish the genetic basis for phenotypic variation and may provide some insights. Specific regions which can be regarded as signature of genome can be used for species identification. Genome comparison analysis of closely related species is currently used for precise strain determination. Our major objective of the study was to analyze the available whole genome sequences of ten bacillus genomes using genome size, base composition, patterns, pairwise and multiple whole genome alignment etc.

**3.1 Comparative Genome Wide Study**

The genome data can be easily downloaded from the major sequence databases i.e NCBI, DDBJ and EMBL. We have obtained the data primarily from NCBI and details are furnished below. We selected whole genome sequences of 10 Bacilli strain; *Firmicutess* Division, *Bacilli* Class, *Bacillales* Order, *Bacillaceae* Famliy, *Bacillus* Genus from NCBI Genome database. For each species the genome sequences of a particular strain has been considered. Selection of a strain for particular species is based on the length of the chromosome and the strain having longest chromosome was selected for full blown analysis. List of the representative strains with their accession number is tabulated below:

|  |  |  |
| --- | --- | --- |
| **Accession Id** | **Species** | **Strain** |
| NC\_009725.1 | *Bacillus amyloliquefaciens* | FZB42 |
| NC\_005945.1 | *Bacillus anthracis* | str. Sterne |
| NC\_011725.1 | *Bacillus cereus* | B4264 |
| NC\_006582.1 | *Bacillus clausii* | KSM-K16 |
| NC\_002570.2 | *Bacillus halodurans* | C-125 |
| NC\_006322.1 | *Bacillus licheniformis* | ATCC 14580 |
| NC\_009848.1 | *Bacillus pumilus* | SAFR-032 |
| NC\_000964.3 | *Bacillus subtilis* | subsp. Subtilis str. 168 |
| NC\_008600.1 | *Bacillus thuringiensis* | str Al Hakam |
| NC\_010184.1 | *Bacillus weihenstephanensis* | KBAB4 |

**TABLE 1:** The complete genome sequences of the 10 bacillus species; specific strains were downloaded from NCBI genome server (<http://www.necbi.nlm.nig.gov> ) and their respective Accession Ids

Sequence data is available in public domains in different formats. All formats have more or less identical names but with different extensions to highlight the contents and types of data. For our purpose whole genome sequences were obtained in Fasta format (‘.fna’, ‘.faa’, ‘.ffn’) and Genbank format (‘.gbk’). Apart from the representative strain data of the bacillus family, we have also considered the whole genome sequence of a model species- E.coli genome [accession ID ‘AC\_000091.1’], using it as on outgroup for supporting the analysis.

**3.2 Genome Size Comparison:**

The genome is the entirety of an organism's hereditary information. Genome size is the total number of DNA base pairs in one copy of a haploid genome. The genome size is positively correlated with the morphological complexity among prokaryotes and lower eukaryotes. It is typically measured in terms of mass in pictograms (pg) (trillionths (10−12) of a gram, or less frequently in Daltons (Da) or as the total number of nucleotide base pairs typically in Megabases (millions of base pairs, Mb or Mbp). One picogram equals 978 megabases. Genome size comparison of all the 10 Bacillus species was performed in our study to gain insight to the data.

Implementation/Run:

The sequence lengths of all the 10 genomes were obtained from the respective NCBI Genbank file and a graph plot between the 10 genome sizes (y axis) and 10 genome sequences (x axis) was constructed using Microsoft excel.

**3.4 Whole Genome Pair Wise Sequence Alignment: MUMmer 3.0**

Genome sequence alignments form the basis of much research. Genome alignment depends on various mundane but critical choices, such as how to mask repeats and which score parameters to use. Genome alignment algorithms are often described as glocal; that is, they try to maximize local alignment while trying to include the start/end of one of the pairs. The benefit is that one can visualize large "blocks" of genome structure such as synteny or large-scale rearrangements: duplication, deletion, inversion.

**MUMmer**

MUMmer [51-53] is an open source bioinformatics software system for pairwise sequence alignment of whole genomes based on the suffix tree data structure for efficient pattern matching. It takes first input sequence as the reference sequence and other one as the query sequence. It then locates the MUMs or Maximal Unique Matches between the reference and query sequences. MUMmer generated a list of exact matches which are displayed in the form of a dot plot or use these exact matches as alignment anchors to generate pair-wise alignments similar to BLAST output, using a seed and extend strategy. It is a very powerful alignment and visualization tool for whole genome pairwise alignment.

Implementation/Run:

Mummer 3.0 is downloaded from <http://sourceforge.net/projects/mummer/> and installed on the LINUX system. Mummer program was run on terminal where one reference and one query sequence given as input. Files of organism with“.fna” extensions are used for the input purpose. Mums are calculated with option - l=25 which limits reported mums of minimum length of 25 nucleotides in output. Option -c was used to report the reverse complemented query positions relative to the reverse of the query sequence in output. Output comes as stdout which is saved in the same directory and mummerplot program run on the saved Output which result in various files with different extension. Files with extension “.gp” is used as input to GNUplot to generate the graphical display of matching and non-matching region between two sequence data .The whole process is iterated to generate alignments between every pairs of genome and results are saved as graphs in postscript files.

* 1. **Whole Genome Multiple Sequence Alignment: MAUVE**

Mauve is a system for efficiently constructing multiple genome alignments in the presence of large-scale evolutionary events such as rearrangement and inversion. It uses the anchored alignment technique to rapidly align genomes and allows the order of alignment anchors to be rearranged in each genome permitting identification of genome rearrangements. Mauve finds the inexact matches using a seed-and-extend method, where each seed match conforms to a pattern of matching nucleotides, thus producing an ungapped alignment. It then applies minimum weight criteria to remove random matches. The weight of a Locally Collinear Blocks (LCB) is defined as the sum of the lengths of matches in that LCB. Mauve removes matches composing low-weight LCBs from the set of alignment anchors before completing the alignment. A finalized gapped global alignments is then completed by applying ClustalW ClustalW progressive global alignment algorithm to each LCB

Implementation/Run:

Mauve Genome Alignment Software is downloaded from <http://asap.ahabs.wisc.edu/> and installed on windows system. All the 10 genomes are uploaded in Genbank (.gbk) format in the mauve window and were run for progressive alignment algorithm. Mauve automatically selects the first genome as the reference genome. However this can be later changed accordingly. The alignment display is organized into one horizontal "panel" per input genome sequence. Each genome's panel contains the name of the genome sequence, a scale showing the sequence coordinates for that genome, and a single black horizontal center line. Colored block outlines appear above and possibly below the center line. Each of these block outlines surrounds a region of the genome sequence that aligned to part of another genome, and is presumably homologous and internally free from genomic rearrangement. Main output files .mauve and .alignment are formed which contain the alignment.

**3.6 Genome Signature – HGT Detection**

A genome signature or pattern is a characteristic that may be different for different organism while being pervasive in a genome. It generally shows the statistical characteristic of a given genome. Genome signature must satisfy two conditions: The pervasiveness and differentiating ability. Horizontally transferred genes can be detected in a genome based on the genomic signature analysis. It allows identifying genomics regions and/or genes that exhibit atypical features compared to the rest of the sequence. GOHTAM is an online tool. This website allows the detection of horizontal transfers based on a combination of parametric methods - window-based signature method (tetrancleotide) and codon usage method. It proposes an origin by researching neighbors in a bank of genomic signatures. This bank is also used to research an origin to DNA fragments from metagenomics studies

Implementation/Run:

Accessing GOHTAM online at <http://gohtam.rpbs.univ-paris-diderot.fr/>. Uploading Bacillus pumilus whole genome in genbank format and using combination of both the implemented algorithms. All the atypical regions in the input genome are listed along with all the atypical genes. Each detected region signature is compared with the signatures of the bank and the 10 closest neighbors are displayed with a confidence rating depending on the length of both query and reference sequences and the distance between the two signatures.

**STUDY II: PHYLOGENOMIC STUDY OF BACILLUS GENOMES**

Phylogenetic classification scheme organizes our knowledge in a way that maximizes information content by being both descriptive and predictive. Until now the phylogenetic analysis was based on the available genes from an organism. However the presence of reticulate events such as horizontal gene transfer, hybridization, duplication, and speciation and recombination largely affect gene based phylogenetic analysis. However genome-scale phylogenetics has the potential power to overcome this incongruence observed in gene-scale phylogenetics.

**3.7 Single Gene Tree Construction**

Ribosomal Tree of Life: For many years, the molecules of greatest interest were ribosomal RNA (rRNA). This is because they perform the same function in every cell, and so changes in sequences are not biased by differing functional selection in different taxa. Also, rRNA has domains that evolve at very different speeds. This occurs because the structure of rRNA is maintained by base pairing. The base paired regions evolve extremely slowly because a single base change will disrupt base pairing, and this disrupts the function of the molecule. Therefore only very rare simultaneous double mutations that do not disrupt base pairing are likely to be fixed by evolution. Conversely, the regions of rRNA molecules that are not base-paired evolve much more rapidly. Therefore, a comparison of rRNA sequences can reveal evolutionary relationships over greatly different time scales.

Implementation/Run:

Ribosomal RNA gene sequences (16s rRNA, 23s rRNA) were obtained for all the 10 bacillus genomes of interest from NCBI. Phylogeny.fr platform (<http://www.phylogeny.fr/>) was used to create gene trees. It is a free, simple to use web service dedicated to reconstructing and analysing phylogenetic relationships between molecular sequences. Phylogeny.fr runs and connects various bioinformatics programs to reconstruct a robust phylogenetic tree from a set of sequences. The output tree can be saved in various formats- nexus, netwick etc. The tree formed can be visualized using various integrated online visualization tools.

**Figure 1:** Flowchart for single gene tree construction

**RESULTS**

**&**

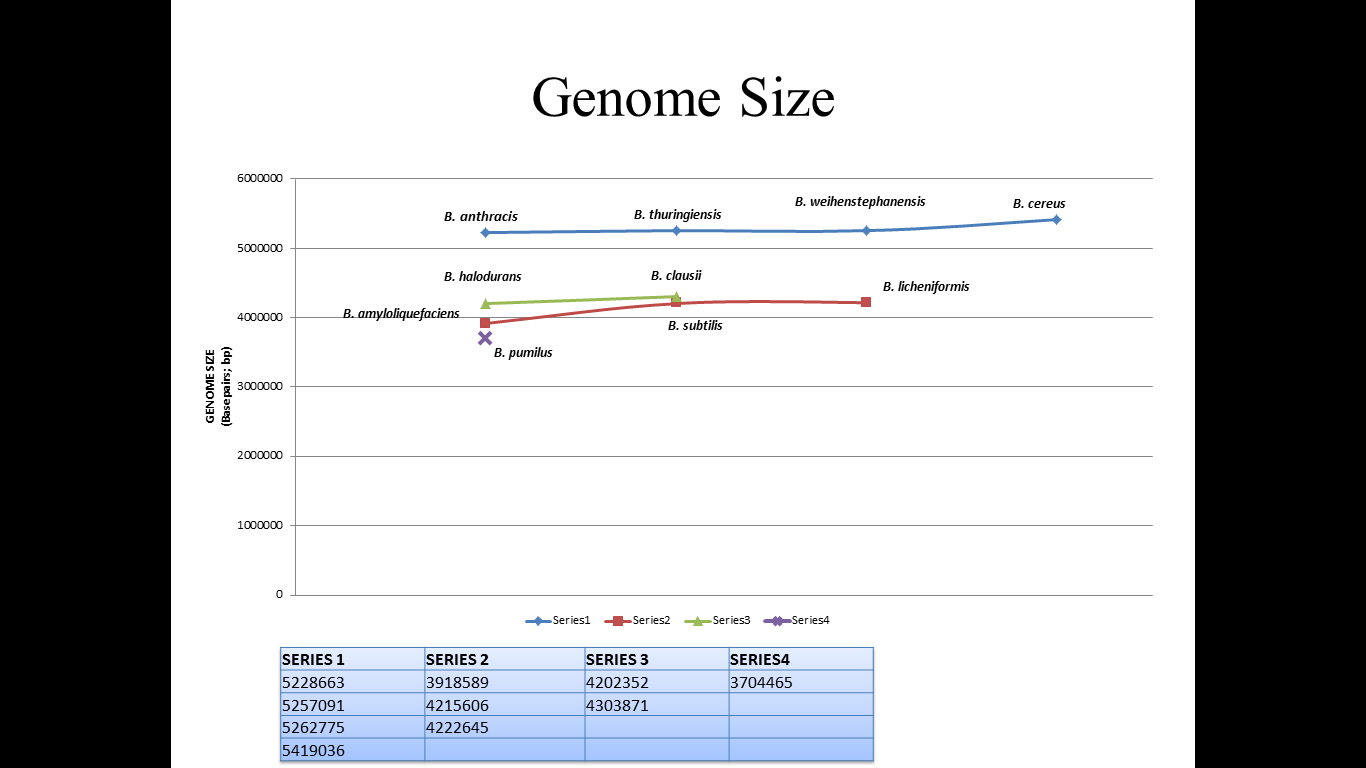
**DISCUSSION**

The raw one- dimensional genomic sequences, which are nothing but long character strings, can be processed or analyzed using existing powerful algorithms from the field of computational sciences. Bioinformatics/ Computational Biology, a field which deeply focuses on application of IT tools and techniques to sequence structure expression and other types of data is increasing getting importance in biomedical field. Comprehensive analysis of the Bacillus genome data was performed using various measures and the details of these measures are described in materials and methods. The objective of the analysis is to see whether any pattern or genomic signature emerges from the vast genome data and to study the phylogenomics of the Bacillus genomes to find any incongruence. This may also help us to understand the genome organization to certain extent and also may reveal the hidden structure of the genome. We have used all the available genome data pertaining to Bacillus family and the results are provided below. We have used the publicly available software, our own perl codes and pre- implemented algorithms. The graphs and figures are generated using the freely available GNU Plot software

**STUDY I: COMPARATIVE GENOMICS OF BACILLUS GENOMES**

**4.1 Genome Size Comparison**

Genome size comparison was performed for all the 10 bacillus genomes. The sequence lengths, in basepairs, of all the 10 genomes was obtained from the respective NCBI Genbank file. A graph was plot using GNUplot all the between the genome size (y axis) and 10 genome sequences (x axis).



