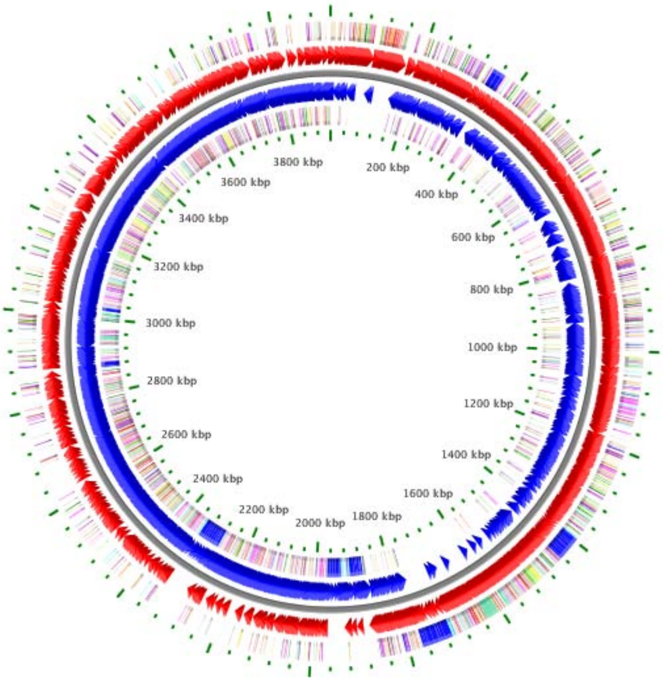
Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science

Bioinformatics assembly and analysis and annotation of the Bacillus amyloliquefaciens strain 5036 genome

Shahid Manzoor



Department of Animal Breeding and Genetics
Examensarbete 323
Uppsala 2010

Master’s Thesis, 30 HEC

One-Year Master’s Programme in Biology
– Bioinformatics

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science
Department of Animal Breeding and Genetics

Bioinformatics assembly and analysis and annotation of the Bacillus amyloliquefaciens strain 5036 genome

Shahid Manzoor

Supervisors:
Erik Bongcam-Rudloff, SLU, Department of Animal Breeding and Genetics
Johan Meijer, SLU, Department of Plant Biology and Forest Genetics
Examiner:
Göran Andersson, SLU, Department of Animal Breeding and Genetics

Credits: 30 HEC
Course title: Degree project in Animal Science
Course code: BI1021
Programme: One-Year Master's Programme in Biology
- Bioinformatics
Level: Advanced, A2E

Place of publication: Uppsala
Year of publication: 2010
Cover picture: Shahid Manzoor
Name of series: Examensarbete 323
Department of Animal Breeding and Genetics, SLU
On-line publication: http://epsilon.slu.se

Key words: Bacillus amyloliquefaciens UCMB-5036, Rhizosphere, Colonization, PGPR, Next generation sequencing, Paired-end Reads, Mapping assembly, *De novo* assembly

Table of Contents

Abbreviations:

UCM	Ukrainian Collection of Microorganisms
PGPR	Plant Growth Promoting Rhizobacteria
ISR	Induced Systemic Resistance
SAR	Systemic Acquired Resistance
PR gene	Pathogenesis-related gene
JA	Jasmonic Acid
Et	Ethylene
IST	Induced Systemic Tolerance
IAA	Indole acetic acid
VOCs	Volatile Organic Compounds
ACC	1-Aminocyclopropane-1-carboxylate
DNA	Deoxyribonucleic acid
HTML	Hyper text markup language
MIRA	Mimicking Intelligent Reads Assembly
ASCII	American Standard Code for Information Interchange
CAF	Common Assembly Format
HCR	High confidence region
LCR	Low confidence region
BASys	Bacterial Annotation System
BLAST	Basic Local Alignment Search Tool
rRNA	Ribosomal RNA

Abstract:

Plants are essential for both human and animal life, several agricultural practices have been adopted or are being developed in order to improve crop production by preventing or mitigating effects of biotic and abiotic stresses faced by plants. Biocontrol is a promising technique being cost effective, environment friendly and able to target pathogens difficult to control through traditional means. Bacteria that are associated with plant roots and exert beneficial effects on plant development are known as plant growth-promoting rhizobacteria (PGPR). Compared to plant growth-promoting Pseudomonas rhizobacteria, Bacillus species are advantageous due to their ability to form spores, which enable survival at high temperature, extreme pH, drought or mechanical and chemical stress and thus ideal for use as commercial biocontrol products. A PGPR may serve both as a biofertilizer and as a biological antagonist to pathogens. PGPR can be used for seed treatment supporting colonization in the rhizosphere on emerging roots. The mechanisms of protection may vary among different bacteria and include alteration of plant cell walls, production of biofilm on plant roots that prevent infection, competition for nutrients and space for growth and production of antibiotics and other damaging compounds to the pathogens. The *Bacillus amyloliquefaciens* strain *UCMB-5036* has potential to serve as a biocontrol agent. This genome was assembled from short paired-end reads of 75bp size generated by Illumina multiplexed technology using a mapping assembly technique against the already published complete genome of *B. amyloliquefaciens FZB42.* The total size of draft genome is **3918701 bp**, containing 4209 genes and 29 identified ribosomal *RNA* genes. After annotating the draft genome, it was observed that most of the genes responsible for colonization and plant growth-promotion in *FZB42* also are present in *UCMB-5036*. *De novo* assembly for unmapped reads indicate the presence of some novel genes i.e. genes which are present in *UCMB-5036* but not in reference genome of *FZB42.*

1 Background:

Plants especially field crops, are essential for our survival. However, the sessile nature of plants and the changing environment provide challenges in the form of abiotic stresses, e.g., drought and frost, and biotic stresses, (pathogens and insect pests) those plants must handle. To maximize crops yield several methods available or being developed to conquer these challenges.

In the past, before exploring different approaches to decrease the loss of production due to plant disease the losses were often more then *10 %* of the total field crop production (Strange and Scott, 2005). Several efforts including building of new agronomical practices, developing resistance breeding through traditional or genetic engineering have been developed to improve yield. Chemical insecticides and fungicides are today the major tool to control pathogens and pests and secure crop production. However, uses of these pesticides are hazardous, provide an extra strain on the environment and are often costly.

In contrast to chemical treatment, biocontrol where other organisms are used to control pathogens and insect pests, has a great potential to increase crop production. There are many advantages with biocontrol strategies over chemical treatment being more environments friendly and more effective against certain pathogens that are difficult to control by traditional means. For example biocontrol bacteria that colonize roots can outcompete soil borne pathogens difficult to control by chemicals. There are different ways to implement the biocontrol treatments like the use of natural enemies in the form of parasitoids or predators to insect pests, entomopathogens to insects or use beneficial microorganisms by spraying plants or by treating seeds. All these methods are used to reach the same goal that is to prevent insect pests or pathogen from damaging plants.

In this project the main focus was to assemble the genome of the *Bacillus amyloliquefaciens* strain, *UCMB-5036* using the *B.amyloliquefaciens* strain *FZB42* as a reference sequence. The *UCMB-5036* strain has potential as a biocontrol agent for rapeseed crops and wheat. The genome would then be analyzed and genes annotated to allow analysis of various aspects such as genes important for colonization and plants growth promotion etc. *DNA* sequence data was generated through the Illumina system in paired-end reads format with reads of maximum 75bp length.

1.1 Biocontrol:

Biocontrol is the reduction or control of unwanted organisms (pest insects, pathogens) by using other beneficial organisms. The most common example of biocontrol by using insects is the use of predatory insects like wasps that remove the unwanted pests. Different ways can be adopted to apply biocontrol like by releasing predatory insects in the field (Wang et al., 1999), and by seed treatment support bacteria colonization in the soil. One of the prominent examples of biocontrol is the use of Bacillus thuringiensis, which kills insects by producing toxin acting in the insect gut (Roh et al., 2007). Bacteria are the important and most strong candidates for the use of biocontrol agent, by acting as biofertilizers and as antagonists (biopesticides) of recognized root pathogens. Bacteria can colonize plants internally (endophytes) or, on plant surfaces (epiphytes), on the aerial parts (Phylosphere) or on belowground tissues (rhizosphere).

1.1.1 Role of Bacteria in biocontrol:

Bacteria can survive in a vast range of habitats such as animal intestines, soil etc. Due to their diversified nature bacteria can appear in different forms of life contrasted by symbionts and pathogens. Pathogens have the ability to deteriorate plant growth by consuming plant tissues and nutrients and symbionts have the potential to promote plant growth by facilitating uptake of nutrients. Different bacterial strains may vary in the protection mechanisms and can possibly use different methods in-parallel against pathogens like, enhance protection against pathogens by altering the plant cell wall (Benhamou et al., 1996), by producing biofilm on the plant roots which makes the plants less sensitive to infection (Bais et al., 2004; Rudrappa et al., 2008), by competing for nutrients and space for growth (Handelsman and Stabb, 1996), by the production of antibiotics and other damaging compounds to the pathogens (Raaijmakers et al., 2002; Whipps, 2001). Some bacteria can also stimulate plant growth by producing plant hormones (Timmusk et al., 1999; Martens et al., 1993).

1.1.2 Biocontrol with Bacillus:

Bacillus is a genus of gram-positive rod shaped bacteria, capable of growth in the presence of oxygen, and forms a unique type of resting cell called an endospore (Reva et al., 2004). The Bacillus genus has many members having potential to serve as biocontrol agent but the most important one is *B. amyloliquefaciens.* As compared to Pseudomonas rhizobacteria, Bacillus species were originally considered as typical soil bacteria, despite their well-established advantages for beneficial action on plant growth and biocontrol (Kloepper et al., 2004, Compant at el., 2005). Pseudomonas and Bacillus are both important for biocontrol but Bacillus species are more suitable and advantageous due to the presence of unique characteristic, that is spore formation, which make its survival possible at high temperature, extreme pH, drought, mechanical and chemical stress for long time periods. They also produce different kinds of antibiotic compounds like Iturin and Zwittermycin (Romero et al., 2007; Raaijmakers et al., 2002;Leifert et al., 1995) important for biocontrol and a number of different metabolites like biosurfactants (Edwards et al., 1992), chitinase and other enzymes which are responsible for degrading the fungal cell wall and improves the biocontrol efficiency (Priest et al., 1977; Pelletier et al., 1990).

Table I. Bacillus species known to mediate biocontrol. (Dunn et al., 2003; Schisler et al., 2004; Rudrappa et al., 2008; van Loon et al., 1998)

Bacillus Species
Bacillus amyloliquefaciens
Bacillus subtilis
Bacillus polymoxa
Bacillus licheniformis
Bacillus cereus
Bacillus pumilis
Bacillus fluorescens
Bacillus putida Bacillus chlororaphis
Bacillus agglomerans
Bacillus cloacae
Bacillus marcescens

There are several Bacillus base commercial products available on the market for biocontrol. For example, Kodiak is a biocontrol product used in the USA for the protection cotton to fungal disease (Jacobsen et al., 2004).

Table II for Bacillus based commercial biocontrol products. (Schisler et al., 2004)

Bacterial strain	Primary target	Product name
B. subtilis QST 713	Fungi and bacteria on fruits and vegetables	Serenade
B.subtilis GB03	Fungi on cotton and soybeans	Kodiak
B.licheniformis	Fungi on turf	Ecoguard
B.amyloliquefasciens B.subtilis GB122	Fungi on bedding plants	BioYield
B.subtilis MB1600	Fungi on cotton and soybeans	Subtilex
B.subtilis MB1600 Rhizobium	Fungi on soybeans	Hi Stick
B.pumilis GB34	Fungi on soybeans	Yield Shield

1.1.3 PGPR:

Different species of naturally occurring soil microorganisms colonize the rhizosphere and promote growth and productivity in crop plants (Kloepper et al., 1991; Lutenberg et al., 1991; Ryu et al., 2004). Rhizosphere is the narrow region of soil that is influenced by root secretions and associated soil microorganisms. Bacteria having potential to colonize this region are referred to as rhizobacteria, and obtain nutrients from root exudates. Plant growth-promoting rhizobacteria (PGPR) has the ability to promote plant growth by producing phytohormones like auxins, abscisic acid, gibberellins, ethylene and cytokinins (Varma et al., 2004). These hormones affect plant life in different ways like proliferation, cell and root elongation (Varma et al., 2004).

The biofertilisation is also an interesting field of biocontrol, that is using rhizobacteria to increase available nutrients in the soil (Bloemberg et al., 2001), for example nitrogen fixtures in root nodules help the plants to obtain nitrogen in exchange for nutrients (Denison et al., 2004).

