= Metagenomics =

Metagenomics is the study of genetic material recovered directly from environmental samples . The broad field may also be referred to as environmental genomics , ecogenomics or community genomics . While traditional microbiology and microbial genome sequencing and genomics rely upon cultivated clonal cultures , early environmental gene sequencing cloned specific genes (often the 16S rRNA gene) to produce a profile of diversity in a natural sample . Such work revealed that the vast majority of microbial biodiversity had been missed by cultivation @-@ based methods . Recent studies use either " shotgun " or PCR directed sequencing to get largely unbiased samples of all genes from all the members of the sampled communities . Because of its ability to reveal the previously hidden diversity of microscopic life , metagenomics offers a powerful lens for viewing the microbial world that has the potential to revolutionize understanding of the entire living world . As the price of DNA sequencing continues to fall , metagenomics now allows microbial ecology to be investigated at a much greater scale and detail than before .

= = Etymology = =

The term " metagenomics " was first used by Jo Handelsman , Jon Clardy , Robert M. Goodman , Sean F. Brady , and others , and first appeared in publication in 1998 . The term metagenome referenced the idea that a collection of genes sequenced from the environment could be analyzed in a way analogous to the study of a single genome . Recently , Kevin Chen and Lior Pachter (researchers at the University of California , Berkeley) defined metagenomics as " the application of modern genomics technique without the need for isolation and lab cultivation of individual species " .

= = History = =

Conventional sequencing begins with a culture of identical cells as a source of DNA . However , early metagenomic studies revealed that there are probably large groups of microorganisms in many environments that cannot be cultured and thus cannot be sequenced . These early studies focused on 16S ribosomal RNA sequences which are relatively short , often conserved within a species , and generally different between species . Many 16S rRNA sequences have been found which do not belong to any known cultured species , indicating that there are numerous non @-@ isolated organisms . These surveys of ribosomal RNA (rRNA) genes taken directly from the environment revealed that cultivation based methods find less than 1 % of the bacterial and archaeal species in a sample . Much of the interest in metagenomics comes from these discoveries that showed that the vast majority of microorganisms had previously gone unnoticed .

Early molecular work in the field was conducted by Norman R. Pace and colleagues , who used PCR to explore the diversity of ribosomal RNA sequences . The insights gained from these breakthrough studies led Pace to propose the idea of cloning DNA directly from environmental samples as early as 1985 . This led to the first report of isolating and cloning bulk DNA from an environmental sample , published by Pace and colleagues in 1991 while Pace was in the Department of Biology at Indiana University . Considerable efforts ensured that these were not PCR false positives and supported the existence of a complex community of unexplored species . Although this methodology was limited to exploring highly conserved , non @-@ protein coding genes , it did support early microbial morphology @-@ based observations that diversity was far more complex than was known by culturing methods . Soon after that , Healy reported the metagenomic isolation of functional genes from " zoolibraries " constructed from a complex culture of environmental organisms grown in the laboratory on dried grasses in 1995 . After leaving the Pace laboratory , Edward DeLong continued in the field and has published work that has largely laid the groundwork for environmental phylogenies based on signature 16S sequences , beginning with his group 's construction of libraries from marine samples .

In 2002 , Mya Breitbart , Forest Rohwer , and colleagues used environmental shotgun sequencing (

see below) to show that 200 liters of seawater contains over 5000 different viruses . Subsequent studies showed that there are more than a thousand viral species in human stool and possibly a million different viruses per kilogram of marine sediment , including many bacteriophages . Essentially all of the viruses in these studies were new species . In 2004 , Gene Tyson , Jill Banfield , and colleagues at the University of California , Berkeley and the Joint Genome Institute sequenced DNA extracted from an acid mine drainage system . This effort resulted in the complete , or nearly complete , genomes for a handful of bacteria and archaea that had previously resisted attempts to culture them .