**FIGURE 3 :** Graph between Genome size (y axis) and the 10 genome sequences of interest (x axis

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **SERIES 1** | | **SERIES 2** | | **SERIES 3** | | **SERIES4** | |
| Bacillus anthracis | 5228663 | Bacillus amyloliquefaciens | 3918589 | Bacillus halodurans | 4202352 | Bacillus pumilus | 3704465 |
| Bacillus thuringiensis | 5257091 | Bacillus subtilis | 4215606 | Bacillus clausii | 4303871 |  |  |
| Bacillus weihenstephanensis | 5262775 | Bacillus licheniformis | 4222645 |  |  |  |  |
| Bacillus cereus | 5419036 |  |  |  |  |  |  |

**Table2:** Clusters of bacillus species on the basis of genome size.

It can be observed that *Bacillus pumilus* has the smallest genome amongst all the ten bacillus genomes. It is clearly visible from the graph that the nine input genomes (i.e. apart from *Bacillus pumilus*) fall in three clearly distinguishable clusters according to their genome size:

Cluster 1: *Bacillus amyloliquefaciens, Bacillus subtilis and Bacillus licheniformis*

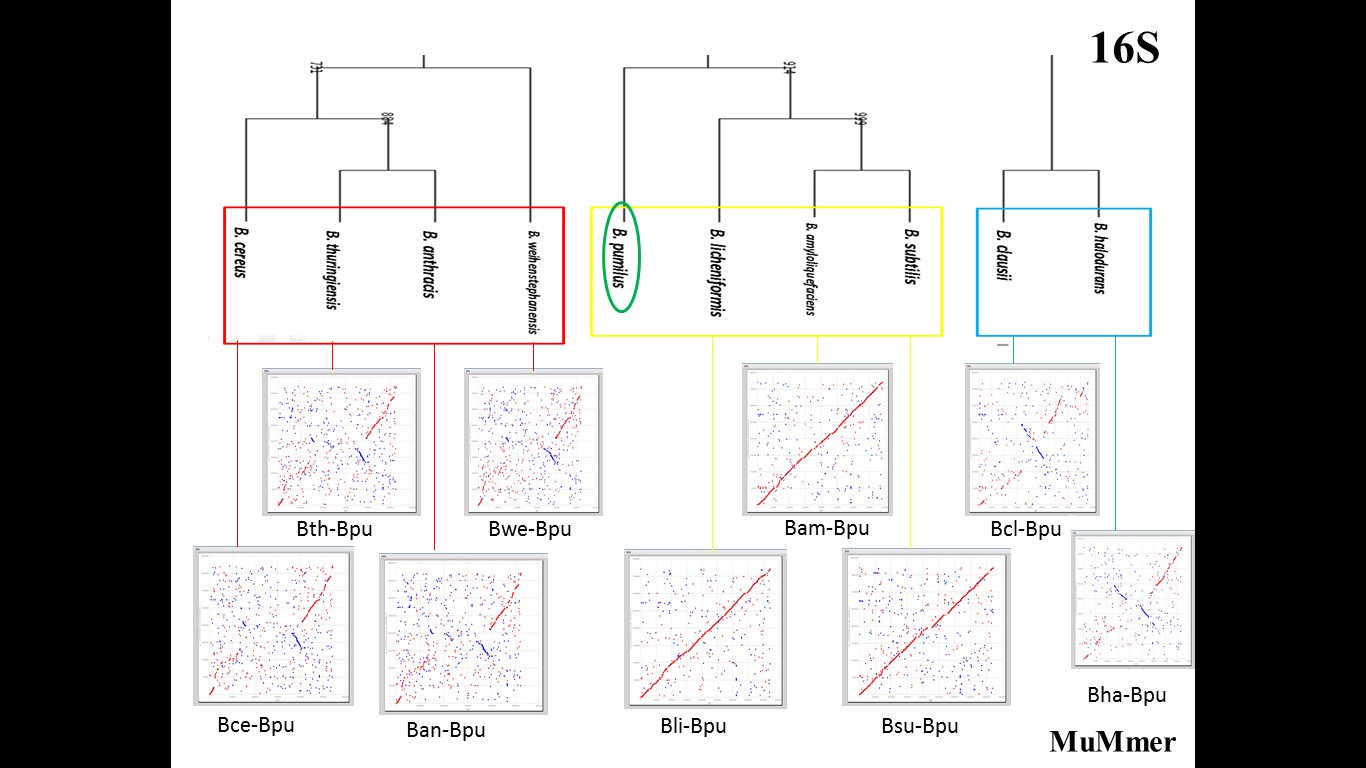
Cluster 2: *Bacillus halodurans and Bacillus clausii*

Cluster3: *Bacillus anthracis, Bacillus amyloliquefaciens,* *Bacillus weihenstephanensis and Bacillus cereus*

*Bacillus pumilus* having the smallest size falls near cluster 1 genome sequences. This classification is in accordance with the 16s ribosomal rna classification of these 10 bacillus species. Thus, according to the genomic size comparison, it can be stated that *bacillus pumilus* should be classified with *Bacillus* subtilis group species.

**4.3** **Whole Genome Pair Wise Sequence Alignment: MUMmer 3.0**

We have analyzed all the ten genomes, using mummer taking two sequences at one time. The mummer plots are shown in following figure (Mummer graphs). It is evidently clear, that the four species, i.e., bam, bli, bpu and bsu shows near perfect similarity and this four can be clustered together and may form a group or sub-class. The other group consisting of four species, ie., bans, bce, bth and bwe. It is interesting to see the two species bcl and bha do not have any similarity with either of the above mentioned group. How these two member species which do not have any significant sequence similarity is part of the bacillus family is striking and need further investigation. In conclusion we see two distinct groups or sub-class arising out of our analysis. For detailed graphs, see APPENDIX II

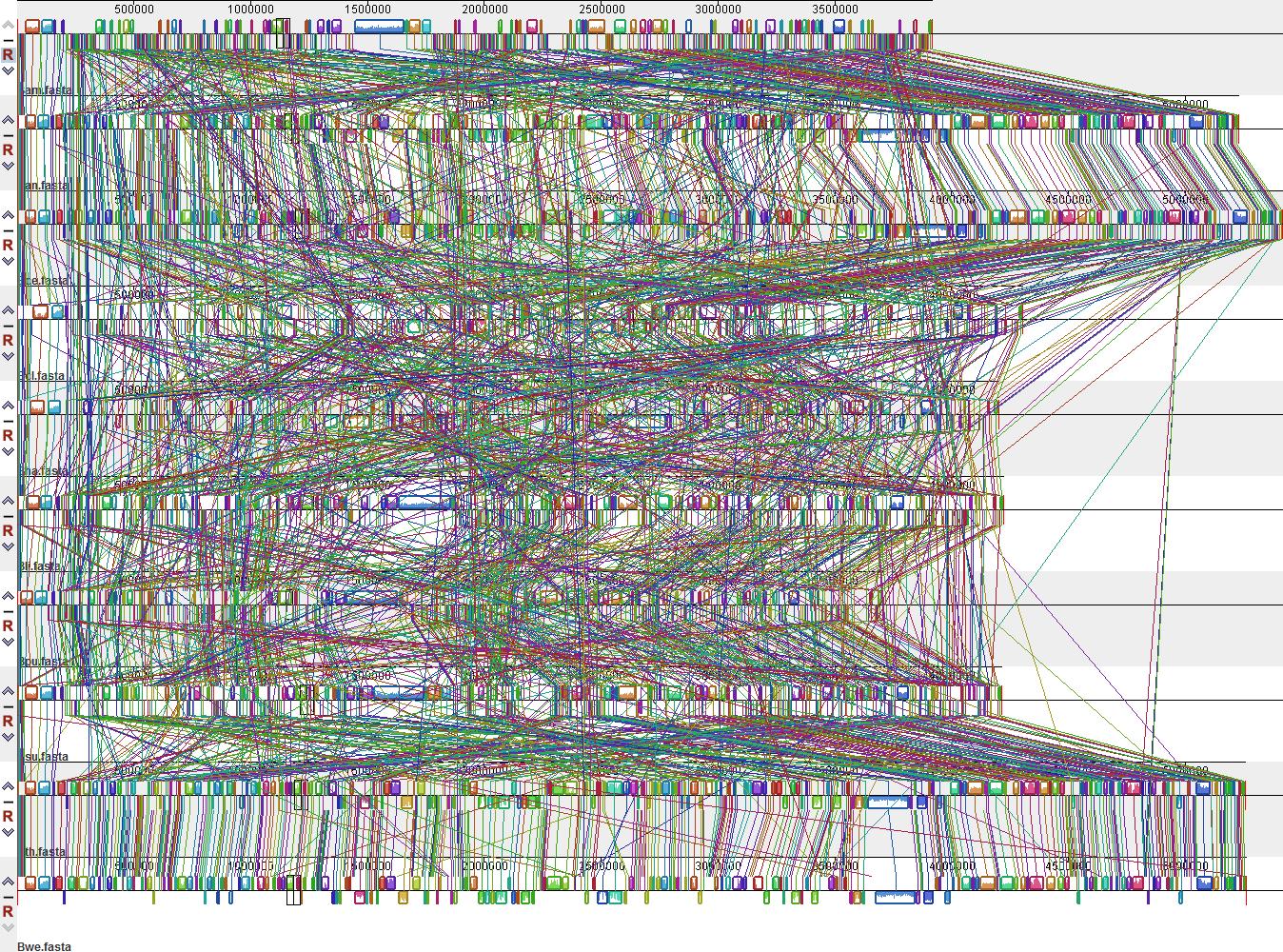


**FIGURE 5:** Comparison of the traditional 16s rRNA based phylogenetic tree (above) and pairwise sequence whole genome alignment results from MUMmer (below).

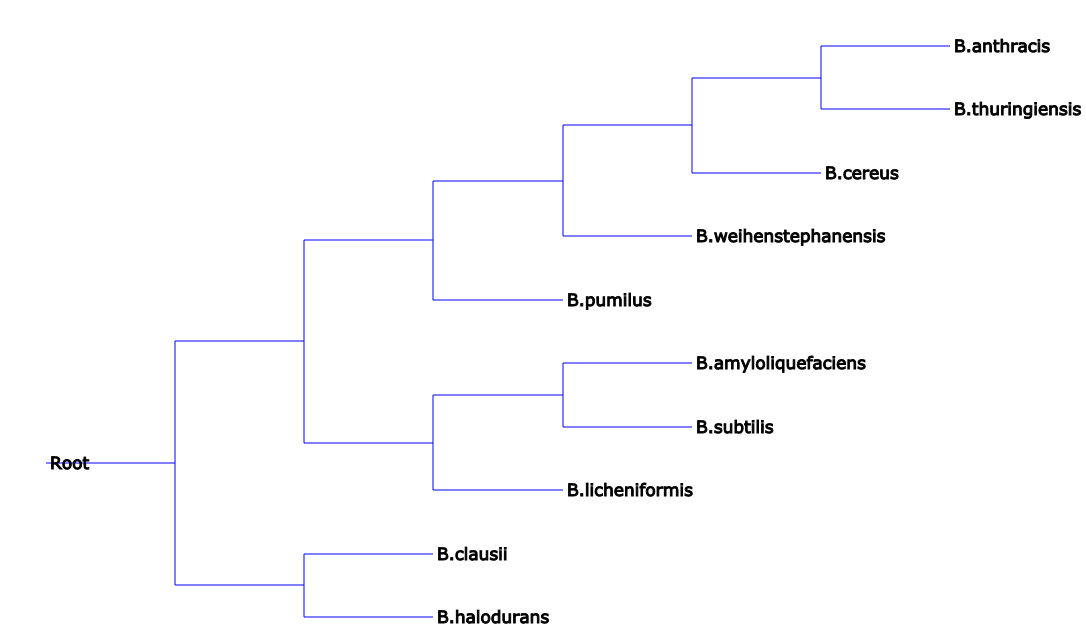
Figure shows the comparison between the (above) clusters obtained from 16s rRNA classification (below) MUMmer pairwise sequence alignment obtained taking Bacillus pumilus as reference organism. This result supports the 16s ribosomal rna evolutionary tree classification.

* 1. **Whole Genome Multiple Sequence Alignment: MAUVE**

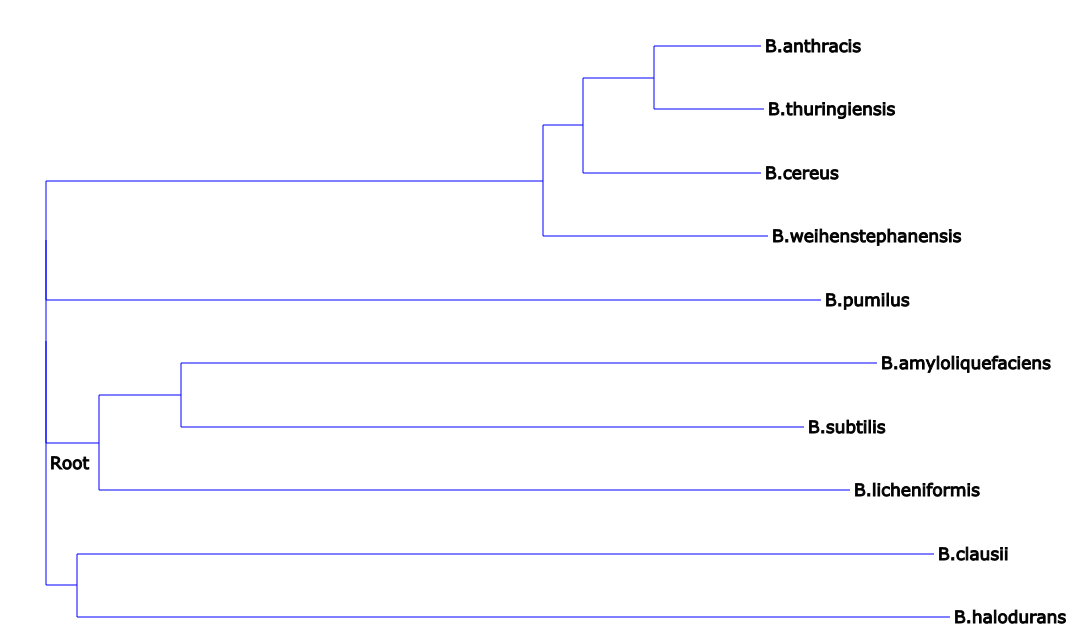
Whole genome multiple sequence alignment is a computationally intensive task. Normal methods of global and local alignment fail in the case of whole genomes. Thus, glocal alignment is used, following seed and extends strategy. All the 10 bacillus genomes werr uploaded in MAUVE and whole genome alignment was obtained. Figure6: since the alignment obtained looks very meshed up, a guide tree was constructed on the basis of the alignment similarity between the 10 genomes. From figure7 & 8, it can be concluded that mauve was unable to cluster Bacillus pumilus, however this means that it lies between the bacillus cereus cluster and bacillus subtilis cluster. This result is opposite of what we get from MUMMer and 16s rRNA phylogenetic tree.



**Figure6:** Whole genome alognment output from MAUVE



**FIGURE 7:** Phylogram obtained from whole genome alignment from MAUVE



**FIGURE 8:** Phylogram obtained from whole genome alignment from MAUVE

**4.5 Genome Signature – HGT Detection**

Every specie genome has a characteristic genome signature. If a part/region/stretch of nucleotides in a genome exhibit different genomic signature pattern than the genomic signature of the complete genome, then that region is called atypical and is considered to be of external origin and could have entered the host genome via reticulation events: horizontal gene transfer, recombination, hybridization etc. However this strategy is applicable only for recent transfers because transferred regions ameliorate over time and match the host genome, which makes then difficult to identify. As observed from the classification done by MAUVE, Bacillus pumilus is left unclustered. This result is opposite of what is obtained from MUMmer pairwise alignment and 16s rRNA based classification. Thus further investigations were done on Bacillus pumilus. The Genbank file was uploaded in **G**enomic **O**rigin of **H**orizontal **T**ransfers, **A**lignment & **M**etagenomics online tool (GOHTAM) for identification of atypical regions in this genome. Probable Horizontally transferred genes were identified in via two algorithms implemented in GOHTAM website:

* Overlapping Sliding Windows/(tetranucleotide) Algorithm
* Codon Usage Algorithm
* Combination of above two algorithms

These algorithms are based on the genomic signature identification in the complete genome. It highlights the atypical genome signatures in a genome and classifies it to be a probable horizontally transferred region. Overlapping sliding windows algorithm identifies the atypical regions in the input genome and codon usage algorithm predicts the genes present in that atypical stretch. The predicted atypical regions are then searched for possible neighbors in the genome signature database maintained by GOHTAM. The similarity and confidence scores are generated for the 10 nearest neighbors of each atypical gene. As shown in table below, when only Sliding window (tetra-nucleotide) Algorithm was applied on Bacillus pumilus genome, 44atypical regions were identified, when only Codon usage Algorithm was applied on Bacillus pumilus genome, 83 atypical genes were identified. However upon application of both the algorithms, 147 atypical genes were found in 82 atypical regions.

|  |  |
| --- | --- |
| **ALGORITHM USED BY GOHTAM** | **HGT’s PREDICTED** |
| Sliding window (tetra-nucleotide Algorithm) | 44 atypical regions predicted |
| Codon usage algorithm | 83 atypical genes obtained from codons |
| Combination of both algorithms | 147 atypical genes (in 82 atypical regions) |

**TABLE** 3: Horizontally transferred gene and regions predicted via algorithms implemented in GOHTAM server.

Thus, a total of 147 genes were predicted in 82 atypical regions by the combined algorithm [Appendix III]. This dataset was further used for finding similar regions in other organism’s genome; i.e. finding which other organisms’ have the similar genomic signatures as in the 82 atypical regions.

