भाकृअनुप
ICAR

CAAST

Practical Manual
Students' Winter School

Course Director

Dr. C. Viswanathan

Course Coordinators

Dr. A. Kumar
Dr. K. Annapurna

Genomics of Plant Pathogens and Agriculturally Important Microbes

December 19th to 31th, 2018
at
Division of Plant Pathology
and
Division of Microbiology
ICAR-IARI, New Delhi

Organized by
National Agricultural Higher Education Project
Center for Advanced Agricultural Science
and Technology (CAAST)

भारतीय कृषि अनुसंधान संस्थान
IARI 1905

भाकृअनुप
ICAR

NAHEP sponsored

Students' Winter School

On

"Genomics of Plant Pathogens and Agriculturally Important Microbes"

December 19 -31, 2018

Course Director

Dr. C. Viswanathan,Head (Acting) and Principal Investigator, **NAHEP-Centre for Advanced Agricultural Science and Technology (CAAST)**, Division of Plant Physiology,
ICAR-Indian Agricultural Research Institute, Pusa Campus,
New Delhi-110012,
Email: viswanathan@iari.res.in
Phone:91-11-25842815, 09013885245

Course Coordinators

Dr. A. Kumar, Principal Scientist,
Division of Plant Pathology,
ICAR-Indian Agricultural Research Institute,
Pusa Campus,New Delhi-110012
Email: kumar@iari.res.in;
kaundy@yahoo.com
Phone: 09540829009

Dr. K. Annapurna, Head,
Division of Microbiology,
ICAR-Indian Agricultural Research Institute,
Pusa Campus
New Delhi 110 012
Email: annapurna96@yahoo.co.in
Phone: 011-25847649

INDIAN AGRICULTURAL
RESEARCH INSTITUTE

Division of Plant Pathology& Division of Microbiology
ICAR-Indian Agricultural Research Institute, Pusa Campus, New Delhi-110012

About NAHEP-CAAST at IARI, New Delhi

Centre for Advanced Agricultural Science and Technology (CAAST) is a new initiative and student centric sub-component of World Bank sponsored **National Agricultural Higher Education Project (NAHEP)** granted to The Indian Council of Agricultural Research, New Delhi to provide a platform for strengthening educational and research activities of post graduate and doctoral students. The ICAR-Indian Agricultural Research Institute, New Delhi was selected by the NAHEP-CAAST programme. NAHEP sanctioned Rs 19.98 crores for the project of IARI on "**Genomic assisted crop improvement and management**" under CAAST programme. The project at IARI specifically aims at inculcating genomics education and skills among the students and enhancing the expertise of the faculty of IARI in the area of genomics.

Objectives:

a) **To develop online teaching facility and online courses for enhancing the teaching and learning efficiency, and scientific communications skills**

b) **To develop and/or strengthen state-of-the art next-generation genomics and phenomics facilities for producing quality PG and Ph.D. students**

c) **To develop collaborative research programmes with institutes of international repute and industries in the area of genomics and phenomics**

d) **To enhance the skills of faculty and PG students of IARI and NARES**

e) **To generate and analyze big data in genomics and phenomics of crops, microbes and pests for genomics augmentation of crop improvement and management**

IARI's CAAST project is unique as it aimed at providing funding and training support to the M.Sc. and Ph.D. students from different disciplines who are working in the area of genomics. It will organize lectures and training programmes, and send IARI students and covering students from several disciplines. It will provide opportunities to the students and faculty to gain international exposure. Further, the project envisages developing a modern lab named as **Discovery Centre** that will serve as a common facility for students' research at IARI.

Acknowledgments

1. Secretary DARE and Director General ICAR, New Delhi
2. Deputy Director General (Education), ICAR, New Delhi
3. Assistant Director General (HRD), ICAR, New Delhi
4. National Coordinator, NAHEP, ICAR, New Delhi
5. CAAST Team, ICAR-IARI, New Delhi
6. P. G. School, ICAR-IARI, New Delhi
7. Director, ICAR-IARI, New Delhi
8. Dean and Joint Director (Education), ICAR-IARI, New Delhi
9. Joint Director (Research), ICAR-IARI, New Delhi
10. Head, Division of Plant Pathology, ICAR-IARI, New Delhi
11. Head, Division of Microbiology, ICAR-IARI, New Delhi
12. AKMU, ICAR-IARI, New Delhi
13. Eurofins Genomics India Pvt. Ltd, Bengaluru
14. SpincoBiotech Pvt. Ltd., Chennai
15. ICAR-National Research Centre on Plant Biotechnology, New Delhi
16. School of Life Sciences, Jawaharlal Nehru University, New Delhi
17. Department of Plant Pathology and Microbiology, Texas A&M University, Texas
18. ICAR-Indian Agricultural Statistical Research Institute, New Delhi
19. QIAGEN lndia Pvt. Ltd, New Delhi
20. Imperial Life Sciences, Gurugram

PREFACE

Global food production needs to keep pace with ever growing human population of 7 billion that is expected to touch 10 billion by 2050. With shrinking cultivable area and consequent ‘agricultural habitat loss’ for sustained crop production, one of the approaches for ensuring, sustaining and enhancing the agricultural productivity and nutritional security is by reducing the losses due to biotic and abiotic stress factors. Biotic factors like pest & diseases, multitude of climate and environmental related abiotic factors are among the major constraints that threaten global agriculture. Traditionally plant stresses, especially biotic ones are managed by deploying resistant cultivars and application of chemical molecules. These approaches, though very effective, are not universally adopted in all situations. While crop resistance is not durable, the chemicals are not a sustainable solution as a long term strategy. Therefore, novel and innovative approaches are, indeed, essential for mitigating the crop losses. Plant associated microorganism are known to play a vital role in shaping and guiding plant growth, development and confers defense against biotic and abiotic stresses. Microbial contributions in nutrient availability to plants and soil health is a well-researched area. In the recent years microbe assisted crop production is gaining momentum as a supplementary strategy in agriculture that is expected to make major impact in clean agricultural production. However, the vast diversity of microbial communities in plant associated niches is not exploited properly for want of appropriate technologies.

The cracking of first microbial genome by Craig Venter in the year 1996 has culminated in the birth of science of genomics. In the last two decade, ‘omics science’ and the genomic data has enabled us to understand diverse plant associated microbial communities, pathogens of crop plants and their behavior on plant associated niches. The exponential growth of genome related information and the associated “Omics tools” provided an opportunity for the plant pathologist and microbiologist to understand the population genetics of microorganisms and their host interactions at cellular and genome level.

The training manual is prepared with three major heading such as i. ***Milestones and terminologies used in genomics,*** **ii.** ***Methods and strategies for whole genome sequencing of plant pathogens*** **and Microorganisms and iii.** ***The Applications of genome sequence data of plant pathogens*** **and agriculturally important microbes.** We are of the opinion that the manual will be a useful guide to all students of plant pathology and microbiology as well as an information resource for the needy in future. We convey our sincere thanks to all those who have contributed in the preparation of this manual

Date: 10 December 2018

A. Kumar, Ph D
K. Annapurna, Ph D
C. Viswanathan, Ph D

FOREWORD

The ICAR-IARI, New Delhi has made significant contributions in developing crop protection and production technologies for all major crops in India. The institute has core strength in the area of genomics and modern research facilities for conducting advanced genomics programmes. Recently the institute has deciphered the whole genome sequences of agriculturally important free living nitrogen fixing diazotrophic microorganisms as *Pseudomonas stutzeri* and *Bacillus* species, *Magnaporthe oryzae* inciting blast in rice*, Tilletia indica* causing Karnal Bunt*, Cochliobolus sativus* causing spot blotch of wheat, *Puccinia striiformis* causing yellow rust of wheat*, Fusarium fujikuroii* inciting bakanae disease of rice*, Ralstonia solanacearum* causing wilt*, Meloidogyne graminicola* infecting riceand several plant viruses. Besides, metagenome analysis of plant microbiomes of major crops is also generated and published. At global level, a total of 312,877 whole genome sequencing projects encompassing most of the plant pathogens and microbes are underway that includes 278953 Prokaryotes, 24,987 Eukaryotes, and 8,937 viruses as on 23 August 2018 (https://gold.jgi.doe.gov/). Further, the Earth Microbiome Project proposed to analyze 200,000 samples from microbial communities using several genomic tools like amplicon sequencing, metagenomics, and metabolomics to produce a Global Gene Atlas describing protein space, environmental metabolic model for each biome, approximately 500,000 reconstructed microbial genomes, a global metabolic model, and a data-analysis portal for visualization of processed information. In order to harness the potential of these genome information, we need to create appropriate infrastructure facilities and human resources to face the challenges in the coming decades. With this background the Centre for Advanced *Agricultural Science and Technology* (CAAST) under NAHEP is organizing **2-weeks Students' Winter School** (SWS) on "**Genomics of plant pathogens and agriculturally important microbes" for the benefit of students of IARI, New Delhi. The SWS initiative itself is first-of-its-kind in the country. I am sure that the training will be useful to all PG students of IARI.**

Date: 17 December 2018

Dean and Joint Director (Education)

ICAR-IARI, New Delhi

Programme

Venue for lectures: Auditorium, Division of Plant Pathology
ICAR-Indian Agricultural Research Institute, New Delhi

Venue for practicals: PG Laboratory, Division of Plant Pathology
ICAR-Indian Agricultural Research Institute, New Delhi

	Wednesday, 19 Dec 2018
09:30-10:00h	**Registration**
10:00-10:10h	**Welcome address** **Dr. K. Manjaiah, Associate Dean, PG School**
10:10-10:20h	**About NAHEP-CAAST** **C. Viswanathan, PI cum Course Director**
10:20-10:30h	**About the Students Winter School** **Dr. A. Kumar**
10:30-10:45h	**Ice breaking; student's introduction**
10:45-10:55h	**Training inauguration & Inaugural address** **Dr. Rashmi Aggarwal, Dean (I/C), IARI**
10:55-11:00h	**Vote of thanks** **Dr. K. Annapurna**
11:00-11:30h	**Tea**
11:30-11:45h	**Pre-training evaluation** **Dr. A.Kumar and K. Annapurna**
12:00-13:00h	**Lecture 1: Genotyping of plant pathogens and Microbes: Strategies and Methods** **Dr. A. Kumar**
14:00-17:00h	**Practical 1 : DNA isolation from pathogenic fungi and bacteria** **Dr. A. Kumar /Dr. Deepa Kamil** **DNA isolation from PGP microorganisms** **Dr. K. Annapurna / Dr. V. Govindasamy/** **Dr. Ramakrishnan**
	Thursday, 20 Dec 2018
10:00-11:00h	**Lecture 2: Functional Genomics of agriculturally important microbes** **Dr. K. Annapurna, IARI, Delhi**
11:00-11:15h	**Tea**
11:30-13:00h	**Lecture 3: Small genome sequencing: experimental strategies and recent approaches** **Dr. Kishore Gaikwad, NRCPB, Delhi**

14:15-17:15h	Practical 1 (Continued) : DNA isolation from pathogenic fungi and bacteria Dr. A. Kumar /Dr. Deepa Kamil DNA isolation from PGP microorganisms Dr. K. Annapurna / Dr. V. Govindasamy/ Dr. Ramakrishnan
	Friday, 21 Dec 2018
10:00-11:00h	Lecture 5: Single Molecular Real Time Sequencing (SMRT): a revolutionary genome sequencing technology in 21st century Mr. Rakshit Chaudhary, PacBio, SpincoBiotech, Chennai
11:00-11:15h	Tea
11:30-13:00h	Lecture 4: Bioinformatic analyses of whole-genome sequence data Dr. Jyothi Malik, Qiagen, Delhi
14:15-17:15h	Practical 2: PCR Primer designing and validation (*Ralstonia solanacearum/Magnaporthe*) Drs. A. Kumar and Deeba Kamil PCR Primer designing and validation (PGP microorganisms) Drs. K. Annapurna , V. Govindasamy IARI, Delhi Practical 3: RAW sequence handling and curation, and assembly (*Ralstonia solanacearum/Magnaporthe*) Drs. A.Kumar and Deepa Kamil RAW sequence handling and curation, and assembly (PGP microorganisms) Drs. V. Govindasamy, B. Ramakrishnan and K. Annapurna, IARI, New Delhi
	Saturday, 22 Dec 2018
10:00-11:00h	Lecture 6: Nanopore - Next Generation hassle free Genome Sequencing Dr. Paras Yadav, ILS, Delhi
11:00-11:15h	Tea
11:30-13:00h	Lecture 7: Functional genomics of fungal pathogens Dr. Rashmi Aggarwal, IARI, New Delhi
14:15-17:15h	Practical 4: Gene annotation and preparation of data for accessioning Drs. A. Kumar, V. Govindasamy, S. Subramanian, and K. Annapurna, IARI, New Delhi
	Sunday, 23 Dec 2018 Holiday *Monday, 24 Dec 2018*
10:00-11:00h	Lecture 8: Genome Assembly: Concepts and updates Dr. Arpita Ghosh, Eurofins, New Delhi
11:00-11:15h	Tea

11:30-13:00h	Lecture 9: Methods to decode plant viral genome and their pathogenesis
	Dr. Supria Chakraborthi, JNU, Delhi
14:15-17:15h	Practical 5: Molecular phylogeny
	Dr. Anirban Roy and A. Kumar, IARI, New Delhi
Tuesday, 25 Dec 2018 Holiday	
Wednesday, 26 Dec 2018	
10:00-11:00h	Lecture 10: Strategies for transcriptome sequencing
	Dr. C. Viswanathan, IARI, New Delhi
11:00-11:15h	Tea
11:30-13:00h	Lecture 11: Databases for microbial and pathogenomics
	Dr. B. Ramakrishnan , IARI, New Delhi
14:15-17:15h	Practical 7: Bacterial transformation using electroporation for tagging with gfp gene
	Dr. A.Kumar, IARI, New Delhi
Thursday, 27 Dec 2018	
10:00-11:00h	Lecture 12: Whole genome based molecular phylogeny
	Dr. Anirban, IARI, New Delhi
11:00-11:15h	Tea
11:30-13:00h	Lecture 13: Viral Genomics
	Dr. V. K. Baranwal, IARI, New Delhi
14:15-17:15h	Practical 8: RNA isolation for RNA seq and qPCR
	Drs. A. Kumar and V. Govindasamy, IARI, New Delhi
Friday, 28 Dec 2018	
10:00-11:00h	Lecture 14: Application of qPCR in plant pathology and microbiology
	Dr. A.Kumar, IARI, New Delhi
11:00-12:00h	Practical 9: qPCR for validation of RNA seq data
	Dr. A. Kumar IARI, New Delhi
14:15-16:15h	Visit to Phenomics facility,
	Dr. C. Viswanathan
	Visit to Sequencing facility,
	Dr. B. Ramcharan, NRCPB, Delhi
	Visit to Bioinformatic facility,
	Dr. A. R. Rao
Saturday, 29 Dec 2018	
10:00-11:00h	Lecture 15: Gene finding strategies and their validation *in silico*
	Dr. Dinesh Kumar, IASRI, New Delhi
11:00-11:15h	Tea

11:30-13:00h	**Lecture 16: Data mining and machine learning tools for whole genome sequencing**
	Dr. A. R. Rao, IASRI, New Delhi
14:15-17:15h	**Practical 10: qPCR for absolute quantitation of microorganisms in environmental samples**
	Dr. B. Ramakrishnan, IARI, New Delhi
Sunday 30 Dec 2018 Holiday	
Monday, 31 Dec 2018	
10:00-11:00h	**Lecture 17: Nematode genomics-an update**
	Dr. Umarao, IARI, New Delhi
11:00-11:15h	**Tea**
11:30-12:30h	**Interaction with students**
12:30 -13:00h	**Post training evaluation**
	Drs. A. Kumar and K. Annapurna, IARI, New Delhi
14:15-17:15h	**Valediction and certificate distribution**
	CAAST-Team

Contents

1. Glossary of terms used in Omics-Science

Kuleshwar Prasad Sahu and A. Kumar
Division of Plant Pathology, ICAR-Indian Agricultural Research Institute,
New Delhi-110012

Accession number: An accession number is a unique identifier given to a sequence when it is submitted to one of the DNA repositories (GenBank, EMBL, and DDBJ). The initial deposition of a sequence record is referred to as version 1. If the sequence is updated, the version number is incremented but the accession number will remain constant.

aCGH: A technique involving the competitive hybridization of "test" and "reference" DNA probes to target genomic (or cDNA clones) immobilized on a microarray. Most often used for the detection of copy number variation (CNV), aCGH also has applications in gene annotation and diagnostics.

AGP: A file that describes how primary sequences can be assembled to make a non-redundant, contiguous sequence. The sequence being assembled may be a contig or a chromosome. This file describes the portion of the component sequence used in the contig, in addition to the location on the contig of the component sequence.

Algorithm: Formal set of instructions that tell a computer how to solve a problem or execute a task. A computer program typically consists of several algorithms.

Allele: One of two or more forms of a gene or a genetic region (generally containing a group of genes). A population or species of organisms typically includes multiple alleles at each locus distributed among various individuals; except very rarely, each individual can have only two alleles at a given locus. Allelic variation at a locus is measurable as the number of alleles (polymorphism) present, or the proportion of heterozygotes in the population. *See also:* locus, gene expression

Allelic series: A collection of distinct mutations that affect a single locus. Often, these different mutations will produce different phenotypes, thus providing a powerful genetic tool for the dissection of gene function.

Amplification: An increase in the number of copies of a specific DNA fragment; can be *in vivo* or *in vitro*. *See also:* cloning, polymerase chain reaction

Annotation: Adding biological information to genome sequence. This is a very complex task, and the process for doing this is rapidly evolving. Several groups are doing automated computational annotation of several genomes. Features that are added to the genome often include gene models, SNPs, and STSs.

Antisense: Nucleotide sequence exactly opposite to an mRNA molecule made through transcription; under given circumstances an antisense oligonucleotide binds to the mRNA molecule to prevent a protein from being made. *See also:* transcription

Arrayed library: Individual primary recombinant clones (hosted in phage, cosmid, YAC, or other vector) that are placed in two-dimensional arrays in microtiter dishes. Each primary clone can be identified by the identity of the plate and the clone location (row and column) on that plate. Arrayed libraries of clones can be used for many applications, including screening for a specific gene or genomic region of interest. *See also:* library, genomic library, gene chip technology

Assembly: Putting sequenced fragments of DNA into their correct chromosomal positions.

Autoradiography: A technique that uses X-ray film to visualize radioactively labeled molecules or fragments of molecules; used in analyzing length and number of DNA fragments after they are separated by gel electrophoresis.

Bacterial artificial chromosome (BAC): A vector used to clone DNA fragments (100- to 300-kb insert size; average, 150 kb) in *Escherichia coli* cells. Based on naturally occurring F-factor plasmid found in the bacterium *E. coli*. .

Base pair (bp): Two nitrogenous bases (adenine and thymine or guanine and cytosine) held together by weak bonds. Two strands of DNA are held together in the shape of a double helix by the bonds between base pairs.

Base sequence analysis: A method, sometimes automated, for determining the base sequence.

Base sequence: The order of nucleotide bases in a DNA molecule; determines structure of proteins encoded by that DNA.

Base: One of the molecules that form DNA and RNA molecules. *See also:* nucleotide, base pair, base sequence

BES: BAC end sequence. The ends of BACs are sequenced and the clone association information is retained. In this way, BAC clones that do not have insert sequence can be integrated with other BAC clones, or with WGS assemblies.

Bioinformatics: The science of managing and analyzing biological data using advanced computing techniques. Especially important in analyzing genomic research data. *See also:* informatics

Biotechnology: A set of biological techniques developed through basic research and now applied to research and product development. In particular, biotechnology refers to the use by industry of recombinant DNA, cell fusion, and new bioprocessing techniques.

BLAST: Basic Local Alignment Search Tool. Computer program that, when given any nucleotide or amino acid (protein) sequence, returns similar sequences retrieved from a chosen query database, usually the non-redundant database which houses all submitted DNA or Protein sequences without duplication of a given sequence. This algorithm has been extended and now includes a suite of programs including megablast and discontiguous megablast.

BLAT: A hashing algorithm developed by Jim Kent to allow rapid searching of large amounts of genome sequence. A hashing algorithm divides the database into words of a prescribed size (often 12-14 bases). The locations of these words are stored in memory. The query sequence is scanned for exact matches to words stored in memory. These types of algorithms tend to be very fast and effective for closely related sequences, but fail as sequences diverge. In addition to nucleotide BLAT, translated BLAT allows for comparison of protein sequences. This sequence aligner also allows for accurate alignment of transcribed sequences by looking at splice site information.

CAGE: Cap Analysis Gene Expression. Technique for identifying transcription start sites and quantifying promoter usage in eukaryotic genomes. The method is based on the isolation and concatenation of short sequence tags (~21 bp) from the 5' ends of individual mRNAs into longer DNA molecules that are subsequently sequenced. Transcriptional start sites are determined via mapping of tags to a reference genome

Candidate gene: A gene located in a chromosome region suspected of being involved in a given trait or function. *See also:* positional cloning, protein

Capillary array: Gel-filled silica capillaries used to separate fragments for DNA sequencing. The small diameter of the capillaries permit the application of higher electric fields, providing high speed, high throughput separations that are significantly faster than traditional slab gels.

cDNA library: A collection of DNA sequences that code for genes. The sequences are generated in the laboratory from mRNA sequences. *See also:* messenger RNA

CDS: Coding sequence. This is the portion of an mRNA or genomic sequence that encodes for a protein sequence.

Centimorgan (cM): A unit of measure of recombination frequency. One centimorgan is equal to a 1% chance that a marker at one genetic locus will be separated from a marker at a second locus due to crossing over in a single generation. *See also:* megabase

Chimera (pl. chimaera): An organism that contains cells or tissues with a different genotype. These can be mutated cells of the host organism or cells from a different organism or species.

Chimeraplasty: An experimental targeted repair process in which a desirable sequence of DNA is combined with RNA to form a chimeraplast. These molecules bind selectively to the target DNA. Once bound, the chimeraplast activates a naturally occurring gene-correcting mechanism. Does not use viral or other conventional gene-delivery vectors. *See also:* cloning vector

ChIP/chip: The hybridization of ChIP purified DNA to microarrays containing genomic DNA sequences to achieve genome-wide identification of protein-DNA interactions.

ChIP/SAGE: The preparation of small tags from ChIP purified DNA and their subsequent SAGE analysis to achieve genome-wide identification of protein-DNA interactions.

ChIP/SEQ: A technique involving size selection, high throughput sequencing (typically using next generation sequencing technologies that produce millions of reads in a run) and mapping of ChIP purified DNA onto a reference genome to achieve genome-wide identification of protein-DNA interactions.

ChIP: Chromatin Immunoprecipitation. A method for identifying protein-DNA interactions. Genomic DNA and associated proteins are cross-linked, sheared, and immunoprecipitated with antibodies that recognize specific DNA proteins. Purified DNA fragments are then assayed by various techniques to determine the association of specific sequences with the protein of interest

Chromatin immunoprecipitation (ChIP): Method used to determine the location in a genome of DNA binding sites recognized by a particular protein of interest.

Chromosome painting: Attachment of certain fluorescent dyes to targeted parts of the chromosome.

Cloning vector: DNA molecule originating from a virus, a plasmid, or the cell of a higher organism into which another DNA fragment of appropriate size can be integrated without loss of the vector's capacity for self-replication; vectors introduce foreign DNA into host cells, where the DNA can be reproduced in large quantities. Examples are plasmids, cosmids, and yeast artificial chromosomes; vectors are often recombinant molecules containing DNA sequences from several sources.

Cloning: Using specialized DNA technology to produce multiple, exact copies of a single gene or other segment of DNA to obtain enough material for further study. The resulting cloned (copied) collections of DNA molecules are called clone libraries. A second type of cloning exploits the natural process of cell division to make many copies of an entire cell. The genetic makeup of these cloned cells, called a cell line, is identical to the original cell. *See also:* cloning vector

Code: *See:* genetic code

Codon: *See:* genetic code

Comparative genomics: The study of genetics by comparisons with other organisms.

Complementary DNA (cDNA): DNA that is synthesized in the laboratory from a messenger RNA template.

Complementary RNA (cRNA): Synthetic transcripts of a specific DNA molecule or fragment made by an *in vitro* transcription system.

Complementary sequence: Nucleic acid base sequence that can form a double-stranded structure with another DNA fragment by following base-pairing rules (A pairs with T and C with G). The complementary sequence to GTAC for example, is CATG.

Complete Genome Sequence: High-quality, low error, gap-free DNA sequence of an entire genome of an organism.

Component: A sequence used to construct a larger sequence (a sequence contig or a scaffold). Typically this is a sequence found in GenBank/EMBL/DDBJ, often a clone sequence or a WGS contig but occasionally a PCR product.

Computational biology: Development and application of data-analysis and theoretical methods, mathematical modeling, and computational simulation techniques to the study of biological systems. *See also:* bioinformatics

Conserved hypothetical proteins: The (often large) fraction of genes in sequenced genomes encoding proteins that are found in organisms from several phylogenetic lineages but have not been functionally characterized and described at the protein chemical level. These structures may represent up to half of the potential protein coding regions of a genome.

Conserved sequence: A base sequence in a DNA molecule (or an amino acid sequence in a protein) that has remained essentially unchanged throughout evolution.

Constitutive ablation: Gene expression that results in cell death.

Contig map: A map depicting the relative order of a linked library of overlapping clones representing a complete chromosomal segment.

Contig: This is short for contiguous sequence. When two sequences overlap at their ends. The sequences can be collapsed into a single, non-redundant sequence.

Copy Number Variation (CNV): Large-scale structural changes in DNA that vary from individual to individual. These include insertions, deletions, duplications and complex multi-site variants that range from kilobases to megabases in size. CNV can influence gene expression, phenotypic variation and alter gene dosage, and in certain instances may be associated with developmental disorders, cause disease or confer susceptibility to complex disease traits.

Cosmid: Cloning vector that typically contains insert sizes of 60-120kb. These vectors are hybrids of lambda phages and plasmids and can be propagated as plasmids or packaged like phage. The name comes from the fact that these vectors retain the phage cos sites that are used for lambda head stuffing. These are generally maintained in multiple copies in E. coli.

Cytological map: A type of chromosome map whereby genes are located on the basis of cytological findings obtained with the aid of chromosome mutations.

Data warehouse: A collection of databases, data tables, and mechanisms to access the data on a single subject.

