Summary

The concept of metagenomics is based on the knowledge that 99% of organisms in nature are recalcitrant to culturablity. Hence, there is a need for a culture-independent approach to directly "capture" wealth from the unexplored microbes present in nature. This can be done by (a) library dependent approach by cloning, and making a library of DNA inserts in plasmids etc. for important biocatalysts or/and (b) a library-independent, approach by doing shotgun sequencing to determine the microbial, functional and the metabolic profiles.

In the present study, the choice of habitat was soil. The soil serves as richest reservoirs of microbial genomic diversity. The soil samples were collected from various habitats *viz.*, IMTECH soil (enriched with saw dust), local milk industry (effluent), Palampur (Jhatingari forest), Siswan Dam, Kaziranga Soil, Mangrove sites in Eastern India Soil (pneumatophore soil and rhizosphere). These sources were selected considering their richness of various microbes which may have been unexplored due to lack of culture conditions in these specialized environmental niches. The goal was to directly clone the collective genomes of all microorganisms present in a habitat at a given time point. For our studies we carried both functional and sequence based metagenomics of soil samples.

The two main factors in soil DNA isolation methodology to consider in soil DNA isolation methodology are the extracting of high purity target DNA and the size of soil DNA fragments. The direct lysis method was adapted involving the soil sample DNA isolated directly using mechanical lysis methods incorporated with chemical approach. Observations suggest that the direct lysis of soil is efficient for greater DNA yield, and also amenable to smaller inserts suitable for plasmid library constructions. Metagenomic plasmid libraries were prepared of average insert size of (2-5kb) using blunt-end cloning in plasmid vectors, to be maintained into suitable *E.coli* strains. These libraries were functionally screened for cellulases. Cellulases are the important enzymes of vast commercial potential in the food, feed and pulp, detergent and most recently in biofuels industry. The functional screening for activity was plate based screening. This is a rapid and selective semi-quantitative method to determine cellulose utilisation by metagenomic libraries made from complex ecosystems like soil. Endoglucanase activities are detected easily by examination of "halos" on solid agar plates using CMC as the substrate, followed by Congo Red dye staining. Putative colonies would show zones of clearance on the Congo Red plates.

Summary

The concept of metagenomics is based on the knowledge that 99% of organisms in nature are recalcitrant to culturablity. Hence, there is a need for a culture-independent approach to directly "capture" wealth from the unexplored microbes present in nature. This can be done by (a) library dependent approach by cloning, and making a library of DNA inserts in plasmids etc. for important biocatalysts or/and (b) a library-independent, approach by doing shotgun sequencing to determine the microbial, functional and the metabolic profiles.

In the present study, the choice of habitat was soil. The soil serves as richest reservoirs of microbial genomic diversity. The soil samples were collected from various habitats *viz.*, IMTECH soil (enriched with saw dust), local milk industry (effluent), Palampur (Jhatingari forest), Siswan Dam, Kaziranga Soil, Mangrove sites in Eastern India Soil (pneumatophore soil and rhizosphere). These sources were selected considering their richness of various microbes which may have been unexplored due to lack of culture conditions in these specialized environmental niches. The goal was to directly clone the collective genomes of all microorganisms present in a habitat at a given time point. For our studies we carried both functional and sequence based metagenomics of soil samples.

The two main factors in soil DNA isolation methodology to consider in soil DNA isolation methodology are the extracting of high purity target DNA and the size of soil DNA fragments. The direct lysis method was adapted involving the soil sample DNA isolated directly using mechanical lysis methods incorporated with chemical approach. Observations suggest that the direct lysis of soil is efficient for greater DNA yield, and also amenable to smaller inserts suitable for plasmid library constructions. Metagenomic plasmid libraries were prepared of average insert size of (2-5kb) using blunt-end cloning in plasmid vectors, to be maintained into suitable *E.coli* strains. These libraries were functionally screened for cellulases. Cellulases are the important enzymes of vast commercial potential in the food, paper and pulp, detergent and most recently in biofuels industry. The functional screening for activity was plate based screening. This is a rapid and selective semi-quantitative method to determine cellulose utilisation by metagenomic libraries made from complex ecosystems like soils. Endoglucanase activities are detected easily by examination of "halos" on solid agar plates using CMC as the substrate, followed by Congo Red dye staining. Putative colonies would show zones of clearance on the Congo Red plates.