**FIGURE 9:** Workflow result for the identification of probable HGTs from GOHTAM.

Out of the 82 atypical regions, only those were kept which had matching genomic signatures in the bacillus genomes of our interest. Thus, 14 Atypical regions, out of 29, containing hits from 10 Bacillus species of our interest

|  |  |  |  |
| --- | --- | --- | --- |
| **S.NO** | **HGT REGION** | **ORGANISMS** | **STATUS** |
| 1 | 25 | B. thuringenesis; Bacillus sp. BS-01 | ACCEPTED |
| 2 | 26 | Bacillus methanolicus MGA3, Bacillus phage Spbeta, Bacillus pumilus | Same species, thus excluded |
| 3 | 27 | Bacillus subtilis subsp. Natto, Bacillus phage Spbeta, Bacillus subtilis subsp. Natto | ACCEPTED |
| 4 | 29 | Bacillus thuringiensis BMB171, Bacillus cereus E33L, Bacillus thuringiensis serovar konkukian str. 97-27, Bacillus sp. 1310(2010), Bacillus weihenstephanensis KBAB4, Bacillus megaterium QM B1551, Bacillus cereus AH820 | ACCEPTED |
| 5 | 36 | Bacillus thuringiensis, Bacillus phage Fah | ACCEPTED |
| 6 | 38 | Bacillus subtilis | NOT CREADIBLE RESULTS |
| 7 | 46 | Bacillus sp. BS-01, Bacillus phage Gamma, Bacillus phage Cherry, Bacillus megaterium QM B1551, Bacillus phage Fah, Bacillus cereus G9842, Bacillus cytotoxicus NVH 391-98 | VERY LOW SIMILARITY SCORE |
| 8 | 47 | Bacillus thuringiensis BMB171, Bacillus phage Spbeta, Bacillus cereus E33L, Bacillus weihenstephanensis KBAB4, Bacillus phage Fah | ACCEPTED |
| 9 | 65 | Bacillus cereus VPC1401, Bacillus cereus | ACCEPTED |
| 10 | 68 | Bacillus cereus H3081.97, Bacillus sp. JAMB750 | VERY LOW SIMILARITY SCORE |
| 11 | 72 | Bacillus thuringiensis serovar konkukian str. 97-27, Bacillus cereus E33L, Bacillus thuringiensis BMB171 | ACCEPTED |
| 12 | 73 | Bacillus cereus VPC1401, Bacillus cereus H3081.97, Bacillus thuringiensis, Bacillus cereus, Bacillus thuringiensis serovar tenebrionis, Bacillus cereus 03BB102 | ACCEPTED |
| 13 | 78 | Bacillus cereus AH820, Bacillus cereus, Bacillus cereus, Bacillus cereus E33L, Bacillus megaterium QM B1551, Bacillus cereus Q1, Bacillus thuringiensis serovar chinensis CT-43, Bacillus weihenstephanensis KBAB4 | ACCEPTED |
| 14 | 80 | B. thuringenesis | ACCEPTED |

Table 3: Atypical regions in Bacillus pumilus obtained via GOHTAM.

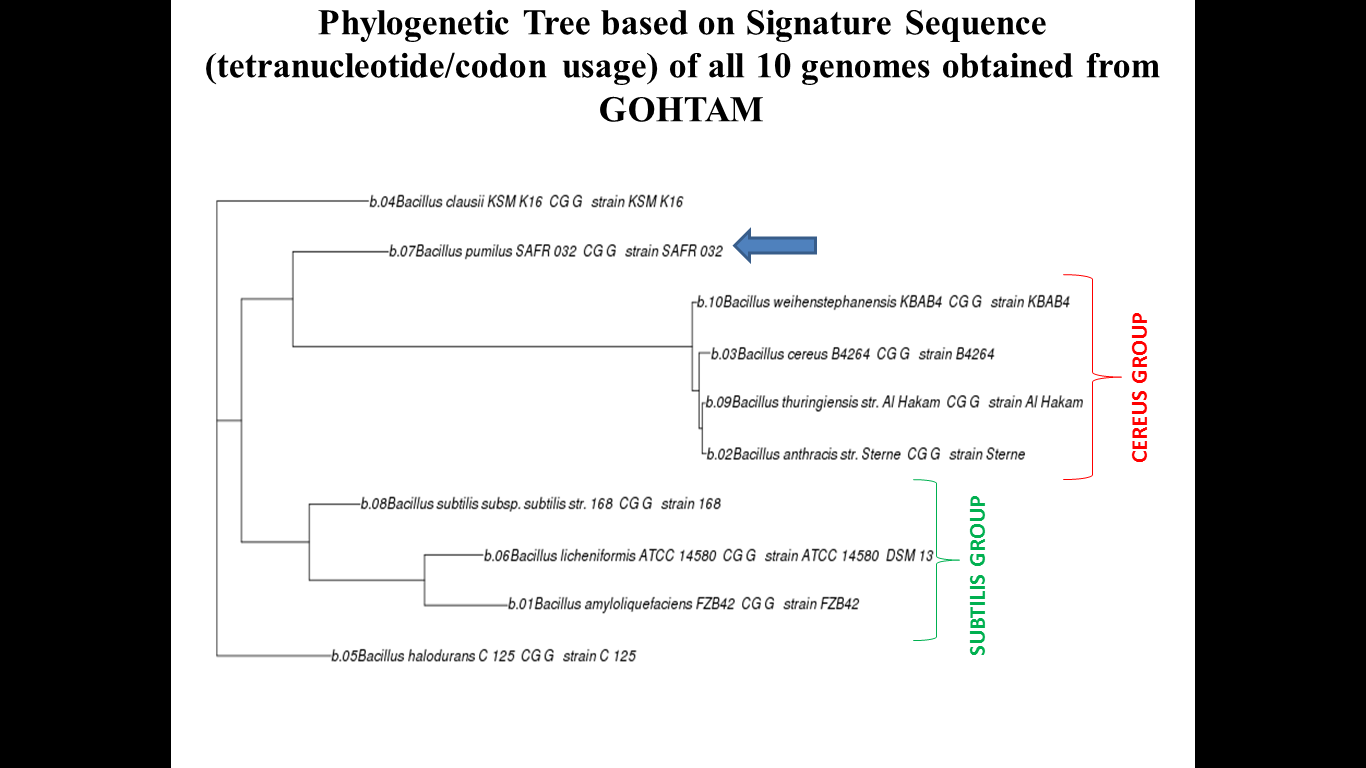
From the above table it can be seen that of all the atypical regions identified, only one, HGT REGION 27 contained a credible hit from *Bacillus subtilis* group, rest all were from *Bacillus cereus*. Thus, it can be concluded that though the pairwise whole genome alignment from MUMmer predicts that *Bacillus* pumilus genome has a very high identity towards species in *Bacillus subtilis* group, the HGT analysis of the Bacillus pumilus genome on the basis of Genomic signature criteria reveals its closeness towards species in *Bacillus cereus* group. Thus it can be said that *Bacillus subtilis* group species is more primitive than *Bacillus cereus* and *Bacillus pumilus*.

Excluding all the organisms other than the nine organisms of our interest, and excluding all the hypothetical proteins, finally we were left with 28 genes in 8 atypical regions. These are the probable horizontally transferred genes in *bacillus pumilus* genome.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **HGT REGION NO** | **HGT REGION LENGTH** | **GOHTAM ID** | **REGION** | **GENE IN THIS REGION** |
| 25 | 588250, 602750 | > BPUM\_0549 | 588295..589413 | spore transcriptional regulator YopK |
| > BPUM\_0550 | complement(589733..590119) | transcriptional regulator |
| > BPUM\_0554 | 593151..594485 | FtsK/SpoIIIE family protein |
| > BPUM\_0555 | 594559..595659 | transcriptional regulator |
| > BPUM\_0561 | 601232..602017 | DNA (cytosine-5-)-methyltransferase |
| 27 | 613250, 618921 | > BPUM\_0576 | 614391..615368 | C40 family peptidase |
| > BPUM\_0580 | 616312..617589 | integrase |
| > BPUM\_0581 | complement(617863..618921) | AraC family transcription regulator |
| 29 | 727814, 736750 | > BPUM\_0652 | 727814..729304 | ATP-binding protein |
| > BPUM\_0653 | 729297..732107 | endonuclease |
| > BPUM\_0656 | complement(734880..736526) | DNA (cytosine-5-)-methyltransferase |
| 36 | 1201702, 1208528 | > BPUM\_1125 | complement(1202715..1203653) | acetyl-CoA carboxylase carboxyltransferase subunit alpha |
| > BPUM\_1126 | complement(1203686..1204132) | MarR family transcriptional regulator |
| > BPUM\_1130 | 1206948..1207778 | acetylxylan esterase |
| > BPUM\_1131 | 1207866..1208528 | chloramphenicol O-acetyltransferase |
| 47 | 1767750, 1771750 | > BPUM\_1721 | complement(1770270..1770851) | flavodoxin |
| > BPUM\_1722 | 1770975..1771319 | MarR family transcriptional regulator |
| 72 | 3196250, 3209454 | > BPUM\_3222 | complement(3196510..3198342) | glycosyltransferase |
| > BPUM\_3223 | complement(3198329..3199507) | aminotransferase |
| > BPUM\_3224 | complement(3199508..3200788) | carbamoylphosphate synthase large subunit short form |
| > BPUM\_3225 | complement(3200801..3202804) | CDP-glycerol glycerophosphotransferase |
| > BPUM\_3226 | complement(3202901..3203962) | Dtdp-glucose 4,6-dehydratase |
| > BPUM\_3227 | complement(3203982..3204866) | glucose-1-phosphate thymidylyltransferase |
| > BPUM\_3228 | complement(3205081..3207336) | CDP-glycerol glycerophosphotransferase |
| > BPUM\_3229 | complement(3207424..3209454) | poly(glycerol-phosphate) alpha-glucosyltransferase |
| 73 | 3215244, 3218858 | > BPUM\_3234 | 3215244..3218858 | glycosyltransferase |
| 80 | 3602621, 3607338 | > BPUM\_3636 | 3602621..3603391 | ABC transporter ATP-binding protein |
| > BPUM\_3641 | 3606160..3607338 | aspartate phosphatase response regulator K |

TABLE 4: Probable horizontally transferred genes in Bacillus pumilus obtained via GOHTAM

Signature sequence is used to infer the phylogenetic tree (Chapus et al. BMC Evol Biol 2005). NJ phylogenetic tree is generated for input sequences on the basis of tetranucleotide signatures/ codon usage.

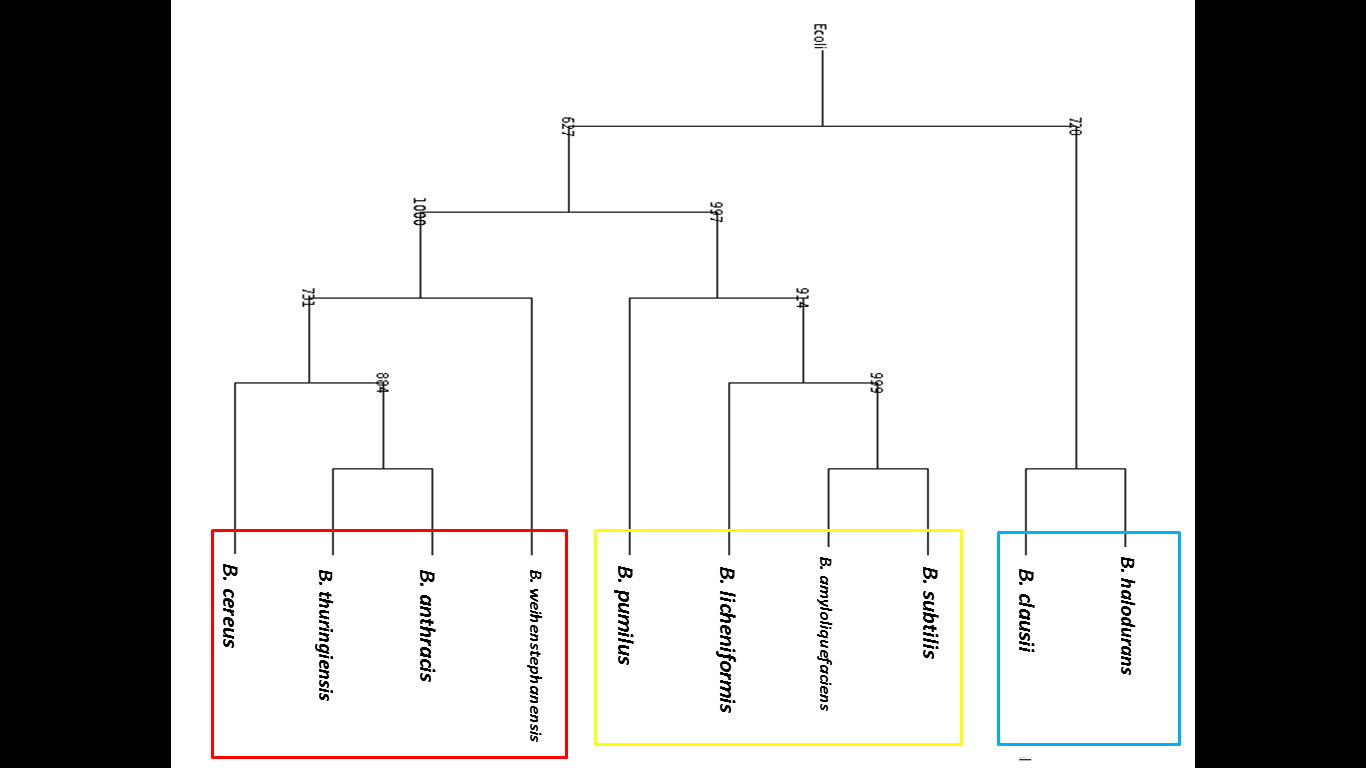


**FIGURE 10:** Phylogenetic tree based on signature sequence (tetranucleotide/codon usage) of 10 genomes from GOHTAM.

From figure , it can be observed that classification on the basis of genomic signatures reveals an totally new classification of our bacillus genomes. Bacillus pumilus was clustered with species of bacillus cereus group rather than with bacillus subtilis group. (Highlighted by blue arrow)

**4.6 Single Gene Tree Construction**

Ribosomal RNA (rRNA) performs the same function in every cell, and so changes in their genetic sequences are not biased by differing functional selection in different taxa. Also, extent of mutation is very less in these genes, they are thus used for construction of the tree of life i.e. they are considered as markers to study evolution. We obtained 16s ribosomal RNA gene sequences for all the 10 bacillus genomes from NCBI and constructed phylogenetic trees as explained in methodology section. E.coli was taken as n outgroup for construction of the tree. Figure:



**FIGURE 11:** 16s rRNA based phylogenetic tree

From figure it is observed that the 10 bacillus species are classified into three distinct clusters:

Cluster 1: Bacillus cereus group (*Bacillus anthracis, Bacillus amyloliquefaciens,* *Bacillus weihenstephanensis and Bacillus cereus*)

Cluster 2: Bacillus subtilis group (*Bacillus amyloliquefaciens, Bacillus subtilis and Bacillus licheniformis)* and *Bacillus pumilus*