There are different factors like genetic factors, growth substrates, indigenous bacteria along with abiotic factors such as humidity, pH and temperature that can influence the bacterial colonization (Garbeva et al., 2004; Smith et al., 1999; Varma et al., 2004). In plant root exudates the mainly included compounds are carbohydrates, amino acids and organic acids and the amount and composition of these metabolites change with the plant species (Lugtenberg et al., 2001; Nelson et al., 2004). Accordingly the different requirement of nutrients by bacteria may explain why bacteria are plant species specific for their colonization (Dunn et al., 2003). Rhizobacteria has the ability to alter the composition and amount of plant exudates (Lugtenberg et al., 2001).

1.1.4 Competition:

This is another way by which rhizobacteria can provide benefit to the plants because by establishment of colonies on important places like junctions between epidermal cells will limit the growth of harmful microorganisms by increasing competition for growth space and nutrients. These places are important due to the presence of more plant exudates. Bacillus also has an advantage in nutrient competition with other pathogens due to the formation of siderophores (Whipps, 2001), which solubilises iron and enhance transport into the bacteria by special receptors. This ability makes rhizobacteria able to utilize more iron as compared to other microorganisms in the competition for nutrients.

1.1.5 ISR:

Plants also develop inducible resistance mechanism to tolerate abiotic stress and resist pathogens, which can be of two types, Induced Systemic Resistance (ISR) and Systemic Acquired Resistance (SAR). Induced Systemic Resistance (ISR) is the process in which specific bacteria enhance the plant defense resulting in the reduction of severity of pathogenic diseases. Particular strains of species B. amyloliquefaciens, B. sutilis, B. pasteurii, B.cereus, B. pumilus, B. mycoides show significant reduction of severity of different plant diseases in a range of plant species (Kloepper et al., 2004). ISR is related with ultra structural changes and cytochemical alteration in plants during pathogen attack (Kloepper et al., 2004), which also induced protection against insects. Different Bacillus and Pseudomonas strains have potential to induced systemic resistance (ISR) in plants (van Loon et al., 1998; Iavicoli et al., 2003; Kloepper et al., 2004). In bean, B. pumilus SE34 induced ISR against the root-rot fungus F. Oxysporum f.sp. Pisi (Benhamou et al., 1996). It is not a direct defense, only activated after the plant attacked by pathogen or pest insects (Conrath et al., 2006). ISR induced by bacteria can depend on pathogenesis related (PR) proteins or ethylene (Ryu et al., 2004).

Bacteria produce salicylic acid (SA), a phytohormones which makes plants tolerant against pathogens by inducing SAR. SAR defense mechanism provide long-term systemic resistance to subsequent pathogens attack, and is dependent on SA which mediates activation of distal tissues producing PR proteins target different infectious organisms. Whereas, ISR signaling requires functional Jasmonic acid (JA) and Ethylene (Et), and does not depend on SA.

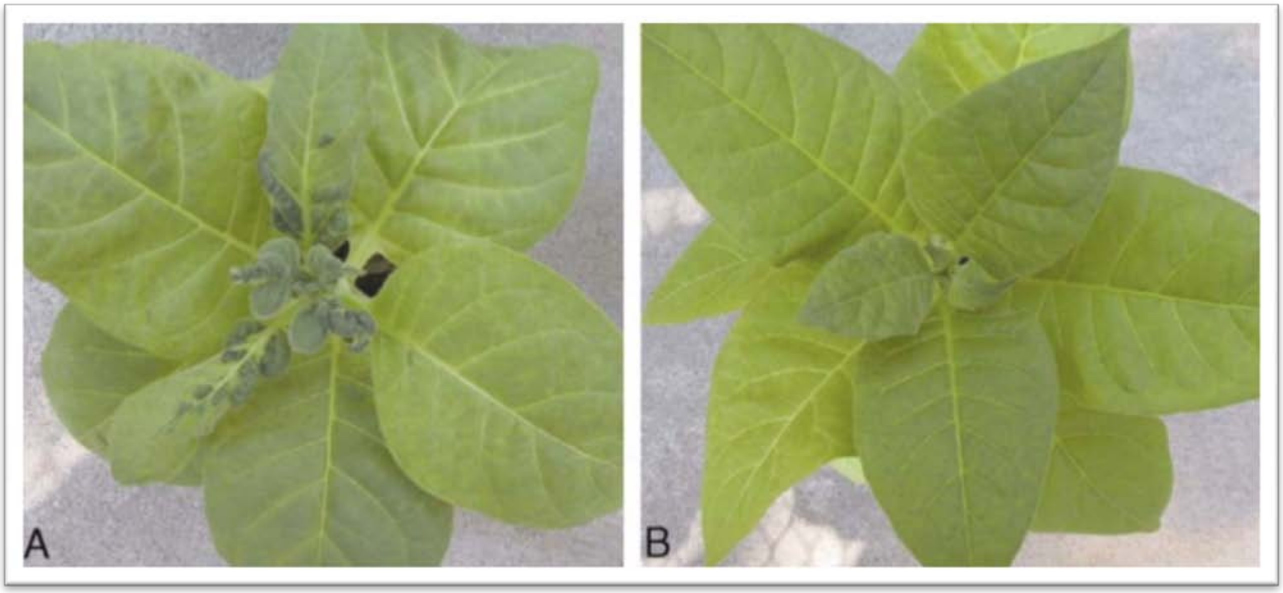


Figure I: Systemic protection against Cucumber mosaic virus on tobacco evoked by *Bacillus pumilus* strain *SE34*. A shows the non-bacterized, virus inoculated control. B shows protection resulting from treatment of tobacco at the time of transplanting with strain *SE34*. (Kloepper et al., 2004)

1.1.6 Primed State:

Primed state of a plant is the physiological condition in which plants respond better against both, biotic and abiotic stress. Priming in plants can also be attained through SAR. It is a plant defense state induced by infection with necrotizing pathogens and provides resistance against a range of harmful attackers (Ryals et al., 1996; Sticher et al., 1997). SAR induction requires deposition of the endogenous signaling molecule SA, which activates a large set of PR genes (Durrant and Dong, 2004).

1.1.7 IST:

As mentioned earlier some PGPR evoke physical or chemical changes associated to plant defense, a few reports have been published on that PGPR also evoke tolerance to abiotic stresses (drought, salt and nutrient deficiency or excess). A term ' Induced systemic tolerance' (IST) used for PGPR-induced physical and chemical changes in plants to augment tolerance to abiotic stress (Yang et al., 2009).

An important abiotic stress (drought) limits the plant growth and productivity. Earlier studies showed that the PGPR Paenibacillus polymyxa augment drought tolerance in Arabidopsis thaliana (Timmusk et al., 1999). Achromobacter piechaudii ARV8 is a PGPR strain that produces 1-aminocyclopropane-1-carboxylate (ACC) deaminase that develop drought tolerance in pepper and tomato plants (Mayak et al., 2004). The plant hormone ethylene endogenously regulates plant homeostasis in stress conditions like drought, which results in reduced root and shoot growth (Glick et al., 2007). So far, bacterial ACC releases plant stress and allow normal growth by degrading the ethylene precursor ACC (Glick et al., 2007).

Soil salinity is another important limiting factor for agricultural crops. In tomato seedlings the ethylene content exposed to high salt was reduced by applying Achromobacter piechaudii, showing functionality of bacterial ACC deaminase (Mayak et al., 2004). A new study also noted that the Bacillus subtilis GB03 strain augment tolerance to salt stress in Arabidopsis thaliana (Zhang et al., 2008), it produced some volatile organic compounds (VOCs), involved in IST (Kloepper et al., 2007)(Fig. II).

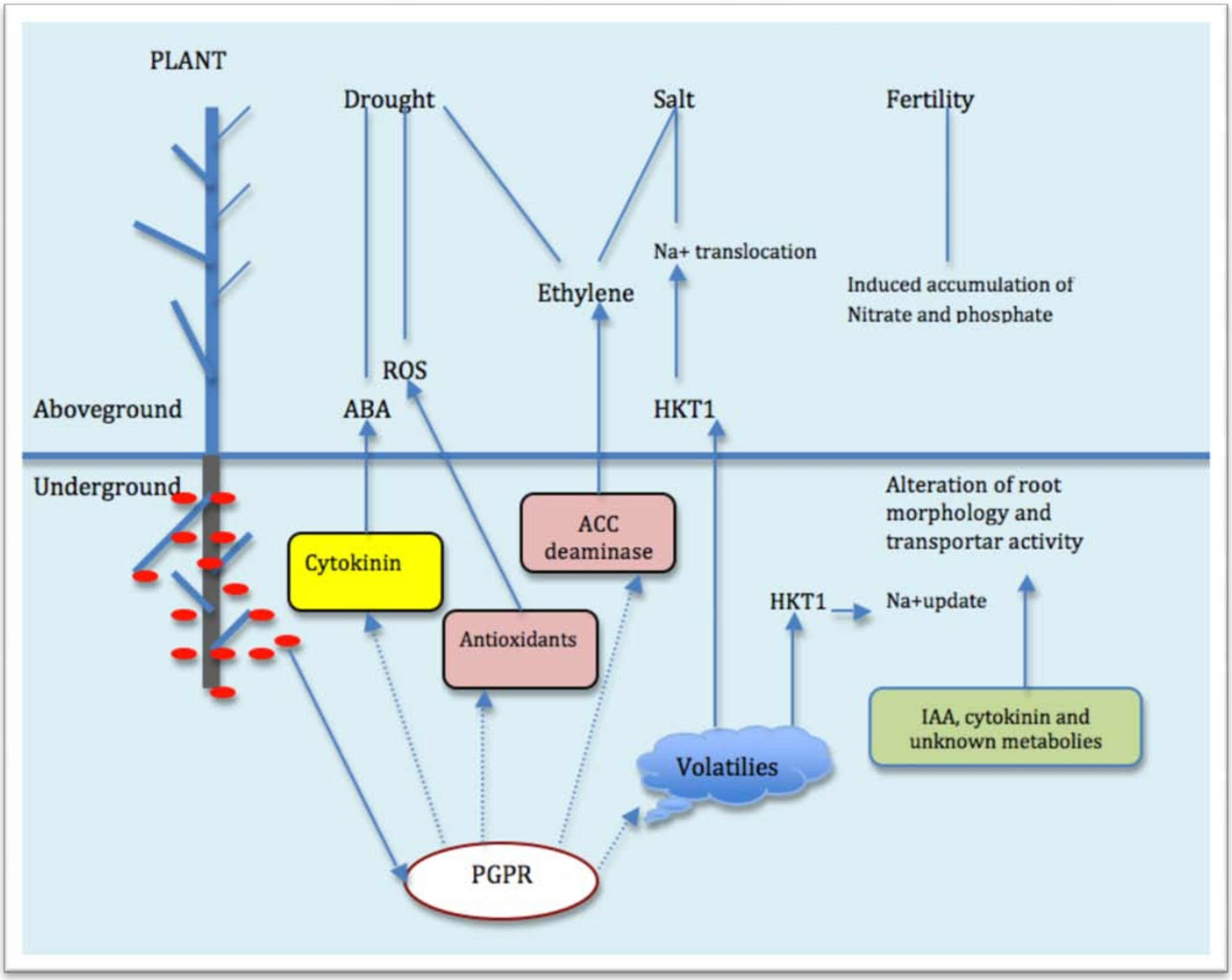


Figure II: IST against abiotic stresses (drought, salinity and fertility). IST elicited by PGPR against drought, salt and fertility stresses in roots and shoots. Broken arrows indicate bioactive compounds secreted by PGPR; solid arrows indicate plant compounds affected by bacterial compounds. Some PGPR strains, indicated in red on the plant roots, produce cytokinin and antioxidants such as catalase, which result in ABA accumulation and ROS degradation, respectively (Figueiredo et al., 2008, Kohler et al., 2008). Degradation of the ethylene precursor ACC by bacterial ACC deaminase releases plant stress and rescues normal plant growth under drought and salt stresses (Kohler et al., 2008, Mayak et al., 2004). The volatiles emitted by PGPR down regulate hkt1 expression in roots but up regulate it in shoot tissues, orchestrating lower Na+ levels and recirculation of Na+ in the whole plant under high salt conditions (Zhang et al., 2008). Production by PGPR of IAA or unknown determinants can increase root length, root surface area and the number of root tips, leading to enhance uptake of nitrate and phosphorous (Gyaneshwar et al., 2002, Adesemoye et al., 2008).

Uptake of sufficient nutrients from soil is also another important abiotic stress faced by plants. Different environmental factors inhibit uptake efficiency of plants for example, phosphorous is very reactive with iron, calcium and aluminum in soils, which can causative to precipitation of up to 90% of soil phosphorous (Gyaneshwar et al., 2002) , as a result making it unavailable to plants.