Deletion map: A description of a specific chromosome that uses defined mutations --specific deleted areas in the genome-- as 'biochemical signposts,' or markers for specific areas.

Deletion: A loss of part of the DNA from a chromosome; can lead to a disease or abnormality.

Diploid: A full set of genetic material consisting of paired chromosomes, one from each parental set.

Directed evolution: A laboratory process used on isolated molecules or microbes to cause mutations and identify subsequent adaptations to novel environments.

Directed mutagenesis: Alteration of DNA at a specific site and its reinsertion into an organism to study any effects of the change.

Directed sequencing: Successively sequencing DNA from adjacent stretches of chromosome.

DNA (deoxyribonucleic acid): Molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between base pairs of nucleotides. The four nucleotides in DNA contain the bases adenine (A), guanine (G), cytosine (C), and thymine (T). A pairs with T and C pairs with G.

DNA annotation: *See* genome annotation

DNA assembly: *See* genome assembly

DNA probe: *See:* probe

DNA repair genes: Genes encoding proteins that correct errors in DNA sequencing.

DNA sequence: The relative order of base pairs, whether in a DNA fragment, gene, chromosome, or an entire genome.

Domain: A discrete portion of a protein with its own function. The combination of domains in a single protein determines its overall function.

Double helix: The twisted-ladder shape that two linear strands of DNA assume when complementary nucleotides on opposing strands bond together.

Draft sequence: This term has had several definitions, but generally refers to sequence that is not yet finished but is of generally high quality. In terms of clone based project, Draft sequence refers to a project in which greater than 90% of the bases are of high quality. DNA sequence that, while incomplete, offers a virtual road map to many of the known genes. A Draft sequence data are mostly in the form of large-sized base pair fragments whose approximate chromosomal locations are known. However, these sequences, in conjunction with other data are a useful substrate for genome assembly and annotation.

Electrophoresis: A method of separating large molecules (such as DNA fragments or proteins) from a mixture of similar molecules. An electric current is passed through a medium containing the mixture, and each kind of molecule travels through the medium at a different rate, depending on its electrical charge and size. Agarose and acrylamide gels are the media commonly used for electrophoresis of proteins and nucleic acids.

Electroporation: A process using high-voltage current to make cell membranes permeable to allow the introduction of new DNA; commonly used in recombinant DNA technology. *See also:* transfection

End Sequence Profiling (ESP): A method for detecting genome-level variation. End sequences of clones from a library of interest are mapped onto a reference genome (in silico). Analysis of end sequence density and end sequence pair plots can reveal regions containing translocations, inversions, deletions/insertions and other complex structures. See also Paired End Mapping (PEM).

Endonuclease: *See:* restriction enzyme

Enzyme: A protein that acts as a catalyst, speeding the rate at which a biochemical reaction proceeds but not altering the direction or nature of the reaction.

e-PCR: Electronic PCR. A program that searches a given sequence for the presence of primer pairs. These primers must be in the proper orientation and a specified distance apart to define a match. There are currently two versions of this program. Forward e-PCR takes a sequence as a query and searches a database of sequence tag sites (STSs) while reverse e-PCR uses an STS to search a sequence for the presence of the primers in the correct orientation at the specified distance.

Epigenome: Set of chemical compounds that modify, or mark, the genome in a way that tells it what to do, where to do it, and when to do it. The marks, which are not part of the DNA itself, can be passed on from cell to cell as cells divide, and from one generation to the next.

Epigenomics: Study of the complete set of epigenetic modifications on the genetic material of a cell, known as the epigenome. Epigenetic modifications are reversible modifications on a cell's DNA or histones that affect gene expression without altering the DNA sequence.

Epistasis: Phenomenon where the effects of one gene are modified by one or several other genes. The gene whose phenotype is expressed is called epistatic, while the phenotype altered or suppressed is called hypostatic.

EST: Expressed sequence tag. These are single-pass sequences of cDNA clones. Databases of EST sequences are highly redundant but quite useful for gene identification. There are many efforts to cluster EST sequences to remove the redundancy and low-quality sequences.

ExoFish: A technique that utilizes Whole Genome Shotgun (WGS) reads from the pufferfish, *Tetraodan nigroviridis*, to identify potential coding sequences in mammalian genomes based on homology. This technique was first used to annotate the Human Genome.

Exogenous DNA: DNA originating outside an organism that has been introduced into the organism.

Exon: The protein-coding DNA sequence of a gene. *See also:* intron

Exonuclease: An enzyme that cleaves nucleotides sequentially from free ends of a linear nucleic acid substrate.

Expressed sequence tag (EST): A short strand of DNA that is a part of a cDNA molecule and can act as identifier of a gene. Used in locating and mapping genes.

Expression quantitative trait locus (eQTL): Genomic locus that regulates expression levels of mRNAs or proteins. *See also:* quantitative trait locus

Fingerprint: The pattern of bands produced by a clone when restricted by a particular enzyme, such as HindIII. Clones that are related will have fingerprint bands in common. The more bands in common, the greater the degree of overlap.

Finished Sequence: A clone insert that has been sequenced with an error rate of 0.01%. These sequence records generally have no gaps.

Fluorescence in situ hybridization (FISH): A physical mapping approach that uses fluorescein tags to detect hybridization of probes with metaphase chromosomes and with the less-condensed somatic interphase chromatin.

Fosmid: A cloning system based on the *E. coli* F factor. These clones have an average insert size of 40 Kb, with a very small standard deviation.

Full gene sequence: The complete order of bases in a gene. This order determines which protein a gene will produce.

Functional annotation: Process of attaching biological information (e.g., biochemical function, biological function, involved regulation and interactions, and expression) to genomic elements. *See also:* genome annotation

Functional genomics: Study of sequencing data to describe gene (and protein) functions and interactions. Unlike genomics, functional genomics focuses on dynamic aspects such as gene transcription, translation, and protein-protein interactions, as opposed to the static aspects of genomic information such as DNA sequence or structures.

Gap: A region of the genome for which no sequence is currently available. There are two types of gaps: heterochromatic and euchromatic. Heterochromatic gaps consist largely of highly repetitive sequence, while euchromatic gaps are more likely to contain genes. Gaps may occur both within and between genomic scaffolds.

GC-rich area: Many DNA sequences carry long stretches of repeated G and C which often indicate a gene-rich region.

Gene amplification: Repeated copying of a piece of DNA. *See also:* gene

Gene chip technology: Development of cDNA microarrays from a large number of genes. Used to monitor and measure changes in gene expression for each gene represented on the chip.

Gene expression: Process by which a gene's coded information is converted into structures present and operating in the cell. Expressed genes include those transcribed into messenger RNA (mRNA) and then translated into proteins, as well as those transcribed into RNA but not translated into proteins [e.g., transfer (tRNA) and ribosomal RNA (rRNA)].

Gene family: Group of closely related genes that make similar products.

Gene function: Biochemical reaction, protein-protein interaction, metabolic or signaling pathway association, cellular localization, phenotype, and changes in protein function that are mediated by shifts in protein structure.

Gene library: *See:* genomic library

Gene mapping: Determination of the relative positions of genes on a DNA molecule (chromosome or plasmid) and of the distance, in linkage units or physical units, between them.

Gene pool: All the variations of genes in a species. *See also:* allele, gene, polymorphism

Gene prediction: Predictions of possible genes made by a computer program based on how well a stretch of DNA sequence matches known gene sequences

Gene product: Biochemical material, either RNA or protein, resulting from expression of a gene. The amount of gene product is used to measure a gene's level of expression (transcription).

Gene regulatory network: Intracellular network of regulatory proteins that control the expression of gene subsets involved in particular cellular functions. A simple network would consist of one or more input signaling pathways, regulatory proteins that integrate the input signals, several target genes (in bacteria a target operon), and the RNA and proteins produced from those target genes.

Gene targeting: This is a specific type of transgenesis that targets a particular gene. If a mutated copy of a gene is electroporated into a cell, the inserted DNA will find the endogenous copy of itself and recombination will occur with some frequency (1-25%). If this event occurs in embryonic stem cells, cells carrying the new copy of the gene can be used to generate embryos that can be assessed for the phenotypic consequences of the mutation. This technique is used frequently in mice to study loss-of - function mutations.

Gene transfer: Incorporation of new DNA into and organism's cells, usually by a vector such as a modified virus. *See also:* mutation, vector

Gene trapping: This strategy uses transgenesis to introduce DNA carrying a reporter gene (lacZ or GFP) flanked by various genomic signals (splice donor or acceptor sites, promoters, etc.). Expression of the reporter gene indicates that the DNA has integrated into a region of the genome containing a gene. The gene that has been trapped can be recovered using the DNA sequences associated with the reporter construct. Often, the introduction of the gene trapping vector inactivates the gene into which it was introduced.

Gene: Fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides, located in a particular position on a particular chromosome that encodes a specific functional product (i.e., a protein or RNA molecule). Multiple variants (see allele) can exist in a population.

Genetic code: The sequence of nucleotides, coded in triplets (codons) along the mRNA that determines the sequence of amino acids in protein synthesis. A gene's DNA sequence can be used to predict the mRNA sequence, and the genetic code can in turn be used to predict the amino acid sequence.

Genetic engineering: Altering the genetic material of cells or organisms to enable them to make new substances or perform new functions.

Genetic informatics: *See:* bioinformatics

Genetic map: *See:* linkage map

Genetic marker: A gene or other identifiable portion of DNA whose inheritance can be followed. *See also:* chromosome, DNA, gene, inherit

Genetic material: *See:* genome

Genetic mosaic: An organism in which different cells contain different genetic sequence. This can be the result of a mutation during development or fusion of embryos at an early developmental stage.

Genetics: The study of inheritance patterns of specific traits.

Genome annotation: Process of identifying elements in the genome and attaching biological information to these elements. Automatic annotation tools perform this process by computer analysis, as opposed to manual annotation (i.e., curation), which involves human expertise. Ideally, these approaches co-exist and complement each other in the same annotation pipeline. *See also:* functional annotation, structural annotation

Genome assembly: Process of taking a large number of short DNA sequences and putting them back together to create a representation of the original chromosomes from which the DNA originated. In a shotgun sequencing project, the entire DNA from a source is first fractured into millions of small pieces. These pieces are then "read" by automated sequencing machines, which can read up to 1,000 nucleotides or bases at a time. A genome assembly algorithm works by taking all the pieces and aligning them to one another, and detecting all places where two of the short sequences, or reads, overlap. These overlapping reads can be merged, and the process continues.

Genome engineering: Techniques for the targeted, specific modification of the genetic information (or genome) of living organisms.

Genome project: Research and technology-development effort aimed at mapping and sequencing the genome of organisms.

Genome sequence: Order of nucleotides or bases within DNA molecules that make up an organism's entire genome. The four bases are adenine, guanine, cytosine, and thymine, represented as A, G, C, and T.

Genome: All the genetic material in the chromosomes of a particular organism. Most prokaryotes package their entire genome into a single chromosome, while eukaryotes have different numbers of chromosomes. Genome size generally is given as total number of base pairs.

Genome-wide association study (GWAS): Examination of many common genetic variants in different organisms to see if any variant is statistically associated with a trait. GWAS are used to identify candidate genes or sequence variants that may link to a condition or property of interest.

Genomic library: A collection of clones made from a set of randomly generated overlapping DNA fragments that represent the entire genome of an organism. *See also:* library, arrayed library

Genomics: The study of genes and their function.

Genotype: An organism's genetic constitution, as distinguished from its physical characteristics (phenotype).

Haploid: A single set of chromosomes (half the full set of genetic material) present in the egg and pollen cells of plants. *See also:* diploid

Haplotype (haploid genotype): A set of closely linked genetic markers present on one chromosome that tend to be inherited together. A haplotype may also refer to a set of single nucleotide polymorphisms (SNPs) on a single chromatid that are statistically associated with one another.

Haplotype: A segment of DNA containing closely linked gene variations that are inherited as a unit.

Heterozygous: Having two different alleles for a single trait, or having two different alleles at a single gene or genetic locus. An allele can be dominant, co-dominant, or recessive.

High throughput: Done on a massive, automated scale.

Highly conserved sequence: DNA sequence that is very similar across several different types of organisms. *See also:* gene, mutation

High-throughput sequencing: A fast method of determining the order of bases in DNA. *See also:* sequencing

Histone: Protein that provides structural support to a chromosome. For very long DNA molecules to fit into the cell nucleus, they wrap around complexes of histone proteins, giving the chromosome a more compact shape. Some histones variants are associated with the regulation of gene expression.

Homeobox: A short stretch of nucleotides whose base sequence is virtually identical in all the genes that contain it. Homeoboxes have been found in many organisms. In the fruit fly, a homeobox appears to determine when particular groups of genes are expressed during development.

Homolog: A member of a chromosome pair in diploid organisms or a gene that has the same origin and functions in two or more species.

Homologous chromosome: Chromosome containing the same linear gene sequences as another, each derived from one parent.

Homologous recombination: Swapping of DNA fragments between paired chromosomes.

Homology: Similarity in DNA sequence or structure based on descent from a common ancestor.

Horizontal gene transfer: Exchange of genetic material between two different organisms (typically different species of prokaryotes). This process gives prokaryotes the ability to obtain novel functionalities or cause dramatic changes in community structure over relatively short periods of time.

HTGS: High Throughput Genome Sequence. This is a term to distinguish all genomic sequence generated in a high-throughput manner. In order to release data more rapidly, it became standard for all sequence centers to submit unfinished sequence into public repositories. This sequence is deposited into the HTG division of GenBank/EMBL/DDBJ.

Hybridization: The process of joining two complementary strands of DNA or one each of DNA and RNA to form a double-stranded molecule.

In silico: Using computers to simulate and investigate natural processes.

In situ hybridization: Use of a DNA or RNA probe to detect the presence of the complementary DNA sequence in cloned bacterial or cultured eukaryotic cells.

In situ: In a natural environment.

In vitro: Outside of a living organism.

In vivo: Within a living organism.

Insertion: A chromosome abnormality in which a piece of DNA is incorporated into a gene and thereby disrupts the gene's normal function. *See also:* chromosome, DNA, gene, mutation

Interactome: Molecular interactions of a cell, typically used to describe all protein-protein interactions or those between proteins and other molecules.

Interference: One crossover event inhibits the chances of another crossover event. Also known as positive interference. Negative interference increases the chance of a second crossover. *See also:* crossing over

Intron: DNA sequence that interrupts the protein-coding sequence of a gene; an intron is transcribed into RNA but is cut out of the message before it is translated into protein. *See also:* exon

Junk DNA: Stretches of DNA that do not code for genes; most of the genome consists of so-called junk DNA which may have regulatory and other functions. Also called non-coding DNA.

Karyotype: A photomicrograph of individual chromosomes arranged in a standard format showing the number, size, and shape of each chromosome type; used in low-resolution physical mapping.

Kilobase (kb): Unit of length for DNA fragments equal to 1000 nucleotides.

Knockout: Deactivation of specific genes; used in laboratory organisms to study gene function. *See also:* gene, locus

Lateral gene transfer: Exchange of genetic material between two different organisms (typically different species of prokaryotes). This process gives prokaryotes the ability to obtain novel functionalities or cause dramatic changes in community structure over relatively short periods of time.

Library: An unordered collection of clones (i.e., cloned DNA from a particular organism) whose relationship to each other can be established by physical mapping. *See also:* genomic library, arrayed library

Linkage map: A map of the relative positions of genetic loci on a chromosome, determined on the basis of how often the loci are inherited together. Distance is measured in centimorgans (cM).

Linkage: The proximity of two or more markers (e.g., genes, RFLP markers) on a chromosome; the closer the markers, the lower the probability that they will be separated during DNA repair or replication processes (binary fission in prokaryotes, mitosis or meiosis in eukaryotes), and hence the greater the probability that they will be inherited together.

Localize: Determination of the original position (locus) of a gene or other marker on a chromosome.

Loci: Chromosomal locations of genes or genetic markers. (Singular: locus)

Long-Range Restriction Mapping: Restriction enzymes are proteins that cut DNA at precise locations. Restriction maps depict the chromosomal positions of restriction-enzyme cutting sites. These are used as biochemical "signposts," or markers of specific areas along the chromosomes. The map will detail the positions where the DNA molecule is cut by particular restriction enzymes.

Macrorestriction map: Map depicting the order of and distance between sites at which restriction enzymes cleave chromosomes.

Mate pair: The sequence obtained from opposite ends of a particular clone are referred to as mate pairs. Knowing that two sequences are derived from the same clone allows these sequences to be linked, even if the full insert of the clone is unavailable. This is key to WGS assemblies.

Megabase (Mb): Unit of length for DNA fragments equal to 1 million nucleotides and roughly equal to 1 cM. *See also:* centimorgan

Messenger RNA (mRNA): RNA that serves as a template for protein synthesis. *See also:* genetic code, transcription, translation

Metabolomics: Type of global molecular analysis that involves identifying and quantifying the metabolome—all metabolites present in a cell at a given time.

Metadata: Data that describe specific characteristics and usage aspects (e.g., what data are about, when and how data were created, who can access the data, and available formats) of raw data generated from different analyses.

Metagenome: Genetic material recovered directly from environmental samples.

Metagenomics: Study of the collective DNA isolated directly from a community of organisms living in a particular environment.

Metaomics: High-throughput, global analysis of DNA, RNA, proteins, or metabolites isolated directly from a community of organisms living in a particular environment.

Metaproteomics: High-throughput, global analysis of proteins isolated directly from a community of organisms living in a particular environment. Metaproteomics can reveal which genes are actively translated into functional proteins by a community.

Metatranscriptome: Transcriptome of a group of interacting organisms or species.

Metatranscriptomics: High-throughput, global analysis of RNA isolated directly from a community of organisms living in a particular environment. Metatranscriptomics can reveal which genes are actively expressed by a community.

Microarray: Analytical technique used to measure the mRNA abundance (gene expression) of thousands of genes in one experiment. The most common type of microarray is a glass slide onto which DNA fragments are chemically attached in an ordered pattern. As fluorescently labeled nucleic acids from a sample are applied to the microarray, they bind the immobilized DNA fragments and generate a fluorescent signal indicating the relative abundance of each nucleic acid in the sample.

Microbiome: A community of microorganisms that inhabit a particular environment. For example, a plant microbiome includes all the microorganisms that colonize a plant's surfaces and internal passages.

Microinjection: A technique for introducing a solution of DNA into a cell using a fine microcapillary pipet.

Micronuclei: Chromosome fragments that are not incorporated into the nucleus at cell division.

Modeling: Use of statistical and computational techniques to create working computer-based models of biological phenomena that can help to formulate hypotheses for experimentation and predict outcomes of research.

Molecular biology: The study of the structure, function, and makeup of biologically important molecules.

Molecular genetics: The study of macromolecules important in biological inheritance.

Molecular machine: Highly organized assembly of proteins and other molecules that work together as a functional unit to carry out operational, structural, and regulatory activities in the cells.

Multiplexing: A laboratory approach that performs multiple sets of reactions in parallel (simultaneously); greatly increasing speed and throughput.

Mutation: Any heritable change in DNA sequence. A sequence variation that deviates from the reference, sequence. This variation can be a SNP, an insertion of sequence, or a deletion of sequence. There can be a great deal of sequence variation between individuals in a population. For example, different humans may have as many as 1 base pair difference every 1000 bp. In practice, mutations are distinguished from variation because they have phenotypic consequences. Mutations in the Pax6 gene that lead to a loss of the function of that gene lead to the ***eyeless*** mutation in flies, the *Small eye* mutation in mice, and aniridia in humans.

N50: The contig/scaffold length at which have of the bases in a given assembly reside. This provides a measure of continuity. For instance, a scaffold N50 of 15 Mb means that at least half of the bases in the assembly are in a contig that is at least 15 Mb.

Nitrogenous base: A nitrogen-containing molecule having the chemical properties of a base. DNA contains the nitrogenous bases adenine (A), guanine (G), cytosine (C), and thymine (T). *See also:* DNA

Northern blot: A gel-based laboratory procedure that locates mRNA sequences on a gel that are complementary to a piece of DNA used as a probe. *See also:* DNA, library

Nuclear magnetic resonance (NMR): Technique used to study molecular structure by analyzing the absorption of electromagnetic resonance at a specific frequency in atoms subjected to strong magnetic field.

Nucleic acid: A large molecule composed of nucleotide subunits. *See also:* DNA

Nucleolar organizing region: A part of the chromosome containing rRNA genes.

Nucleotide: A subunit of DNA or RNA consisting of a nitrogenous base (adenine, guanine, thymine, or cytosine in DNA; adenine, guanine, uracil, or cytosine in RNA), a phosphate molecule, and a sugar molecule (deoxyribose in DNA and ribose in RNA). Thousands of nucleotides are linked to form a DNA or RNA molecule. *See also:* DNA, base pair, RNA

Oligo: *See:* oligonucleotide

Oligogenic: A phenotypic trait produced by two or more genes working together.

Oligonucleotide: Short nucleic acid polymer, typically with 50 or fewer bases. *See also:* nucleotide

Omics: Collective term for a range of new high-throughput biological research methods (e.g., transcriptomics, proteomics, and metabolomics) that systematically investigate entire networks of genes, proteins, and metabolites within cells.

Open reading frame (ORF): The sequence of DNA or RNA located between the start-code sequence (initiation codon) and the stop-code sequence (termination codon).

Operon: A set of genes transcribed under the control of an operator gene.

Optical Mapping: A light microscope based technique in which images of single DNA molecules undergoing restriction enzyme digest are recorded and used for the construction of physical maps of large pieces of

DNA. Optical maps often serve as scaffolds for the precise alignments of sequence contigs that are generated during genome sequencing projects.

Ortholog: Similar gene or gene segments appearing in the genomes of different species but resulting from speciation and mutation.

P1-derived artificial chromosome (PAC): One type of vector used to clone DNA fragments (100- to 300-kb insert size; average, 150 kb) in *Escherichia coli* cells. Based on bacteriophage (a virus) P1 genome. *See also:* cloning vector

Paired End Mapping (PEM): A method for detecting genome-level variation. Paired ends from size-selected sheared genomic DNA fragments are subjected to high-throughput sequencing and then mapped onto a reference genome (in silico). Analysis of paired end spans can reveal regions containing translocations, inversions, deletions/insertions and other complex structures. See also End Sequence Profiling.

Pedigree: A family tree diagram that shows how a particular genetic trait has been inherited. *See also:* inherit

Peptide: Two or more amino acids joined by a bond called a "peptide bond." *See also:* polypeptide

Phage: A virus for which the natural host is a bacterial cell.

Phenocopy: A trait not caused by inheritance of a gene but appears to be identical to a genetic trait.

Phenology: Study of recurring biological phenomena.

Phenomics: Collective study of multiple phenotypes (e.g., all phenotypes associated with a particular biological function).

Phenotype: An observable characteristic displayed by an organism. These characteristics can be controlled by genes, by the environment, or a combination of both. The characteristic can be directly observable, such as having brown eyes. In some cases, the phenotype will be measurable, such as having high blood pressure.

Phylogenetics: Study of evolutionary relationships among groups of organisms (e.g., species, populations), based on their DNA sequences.

Phylogenomics: Comparison and analysis of entire genomes, or large portions of genomes, to determine the relationship of the function of genes to their evolution.

Physical map: A map of the locations of identifiable landmarks on DNA (e.g., restriction-enzyme cutting sites, genes), regardless of inheritance. Distance is measured in base pairs.

Plasmid: Autonomously replicating extra-chromosomal circular DNA molecules, distinct from the normal bacterial genome and nonessential for cell survival under nonselective conditions. Some plasmids are capable of integrating into the host genome. A number of artificially constructed plasmids are used as cloning vectors.

Pleiotropy: One gene that causes many different physical traits such as multiple disease symptoms.

Polymerase chain reaction (PCR): A method for amplifying a DNA base sequence using a heat-stable polymerase and two 20-base primers, one complementary to the (+) strand at one end of the sequence to be amplified and one complementary to the (-) strand at the other end. Because the newly synthesized DNA strands can subsequently serve as additional templates for the same primer sequences, successive rounds of primer annealing, strand elongation, and dissociation produce rapid and highly specific amplification of the desired sequence. PCR also can be used to detect the existence of the defined sequence in a DNA sample.

Polymerase, DNA or RNA: Enzyme that catalyzes the synthesis of nucleic acids on preexisting nucleic acid templates, assembling RNA from ribonucleotides or DNA from deoxyribonucleotides.

Polypeptide: A protein or part of a protein made of a chain of amino acids joined by a peptide bond.

Population genetics: Study of allele frequency distribution and change under the influence of the four main evolutionary processes: natural selection, genetic drift, mutation, and gene flow. Population genetics also encompasses the factors of recombination, population subdivision, and population structure and attempts to explain such phenomena as adaptation and speciation.

Positional cloning: A technique used to identify genes, usually those that are associated with diseases, based on their location on a chromosome.

Premature chromosome condensation (PCC): A method of studying chromosomes in the interphase stage of the cell cycle.

Primer: Short preexisting polynucleotide chain to which new deoxyribonucleotides can be added by DNA polymerase.

Probe: Single-stranded DNA or RNA molecules of specific base sequence, labeled either radioactively or immunologically, that are used to detect the complementary base sequence by hybridization.

Promoter: A DNA site to which RNA polymerase will bind and initiate transcription.

Protein complex: Aggregate structure consisting of multiple protein molecules.

Protein expression: Subcomponent of gene expression. It consists of the stages after DNA has been transcribed to mRNA. The mRNA is then translated into polypeptide chains, which are ultimately folded into proteins.

Protein: Large molecule composed of one or more chains of amino acids in a specific order; the order is determined by the base sequence of nucleotides in the gene that codes for the protein. Proteins maintain distinct cell structure, function, and regulation.

Proteome: Collection of proteins expressed by a cell at a particular time and under specific conditions.

Proteomics: Large-scale analysis of the proteome to identify which proteins are expressed by an organism under certain conditions. Proteomics provides insights into protein function, modification, regulation, and interaction.

Pseudogene: A sequence of DNA similar to a gene but nonfunctional; probably the remnant of a once-functional gene that accumulated mutations.

Purine: A nitrogen-containing, double-ring, basic compound that occurs in nucleic acids. The purines in DNA and RNA are adenine and guanine. *See also:* base pair

Pyrimidine: A nitrogen-containing, single-ring, basic compound that occurs in nucleic acids. The pyrimidines in DNA are cytosine and thymine; in RNA, cytosine and uracil. *See also:* base pair

Quantitative trait locus: Stretch of DNA containing or linked to the genes that underlie a quantitative trait. *See also:* expression quantitative trait locus

Recombinant clone: Clone containing recombinant DNA molecules. *See also:* recombinant DNA technology

Recombinant DNA molecules: A combination of DNA molecules of different origin that are joined using recombinant DNA technologies.