Cluster 3: *Bacillus halodurans and Bacillus clausii*

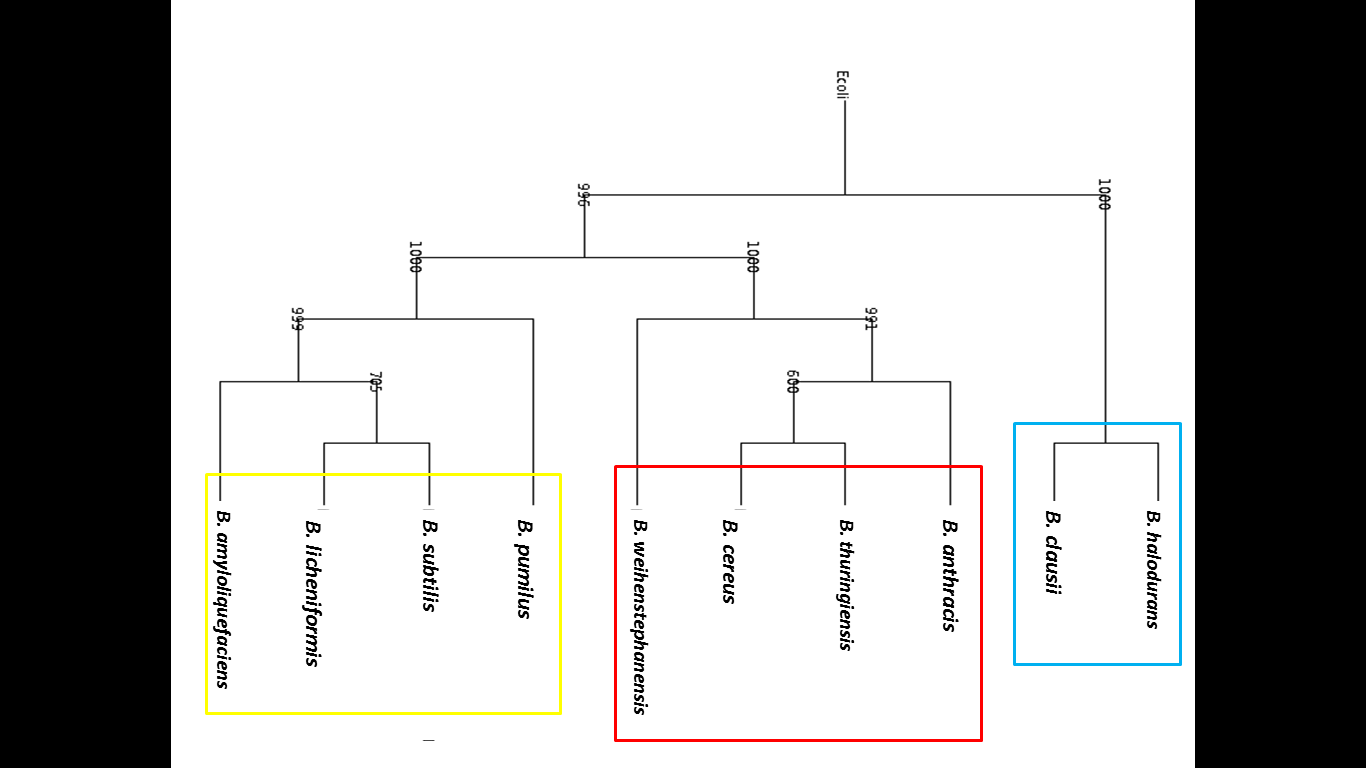


FIGURE12: 23s rRNA based phylogenetic tree

. For the production of more enriched classification, 23s rRNA sequences were also obtained for all the 10 bacillus genomes and similar phylogenetic methods were applied to produce a phylogenetic tree. From the figure, it is observed that the clustering done by 23s rRNA based phylogenetic tree was identical to that done by 16s rRNA tree. The only difference was the positions of the clusters 1 and 2, which are opposite in both trees.

**CONCLUSION**

**&**

**FUTURE PERSPECTIVE**

Evolution plays a vital role in diversity and adaptation. Besides, single point mutations chunks of DNA are also get transferred within genome or across the genome. These changes give new biological functions for the acquired organism and the evolution makes further changes. Tracing the history of evolution and development is a difficult task as we make several assumptions about evolutionary events which may be right or wrong and always a topic of hot debate or discussion. Nevertheless, the exercise that we do in relation to evolution is always interesting and useful in getting insights into some of the properties or functions that we are observing today.

The vast amount of genome data helps us to get a useful picture of bacillus genomes in terms of size, geometry etc. The detailed analysis of the data may give us new insights about the genome structure and evolution. Many hypotheses have been proposed to explain such observation. Methods for analysing genome sequence features enable us to make assessments about genomic organisation and heterogeneity. Some of the measures aim at studying the genomic signature and their evolutionary implications, sequence heterogeneity within and across genomes, codon usage patterns and their relative fluctuations from the average, stand asymmetry, nucleotide transitions etc. We have carried out our comprehensive analysis keeping in mind the above salient features specifically for the ten Bacillus genomes.

Our results shows that there is definitely two distinct sub-classes, within this family i.e one with B.anthracis, B.weihenstephanensis, B.thuringiensis, B.cereus and otherwith B.amyloliquefaciens, B.clausii, B.halodurans, B.licheniformis, B.pumilus,and B.subtilis. The members of the two sub-classes are in general more or less the same except in few cases where they differ by one or two. Results indicate that there is a possibility that the Bacillus family can be categorised into two distinct classes. Several measures pinpoint such grouping and also are agreement with our analysis with 16S RNA and 23S RNA sequences. Which class arisen first may be a subject of further investigation. The significance of these two sub-classes may enable researchers to distinctly deal with these groups while studying any biological process of intrest. To the best of our knowledge grouping of these bacillus genomes based on genomic data alone is itself a novel attempt.

1. Traditional phylogenetic analysis (Tree of life) was done using the16s ribosomal RNA marker sequences from the 10 bacillus species taking ecoli as outgroup. 3 distinct clusters were observed in the resulting phylogenetic tree:

Cluster 1: Bacillus cereus group (*Bacillus anthracis, Bacillus*

*thuringiensis, Bacillus w and Bacillus cereus*)

Cluster 2: Bacillus subtilis group (*Bacillus amyloliquefaciens,*

*Bacillus subtilis and Bacillus licheniformis)* and *Bacillus*

*pumilus*

Cluster 3: *Bacillus halodurans and Bacillus clausii*

1. Pair wise alignment of complete genomes of the 10 bacillus species of interest was done using MUMmer. Bacillus Genome family based on molecular genomic analysis forms 3 clusters; which is identical to the traditional 16S clustering.
2. Detailed multiple sequence alignment of whole genomes using MAUVE predicted that bacillus pumilus was not clustered and was placed in between the two clusters: bacillus subtilis cluster and bacillus cereus cluster; opposite to the results obtained from the traditional phylogenetic analysis.
3. Genomic signature based phylogenetic analysis clustered Bacillus pumilus near to, and not with, bacillus cereus cluster than with Bacillus cluster, as obtained from 16s rRNA tree.

Our analysis provides some possible clues and direction about the evolutionary path within the bacillus genomes under study, but still comprehensive analysis is required to establish to get the insight to the truth.However, we were unable to provide any significant explanation for this occurrence. The nucleotide composition of Bacillus pumilus is similar to Bacillus subtilis cluster, as observed from 16s rRNA phylogenetic analysis and MUMmer result. However, multiple sequence alignment positions Bacillus pumilus unclustered. Also, genomic signature based analysis revealed probable horizontal gene transfer between Bacillus pumilus and Bacillus cereus cluster species. Thus, the direction of evolution could not be clearly established.

**REFERNCES**

1. Lecointre G, Philippe H, Vân Lê HL, Le Guyader H. Species sampling has a major impact on phylogenetic inference. Mol Phylogenet Evol. 1993 Sep;2(3):205-24.
2. Ong CK, Nee S, Rambaut A, Harvey PH. Inferring the population history of an epidemic from a phylogenetic tree. J Theor Biol. 1996 Sep 21;182(2):173-8.
3. Darwin, C. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. 1859 John Murray, London. [1st ed.]
4. Cracraft, J. and Donoghue, M. J. (eds). Assembling the Tree of Life, 2004 Oxford University Press, Oxford.
5. Yates, T. L., Salazar-Bravo, J., and Dragoo, J. W. In Assembling the Tree of Life (Cracraft, J. and Donoghue, M. J., eds), pp. 7–17, 2004 Oxford University Press, Oxford.
6. Dawkins, R. (2003) A Devil’s Chaplain, Houghton Mifflin, New York.
7. Fleischmann, R. D., Adams, M. D., White, O., et al. (1995) Whole-genome random sequencing and assembly of Haemophilus nfluenza Rd. Science 269, 496–512.
8. Goffeau, A., Barrell, B. G., Bussey, H., et al. (1996) Life with 6000 genes. Science 274, 546, 563–567.
9. Liolios, K., Tavernarakis, N., Hugenholtz, P., and Kyrpides, N. C. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. 2006 Nucleic Acids Res. 34, D332–D334.
10. Rokas, A., Kruger, D., and Carroll, S. B. Animal evolution and the molecular signature of radiations compressed in time. 2005 Science 310, 1933–1938.
11. Driskell, A. C., Ane, C., Burleigh, J. G., McMahon, M. M., O’Meara, B. C., and Sanderson, M. J. Prospects for building the tree of life from large sequence databases. 2004 Science 306, 1172–1174.
12. Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., and Bork, P. Toward automatic reconstruction of a highly resolved tree of life. 2006 Science 311, 1283–1287.
13. Takezaki, N., Figueroa, F., Zaleska-Rutczynska, Z., Takahata, N., and Klein, J. The phylogenetic relationship of tetrapod, coelacanth, and lungfish revealed by the sequences of 44 nuclear genes. 2004 Mol. Biol. Evol. 21, 1512–1524.
14. Jennings,W. B. and Edwards, S. V. Speciational history of Australian grass finches (Poephila) inferred from thirty gene trees. 2005 Evolution 59, 2033–2047.
15. Rokas, A., Williams, B. L., King, N., and Carroll, S. B. Genome-scale approaches to resolving incongruence in molecular phylogenies. 2003 Nature 425, 798–804.
16. Gogarten, J. P. and Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. 2005 Nat. Rev. Microbiol. 3, 679–687
17. Satta,Y., Klein, J., and Takahata, N. DNA archives and our nearest relative: the trichotomy problem revisited. 2000 Mol. Phylog. Evol. 14, 259–275.
18. Edwards S.V. Is a new and general theory of molecular phylogenetics emerging? 2009 Evolution 63: 1–19.
19. Morrison D.A. Multiple sequence alignment for phylogenetic purposes. 2006 Australian Systematic Botany 19: 479–539.
20. Morrison, D.A. A framework for phylogenetic sequence alignment. 2009 Plant Systematics and Evolution 282: 127–149.
21. Müller, K.F. Incorporating information from length-mutational events into phylogenetic analysis. 2006 Molecular Phylogenetics and Evolution 38: 667–676.
22. Ochoterena H. Homology in coding and non-coding DNA sequences: a parsimony perspective. 2009 Plant Systematics and Evolution 282: 151–168.
23. Morrison D.A. Phylogenetic analyses of parasites in the new millennium. 2006. Advances in Parasitology 63: 1–124.
24. Gogarten, J.P.,W.F. Doolittle & J.G. Lawrence. Prokaryotic evolution in light of gene transfer. 2002 Mol. Biol. Evol. 19: 2226–2238.
25. Zhaxybayeva, O. & J.P. Gogarten. 2004. Cladogenesis, coalescence and the evolution of the three domains of life. Trends Genet. 20: 182–187.
26. Gogarten, J., G. Fournier & O. Zhaxybayeva. Gene transfer and the reconstruction of life’s early history from genomic data.2008 Space Sci. Rev. 135: 115– 131.
27. Raymond, J., O. Zhaxybayeva, J.P. Gogarten, et al. Whole-genome analysis of photosynthetic prokaryotes. 2002 Science 298: 1616–1620.
28. Xiong, J. & C.E. Bauer. Complex evolution of photosynthesis. 2002 Annu. Rev. Plant Biol. 53: 503–521.
29. Fournier, G.P. & J.P. Gogarten. Evolution of acetoclastic methanogenesis in Methanosarcina via horizontal gene transfer from cellulolytic Clostridia. 2008 J. Bacteriol. 190: 1124–1127.
30. Gogarten, J.P. The early evolution of cellular life. 1995 Trends Ecol. Evol. 10: 147–151.
31. Huang, J. & J.P. Gogarten. Concerted gene recruitment in early plant evolution. 2008 Genome Biol. 9: R109.
32. Huang, J. & J.P. Gogarten. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? . 2007. Genome Biol. 8: R99.
33. Huang, J.L. & J.P. Gogarten. Ancient horizontal gene transfer can benefit phylogenetic reconstruction. 2006. Trends Genet. 22: 361–366.
34. Huang, J., Y.Xu & J.P. Gogarten. The presence of a haloarchaeal type tyrosyl-tRNAsynthetasemarks the opisthokonts as monophyletic. 2005 Mol. Biol. Evol. 22: 2142–2146.
35. Maynard Smith, J., Smith, N. H., O’Rourke, M. & Spratt, B. G. How clonal are bacteria? 1993, Proc. Natl. Acad. Sci. USA 90, 4384–4388.
36. Henze, K., Badr., A., Wettern, M., Cerff, R. & Martin, W. (1995) Proc. Natl. Acad. Sci. USA 92, 9122–9126.
37. Gupta, R. S., Aitken, K., Falah, M. & Singh, B. A nuclear gene of eubacterial origin in Euglena gracilis reflects cryptic endosymbioses during protist evolution.1994 Proc. Natl.Acad. Sci. USA 79, 2895–2899.
38. Lake, J. A. Mapping evolution with ribosome structure: intralineage constancy and interlineage variation.1982 Proc. Natl. Acad. Sci. USA 79, 5948–5952.
39. Feng, D.-F., Cho, G. & Doolittle, R. F. Determining divergence times with a protein clock: update and reevaluation. 1997 Proc. Natl. Acad. Sci. USA 94, 13028–13033.
40. Brown, J. R. & Doolittle, W. F. Archaea and the prokaryote-to-eukaryote transition 1997 Microbiol. Mol. Biol. Rev. 61, 456–502.
41. Ribeiro, S. & Golding, G. B. The mosaic nature of the eukaryotic nucleus. 1998 Mol. Biol. Evol. 15, 779–788.
42. Koonin, E. V., Mushegian, A. R., Galperin, M. Y. & Walker, D. R. Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. 1997 Mol. Microbiol. 25, 619–637.
43. Rivera, M. C., Rain, R., Moore, J. E. & Lake, J. A. Genomic evidence for two functionally distinct gene classes. 1998 Proc. Natl. Acad. Sci. USA 95, 6239–6244.
44. Vernikos GS, Thomson NR, Parkhill J: Genetic flux over time in the Salmonella lineage. 2007 Genome Biol , 8:R100.
45. Lawrence JG, Ochman H: Amelioration of bacterial genomes: rates of change and exchange. 1997 J Mol Evol, 44:383-97.
46. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: Toward automatic reconstruction of a highly resolved tree of life. 2006, Science, 311:1283-7.
47. Sorek R, Zhu Y, Creevey C, Francino M, Bork P, Rubin E: Genomewide experimental determination of barriers to horizontal gene transfer. 2007 Science, 318:1449-52.
48. Huang J, Gogarten J: Ancient horizontal gene transfer can benefit phylogenetic reconstruction. 2006 Trends Genet, 22:361-6.
49. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD. “Multiple sequence alignment with the Clustal series of programs”. 2003. Nucleic Acids Res 31 (13): 3497–3500.
50. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. “ClustalW and ClustalX version 2”. 2007. Bioinformatics 23 (21): 2947–2948. DOI:10.1093/bioinformatics/btm404.
51. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. “The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools”. 1997. Nucleic Acids Research 25 (24): 4876–4882
52. Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
53. A.L. Delcher, A. Phillippy, J. Carlton, and S.L. Salzberg ;”Fast Algorithms for Large-scale Genome Alignment and Comparision.”, 2002 Nucleic Acids Research, Vol. 30, No. 11 2478-2483.
54. A.L. Delcher, S. Kasif, R.D. Fleischmann, J. Peterson, O. White, and S.L. Salzberg; Alignment of Whole Genomes, 1999 Nucleic Acids Research, 27:11, 2369-2376.
55. S. Kurtz ; Reducing the Space Requirement of Suffix Trees,1999 Software-Practice and Experience, 29(13): 1149-1171
56. Aaron E. Darling, Bob Mau, and Nicole T. Perna. ProgressiveMauve: Multiple Genome Alignment with Gene Gain, Loss, and Rearrangement. 2010, PloS One. 5(6):e11147.
57. Aaron C.E. Darling, Bob Mau, Frederick R. Blatter, and Nicole T. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. 2004 Genome Research. 14(7):1394-1403.
58. Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B. Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Mol Biol Evol. 1999 Oct;16(10):1391-9.
59. Becq J, Churlaud C, Deschavanne P. A benchmark of parametric methods for horizontal transfers detection. PloS One. 2010 Apr 1;5(4):e9989.
60. Chapus C, Dufraigne C, Edwards S, Giron A, Fertil B, Deschavanne P. Exploration of phylogenetic data using a global sequence analysis method. BMC Evol Biol. 2005 Nov 9;5:63.
61. Dufraigne C, Fertil B, Lespinats S, Giron A, Deschavanne P. Detection and characterization of horizontal transfers in prokaryotes using genomic signature. Nucleic Acids Res. 2005 Jan 13;33(1):e6.
62. Mallet LV, Becq J, Deschavanne P. Whole genome evaluation of horizontal transfers in the pathogenic fungus Aspergillus fumigatus. BMC Genomics. 2010 Mar 12;11:171.
63. Nakamura Y, Itoh T, Matsuda H, Gojobori T. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet. 2004 Jul;36(7):760-6. Erratum in: Nat Genet. 2004 Oct;36(10):1126.
64. Ochman H, Lerat E, Daubin V. Examining bacterial species under the specter of gene transfer and exchange. Proc Natl Acad Sci U S A. 102 Suppl 1:6595-9. Epub 2005 Apr 25. Review.