Some PGPR provide plant growth promotion through solubilization and increased uptake of phosphate (Gyaneshwar et al., 2002). Some PGPR associated with the promotion of root development (Mantelin et al. 2004) and change root architecture by producing phytohormones like indole acetic acid (IAA) (Kloepper et al., 2007)(Fig II), resulting in increased root surface area and tips. The root surface area and tips are used for nutrient uptake, meaning that PGPR enhance nutrient uptake by stimulating root development.

2 Aims:

The over all aim of this project was primarily to assemble the genome of the *Bacillus amyloliquefaciens* strain *UCMB-5036* and secondly to study the similarities and differences relative to the known genome of *Bacillus amyloliquefaciens FZB42* available at *NCBI*. B. *amyloliquefaciens FZB42* is a plant growth-promoting rhizobacteria, which has potential use as a biocontrol agent. So in this project our attention was mostly directed at comparative studies aiming to identify coding sequences in the *UCMB-5036* strain that could be involved in colonization and ultimately promote plant growth by providing protection of plants to pathogens by suppressing disease.

3 Material and Methods:

3.1 Data Isolation:

The *Bacillus amyloliquefaciens* strain, *UCMB-5036,* is an important tool for studies of bacteria-plant association and the role of bacterial in plant growth promotion through stimulating plant growth and suppressing soil borne pathogens by producing secondary metabolites. This strain was identified as belonging to the B. amyloliquefaciens group based on phenotypic analysis (Reva et al., 2004). Bacillus have several advantages over other bacteria for biocontrol , in that they are easy to cultivate and store. *UCMB-5036* is a UCM (Ukranian Collection of Microorganisms) strain, isolated from cotton plant.

3.2 Genome Sequencing:

The bacterial genome of strain *UCMB-5036* was sequenced with the A-C system at academic hospital, Uppsala, through multiplexed sequencing method, using Illumina sequencing technology on a fully automated Genome Analyzer. A total of 65,44,152 short paired-end reads of length 75 bp were generated, with the average length of inserts size 230 bp. Paired-end library size ranging from 300 – 400 bp including adaptor and primer sequences, which was estimated from electropherogram summary (Figure IIIa, IIIb, IIIc) with ladder peak table (Table I). All this information about the sequenced paired-end reads were provided by the vender company.

Electrophoresis summary

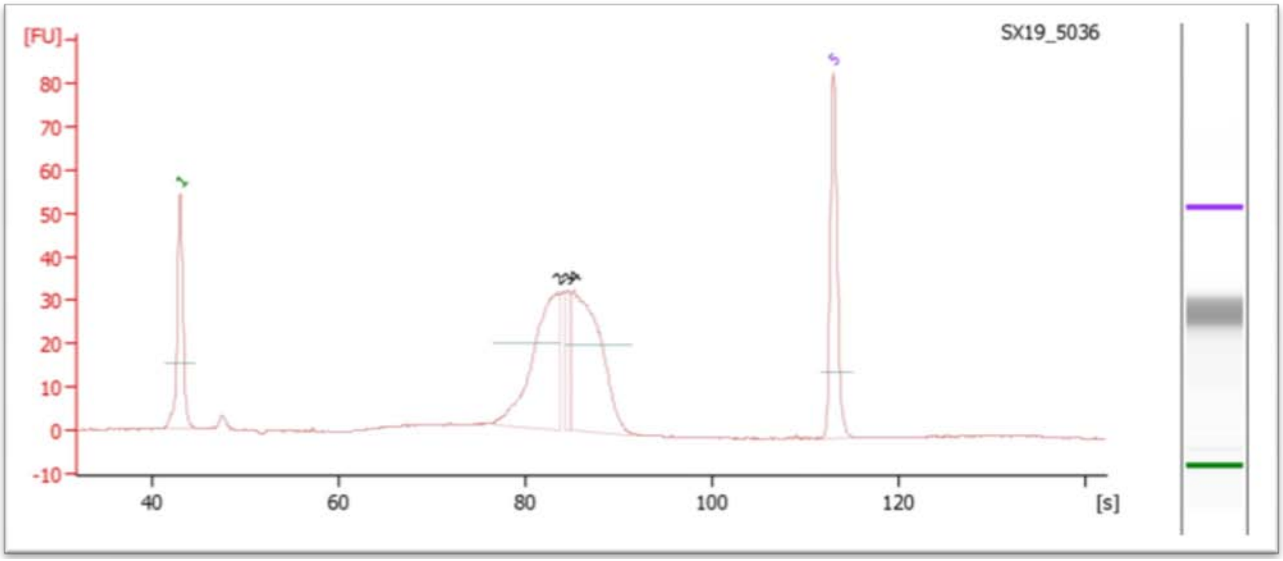


Figure IIIa: shows detection of sequence peaks at 80-85sec elution times for 5036.

The major advantage of Illumina multiplexed sequencing technology is it increases experimental throughput and reduce time and cost.

Ladder

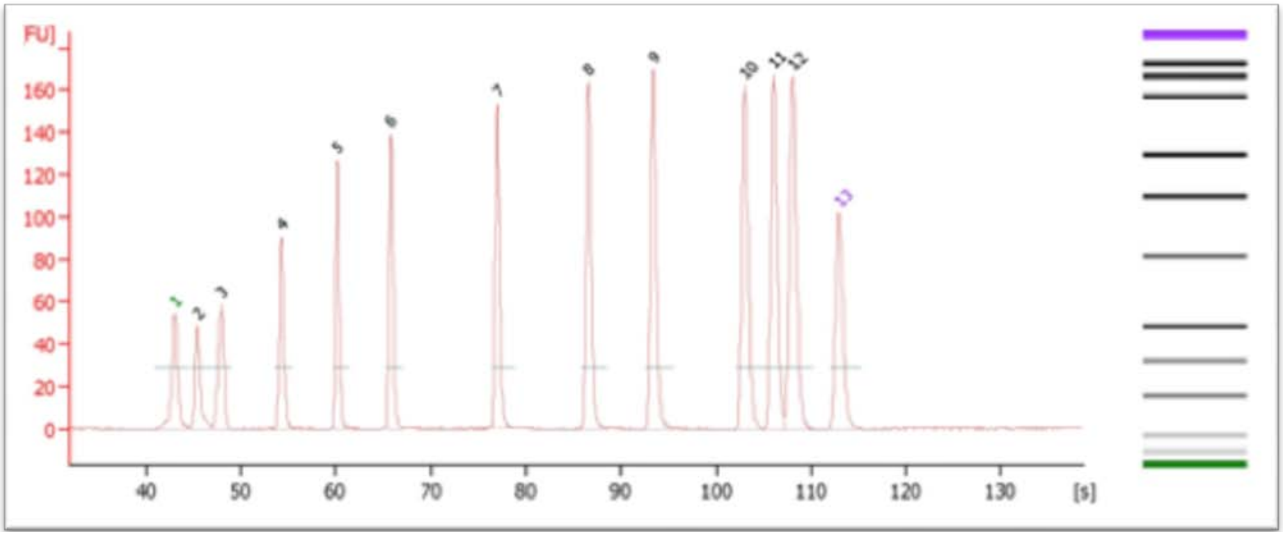


Figure IIIb: Peak 7 and 8 near elution time 80-85secs corresponds to fragment length 300-400bp as described in Table III.

Table III: Peak table for ladder indicating fragment length of paired end reads.

Peak	Size (bp)	Concentration (ng/µl)	Molarity (nmol/l)	Observations
1	15	4.20	424.2	Lower Marker
2	25	4.00	242.4	Ladder Peak
3	50	4.00	121.2	Ladder Peak
4	100	4.00	60.6	Ladder Peak
5	150	4.00	40.4	Ladder Peak
6	200	4.00	30.3	Ladder Peak
7	300	4.00	20.2	Ladder Peak
8	400	4.00	15.2	Ladder Peak
9	500	4.00	12.1	Ladder Peak
10	700	4.00	8.7	Ladder Peak
11	850	4.00	7.1	Ladder Peak
12	1,000	4.00	6.1	Ladder Peak
13	1,500	2.10	2.1	Upper Marker

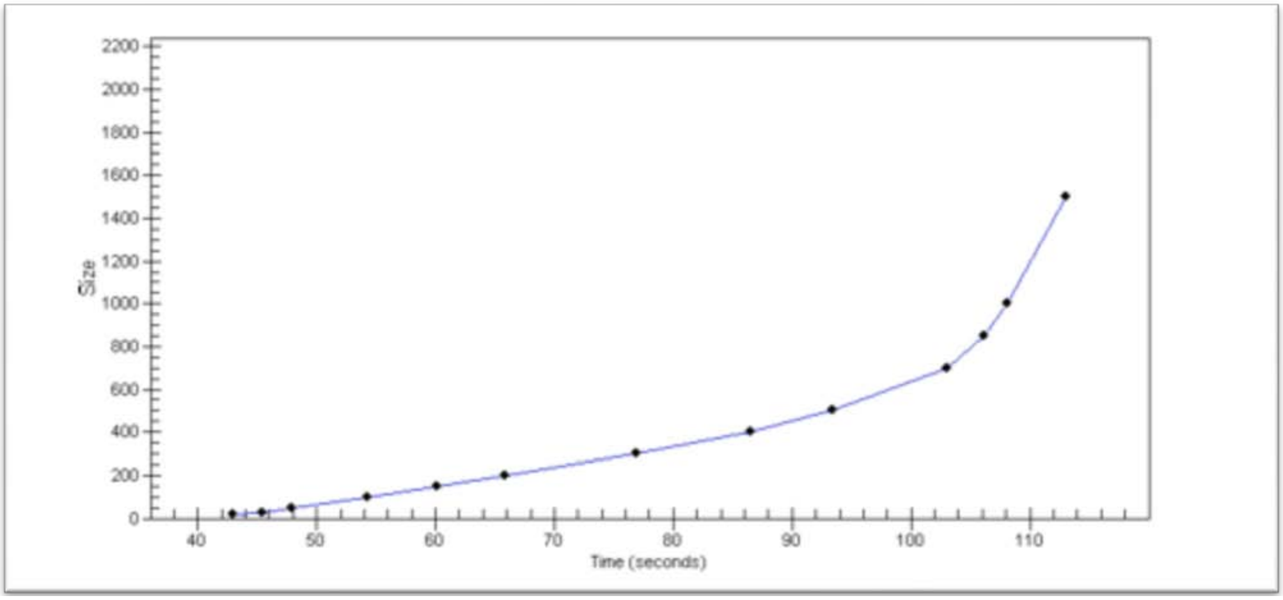


Figure IIIc: Standard curve plot of sequencing illustrating reads fragment length is approximately 300-400bp at the interval of 80 sec.

3.2.1 Paired-end Reads:

For our project we decided to use short paired-end reads because, paired-end libraries are important for the detection of large and small deletions, insertions, inversions and other rearrangements. These reads also provide great advantage to identify repetitive sequence elements (Illumina-Sequencing Technology, http://www.illumina.com/).

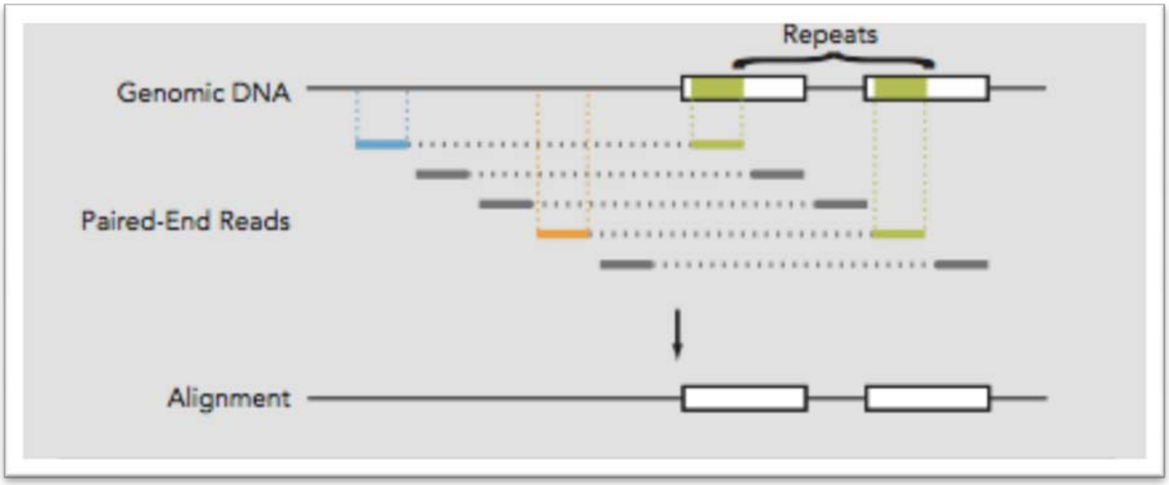


Figure IV: Unique Alignment of paired-end reads in repeats. Reads in repeats (green) can be unambiguously aligned in complex genomes. Each read is associated with a paired read (blue or orange) and the separation between read pairs is known from the fragment size of the input DNA.