Recombinant DNA technology: Procedure used to join together DNA segments in a cell-free system (an environment outside a cell or organism). Under appropriate conditions, a recombinant DNA molecule can enter a cell and replicate there, either autonomously or after it has become integrated into a cellular chromosome.

Recombination: The process by which progeny derive a combination of genes different from that of either parent. In higher organisms, this can occur by crossing over. *See also:* crossing over, mutation

RefSeq: Reference Sequence. The goal of the RefSeq project is to produce a reference sequence for all naturally occurring molecules from the central dogma (DNA, RNA, and Protein).

Regulatory elements: Segments of the genome (e.g., regulatory regions, genes that encode regulatory proteins, or small RNAs) involved in controlling gene expression.

Regulatory region or sequence: Segment of DNA sequence to which a regulatory protein binds to control expression of a gene or group of genes that are expressed together.

Resolution: Degree of molecular detail on a physical map of DNA, ranging from low to high.

Restriction enzyme, endonuclease: A protein that recognizes specific, short nucleotide sequences and cuts DNA at those sites. Bacteria contain over 400 such enzymes that recognize and cut more than 100 different DNA sequences. *See also:* restriction enzyme cutting site

Restriction fragment length polymorphism (RFLP): Variation between individuals in DNA fragment sizes cut by specific restriction enzymes; polymorphic sequences that result in RFLPs are used as markers on both physical maps and genetic linkage maps. RFLPs usually are caused by mutation at a cutting site. *See also:* marker, polymorphism

Restriction-enzyme cutting site: A specific nucleotide sequence of DNA at which a particular restriction enzyme cuts the DNA. Some sites occur frequently in DNA (e.g., every several hundred base pairs); others much less frequently (rare-cutter; e.g., every 10,000 base pairs).

Retroviral infection: The presence of retroviral vectors, such as some viruses, which use their recombinant DNA to insert their genetic material into the chromosomes of the host's cells. The virus is then propagated by the host cell.

Reverse transcriptase: An enzyme used by retroviruses to form a complementary DNA sequence (cDNA) from their RNA. The resulting DNA is then inserted into the chromosome of the host cell.

RFLP: Restriction fragment length polymorphism. A type of polymorphism detectable in a genome by the size differences in DNA fragments generated by restriction enzyme analysis.

RH mapping: Radiation Hybrid mapping. A physical mapping method that estimates linkage and distance relative to radiation-induced chromosome breaks. This is analogous to genetic mapping.

RNA (ribonucleic acid): Molecule that plays an important role in protein synthesis and other chemical activities of the cell. RNA's structure is similar to that of DNA. Classes of RNA molecules include messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA), and other small RNAs, each serving a different purpose.

RNA-Seq: Use of high-throughput sequencing technologies to sequence complementary DNA (cDNA) and obtain information about a sample's RNA content.

SAGE (Serial Analysis of Gene Expression): A technique for identifying and quantifying transcripts from eukaryotic genomes. This method is based on the isolation and concatenation of short sequence tags (~14 bp) from individual mRNAs into longer DNA molecules that are subsequently sequenced. A tag's gene origin is determined via mapping of the tag to a reference genome. See also CAGE (Cap Analysis Gene Expression).

Sanger sequencing: A widely used method of determining the order of bases in DNA. *See also:* sequencing, shotgun sequencing

Satellite: A chromosomal segment that branches off from the rest of the chromosome but is still connected by a thin filament or stalk.

Scaffold: In genomic mapping, a series of contigs that are in the right order but not necessarily connected in one continuous stretch of sequence. *See also:* supercontig

Segmental Duplication: A region of genomic DNA ranging from 1 to 400kb that may be found at more than one site in the genome. Segmental duplications often share >90% sequence identity. See also Copy Number Variation (CNV).

Sequence assembly: A process whereby the order of multiple sequenced DNA fragments is determined.

Sequence: *See:* base sequence

Sequencing technology: The instrumentation and procedures used to determine the order of nucleotides in DNA.

Sequencing: Determination of the order of nucleotides (base sequences) in a DNA or RNA molecule or the order of amino acids in a protein.

Shotgun method: Sequencing method that involves randomly sequenced cloned pieces of the genome, with no foreknowledge of where the piece originally came from. This can be contrasted with "directed" strategies, in which pieces of DNA from known chromosomal locations are sequenced. *See also:* library, genomic library

Simulation: Combination of multiple models into a meaningful representation of a whole system that can be used to predict how the system will behave under various conditions. Simulations can be used to run *in silico* experiments to gain first insights, form hypotheses, and predict outcomes before conducting more expensive physical experiments.

Single nucleotide polymorphism (SNP): DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. *See also:* mutation, polymorphism,

SNP: Single Nucleotide Polymorphism. A single base difference found when comparing the same DNA sequence from two different individuals.

Southern blotting: Transfer by absorption of DNA fragments separated in electrophoretic gels to membrane filters for detection of specific base sequences by radio-labeled complementary probes.

SSAHA: A hashing algorithm developed for rapid searching of large amounts of genome sequence. This program is similar to BLAT but does not use splice information to align mRNA sequences, nor can it perform translated searches.

SSLP: Simple sequence length polymorphisms. Common examples of these in mammalian genomes include runs of dinucleotide or trinucleotide repeats (CACACACACACACACACA).

Structural annotation: Process of identifying gene elements such as coding regions, gene structure, regulatory motifs, and open reading frames (ORFs). *See also:* genome annotation

Structural genomics: The effort to determine the 3D structures of large numbers of proteins using both experimental techniques and computer simulation

STS: Sequence Tag Site. In general, short sequences (200-500 bp) are produced throughout a genome. Oligonucleotide primers are generated such that this sequence can be amplified using PCR to produce a discrete band when analyzed by electrophoresis. STS markers can be polymorphic or monomorphic. They are critical to integrating non-sequence based maps (such as genetic or RH) with sequence based maps.

Supercontig (Scaffold): A supercontig is formed when an association can be made between two contigs that have no sequence overlap. This commonly occurs using information obtained from paired plasmid ends. For example, both ends of a BAC clone are sequenced. It can be inferred that these two sequences are approximately 150-200 Kb apart (based on the average size of a BAC). If the sequence from one end is found in a particular sequence contig, and the sequence from the other end is found in a different sequence contig, the two sequence contigs are said to be linked. In general, it is useful to have end sequences from more than one clone to provide evidence for linkage.

Synchrotron: Research facility that accelerates charged particles and uses an increasing magnetic field to keep the particles in a circular path. Electromagnetic radiation emitted by the high- energy, accelerated particles can be used in a variety of scientific applications.

Syngeneic: Genetically identical members of the same species.

Synteny: Genes occurring in the same order on chromosomes of different species. *See also:* linkage, conserved sequence

Synthetic biology: Field of biological research and technology that combines science and engineering with the goal of designing and constructing new biological functions and systems not found in nature. Essential synthetic biology tools include DNA sequencing, fabrication of genes, modeling how synthetic genes behave, and precisely measuring gene behavior.

Tandem repeat sequences: Multiple copies of the same base sequence on a chromosome; used as markers in physical mapping. *See also:* physical map

Targeted mutagenesis: Deliberate change in the genetic structure directed at a specific site on the chromosome. Used in research to determine the targeted region's function. *See also:* mutation, polymorphism

Tiling (Targeting Induced Local Lesions in Genomes): A reverse genetics technique that permits the directed identification of mutations in genes of interest. A chemical mutagen is used to induce lesions in an individual's gametes, from which the DNA is recovered and subsequently analyzed by gene specific PCR and enzyme digest to identify genes with mutations.

TPF: Tiling Path File. This is a simple file that simply lists the order of clones along a chromosome. These files are often used in genome assemblies in an effort to convey mapping information to the assembly program.

Transcript: RNA molecule (mRNA) generated from a gene's DNA sequence during transcription.

Transcription factor: Protein that binds to regulatory regions in the genome and helps control gene expression.

Transcription: Synthesis of an RNA copy of a gene's DNA sequence; the first step in gene expression. *See also:* translation

Transcriptome: Set of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA produced in one or a population of cells.

Transcriptomics: Global analysis of expression levels of all RNA transcripts present in a cell at a given time.

Transfection: The introduction of foreign DNA into a host cell. *See also:* cloning vector

Transfer RNA (tRNA): RNA that transports amino acids to ribosomes for incorporation into a polypeptide undergoing synthesis.

Transformation: A process by which the genetic material carried by an individual cell is altered by incorporation of exogenous DNA into its genome.

Transgenesis: The introduction of exogenous DNA into a cell. Typically, this term refers to the introduction of a gene into an embryo or other eukaryotic cell. In general, this DNA will insert into the genome at random, although specific loci can be targeted. The size of the DNA molecule introduce can be small (a few basepairs) to quite large (over 100 Kb).

Transgenic: An experimentally produced organism in which DNA has been artificially introduced and incorporated into the organism's germ line. *See also:* cell, DNA, gene, nucleus

Translation: The process in which the genetic code carried by mRNA directs the synthesis of proteins from amino acids. *See also:* transcription

Transposable element: A class of DNA sequences that can move from one chromosomal site to another.

Western blot: A technique used to identify and locate proteins based on their ability to bind to specific antibodies. *See also:* Northern blot, protein, Southern blotting

Whole Genome Assembly Comparison (WGAC): A computational method that detects identity between long stretches of genomic sequence, revealing regions of segmental duplication. A method complementary to Whole Genome Shotgun Sequence Detection (WSSD).

Whole Genome Shotgun Sequence Detection (WSSD): A computational method for the comparison of whole genome shotgun sequence (WGS) to a reference genome, commonly used for the detection of segmental duplications. A method complementary to Whole Genome Alignment Comparison (WGAC).

Whole Genome Shotgun Sequencing (WGS): Whole Genome Shotgun. A sequencing method by which an entire genome is cut into chunks of discrete sizes (usually 2, 10, 50 and 150 Kb) and cloned into an appropriate vector. The ends of these clones are sequenced. The two ends from the same clone are referred to as mate pairs. The distance between two mate pairs can be inferred if the library size is known and should have a narrow window of deviation.

Wild type: The form of an organism that occurs most frequently in nature.

YAC: Yeast artificial chromosome. These cloning vectors were developed using yeast centromere and telomere sequences. The average insert size of these clones ranges from 100-1000 Kb. These clones can span large portions of the genome but can be highly unstable.

Zinc-finger protein: A secondary feature of some proteins containing a zinc atom; a DNA-binding protein.

2. Major milestones in genomic research - Historical timeline

M. Ashajyothi and A. Kumar
Division of Plant Pathology, ICAR-Indian Agricultural Research Institute
New Delhi-110012

Introduction

Genomics is an interdisciplinary and interesting field of science which deals with study of structure, function, mapping and evolution of genomes. Genomics also involves the sequencing and analysis of genomes. Advances in genomics have triggered a revolution in discovery-based research to understand even the most currently complex biological systems. Availability of whole genome sequences of various organisms facilitated to conduct research in various aspects of biology to solve the hidden complexity of individual genomes.

Etymology

The term "*genome"* is attributed to Hans Winkler from early 1920s whereas the term "*genomics"* was coined by Thomas Tom Roderick, a geneticist at the Jackson Laboratory, in a meeting held in Maryland on the mapping of the human genome in 1986 (Yadav, 2007)

The area of genomics research mainly consists of:

Functional genomics

Functional genomics answers the questions about function of DNA at the gene level, RNA transcripts and protein products and gene expression by using high throughput methods rather than by utilizing traditional gene for gene approach. The two most important tools to study functional genomics are microarrays and Bioinformatics

Structural genomics

Structural genomics is useful to describe the 3-dimensional structure of every protein encoded by a given genome. It also helps in construction of high resolution genetic and physical maps (Brenner and Levitt, 2000). Consequently, structure determination promises to be an increasingly effective and efficient means of detecting homology, and thus suggesting molecular function for proteins.

Comparative genomics

Whole or parts of genomes from various genome projects are compared to study basic biological similarities and differences as well as the evolutionary relationships between organisms (Touchman, 2010). The principle of comparative genomics is that common features of two organisms will often be encoded within the DNA that is evolutionarily conserved between them (*Hardison, 2003)*

Epigenomics

This deals with the study of complete set of epigenetic modifications. Till date the two most important epigenetic modifications found are DNA methylation and histone modification. These will play great role in gene expression and cellular processes like tumorigenesis *etc*. (Francis, 2011)

Meta genomics

It is the study of genome derived directly from environmental samples. Meta genomics facilitate to see the diversity of organisms in microscopic view. Since the discovery of genetic material, the time line of historical events of genomics research is presented in Table 1 and Table 2.

Table 1. Mile stones in genomic research

Year	**Mile stones in genomic research**
1859	Charles Darwin wrote "On the Origin of Species by Means of Natural Selection".
1865	Gregor Mendel's experiments on peas demonstrate that heredity is transmitted in discrete units.
1869	Frederick Miescher isolates DNA from cells for the first time and calls it "nuclein".
1952	Alfred Hershey and Martha Chase show that only the DNA of a virus needs to enter a bacterium to infect it, providing strong support for the idea that genes are made of DNA
1953	James D Watson and Francis H Crick solved 3-dimensional structure of DNA double helix.
1965	Robert Holley and colleagues were able to produce the first whole nucleic acid sequence, that of alanine tRNA from *Saccharomyces cerevisiae* (Holley, 1965)
1965	Fred Sanger and colleagues developed a related technique based on the detection of radio-labelled partial-digestion fragments after two-dimensional fractionation (Sanger *et al*., 1965)
1972	Walter Fiers' laboratory was able to produce the first complete protein-coding gene sequence of the coat protein of bacteriophage MS2 by using 2-D fractionation method (Heather and Chain, 2016)
1975; 1977	Two main DNA sequencing methods that contributed for 1st generation sequencing i.e. Chain termination method by Fredrick Sanger and chemical cleavage by Maxam and Gilbert (Maxam and Gilbert, 1977; Sanger and Coulson, 1975)
1977	Sanger and colleagues sequenced the first DNA genome of bacteriophage φX174 by using chain termination method (Sanger *et al*., 1977)
1985	Development of Polymerase chain reaction (PCR), Electrophoresis and recombinant DNA technologies further revolutionized genomics by providing means of generating the high concentrations of pure DNA species required for sequencing (Saiki *et al*., 1985).
1986	Dideoxy sequencers – such as the ABI PRISM range developed from Leroy Hood's research, produced by Applied Biosystems.
1987	Second generation sequencing technologies used Pyrosequencing technique, pioneered by Pal Nyren and colleagues that could be performed using natural nucleotides and observed in real time.
1995	The first organism whose entire genome sequenced was *Haemophilus influenza* (*Fleischmannet al., 1995)*.
1996	First eukaryotic organism *Saccharomyces cerevisiae* whole genome sequenced.
1998	The first *multicellular* eukaryote, and animal to have its whole genome sequenced was the nematode worm: *Caenorhabditis elegans* in 1998 (*The C. elegans sequencing consortium. 1998)*
2001	First draft of Human genome published years ahead of schedule by using short gun sequencing technology.
2008	The first high-throughput sequencing (HTS) machine 454 / GS 20 pyrosequencer introduced a greater number of reads as well as better quality data with microfabrication and high-

	resolution imaging.
2008	The first 3rd generation SMS (Single Molecule Sequencing) technology was developed in the lab of Stephen Quake later commercialized by Helicos BioSciences (Braslavsky *et al.*, 2003).
2009	Illumina has announced launching their own Personal Full Genome Sequencing Service at a depth of 30× for $48,000 per genome.
2010	Life Technologies commercialized Ion Torrent's semiconductor sequencing technology in the form of the bench top Ion PGM sequencer.
2011	Complete Genomics, Illumina, Sequenom, Helicos Biosciences, Pacific Biosciences, ION Torrent Systems, Halcyon Molecular, NABsys, IBM, and GE Global all are going head to head in the race to commercialize whole genome sequencing (Mukhopadhyay, 2009)
2012	10x Genomics microfluidic platform for direct determination of diploid genome sequences
2014	Oxford Nanopore Technologies is the first company offering nanopore sequencer platforms like GridION and MinION which is a small, pocket sized portable device for biological analysis

Table 2. Time line of High-Throughput NGS Technologies development in past 10 years (Reuter *et al.*,2015)

NGS	**NGS Platform**	**Run time**	**Read length (bp)**	**Sequence by**	**Detection by**
Roche	GS FLX Titanium XL+	23h	700	Synthesis	Pyrophosphate detection
	GS Junior System	10h	400	Synthesis	Pyrophosphate detection
Life technologies	Ion torrent	4h	200-400	Synthesis	Proton release
	Proton	4h	125	Synthesis	Proton release
	ABI/SoLiD	10days	75+35	Ligation	Fluorescence detection
Illumina/Solexa	HiSeq 2000/2500	12days	2x100	Synthesis	Fluorescence detection
	MiSeq	65h	2x300	Synthesis	Fluorescence detection
Pacific Biosciences	RS II	2days	>10kb	Single molecule synthesis	Fluorescence detection
Helicos	Helicose	10days	~30	Single molecule synthesis	Fluorescence detection
Oxford Nanopore technologies	MinION	1min – 48h	200kb	Strand sequencing	Fluctuation in ionic current
	GridION	1min – 48h	200kb	Strand sequencing	Fluctuation in ionic current

Selected references

1. Braslavsky, B., Hebert, E., Kartalov, S.R. et al. 2003. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. U. S. A.* 100: 3960–3964.

2. Brenner, S, E. and Levitt, M. 2000. Expectations from structural genomics. *Protein Sci.* 9:197–200.
3. *Fleischmann, R., Adams, M., White, O., et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 269 (5223): 496–512.*
4. Francis, R, C. 2011. Epigenetics: the ultimate mystery of inheritance. *New York: W.W. Norton.* ISBN 978-0-393- 07005-7.
5. *Hardison, R.C. 2003. Comparative genomics.PLoS Biology.1 (2): 58.*
6. Heather, J.M. and Chain, B. 2016. The sequence of sequencers: The history of sequencing DNA. *Genomics.* 107: 1–8.
7. Holley, RW. 1965. Structure of a ribonucleic acid. *Science.* 147: 1462–1465.
8. Maxam, A.M. and Gilbert, W. 1977. A new method for sequencing DNA. *Proc. Natl. Acad. Sci.U.S.A*. 74.
9. Mukhopadhyay, R. 2009. "DNA sequencers: the next generation". *Anal. Chem.* 81 (5): 1736–40.
10. Reuter, J.A., Spacek, D.V. and Snyder, M.P. 2015. High-Throughput Sequencing Technologies. *Mol. Cell.*58(4):586-97
11. Saiki, R.K., Scharf, S.F., Faloona, Mullis, K.B. and Horn, G.T.1985. Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia, *Science.* 230: 1350–1354.
12. Sanger, F. and Coulson, A. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* 94: 441–448.
13. Sanger, F., Air, G, M., Barrell, B.G.N., et al.1977. Nucleotide sequence of bacteriophage phi X174 DNA, *Nature.* 265: 687–69.
14. Sanger, F., Brownlee, G. and Barrell, B. 1965. A two-dimensional fractionation procedure for radioactive nucleotides. *J. Mol. Biol.* 13: 373.
15. *The C. elegans sequencing consortium. 1998. Genome Sequence of the Nematode C. elegans: A platform for investigating biology. Science. 282 (5396): 2012–2018.*
16. Touchman, J. 2010. Comparative Genomics. *Nature Education Knowl.* 3(10): 13.
17. Yadav, S, P. 2007. The Wholeness in Suffix -omics, -omes, and the Word Om. *J. of Biomol. Techn.* 18: 277.

Do you know?

Frederick Sanger was a British biochemist who twice won the Nobel Prize in Chemistry, one of only two people to have done so in the same category, the fourth person overall with two Nobel Prizes, and the third person overall with two Nobel Prizes in the sciences. In 1958, he was awarded a Nobel Prize in Chemistry "for his work on the structure of proteins, especially that of insulin". In 1980, Walter Gilbert and Sanger shared half of the chemistry prize "for their contributions concerning the determination of base sequences in nucleic acids". The other half was awarded to Paul Berg "for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant DNA".

3. Plant viral genome- Structure and function

V.K. Baranwal
Plant Virology Unit, Division of Plant Pathology
ICAR-Indian Agricultural Research Institute, New Delhi-110012

Introduction

For a long time, single stranded RNA was considered universal genetic material of all plant viruses. It is now known that the plant viral genome may consist of any of the four types of nucleic acid: single stranded RNA (ssRNA, about 75% of plant viruses), double stranded RNA (dsRNA e.g. Reoviruses), single stranded DNA (ssDNA, e.g. Geminiviruses) or double stranded DNA (dsDNA, e.g. Caulimovirus and Badnaviruses).

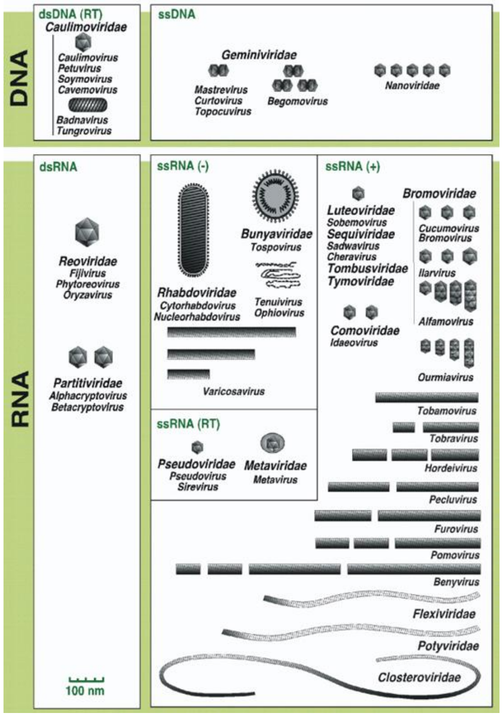


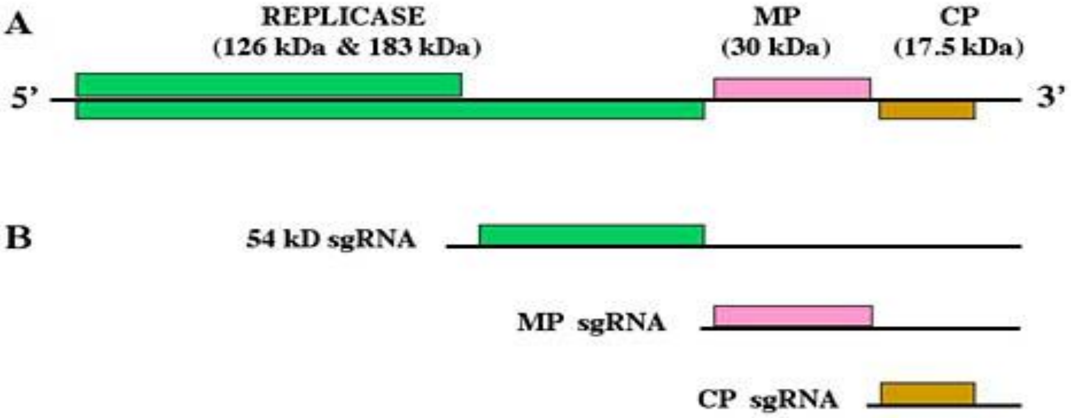
Fig 1. Families and genera of viruses infecting plants

Plant virus genomes are small (~ 4-20 kb). They make very efficient use of the limited amount of genomic nucleic acid they possess and code for only a few genes in a very compact manner. Viral genome may be circular (as in all known plant DNA viruses) or linear. The genome may occupy a single segment (e.g. in the genera *Potyvirus* and *Tobamovirus*) or may be distributed on two or more separate segments (11 in some members of the genus *Nano virus*). Individual components vary in size from about 1kb (*Nanovirus* components) to about 20 kb (genus *Closterovirus*) **(Fig 1)**.

The RNA strand of a single-stranded genome may be either a sense strand (plus strand), which can function as messenger RNA (mRNA), or an antisense strand (minus strand), which is complementary to the sense strand and cannot function as mRNA for protein translation. Sense viral RNA alone can replicate if injected into cells, since it can function as mRNA and initiate translation of virus-encoded proteins. Antisense RNA, on the other hand, has no translational function and cannot per se produce viral components. There are also some economically important viruses with antisense genomic RNA species (*tospoviruses*). This means that the sense strand and antisense strand can encode for proteins.   The eukaryotic host cell translation machinery which the viruses use to synthesize its proteins in general uses monocistronic mRNAs. So there is a problem in making more than one type of protein from a single mRNA. Hence the 3′-proximal genes of some viruses is expressed from intermediates known as subgenomic RNAs.

The number of genes coded by viral genomes vary considerably. It ranges from 1 (in satellite Tobacco necrosis virus) to 12 (in Clostero and Reoviruses). But most plant viruses have one or more genes associated with the three functions: replication of the nucleic acid, cell-to-cell movement of the virus and a structural protein that is assembled into the virus particle (usually called the coat or capsid protein). Additional genes present may have a regulatory function (NSs gene of tospoviruses which is an RNA silencing suppressor) or required for vector transmission (G1/G2 of tospoviruses for thrips transmission). Some others may act as enzymes. Proteases are coded by those viruses whose genome is transcribed as a single polyprotein mRNA. The proteases later help in cleaving the polyprotein into individual proteins (potyviruses). The genome organization of a typical member of ssRNA, *Tobacco mosaic virus* (TMV) is represented here **(Fig 2)**. TMV the type member of a large group of viruses within the genus *Tobamovirus*. The rod-shaped virus particles (virions) of TMV measure about 300 nm x 15 nm. This single-stranded +ve sense RNA encodes four genes: two replicase-associated proteins that are directly translated from the TMV RNA, and the movement protein (MP) and a coat protein (CP) that are translated from subgenomic RNAs by the host ribosomes.

In addition to coding regions for proteins, genomic nucleic acid contains nucleotide sequences with recognition and control function that are important for virus replication. These are found mainly in the 5' and 3' non-coding sequences of the ssRNA viruses, however, they may also occur internally even in coding sequences. Sequences of plant viral genomes are available on NCBI database.