**APPENDIX**

**APPENDIX I**

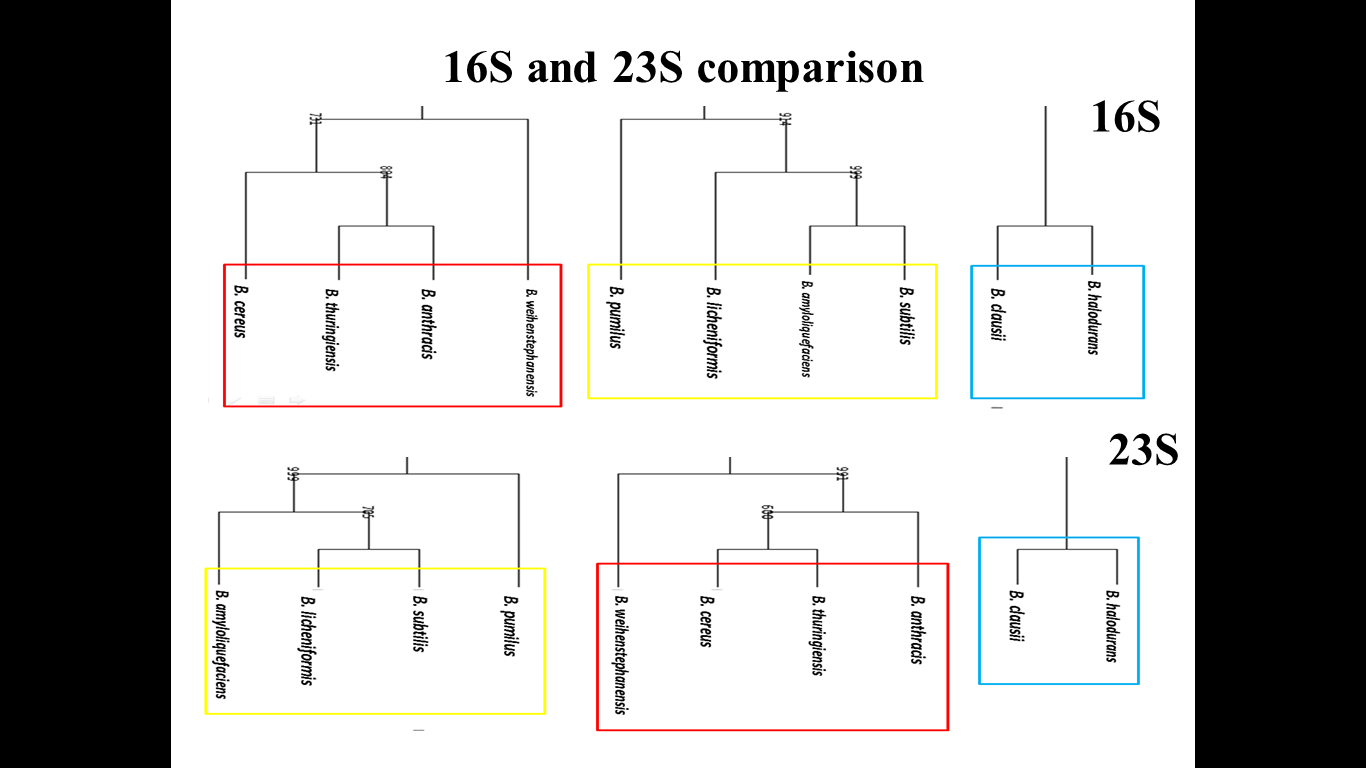


FIGURE: Comparison between 16s rRNA and 23s rRNA based phylogenetic tree obtained from PHYLIP