Clones containing the cellulase gene were confirmed by sequencing. A positive cellulase gene obtained from functional screening was extensively studies for its promoters, RBS, structural homologs, disulphide bond patterns, and protein folds and functions. Analysis revealed from the domain and superfamily prediction showed the sequence had a strong match to Cellulase (endoglucanase) of the Glycosyl hydrolase family 5 and belonging to the transglycosidases superfamily (SCOP). The neighbouring genes and the related species homologs were studied. The ORFs with neighbouring genes presented details with respect to the evolutionary significance. A predictive structural model including catalytic residues and conserved signature sequence was proposed. Phylogenetic analysis revealed the endoglucanase gene obtained from functional screening was quite unique. The close relationship were obtained with *Paludibacter* species although the branching in the phylogenetic tree was distinctively separate. Sequence alignments of metagenomic ORF were obtained with anaerobes like *Paludibacter, Prevotella buccae* and *Bacteroides* sps. The anaerobes utilize cellulase in a multienzyme fashion. No *Paludibacter* of functional cellulolytic activity has been reported till date. The metagenome gene fragment of cellulase may have co-speciated in the soils. Thus we report for the first time an endoglucanase gene of the conserved motif belonging to the glycosyl hydrolase family 5 derived from metagenomic library homologous to Cytophaga-Flavobacteria-Bacteroides (CFB) group bacteria particularly *Paludibacter* species.

Evident also from the functional screenings was that the overall rate of discovery of novel cellulase enzymes from plasmid libraries is limited. The reason could be that for a complete heterologous expression of a gene from metagenomic source the desired gene should contain (a) unique and subtle structural feature of gene sequence (b) stability and translational efficiency of mRNA (c) ease of protein folding for efficient activity and presence of signal sequences for targeting the protein (d) non toxic to the heterologous host (e) and a codon usage frequency with the heterologous host. Functional screening is preferred because it does not require any sophisticated apparatus and simply assayed on substrate agar plates. The metagenomic library can be easily screened and the phenotypic activity of the positive clone can be easily identified visually. Thus, soil being rich in innumerable biotechnologically important enzymes/biocatalysts may be missed out in biased targeted screening of single set enzymes. Since, the aim of this study was the screening for potentially useful genes which might have important role in enzymatic processes from the soil metagenomes, the use of techniques for DNA sequencing enabled the sequencing of fairly large number of gene

fragments. The Primer Walking approach was adapted for the primary sequencing of small regions (2-5 kb) of clones with the aim of mining useful genes from the metagenomic library clones. Based on the ORF information, many important soil metagenome derived chaperones, kinases, hydrolases, transcription regulators, replication and repair proteins, oxidases, nucleases, polymerases, oxidoreductases, ABC transporters, carrier proteins, hydratases, phage proteins, transferases, reductases, dehydrogenases, deaminases, synthases, methylases, helicases, isomerases, lyases and ligases were predicted. Eighty two sequences were studied in detail at the level of open reading frame information, biotechnological applications, remote homology search and fold and functional assignments. The discovery of (Domain of unknown function) DUFs emphasizes the importance of metagenomic screenings. The metagenomic data cross validated with the Pfam, PDB and SCOP helped in identifying the description of the DUFs.

We also carried out shotgun sequencing of Kaziranga metagenome to carry out *In silico* based screening of query glycosyl hydrolase genes from 60 Mb sequencing data. Out of 60,000,000 bases of the soil metagenome the total numbers of hits obtained for GH family were 139. This means that the nature has evolved diverse cellulose hydrolysis enzymes for the most available biopolymer in earth. And therefore this diversity of GH families resulted in limited number of hits. It is pertinent to note that functional screening of small insert libraries resulted only in one celluase, however just 60 Mb of metagenome sequence yielded as many as 139 hits of cellulases to diverse families. Hence it is important to include metagenome sequencing approach along with functional screening of cloned metagenomic DNA.

Sequencing of whole communities from environmental soil DNA using next-generation sequencing method was also performed. For the first round, 60Mb sequence information using pyrosequencing for Kaziranga soil DNA was compared with known metagenomes Luquillo experimental forest soil, Waseca Farm, Acid Mine Drainage and the Sargasso Sea. The study focuses on organismal and functional profile of Kaziranga metagenome by whole genome sequencing. The aim was to obtain comprehensive view of the metagenomic/gene content in the soil metagenome. Sequence analysis further revealed in this soil environment, *Proteobacteria, Actinobacteria, Acidobacteria, Bacteroidetes, Firmicutes* and *Cyanobacteria* form dominant phylas. More than 60 genera were found to be pre-dominant in the Kaziranga soil metagenomes, in comparison with other metagenomes and remarkably, 72 genera were found