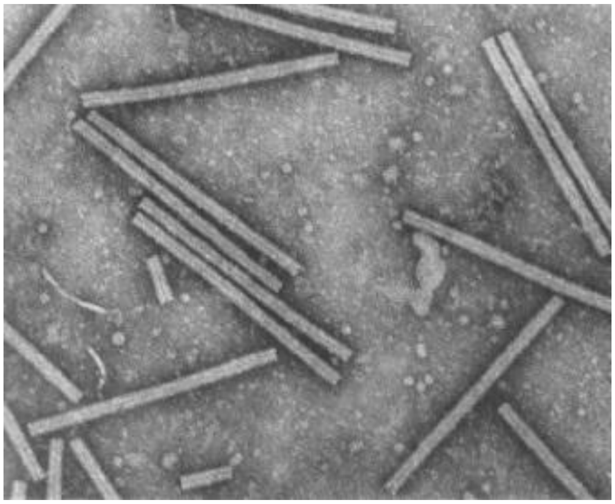


Fig 2. *Tobacco mosaic virus* (TMV) genomic and subgenomic RNAs (A) The 6400 nucleotide TMV genomic RNA acts as a messenger RNA for the expression of the 126 kDa and 183 kDa replicase-associated proteins. (B) The other genes on TMV RNA are expressed from subgenomic RNAs (sgRNAs) during the replication cycle. The rod-shaped TMV virions (C) are composed of the CP and the genomic RNA. From Scholthof, K.B.G. (2000)

Selected references

1. Scholthof, K.B.G. 2000. Tobacco mosaic virus. *The Plant Health Instructor*. DOI: 10.1094/PHI-I-2000-1010-01 Updated 2005
2. Hull, R. 2002. *Matthews' Plant Virology,* Fourth Edition, Academic Press, London.

Do you know?

Largest genomes belongs to a very small creature, *Amoeba dubia*. This protozoan genome has 670 billion units of DNA, or base pairs. The genome of a cousin, *Amoeba proteus*, has a mere 290 billion base pairs, making it 100 times larger than the human genome. Human genome is 3000.000.000 base pair, if we publish it as a book it needs 1000 books of 1000 pages!

4. Research strategies for whole genome sequencing of fungi

Arpita Ghosh
Eurofins Genomics India Pvt. Ltd, Bengaluru-560048

Introduction

A genome is an organism's complete set of DNA, including all of its genes. All living things have a unique genome. The technique that allows researchers to read and decipher the genetic information found in the DNA of any living organism is called genomic sequencing.The genomes of more than 1000 fungal species are already publically available and it's growing steadily. The fungal kingdom comprises some of the most devastating plant pathogens (Aylward *et al.,* 2017).Fungal genomics has enabled to rapidly develop tools to study pathogen biology and genetics (Moller *et al.,* 2017). In a field where delayed action has profound consequences for livelihoods and food, genome sequences provide us with essential tools to prepare for the emergence of new plant pathogens and future disease outbreaks.

Brief methodology and techniques involved

A. Isolation, Library preparation and sequencing

- The fungus is characterized for internal transcriber region using ITS-4 and ITS-5 markers to confirm the fungus genus and species.
- Once the fungus is confirmed gDNA is isolated from the fungus. This gDNA is used in library preparation as mentioned below

i. Paired End (PE) Short Gun Library preparation

- The paired-end sequencing library is prepared using Illumina Library Preparation Kit. The insert size of the PE sequencing libraries is ~300bp -550 bp.
- Each Sequencing library will be individually indexed/barcoded for sequencing

ii. Mate Pair (MP) Library preparation

- Preparation of Mate Pair library with jumping distance of 3KB & 8KB average insert size.
- Each Sequencing library will be individually indexed/barcoded for sequencing

iii. PacBio Sequel Library preparation

- Preparation of SMRT bell libraries using C4 chemistry (DNA sequencing Reagent 4.0) and 1 × 240 minute movies were captured for each SMRT cell using the PacBio Sequel (Pacific Biosciences) sequencing platform.

B. Data generation

Data generation depends on the genome size & complexity which is to be sequenced and also whether the reference genome is available or not. Various approaches are used such as Hybrid approach of long and short reads, various libraries types are used such as MP& PE.

C. Types of genome assembly

i. de novo Genome Assembly

Genome assembly refers to the process of taking a large number of short DNA sequences and putting them back together to create a representation of the original chromosomes from which the DNA originated. The word *'de novo' means starting from the beginning. The following are the basic steps involved.*

- ***Quality Filtration***

The machine generated raw reads are filtered to obtain high quality clean reads using Quality Trimming tools to remove adapter sequences, ambiguous reads (reads with unknown nucleotides "N" larger than 5%), and low-quality sequences (reads with more than 10% quality threshold (QV) < 20 phred score(for phred scale phred33)). Few quality filtration tools are Trimmomatic, *Cutadapt.*

- ***de novo Genome Assembly***

The purpose of assembly is to process for assembling the quality trimmed reads into draft contigs (contigs are simply reads that have been assembled together). Assemblies can be produced which have less gaps, less or no mis-assemblies, less errors by tweaking the input parameters. Genome Assemblers widely used are velvet, SPAdes, Allpath-LG, MaSurCa, SOAP de novo. This step is performed to optimize the generated assemblies by combining overlapping contigs and introducing appropriate gaps. A scaffold can be defined as a portion of the genome sequence reconstructed from end-sequenced whole-genome shotgun clones. Some of the scaffolders are SSPACE, PB Jelly.

- ***Prediction genes***

The assembled scaffolds are then searched for their coding potential to get the protein coding gene sequences. The widely used tools for this purpose are Prodigal, Augustus, GeneMark and Maker-P. The predicted genes areannotated based on sequence homology to known proteins by performing a BlastX against protein databases such as non-redundant protein database at NCBI, UniprotKB, etc. The GO (Gene Ontology) mapping provides ontology of defined terms representing gene product properties which are grouped into three main domains: Biological process, Molecular function and Cellular component.

ii. Reference based Genome Analysis

This method is executed, if there is a well assembled reference genome available for analysis. These high quality reads are first aligned to the reference genome. The genomes have to be indexed as per the aligner tool used for mapping the reads. The alignment tool needs to be specified if the reads are paired end or single end. Most popular aligning tools are bwa, Bowtie, TopHat

Selected references

1. Kumar, A., Sheoran, N., Prakash, G., Ghosh, A., Chikara, S.K., Rajashekara, H., Singh, U.D., Aggarwal, R. and Jain, R.K. 2017. Genome Sequence of a Unique *Magnaporthe oryzae* RMg-Dl Isolate from India That Causes Blast Disease in Diverse Cereal Crops, Obtained Using PacBio Single-Molecule and Illumina HiSeq2500 Sequencing. *Genome announcements*, *5*(7), pp.e01570-16.

2. Kumar, A., Pandey, V., Singh, M., Pandey, D., Saharan, M.S. and Marla, S.S. 2017. Draft genome sequence of Karnal bunt pathogen (*Tilletia indica*) of wheat provides insights into the pathogenic mechanisms of quarantined fungus. *PloS one*, *12*(2), p.e0171323.
3. Aylward, J., Steenkamp, E.T., Dreyer, L.L., Roets, F., Wingfield, B.D. and Wingfield, M.J. 2017. A plant pathology perspective of fungal genome sequencing, *IMA Fungus* 8(1):1-15
4. Hittalmani, S., Mahesh, H.B., Mahadevaiah, C. and Prasannakumar, M.K. 2016. De novo genome assembly and annotation of rice sheath rot fungus *Sarocladium oryzae* reveals genes involved in Helvolic acid and Cerulenin biosynthesis pathways. *BMC genomics*, *17*(1): 271.
5. Moller, M. and Stukenbrock, E.H. 2017. Evolution and genome architecture in fungal plant pathogens. *Nature Reviews Microbiology*15**:** 756–771

Do you know?

Smallest genome identified is from a Viroid family. Viroids are the smallest known pathogenic agents, and one of the smallest belongs to the Grapevine yellow speckle viroid with 220 nucleotide.

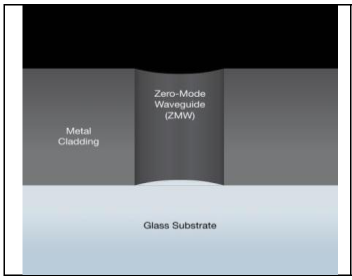
5. SMRTsequencing technology: ARevolutionary sequencing technology in 21st century

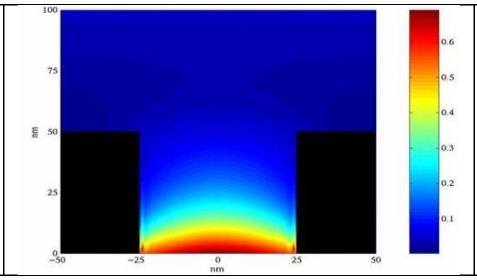
Rakshit Chaudhary
SpincoBiotech Pvt. Ltd., Chennai

Overview: Long Reads from Single Molecules

The PacBio RS II and Sequel Systems are sequencing instruments with single-molecule resolution. PacBio's Single Molecule, Real-Time (SMRT) sequencing technology provides simultaneous and uninterrupted observation of the process of natural DNA synthesis by thousands of individual DNA polymerase molecules, each working in a continuous, processive manner.

SMRT Sequencing is performed on SMRT Cells, each containing 150,000 Zero-Mode Waveguides (ZMWs). The ZMWs are tiny holes in a 100 nm metal film deposited on a silicon dioxide substrate **(Fig. 1)**. Each ZMW is illuminated from below, providing a window for observing a confined reaction volume in real time **(Fig. 2)**.





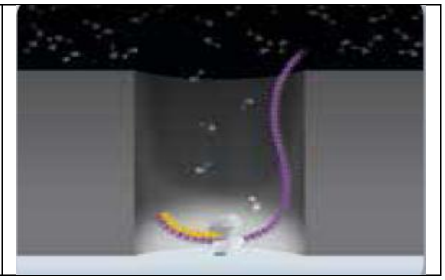


Fig. 1. Individual ZMW Fig. 2. Detection Volume   Fig. 3. Immobilized Polymerase

A DNA polymerase-template complex is immobilized via streptavidin binding at the bottom surface of the ZMW to allow optimal detection of a **sequencing-by-synthesis** reaction in real time regardless of template size **(Fig. 3)**.

To detect incorporation events and determine base identity, SMRT Sequencing uses nucleotide analogs that have fluorescent dyes linked to the polyphosphate **(Fig. 4)**. Each of the four nucleotides, A, C, G and T, is labeled with a different fluorescent dye to allow discrimination during incorporation. The fluorophores are attached to the polyphosphate, rather than to the nucleobase or sugar.

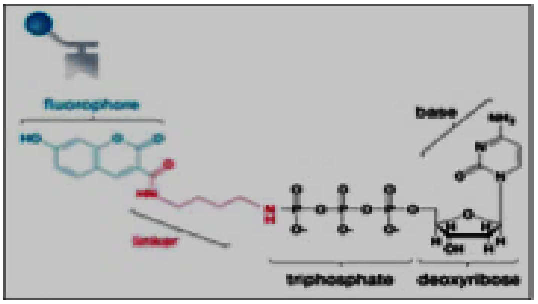


Fig. 4. Fluorophores Attached to the Polyphosphate Create Phospholinked Nucleotide Analogues

With an active polymerase-template complex immobilized at the bottom of a ZMW, phospholinked nucleotide analogues are introduced onto the SMRT Cell at high concentrations. High concentrations support enzyme speed, accuracy and processivity. As the DNA polymerase incorporates complementary bases, each nucleotide is held within the detection volume for tens of milliseconds, orders of magnitude longer than unincorporated analogues, which are diffusing in and out of the detection volume. During this time, the bound fluorophore emits fluorescent light whose colour corresponds to the base identity. As part of the natural incorporation cycle, the polymerase cleaves the polyphosphate, and the fluorophore diffuses out of the polymerase active site. The nucleobase is incorporated, elongating a growing nascent strand of native DNA. Following each base incorporation, the signal returns to baseline and the process repeats as the polymerase moves to the next complementary base **(Fig. 5)**.

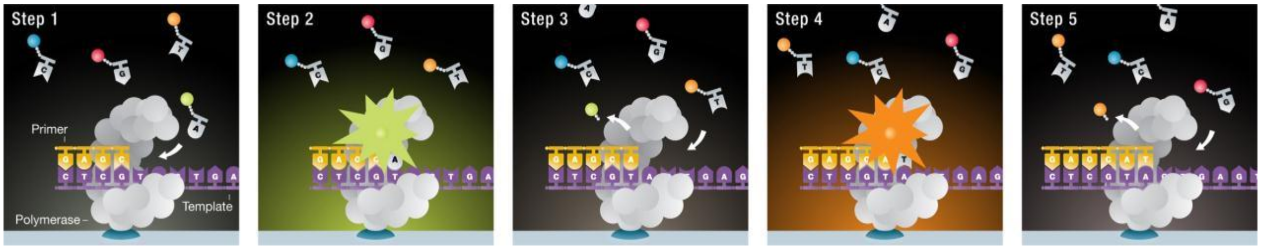
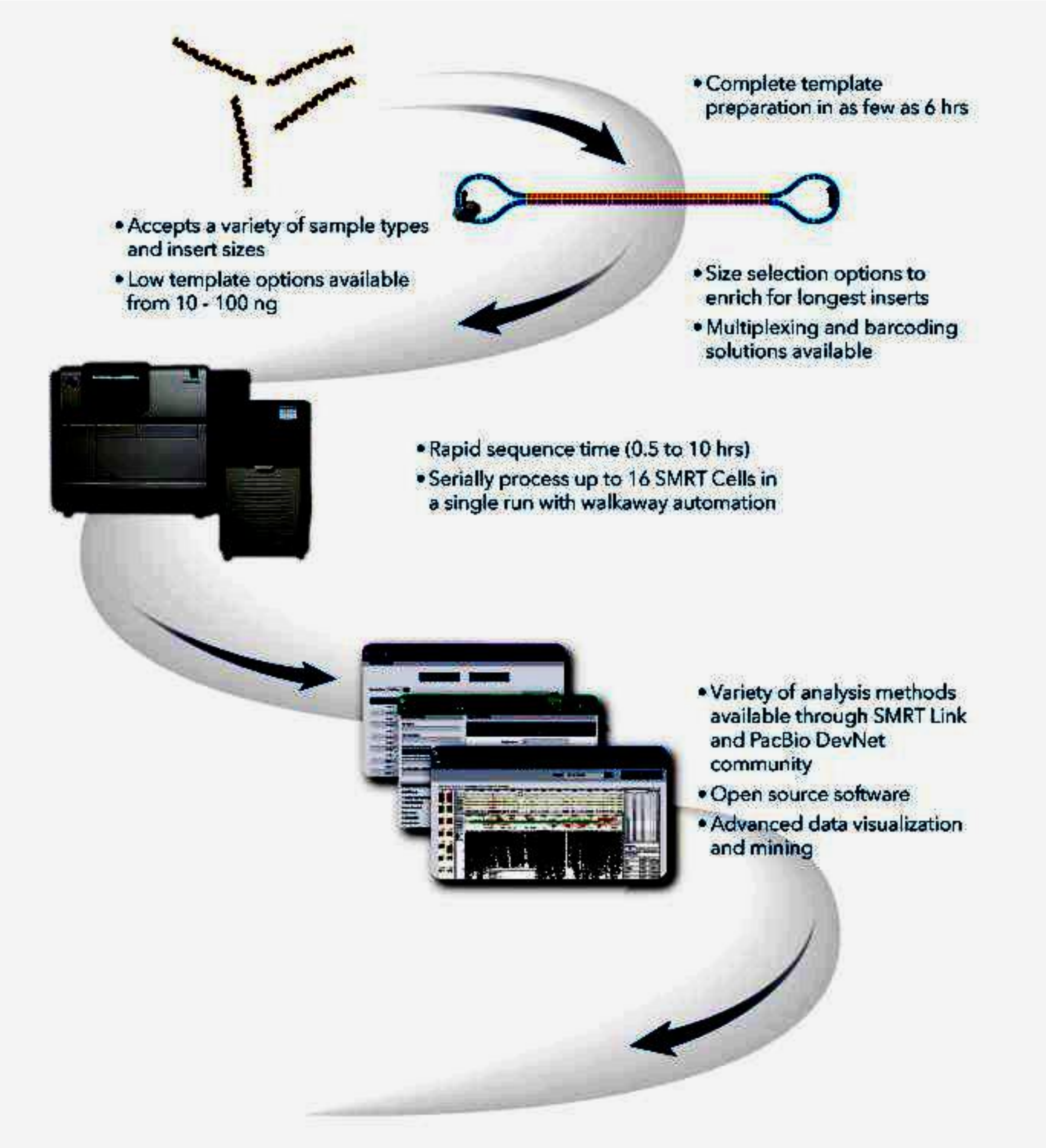


Fig. 5. Processive Synthesis with Phospholinked Nucleotides

Step 1: Fluorescent, phospholinked nucleotide analogues are introduced into the ZMW.
Step 2: The base being incorporated is held in the detection volume for tens of milliseconds, producing an extended emission of light.
Step 3: The phosphate chain is cleaved, releasing the attached dye molecule.
Steps 4-5: The process repeats as the next complementary base is incorporated.

Products and Workflow

The PacBioworkflow, consumables and software provide a complete set of tools to perform cutting-edge molecular biology.


• Complete template preparation in as few as 6 hrs
• Accepts a variety of sample types and insert sizes
• Low template options available from 10 - 100 ng
• Size selection options to enrich for longest inserts
• Multiplexing and barcoding solutions available
• Rapid sequence time (0.5 to 10 hrs)
• Serially process up to 16 SMRT Cells in a single run with walkaway automation
• Variety of analysis methods available through SMRT Link and PacBio DevNet community
• Open source software
• Advanced data visualization and mining

Acknowledgement

Thanks to Division of Plant Pathology, ICAR-IARI, New Delhi for the opportunity to deliver a lecture and Spinco Biotech, the Channel Partner of Pacific Biosciences in India and Pacific Biosciences, San Francisco, CA, USA for the support.

Selected references

1. Bernardinia,F., Galizia,R., Menichellia,M.,Papathanosb, P.A., Dritsoub,V., Maroisc, E., Crisantia,A. and Windbichlera, N. 2014. Site-specific genetic engineering of the *Anopheles gambiae* Y chromosome. *Proceedings of the National Academy of Sciences.*111 (21), 7600–7605.

2. Biller, S.J., Berube, P.M., Berta-Thompson, J.W., Kelly, L., Roggensack, S.E., Awad, L., Roache-Johnson,K.H., Ding, H., Giovannoni, S.J., Rocap, G., Moore, L.R. and Chisholm, S.W.2014. Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *Scientific Data.*140034 doi:10.1038/sdata.2014.3
3. Byrd, A.L., Perez-Rogers, J.F., Manimaran, S., Castro-Nallar, E., Toma,I., McCaffrey,T., Siegel, M., Benson, G., Crandall, K.A. and Johnson, W.E. 2014. Clinical PathoScope: rapid alignment and filtration for accurate pathogen identification in clinical samples using unassembled sequencing data. *BMC Bioinformatics.* 15262.
4. Chen, Y., Frazzitta, A.E., Litvintseva, A.P., Fang, C., Mitchell, T.G., Springer, T.J., Ding, Y., Yuan, G. And Perfect, J.R. 2015. Next generation multilocus sequence typing (NGMLST) and the analytical software program MLSTEZ enable efficient, cost-effective, high-throughput, multilocus sequencing typing. *Fungal Genetics and Biology*. 7564–71.
5. Doolan, K. M. and Colby, D. W. 2015. Conformation dependent epitopes recognized by prion protein antibodies probed using mutational scanning and deep sequencing. *Journal of Molecular Biology*. 427 (2), 328–340.
6. Freedman, Z. and Zak, D.R. 2014. Atmospheric N deposition increases bacterial laccase-like multicopper oxidases: implications for organic matter decay. *Applied and Environmental Microbiology*. 80 (14), 4460–4468.
7. Freedman, Z.B. and Zak, D.R. 2015. Atmospheric N deposition alters connectance, but not functional potential among saprotrophic bacterial communities. *Molecular Ecology*. 24: 3170–3180.

Do you know?

***Ophioglossum reticulatum*, a species of fern also known as Adders-tongue, has the largest number of chromosomes with more than 1,260 (630 pairs).**

6. Sequencing of small genomes: Strategies and applications

Kishor Gaikwad
ICAR-National Research Centre on Plant Biotechnology, New Delhi-110012

Introduction

A genome is a collection of the entire cellular DNA in a manner that allows for proper packaging into chromosomes and specific expression that allows a cell to pass on its genetic information to the next generation. The genomes are present in the nucleus, chloroplast and mitochondria in plants and fungi. The timing, specificity and also the amount of expression of these genes are determined by the sequence of the gene itself. Thus it is important to know the sequence of the gene and its adjoining regions on the genome and for that matter the entire genome to understand its complexity and structure. The genomes of all the flora and fauna on this planet are highly variable in terms of size and complexity. The genomes of these organisms vary from Kilo bases in viruses to Mega Bases in the fungal genomes.

Sequencing strategies for pathogen genome

Sequencing of any such genome requires the knowledge about the organism's chromosome number and genome size. The bacterial genomes generally harbor a single circular genome whereas the fungi carry genomes in multiple chromosomes. Decoding these genomes requires a multiple experimental approach. Generally two approaches are followed to sequence any genome. These include the WGS approach and the BAC approach. The Whole Genome Shotgun (WGS) is more suitable for smaller genomes but BAC approach also can be utilized to achieve a chromosome based data. Multiple sequencing chemistries are available for sequencing this genome. A mix of new and old sequencing chemistries should provide the information coded in the genome. So, the gold standard provided by Sanger sequencing and the current chemistries like, Illumina, SMRT and Oxford nanopore can be used in combination to achieve the sequencing of such genomes.

1. Chromosomes can be separated using FACS and then they can be sequenced separately to generate proper sequence data.
2. For sequencing of bacterial genomes, the chemistries like Illumina MiSeq, SMRT and Oxford nanopore can be employed
3. For fungal genomes, a mix of Illumina (HiSeq and MiSeq) plus PacBio (SMRT) can be used
4. Some higher fungal genomes may require a development of physical maps, that can be generated through BioNano technology

Genome assembly and annotation

1. The sequence once completed will contain millions of bp of DNA, which will need to be assembled either de novo or in a reference based methods.
2. The bacterial genomes can be assembled in CLC software, whereas several freely available programs like Soap De Novo and NewBler can take care of fungal genomes.
3. Then there are scaffolding tools that will merge the contigs to generate scaffolds that match the chromosomes.
4. The downstream analysis includes gene prediction and annotation, repeat discovery, variant analysis *etc.*
5. Some of the tools used are GENEMARK, FGENESH and AUGUSTUS.

6. Once data is assembled and annotated, it needs to be submitted to the NCBI so that a number is assigned which becomes a reference for future studies**(Fig.1)**.

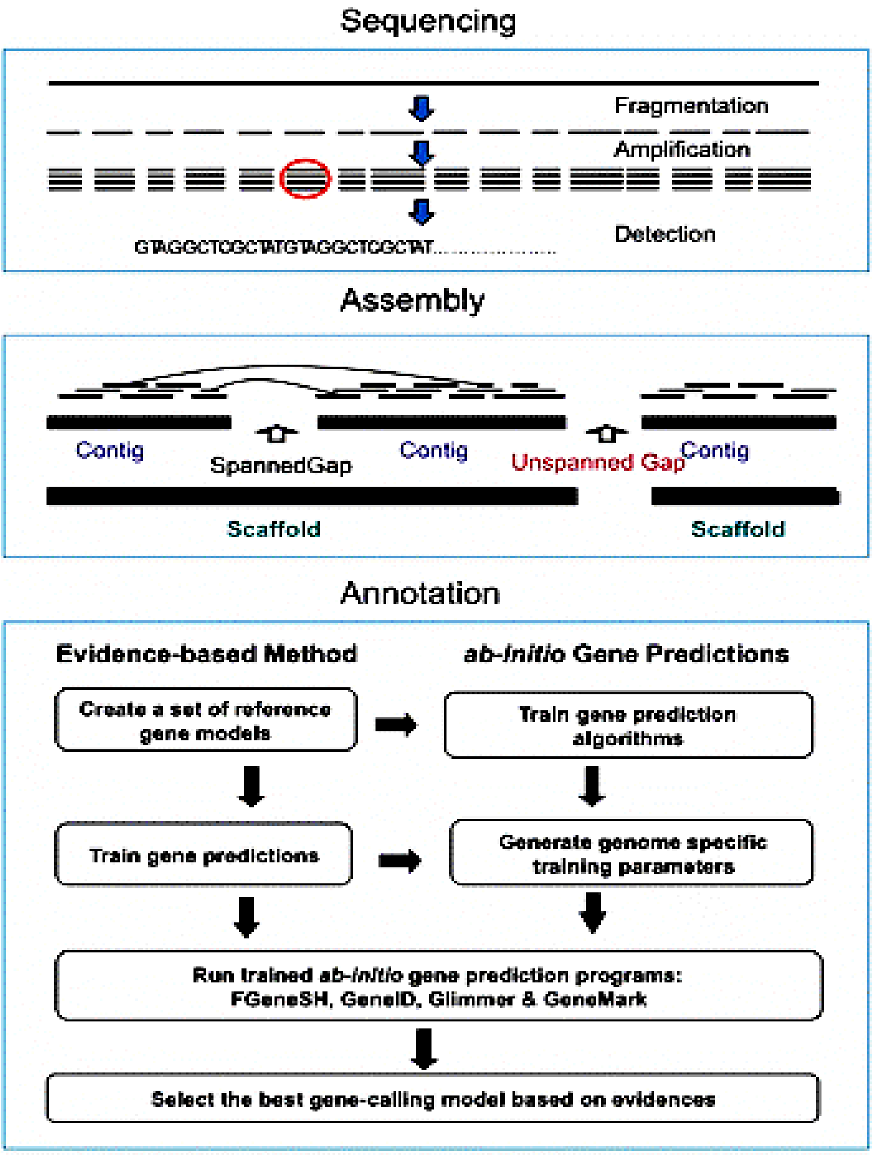

Sequencing
Fragmentation
Amplification
Detection
GTAGGCTOGCTATGTAGGCTOGCTAT......................
Assembly
Contig
SpannedGap
Contig
Unspanned Gap
Contig
Scaffold
Scaffold
Annotation
Evidence-based Method
ab-Initio Gene Predictions
Create a set of reference gene models
Train gene prediction algorithms
Train gene predictions
Generate genome specific training parameters
Run trained ab-initio gene prediction programs: FGeneSH, GeneID, Glimmer & GeneMark
Select the best gene-calling model based on evidences

Fig. 1. Experimental and bioinformatic pipeline for whole genome sequencing of plant pathogens

Conclusion

1. Small genomes are generally easier to sequence and assemble owing to smaller genome size
2. However, newer computational tools are required as these are quite different from the higher organism genomes. For example, the gene density of a fungal genome is very high as compared to plants and animals, hence improper prediction may result in errors.
3. A hybrid approach of sequencing i.e. combination of short and long reads is the best way forward to attain a properly characterized genome

Selected references

1. Adrian, T. 2014. Fungal genomics, *Briefings in Functional Genomics*, 13: 421-423.
2. McMurdie, P.J. and Holmes, S. 2013. Phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PloS one*, *8*(4), e61217.
3. Puppulin, L., Pezzotti, G., Sun, H., Hosogi, S., Nakahari, T., Inui, T., Kumamoto, Y., Tanaka, H. and Marunaka, Y. 2017. Raman micro-spectroscopy as a viable tool to monitor and estimate the ionic transport in epithelial cells. *Scientific Reports*, *7*(1), 3395.
4. Xu, J. R., Peng, Y. L., Dickman, M. B. and Sharon, A. 2006. The dawn of fungal pathogen genomics. *Annu. Rev. Phytopathol.*, *44*, 337-366.