**APPENDIX II**

**MUMmer OUTPUT**

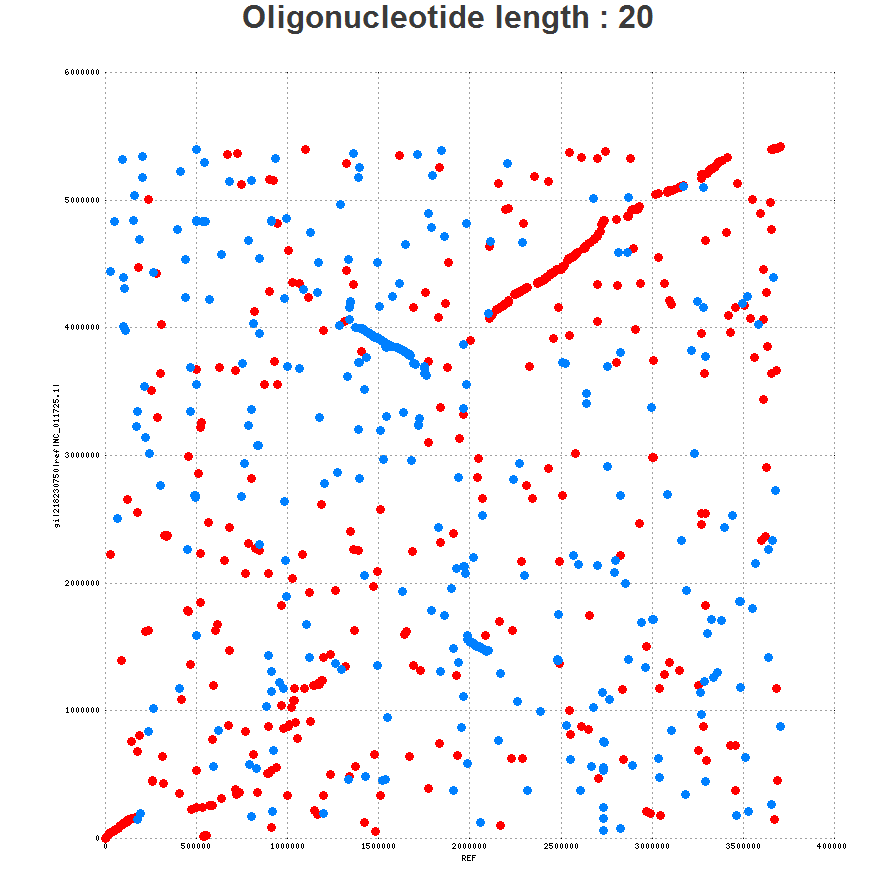


Figure: Bacillus pumilus and Bacillus cereus output from MUMmer 3

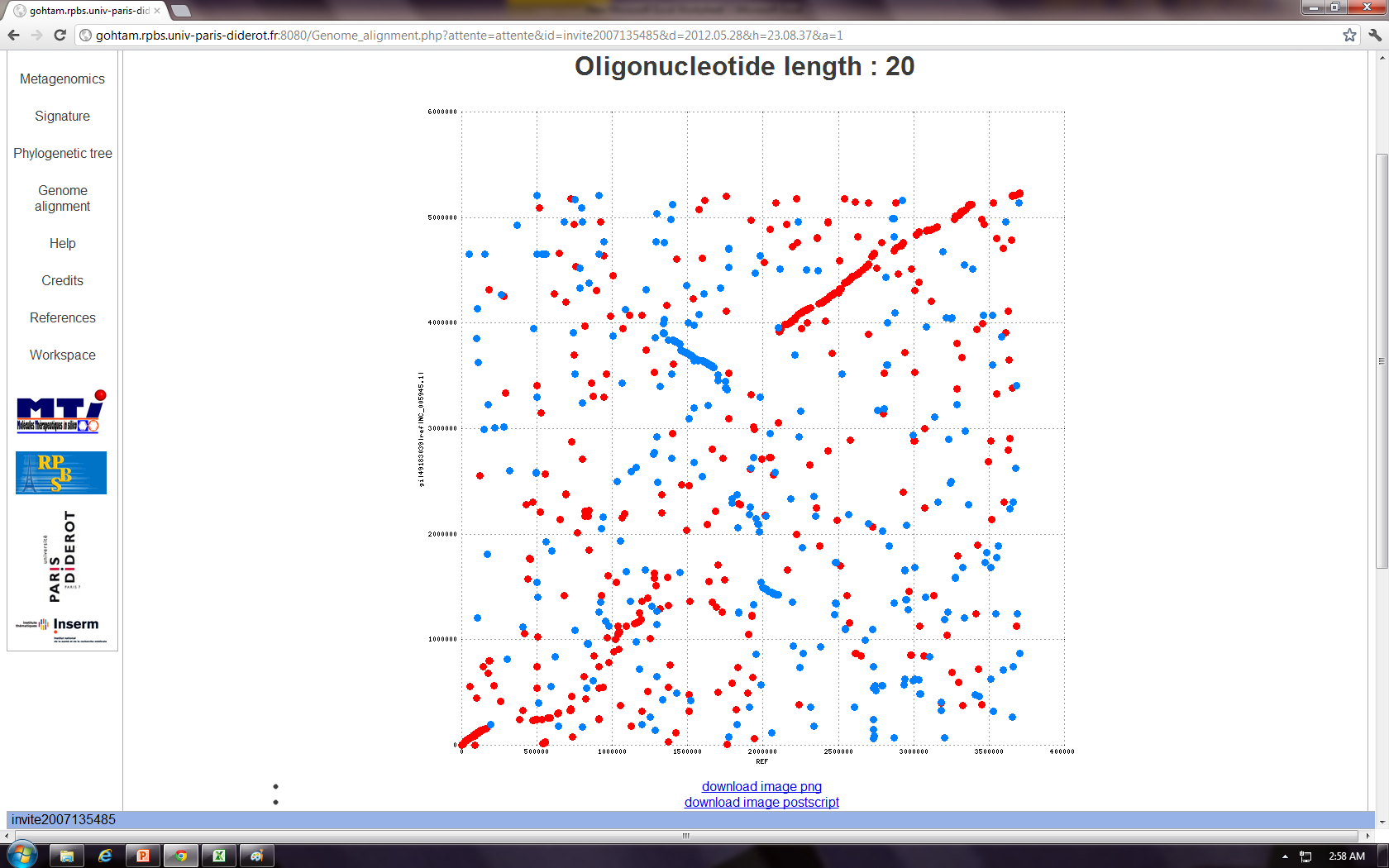


Figure: Bacillus pumilus and Bacillus anthracis output from MUMmer 3

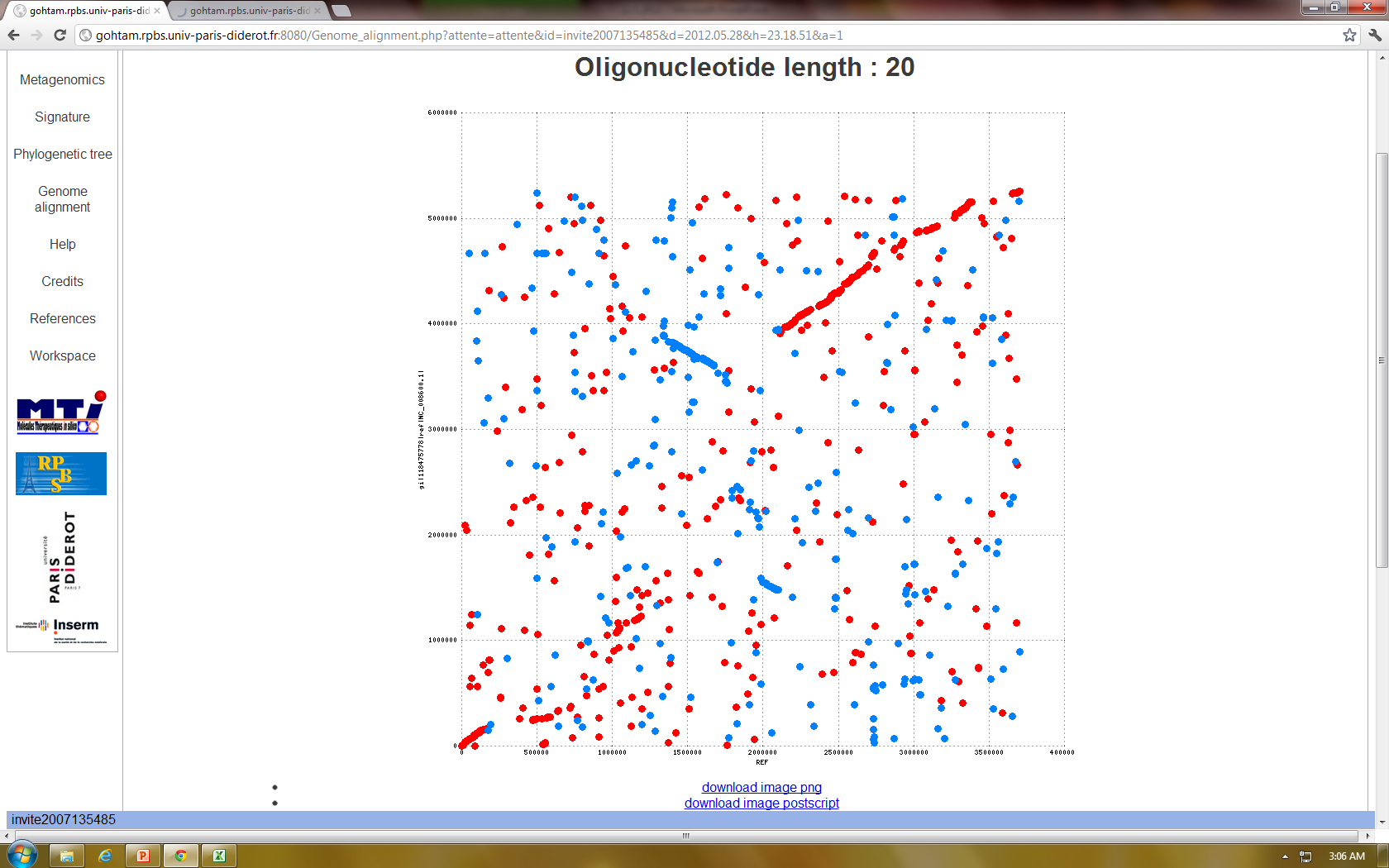


Figure: Bacillus pumilus and Bacillus thuringenesis output from MUMmer 3

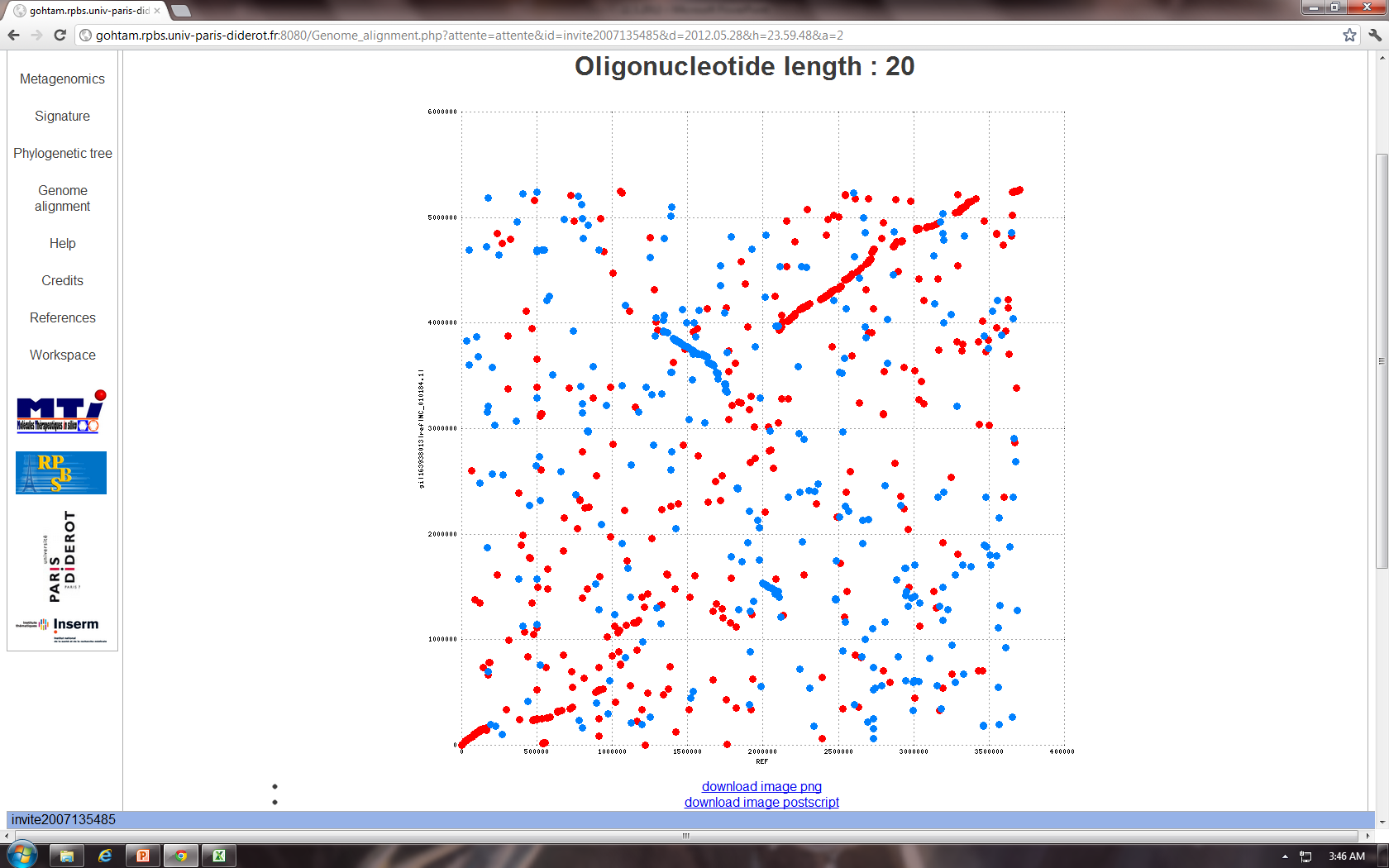


Figure: Bacillus pumilus and Bacillus we output from MUMmer 3

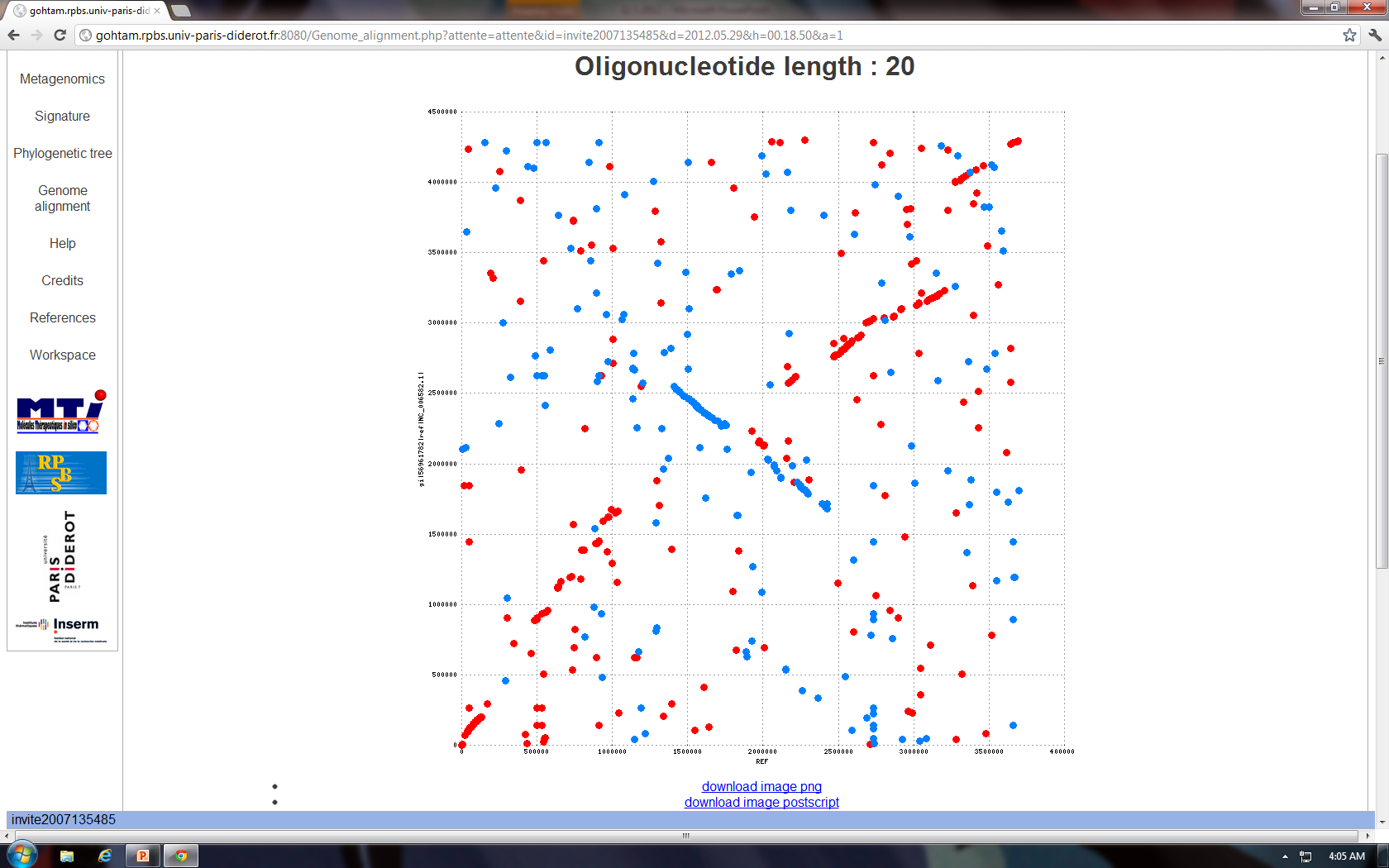


Figure: Bacillus pumilus and Bacillus clausii output from MUMmer 3