3.3 Genome assembly and annotation:

There are two fundamentally different approaches for sequence assembly from short read data (Pop et al., 2008). The first is known as *de novo* assembly in which reads are assembled together to form a new sequence which is previously unknown. Second is mapping assembly, in which reads are assembled against a template sequence. Idea behind is that to assemble a new sequence, which is similar but not necessarily identical to the template sequence. For our project mainly we used mapping assembly against already published genome of *B. amyloliquefaciens* strain *FZB42* as reference sequence and for finishing (filling gaps) the consensus sequence also performed *de novo* assembly and calculated anchors from 104 contigs with length greater than 0.5 kbp against reference genome.

There are several genome assemblers available publicly, such as *Mosaik, MIRA3*, and *MAQ*. We chose *MIRA3* (Chevreux et al. 1999) for our project because of its distinct features for assembly like, handling repetitive elements, built-in clipping and masking of bad sequences, tagging of regions of interest in the consensus sequence, and support for a range of input and output formats.

3.3.1 Mapping Assembly:

For mapping the genome of *UCMB-5036,* first data was preprocessed according to assembler requirement and to remove the adaptor sequence contamination from the read data.

3.3.2 Read Data Preparation:

I) Format Conversion

The read data files generated from Illumina Genome Analyzer were initially in *scarf ASCII* format containing sequence identifiers, reads, and their quality values ranging from 64-104 *ASCII* characters. We converted *scarf ASCII* format to *FASTQ* format to fulfill the requirement of assembler, using a simple Perl script.

II) Adaptor Sequence Screening

For *DNA* sequencing primers and adaptor are required, which may contaminate reads with adaptor and primer sequence during *DNA* sequencing process. To avoid the effect of adaptor and primer sequence on the alignment, we performed screening of read data for adaptor sequence. We used *ssaha2* (from Sanger Center) for screening, using the following parameters:

```
“-output ssaha2 -kmer 8 -skip 1 -seeds 1 -score 12 -cmatch 9 -ckmer”
```

Table IV. Shows the adaptor and primer sequences provided by the vender.

	Sequence
Adaptor 1	5' GATCGGAAGAGCACACGTCT 3'
Adaptor 2	5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT 3'
Primer 1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Primer 2	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT

MIRA3 also uses simple key-value straindata file for assembling, that was generated from the read data *FASTQ* file by executing the following grep command:

```
“ grep "@" file.fastq     | sed -e 's/@//'     | cut -f 1    | cut
-f 1 -d ' '    | sed -e 's/$/ bas_5036/' “
```

3.3.3 Mapping Reads to the Reference Sequence:

Processed read data was mapped against the *B. amyloliquefaciens FZB42* genome (available at *NCBI*), using *MIRA3* assembler. For reference sequence GenBank format (gbf) was used because after assembly *MIRA3* can utilize annotation information, present in gbf file. The file produced by *ssaha2* was also provided to the assembler, which contained information for adaptor sequences contaminated reads. Mapping assembly performed using the following parameters:

“ --project=fz5036 --job=mapping,genome,accurate,solexa -AS:nop=2 -SB:lsd=yes:bsn=FZB42:bft=gbf:lb=yes:bbq=30 SOLEXA_SETTINGS -CO:msr=no -GE:uti=no:tismin=180:tismax=280 -CL:msvs=yes:qc=yes -LR:lsd:yes:ft=fastq:fqqo=64 -AL:mrs=70 “

Of the total reads, 90.20% mapped against the reference genome with average sequencing coverage of 69-fold across the entire genome.

3.3.4 Assembly Visualization:

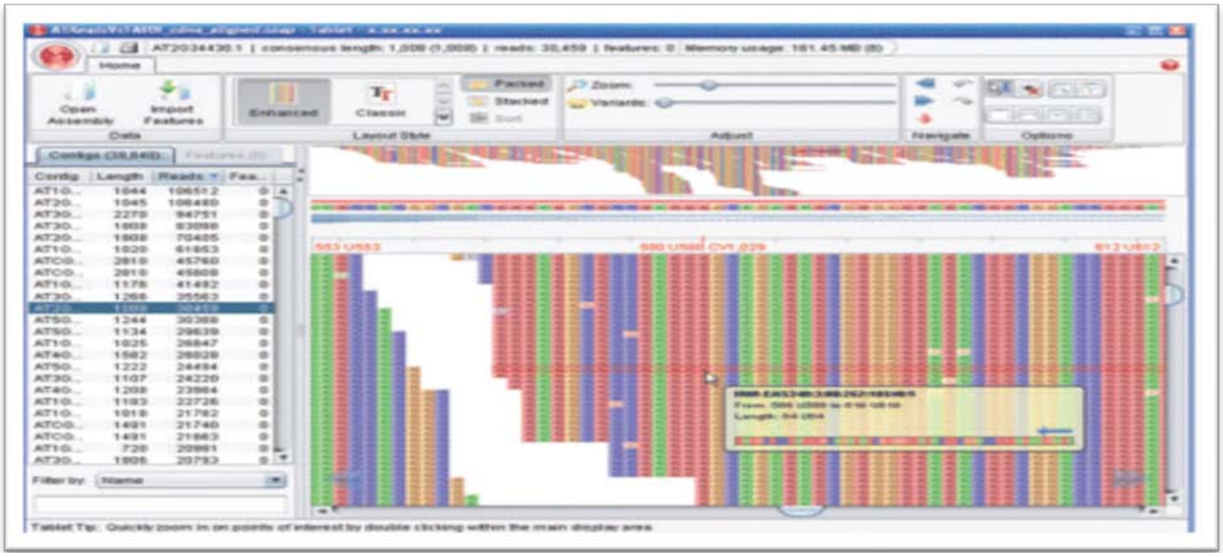
MIRA3 has produced output in different formats e.g. *CAF*, *FASTA*, *ACE*, *HTML* and simple *TEXT*. There are a number of visualization tools available for viewing these output files for example, Consed (Gordon et al., 1998), Hawkeye (Schatz et al., 2007), EagleView (Huang and Marth, 2008), MapView (Bao et al., 2009), SAMtools' tview (Li et al., 2009) and Maqview (http://maq.sourceforge.net/maqview.shtml). We used Tablet visualization tool (Milne et al. 2009) to visualized the aligned reads from *ACE* file produced by MIRA3, because it is a lightweight, high performance and memory efficient assembly viewer.

3.3.4.1 Performance Comparison:

For handling the assembly data in viewers there are basically two different approaches, the first one is memory based, in which all the data is loaded into memory, or disk cached. These applications are much faster for viewing the data but the size of data is limited by the amount of available memory. The second one is the cache-based approach that can display large data set by using less memory at the cost of speed. On the other hand Tablet is the product of hybrid approach (Combination of both approaches memory based and cache based). It only keeps the skeleton layout of the reads in memory, which contains limited information on each read like internal ID, its position against the consensus or reference and its length. The nucleotide data with its supplementary information (read's name and its orientation) is kept in an indexed disk cache and only accessed when required. So this hybrid approach offers maximum functionality by using relatively low memory.

In comparison for data indexing/loading times and memory utilization between different tools for an assembly file containing approximately 2.9 million Illumina Solexa reads of length 51 bp, the cache based viewers (Maqview, Mapview, tview) constantly use memory between 35MB to 70MB for viewing the data, with indexing times varying from 10 s to 50 s, on the other hand memory based viewers use memory and time accordingly Hawkeye (5500 MB; 107 s), Consed (2600 MB; 73 s) and

EagleView (2450 MB; 98 s). In contrast Tablet loads data in 25 s, and uses only 175 MB of memory (Iain Milne et al., 2009). Out put of alignment file with tablet is shown in Figure V.



AT2G34430.1 | consensus length: 1,008 (1,008) | reads: 30,459 | features: 0 | Memory usage: 161.45 MB (8)

Contig	Length	Reads	Fea...
AT1G...	1044	106512	0
AT2G...	1045	108480	0
AT3G...	2270	94751	0
AT3G...	1809	83098	0
AT2G...	1809	70405	0
AT1G...	1020	61653	0
ATCG...	2810	45760	0
ATCG...	2810	45808	0
AT1G...	1176	41482	0
AT3G...	1266	35563	0
AT2G...	1008	30459	0
AT5G...	1244	30368	0
AT5G...	1134	29639	0
AT1G...	1025	26847	0
AT4G...	1582	26028	0
AT5G...	1222	24494	0
AT3G...	1107	24220	0
AT4G...	1208	23984	0
AT1G...	1183	22726	0
AT1G...	1018	21782	0
ATCG...	1491	21740	0
ATCG...	1491	21663	0
AT1G...	720	20981	0
AT2G...	1808	20793	0

Tablet Tip: Quickly zoom in on points of interest by double clicking within the main display area

Figure V: Output of tablet visualization tool. Tablet showing Illumina Solexa reads against reference sequence.

3.3.5 Genome Annotation:

A consensus sequence was finally produced from the *MIRA3* alignment of short reads of *UCMB-5036* against *FZB42* reference genome. We called this consensus sequence the '*draft genome'*, submitted to an annotating pipeline, *BASys* 'Bacterial Annotation System' (Van Domselaar GH et al. 2005), to predict functional elements on the genome, and to attach biological information to such elements. *BASys* is a web based annotation tool, which provide automated, in-depth annotation of bacterial genomic sequences. It uses more than 30 different tools and databases to produce approximately 60 separate annotations for each gene (http://basys.ca/basys/cgi/programs_and_dbs.pl). *BASys* provides genomic information in textual, html and graphical forms.

BASys annotation pipeline complete the process in three parts:

(a) A front-end web interface used for submitting raw genomic sequence, for scheduling annotation and monitoring or reporting the output results.

(b) Second annotation engine perform sequence analysis and generate the annotation.

(c) Third is the reporting system responsible for generating HTML, graphical and textual outputs.

3.3.6 De *novo* Assembly:

Total of 6.5 millions paired-end reads (65,44,152 reads), generated from Illumina Genome Analyzer were used for *de novo* assembly. All the reads were input into the *MIRA3* assembler for building the *de novo* assembly using the following parameters:

“-project=ucmb-5036 --job=denovo,genome,accurate,solexa COMMON_SETTINGS -GE:kcim=yes -SK:bph=20:hss=8 -OUT:ora=yes:org=yes -AS:sd=yes:sep=yes SOLEXA_SETTINGS -DP:ure=yes

-CO:msr=no -GE:uti=no:tismin=180:tismax=280 -LR:lsd:yes:ft=fastq:fqqo=64:rns=solexa -AL:mrs=65 -ED:ace=yes”

In *de novo* assembly total 2208 contigs with different sizes were produced, which were filtered to discard the small contigs through *MIRA3* utility (Convert _Project) from the *CAF* file. Total 107 contigs with length greater than 0.5 kbp were generated in *FASTA* format. These 107 contigs were used to calculate anchors against the reference sequence, using a multiple genome comparison program *Murasaki* ((Popendorf et al., 2007).

“Anchors are short well conserved subsequences between each contig and the reference sequence”

3.4 *De novo* Assembly of Unmapped reads:

In mapping assembly, 9.34 % (611550 reads) of the total reads were not mapped against the reference sequence due to different reasons as follows:

I) After clipping the bad quality bases the reads remain too short to be mapped by the assembler.

II) Some reads appeared to be chimeric and/or contaminated ones.

III) Some reads cannot be mapped due to the non-similarity with the reference sequence.

The identifiers of all the above-mentioned unaligned reads were stored into a separate debris file by the assembler. We were interested to perform *de novo* assembly for these unaligned reads, to do this we extracted all debris reads from read data file. Steps as follows:

I) Read data file was loaded into *MySQL* database, by converting *scarf ASCII* format to *scarf* numeric, because there was problem with *MySQL* Database Management System to load qualities values in *scarf ASCII* format into the database.

II) Debris file containing identifiers of unaligned reads was loaded into a separate table in the database.

III) *SQL* query was designed to extract the complete reads from the read data file against the identifiers of debris file.

IV) All extracted reads were exported from the database into the text file, and again converted into the *FASTQ* format. All these formats conversions were performed with the Perl scripts.

Extracted debris reads were input into the MIRA3 assembler for *de novo* assembly, using the following parameters:

“--project=ucmb-5036 --job=denovo,genome,accurate,solexa SOLEXA_SETTINGS -GE:uti=no:tismin=180:tismax=280 -LR:lsd:yes:ft=fastq:fqqo=64 -AL:mrs=70”

A total of 26426 contigs were produced for unaligned reads of *UCMB-5036* genome. 104 contigs with the length of greater than 0.5 kbp were used for local alignment, using *BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi)* for similarity search against *NCBI* genome database.