Do you know?

As sequencing technology continues to improve, however, a new generation of effective fast turnaround benchtop sequencers has come within reach of the average academic laboratory. On the whole, genome sequencing approaches fall into two broad categories, *shotgun* and *high-throughput* (or *next-generation*) sequencing

7. Strategies for whole genome sequencing of plant viruses

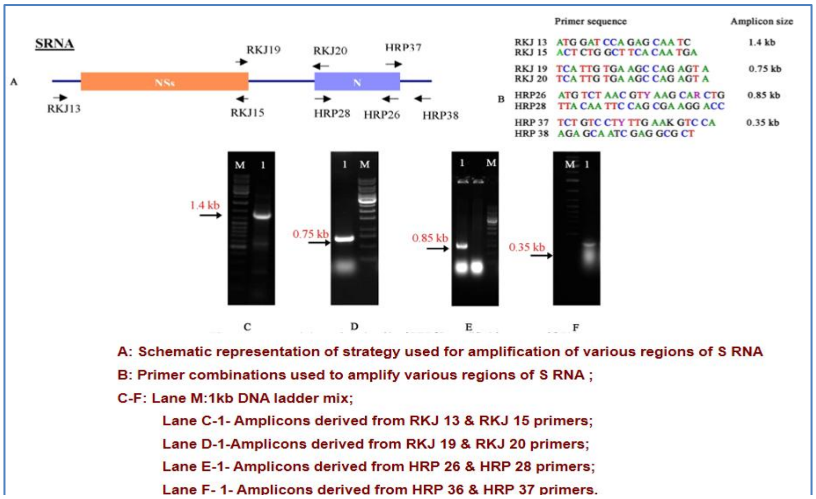
V. K. Baranwal
Plant Virology Unit, Division of Plant Pathology
ICAR-Indian Agricultural Research Institute, New Delhi-110012

Introduction

Our understanding of plant viral genomes have increased rapidly in parallel with development of molecular biological techniques. The first viral genome sequenced was that of *Cauliflower mosaic virus* (CaMV) in 1980 followed by *Tobacco mosaic virus* (TMV) in 1982.There has been an explosion of plant viral genome data in the last three decades. The data generated can be used to compare novel strains with other viruses, help to understand the genetic basis of important phenotypic characteristics, such as antigenic determinants and help answer fundamental questions related to the evolution of viruses. Overview of different WGS methods that are used in virology are summarized here with their advantages and disadvantages.

Using PCR to amplify viral genomes

This is the most widely used method for genome amplification of viruses. The genome can be amplified by PCR either as a single long fragment covering the length of the virus genome or as multiple small amplicons using overlapping primers that are complementary to a known nucleotide sequence.



	Primer sequence	Amplicon size
RKJ 13	ATG GAT CCA GAG CAA TC	1.4 kb
RKJ 15	ACT CTG GCT TCA CAA TGA	
RKJ 19	TCA TTG TGA AGC CAG AGT A	0.75 kb
RKJ 20	TCA TTG TGA AGC CAG AGT A	
HRP26	ATG TCT AAC GTY AAG CAR CTG	0.85 kb
HRP28	TTA CAA TTC CAG CGA AGG ACC	
HRP 37	TCT GTC CTY TTG AAK GTC CA	0.35 kb
HRP 38	AGA GCA ATC GAG GCG CT	

A: Schematic representation of strategy used for amplification of various regions of S RNA

B: Primer combinations used to amplify various regions of S RNA ;

C-F: Lane M:1kb DNA ladder mix;

Lane C-1- Amplicons derived from RKJ 13 & RKJ 15 primers;

Lane D-1-Amplicons derived from RKJ 19 & RKJ 20 primers;

Lane E-1- Amplicons derived from HRP 26 & HRP 28 primers;

Lane F- 1- Amplicons derived from HRP 36 & HRP 37 primers.

Schematic strategy used for amplification of SRNA of *Ground nut bud necrosis virus* by using overlapping primers.

For RNA viruses, Reverse transcription of the viral RNA to a cDNA strand (cDNA) is necessary prior to the PCR. During reverse transcription, a primer is used by reverse transcriptase (RNA-dependent DNA polymerase) to initiate the synthesis of a cDNA from the RNA. Specific sets of primers are designed for the detection of each particular virus. Primers designed to conserved regions or genes found in the viral genome hybridize with multiple members of a particular viral family. Polythymine (T_{16}) primers can be used produce cDNA from entire viral genome for those viruses that have a poly A tail. The reverse transcription step is not necessary to amplify viral genome composed of DNA.

Rolling circle amplification

Rolling-circle amplification (RCA) using the DNA polymerase of bacteriophage phi29 is used for the amplification of viruses having circular DNA genome populations. Phi29 polymerase possesses several features, such as strand displacement activity, proof-reading activity and generation of very long synthesis products, which make it suitable for the efficient amplification of circular DNA molecules from complex biological samples. A specific primer or random hexamer primer is used to initiate the reaction. The polymerase incorporates nucleotides complementary to the template strand. The synthesis is completed at the end of the template in case of a linear template molecule. In case of a circular template molecule, however, the polymerase reaches the primer binding site again, displaces the newly synthesized strand and goes on with DNA synthesis for several rounds (Rolling circle mechanism).

By this mechanism, a large DNA molecule consisting of repeated copies of the template sequence is produced, leading to a more efficient amplification of a circular target sequence as compared to a linear template. The isothermal amplification is carried out overnight at 30^0C. The RCA product appears as a high molecular weight band on agarose gel. It is then digested with a single cutting restriction enzyme that can release the unit genome length which is later cloned (Sharma *et al*., 2015).

Viral genome sequencing using next generation sequencing technologies

The cloned viral genome is sequenced most commonly by di-deoxy chain termination method devised initially by Frederic Sanger in 1970s. The advent of new deep sequencing techniques allow identification of novel viruses and enable better examination of diversity and the analysis of virus populations that contain nucleotide variants at low frequencies.

Plant viruses can be identified by directly sequencing the total RNA. But the presence of contaminating host RNA complicates the identification of viruses and/ or viroids. Different methods are used to enrich viral nucleic acids to maximize the detection of viruses in samples (i) Host ribosomal RNAs (rRNA) depletion (ii) Sequencing double-stranded RNA (dsRNA) (ii)Isolation of RNA from virus like particles (VLPs) (iii) sequencing small RNA. Among these small RNA sequencing is the most common method used for virus and viroid identification.

Virus identification by sequencing small RNA

Plants do not produce dsRNA. Hence the dsRNA replicative intermediates of viruses (both RNA and DNA) as well as viroids and satellite viruses are subject to RNA silencing by plants. This gives rise to virus/ viroid-derived small interfering RNAs (siRNAs). Even though the viral siRNAs are only 21 to 24 nt long they overlap each other extensively. The reads are then assembled into large sequence contigs which are later used for virus discovery. The details of this method are given in the next chapter. Virus discovery by deep sequencing and assembly of total small RNAs (vdSAR) is the most widely used approach for plant virus discovery at present.

Selected references

1. Cottam, E.M., Wadsworth, J., Knowles, N.J. and King, D.P. 2009. Full Sequencing of Viral Genomes: In: Caugant D. (Eds) Molecular Epidemiology of Microorganisms. *Methods in Molecular Biology*™ (Methods and Protocols), vol 551. Humana Press, Totowa, NJ.
2. Saritha, R.K. and Jain, R.K. 2007. Nucleotide sequence of S and M RNA segments of *Groundnut bud necrosis virus* isolate from *Vigna unguiculata* in India. *Arch. Virol.* 152: 1195-1200.

3. Sharma, S.K., Kumar, P.V and Baranwal, V.K. 2014. Immunodiagnosis of episomal Banana streak MY virus using polyclonal antibodies to an expressed putative coat protein.*J Virol Methods.* 207:86-94.
4. Sharma, S.K., Kumar, P.V., Geetanjali A.S., Pun, K.B. and Baranwal, V.K. 2015. Subpopulation level variation of banana streak viruses in India and common evolution of banana and sugarcane badnaviruses.*Virus Genes*, 50(3): 450-65

Do you know?

Epigenomics is the study of the complete set of epigenetic modifications on the genetic material of a cell, known as the epigenome. Epigenetic modifications are reversible modifications on a cell's DNA or histones that affect gene expression without altering the DNA sequence. Two of the most characterized epigenetic modifications are DNA methylation and histone modification. Epigenetic modifications play an important role in gene expression and regulation, and are involved in numerous cellular processes such as in differentiation/development and tumorigenesis. The study of epigenetics on a global level has been made possible only recently through the

8. Methods to decode plant viral genome and their pathogenesis

S. Chakraborty
School of Life Sciences, Jawaharlal Nehru University, New Delhi -110067

Introduction

Plant viruses are nucleoprotein molecules have either RNA or DNA as their genome covered by protective coat made of either protein or lipoprotein. The nature of genome is either circular or linear and single stranded or double stranded. Due to smaller size and limited coding capacities, viruses encode a few number of genes that are essential for their pathogenesis. In addition, viruses being obligate parasites require the assistance of host factors to establish successful infection. Therefore, efforts are required to diagnose plant viruses, isolate their genomes, identify their genetic organizations and study function of viral genes. Some of the commonly used methods used for decoding plant viral genomes and their pathogenesis are briefly discussed below

Cloning and sequencing of viral genomes

It is the pre-requisite step for identifying and understanding any virus. It involves conventional cloning procedure from infected samples using standard procedures (PCR, RCA based methods) followed by sequencing. Because of the advances made in sequencing technology especially the advent of NGS, sequencing of viral genomes have improved substantially and helped us to discover many new plant viruses (Blawid *et al*., 2015).

Once the genome information is available then comes the decoding of the information present in the genome in simple terms understanding the coding and non-coding part of the viral genome. Over the past few decades advances in the area of molecular biology and biochemistry led to the enhanced understanding of functions of genes.

Yeast two hybrid

It is an indispensible tool these days to study the interaction between two proteins. This technique is very essential in the context of virology as it is a known fact that viruses rely heavily on host factors to establish pathogenesis and for this purpose the viral proteins have got the tendency to interact with numerous host factors. Therefore technique like yeast two hybrid is essential to identify the interacting partners in the host, by doing so one can assume the function of the viral proteins and validate substantially their role during pathogenesis. It is a protein-fragment complementation assay which involves fusing the proteins (whose interaction has to be studied) to one half of Activation domain (AD) and another to DNA binding domain (BD) of yeast transcription factor**,** transcription of reporter gene occurs upon the interaction between the proteins (Kushwaha *et al*., 2017).

Yeast one hybrid

It is a variant of yeast two hybrid and it involves studying the interaction between DNA and protein.

Electron microscopy (EM)

Electron microscopy has wide range of applications in plant virology as it can provide high resolution information about cells, organelles, virion particle etc. It also can study the localization of viral proteins and their impact on the ultrastructure of organelles (Bhattacharyya *et al*., 2015). With the advances made in the area Transmission Electron Microscopy (TEM) led to a blooming area called as cryoEM. Recently CryoEM assisted in gaining insight into the structure of coat protein of *African Cassava Mosaic Virus* (ACMV). It helped in understanding of interaction of coat proteins within and between capsomeres, coat protein-DNA interaction and the position of key residues that are important for interaction with whitefly proteins (Hipp *et al*., 2017).

Confocal microscopy

Confocal microscopy largely assists in studying localization viral proteins, alteration of host proteins in presence of viral infection, studying interaction between two proteins (Bimolecular Fluorescence Complementation-BiFC) etc under *in vivo* conditions without denaturing the natural cellular milieu. Because of aforesaid applications, confocal microscopy has become an indispensible tool in viral pathogenesis. BiFC is also a protein-fragment complementation assay in which fluorescent proteins are fragmented and fused with proteins whose interaction has to be studied (Kushwaha *et al*., 2017). Similarly, Fluorescence Resonance Energy Transfer (FRET analysis) is also used for studying invitro interaction between two protein partners.

Protoplast based assays

Protoplasts have wide applications in the field of plant virology. It helps to study viral replication, localization of viral proteins, to study interaction between viral proteins and the host proteins, PTGS suppressor activity of viral proteins etc. Protoplast based assays are less time consuming at the same time it requires skill and expertise to maintain.

Suppressor assay

It is a transient gene expression assay to assess the gene silencing suppressor activity of a viral protein. This method is fast, flexible and reproducible. It involves co-infiltrating *Agrobacterium* culture into the 16c transgenic lines of *Nicotiana benthamiana* leaves harboring the Green Fluorescent Protein (GFP) along with viral proteins whose silencing suppressor activity to be tested using vacuum infiltration or syringe. 16c transgenic lines are GFP over expressing lines. Transiently expressing GFP in GFP over expressing line will trigger gene silencing resulting in reduced GFP fluorescence in the infiltrated region. The GFP fluorescence will be restored only if the co-infiltrated protein has silencing suppressor activity (Stephan *et al*., 2011).

Protein expression system

Expressing the viral ORFs by recombinant DNA technology enables one to understand the in vitro properties which in turn help us to decode the viral genome. Protein expression and purifying systems are available for this purpose. Protein expression and purification comprises of three modules: expression host, expression vectors and purification systems. Protein expression hosts include organisms such as bacteria, yeast and insect cell lines that are manipulated for the expression of heterogeneous proteins. Bacterial based expression systems are widely preferred mainly for their ease of maintenance and affordability. There are some popular protein expression vectors which are commercially available that can help in fishing out the protein of interest from other proteins present in expression hosts. These commercial vectors generate recombinant protein with tags such as pET expression system has $(His)_6$ and SUMO, pMAL has MBP (Maltose Binding Protein), pGEX has GST (Glutathione S Transferase). Third module purification system include largely chromatographic based techniques such as affinity chromatography, size exclusion chromatography, ion-exchange based chromatography etc. In spite of the constant evolution and advancement in the area of protein expression and purification, purifying viral proteins are still a major challenge.

RACE (Rapid Amplification of cDNA Ends)

It is a technique to identify or map the transcript from viral genome. Not just coding region but it is also helpful in identify and map the non-coding region. RACE has evolved into high throughput because of the advances in next generation sequencing and it is referred as RACE-seq. RACE-seq is highly time efficient, cost effective and more sensitive.

Ribosome profiling

The efficiency of this technique was demonstrated in the viruses HCMV (Human cytomegalovirus) and Karposi's sarcoma-associated Herpes virus. This technique involves classical ribosome footprinting followed by RNA-seq. Many putatively unidentified ORF identified and verified by mass spectrometry in both HCMV and KSHV. Ribosome profiling also provided information on leaky scanning, non-ATG initiation (CUG initiation) and re-initiation of HCMV and KSHV genomes. Ribosome profiling also helps us to study the differential translation during viral infection (Stern-Ginossar, 2015). This technique may unravel many unidentified ORFs in plant viruses too.

Conclusions

Viruses are among the major threat to global food security. Devising potent antiviral strategy is essential to avoid crop loss because of viral pathogenesis. Potent antiviral strategy can only be made after proper understanding of the viruses, their interaction with the respective host and the environment.

Selected references

1. Bhattacharyya, D., Gnanasekaran, P., Kumar, R.K., Kushwaha, N.K., Sharma, V.K., Yusuf, M.A. and Chakraborty, S. 2015. A geminivirus betasatellite damages the structural and functional integrity of chloroplasts leading to symptom formation and inhibition of photosynthesis. *J Exp Bot.* 66: 5881-95.
2. Blawid, R., Silva, J.M.F. and Nagata, T. 2017. Discovering and sequencing new plant viral genomes by next-generation sequencing: description of a practical pipeline. *Ann App Biol*. 170: 301-314.
3. Hipp, K., Grimm, C., Jeske, H. and Bottcher, B. 2017. Near-atomic resolution structure of plant geminivirus determined by electron cryomicroscopy. *Structure*. 25: 1303-1309.
4. Kushwaha, N.K., Bhardwaj, M. and Chakraborty, S. 2017. The replication initiator protein of geminivirus interacts with host monoubiquitination machinery and stimulates transcription of the viral genome. *PLoS Pathog.* 13: e1006587.
5. Stephan, D., Slabber, C., George, G., Ninov, V., Francis, K.P. and Burger, J.T. 2011. Visualization of plant viral suppressor silencing activity in intact leaf lamina by quantitative fluorescent imaging. *Plant Methods*. 7:25.
6. Stern-Ginossar, N. 2015. Decoding viral infection by ribosome profiling. *J Virol.* 89: 6164-6166.

9. Plant Viruses Genomics: Redefining with current molecular approaches

V K Baranwal
Division of Plant Pathology
ICAR-Indian Agricultural Research Institute, New Delhi-110012

Introduction

Viruses are small obligate intracellular parasites, comprising of a nucleic acid core surrounded by a protective protein coat. They are unique pathogens with the ability to infect all types of life forms from animals and plants to microorganisms like bacteria. Over 3700 viruses and viroids have been recognized and approximately a third of these are plant viruses (ICTV 2015). Plant viruses are major constraints in Indian agriculture as they damage crops through numerous diseases, resulting in reduced plant vigor, huge yield losses, unmarketable crops or even plant death. The majority of plant viruses are composed of single-stranded RNA (ssRNA, about 75% of plant viruses) with the virus genome in the sense orientation (ie. members of families *Bromoviridae*, *Closteroviridae*, *Luteoviridae* and *Potyviridae*). Other plant viruses possess single-stranded (ss) RNA genomes in an ambisense or antisense orientation (i.e. members of families *Bunyaviridae* and *Rhabdoviridae*), double-stranded (ds)RNA genomes (ie. members of families *Reoviridae*, *Endornaviridae*), ssDNA genomes (family *Geminiviridae*), or ds DNA genomes (family *Caulimoviridae)*. Plant virus genomes are small (~4-20 kb). They make very efficient use of the limited amount of genomic nucleic acid they possess and code for only a few genes in a very compact manner. Viral genome may be circular (as in all known plant DNA viruses) or linear. The entire genome may occupy either one nucleic acid molecule (monopartite genome, in the genera *Potyvirus* and *Tobamovirus*) or several nucleic acid segments (multipartite genome, 11 in some members of the genus *Nanovirus*).

Compared to the other plant pathogens like fungus and bacteria, which are studied since long, the study on viruses are relatively new and much more difficult to manage. Traditionally, plant virus studies in India were based on transmission, host reactions, particle morphology and serology. Studies on plant viral genome sequence redefined plant virus identification to a finest level. A large number of viral sequences are available in databases which laid foundation for designing modern molecular diagnostic tools for control of plant viruses. Some of the methods used for whole genome sequencing of plant viruses have been summarized here.

Polymerase chain reaction (PCR) for amplification of viral genomes

PCR is the most commonly used method for genome amplification of viruses. Specific primers can be designed using virus sequences available in NCBI GenBank. PCR can amplify the whole genome either as a single long fragment or as multiple small amplicons using overlapping primers that are complementary to a known nucleotide sequence. For RNA viruses, reverse transcription of the viral RNA into c-DNA is a mandatory step prior to PCR. Reverse transcriptases (RTs) use a RNA template and a short primer complementary to the 3' end of the RNA to direct the synthesis of the first strand cDNA. Specific primers can be designed for detection of a particular virus. Primers designed to conserved regions or genes found in the viral genome hybridize with multiple members of a particular viral family. Oligo $(dT)_{18}$ primers can be used to produce cDNA from entire viral genome for those viruses that have a poly A tail.

Rolling circle amplification (RCA)

RCA is particularly used for the amplification of viruses with circular DNA genomes using exo-resistant random hexamer primers or specific primers and utilizing the strand displacement activity of *Phi*29 DNA polymerase. RCA is a sequence independent amplification, carried out overnight at isothermal temperature 30°C. As random hexamers are employed in RCA, the prior sequence information of the targeted viral genomes is not required. Thus, it has the potential to amplify novel circular viral genomes. The RCA product after restriction digestion needs to be sequenced for confirmation of viral origin. Random primed RCA was employed to identify the shorter Banana streak OL virus variants causing the leaf streak disease of banana in India (Baranwal *et al*., 2014). Although, RCA is an efficient amplification technique, relatively fast and easy to perform, there are some limitations. RCA is less suitable for larger genomes as the amplification efficiency decreases with the length of the circular DNA template. The probability of strand breaks increases with the length of the DNA molecule, resulting in termination of the RCA.

RACE (Rapid Amplification of cDNA Ends)

It is a powerful PCR based technique used to identify or map the transcript from viral genome. It is an efficient approach for obtaining full-length cDNA when only partial sequences are available. It is based on the amplification of nucleic acid sequences from a messenger RNA template between a defined internal site and unknown sequences at either the 3' or the 5' -end of the mRNA. 3' RACE has the benefit of the natural poly(A) tail in mRNA as a generic priming site for PCR amplification. In this procedure, mRNAs are converted into cDNA using RT enzyme and an oligo-dT adapter primer. Specific cDNA is then directly amplified by PCR using a gene-specific primer (GSP) that anneals to a region of known exon sequences and an adapter primer that targets the poly(A) tail region. This permits the capture of unknown 3'-mRNA sequences that lie between the exon and the poly(A) tail. 5' RACE, or "anchored" PCR, facilitates the isolation and characterization of 5' ends from low-copy messages. The first step involves the synthesis of first strand cDNA using a gene-specific antisense oligonucleotide (GSP1), following which homo-polymeric tails are added to the 3' ends of the cDNA using TdT (Terminal deoxynucleotidyl transferase). The tailed cDNA is then amplified by PCR using a mixture of three primers: a nested gene-specific primer (GSP2), which anneals 3' to GSP1; and a combination of a complementary homopolymer-containing anchor primer and corresponding adapter primer which permit amplification from the homo-polymeric tail. This allows amplification of unknown sequences between the GSP2 and the 5'-end of the mRNA. RACE has evolved into high throughput because of its advances in next generation sequencing and it is referred as RACE-seq. RACE-seq is highly time efficient, cost effective and more sensitive.

Next generation sequencing for exploring unknown viruses

Next generation high-throughput sequencing technology has revolutionized detection of unknown disease associated viruses and discovery of novel viruses in an unbiased manner without antibodies or prior knowledge of the virus sequences. Entire viral genome can be sequenced from symptomatic or asymptomatic plants through different approaches in next generation sequencing of either total nucleic acids or small interfering RNAs (siRNAs). The high throughput sequencing evolved from older Sanger sequencing; from single reads to millions of reads in a random massive parallel sequencing strategy. This shifted the paradigm from "one assay to detect one virus" to one assay to uncover the "virome" of the plant. The concept of "virome" is more relevant to those crops possessing multiple infections of viruses or viroids and do not require prior information on the virus. Viral genome variability, evolution within the host and virus defence mechanism in plants can also be easily understood by massive parallel sequencing (Prabha *et al*., 2013).

Selected References

1. Baranwal, V.K., Sharma, S.K., Khurana, D. and Varma, R. 2014. Sequence analysis of shorter than genome length episomal Banana streak OL virus like sequences isolated from banana in India. Virus Genes, 48(1):120-7.
2. Prabha, K., Baranwal, V.K. and Jain, R.K. 2013. Applications of next generation high throughput sequencing technologies in characterization, discovery and molecular interaction of plant viruses. Indian J. Virol. 24:157-165.

Do you know?

The Illumina dye sequencing method is based on reversible dye-terminators and was developed in 1996 at the Geneva Biomedical Research Institute by Pascal Mayer and Laurent Farinelli.In this method, DNA molecules and primers are first attached on a slide and amplified with polymerase so that local clonal colonies, initially coined "DNA colonies", are formed. To determine the sequence, four types of reversible terminator bases (RT-bases) are added and non-incorporated nucleotides are washed away

10.Complete genome sequencing of the Plant pathogenic bacteria: Whole-genome Sequencing, Assembly, and Annotation

Aravind Ravindran and A. Kumar*
Department of Plant Pathology and Microbiology, Texas A&M University
College Station, TX, 77843-2132
[*]Division of Plant Pathology, ICAR-Indian Agricultural Research Institute
New Delhi-110012

Introduction

Complete bacterial genome sequences were first published in 1995 (two decades ago), the science of bacteria has dramatically changed. Rapid progress in next-generation DNA sequencing helps to complete the sequence of a bacterial genome from several days to few hours. The powerful combination high-throughput sequencing technology and the simultaneous development of bioinformatics tools has transformed our understanding of how bacteria function, evolve and interact with each other, with their hosts, and with their surroundings (Loman and Pallen, 2015). It is now within reach for individual research groups in the eco-evolutionary and conservation community to generate genome sequences for any organism of choice. The dramatic cost's reduction of sequencing has made to do bacterial genome sequencing more accessible and affordable to large number of labs and research groups.

Methodology and bioinformatic techniques involved:

Whole-genome sequencing:High quality of genomic DNA was extracted from bacterial culture and used for high-throughput DNA sequencing methods. First, 454 pyrosequencing (http://www.454.com/) of single and paired-end reads of genomic DNA fragments were generated on a 454 GS- FLX Titanium sequencer (454 Life Sciences, Branford, CT). Subsequently, Illumina sequencing (http://www.illumina.com/) generated single and paired-end sequence reads of genomic DNA, and was conducted using an Illumina Genome Analyzer IIx (Illumina, Hayward, CA).