Beginning in 2003 , Craig Venter , leader of the privately funded parallel of the Human Genome Project , has led the Global Ocean Sampling Expedition (GOS) , circumnavigating the globe and collecting metagenomic samples throughout the journey . All of these samples are sequenced using shotgun sequencing , in hopes that new genomes (and therefore new organisms) would be identified . The pilot project , conducted in the Sargasso Sea , found DNA from nearly 2000 different species , including 148 types of bacteria never before seen . Venter has circumnavigated the globe and thoroughly explored the West Coast of the United States , and completed a two @-@ year expedition to explore the Baltic , Mediterranean and Black Seas . Analysis of the metagenomic data collected during this journey revealed two groups of organisms , one composed of taxa adapted to environmental conditions of ' feast or famine ' , and a second composed of relatively fewer but more abundantly and widely distributed taxa primarily composed of plankton .

In 2005 Stephan C. Schuster at Penn State University and colleagues published the first sequences of an environmental sample generated with high @-@ throughput sequencing , in this case massively parallel pyrosequencing developed by 454 Life Sciences . Another early paper in this area appeared in 2006 by Robert Edwards , Forest Rohwer , and colleagues at San Diego State University .

= = Sequencing = =

Recovery of DNA sequences longer than a few thousand base pairs from environmental samples was very difficult until recent advances in molecular biological techniques allowed the construction of libraries in bacterial artificial chromosomes (BACs) , which provided better vectors for molecular cloning .

= = = Shotgun metagenomics = = =

Advances in bioinformatics , refinements of DNA amplification , and the proliferation of computational power have greatly aided the analysis of DNA sequences recovered from environmental samples , allowing the adaptation of shotgun sequencing to metagenomic samples . The approach , used to sequence many cultured microorganisms and the human genome , randomly shears DNA , sequences many short sequences , and reconstructs them into a consensus sequence . Shotgun sequencing reveals genes present in environmental samples . Historically , clone libraries were used to facilitate this sequencing . However , with advances in high throughput sequencing technologies , the cloning step is no longer necessary and greater yields of sequencing data can be obtained without this labour @-@ intensive bottleneck step . Shotgun metagenomics provides information both about which organisms are present and what metabolic processes are possible in the community . Because the collection of DNA from an environment is largely uncontrolled , the most abundant organisms in an environmental sample are most highly represented in the resulting sequence data . To achieve the high coverage needed to fully resolve the genomes of under @-@ represented community members , large samples , often prohibitively so , are needed . On the other hand , the random nature of shotgun sequencing ensures that many of these organisms , which would otherwise go unnoticed using traditional culturing techniques , will be represented by at least some small sequence segments .

= = = High @-@ throughput sequencing = = =

The first metagenomic studies conducted using high @-@ throughput sequencing used massively parallel 454 pyrosequencing . Three other technologies commonly applied to environmental sampling are the Ion Torrent Personal Genome Machine , the Illumina MiSeq or HiSeq and the Applied Biosystems SOLiD system . These techniques for sequencing DNA generate shorter fragments than Sanger sequencing ; Ion Torrent PGM System and 454 pyrosequencing typically produces ~ 400 bp reads , Illumina MiSeq produces 400 @-@ 700bp reads (depending on whether paired end options are used) , and SOLiD produce 25 @-@ 75 bp reads . Historically , these read lengths were significantly shorter than the typical Sanger sequencing read length of ~ 750 bp , however the Illumina technology is quickly coming close to this benchmark . However , this limitation is compensated for by the much larger number of sequence reads . In 2009 , pyrosequenced metagenomes generate 200 ? 500 megabases , and Illumina platforms generate around 20 ? 50 gigabases , but these outputs have increased by orders of magnitude in recent years . An additional advantage to high throughput sequencing is that this technique does not require cloning the DNA before sequencing , removing one of the main biases and bottlenecks in environmental sampling .

= = Bioinformatics = =

The data generated by metagenomics experiments are both enormous and inherently noisy , containing fragmented data representing as many as 10 @,@ 000 species . The sequencing of the cow rumen metagenome generated 279 gigabases , or 279 billion base pairs of nucleotide sequence data , while the human gut microbiome gene catalog identified 3 @.@ 3 million genes assembled from 567 @.@ 7 gigabases of sequence data . Collecting , curating , and extracting useful biological information from datasets of this size represent significant computational challenges for researchers .