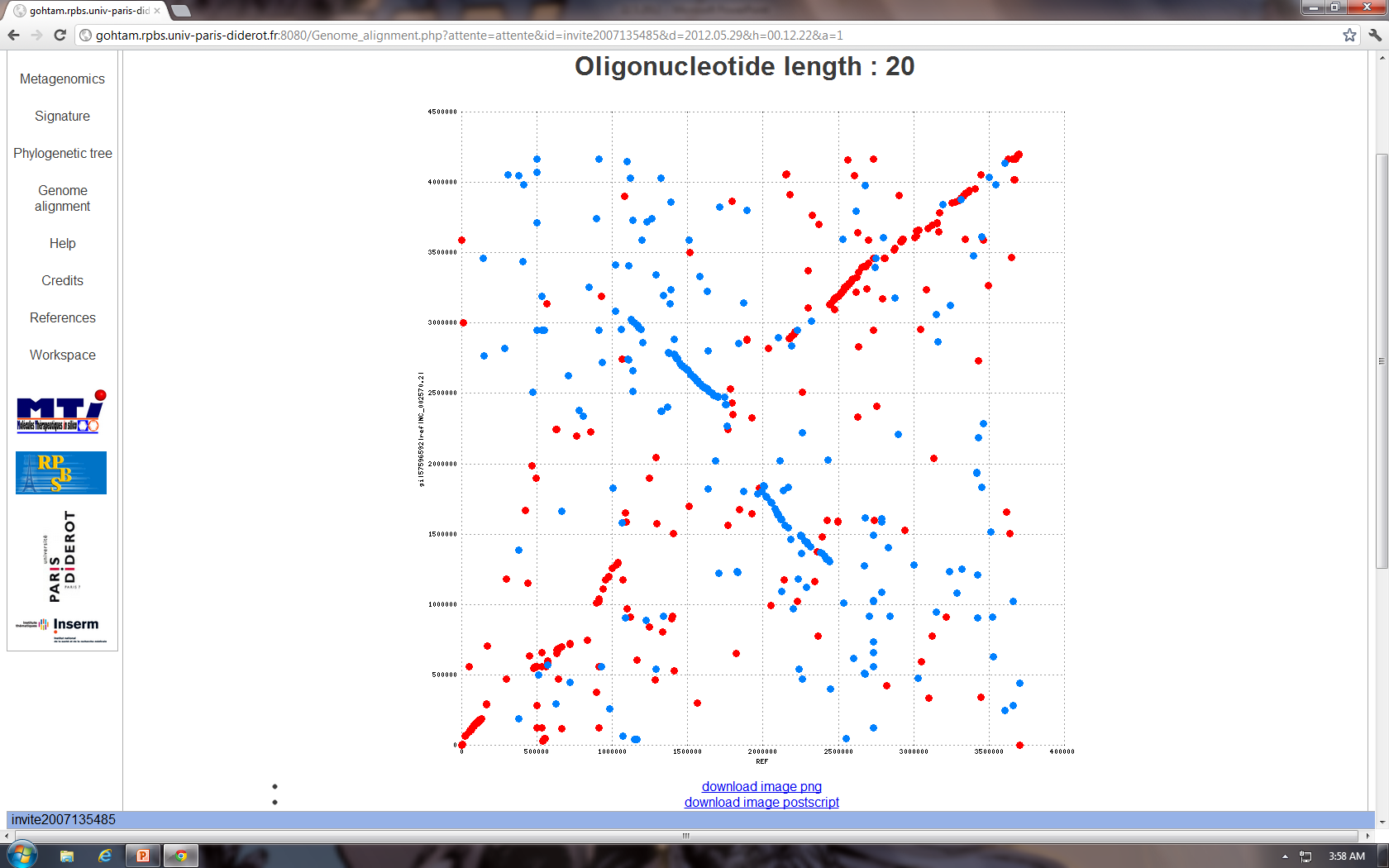


Figure: Bacillus pumilus and Bacillus ha output from MUMmer 3

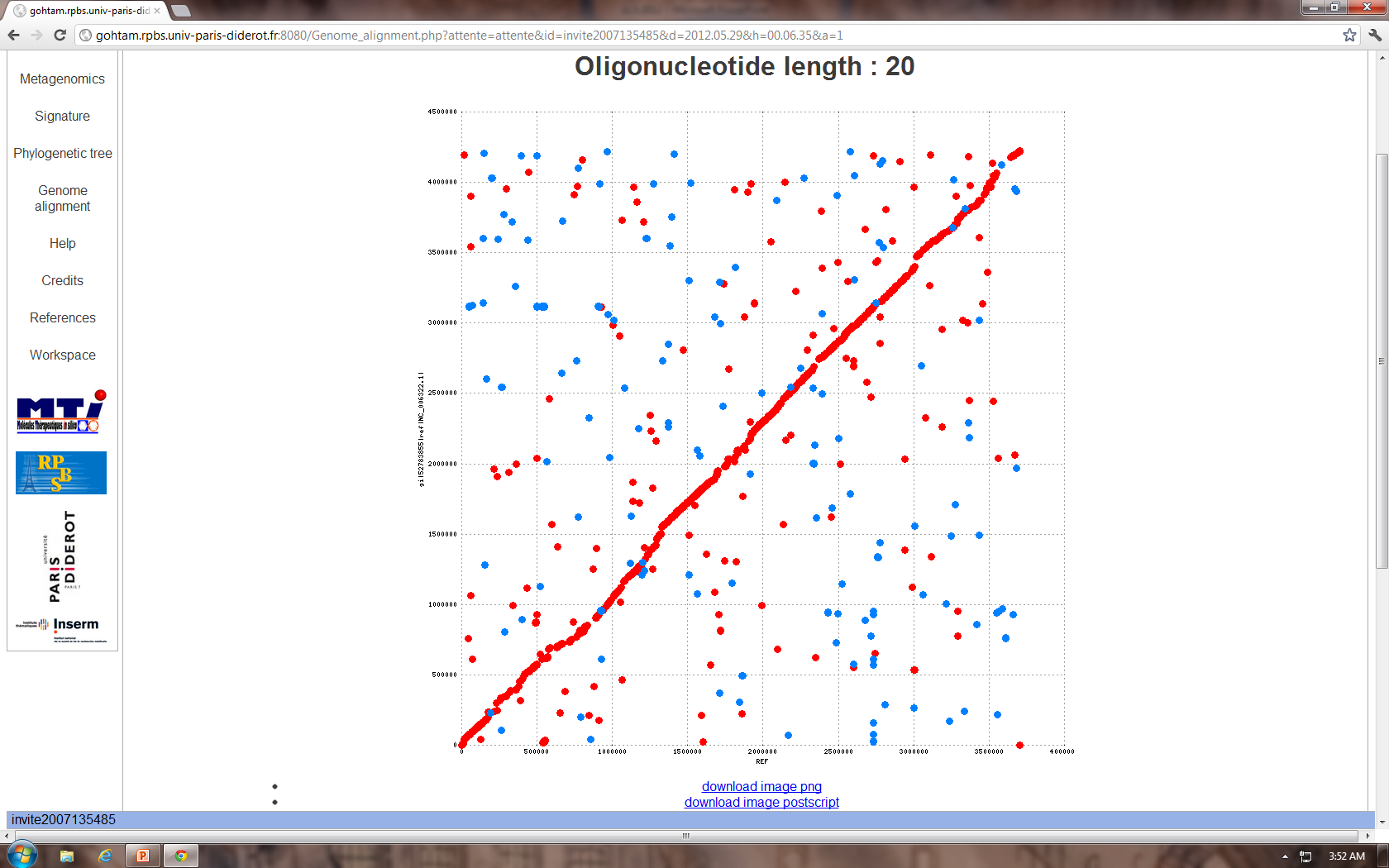


Figure: Bacillus pumilus and Bacillus li output from MUMmer 3

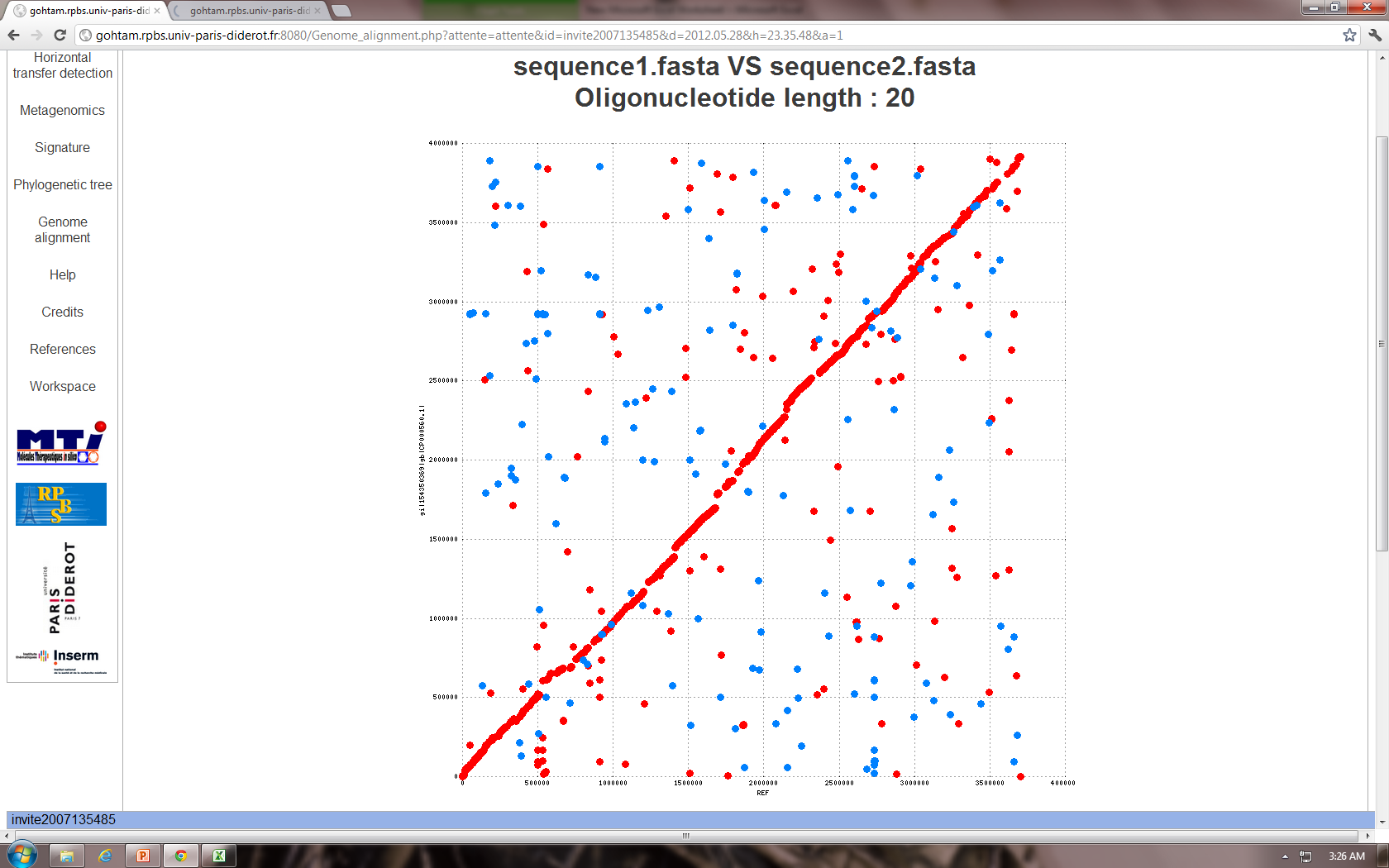


Figure: Bacillus pumilus and Bacillus amyloloquefaciens output from MUMmer 3

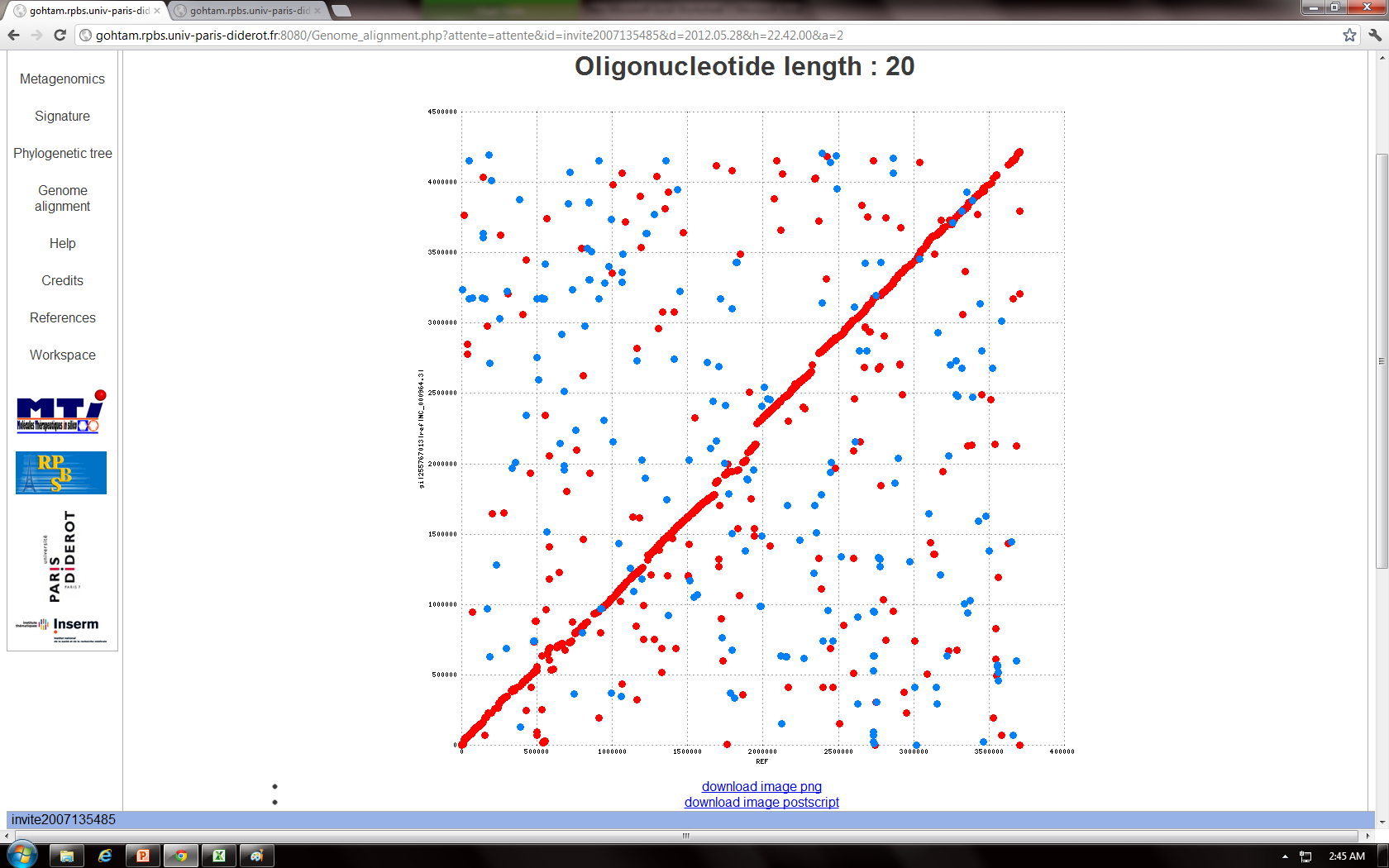


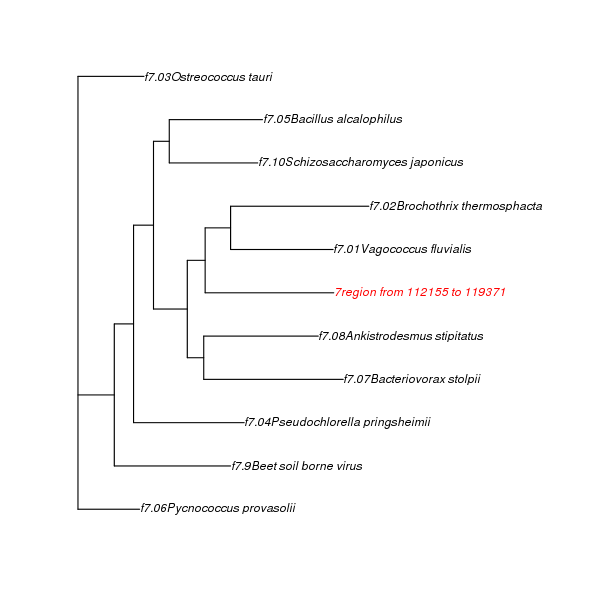
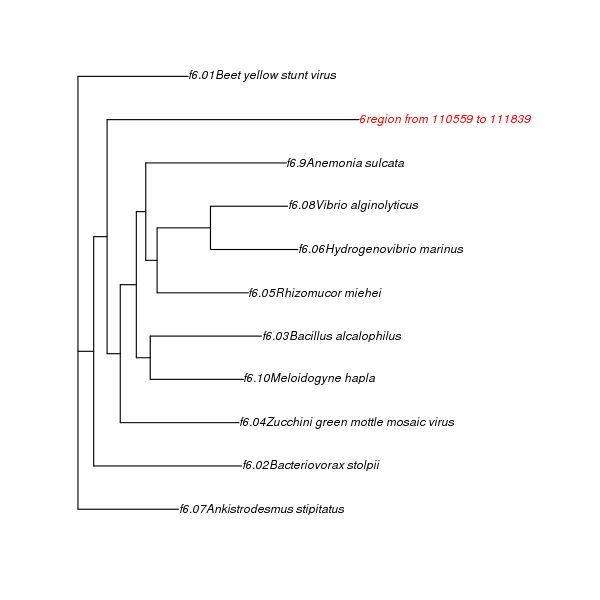
Figure: Bacillus pumilus and Bacillus subtilis output from MUMmer 3