Assembly: The genome sequences were assembled into contigs and scaffolds using the 454 de novo assembler Newbler 2.0. CLCbio Genomics Workbench version 4.9 (Boston, MA, USA) was used for de novo assembly of reads; all other parameters were set at default values except for similarity set at 0.9. Align Illumina contigs against the 454 scaffolds to confirm the orientations and integrity of the assembled sequences and link contigs together within the scaffold**(Fig.1)**.

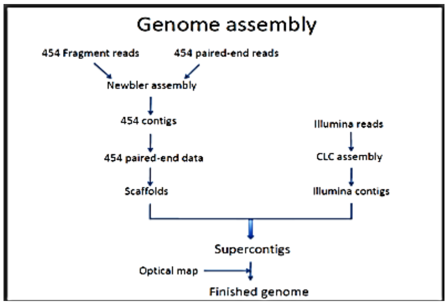
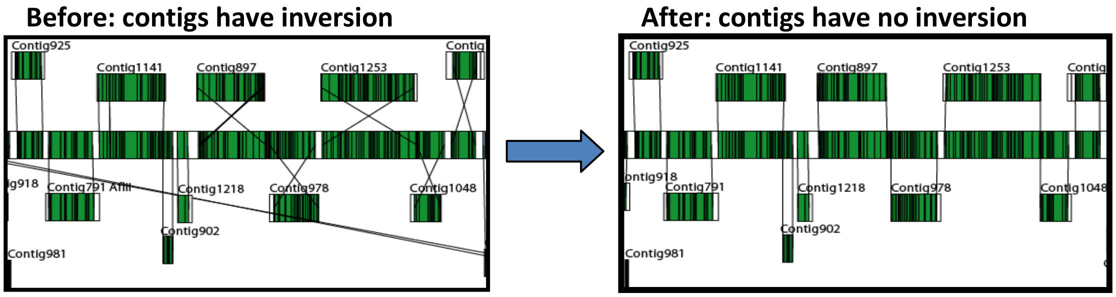


Fig. 1. Genome assembly of the bacteria using Roche 454 and illumina.

OpGen technologies (Madison, WI) generated a *de novo* assembled optical map using restriction sites of the genome. Optical mapping is a *de novo*process that generates whole genome, ordered, restriction maps with no requirement for previous/reference genome sequence information **(Fig. 2)**. *In silico* restriction maps of the scaffolds were constructed and aligned to the optical map according to their restriction fragment pattern by using MapSolver v.3.1 software (OpGen Technologies, Inc.).



Before: contigs have inversion

After: contigs have no inversion

Fig. 2. Illumina assembled contigs were aligned against the optical map of the genome according to their restriction fragment pattern

PCR primers were designed to close the gaps by amplified the PCR products and sequenced; these sequences were helpful in joining between the contigs/scaffolds. Finally, whole genome sequence of the bacteria was completely assembled into a circular genome (Ravindran *et al*., 2015).

Annotation and curation: The completely assembled was submitted to the Integrated Microbial Genomes Expert Review (IMG/ER) system (https://img.jgi.doe.gov/cgi-bin/er/main.cgi) for annotation by gene calling (Markowitz *et al*.,2009). The IMG annotation system predicted functions of the genomes based on matches to a combination of functional annotation databases (COG, Pfam, TIGRfam, InterPro, Gene Ontology, and KEGG). The annotated genomes were identified for the short, long, broken and interrupted genes and then to be corrected using Artemis. The Final whole-genome was submitted in IMG/ER and GenBank and further used for comparative genome studies.

Conclusion

For this analysis, we focused on bacteria with complete genomes and excluded draft genomes due to their limitations (Palmer and McCombie, 2002). It is unfortunate with the extensive availability of draft genomes that the more labor and time-intensive task of complete genomes has been skipped. While these sequences provide a wealth of data for researchers, they also point to the limitations of draft sequence versus complete sequence. The limitations of draft sequence can be grouped into three main areas: problems relating to the incompleteness of the genome sequence, problems relating to the discontinuity of the data and problems caused by the greater likelihood of errors in a draft sequence.

Acknowledgement

I thank Dennis C. Gross for giving this opportunity to work on project to complete genome sequencing of the bacteria. I gratefully acknowledge the Texas A & M AgriLife Research for providing for assistance and access to the CLC Genomics Workbench program.

Selected references

1. Loman, N.J. and Pallen, M.J. 2015. Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology.*13:787–794.
2. Markowitz, V.M., Mavromatis, K., Ivanova, N.N., Chen, I.M., Chu, K. and Kyrpides, N.C. 2009. IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics. 25:2271–2278.
3. Palmer, L.E. and McCombie, W.R. 2002. On the importance of being finished. Genome Biology. 3(10): comment2010.1–comment2010.4.
4. Ravindran, A., Jalan, N., Yuan, J.S., Wang, N. and Gross, D.C. 2015. Comparative genomics of *Pseudomonas syringae* pv. *syringae* strains B301D and HS191and insights into intrapathovar traits associated with plant pathogenesis. *MicrobiologyOpen.* 4(4): 553-573.

Do you know?

Sequence assembly refers to aligning and merging fragments of a much longer DNA sequence in order to reconstruct the original sequence.This is needed as current DNA sequencing technology cannot read whole genomes as a continuous sequence, but rather reads small pieces of between 20 and 1000 bases, depending on the technology used. Third generation sequencing technologies such as PacBio or Oxford Nanopore routinely generate sequencing reads >10 kb in length

11. Biological sample preparation for whole genome sequencing

Filamentous fungus

Neelam Sheoran, G. Prakash, Deeba Kamil and A. Kumar
Division of Plant Pathology, ICAR-Indian Agricultural Research Institute
New Delhi-110012

Introduction

One of the most crucial aspects of whole genome sequencing is preparation of biological samples. Not only the fungal mycelial sample must be free from any other microbial contamination but also must be genetically pure.Rice blast disease caused by *Magnaporthe oryzae* continue to plague rice cultivation in several paddies around the world leading to loss of productivity and escalation of cost of cultivation due to heavy reliance on agrochemicals. The pathogen is a major pathological threat to variety of important cereals including wheat, pearl millet, barley and rye, apart from its main host, rice (Talbot *et al*., 2001; Valent and Chumley, 1991). In the recent years, *M. oryzae* has become a model organism to understand plant-pathogen interaction (Talbot, 2003; Valent and Chumley, 1991). Isolation of this pathogen from infected host plant is very crucial and challenging task. To explore the behavior of this pathogen it is very important to adopt the potential methodology to isolate this pathogen from infected tissues. In the study we will try to isolate *Magnaporthe oryzae* from different host plants.

Materials and Method

Reagents: Blast samples, Sterile scissors, Needles and forceps, Sterile cotton, Sterile distilled water, Micropipettes, Microfuge tubes, Sterile filter paper discs, Shaker-Incubator, Oat meal agar, Rice Straw Extract medium, Potato dextrose broth, Sodium hypochlorite solution, STE Buffer, Proteinase K, SDS, Chloroform, Isoamyl alcohol, Isopropanol

Preparation of fungal mycelium

Method: Monoconidial culture of the fungus by single spore isolation technique (Rajashekara *et al*., 2016)

Isolation of fungus from infected plant tissues

1. Isolation of *M. oryzae*pathogen from blast infected samples (leaf and panicle) is carried out under aseptic conditions by spore drop method (Rajashekara *et al.,* 2016)
2. Select the ashy grey infected lesions for leaf blast samples and brown or black colored lesions for neck blast samples.
3. Cut the infected tissues into approximately 3-5 mm bits.
4. Surface sterilize the cut bits of leaf/neck blast samples with 1% sodium hypochlorite (or) 95% ethyl alcohol solution and wash three times with sterilized distilled water.
5. Place the surface sterilized leaf and neck blast lesions over sterilized moist cotton set-up in separate petriplates and incubate for 24-48 h at 25±1°C temperature.
6. Keep the incubated samples in the Lesion Print (LP) set-up, *i.e.* blast infected tissues transferred to the sterilized moist cotton stuck over the inner surface of the upper lid of small Petriplate (5.5cm diameter) and lower lid contains rice straw extract agar medium (RSEA).

7. Incubate the Lesion Print (LP) set-up at 25±1°C for 3 to 4 days and the single spore fungal colonies are visible on the medium. These colonies are further inoculated on Oat Meal Agar (OMA) medium by pour-plating method for *M.oryzae* spores.
8. Confirm the culture morphology and spore characters under microscope.
9. Transfer the single spore colony developed in RSEA medium to Oat Meal Agar (OMA) slants for short term storage and on sterilized filter paper discs (stored at -20°C) for long term preservation.
10. Proceed with DNA extraction using DNeasy Plant mini kit (QIAGEN) as per the protocol.
    OR
11. Proceed with CTAB+SDS method (Goes-Neto *et al*., 2005)

Mass multiplication of fungus

1. Inoculate single spore colony of *M.oryzae* pathogen in Potato Dextrose broth and incubate at 25-28°C with shaking at 120rpm.
2. Harvest the mycelial balls after 3-4 days by centrifugation at 8000 rpm followed by 3-4 washings with sterile distilled water
3. Filter the mycelium on a sterile Whatman no.1 paper and air dry the fungal mycelial mass to remove excess of moisture and grind the mycelial mats in liquid nitrogen and store immediately at -80°C. This can be used isolation of DNA

Total Genomic DNA isolation

1. Revive the stored culture of *M. oryzae* on suitable medium for mycelia growth (OMA).
2. Agar disc with fungal mycelium are further to be inoculated in Potato dextrose broth and incubate at 25-28°C with shaking 120rpm.
3. After 3-4 days, harvest the mycelial balls in the broth using centrifugation at 8000rpm followed by 3-4 washing with sterile distilled water.
4. Proceed for DNA extraction using DNeasy Plant mini kit (QIAGEN)/ MasterPure Complete DNA and RNA Purification Kit (Illumina)/ other commercially available DNA extraction kits as per the protocol.

'OR'

1. Grind mycelium (200-300 mg) with600 µlpre-warmed (65°C) STE buffer aided by autoclaved abrasive glass powder.
2. Incubate macerated mycelium with 5 µl of Proteinase K (20 mg mL^{-1}) at 37°C for 30 min.
3. Further, incubate with 75 µl of 20% SDS at 65°C for 1 h for complete lysis.
4. Clean the lysate twice with chloroform: Isoamyl alcohol (24:1).
5. Precipitate the DNA with 0.6 volume of isopropanol.
6. Dissolve the DNA in 50 µl deionized water.
7. Check the genomic DNA integrity on 0.8% agarose gel by loading 2 µl of extracted DNA and run at 110 V for 30 min.
8. Quantification and quality analysis of extracted genomic DNA can also be done using 1 µl of sample to determine the A260/280 ratio (Nanodrop 2000) and concentration (Qubit® 3.0 Fluorometer). Check the quality of genomic DNA in 0.7% agarose gel

Selected references

1. Goes-Neto, A., Loguercio-Leite, C. and Guerrero, R.T. 2005. DNA extraction from frozen field-collected and dehydrated herbarium fungal basidiomata: Performances of SDS and CTAB based methods. *Biotemas* 18, 19-32
2. Kumar, A., Sheoran, N., Prakash, G., Ghosh, A., Chikara, S.K., Rajashekara, H., Singh, U.D., Aggarwal, R. and Jain, R.K. 2017. Genome sequence of a unique Magnaporthe oryzae RMg-Dl isolate from India that causes blast disease in diverse cereal crops, obtained using PacBio single molecule and Illumina HiSeq2500sequencing. *Genome Announc* 5:e01570-16. https://doi.org/10.1128/genomeA.01570-16.
3. Rajashekara, H., Prakash, G., Pandian, R.T.P., Sarkel, S., Dubey, A., Sharma, P., Chowdary, V., Mishra, D., Sharma, T.R. and Singh U. D. (2017). An efficient technique for isolation and mass multiplication of *Magnaporthe oryzae* from blast infected samples. *Indian Phytopathology*. 69(4s): 260-265.
4. Talbot, N. J. and Foster, A.J. 2001. Genetics and genomics of the rice blast fungus *Magnaporthe grisea*: developing an experimental model for understanding fungal diseases of cereals. *Adv. Bot. Res.* 34, 263–87.
5. Valent, B. and Chumley, F.G. 1991. Molecular genetic analysis of the rice blast fungus *Magnaporthe grisea. Annu. Rev. Phytopathol.* 29, 443–67.

Do you know?

Assembly can be broadly categorized into two approaches: *de novo* assembly, for genomes which are not similar to any sequenced in the past, and comparative assembly, which uses the existing sequence of a closely related organism as a reference during assembly Relative to comparative assembly, *de novo* assembly is computationally difficult (NP-hard), making it less favorable for short-read NGS technologies. Within the *de novo* assembly paradigm there are two primary strategies for assembly, Eulerian path strategies, and overlap-layout-consensus (OLC) strategies. OLC strategies ultimately try to create a Hamiltonian path through an overlap graph which is an NP-hard problem. Eulerian path strategies are computationally more tractable because they try to find Eulerian path through a deBruijn graph!

12. Biological sample preparation for whole genome sequencing

Gram Positive and Gram Negative Bacteria

A. Kumar, Neelam Sheoran, *V. Govindasamy, *B. Ramakrishnan, *K. Annapurna
Division of Plant Pathology & *Division of Microbiology
ICAR-Indian Agricultural Research Institute
New Delhi-110012

Introduction

Gram negative bacterium *Ralstonia solanacearum,* a devastative bacterial wilt and rot causing pathogen causes huge losses in many important crops of economic value worldwide. In India alone more than 130 plant species belonging to 47 genera have been reported to be infected by this pathogen. *Ralstonia* causes huge losses in Solanaceae plants like potato, chilli, tomato.

For whole genome sequencing high quality high molecular weight DNA is essential.

Material and Methods

Reagents: Cassamino Peptone Glucose Agar; Micropipettes; Microfuge tubes; Incubator; Centrifuges, Deep Freezers

Sample collection and Isolation

1. Bacterial wilt affected plant samples are collected from field and processed for isolation of bacterium.
2. In order to isolate *Ralstonia solanacearum* from plants, excise stem pieces (2 to 3 cm long) from infected plants, wash five times in sterilized deionized water and blot dry them 15 min on an autoclaved paper towel.
3. Place the stem pieces in test tubes containing 5 ml of sterile water for approximately 5 to 10 min.
4. Afterward, streak a loopful of cell suspension onto Cassamino acid-peptone-glucose (CPG) agar (Cassamino acid, 1 g liter^{-1}; peptone, 10 g liter^{-1}; glucose, 10 g liter^{-1}; and agar, 15 g liter^{-1}; pH 7.2) amended with 2, 3, 5 triphenyl tetrazolium chloride (1%) and incubate for 36 to 48 h at 28 to 30°C.
5. Check for the appearance of fluidal white colonies with a pink center and store in 30% glycerol at −80°C for long-term storage.

The following protocol is universal DNA isolation method for Gram negative and Gram positive bacteria

Total Genomic DNA

1. Single colony of *Ralstonia*is inoculated in 5ml of CPG broth.
2. Incubate at 28°Cfor 36-48 h with shaking.
3. Use this culture for DNA extraction.
4. Grow isolate of interest o/n in 5 ml LB at 25°C, with shaking
5. Pipette 1.5 ml of o/n culture in 1.5 ml tube
6. Centrifuge 2 min at 14,000 rpm at Room Temperature

7. Discard supernatant and wash pellet 3 times with sterile water
8. Add 550µl of TE buffer + lysozyme
9. Resuspended by pipetting
10. Incubate suspension for 30 min at 37 °C
11. Add 76 µl of 10 % SDS + proteinase K
12. Mix contents by flipping tube
13. Incubate 15 min at 65 °C
14. Add 100 µl of 5 M NaCl and mix by flipping tube
15. Add 80 µl of CTAB/NaCl
16. Mix by flipping tube
17. Incubate for 10 min at 65 °C
18. Add 700 µl of Chloroform/Isoamyl alcohol
19. Mix contents by flipping tube for at least 15 sec
20. Centrifuge 5 min at 14,000 rpm at RT
21. Transfer aqueous layer (without disturbing or carrying over any of the white middle layer) to new 1.5 ml tube. This step can be repeated several times when proteins are taken up with the aqueous layer
22. Add equal volume of Isopropanol and invert several times to mix
23. Centrifuge for 15 min at 4 °C
24. Gently drain of the supernatant and carefully add approx. 1 ml ice cold 70 % ethanol
25. Collect DNA by centrifugation for several minutes at 14,000 rpm at RT
26. Carefully remove the supernatant and evaporate the remaining ethanol in the laminar flow cabinet
27. Dissolve DNA in 50-100 µl 10mM Tris (pH 8.0)
28. Dissolve the DNA by incubating over night at 4 °C
29. To remove contaminating RNA from preparation, add 6 µl of RNase to DNA solution, incubate for 30 min at 37 °C
30. Store DNA at -20 °C

Check the quantity of DNA in Nanodrop.Calculate A260/A280 ratio to check the purity of the DNA preparation (Reading from 1.8-2.0 indicate its good quality). Check the quality of genomic DNA in 0.7% agarose gel.

A minimum quantity of 400-500 µg per µl is needed for genomic library preparation for whole genome sequencing

Selected references

1. Kumar, A., Prameela, T.P., Bhai, R.S., Siljo, A., Anandaraj, M. and Vinatzer, B. A.2014. Host specificity and genetic diversity of race 4 strains of *Ralstonia solanacearum, Plant Pathology* (BSPP).63: 1138-1148.
2. Sakthivel, K., Kumar, A., Devendrakumar, C., Vibhuti, M., Neelam, S., Gautam, R.K., Kumar, K., Roy, S.D. and Vinatzer, B.A.2015. Diversity of *Ralstonia solanacearum* strains on the Andaman Islands in India.*Plant Disease*. 100 (4): 732-738.

13. Genomic library preparation for whole genome sequencing of *Magnaporthe oryzae* using SMRT and SBS technology

Neelam Sheoran, G. Prakash and A.Kumar
Division of Plant Pathology, ICAR-Indian Agricultural Research Institute
New Delhi-110012

Introduction

The second-generation sequencing (SGS) technologies have offered vast improvements over Sanger sequencing, their limitations, especially their short read lengths, make them poorly suited for some particular biological problems, including assembly and determination of complex genomic regions, gene isoform detection, and methylation detection. Single-molecule real-time (SMRT) sequencing, developed by Pacific BioSciences (PacBio), offers an alternative approach to overcome many of these limitations. Single-molecule real-time sequencing developed by Pacific BioSciences offers longer read lengths than the second-generation sequencing (SGS) technologies, making it well-suited for unsolved problems in genome, transcriptome and epigenetics research.

Objective

To prepare genomic library for whole genome sequencing of *Magnaporthe oryzae* using SMRT and SBS technology

Material

1. Illumina TruSeq Nano DNA HT Library Preparation Kit for Illumina platform
2. Hairpin adaptor protocol for ultra-long read sequencing for PacBio RSII platform
3. Illumina HiSeq 2500 platform
4. PacBio RSII platform

Protocol

1. Check the genomic DNA integrity on 0.8% agarose gel by loading 2 µl of extracted DNA and run at 110 V for 30 min.
2. Quantification and quality analysis of extracted genomic DNA can also be done using 1 µl of sample to determine the A260/280 ratio (Nanodrop 2000) and concentration (Qubit® 3.0 Fluorometer)

***Preparation of 2 x 125 HiSeq 2500 library**

1. Prepare the paired-end sequencing library using Illumina TruSeq Nano DNA HT Library Preparation Kit.
2. 200ng of gDNA to be fragmented by Covaris to generate a mean fragment distribution of 3500bp. Covaris shearing generates dsDNA fragments with 3' or 5' overhangs.
3. The fragments are then subjected to end-repair. This process converts the overhangs resulting from fragmentation into blunt ends using End Repair Mix.
4. The 3' to 5' exonuclease activity of this mix removes the 3' overhangs and the 5' to 3' polymerase activity fills in the 5' overhangs.
5. A single 'A' nucleotide is to be added to the 3' ends of the blunt fragments to prevent them from ligating to one another during the adapter ligation reaction.

6. A corresponding single 'T' nucleotide on the 3' end of the adapter provides a complementary overhang for ligating the adapter to the fragment. This strategy ensures a low rate of chimera (concatenated template) formation.
7. Ligate the Indexing adapters to the ends of the DNA fragments, preparing them for hybridization onto a flow cell.
8. The ligated products are to be purified using SP beads supplied in the kit.
9. Further proceed with PCR amplification of the size-selected product as described in the kit protocol.

***Preparation of PacBio DNA library**

1. Use high molecular weight DNA to prepare 1 SMRT bell library of 5-8kb for sequencing on PacBio platform using Hairpin adaptor protocol for ultra-long read sequencing.
2. The library is to be individually indexed for sequencing on PacBio RSII platform.

This illumina library prepared from sample using TruSeq Nano DNA HT Library sample preparation Kit for sequencing on HiSeq 2500 using 2 x 125 bp chemistry will generate 10 Gb of data. The PacBio bell library prepared from the high molecular weight DNA using Hairpin adaptor protocol for ultra-long read sequencing on PacBio RS II platform will generate ~400 – 500 Mbp data using one SMRT cell.

Selected references

1. Bao, J., Chen, M., Zhong, Z., Tang, W., Lin, L., Zhang, X., Jiang, H., Zhang, D., Miao, C., Tang, H., Zhang, J., Lu, G., Ming, R., Norvienyeku, J., Wang, B. and Wang Z. 2017. PacBio Sequencing Reveals Transposable Element as a Key Contributor to Genomic Plasticity and Virulence Variation in *Magnaporthe oryzae*. *Mol. Plant*. doi: 10.1016/j.molp.2017.08.008.
2. Kumar, A., Sheoran, N., Prakash, G., Ghosh, A., Chikara, S.K., Rajashekara, H., Singh, U.D., Aggarwal, R. and Jain, R.K. 2017. Genome sequence of a unique Magnaporthe oryzae RMg-Dl isolate from India that causes blast disease in diverse cereal crops, obtained using PacBio single molecule and Illumina HiSeq2500sequencing. *Genome Announc* 5:e01570-16. https://doi.org/10.1128/genomeA.01570-16.
3. Shirke, M.D., Mahesn, H.B and Gowda, M. 2016. Genome-Wide Comparison of *Magnaporthe* Species Reveals a Host-Specific Pattern of Secretory Proteins and Transposable Elements. *PLoS One* 11(9): e0162458. doi:10.1371/journal.pone.0162458.

Do you know?

The term microbiome was 'coined' by Nobel laureate-microbiologist Joshua Lederberg in a 2001

14. Gene finding strategies and annotation of sequence reads

Neelam Sheoran and A.Kumar
Division of Plant Pathology, ICAR-Indian Agricultural Research Institute
New Delhi-110012

The annotation process can be conceptually divided into two phases: a 'computational phase' where several lines of evidence from other genomes or from species-specific transcriptome data are used in parallel to create initial gene and transcript predictions. In a second 'annotation phase', all information is then synthesized into a gene annotation, following a set of rules determined by the annotation pipeline which is as follows:

1. After the initial assembly, contigs are joined to form longer stretches of sequence known as scaffolds.
2. For prokaryotic genomes (say *Ralstonia solanacearum*) the prediction of genes is done from these assembled scaffolds with the help of softwares such as Prodigal (Prodigal: prokaryotic gene recognition and translation initiation site identification)https://github.com/hyattpd/Prodigal.
3. Functional annotation of the genes is performed using BLASTx program, which is a part of NCBI blast-2.3.0+ standalone tool. BLASTx find the homologous sequences for the genes against NR (non-redundant protein database).
4. Gene ontology (GO) annotations of the genes are determined by the Blast2GO (B2G) program (https://www.blast2go.com/blast2go-pro/download-b2g). GO terms are assigned to genes for functional categorization. The genes are categorized into different categories such as biological process, molecular functions, and cellular component.
5. B2G is a Java application made available by Java Web Start. It is platform independent and has no further requirements than an Internet connection.
6. After obtaining GO annotation for each gene, The Web Gene Ontology Annotation Plot (WEGO) software(wego.genomics.org.cn/) is used to display GO functional classification. The main purpose of the WEGO is to visualize the annotation of sets of genes, comparing the provided gene datasets and plotting the distribution of GO annotation results into a histogram.
7. For phylogenetic analysis with the available genomes, AAF (Alignment and assembly free) phylogeny tool(https://omictools.com/alignment-and-assembly-free-tool) is used. The phylogenic tree is constructed using Mega6.

Selected references

1. Kanehisa, M. and Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 28: 27-30.
2. Primmer, C.R., Papakostas, S., Leder, E.H., Davis, M.J. and Ragan, M.A. 2013. Annotated genes and non-annotated genomes: cross-species use of Gene Ontology in ecology and evolution research. *Mol Ecol*. 22: 3216–3241.

15. Isolation of RNA and preparation of cDNA for gene expression analysis

Neelam Sheoran,Deeba Kamil, *V. Govindasamy and A. Kumar
Division of Plant Pathology &*Division of Microbiology,
ICAR-Indian Agricultural Research Institute, New Delhi-110012

Introduction

The key to successful purification of intact RNA from cells and tissues is speed. Cellular RNases should be inactivated as quickly as possible at the very first stage in the extraction process. TRIzol™ Reagent is a monophasic solution of phenol, guanidine isothiocyanate, and other proprietary components which facilitate the isolation of a variety of RNA species of large or small molecular size. TRIzol™ Reagent maintains the integrity of the RNA due to highly effective inhibition of RNase activity while disrupting cells and dissolving cell components during sample homogenization. The yield of total RNA depends on the tissue or cell source, but it is generally in the range of 4-7 µg/mg of staring tissue. The A_{260}/A_{280} ratio of the extracted RNA is generally 1.9-2.0.

Methodology and technique involved

a. **Preparation of fungal mycelium: Refer chapter on biological sample preparation**

b. **Isolation of total RNA**

Precautions

1. Take precautions to avoid RNase contamination when preparing and handling RNA.
2. Pestle, mortar, tubes should be DEPC treated (0.1%) before use II.