= = = Sequence pre @-@ filtering = = =

The first step of metagenomic data analysis requires the execution of certain pre @-@ filtering steps , including the removal of redundant , low @-@ quality sequences and sequences of probable eukaryotic origin (especially in metagenomes of human origin) . The methods available for the removal of contaminating eukaryotic genomic DNA sequences include Eu @-@ Detect and DeConseq .

= = = Assembly = = =

DNA sequence data from genomic and metagenomic projects are essentially the same , but genomic sequence data offers higher coverage while metagenomic data is usually highly non @-@ redundant . Furthermore , the increased use of second @-@ generation sequencing technologies with short read lengths means that much of future metagenomic data will be error @-@ prone . Taken in combination , these factors make the assembly of metagenomic sequence reads into genomes difficult and unreliable . Misassemblies are caused by the presence of repetitive DNA sequences that make assembly especially difficult because of the difference in the relative abundance of species present in the sample . Misassemblies can also involve the combination of sequences from more than one species into chimeric contigs .

There are several assembly programs , most of which can use information from paired @-@ end tags in order to improve the accuracy of assemblies . Some programs , such as Phrap or Celera Assembler , were designed to be used to assemble single genomes but nevertheless produce good results when assembling metagenomic data sets . Other programs , such as Velvet assembler , have been optimized for the shorter reads produced by second @-@ generation sequencing through the use of de Bruijn graphs . The use of reference genomes allows researchers to improve the assembly of the most abundant microbial species , but this approach is limited by the small subset of microbial phyla for which sequenced genomes are available . After an assembly is created

, an additional challenge is " metagenomic deconvolution " , or determining which sequences come from which species in the sample .

= = = Gene prediction = = =

Metagenomic analysis pipelines use two approaches in the annotation of coding regions in the assembled contigs . The first approach is to identify genes based upon homology with genes that are already publicly available in sequence databases , usually by simple BLAST searches . This type of approach is implemented in the program MEGAN4 . The second , ab initio , uses intrinsic features of the sequence to predict coding regions based upon gene training sets from related organisms . This is the approach taken by programs such as GeneMark and GLIMMER . The main advantage of ab initio prediction is that it enables the detection of coding regions that lack homologs in the sequence databases ; however , it is most accurate when there are large regions of contiguous genomic DNA available for comparison .

= = = Species diversity = = =

Gene annotations provide the " what " , while measurements of species diversity provide the " who " . In order to connect community composition and function in metagenomes , sequences must be binned . Binning is the process of associating a particular sequence with an organism . In similarity @-@ based binning , methods such as BLAST are used to rapidly search for phylogenetic markers or otherwise similar sequences in existing public databases . This approach is implemented in MEGAN . Another tool , PhymmBL , uses interpolated Markov models to assign reads . MetaPhlAn and AMPHORA are methods based on unique clade @-@ specific markers for estimating organismal relative abundances with improved computational performances . In composition based binning , methods use intrinsic features of the sequence , such as oligonucleotide frequencies or codon usage bias . Once sequences are binned , it is possible to carry out comparative analysis of diversity and richness utilising tools such as Unifrac .

= = = Data integration = = =

The massive amount of exponentially growing sequence data is a daunting challenge that is complicated by the complexity of the metadata associated with metagenomic projects . Metadata includes detailed information about the three @-@ dimensional (including depth , or height) geography and environmental features of the sample , physical data about the sample site , and the methodology of the sampling . This information is necessary both to ensure replicability and to enable downstream analysis . Because of its importance , metadata and collaborative data review and curation require standardized data formats located in specialized databases , such as the Genomes OnLine Database (GOLD) .