**APPENDIX III**

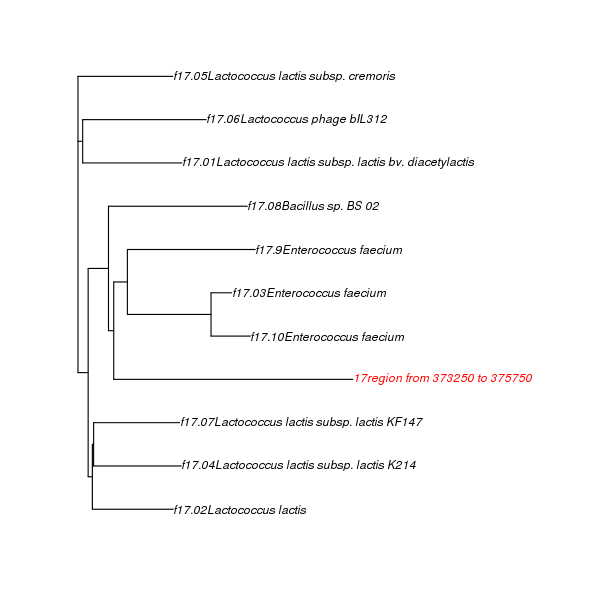
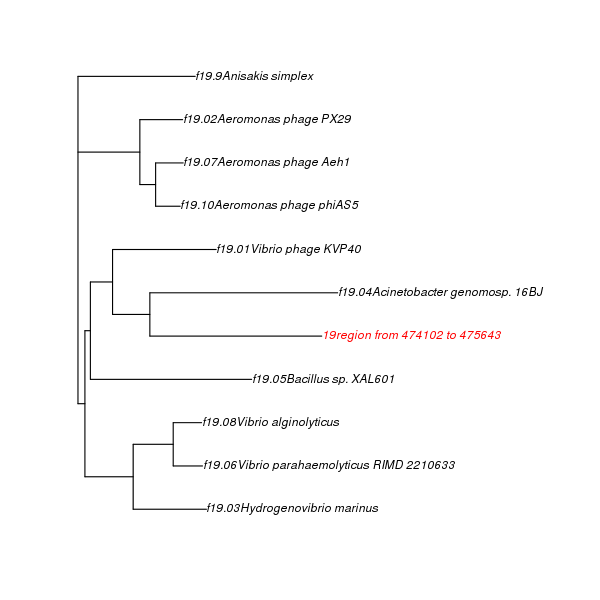
**BACILLUS SPECIES**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| fasta 6 (length:1281) region\_pic\_110559\_111839 \_ 1281 pb // Euclidean distance to genome : 295.62029987 A.U.// | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Distance (A.U.Euclidean) | | | rRNA | | Subject | | | | | | Strain | | | | | | | Reference length (pb) | | | | | | Origin | | | | | | Taxonomy | | | [similarity](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | | | | | | | [confidence](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | | | | |
| 237 | | | no | | [Beet yellow stunt virus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=35290) | | | | | |  | | | | | | | 10545 | | | | | | genomic | | | | | | Viruses | | | 1/5 | | | | | | | 3.0/5 | | | | |
| 238 | | | no | | [Bacteriovorax stolpii](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=960) | | | | | |  | | | | | | | 6372 | | | | | | genomic | | | | | | Bacteria | | | 1/5 | | | | | | | 3.0/5 | | | | |
| 240 | | | no | | [Bacillus alcalophilus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1445) | | | | | |  | | | | | | | 33139 | | | | | | genomic | | | | | | Bacteria | | | 1/5 | | | | | | | 3.0/5 | | | | |
| 241 | | | no | | [Zucchini green mottle mosaic virus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=111418) | | | | | |  | | | | | | | 6513 | | | | | | genomic | | | | | | Viruses | | | 1/5 | | | | | | | 3.0/5 | | | | |
| 243 | | | no | | [Rhizomucor miehei](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=4839) | | | | | |  | | | | | | | 29225 | | | | | | genomic | | | | | | Eukaryota | | | 1/5 | | | | | | | 3.0/5 | | | | |
| 243 | | | no | | [Hydrogenovibrio marinus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=28885) | | | | | |  | | | | | | | 32840 | | | | | | genomic | | | | | | Bacteria | | | 1/5 | | | | | | | 3.0/5 | | | | |
| 244 | | | no | | [Ankistrodesmus stipitatus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=3109) | | | | | |  | | | | | | | 6002 | | | | | | chloroplast | | | | | | Eukaryota | | | 1/5 | | | | | | | 3.0/5 | | | | |
| 244 | | | no | | [Vibrio alginolyticus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=663) | | | | | |  | | | | | | | 286704 | | | | | | genomic | | | | | | Bacteria | | | 1/5 | | | | | | | 3.0/5 | | | | |
| 245 | | | no | | [Anemonia sulcata](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=6108) | | | | | |  | | | | | | | 4197 | | | | | | genomic | | | | | | Eukaryota | | | 1/5 | | | | | | | 3.0/5 | | | | |
| 245 | | | no | | [Meloidogyne hapla](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=6305) | | | | | |  | | | | | | | 7188 | | | | | | genomic | | | | | | Eukaryota | | | 1/5 | | | | | | | 3.0/5 | | | | |
| fasta 7 (length:7217) region\_pic\_112155\_119371 \_ 7217 pb // Euclidean distance to genome : 248.154322807 A.U.// | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Distance (A.U.Euclidean) | | | rRNA | | Subject | | | | | | | Strain | | | | | Reference length (pb) | | | | | | | | Origin | | | | | | | Taxonomy | | | | | [similarity](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | | | | | | [confidence](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | |
| 172 | | | no | | [Vagococcus fluvialis](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=2738) | | | | | | |  | | | | | 4542 | | | | | | | | genomic | | | | | | | Bacteria | | | | | 2/5 | | | | | | 4.0/5 | |
| 176 | | | no | | [Brochothrix thermosphacta](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=2756) | | | | | | |  | | | | | 4629 | | | | | | | | genomic | | | | | | | Bacteria | | | | | 2/5 | | | | | | 4.0/5 | |
| 176 | | | no | | [Ostreococcus tauri](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=70448) | | | | | | |  | | | | | 71666 | | | | | | | | chloroplast | | | | | | | Eukaryota | | | | | 2/5 | | | | | | 4.0/5 | |
| 180 | | | no | | [Pseudochlorella pringsheimii](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=3072) | | | | | | |  | | | | | 17867 | | | | | | | | chloroplast | | | | | | | Eukaryota | | | | | 2/5 | | | | | | 4.0/5 | |
| 185 | | | no | | [Bacillus alcalophilus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1445) | | | | | | |  | | | | | 33139 | | | | | | | | genomic | | | | | | | Bacteria | | | | | 2/5 | | | | | | 4.0/5 | |
| 185 | | | no | | [Pycnococcus provasolii](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=41880) | | | | | | |  | | | | | 80211 | | | | | | | | chloroplast | | | | | | | Eukaryota | | | | | 2/5 | | | | | | 4.0/5 | |
| 186 | | | no | | [Bacteriovorax stolpii](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=960) | | | | | | |  | | | | | 6372 | | | | | | | | genomic | | | | | | | Bacteria | | | | | 2/5 | | | | | | 4.0/5 | |
| 187 | | | no | | [Ankistrodesmus stipitatus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=3109) | | | | | | |  | | | | | 6002 | | | | | | | | chloroplast | | | | | | | Eukaryota | | | | | 2/5 | | | | | | 4.0/5 | |
| 191 | | | no | | [Beet soil-borne virus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=46436) | | | | | | |  | | | | | 12295 | | | | | | | | genomic | | | | | | | Viruses | | | | | 2/5 | | | | | | 4.0/5 | |
| 191 | | | no | | [Schizosaccharomyces japonicus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=4897) | | | | | | |  | | | | | 23372 | | | | | | | | genomic | | | | | | | Eukaryota | | | | | 2/5 | | | | | | 4.0/5 | |
| fasta 17 (length:2501) region\_pic\_373250\_375750 \_ 2501 pb // Euclidean distance to genome : 339.628383212 A.U.// | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Distance (A.U.Euclidean) | | rRNA | | Subject | | | | | | | | | Strain | | | | | | Reference length (pb) | | | | | | Origin | | | | | | Taxonomy | | | | | [similarity](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | | | | | | [confidence](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | | |
| 144 | | no | | [Lactococcus lactis subsp. lactis bv. diacetylactis](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=44688) | | | | | | | | |  | | | | | | 21728 | | | | | | plasmid | | | | | | Bacteria | | | | | 3/5 | | | | | | 3.5/5 | | |
| 146 | | no | | [Lactococcus lactis](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1358) | | | | | | | | |  | | | | | | 64980 | | | | | | plasmid | | | | | | Bacteria | | | | | 3/5 | | | | | | 3.5/5 | | |
| 146 | | no | | [Enterococcus faecium](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1352) | | | | | | | | |  | | | | | | 65029 | | | | | | plasmid | | | | | | Bacteria | | | | | 3/5 | | | | | | 3.5/5 | | |
| 147 | | no | | [Lactococcus lactis subsp. lactis K214](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=889971) | | | | | | | | |  | | | | | | 29871 | | | | | | plasmid | | | | | | Bacteria | | | | | 3/5 | | | | | | 3.5/5 | | |
| 150 | | no | | [Lactococcus lactis subsp. cremoris](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1359) | | | | | | | | |  | | | | | | 42810 | | | | | | plasmid | | | | | | Bacteria | | | | | 3/5 | | | | | | 3.5/5 | | |
| 151 | | no | | [Lactococcus phage bIL312](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=151539) | | | | | | | | |  | | | | | | 15179 | | | | | | genomic | | | | | | Viruses | | | | | 3/5 | | | | | | 3.5/5 | | |
| 151 | | no | | [Lactococcus lactis subsp. lactis KF147](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=684738) | | | | | | | | |  | | | | | | 37510 | | | | | | plasmid | | | | | | Bacteria | | | | | 3/5 | | | | | | 3.5/5 | | |
| 152 | | no | | [Bacillus sp. BS-02](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=887059) | | | | | | | | |  | | | | | | 16543 | | | | | | plasmid | | | | | | Bacteria | | | | | 3/5 | | | | | | 3.5/5 | | |
| 152 | | no | | [Enterococcus faecium](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1352) | | | | | | | | |  | | | | | | 21344 | | | | | | plasmid | | | | | | Bacteria | | | | | 3/5 | | | | | | 3.5/5 | | |
| 153 | | no | | [Enterococcus faecium](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1352) | | | | | | | | |  | | | | | | 52890 | | | | | | plasmid | | | | | | Bacteria | | | | | 3/5 | | | | | | 3.5/5 | | |
| fasta 19 (length:1542) region\_pic\_474102\_475643 \_ 1542 pb // Euclidean distance to genome : 260.875140255 A.U.// | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Distance (A.U.Euclidean) | | rRNA | | | Subject | | | | | Strain | | | | | | Reference length (pb) | | | | | | | Origin | | | | | | Taxonomy | | | | | | [similarity](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | | | | | [confidence](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | | | | |
| 211 | | no | | | [Vibrio phage KVP40](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=75320) | | | | |  | | | | | | 244834 | | | | | | | genomic | | | | | | Viruses | | | | | | 1/5 | | | | | 3.5/5 | | | | |
| 234 | | no | | | [Aeromonas phage PX29](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=926067) | | | | |  | | | | | | 222006 | | | | | | | genomic | | | | | | Viruses | | | | | | 1/5 | | | | | 3.5/5 | | | | |
| 236 | | no | | | [Hydrogenovibrio marinus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=28885) | | | | |  | | | | | | 32840 | | | | | | | genomic | | | | | | Bacteria | | | | | | 1/5 | | | | | 3.5/5 | | | | |
| 237 | | no | | | [Acinetobacter genomosp. 16BJ](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=70347) | | | | |  | | | | | | 4089 | | | | | | | genomic | | | | | | Bacteria | | | | | | 1/5 | | | | | 3.5/5 | | | | |
| 237 | | no | | | [Bacillus sp. XAL601](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=460016) | | | | |  | | | | | | 6936 | | | | | | | genomic | | | | | | Bacteria | | | | | | 1/5 | | | | | 3.5/5 | | | | |
| 238 | | no | | | [Vibrio parahaemolyticus RIMD 2210633](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=223926) | | | | |  | | | | | | 5165771 | | | | | | | genomic | | | | | | Bacteria | | | | | | 1/5 | | | | | 3.5/5 | | | | |
| 239 | | no | | | [Aeromonas phage Aeh1](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=227470) | | | | |  | | | | | | 233234 | | | | | | | genomic | | | | | | Viruses | | | | | | 1/5 | | | | | 3.5/5 | | | | |
| 239 | | no | | | [Vibrio alginolyticus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=663) | | | | |  | | | | | | 286704 | | | | | | | genomic | | | | | | Bacteria | | | | | | 1/5 | | | | | 3.5/5 | | | | |
| 242 | | no | | | [Anisakis simplex](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=6269) | | | | |  | | | | | | 28064 | | | | | | | genomic | | | | | | Eukaryota | | | | | | 1/5 | | | | | 3.5/5 | | | | |
| 242 | | no | | | [Aeromonas phage phiAS5](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=879630) | | | | |  | | | | | | 225268 | | | | | | | genomic | | | | | | Viruses | | | | | | 1/5 | | | | | 3.5/5 | | | | |
| fasta 25 (length:14501) region\_pic\_588250\_602750 \_ 14501 pb // Euclidean distance to genome : 281.099062517 A.U.// | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Distance (A.U.Euclidean) | rRNA | | | Subject | | | Strain | | | | | | Reference length (pb) | | | | | | Origin | | | | | | | Taxonomy | | | | | | [similarity](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | | | | | [confidence](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | | | | | |
| 112 | no | | | [Petrotoga mobilis SJ95](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=403833) | | |  | | | | | | 2169548 | | | | | | genomic | | | | | | | Bacteria | | | | | | 4/5 | | | | | 4.0/5 | | | | | |
| 113 | no | | | [Staphylococcus sciuri](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1296) | | |  | | | | | | 17108 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 4/5 | | | | | 4.0/5 | | | | | |
| 114 | no | | | [Bacillus thuringiensis](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1428) | | |  | | | | | | 55712 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 4/5 | | | | | 4.0/5 | | | | | |
| 115 | no | | | [Lactococcus lactis subsp. cremoris](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1359) | | |  | | | | | | 42810 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 4/5 | | | | | 4.0/5 | | | | | |
| 115 | no | | | [Staphylococcus aureus subsp. aureus ST398](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=523796) | | |  | | | | | | 14362 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 4/5 | | | | | 4.0/5 | | | | | |
| 117 | no | | | [Lactococcus lactis subsp. lactis](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1360) | | |  | | | | | | 28546 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 4/5 | | | | | 4.0/5 | | | | | |
| 118 | no | | | [Lactococcus lactis](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1358) | | |  | | | | | | 64980 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 4/5 | | | | | 4.0/5 | | | | | |
| 118 | no | | | [Lactococcus lactis subsp. lactis bv. diacetylactis](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=44688) | | |  | | | | | | 53876 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 4/5 | | | | | 4.0/5 | | | | | |
| 119 | no | | | [Lactococcus lactis subsp. lactis](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1360) | | |  | | | | | | 20331 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 4/5 | | | | | 4.0/5 | | | | | |
| 119 | no | | | [Bacillus sp. BS-01](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=745376) | | |  | | | | | | 16492 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 4/5 | | | | | 4.0/5 | | | | | |
| fasta 26 (length:6501) region\_pic\_605750\_612250 \_ 6501 pb // Euclidean distance to genome : 215.495467287 A.U.// | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Distance (A.U.Euclidean) | | rRNA | | Subject | | Strain | | | | | | | | Reference length (pb) | | | | | | | Origin | | | | Taxonomy | | | | | | | | [similarity](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | | | | | | | [confidence](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | | | | |
| 132 | | no | | [Bacillus methanolicus MGA3](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=796606) | |  | | | | | | | | 19167 | | | | | | | plasmid | | | | Bacteria | | | | | | | | 3/5 | | | | | | | 4.0/5 | | | | |
| 133 | | no | | [Bacillus phage SPbeta](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=66797) | |  | | | | | | | | 134416 | | | | | | | genomic | | | | Viruses | | | | | | | | 3/5 | | | | | | | 4.0/5 | | | | |
| 134 | | no | | [Lactococcus lactis subsp. lactis bv. diacetylactis](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=44688) | |  | | | | | | | | 53876 | | | | | | | plasmid | | | | Bacteria | | | | | | | | 3/5 | | | | | | | 4.0/5 | | | | |
| 135 | | no | | [Lactococcus lactis subsp. cremoris SK11](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=272622) | |  | | | | | | | | 47208 | | | | | | | plasmid | | | | Bacteria | | | | | | | | 3/5 | | | | | | | 4.0/5 | | | | |
| 136 | | no | | [Lactococcus lactis subsp. lactis](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1360) | |  | | | | | | | | 48978 | | | | | | | plasmid | | | | Bacteria | | | | | | | | 3/5 | | | | | | | 4.0/5 | | | | |
| 136 | | no | | [Schizosaccharomyces pombe 972h-](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=284812) | |  | | | | | | | | 12571822 | | | | | | | genomic | | | | Eukaryota | | | | | | | | 3/5 | | | | | | | 4.0/5 | | | | |
| 137 | | no | | [Bacillus pumilus](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1408) | |  | | | | | | | | 6817 | | | | | | | plasmid | | | | Bacteria | | | | | | | | 3/5 | | | | | | | 4.0/5 | | | | |
| 137 | | no | | [Lactococcus lactis subsp. lactis](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1360) | |  | | | | | | | | 20331 | | | | | | | plasmid | | | | Bacteria | | | | | | | | 3/5 | | | | | | | 4.0/5 | | | | |
| 138 | | no | | [Bulleidia extructa W1219](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=679192) | |  | | | | | | | | 1419598 | | | | | | | genomic | | | | Bacteria | | | | | | | | 3/5 | | | | | | | 4.0/5 | | | | |
| 138 | | no | | [Enterococcus faecium](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1352) | |  | | | | | | | | 63135 | | | | | | | plasmid | | | | Bacteria | | | | | | | | 3/5 | | | | | | | 4.0/5 | | | | |
| fasta 27 (length:5672) region\_pic\_613250\_618921 \_ 5672 pb // Euclidean distance to genome : 214.448755603 A.U.// | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Distance (A.U.Euclidean) | | | rRNA | | | Subject | | | Strain | | | | | | Reference length (pb) | | | | | | Origin | | | | | | | Taxonomy | | | | | | [similarity](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | | | | | [confidence](http://gohtam.rpbs.univ-paris-diderot.fr:8080/help2.php) | | | | |
| 134 | | | no | | | [Bulleidia extructa W1219](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=679192) | | |  | | | | | | 1419598 | | | | | | genomic | | | | | | | Bacteria | | | | | | 3/5 | | | | | 4.0/5 | | | | |
| 143 | | | no | | | [Bacillus subtilis subsp. natto](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=86029) | | |  | | | | | | 85603 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 3/5 | | | | | 4.0/5 | | | | |
| 144 | | | no | | | [Bacillus phage SPbeta](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=66797) | | |  | | | | | | 134416 | | | | | | genomic | | | | | | | Viruses | | | | | | 3/5 | | | | | 4.0/5 | | | | |
| 146 | | | no | | | [Listeria monocytogenes 08-5578](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=653938) | | |  | | | | | | 77054 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 3/5 | | | | | 4.0/5 | | | | |
| 147 | | | no | | | [Listeria monocytogenes](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1639) | | |  | | | | | | 57780 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 3/5 | | | | | 4.0/5 | | | | |
| 149 | | | no | | | [Bacillus subtilis subsp. natto](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=86029) | | |  | | | | | | 65774 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 3/5 | | | | | 4.0/5 | | | | |
| 149 | | | no | | | [Streptococcus uberis 0140J](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=218495) | | |  | | | | | | 1852352 | | | | | | genomic | | | | | | | Bacteria | | | | | | 3/5 | | | | | 4.0/5 | | | | |
| 149 | | | no | | | [Enterococcus faecium TX0133a01](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=749522) | | |  | | | | | | 3073810 | | | | | | genomic | | | | | | | Bacteria | | | | | | 3/5 | | | | | 4.0/5 | | | | |
| 150 | | | no | | | [Eubacterium cylindroides T2-87](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=717960) | | |  | | | | | | 1966750 | | | | | | genomic | | | | | | | Bacteria | | | | | | 3/5 | | | | | 4.0/5 | | | | |
| 150 | | | no | | | [Enterococcus faecium](http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1352) | | |  | | | | | | 63135 | | | | | | plasmid | | | | | | | Bacteria | | | | | | 3/5 | | | | | 4.0/5 | | | | |

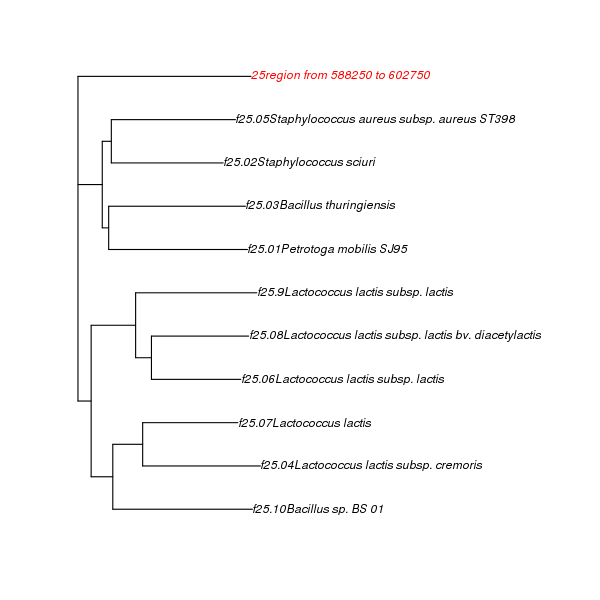
**Phylogenetic Trees obtained from GOHTAM.**

****

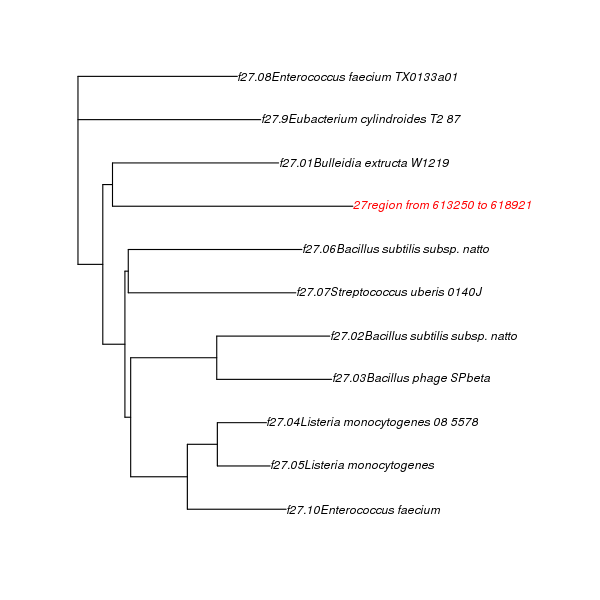
**Region: 6 Region 7:**

****

**Region 17: Region 19**



**Region 25 Region 26**

  
**Region 27**