I. Homogenizing samples

1. Perform homogenization at room temperature.
2. The sample volume should not exceed 10% of the volume of TRIzol® Reagent used for homogenization.
3. Add 1 mL TRIzol® Reagent per 50–100 mg of tissue sample. [Note: Be sure to use the indicated amount of TRIzol® Reagent, because an insufficient volume can result in DNA contamination of isolated RNA]

II. Phase separation

1. Incubate the homogenized sample for 5 min at room temperature to permit complete dissociation of the nucleoprotein complex.
2. Add 0.2 mL of chloroform per 1 mL of TRIzol® Reagent used for homogenization. Cap the tube securely.
3. Shake tube vigorously by hand for 15 sec.
4. Incubate for 2–3 min at room temperature.
5. Centrifuge the sample at 12,000 rpm for 20 min at 4°C. [Note: The mixture separates into a lower red phenol chloroform phase, an interphase, and a colorless upper aqueous phase. RNA remains exclusively in the aqueous phase. The upper aqueous phase is ~50% of the total volume.]
6. Remove the aqueous phase of the sample by angling the tube at 45°C and pipetting the solution out. Avoid drawing any of the interphase or organic layer into the pipette when removing the aqueous phase.

7. Place the aqueous phase into a new tube and proceed to RNA precipitation.

III. RNA precipitation

8. Add 0.5 mL of 100% isopropanol to the aqueous phase, per 1 mL of TRIzol® Reagent used for homogenization.
9. Incubate at room temperature for 10 min.
10. Centrifuge at 12,000 rpm for 20 min at 4°C. [Note: The RNA is often invisible prior to centrifugation, and forms a gel-like pellet on the side and bottom of the tube.]
11. Proceed to RNA wash.

IV. RNA wash

12. Remove the supernatant from the tube, leaving only the RNA pellet.
13. Wash the pellet, with 1 mL of 75% ethanol per 1 mL of TRIzol® Reagent used in the initial homogenization.
14. Vortex the sample briefly, then centrifuge the tube at 7500 × g for 5 min at 4°C. Discard the wash.
15. Vacuum or air dry the RNA pellet for 5–10 min. Do not dry the pellet by vacuum centrifuge. [Note: Do not allow the RNA to dry completely, because the pellet can lose solubility. Partially dissolved RNA samples have an A260/280 ratio < 1.6.
16. Proceed to RNA suspension.

V. RNA resuspension

17. Resuspend the RNA pellet in RNase-free water (20–50 µL) by passing the solution up and down several times through a pipette tip.
18. Incubate in a water bath or heat block set at 55–60°C for 10–15 min.
19. Proceed to downstream application, or store at –80°C.

VI. First strand cDNA synthesis

1. Complementary DNA (cDNA) will be synthesized by using RNA as a template with standard kit like (Verso cDNA kit of Thermo Fisher Scientific)
2. Protocol example of reaction mix preparation
3. The volume of each component is for a 20 µL final reaction.

Components	**Volume needed**	**Final concentration**
5X cDNA synthesis buffer	4 µL	1X
dNTP Mix	2 µL	500 µM each
RNA Primer	1 µL	
RT Enhancer	1 µL	
Verso Enzyme Mix	1 µL	
Water (PCR grade)	Variable	
Template (RNA)	1 - 5 µL	1 ng
Total	20 µL	

4. Then proceed to cDNA synthesis

Programme	Temp	Time	No of cycles
cDNA synthesis	42ºC	30-60 min	1 cycle

Notes

1. The volume of the total reaction should be completed up to 20 µL with water.
2. To remove secondary structure, heat at 70°C for 5 min and place immediately on ice
3. The amount of RNA added as a template should be between 1 pg and 1 µg.
4. Depending on the length of template and degree of secondary structure, the efficiency of the first strand synthesis maybe improved by optimizing temperature and time (42-57°C for 5-60 min)
5. Adjust the final volume to 25 µL using 15.5 µL of nuclease free water.
6. Perform analysis of transcription with the fungal actin primers and amplify target geneusing 94°C for 3 min, 30 cycles of 94°C for 30 sec, 52°C for 40 sec, 72°C for 30 sec with a final extension at 72°C for 7 min
7. Perform electrophoresis on 1.2% agarose gel for the PCR products (25 µl).

Do you know?

Structural genomics seeks to describe the 3-dimensional structure of every protein encoded by a given genome. This genome-based approach allows for a high-throughput method of structure determination by a combination of experimental and modeling approaches. The principal difference between structural genomics and traditional structural prediction is that structural genomics attempts to determine the structure of every protein encoded by the genome, rather than focusing on one particular protein

16. Sequence based molecular phylogeny: basic concepts and application

Anirban Roy
Plant Virology Unit, Division of Plant Pathology
Indian Agricultural Research Institute, New Delhi- 110012

Introduction

The development of new rapid, inexpensive next generation high-throughput sequencingtechnologies over the last 10 years or so is changing the ways we think about the application of sequences to plant virology. The analysis and comparison of sequence data are playing an increasing role in virus classification. Most of these analyses and comparisons are undertaken using Bioinformatics. Bioinformatics, a new field of science includes biology, computer science, statistics and Information Technology. The sudden growth in the quantitative information in biology has resulted in realization of inherent bio-complexity issues which call for innovative tools to convert the information into knowledge. Bioinformatics, in one hand, involves computer specialists and statisticians for development of the tools and new algorithms for organizing and analyzing the data and in other hand helps biologists in understanding the structural and functional genomics, proteomics, protein engineering etc. using those tools (computational biology) in a biologically meaningful manner. In line with the theme of the "Central Dogma", bioinformatics utilizes the prediction approach to find out the sequence similarity in DNA that can lead to structural and functional similarity in protein and thus narrows down the search for understanding the functional role of a protein.

Due to ease of sequencing technology huge number of viruses is being sequenced every day. There are 27,091 full-length virus genomes deposited in GenBank as of 2010. The technology thus allows a larger genomic view of viruses, particularly as populations rather than single entities. The data generated thus is needed attention for proper analysis to illustrate them in more meaningful way. From the first virus genome sequence completed (MS2; 3.6 kb) (Fiers *et al.*, 1976) to that of the recent largest virus (mimivirus; 1.2 Mb) (Le Raoult *et al.*, 2004), the viral genome and its analysis have revealed high-resolution details of the molecular basis of a particular biological system, along with unexpected and surprising details. Viral genomes differ from other organism genomes in complexity, despite their generally smaller sizes and presumed "simplicity". For example, one problem is that the smaller size of the genome dictates a higher density of gene coding, with all six reading frames utilized. Coding regions frequently overlap.

Databases and bioinformatic tools that contain genomic, proteomic, and functional information have become indispensable for virology studies. Bioinformatic analysis on viruses involves the general tasks related to the analysis of any novel sequences, such as identification of open reading frames, gene prediction, base calling and assembly, homology searching, sequence alignment, and motif and epitope recognition.The recent ICTV Reports on Virus Taxonomy "Species demarcation criteria in each genus." These criteria include nucleic acid sequence and amino acid sequence derived from the genome sequence data differentiation for many of the appropriate genera. Molecular phylogenetic analysis helps us identifying and grouping viruses.

Different terminologies those are being routinely used bioinformatics analysis are described below:

Terminologies and concept of sequence analysis

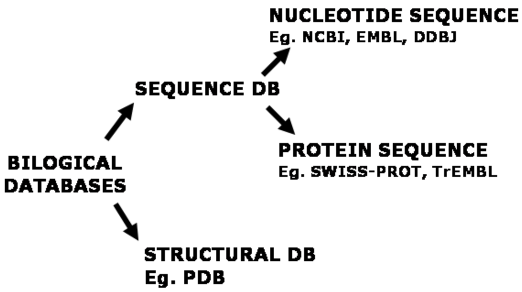
Biological Database:

Organized body of persistent data

Update

Query

Retrieve Components



Ways of submitting DNA sequence

- There are two principal ways of submitting DNA sequences to GenBank and EMBL.
- BankIt
- Sequin
- Webin-Align

Annotation

Refers to commentary or explanation of the information appended to DNA or protein sequences stored in databases.

Annotation can include:

Known information about

- Source, Country, Organism
- protein(s) sequence
- predicted protein structure
- domain(s) of the protein.
- quaternary structure of the protein.
- protein function
- common post-translational modifications of the protein

Data Retrieval

Collection of data from databases

Data Mining

Generation of information from data in databases. E.g. – primer designing, gene finding, phylogenetic relationship study etc.

Gene Finding approaches

- Content based approach: The content based approach relies upon the differences in composition of nucleotide bases between the coding exons and noncoding introns. The periodicity of repeats and compositional complexity of codon triplets differentiate the exons from introns.
- Site based approach: The gene has its own syntax. Start codon,stopcodon, donor and acceptor sequences, noncoding introns, ribosome binding sites, transcription factor binding sites, promoter sites, the poly adenylate sites etc are the specific signatures of genes
- Comparative method: The anonymous sequence is compared with cDNA sequence library.

Phylogenetic relationship study: Terminologies and Concepts

Homology:

This is a state of gene or morphological character that shares a common ancestry with a different gene or morphological character. For molecular sequence data, it is taken to mean that two sequences or even two characters within sequences are descended from a common ancestor.

This term is frequently misused as a synonym for 'similar', as in "two sequences were 70% homologous". This is totally incorrect! Sequences show a certain amount of similarity. From this similarity value, we can probably infer that the sequences are homologous or not.

Homology cannot be measured only we can say whether homology is there or not but we can measure similarity. Homologous sequence must have similarity, but if there is similarity we cannot say there is homology

Homologous Gene Super family

A) Orthologous Gene

Same sequence and same function but found in different taxa. E.g. - DNA polymerase of Goat, DNA Polymerase of Human. Result of lineage Transfer.

B)Paralogous Gene

Found in same taxa. Same sequence but different function. E.g. - Hemoglobin, Myoglobin. Result of a gene duplication.

Alignment

An Alignment is a computational hypothesis which identify positional similarity or identity between bases/Amino Acids. Two ways: Local and Global Alignment.

Sequence Alignment Tools

- BLAST
- FASTA
- BLITZ
- BEAUTY, a modified BLAST

BLAST (Basic Local Alignment Search Tool)

- BLAST is the algorithm used by a family of five programs that will align your query sequence against sequences in a molecular database.
- Statistical methods are applied to judge the significance of matches.
- Alignments are reported in order of significance, as estimated by the applied statistics.
- BLASTN: Compares a nucleotide query sequence against a nucleotide sequence database.
- BLASTP: Compares an amino acid query sequence against a protein sequence database.
- BLASTX: Compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
- TBLASTN: Compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).

What We Know From BLAST

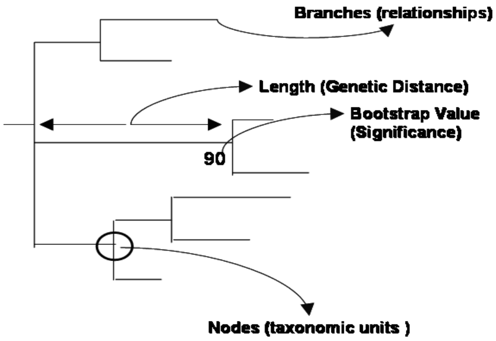
- Sequences that share similarity with query sequence
- Helps to retrieve those sequences

What We Do Not Know From BLAST

- Cannot quantify the sequence similarity
- Cannot tell us about the relationship between all those sequences

Multiple Alignment: Clustal W

Quick pairwise alignment: calculate distance matrix …> guide tree…> Progressive alignment following guide tree

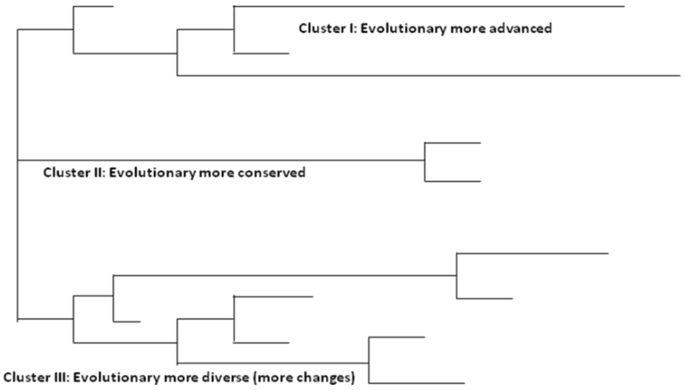


The branching pattern of a tree is called the TOPOLOGY
Representation of relationship through LINE: DENDROGRAM

Types of dendrogram

Phylogram

This is a phylogenetic tree that indicates the relationships between the taxa and also conveys a sense of time or rate of evolution. The temporal aspect of a phylogram is missing from a cladogram.


Cluster I: Evolutionary more advanced
Cluster II: Evolutionary more conserved
Cluster III: Evolutionary more diverse (more changes)

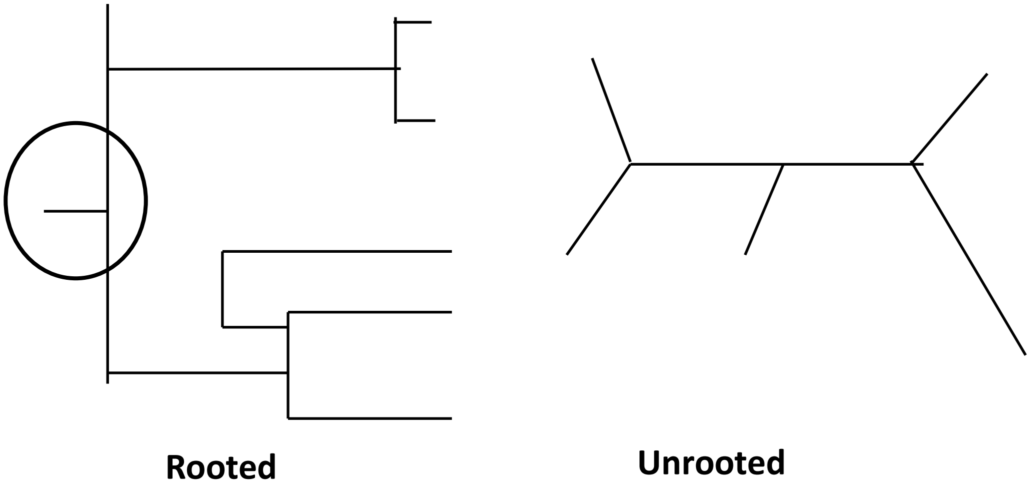
Cladogram

A dendrogram depicting the hypothesized branching order of a number of sequences. Cladograms do not give any indication of temporal change, but phylogram does.
Rectangular Cladogram / Phenogram – Suitable for grouping in taxonomic studies
Slanted Cladogram – Suitable for understanding convergence, divergence or parallelism in evolutionary studies

Presentation of a tree

Rooted tree: Assume that all taxa derived from a common ancestor

Unrooted tree: Assume that all taxa derived not from a common ancestor



Methods for constructing phylogenetic tree:

Character based and distance based method for tree development

Neighbor-joining tree - distance based method

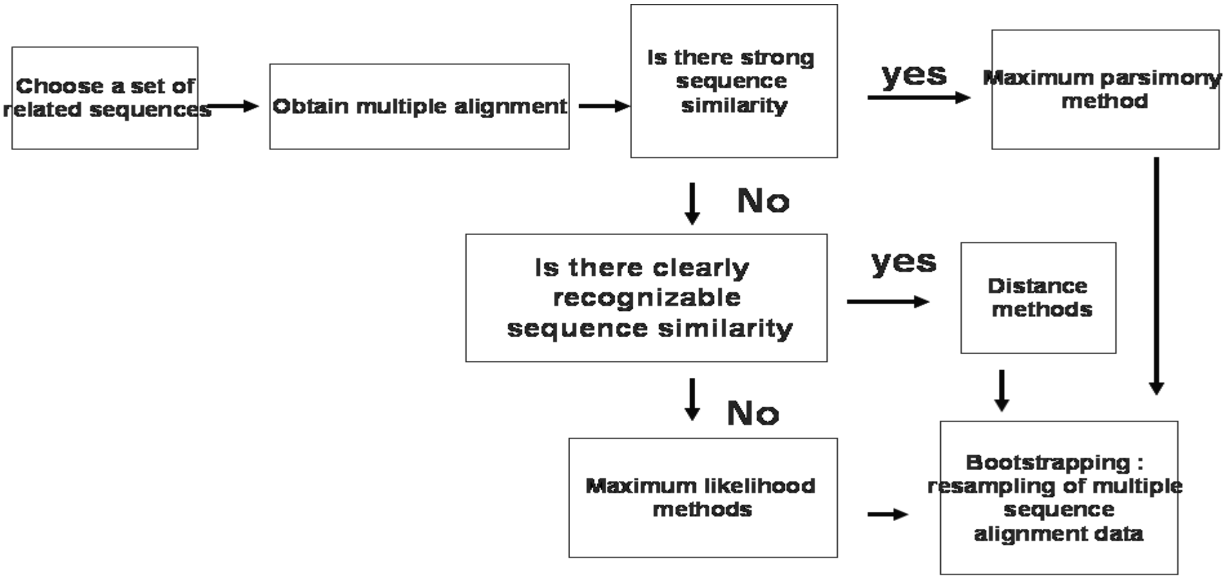
Parsimony tree – character based tree. Use when sequences are quite similar, e.g. – strains of different viruses. Use small numbers of sequences for parsimony analysis.

Bootstrapping:

- The bootstrap is a method for assessing the statistical significance the positions of branches in a phylogenetic tree.
- For each aligned pair, it samples scores from random positions in the alignment, adding the scores.
- When all the pairs have been sampled, it converts the scores to distances and computes a tree.
- This whole process is repeated many times and the frequency with which particular tree features are observed is taken as a measure of the probability that the feature is correct.

When to choose what type of tree:

Different bioinformatics analysis that are being routinely used for virus genomics studies are described below:


Choose a set of related sequences
Obtain multiple alignment
Is there strong sequence similarity
yes
Maximum parsimony method
No
Is there clearly recognizable sequence similarity
yes
Distance methods
No
Maximum likelihood methods
Bootstrapping : resampling of multiple sequence alignment data

Open reading frame (ORF) identification and gene prediction

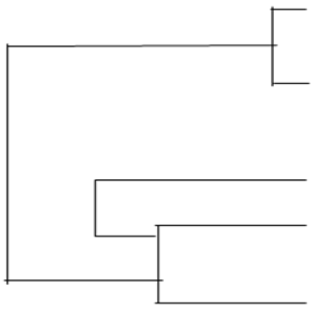
ORF finding is the basis for further homologous search, functional analysis, and identification of viral proteins for possible utilities such as antiviral agents or vaccine targets. From genomic DNA or RNA sequences, ORFs can be identified for candidate genes. National Center for Biotechnology Information (NCBI)'s ORF Finder program (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) is a general ORF prediction tool. The program GeneMark (http://opal.biology.gatech.edu/GeneMark/genemarks.cgi) also provides gene prediction tools for viruses. In addition, the Gene Ontology (GO) (http://www.geneontology.org/) provides a controlled vocabulary for genome annotation.

Homology searching and sequence alignment

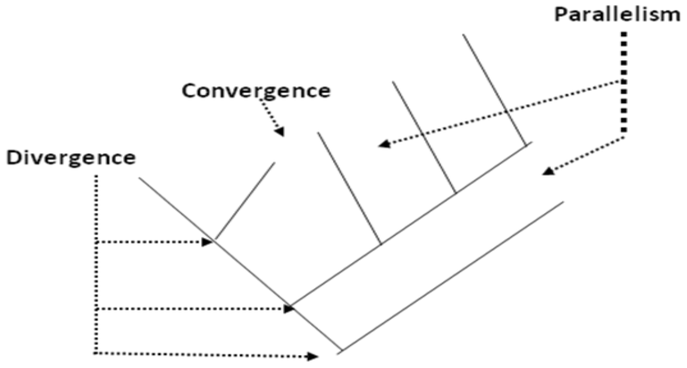
Homology searching against known or already annotated viral genomes, such as using the BLAST program, can also be used for predicting genes in unknown viral genomes. Homology searching is usually the next step for genome annotation and functional analysis after ORF finding in viral genome research. A high degree of homology between an ORF from an unknown genome and a known protein may suggest the new protein's similar function to the known one. A commonly used program for homology searching is BLAST. The program can be used for both nucleotide and amino acid sequence searching.

The sequence alignment program ClustalW (http://www.ebi.ac.uk/clustalw/) has been used extensively in studying viral genomes. Nucleotide and amino acid sequence alignments are important in comparing viral sequences in different species and strains. Such analysis is useful for identifying similarities, comparing conserved and non-conserved regions, establishing evolutionary relationships, and building phylogenetic trees.

A comprehensive list of programs for building phylogenetic trees is available at Phylogeny Programs (http://evolution.genetics.washington.edu/phylip/software.html). Phylogeny packages are grouped nicely at this site, according to the available methods such as maximum likelihood and Bayesian methods, or computer systems on which they work.



Rectangular Cladogram / Phenogram



Slanted Cladogram

Basic sequence analysis steps:

1. After obtaining a sequence purify it from the contamination of vector sequence. Use online service like VecScreen (www.ncbi.nlm.nih.gov/tools/vecscreen/) for the purpose. After an initial idea from VecScreen, carefully see the border region between vector and insert using Bioedit Sequence Alignment Editor Software and remove the vector sequence.
2. Join two or more sequences of a single clone obtained from primer walking by removing the overlapping sequence (use "allow end to slide" option in Bioedit to find the overlapping ends).
3. If it is circular molecule, then find the origin of the sequence (e.g. – in case of begomoviruses it is TAATATT↓ACC)
4. Go to BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and choose a BLAST program to run (e.g. - nucleotide blast for searching a nucleotide database using a nucleotide query). Paste the query sequence and analyse it using either megablast (if you expect a highly similar sequences) ordiscontiguous megablast (when there is more dissimilar sequences you expect in database) or blastn (when there is somewhat similar sequences you expect in database).
5. Select the sequences in database which showed high scores (low E-value) after analysis in BLAST. Retrieve those sequences from database in fasta format.
6. Find out the ORFs in the virus genome using online service ORF finder (www.ncbi.nlm.nih.gov/projects/gorf/gorf.html). If anomaly observed carefully check the sequence as there may be some sequencing error due to repetitive sequence.
7. Annotate the sequence based on their feature (ORFs, any typical feature like stem loop structure etc.).
8. Do a multiple alignment using ClustalW algorithm in Bioedit and develop a sequence identity matrix.
9. Alternatively do the alignment using MEGA software and develop a bootstrapped consensus phylogenetic tree.
10. Draw the genome of the virus using CloneMap software to visualize the genome organization.

Do you know?

Nasuia deltocephalinicola **was recently discovered to have the smallest genome of all bacteria, with 112,091 nucleotides.**

17. Molecular phylogenetic analysis of fungi by MLST: Case study with *Magnaporthe oryzae*

A. Kumar, Neelam Sheoran, G. Prakash, Deeba Kamil, Asharani Patel and Mukesh Kumar
Division of Plant Pathology, ICAR-Indian Agricultural Research Institute
New Delhi-110012

Introduction

Multilocus sequence typing (MLST) characterizes isolates of microbial species using the DNA sequences of internal fragments of multiple housekeeping genes. Upon sequencing of approximately 450-500 bp internal fragments of each gene, the different sequences present within a bacterial species are assigned as distinct alleles and, for each isolate, the alleles at each of the loci define the allelic profile or sequence type. MLST is of paramount importance in the context of surveillance and management of disease outbreaks, because of its ability to quickly type and track infectious diseases. Many studies exemplify the use of MLST under circumstances such as detection of disease outbreaks, estimation of prevalence rates and the origins of virulence factors.

Objective

To genotype *M. oryzae* isolates representing diverse host and geography

Selection of Loci

MLST-based genotyping to be conducted on *M. oryzae* isolates as described by Castillo and Greenberg (2007). This MLST scheme exploits variation in nucleotide sequences of eleven housekeeping genes (Calmodulin- *Cal*; DNA replication ATP-dependent helicase dna2- *Dna2*; DnaJ domain-containing protein- *DnaJ*; Elongation factor Tu- *Ef-Tu*; glyceraldehyde-3-phosphate dehydrogenase- *GAPDH*;guanylate kinase- *GuaK*; DNA ligase 1- *DNALig*; 6-phosphofructo-2-kinase 1- *Pfk*; phosphoglycerate kinase- *Pgk*; transcription initiation factor TFIID subunit 6- *Tif-TFIID*; DNA topoisomerase III- *DNA-TopoIII*) and seven virulence-related genes (ATP-binding cassette "ABC1 protein- *ABC1*; exocyst complex protein -*EXO70*; myosin regulatory light chain cdc4- *Mlc1*; hydrophobin-like protein- *MPG1*; TTK protein kinase *MPS1*; WD repeat-containing protein slp1- *Slp1*; alpha, alpha-trehalose-phosphate synthase 1- *Tps1*). Primers for 11 housekeeping and 7 effector genes are designed using Primer 3 plus software and synthesized. Twenty four isolated from our study includes 11 isolates from leaf region of infected rice plants from Kashmir, 10 isolates from panicles of infected rice plants from Kashmir, one isolate each collected from rice plants of Madhubani, Bihar; finger millet plant from Karnataka and pearl millet plant from Delhi region respectively.

PCR Amplification and Sanger's sequencing

PCR to be carried out as described previously (Kumar *et al*., 2014). Briefly, the reaction mixture (50 µl) contained 100 ng of template DNA, 1× PCR buffer, 1.5 mM $MgCl_2$,50mM each dNTP, 10 pmol of primers and 1 U of Taq DNA polymerase. The PCR thermal profile includes a 9-min denaturation step at 96°C; followed by 30 cycles at 95°C for 1 min, the appropriate annealing temperature for 1 min and extension temperature of 72°C for 2 min; with a final extension step at 72°C for 10 min. PCR optimization for specific size amplicon to be performed and Sanger's sequencing to be done to get partial sequence of each gene. All sequences needs to be end trimmed, edited, annotated and submitted to GenBank.

Phylogenetic analysis

Sequences for all eleven loci are to be concatenated and used for establishing phylogeny. Proceed further for multiple alignment of sequence using CLC workbench to study the variation in each isolate. The evolutionary history inferred using 7609 bp of concatenated sequences representing multiple housekeeping genes. A phylogenetic tree is to be constructed with the concatenated sequences using the Neighbor-Joining method (Saitou and Nei, 1987). Compute the evolutionary distances using the Maximum Composite Likelihood method (Tamura *et al*., 2004). Evolutionary analyses are to be conducted in MEGA 6.04 (Tamura *et al*., 2013).

Case study observation

MLST based genotyping clustered all rice isolates in one group with finger millet and pearl millet isolates clustered separately in two groups. Finger millet isolates was found to be closer with cluster formed by rice isolates whereas Pearl millet isolates formed separate group depicting the variability from other rice and finger millet isolates. No genetic difference between neck blast and leaf blast isolates was observed for rice isolates. Very minor variation among the isolates is observed indicating near isogenic population within the field.