exclusively in the Kaziranga soil. It was further analyzed for functional profiling of the data for determining the majority of metabolic activity responsible for specialized physiology and adaptation of the microbes in this ecosystem. Kaziranga metagenome was comparable to the Luquillo forest soil, as both being from tropical rainforests of the in terrestrial ecozones. Overall, Kaziranga soil, Luquillo experimental forest soil, Waseca Farm soil were highly complex metagenomes. This is unlike the sea and acid mine drainage system, which have specialized metagenomes where the microbial demands are rather limited and hence few vital but specialized pathways suffice for the microbial survival. The study thus highlighted the microbial complexity and nutritional demands of different geographically variable soil metagenomes. This study provides for the first time the microbial and functional insight for the sub-Himalayan biodiversity hotspot viz. Kaziranga from a metagenomics perspective.

We also carried out comparative metagenomics for different soils of India wherein more than 100Mb sequences were obtained. Metagenomic profiling of different sites based on the phylogenetic distribution and metabolic distribution and statistical significance of Kaziranga, Pneumatophore root soil, Common rhizosphere soil, Sawdust enriched soil was performed. The approach of sequencing soil metagenome can be considered as a finest endeavour for profiling the taxonomic and functional microbial diversity in some of the selected Indian soils. The information gained by sequencing provided a comprehensive view contributing to the microgeocataloguing. Interestingly sequence hits were obtained with 89 unique genera in Kaziranga soil metagenome, 141 genera were uniquely found in sawdust-enriched soil, 152 unique genera in the pneumatophore soil and 138 unique genera in rhizosphere soil.

Kaziranaga is the soil belonging to north eastern part of India with tropical-subtropical broad leaf biome. The Kaziranga soil is thus a forest soil. The soil is alluvial in nature and the river Brahmaputra contributes to the silt deposits. The functional abundance is reflective of the soil fertility directly and indirectly contributing to/contributed by the microbial activities. The metagenome of regions may be having nutritional surplus because of the organically rich soil conditions. Moreover the Kaziranga soil is heavily flooded in the rainy season, and the water logged conditions contributes to generation of a unique microbial environment. The presence of sequences of unique marine, thermophilic, acid tolerant, methylotropic, nitrifying and xylan utilizing nature of microbes were clearly obtained. Also, the nitrite and nitrate tolerance, metal and acid and utilizing microbial sequences were detected in higher frequency. The soil also has presence of sequences from microbes producing secondary

etabolites, toxins and antibiotic resistance. The soil had sequences of microbes with otential of heavy metal degradation, recalcitrant aromatic compounds utilization like ndane, benzopyrene degradation etc.which has potential for applications in bioremediations. he soil demonstrated functional hits with nitrogen, potassium and phosphorus, which are mportant components in soil fertility and may have application in agriculture.

he Pneumatophore soils from the mangrove forests of littoral region in eastern India are rich n metagenome population of highly anaerobic and halophilic nature. Comparative statistical nalysis showed high species richness and maximum unique genera among the metagenomes nder study. Pneumatophore metagenome had high species count (species richness) but each pecies was unevenly distributed i.e. the number of each species in the environment was ariable. Comparatively unique sequences from the phyla *Proteobacteria* and Cyanobacteria were predominant in Pneumatophore soil metagenomes. A vast majority of sequence hits rom unique genera were of marine/aquatic origin. Being from the mangrove regions and igh salt conditions, the metagenome features demonstrated relatively higher abundance of ene sequences related to the stress response, regulation and cell signalling, motility and hemotaxis, phages, prophages, transposable elements and plasmids. These features echo the tress conditions prevalent and microbial survival in such environmental conditions. In the icrobiota from pneumatophore soil, sequences of microbes of sulphur-utilizing hemolithotrophs, fatty acid oxidizing, acetate and hydrogen utilizing, chlororespiration/ alorespiration and carboxydotrophy related gene were observed. Beta-glucosidase and beta- garase producing, mannoside and chitin degrading bacterial sequence and also nitrogenase- pecific proteolytic activity producing microbial sequence hits were observed in dominance. he sequences in pneumatophore metagenomes gave hits with members of high temperature olerance, high salt tolerance and high concentration of heavy metals. The cytotoxic yanobacteria sequences were observed possibly role in the eutrophic regions with hosphorus limitations. The metagenome profile of pneumatophore has potential applications degrading the oil spills and petrol like aliphatic hydrocarbon. Microbial sequences with oremediation potential in soil and wastewater/ sludge were also obtained.