Several tools have been developed to integrate metadata and sequence data , allowing downstream comparative analyses of different datasets using a number of ecological indices . In 2007 , Folker Meyer and Robert Edwards and a team at Argonne National Laboratory and the University of Chicago released the Metagenomics Rapid Annotation using Subsystem Technology server (MG @-@ RAST) a community resource for metagenome data set analysis . As of June 2012 over 14 @.@ 8 terabases (14x1012 bases) of DNA have been analyzed , with more than 10 @,@ 000 public data sets freely available for comparison within MG @-@ RAST . Over 8 @,@ 000 users now have submitted a total of 50 @,@ 000 metagenomes to MG @-@ RAST . The Integrated Microbial Genomes / Metagenomes (IMG / M) system also provides a collection of tools for functional analysis of microbial communities based on their metagenome sequence , based upon reference isolate genomes included from the Integrated Microbial Genomes (IMG) system and the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project .

One of the first standalone tools for analysing high @-@ throughput metagenome shotgun data was MEGAN (MEta Genome ANalyzer) . A first version of the program was used in 2005 to

analyse the metagenomic context of DNA sequences obtained from a mammoth bone . Based on a BLAST comparison against a reference database , this tool performs both taxonomic and functional binning , by placing the reads onto the nodes of the NCBI taxonomy using a simple lowest common ancestor (LCA) algorithm or onto the nodes of the SEED or KEGG classifications , respectively .

= = = Comparative metagenomics = = =

Comparative analyses between metagenomes can provide additional insight into the function of complex microbial communities and their role in host health . Pairwise or multiple comparisons between metagenomes can be made at the level of sequence composition (comparing GC @-@ content or genome size) , taxonomic diversity , or functional complement . Comparisons of population structure and phylogenetic diversity can be made on the basis of 16S and other phylogenetic marker genes , or ? in the case of low @-@ diversity communities ? by genome reconstruction from the metagenomic dataset . Functional comparisons between metagenomes may be made by comparing sequences against reference databases such as COG or KEGG , and tabulating the abundance by category and evaluating any differences for statistical significance . This gene @-@ centric approach emphasizes the functional complement of the community as a whole rather than taxonomic groups , and shows that the functional complements are analogous under similar environmental conditions . Consequently , metadata on the environmental context of the metagenomic sample is especially important in comparative analyses , as it provides researchers with the ability to study the effect of habitat upon community structure and function .

Additionally , several studies have also utilized oligonucleotide usage patterns to identify the differences across diverse microbial communities . Examples of such methodologies include the dinucleotide relative abundance approach by Willner et al. and the HabiSign approach of Ghosh et al . Ghosh et al . (2011) also indicated that differences in tetranucleotide usage patterns can be used to identify genes (or metagenomic reads) originating from specific habitats . Additionally some methods as TriageTools or Compareads detect similar reads between two read sets . The similarity measure they apply on reads is based on a number of identical words of length k shared by pairs of reads .

A key goal in comparative metagenomics is to identify microbial group (s) which are responsible for conferring specific characteristics to a given environment . However , due to issues in the sequencing technologies artifacts need to be accounted for like in metagenomeSeq . Others have characterized inter @-@ microbial interactions between the resident microbial groups . A GUI @-@ based comparative metagenomic analysis application called Community @-@ Analyzer has been developed by Kuntal et al. which implements a correlation @-@ based graph layout algorithm that not only facilitates a quick visualization of the differences in the analyzed microbial communities (in terms of their taxonomic composition) , but also provides insights into the inherent inter @-@ microbial interactions occurring therein . Notably , this layout algorithm also enables grouping of the metagenomes based on the probable inter @-@ microbial interaction patterns rather than simply comparing abundance values of various taxonomic groups . In addition , the tool implements several interactive GUI @-@ based functionalities that enable users to perform standard comparative analyses across microbiomes .

= = Data analysis = =

= = = Community metabolism = = =

In many bacterial communities , natural or engineered (such as bioreactors) , there is significant division of labor in metabolism (Syntrophy) , during which the waste products of some organisms are metabolites for others . In one such system , the methanogenic bioreactor , functional stability requires the presence of several syntrophic species (Syntrophobacterales and Synergistia) working together in order to turn raw resources into fully metabolized waste (methane) . Using comparative

gene studies and expression experiments with microarrays or proteomics researchers can piece together a metabolic network that goes beyond species boundaries . Such studies require detailed knowledge about which versions of which proteins are coded by which species and even by which strains of which species . Therefore , community genomic information is another fundamental tool (with metabolomics and proteomics) in the quest to determine how metabolites are transferred and transformed by a community .