4 Results:

4.1 Mapping read data to the reference genome:

Genomic *DNA* was extracted from *Bacillus amyloliquefaciens* strain *UCMB-5036* and whole genome shotgun sequences were obtained using the Illumina Genome Analyzer. A total of 65,44,152 paired-end reads of 75 bp lengths were generated, with the average length of inserts of paired-end reads at 230 bp.

The generated read data was mapped against the *B. amyloliquefaciens FZB42 as reference genome* using MIRA3 assembler. Of the total reads, 5902827 (90.20 %) reads could be mapped to the reference genome with 69-fold average coverage across the whole genome. Consensus sequence of size 3.9 MB was produced; statistics about produced consensus sequence are summarized in table V.

Table V. Summary of consensus sequence of *UCMB-5036*:
Genomic *DNA* was extracted form B. amyloliquefaciens *UCMB-5036*, isolated from cotton plant. The paired-end reads were generated by the Illumina Genome Analyzer. The generated reads data was mapped against the B. amyloliquefaciens *FZB42* as reference genome using MIRA3 Assembler.

Strain	Consensus Sequence (bp)	Total reads (million)	Mapped reads (%)	Avg. Coverage	GC (%)	Max. Coverage
UCMB-5036	3918701	6.5	90.2	69.05	46.49	420

MIRA3 assembler tagged important regions in the consensus sequence, recognizable by finishing tools. These regions include strong and weak repetitive markers (SRMc, WRMc), possible deletions or no coverage in consensus (MCVc) and conflicting bases in the assembly that could not be decided by assembler (IUPc). SNPs were tagged as SROc in the consensus sequence. Statistics regarding the tagged regions are summarized in table VI.

Table VI. Summary of tagged locations in consensus sequence:
MIRA Assembler tagged regions in consensus sequence, important for finishing genome-finishing process. These tags are understandable by different genome finishing tools e.g., gap4, and consed.

MCVc	SRMc	WRMc	SROc	Indels	IUPc	No. Coverage bp
589	23	55	44254	1135	663	162636

4.2 *De novo* assembly for read data and calculates anchors:

The read data generated by the Illumina Genome Analyzer was input into the *MIRA3* assembler for *de novo* assembly. A total of 2208 contigs with the average length of 1,903 bp were produced. The total size of the produced contigs was 4,200 kbp, and that of 107 contigs with length greater than 0.5 kbp is 4,000 kbp, that is, 95.2% of all the contigs we used for the next step of calculating anchors. Statistics regarding the mapped reads and coverage are summarized in table VII.

Table VII. Summary of mapped reads and coverage:

The read data generated by Illumina genome analyzer was input into the assembly software MIRA3 and a set of contigs was produced. Of the all reads, 98.36 % reads were aligned to produced contigs with average coverage of 127-fold

Strain	Total no. reads	Mapped reads no.	Mapped reads %	Avg. Coverage	Max. Coverage
UCMB-5036	65,44,152	64,37,424	98.36	127-fold	2725

Table VIII. Summary of contigs produced by MIRA:

Total of 2208 contigs were produced by *de novo* assembly, with the size 42,00 kbp and that of 107 contigs with length greater than 0.5 kbp is 4,000 kbp.

Strain	No. of total contigs	Total length (kbp)	N50 contig size (bp)	No. of contigs greater than 0.5 kbp	No. of contigs longer than 1 kbp
UCMB-5036	2208	42,00	81604	13	95

107 contigs with length greater than 0.5 kbp were used to calculate anchors against the reference sequence, using multi genome comparison program Murasaki (Popendorf et al. 2007). Anchors are short, well-conserved subsequences between each contig and the reference genome.

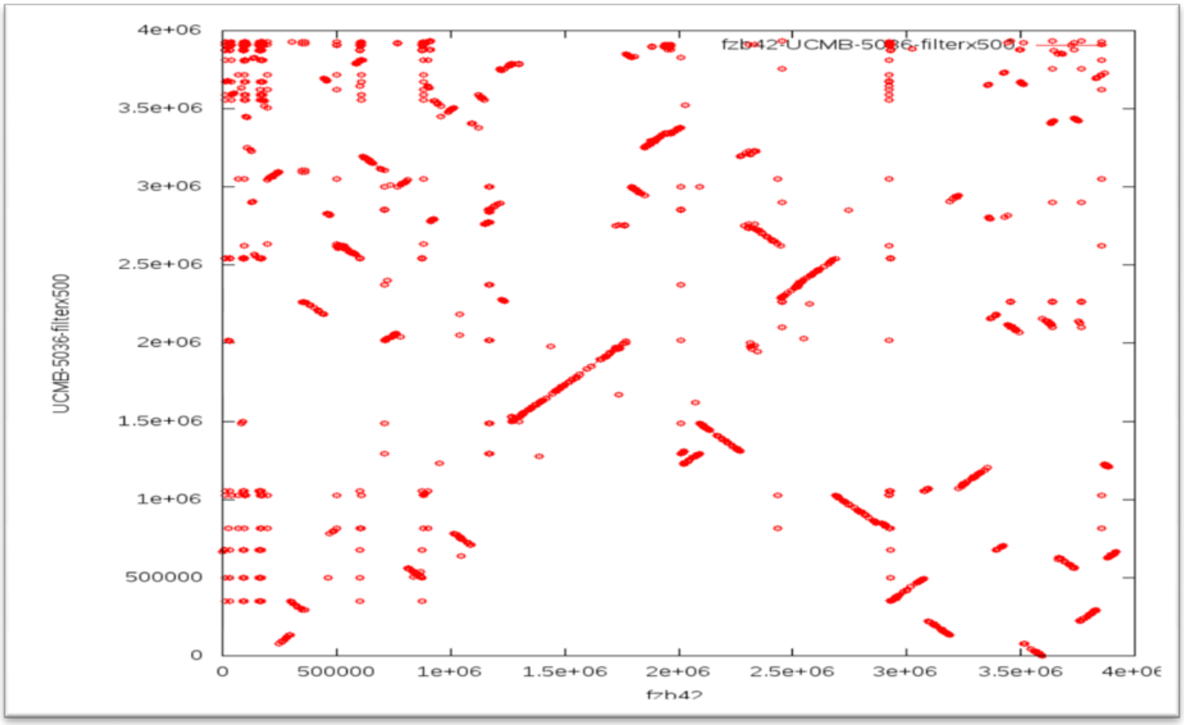


Figure VI. Shows Plotted anchors against reference genome. Total 107 contigs with length greater than 0.5 kbp were used to calculate anchors against the reference genome, using multi genome comparison program *Murasaki*.

4.3 Genome annotation:

The draft genome was input to the *BASys* ('Bacterial Annotation System') web based annotation system freely available at the site (http://wishart.biology.ualberta.ca/basys). *BASys* predicted 4209 genes. Statistics regarding the annotated results are summarized in table IX.

Table IX: Shows the annotated results produced by BASys annotation web server for the draft genome of UCMB-5036

Length:	3918701
Topology:	Circular
Gram Strain	Positive
Number of Genes Identified:	4209
Number of Genes Annotated:	4208

The circular graphical map of the whole genome produced by *BASys* annotation server , shows all the coding sequences within the genome of UCMB-5036 (Fig VII).

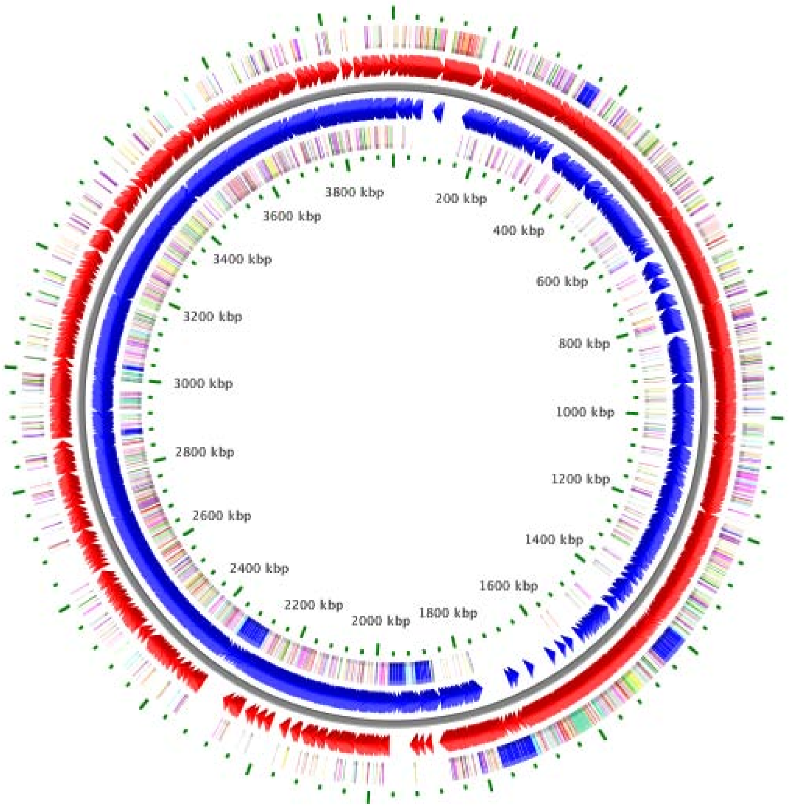


Figure VII. Show map view of the whole genome of UCMB-5036, shows genes encoding proteins and functional RNAs in the genome. Filled circular red and blue lines represent forward and reverse strands, respectively. Whereas, others demonstrating COG functional categories in the genome.

Protein locations and functions annotated for the draft genome of UCMB-5036 by BASys annotation web server are represented in figures III and IV respectively.

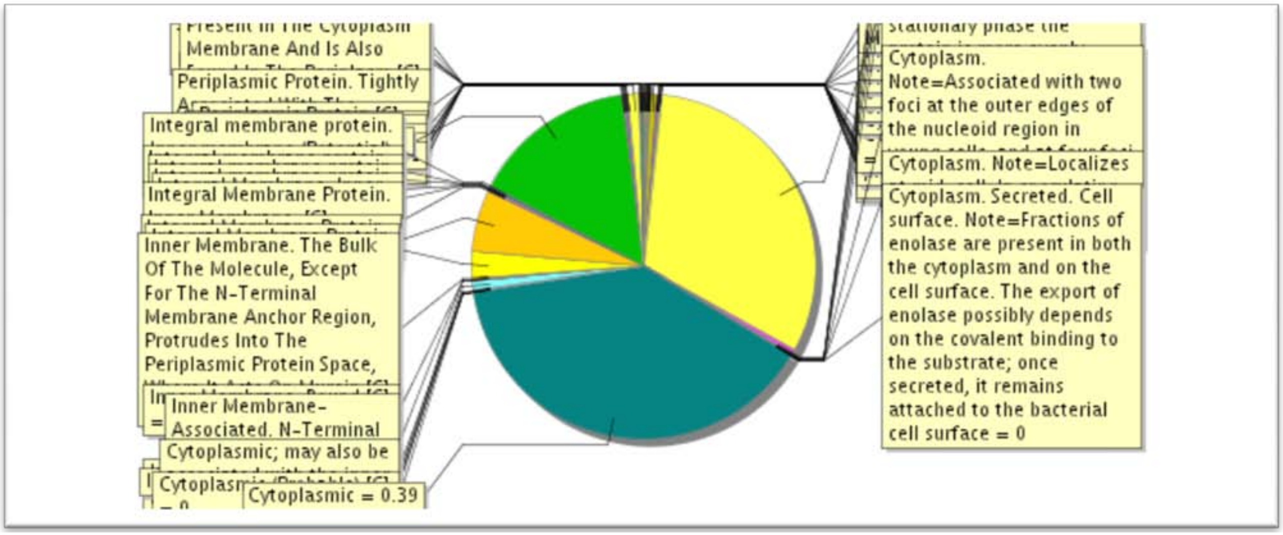


Figure VIII: Shows the locations of proteins identified in the draft genome of *UCMB-5036*.