wdust enriched soil is an example of intervention of micrflora by supplemental sawdust for o years. The sawdust enriched soil contains high carbon content providing a rich source of ergy for the microbial decomposition but low nitrogen content which creates a nutritional mand from the soil. The nutritional balance may thus have been grossly disproportional.

he disproportionality may contribute to the competition among the microbial populations. ompared to other metagenomes, sawdust features had sequences in higher prevalence from etabolism of amino acid and derivatives, defence, diseases and virulence, aromatic ompounds metabolism and nitrogen metabolism. The sawdust enriched microbiota sequence its with metabolism for aromatics compounds may possibly contribute to the microbial daptation to alternate nutritional sources. Aromatic compounds also include recalcitrant ydrocarbons. The unique phyla dominating the metagenome were the Proteobacteria, ctinobacteria and Firmicutes. The enrichment resulted in predominance of sequences of etagenome linked to enzymes like cellulase, amylase, xylanase, mannase and ligninases. equences of microbes producing laccases which have a role in promoting oxidative coupling f lignols for the production of lignins were also observed. The sawdust adsorbs moisture om the atmosphere. The self-aggregation and biofilms like property was observed in certain ique genera sequences of sawdust metagenomes. The sequence hits of microbes emonstrating chemotaxis and magnetotaxis were also reported. The property of echlorination, debromination and dehalogenation were also present in the unique microbiota equence of sawdust enriched soil. Unique sequences of microbial genera producing products pharmacological relevance were also obtained. These include the polyketide synthases sed in chemotherapy, other chemotherapy analogs, anticancer bryostatins, antineoplastics, ntibiotic productions etc. Microbes sequence with bioremediation potential and ineralization of explosives were also obtained. Sequences from pathogenic microorganisms und in the sawdust enriched soil may have application as a bacterial biological control gent to inhibit pathogenic fungi/nematodes in soil.

he rhizosphere soil metagenome was the sample collected from the microbiota attached ound the roots from an uprooted tree. The region has predisposition to natural disasters like gh tides, flood, cyclone, tornado, droughts. The collection site was in the eastern part of dia, where the soil is primarily a complex network of tidal waterways, mudflats and small ets of salt rich wetlands. The species count along with the pneumatophores metagenomes s relatively higher compared to other metagenomes. Exceptionally high *E.coli* sequences dominated in the metagenome. Unique genera sequences containing glycerosphingolipids SL) instead of lipopolysaachrides in the cell envelope were also observed. The pathogens *E.coli* take the advantage of GSL in adhesion to the host for the release of the toxins and ctions. Interestingly most of the bacterial sequences hits were belonging to pathogens. uses were also present in much higher numbers then rest of the metagenomes under study.

t may be mentioned that the Rhizosphere is strongly influenced by the plant roots. The icrobiota majorly produced sequence hits of genera belonging to the marine environment nd alkalihalotolerant in nature. This can be explained as the region was closer to the coastal ine leading to salty and marshy wet soil. The sequence of osmoprotectant glycinebetaine eflects the mechanism adapted by the microbes towards the protection against drastic onditions like drought, high temperature, high salt and high osmotic stress. Since the roots issues and rhizospheric microorganisms in the rhizosphere are mostly inaccessible to iron ue to low solubility, iron acquisition sequences were obtained. Sequence of anoxic hotoautotrophs which are Fe(II) oxidizers, methane oxidizers were also obtained. Majority ere chemoautolithotrophic and sulphur oxidizing. Capability to degrade complex arbohydrates like cellulose, alginate, xylan and chitin was reported by the unique genera resent in the rhizosphere. Additionally, sequences of microbes showed potential of iodegradation and utilization of complex aromatic hydrocarbons like phenanthene and nthracene and halogenated organics such as toxic chlorinated ethanes and polychlorinated iphenyls. Hence, the soil microbiota has a possible role in xenobiotic degradations.

he study to helped understand that soil is composed of myriad, rare and undiscovered icrobial species. Numerous genes and proteins present in the soil may be unexplored till ate. The taxonomic and functional profiling of soil metagenome can serve as a topographical ap of India and this is an an important pioneering study in this direction. Each microbial abitat offers its unique exclusivity. With the help of sequencing, the unique species were ghlighted from the metagenomes under study. Abundance of the particular microbial opulation may provide useful information of the physicochemical character of the soil icrobiota. The information gained by unique microbes over a particular community can ve as soil markers. The usefulness of the soil metagenome can be extrapolated to gineering soil transplantation successfully where the soil deficient is supplemented. ights from this study can be used to construct metagenomic libraries for targeted screening particular enzymes or can help in defining selective media for isolation of the roorganism of interest.