= = = Metatranscriptomics = = =

Metagenomics allows researchers to access the functional and metabolic diversity of microbial communities , but it cannot show which of these processes are active . The extraction and analysis of metagenomic mRNA (the metatranscriptome) provides information on the regulation and expression profiles of complex communities . Because of the technical difficulties (the short half @-@ life of mRNA , for example) in the collection of environmental RNA there have been relatively few in situ metatranscriptomic studies of microbial communities to date . While originally limited to microarray technology , metatranscriptomcs studies have made use of direct high @-@ throughput cDNA sequencing to provide whole @-@ genome expression and quantification of a microbial community , as first employed by Leininger et al . (2006) in their analysis of ammonia oxidation in soils .

= = = Viruses = = =

Metagenomic sequencing is particularly useful in the study of viral communities . As viruses lack a shared universal phylogenetic marker (as 16S RNA for bacteria and archaea , and 18S RNA for eukarya) , the only way to access the genetic diversity of the viral community from an environmental sample is through metagenomics . Viral metagenomes (also called viromes) should thus provide more and more information about viral diversity and evolution . For example a metagenomic pipeline called Giant Virus Finder showed the first evidence of existence of giant viruses in a saline desert .

= = Applications = =

Metagenomics has the potential to advance knowledge in a wide variety of fields . It can also be applied to solve practical challenges in medicine , engineering , agriculture , sustainability and ecology .

= = = Medicine = = =

Microbial communities play a key role in preserving human health , but their composition and the mechanism by which they do so remains mysterious . Metagenomic sequencing is being used to characterize the microbial communities from 15 @-@ 18 body sites from at least 250 individuals . This is part of the Human Microbiome initiative with primary goals to determine if there is a core human microbiome , to understand the changes in the human microbiome that can be correlated with human health , and to develop new technological and bioinformatics tools to support these goals .

Another medical study as part of the MetaHit (Metagenomics of the Human Intestinal Tract) project consisted of 124 individuals from Denmark and Spain consisting of healthy , overweight , and irritable bowel disease patients . The study attempted to categorize the depth and phylogenetic diversity of gastrointestinal bacteria . Using Illumina GA sequence data and SOAPdenovo , a de Bruijn graph @-@ based tool specifically designed for assembly short reads , they were able to generate 6 @.@ 58 million contigs greater than 500 bp for a total contig length of 10 @.@ 3 Gb and a N50 length of 2 @.@ 2 kb .

The study demonstrated that two bacterial divisions , Bacteroidetes and Firmicutes , constitute over 90 % of the known phylogenetic categories that dominate distal gut bacteria . Using the relative

gene frequencies found within the gut these researchers identified 1 @,@ 244 metagenomic clusters that are critically important for the health of the intestinal tract . There are two types of functions in these range clusters : housekeeping and those specific to the intestine . The housekeeping gene clusters are required in all bacteria and are often major players in the main metabolic pathways including central carbon metabolism and amino acid synthesis . The gut @-@ specific functions include adhesion to host proteins and the harvesting of sugars from globoseries glycolipids . Patients with irritable bowel syndrome were shown to exhibit 25 % fewer genes and lower bacterial diversity than individuals not suffering from irritable bowel syndrome indicating that changes in patients ? gut biome diversity may be associated with this condition .

While these studies highlight some potentially valuable medical applications , only 31 @-@ 48 @.@ 8 % of the reads could be aligned to 194 public human gut bacterial genomes and 7 @.@ 6 @-@ 21 @.@ 2 % to bacterial genomes available in GenBank which indicates that there is still far more research necessary to capture novel bacterial genomes .

= = = Biofuel = = =

Biofuels are fuels derived from biomass conversion , as in the conversion of cellulose contained in corn stalks , switchgrass , and other biomass into cellulosic ethanol . This process is dependent upon microbial consortia that transform the cellulose into sugars , followed by the fermentation of the sugars into ethanol . Microbes also produce a variety of sources of bioenergy including methane and hydrogen .