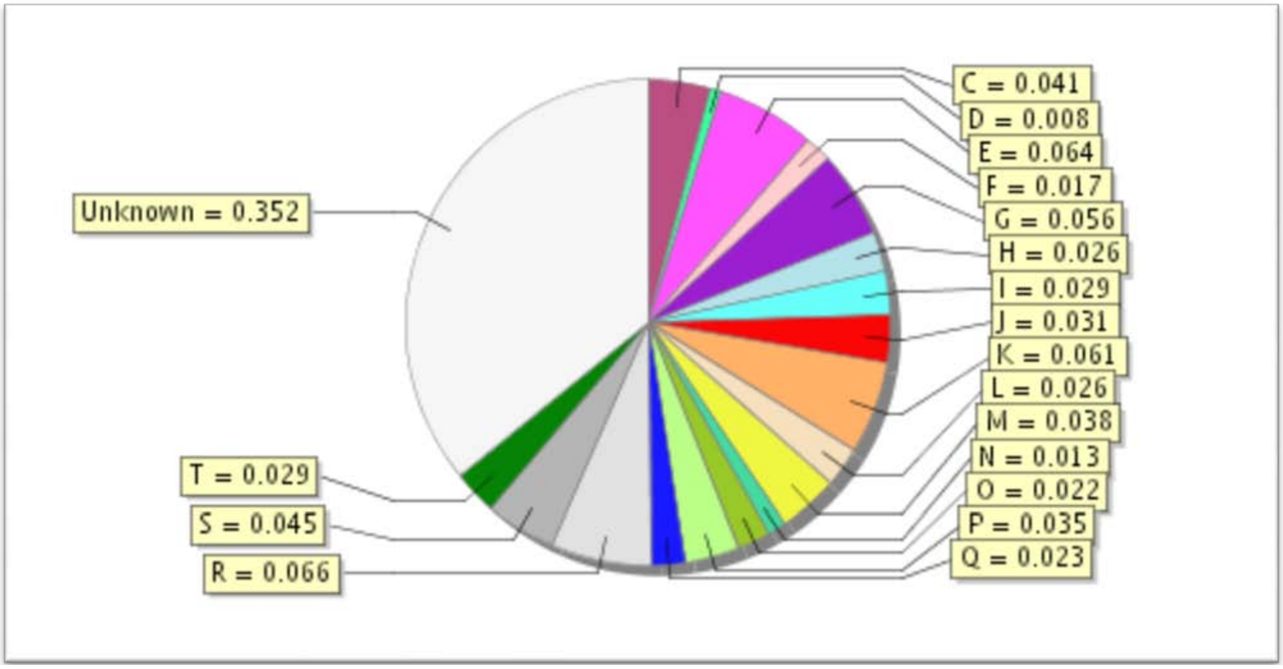


Figure IX: Show the functions of identified proteins in the draft genome of *UCMB-5036*. Letters refer to COG functional categories. C - Energy production and conversion; D - Cell division and chromosome partitioning; E - Amino acid transport and metabolism; F - Nucleotide transport and metabolism; G - Carbohydrate transport and metabolism; H - Coenzyme metabolism; I - Lipid metabolism; J - Translation, ribosomal structure and biogenesis; K - Transcription; L - DNA replication, recombination and repair; M - Cell envelope biogenesis, outer membrane; O - Posttranslational modification, protein turnover, chaperones; P - Inorganic ion transport and metabolism; R - General function prediction only; S - COG of unknown function.

4.4 *De novo* Assembly of Unaligned Reads:

All unaligned reads were extracted from read data file as outlined in material and methods, and served as input into the MIRA3 assembler for *de novo* assembly. Totally 26426 contigs were produced having different sizes and 104 contigs were filtered out with length greater than 0.5 kbp by using MIRA utility (Convert_Project), that were used for similarity search against *NCBI* genome sequences database through *BLAST*. All statistics for contigs produced from unaligned reads are summarized in table X.

Table X. Summary of contigs produced from unaligned reads by *de novo* assembly.
Reads remained unaligned in mapping assembly were input to *de novo* assembler *MIRA3*. From total of 26426 contigs 104 were filtered out with length greater than 0.5 kbp.

Strain	No. of total reads	Unaligned reads (%)	Total No. of Contigs	No. of contigs with length greater than 0.5 kbp	N50 contig size (bp)
UCMB-5036	65,44,152	9.34	26426	104	4079

4.5 Multiple sequence alignment for selected Genes:

Based on phylogenetic relationship, the *Bacillus clausii*, *B. cereus*, *B. halodurans B. subtilis 168* and *B. amyloliquefaciens FZB42* were chosen with *B. amyloliquefaciens UCMB-5036* for multiple sequence alignment of two selected gene sequences (gyrA, cheA). The gyrA gene encodes the *DNA* gyrase subnit A and the cheA gene encodes the two-component sensor histidine kinase CheA, which is crucial for regulating bacterial chemotaxis. Both genes have been shown to be effective for resolving closely related taxa of the B. subtilis group (Kunst et al., 1995).

Multiple sequence alignment was performed with translated sequences of these genes by using ClustalW2 program (Larkin et al., 2007). Multiple alignment of protein sequences is important, because it provides an opportunity to identify conserved sequence regions. Phylograms were constructed from the same program (Fig X, A, B) which showed that *B. amyloliquefaciens UCMB-5036* is more closely related to *B. amyloliquefaciens FZB42.*

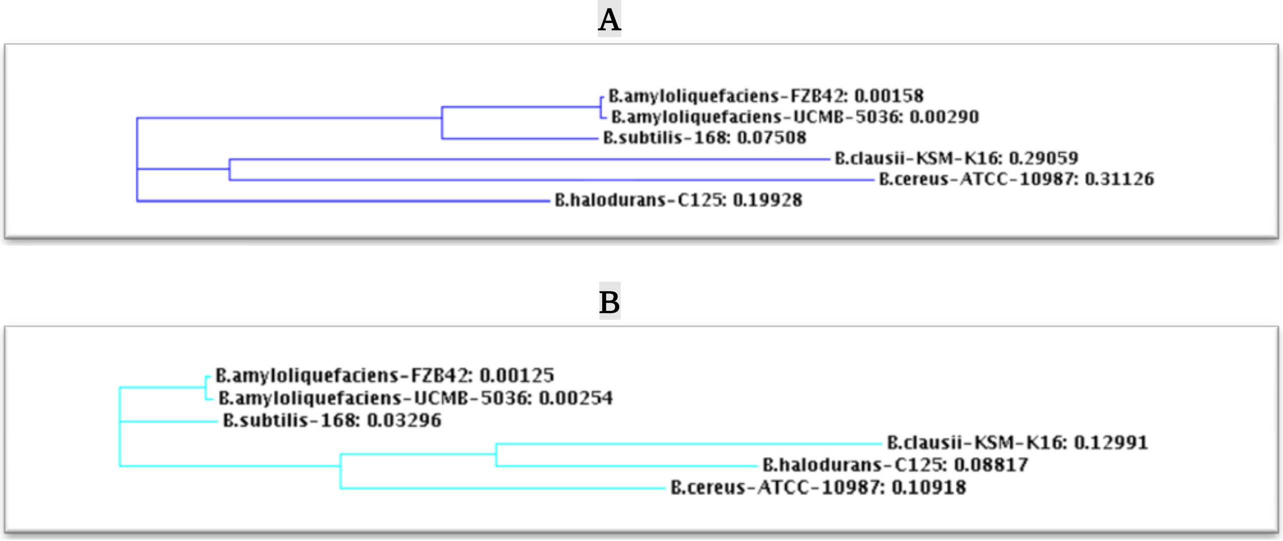
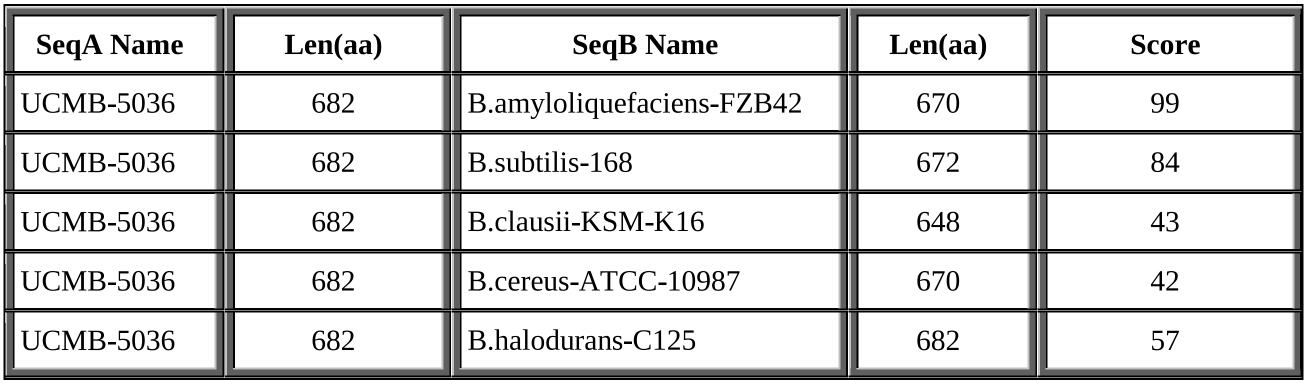

A
B.amyloliquefaciens-FZB42: 0.00158
B.amyloliquefaciens-UCMB-5036: 0.00290
B.subtilis-168: 0.07508
B.clausii-KSM-K16: 0.29059
B.cereus-ATCC-10987: 0.31126
B.halodurans-C125: 0.19928
B
B.amyloliquefaciens-FZB42: 0.00125
B.amyloliquefaciens-UCMB-5036: 0.00254
B.subtilis-168: 0.03296
B.clausii-KSM-K16: 0.12991
B.halodurans-C125: 0.08817
B.cereus-ATCC-10987: 0.10918

Figure X: (A) Phylogram of gene cheA and (B) Phylogram of gene gyrA shows that *UCMB-5036* is closely related to FZB42.

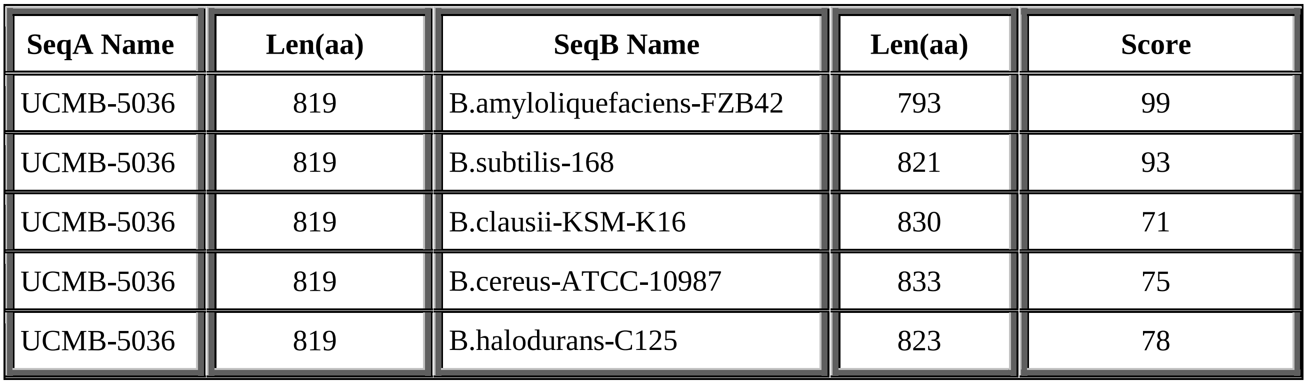
Multiple alignment scores are summarized in table (Xia , XIb) presenting % identity of *amyloliquefaciens UCMB-5036* with other *Bacillus* strains for two selected genes.

Table XIa



SeqA Name	Len(aa)	SeqB Name	Len(aa)	Score
UCMB-5036	682	B.amyloliquefaciens-FZB42	670	99
UCMB-5036	682	B.subtilis-168	672	84
UCMB-5036	682	B.clausii-KSM-K16	648	43
UCMB-5036	682	B.cereus-ATCC-10987	670	42
UCMB-5036	682	B.halodurans-C125	682	57

Table XIb



SeqA Name	Len(aa)	SeqB Name	Len(aa)	Score
UCMB-5036	819	B.amyloliquefaciens-FZB42	793	99
UCMB-5036	819	B.subtilis-168	821	93
UCMB-5036	819	B.clausii-KSM-K16	830	71
UCMB-5036	819	B.cereus-ATCC-10987	833	75
UCMB-5036	819	B.halodurans-C125	823	78

Table XIa for gene cheA and table XIb for gene gyrA show % identity of B. amyloliquefaciens *UCMB-5036* with other Bacillus strains.

4.6 Genes involved in Plant Bacterium Association:

Our reference genome *B. amyloliquefaciens FZB42*, which we used as backbone for mapping assembly, is a plant root-colonizing naturally occurring isolate, distinguished from the model organism *B. sutilis 168* by its abilities to stimulate plant growth and suppress plant pathogens (Idriss *et al.*, 2002). In a report of the *B. amyloliquefaciens FZB42* genome, genes that may contribute to its plant-associated lifestyle were highlighted (Chen et al., 2007). In the analysis of the draft genome of the *UCMB-5036,* we found that all those genes that may contribute to plant-association also seemed to be present in *B. amyloliquefaciens FZB42* genome study. Some of these identified genes are summarized in table (XII).

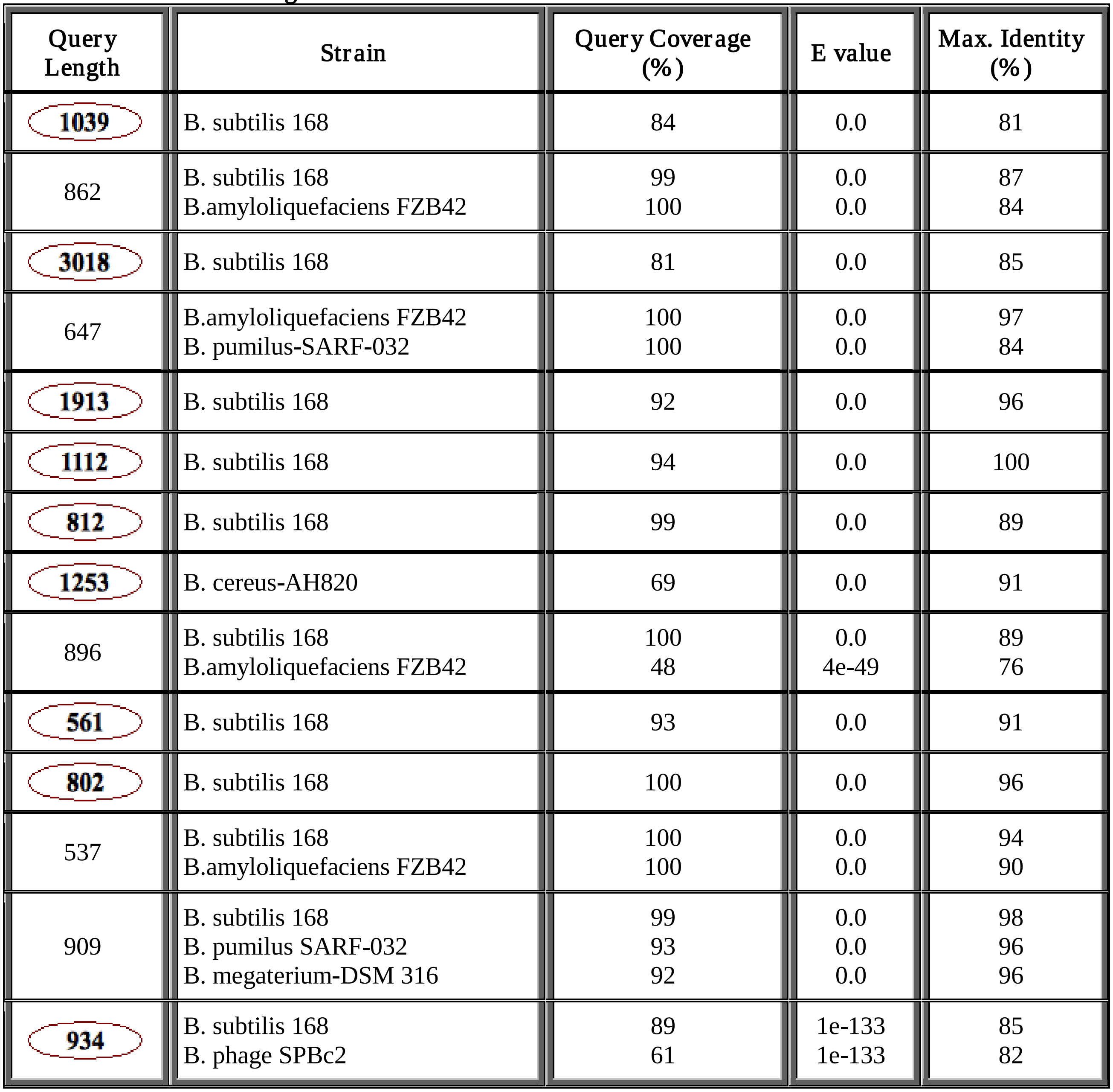
Table XII. Genes probably involved in plant bacterium interactions.

Root colonization, swarming motility and biofilm formation:		
Gene	**Function**	**Identity with FZB42 (%)**
efp	Involved in peptide bond synthesis; alters the affinity of the ribosome for aminoacyl-tRNA	100
spo0A	Involved in initiation of biofilm formation	99.2
abrB	Regulation of stationary/sporulation gene expression; transcriptional control of biofilm formation	98.9
ymcA	Conserved hypothetical protein; Control of community development	100
Bacterial target molecules for general plant immune response:		
flgK	Involved in elicitation of plant basal defense, repression by exudates.	99.6
hag	Involved in elicitation of plant basal defense; Upregulation by root exudates	62.5
tufA	Involved in elicitation of plant basal defense	100
Biofertilization: Mineral availability (Iron, Phosphate)		
yvrC	Putative iron-binding protein.	100
yusV	Putative iron (III) ABC transport ATPase component.	100
phy	3-phytase precursor (Myo-inositol-hexaphosphate 3-phosphohydrolase)	99.2
yclQ	Putative ferrichrome ABC transporter (periplasmic binding protein)	99.4

4.7 Similarity Search of contigs from unaligned reads:

In the process of sequence analysis, to explore and find out maximum information about the mystery of sequence, the first step is always to carry out a primary database search (*BLAST*) in order to compare a novel sequence with those contained in nucleotide and protein databases by aligning the novel sequence with previously characterized genes. 104 contigs of length greater than 0.5 kbp were generated from *de novo* assembly of 9.34% unaligned reads, which we used for similarity search through *BLAST* against *NCBI* genome database (http://blast.ncbi.nlm.nih.gov/Blast.cgi). 50% of total showed no similarity, 19% of total showed low similarity and 31% of total showed high similarity with different Bacillus strains. Similarity results are summarized in table XIII, encircled results are contigs, which showed high similarity with different Bacillus strains other then reference sequence FZB42.

Table XIII show some high similarities *BLAST* results with other Bacillus strains.



Query Length	Strain	Query Coverage (%)	E value	Max. Identity (%)
1039	B. subtilis 168	84	0.0	81
862	B. subtilis 168 B.amyloliquefaciens FZB42	99 100	0.0 0.0	87 84
3018	B. subtilis 168	81	0.0	85
647	B.amyloliquefaciens FZB42 B. pumilus-SARF-032	100 100	0.0 0.0	97 84
1913	B. subtilis 168	92	0.0	96
1112	B. subtilis 168	94	0.0	100
812	B. subtilis 168	99	0.0	89
1253	B. cereus-AH820	69	0.0	91
896	B. subtilis 168 B.amyloliquefaciens FZB42	100 48	0.0 4e-49	89 76
561	B. subtilis 168	93	0.0	91
802	B. subtilis 168	100	0.0	96
537	B. subtilis 168 B.amyloliquefaciens FZB42	100 100	0.0 0.0	94 90
909	B. subtilis 168 B. pumilus SARF-032 B. megaterium-DSM 316	99 93 92	0.0 0.0 0.0	98 96 96
934	B. subtilis 168 B. phage SPBc2	89 61	1e-133 1e-133	85 82

5 Discussion:

In this work we present an almost finished whole genome assembly of the gram-positive bacterium *UCMB-5036*, which is closely related to plant root-colonizing *Bacillus amyloliquefaciens FZB42 that* has ability to enhance plant growth and to suppress plant pathogenic organisms in the soil. From the draft genome annotation *UCMB-5036* showed great resemblance to *FZB42*. By the comparison of both strains, we found that almost all the genes responsible for plant root colonization and growth promotion in *FZB42* are also present in *UCMB-5036*, but further analysis is required to determine the level of conservation of these genes with *FZB42*.

The strain specific sequences cannot be produced through mapping assembly because it only align reads against reference sequence which is useful for getting the similar sequence of strains with polymorphism. Due to this reason we also performed *de novo* assembly for unmapped reads against reference sequence, which produced 104 contigs of length greater than 0.5 kbp and performed Local alignment search against *NCBI* genome database to compare our novel contigs with previously characterized genes. As a result, 50% of totally 104 contigs showed no similarity at all, which can indicate the presence of some novel genes in *UCMB-5036* that were not previously characterized in the genomes database, 19% of total contigs showed low similarity and the remaining 31% showed high similarity with different Bacillus strains, that indicate the presence of some coding sequences in *UCMB-5036,* which are not present in *FZB42*. Further analysis will be performing for confirmation and validation.

From the *BASys* annotation system 4209 genes were identified from the draft genome. We used *RNAmmer* web portal (Lagesen et al. 2007) for the prediction of rRNA genes, which identified 29 rRNA genes in the *UCMB-5036* genome.

The *BASys* annotation system has a limitation that it can not be configured locally with customized settings according to user specific requirements and we also observed that it uses some older versions of software's and databases (nucleotide and proteins) for annotation.

In parallel with mapping assembly we also performed *de novo* assembly for all paired-end reads and calculate anchors from 104 contigs with length greater than 0.5 kbp against reference genome, that will be used for sorting and aligning the contigs along reference sequence to fill the gaps in the consensus sequence of *UCMB-5036* and remaining gaps will be filled by doing PCR, because the combination of *de novo* assembly and reference guided assembly is more successful and validated methodology to produce the bacterial genome assembly, as used to assemble the genome of *B. subtilis natto* (Yukari et al. 2010).

Genome finishing is also a very important step for assembling the genome with no/minimum errors, there are different finishing tools, provide genome-editing functionality in manual and auto mode. *MIRA3* assembler also marks considerable regions, which should be checked by using these finishing tools like gap4. For our project we fail to perform finishing for two reasons first is due to the time constraint and second is the unavailability of a 64-bit version of the caf2gap tool, which was required for *CAF* to *GAP* format conversion as a requirement of gap4 finishing tool. But this will be achieved in the future.

Genome completion of *UCMB-5036* will provide a great gain to microbiology research. It will be very helpful to understand the symbiotic relationship between rhizobacteria and plants, and will support identification of genes responsible for plant growth promotion and control of pathogens e.g. by production of antibiotics.

Future Work:

Complete genome sequence will be used to understand factors and mechanisms that support the symbiotic interaction between beneficial Bacillus amyloliquefaciens strains and plants in the rhizosphere. Of special interest is to identify genes responsible for developing the intricate relationship with the plants with respect to colonization and improved plant vigor through systemic effects like priming of *ISR* in plants. The work will mainly be based on in silico analysis (bioinformatics) but complemented with wet laboratory work when necessary to support the findings.

6 Conclusion:

In this study we have optimized parameters of the *MIRA3* assembler by mapping the *UCMB-5036* genome against the reference sequence to minimize errors. After annotation we found that many genes that participate in plant-bacterium interactions exits in both *UCMB-5036* and the reference genome *FZB42* and also were highly similar. Through multiple sequence alignment of two genes (gyrA, cheA), which previously have been shown to be effective for resolving closely related taxa of the B. subtilis group, we found that the *UCMB-5036* strain is more closely related to *B. amyloliquefaciens* strain *FZB42* as compared to other Bacillus strains. By doing *de novo* assembly of unmapped reads we found some contigs which may represent novel genes present in *UCMB-5036* but absent in the reference sequence (*FZB42*) and some other contigs (encircled in *BLAST* similarity results table XII) can contain genes present in *UCMB-5036* and other Bacillus strains but absent in *FZB42*. However further analysis required to validate this interpretation.

This work will now continue with the identification of key genes involved in the symbiotic interaction between beneficial *Bacillus amyloliquefaciens* strains and plants in the rhizosphere and posterior wet-lab work to corroborate our findings.

7 Acknowledgements:

First of all I would like to say thanks to Higher Education Commission (HEC) of Pakistan for providing this great opportunity of higher education in new emerging field of bioinformatics by awarding scholarship, and UPPMAX Uppsala University for providing computational resources in this project. There are also a number of people who made all this work possible by supporting me academically, technically and financially. So I would like to express my sincere gratitude to the following people: Erik Bongcam-Rudloff, Johan Meijer, Hans Henrik Fuxelius and Oscar Eriksson. At the end I would also like to pay special regards to my parents and beloved wife who all the time pray for me and morally support me during my whole stay in Sweden.

References:

Adesemoye, A.O. et al. (2008) Enhanced plant nutrient use efficiency with PGPR and AMF in an integrated nutrient management system. Can. J. Microbiol. 54, 876–886.

Bais HP, Fall R and Vivanco JM (2004) Biocontrol of Bacillus subtilis against infection of Arabidopsis roots by Pseudomonas syringae is facilitated by biofilm formation and surfactin production. Plant Physiol. 134, 307-319.

Bao,H. et al. (2009) MapView: visualization of short reads alignment on a desktop computer, Bioinformatics, 25, 1554-1555.

Benhamou N, Kloepper JW, Quadt-Hallman A and Tuzun S (1996) Induction of defense-related ultrastructural modifications in pea root tissues inoculated with endophytic bacteria. Plant Physiol. 112, 919-929.