The efficient industrial @-@ scale deconstruction of biomass requires novel enzymes with higher productivity and lower cost . Metagenomic approaches to the analysis of complex microbial communities allow the targeted screening of enzymes with industrial applications in biofuel production , such as glycoside hydrolases . Furthermore , knowledge of how these microbial communities function is required to control them , and metagenomics is a key tool in their understanding . Metagenomic approaches allow comparative analyses between convergent microbial systems like biogas fermenters or insect herbivores such as the fungus garden of the leafcutter ants .

= = = Environmental remediation = = =

Metagenomics can improve strategies for monitoring the impact of pollutants on ecosystems and for cleaning up contaminated environments . Increased understanding of how microbial communities cope with pollutants improves assessments of the potential of contaminated sites to recover from pollution and increases the chances of bioaugmentation or biostimulation trials to succeed .

= = = Biotechnology = = =

Microbial communities produce a vast array of biologically active chemicals that are used in competition and communication . Many of the drugs in use today were originally uncovered in microbes ; recent progress in mining the rich genetic resource of non @-@ culturable microbes has led to the discovery of new genes , enzymes , and natural products . The application of metagenomics has allowed the development of commodity and fine chemicals , agrochemicals and pharmaceuticals where the benefit of enzyme @-@ catalyzed chiral synthesis is increasingly recognized .

Two types of analysis are used in the bioprospecting of metagenomic data : function @-@ driven screening for an expressed trait , and sequence @-@ driven screening for DNA sequences of interest . Function @-@ driven analysis seeks to identify clones expressing a desired trait or useful activity , followed by biochemical characterization and sequence analysis . This approach is limited by availability of a suitable screen and the requirement that the desired trait be expressed in the host cell . Moreover , the low rate of discovery (less than one per 1 @,@ 000 clones screened) and its labor @-@ intensive nature further limit this approach . In contrast , sequence @-@ driven

analysis uses conserved DNA sequences to design PCR primers to screen clones for the sequence of interest . In comparison to cloning @-@ based approaches , using a sequence @-@ only approach further reduces the amount of bench work required . The application of massively parallel sequencing also greatly increases the amount of sequence data generated , which require high @-@ throughput bioinformatic analysis pipelines . The sequence @-@ driven approach to screening is limited by the breadth and accuracy of gene functions present in public sequence databases . In practice , experiments make use of a combination of both functional and sequence @-@ based approaches based upon the function of interest , the complexity of the sample to be screened , and other factors .

= = = Agriculture = = =

The soils in which plants grow are inhabited by microbial communities , with one gram of soil containing around 109 @-@ 1010 microbial cells which comprise about one gigabase of sequence information . The microbial communities which inhabit soils are some of the most complex known to science , and remain poorly understood despite their economic importance . Microbial consortia perform a wide variety of ecosystem services necessary for plant growth , including fixing atmospheric nitrogen , nutrient cycling , disease suppression , and sequester iron and other metals . Functional metagenomics strategies are being used to explore the interactions between plants and microbes through cultivation @-@ independent study of these microbial communities . By allowing insights into the role of previously uncultivated or rare community members in nutrient cycling and the promotion of plant growth , metagenomic approaches can contribute to improved disease detection in crops and livestock and the adaptation of enhanced farming practices which improve crop health by harnessing the relationship between microbes and plants .

= = = Ecology = = =

Metagenomics can provide valuable insights into the functional ecology of environmental communities . Metagenomic analysis of the bacterial consortia found in the defecations of Australian sea lions suggests that nutrient @-@ rich sea lion faeces may be an important nutrient source for coastal ecosystems . This is because the bacteria that are expelled simultaneously with the defecations are adept at breaking down the nutrients in the faeces into a bioavailable form that can be taken up into the food chain .

DNA sequencing can also be used more broadly to identify species present in a body of water , debris filtered from the air , or sample of dirt . This can establish the range of invasive species and endangered species , and track seasonal populations .