Benhamou, N., Kloepper, J. W., Quadt-Hallman, A., and Tuzun, S. 1996. Induction of defense-related ultrastructural modifications in pea root tissues inoculated with endophytic bacteria. Plant Physiol. 112:919-929.

Bloemberg GV and Lugtenberg BJJ (2001) Molecular basis of plant growth promotion and biocontrol by rhizobacteria. Curr. Opin. Plant Biol. 4, 343- 350.

Bostock RM (2005) Signal crosstalk and induced resistance: straddling the line between cost and benefit. Annu. Rev. Phytopathol. 43, 545-580.

Chen et al., 2007 X.H. Chen, A. Koumoutsi, R. Scholz, A. Eisenreich, K. Schneider, I. Heinemeyer, B. Morgenstern, B. Voss, W.R. Hess, O. Reva, H. Junge, B. Voigt, P.R. Jungblut, J. Vater, R. Süssmuth, H. Liesegang, A. Strittmatter, G. Gottschalk and R. Borriss, Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42, *Nat. Biotechnol.* **25** (2007), pp. 1007–1014.

Chevreux, B., Wetter, T. and Suhai, S. (1999): *Genome Sequence Assembly Using Trace Signals and Additional Sequence Information*. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, pp. 45-56.

ClustalW and ClustalX version 2. Bioinformatics 2007 23(21): 2947-2948.

Compant, S. et al. Use of plant-growth promoting bacteria for biocontrol of plant diseases: principles, mechanisms of action, and future prospects. Appl. Environ. Microbiol. 71, 4951–4959 (2005).

Conrath U, Beckers GJM, Flors V, García-Agustín P, Jakab G, Mauch F, Newman M-A, Pieterse CMJ, Poinssot B, Pozo MJ, Pugin A, Schaffrath U, Ton J, Wendehenne D, Zimmerli L and Mauch-Mani B (2006) Priming: getting ready for battle. Mol. Plant-Microbe Interact. 19, 1062-1071.

Costacurta A, Vanderleyden J (1995) Synthesis of phytohor- mones by plant associated bacteria. Crit Rev Microbiol 21:1– 18.

Denison RF and Kiers Et (2004) Lifestyle alternatives for rhizobia: mutualism, parasitism, and forgoing symbiosis. FEMS Microbiol. Lett. 237, 187-193.

Dunn AK, Klimowicz AK and Handelsman J (2003) Use of a promoter trap to identify Bacillus cereus genes regulated by tomato seed exudate and a Rhizosphere resident, Pseudomonas aureofaciens. Appl. Environ. Microbiol. 69, 1197-1205.

Durrant, W. E., and Dong, X. 2004. Systemic acquired resistance. Annu Rev Phytopathol. 42:185-209.

Figueiredo, V.B. et al. (2008) Alleviation of drought stress in the common bean (Phaseolus vulgaris L.) by co-inoculation with Paenibacillus polymyxa and Rhizobium tropici. Appl. Soil Ecol 40, 182–188.

Garbeva P, van Veen JA and van Elsas JD (2004) Microbial diversity in soil: selection microbial populations by plant and soil type and implications for disease suppressiveness. Annu. Rev. Phytopathol. 42, 243-270.

Glick, B.R. et al. (2007) Promotion of plant growth by bacterial ACC deaminase. Crit. Rev. Plant Sci. 26, 227–242.

Gordon,D. *et al*. (1998) Consed: a graphical tool for sequence finishing. *Genome Res*., **8**, 195–202.

Gyaneshwar, P. et al. (2002) Role of soil microorganisms in improving P nutrition of plants. Plant Soil 245, 83–93.

Handelsman J and Stabb EV (1996) Biocontrol of soilborne plant pathogens. Plant Cell 8, 1855-1869.

Huang, W. and Marth, G. (2008) EagleView: a genome assembly viewer for next-generation sequencing technologies. Genome Res., 18, 1538-1543.

Iavicoli A, Boutet E, Buchala A and Metraux J-P (2003) Induced systemic resistance in Arabidopsis thaliana in response to root inoculation with Pseudomonas fluorescens CHA0. Mol. Plant-Microbe Interact. 16, 851-858.

Idriss, E.E.S. *et al*. Extracellular phytase activity of *Bacillus amyloliquefaciens* FZB45 contributes to its plant-growth-promoting effect. Microbiology 148, 2097–2109 (2002). | PubMed | ISI | ChemPort |

Jacobsen B, Zidack N and Larson B (2004) The role of Bacillus based biological control agents in integrated pest management systems: Plant diseases. Phytopathology 94, 1272-1275.

Kloepper JW, Ryu C-M and Zhang S (2004) Induced systemic resistance and promotion of plant growth by Bacillus spp. Phytopathology 94, 1259- 1266.

Kloepper JW, Zablotowick RM, Tipping EM, Lifshitz R (1991) Plant growth promotion mediated by bacterial rhizosphere colonizers. In: Kliester DL, Cregan PG (eds) The rhizo- sphere and plant growth.

Kloepper, J.W. et al. (2007) Photoperiod regulates elicitation of growth promotion but not induced resistance by plant growth-promoting rhizobacteria. Can. J. Microbiol. 53, 159–167.

Kloepper, J.W., Ryu, C.-M. & Zhang, S. Induced systemic resistance and promotion of plant growth by Bacillus spp. Phytopathology 94, 1259–1266 (2004).Kluwer Academic Press, Dordr- echt, The Netherlands, pp 315–326.

Kohler, J. et al. (2008) Plant-growth-promoting rhizobacteria and arbuscular mycorrhizal fungi modify alleviation biochemical mechanisms in water-stressed plants. Funct. Plant Biol. 35, 141–151.

Kunst,F.&Rapoport,G.Salt stress is an environmental signal affecting degradative enzyme synthesis in Bacillus subtilis. J. Bacteriol. 177, 2403–2407 (1995).

Lagesen K, Hallin PF, R�dland E, St�rfeldt HH, Rognes T Ussery DW Rnammer : consistent annotation of rRNA genes in genomic sequences
Nucleic Acids Res. 2007 Apr 22.

Leifert C, Li H, Chidburee S, Hampson S, Workman S, Sigee D, Epton HAS and Harbour A (1995) Antibiotic production and biocontrol activity by Bacillus subtilis CL27 and Bacillus pumilus CL45. J. Appl. Bacteriol. 78, 97-108.

Li,H. et al. (2009) The Sequence Alignment/Map format and SAMTools. Bioinformatics, 25, 2078–2079.
Lugtenberg B, Dekkers L and Bloemberg G (2001) Molecular determinants of Rhizosphere colonization by Pseudomonas. Annu. Rev. Phytopathol. 39, 461-490.

Lutenberg BJJ, De Weger LA, Bennett JW (1991) Microbial stimulation of plant growth and protection from disease. Curr Opin Microbiol 2:457–464.

Mantelin, S. and Touraine, B. (2004) Plant growth-promoting bacteria and nitrate availability impacts on root development and nitrate uptake. J. Exp. Bot. 55, 27–34.

Martens DA and Frankenberger WT (1993) Metabolism of trypthophan in soil. Soil Biol. Biochem. 25, 1679-1686.

Mayak, S. et al. (2004) Plant growth-promoting bacteria confer resistance in tomato plants to salt stress. Plant Physiol. Biochem. 42, 565–572.

Mayak, S. et al. (2004) Plant growth-promoting bacteria that confer resistance to water stress in tomatoes and peppers. Plant Sci. 166, 525– 530.

Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D.
Nelson E (2004) Microbial dynamics and interactions in the spermosphere. Annu. Rev. Phytopathol. 42, 271-309.

Pop M, Salzberg S. Bioinformatics challenges of new sequencing technology. Trends Genet. 2008;24:142–149. [PMC free article] [PubMed]

Popendorf K, Osana Y, Hachiya T, Sakakibara Y: Murasaki: homology detection across multiple large-scale genomes. *Fifth Annual RECOMB Satellite Workshop on Comparative Genomics. San Diego* 2007.

Priest F, Goodfellow M, Shute L and Berkeley R (1987) Bacillus amyloliquefaciens sp. nov., nom. rev. Int. J. Syst. Bacteriol. 37, 69-71.

Raaijmakers J, Vlami M and de Souza J (2002) Antibiotic production by bacterial biocontrol agents. Antonie van Leeuwenhoek 81, 537-547.

Raaijmakers J, Vlami M and de Souza J (2002) Antibiotic production by bacterial biocontrol agents. Antonie van Leeuwenhoek 81, 537-547.

Reva ON, Dixelius C, Meijer J and Priest FG (2004) Taxonomic characterization and plant colonizing abilities of some bacteria related to Bacillus amyloliquefaciens and Bacillus subtilis. FEMS Microbiol. Ecol. 48, 249- 259.

Roh JY, Choi JY, Li MS and BR, Je J (2007) Bacillus thuringiensis as a specific, safe, and effective tool for insect pest control. Microbiol Biotechnol. 17, 547-559.

Romero D, de Vicente A, Rakotoaly RH, Dufour SE, Veening JW, Arrebola E, Cazorla FM, Kuipers OP, Paquot M and Pérez-García A (2007) The Iturin and Fengycin families of lipopeptides are key factors in antagonism of Bacillus subtilis toward Podoshaera fusca. Mol. Plant-Microbe Interact. 20, 430-440.

Rudrappa T, Biedrzycki M, Bais HP (2008) Causes and consequences of plant-associated biofilms. FEMS Microbiol. Ecol. 64, 153-166.

Ryals, J. A., Neuenschwander, U. H., Willits, M. G., Molina, A., Steiner, H.- Y., and Hunt, M. 1996. Systemic acquired resistance. Plant Cell 8:1809- 1819.

Ryu C-M, Murphy JF, Mysore KS, and Kloepper JW (2004) Plant growth- promoting rhizobacteria systemically protect Arabidopsis thaliana against Cucumber mosaic virus by a salicylic acid and NPR1-independent and jasmonic acid-dependent signaling pathway. Plant J. 39, 381-392.

Schatz,M.C et al. (2007) Hawkeye: an interactive visual analytics tool for genome assemblies. Genome BIO., 8, R34.

Schisler D, Slininger P, Behle R and Jackson M (2004) Formulation of Bacillus spp. for biological control of plant diseases. Phytopathology 94, 1267-1271.

Smith KP, Handelsman J and Goodman RM (1999) Genetic basis in plants for interactions with disease-suppressive bacteria. Proc. Natl. Acad. Sci. USA 96, 4786-4790.

Strange RN and Scott PR (2005) Plant disease: a threat to global food security. Annu. Rev. Phytopathol. 43, 83-116.

Tablet--next generation sequence assembly visualization. Bioinformatics. 2010 Feb 1;26(3):401-2. Epub 2009 Dec 4. PubMed PMID: 19965881; PubMed Central PMCID:PMC2815658.

Timmusk S, Nicander B, U. Granhall B and Tillberg E (1999) Cytokinin production by Paenibacillus polymyxa. Soil Biol. Biochem. 31, 1847-1852.

Timmusk, S. and Wagner, G.H. (1999) The plant-growth-promoting rhizobacterium Paenibacillus polymyxa induces changes in Arabidopsis thaliana gene expression: a possible connection between biotic and abiotic stress responses. Mol. Plant Microbe Interact. 12, 951–959.

van Loon LC, Bakker PAHM, Pieterse CMJ (1998) Systemic resistance induced by rhizosphere bacteria. Annu. Rev. Phytopathol. 36, 453-483.

Varma A, Abbot L, Werner D and Hampp R (eds) (2004) Plant Surface Microbiology, Springer, Berlin.

Wang B, Ferro D and Hosmer D (1999) Effectiveness of Trichogramma ostriniae and T.nubilale for controlling the european corn borer Ostrinia nubilalis in sweet corn. Entomol. Exp. Appl. 91, 297-303.

Whipps J (2001) Microbial interactions and growth in the rhizosphere. J. Exp. Bot. 52, 487-511.

Yang J, Kloepper JW, Ryu CM. Rhizosphere bacteria help plants tolerate abiotic stress. Trends Plant Sci. 2009 Jan;14(1):1-4. Epub 2008 Dec 4. PubMed PMID:19056309.

Yukari Nishito, Yasunori Osana, Tsuyoshi Hachiya, Kris Popendorf, Atsushi Toyoda, Asao Fujiyama, Mitsuhiro Itaya, and Yasubumi Sakakibara (2010):Whole genome assembly of a natto production strain

Bacillus subtilis natto from very short read data.Published online 2010 April 16. doi: 10.1186/1471-2164-11-243.PMCID: PMC2867830.

Zhang, H. et al. (2008) Soil bacteria confer plant salt tolerance by tissue-specific regulation of the sodium transporter HKT1. Mol. Plant Microbe Interact. 21, 737–744.