Selected references

1. Castillo, J.A. and Greenberg, J.T. 2007. Evolutionary dynamics of *Ralstonia solanacearum*. *Appl. Environ. Microbiol.* 73:1225-1238.
2. Kumar, A., Prameela, T. P., Suseelabhai, R., Siljo, A., Anandaraj, M., and Vinatzer, B. A. 2014. Host specificity and genetic diversity of race 4 strains of *Ralstonia solanacearum*. *Plant Pathol.* 63:1138-1148.
3. Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol*. 4:406-425.
4. Tamura, K., Nei, M. and Kumar, S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. USA* 101:11030-11035.
5. Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. 2013. Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.* 30:2725-2729.

Do you know?

Paris japonica **has the largest genome of any plant yet assayed, about 150 billion base pairs long. An octoploid and suspected allopolyploid hybrid of four species, it has 40 chromosomes.**

18. Molecular phylogenic analysis of bacteria by MLST: Case study with *Ralstonia solanacearum*

A.Kumar, M. Ashajyothi, Neelam Sheoran and *S. Subramanian
Division of Plant Pathology& Division of Entomology,
ICAR-Indian Agricultural Research Institute, New Delhi-110012

Introduction

Though popular, the phenotypic techniques are not enough for deciphering the population biology of the strains across the geographical locations. Besides, these two independent systems of classification are not in agreement with each other. Perhaps these anomalies prompted the researchers to devise finer tools to reveal the "actual diversity" exist in population of *Ralstonia solanacearum*. Many techniques based on the electrophoretic mobility of the genomic fragments are in use for the analysis of population structure of *R. solanacearum* isolates worldwide. Sequence based discrimination of strains such as Multilocus sequence typing (MLST) and Comparative Genome Hybridization (CGH) which uncovers allelic variants in conserved housekeeping and virulence genes is portable across the laboratories.Multilocus sequence typing approach uses sequences of internal gene fragments and assigns different allele numbers to the sequence at each locus, so it will provide unique allelic profile for each isolate called **Sequence Types** (STs). Based on this approach Castillo and Greenberg (2007) had analyzed the evolutionary forces operating on *R. solanacearum* populations using Multilocus Sequence Typing (MLST) including five housekeeping and three virulence-related genes. *R. solanacearum* to be a diverse pathogen, showing high levels of nucleotide polymorphism and a number of unique alleles in the Chromosome and in the Megaplasmid.

Objective

To genotype *Ralstonia solanacearum* isolates representing diverse geography

Multilocus Sequence Typing

The selected genes are amplified, purified and sequenced. The sequence reads are assembled and compared with the database for assigning the alleles. The combination of the allele numbers is unique for each strain of the bacterium in question. The allele numbers are further compared among the strains in order to decipher the strain migration in the field of molecular epidemiology.

Choice of loci

For the diversity analysis five housekeeping genes, which resides in the chromosome *(pps*A, phosphoenol pyruvate synthase; *gyr*B, DNA gyrase, subunit B; *adk*, adenylate kinase; *gdh*A, glutamate dehydrogenase oxidoreductase; and *gap*A, glyceraldehyde 3-phosphate dehydrogenase oxidoreductase) and three plasmid borne virulence related genes (*hrp*B, regulatory transcription regulator; *fli*C, encoding flagellin protein; and *egl*, endoglucanase precursor) are considered.

PCR Amplification

For PCR amplification, prepare reaction mixture (50µl) containing 50-100ng of template genomic DNA, 1× PCR buffer, MgCl2 1.5mM, each dNTPs 200µM, 100 µM of each primer), and 1 U of *Taq* DNA polymerase.DNA is amplified using an initial denaturation at 95ºC for3 min, followed by 35 cycles of 95ºC for 30s, annealing for 30s and extension 72ºC for 1 min. Reaction is completed with a final extension step of 10 min at 72ºC. All PCR products are electrophoresed through a 1.0 % agarose gel and visualized with UV light after ethidium bromide staining.

Sequence analysis

Sequences are carefully analysed and sequence type assigned for each of the strain by comparing the data sets with www.pamdb.org. The strain relation with the existing collection of strain can be determined by eBurst programme (http://eburst.mlst.net).

Handling sequence data

The sequencing machines would give us the chromatogram indicating the quality of the sequence reads. The sequence reads are carefully observed for any errors in the base using any one of the chromatogram viewers (e.g. DNA baser, Bioedit, Chromos etc). Thus obtained sequence is called as raw sequence. For each gene, two such sequences are obtained which are known as forward sequence and reverse sequence respectively. The forward and reverse sequences are assembled using any one of the programmes that are available in public domain (e.g. DNA baser). The assemble sequences are called Contigs. Such a Contigs are used to determine the allelic variations.

The string of allele numbers (integers) for the housekeeping and virulence genes obtained for a strain is called as sequence type which is specific for a strain of bacterium. For example, the allele numbers obtained for a cardamom strain of *Ralstonia solanacearum* is *pps*A-10, *fli*C-19, *hrp*B-27, *gdh*A- 24, *adk*-1, *gyr*B-26, *egl*-25. The combination of integers (10, 19, 27, 24, 1, 26, and 25) serves the input data for establishing the strain relationship by eBurst programme which is based on eBurst algorithm, a dedicated programme for analysis of microbial MLST data.

Phylogenetic analysis using MLST data

The allelic sequences, thus, obtained from the strains are pooled to construct concatenated sequences which serve input data for establishing phylogeny. The concatenated sequence is nothing but the string of all the loci are assembled in an order (ppsA + fliC + hrpB + gdhA + adk + gyrB + egl) to get large sequence length. This large sequence length is used in the phylogenetic analysis of bacterium in question.

Selected references

1. Castillo, J.A. and Greenberg, J.T. 2007 Evolutionary dynamics of *Ralstonia solanacearum*.*Appl. Environ. Microbiol*.73: 1225-1238.
2. Kumar, A., Prameela, T.P., Bhai, R.S., Siljo, A., Anandaraj, M. and Vinatzer, B. A.2014. Host specificity and genetic diversity of race 4 strains of *Ralstonia solanacearum, Plant Pathology* (BSPP).63: 1138-1148.
3. Sakthivel, K., Kumar, A., Devendrakumar, C., Vibhuti, M., Neelam, S., Gautam, R.K.,   Kumar, K., Roy, S.D. and Vinatzer, B.A. 2015. Diversity of *Ralstonia solanacearum* strains on the Andaman Islands in India, *Plant Disease*. 100 (4): 732-738.

Do you know?

In 1995, Hamilton O. Smith and his team from The Institute for Genomic Research sequenced the first genome of a free living organism *Haemophilus influenzae* having genome size of 1.8 Mb.

19. *In planta* quantification of *Magnaporthe oryzae* usingreal time PCR

Neelam Sheoran, G. Prakash, Asharani Patel and A. Kumar
Division of Plant Pathology, ICAR-Indian Agricultural Research Institute
New Delhi-110012

Introduction

Quantitative polymerase chain reaction (qPCR) can detect slow-growing, difficult-to-cultivate, or uncultivatable microorganisms, and can be used when traditional microbiological techniques are inadequate, ambiguous, time-consuming, difficult, and costly. Real Time PCR is based on the detection of the fluorescence produced by a reporter molecule which increases, as the reaction proceeds. This occurs due to the accumulation of the PCR product with each cycle of amplification. These fluorescent reporter molecules include dyes that bind to the double-stranded DNA (*i.e.* SYBR Green) or sequence specific probes (TaqMan Probes). The procedure follows the general principle of polymerase chain reaction; its key feature is that the amplified DNA is quantified as it accumulates in the reaction in real time after each amplification cycle. The real-time PCR assay can simultaneously detect and quantitate bacterial, fungal and viral pathogens. Real-time PCR can be a fast diagnostic tool and may be useful as an adjunct to identify potential pathogens.

Objective

To detect and quantify the pathogen in infected tissues using RT-PCR

Material required

Real Time PCR reagents; Target specific primers; cDNA of test samples; Micropipettes; Microfuge tubes; PCR tube/plate compatible with thermocycler; RT-PCR thermocycler; Analysis software

Protocol

Primer designing: RTm-PCR primers specific to gfp- gene are designed using online IDT Primer-Quest software available at http://eu.idtdna.comwith the following parameters: optimal length, 25 base pairs; GC content, 50-55%; melting temperature, 60°C; amplicon length, 100 to 160 base pairs; maximum self-complementarity at the 3′ end -five nucleotides, and absence of stable hairpins & dimers. Primer specificity and quality parameters to be checked with the help of Oligo-Analyzer (https://www.idtdna.com/calc/ analyzer*)*.

PCR amplification

Real-time PCR is performed using a 96-well reaction plate (LightCycler® 480 Multiwell Plate 96) and Light Cycler® 96 SW 1.1 (Roche Diagnostics, Switzerland). Each well contains a 20-µl reaction mixture that includes 10 µl of 2× SYBR Green PCR Master Mix (Light Cycler® 480 SYBR Green I Master), final primer concentration of 0.4 µM and three technical replicates with 15 ng of DNA templates. qPCR to be carried out according to the following protocol: denaturation at 95°C for 5 min, 40 repeats of 95°C for 10 s, 61°C for 15 s and 72°C for 15 s. A melting curve analysis will be conducted from 95°C for 10 s, 66°C for 60 s and 97°C for 1 s single time to confirm the amplification of single amplicon. PCAMBgfp vector and wild type *M. oryzae* can be used as positive and negative control in qPCR experiments in order to check the specificity of qPCR based detection assay.

Standardization of absolute quantitation of pathogen biomass

Real time PCR assay for absolute quantitation of *M. oryzae* is to be optimized using real time PCR primer pairs specific for gfp gene. To construct a standard curve, series of concentration of pCAMBgfp viz., 30000, 3000, 300, 30, 3, 0.3, 0.03, 0.003. 0.0003, 0.00003 pg to be mixed with 15ng of DNA isolated from healthy rice leaves. qPCR to be carried out as described above and a standard graph between amplification threshold values (Cq-values) and template DNA concentration prepared. Absolute biomass is quantitation to be estimated using the formula (http://cels.uri.edu/gsc/cndna.html) furnished below.

$$\textbf{Absolute quantification:}\quad \frac{[\text{DNA concentration X } 6.022\text{x}10^{23}]}{[\text{Product PCR size (bp) X } 10^{9} \text{ X } 650]}$$

RTm-PCR based pathogen quantitation by estimation of transgene copy number

Observation

Primers gfpMgF (5'-GGCCGATGCAAAGTGCCGATAAA-3') gfpMgR (5'-AGGGCGAAGAATCTCGTGCTTTCA-3') is expected to generate a specific 142-bp DNA product for *M. oryzaeRMg_Dl::gfp* isolate used in the study. RTm-PCR primers yielded specific signals of *M. oryzaeRMg_Dl::gfp* at 19th cycle whereas no amplification is expected with wild type *M. oryzae RMg_Dl*and water control. As far sensitivity, the RTm-PCR assay could detect pico gram or femto gram quantities of pathogen. The standard curve is constructed based on the DNA of *M. oryzae* versus the Ct value obtained in the RT PCR. Similarly we can quantify the presence of pathogen in infected plants at different time interval during pathogenesis. Sampling can be done at regular time intervals to study the pattern of pathogen behavior within the plant during disease infection.

Selected references

1. Qi, M. and Yand, Y. 2002. Quantification of *Magnaporthe grisea* During Infection of Rice Plants Using Real-Time Polymerase Chain Reaction and Northern Blot/Phosphoimaging Analyses. *Phytopathology.* 92(8): 870-876.
2. Schena, L., Nigro, F., Ippolito, A. and Gallitelli, D. 2004. Real-time quantitative PCR: a new technology to detect and study phytopathogenic and antagonistic fungi. *European Journal of Plant Pathology.* 110(9): 893–908.

Do you know?

The Earth Microbiome Project is a systematic attempt to characterize global microbial taxonomic and functional diversity for the benefit of the planet and humankind. The Earth Microbiome Project (EMP) is a massively collaborative effort to characterize microbial life on this planet. The Earth Microbiome Project was founded in 2010

20. Genome assembly tools

Neelam Sheoran, A.Kumar, G. Prakash, *V. Govindasamy and *K. Annapurna
Division of Plant Pathology& *Division of Microbiology,
ICAR-Indian Agricultural Research Institute, New Delhi-110012

Introduction

The first WGS assemblers, used for bacterial and viral genomes and for BAC clones, were Phrap (Phil Green), TIGR Assembler (Granger Sutton), and Cap3 (X. Huang). With the advent of second generation sequencers, most of which produced very large numbers of relatively short reads, new approaches to assembly had to be developed. Many of these new assemblers take an approach known as the de Bruijn graph to performing assemblies. This approach is attractive as it does not require all reads to be aligned to all other reads and it can compress redundant sequence.

Powerful computer algorithms are then utilized to piece the resulting sequence reads back together into longer continuous stretches of sequence (*contigs*), a process known as *de novo* assembly. For correct assembly, it is important that there is sufficient overlap between the sequence reads at each position in the genome, which requires high sequencing coverage (or read depth). Naturally, for longer sequence reads, more overlap can be expected, reducing the required raw read depth. Usually, longer fragments (several hundred base pairs) are sequenced from both ends (paired-end sequencing) to provide additional information on correct read placement in the assembly. After the initial assembly, *contigs* are typically joined to form longer stretches of sequence (known as *scaffolds*). To achieve this, libraries from long DNA fragments spanning several kilobases (kb) of sequence in the genome are prepared and their endpoints sequenced. Depending on the technology and the specifics of the library preparation, these libraries are (somewhat confusingly) called, for example paired-end, mate-pair or jump libraries.

Objective

To assemble and filter the data generated using two different platforms

Material required

Raw data generated by different platforms
Work station compatible to run Softwares (e.g. SPAdes, VALVET, SSPACE Basic (v2.0) etc)

Protocol

Data generation on HiSeq 2500

Magnaporthe oryzae; Tilletia indica

The data of 2 x 125 bp chemistry will be generated on HiSeq 2500. The raw reads generated filtered using Trimmomatic (v 0.35) with quality value QV > 30 and other contaminants such as adapters will be trimmed.

Parameters to be considered for filtration are as follows

1. Perform adapter trimming.
2. SLIDINGWINDOW: Perform a sliding window trimming of 20 bp, cutting once the average quality within the window falls below a threshold of 20.
3. LEADING: Cut bases off the start of a read, if below a threshold quality of 30
4. TRAILING: Cut bases off the end of a read, if below a threshold quality of 25
5. MINLENGTH: Drop the read if it is below 100 bp length

PacBio Data generated on PacBio RSII

The data of 5-8kb library size will be generated on PacBio platform using hairpin adaptor protocol for ultra-long read sequencing. Quality filtration of PacBio data is not required so data will be filtered for adapter sequences and processed data to be used for downstream analysis.

De Novo Assembly by hybrid approach

High quality paired end reads of illumina HiSeq 2500 and long reads of PacBio can be assembled using hybrid approach by SPAdes (Version: 1.5.2) with default parameter.

Ralstonia solanacearum

The data of 2 x 150bp chemistry will be generated on NextSeq 500. The raw reads generated will be filtered using Trimmomatic (v 0.35) with quality value QV > 30 and other contaminants such as adapters will be trimmed.

Parameters to be considered for filtration are as follows:

1. Adapter trimming to be performed.
2. SLIDINGWINDOW: Perform a sliding window trimming of 20 bp, cutting once the average quality within the window falls below a threshold of 20.
3. LEADING: Cut bases off the start of a read, if below a threshold quality of 30
4. TRAILING: Cut bases off the end of a read, if below a threshold quality of 25
5. MINLENGTH: Drop the read if it is below 100 bp length

De Novo Assembly and Scaffolding

The filtered high quality reads will be assembled into contigs using Velvet (v.1.2.10) on optimized kmer 127. Thus obtained primary assembly will be further optimized by scaffolding tool SSPACE Basic (v2.0), where primary assembly (velvet produced contigs) and paired end reads used.

Selected reference

1. Huang, X., Wang, J., Aluru, S.,Yang, S.P. and Hillier, L. 2003. PCAP: a whole-genome assembly program. *Genome Res.* 13(9): 2164-70.

Do you know?

NGS platform which can sequence directly RNA is NANOPORE!

21. Genetic transformation of bacterium for reporter gene expression

Neelam Sheoran, A. Kumar and *B. Ramakrishnan
Division of Plant Pathology& *Division of Microbiology,
ICAR-Indian Agricultural Research Institute, New Delhi-110012

Introduction

The gene which produces special phenotype and enables us for the differentiation of the cells that have this gene from those cells which do not have the gene is called as reporter gene. Reporter gene produces special protein molecules which are easily detectable as well as quantifiable and some time we can trace its moment in living organism also. There are two types of reporter or marker gene, scorable marker (*e.g*. Green fluorescent protein) and selectable marker (*e.g.* Antibiotics). Transformation is a biological process by which foreign DNA is introduced into a cell. Transformation of bacteria for the reporter gene is important not only for studies in bacteria but also for its interactions with other organism. There are several methods for transformation of reporter gene among them two are most important and widely used; they are tri-parental mating and electroporation. Both methods are described in this chapter.

Methodology

i. Triparental mating

Requirements

Shaking incubator at 37 °C; stationary incubator at 37 °C; micro-centrifuge tubes and sterile spreader; Luria Bertani agar plates (with appropriate antibiotics); LB broth.

Strains

1. *E. coli* with pBKminiTn7 gfp2 Gm10- Donor stain
2. *E. coli* with pUXBF13 Amp100- Helper strain
3. *Ralstonia solanacearum* Rif 50- Recipient strain

Procedure

1. Streaking of above bacterial strains into following LBA plates
   a. *E. coli* with pBKminiTn7 gfp2 Gm10- Donor stain (D) on LBA with Gentamycin 20 plates
   b. *E. coli* with pUXBF13 Amp100- Helper strain (H) on LBA with Ampicillin 100 plates
   c. *Ralstonia solanacearum*Rif 50- Recipient strain (R) on LBA Rifampicin 50 plates
2. Then incubate the above plates on 37°C for 24-48 hours
3. After appearance of colonies, subculture the single colony into 5 ml of LB broth media with respective antibiotic as in the LBA plates
4. Incubate on 37°C for 16-24 hours with shaking
5. Take 1ml of broth culture in a micro centrifuge tube and spin down at 8000g for 2 min
6. Wash 3 times in LB broth and spin down the pallet
7. Re-suspend the pellet in 600µl of LB broth
8. Now take 3 tubes and mix above bacterial suspension on following manner
   a. Mix 200 µl each of Recipient, Donor, and Helper (RDH)
   b. Mix 200 µl each of Donor and Helper and 200 µl of LB broth (DH Control)
   c. Mix 200 µl of Recipient and 400 µl of LB broth (R control)

9. Leave the mixture for 30 minutes on work bench or keep it at 42°C for 30 sec
10. Spin down at 8000g for 2min
11. Remove supernatant leaving around 20-30 µl liquid at the bottom
12. Mix it well to dissolve the pellet.
13. Take LB agar plate and place entire content at the center in the form of a droplet.
14. Allow it to dry on laminar and then incubate for overnight on 37°C
15. After incubation, scrape it and mix with 1ml of LBB
16. Take the LBA plate with double antibiotic, Gm 20+Rif 50 for spreading above mixture. Take 50 µl of mixture and spread with sterile spreader till it dry
17. Incubate at 37°C for 48-72 hours
18. Take observation.

Observation

Appeared colonies on double antibiotic plate will be transformed one. Control plates must not yield any colony. Transformed colonies may be further confirmed for gfp expression by fluorescent microscope.

ii. Electroporation Method

Requirements

Shaking incubator at 37°C; Stationary incubator at 37 °C; Electroporator; Ice bucket; Micro centrifuge tubes; electroporation tubes and sterile spreader; Luria Bertani agar plates (with appropriate antibiotics); LB broth.

Strains and DNA

*Ralstonia solanacearum*Rif 50- Recipient strain, Plasmid DNA having reporter gene (pBKminiTn7 gfp2 Gm10- Donor plasmid, pUXBF13 Amp100- Helper plasmid)

Procedure

Preparation of electro competent cells (ECC)

1. Streak the recipient bacterial strain on LBA Rifampicin 50 plates
2. Then incubate the above plates on 37°C for 24-48 hours
3. After appearance of colonies, subculture the single colony into 5 ml of LB broth media with Rifampicin 50
4. Incubate on 37°C for 16-24 hours with shaking
5. Take 1 ml of culture in each of four 1.5 ml tubes
6. Harvest the cells by centrifugation at 8000g for 2 min
7. Wash the pellet with 1ml of 300mM sucrose, repeat this step for three times
8. Resuspend the pellet in 25 µl of 300mM Sucrose
9. Four tubes pooled together to make 100 µl of electro competent cells aliquot.

Electroporation

1. Add 100ng of each plasmid DNA in the ECC aliquot
2. Mix well and transfer it into electroporation tube
3. Electroporate at 2500V for few milli seconds.
4. Add 1ml of LBB and transfer it into a new 1.5ml tube

5. Allow it to grow for 1 hour at 37°C
6. Harvest the cells by centrifugation at 8000g for 2 min
7. Discard the 800 µl of supernatant and mix the pellet in remaining 200 µl of media
8. Spread 50 µl in LBA+GM20+Rif50 plates (4 plates), also spread control plates without plasmid but subjected to electroporation same as above
9. Incubate at 37°C for 48-72 hours
10. Take observation

Observation: Appeared colonies on double antibiotic plate will be transformed one. Transformed colonies may be further confirmed for gfp expression by fluorescent microscope

Selected reference

1. Neelam Sheoran, Agisha Valiya Nadakkakath, Vibhuti Munjal, Aditi Kundu, Kesavan Subaharan, Vibina Venugopal, Suseelabhai Rajamma, Santhosh J Eapen and Kumar A (2015) Genetic analysis of plant endophytic *Pseudomonas putida* BP25 and chemo-profiling of its antimicrobial volatile organic compounds, *Microbiological Research* (Elsevier), 173: 66–78

Do you know?

The term genomics was coined by T.H. Roderick, a geneticist in 1986 at the Jackson Laboratory

22. Databases for microbial and pathogenomics

B. Ramakrishnan
Division of Microbiology, ICAR-Indian Agricultural Research Institute
New Delhi- 110012

Advancements in DNA sequencing technologies and their applications are rapidly increasing the generation of genomic data. The downward spiraling of costs for sequencing and speedy collection of genomic data and information on genes, protein sequences and structures, metabolites and reactions and pathways will outsmart the human handling and inferring from them in the future. The big data collection efforts necessitate the processing and storage for retrieval. The challenges of database construction, consolidation and sharing information and data mining are huge and demand for innovations. The genome information of more than 176,445 prokaryotes and about 7184 eukaryotes is currently available, and information on several other strains is being processed. Hence, the role of databases for managing and effectively utilizing these genomic data becomes very significant. The online database collection of *Nucleic Acids Research*(http://www.oxfordjournals.org/ourjournals/nar/database/a/) provides comprehensive information on more than 1,550 biological databases. These databases may contain similar information on the genomes but with heterogeneous types of data, often curated differently for various purposes and diverse tools for accessing them. The major categories of these databases are global resources, comprehensive databases, and special-purpose or community databases.

The National Center for Biotechnology Information (NCBI) at the National Institutes of Health in the United States of America and the European Molecular Biology Laboratory/European Bioinformatics Institute (EMBL-EBI) provide the global data collections on genomes of bacteria, archaea, eukaryotic microorganisms, and viruses. The data from GenBank, EMBL Bank or DNA Data Bank of Japan (DDBJ) are accessible by the text-based and sequence-based search engines. The significant feature of NCBI collections is the Reference Sequences or RefSeq, which are non-redundant (NR) databases of nucleotide and protein sequences and their curated subsets. Other important features are (i) BioProject which provides access to the primary data from sequencing projects, (ii) the repository of next generation data in the Sequence Read Archive (SRA), and (iii) the functional genomics data sets in GEO. Among the database search tools, BLAST is well-known for sequences and the sequence similarity against any taxonomic group, and metagenomic data. The NCBI Primer-BLAST helps to design and analyse PCR primers. Information on each taxonomic node of the described species of prokaryotes and several eukaryotes can be obtained from the NCBI Taxonomy database. The access to the genome sequence data on viruses is available in the NCBI Virus Variation resource.

The EMBL-EBI has databases on DNA, RNA, proteins, metabolites, and systems, cross-linked with several tools for accessing information. The important features of EMBL-EBI databases include Universal Protein Resource (UniProt) and its curated knowledge base UniProtKB, and the Metagenomics portal. The recent delugein genomic and metagenomic data has led to the continual changes in the maintenance of databases and the development of new access tools. The Genomes OnLine Database (GOLD) is a login-free, user-friendly open resource, launched in 1997, that has information sources of internal projects of Department of Energy Joint Genome Institute (DOE-JGI), those submitted by the external users and from public databases such as NCBI, EBI, and others. The four-level classification of GOLD includes Study, Biosample/Organism, Sequencing Project and Analysis Project. The overall objective of the sequencing projects is given at the GOLD Study. There are about 33,408 Studies and 215,881 Sequencing Projects in the current version of GOLD (v.7). In the GOLD project, the Digital Object Identifiers (DOIs) for the name, taxon, and exemplar for each organism are obtained from Names4Life (www.namesforlife.com) for accurate cross-referencing and comparative analysis. The GOLD proposes to implement machine learning approaches for metadata curation.

Integrated Microbial Genomes (IMG) data warehouse is a contemporary, comprehensive resource which integrates genomes and relative metadata such as proteomics, RNA-seq data sets and information biosynthetic clusters from bacteria, archaea, eukaryotic microorganisms, and viruses. Besides the IMG, the databases of xBase, Microbial Genome Database (MBGD), and the MicrobesOnline resource provide the means to analyse comparative genomic data. The genomic data on prokaryotic- and eukaryotic-microorganisms (fungi and protozoa) along with human, fruit fly, *Caenorhabditis elegans* and *Arabidopsis thaliana* from the Gene Trek in Prokaryote Space (GTPS) of DDBJ, and RefSeq of NCBI are available in MBGD for comparative analysis. The comparative analysis based on Ortholog analysis of MBGD offers the general or taxon-specific Ortholog groups using a hierarchical clustering program, DomClust.

The biological significance of raw genomic data is derived from the multistep process of the genome annotation. Several approaches are integrated into the databases for improving the quality of annotation and comparative analysis. The RAST (Rapid Annotation using Subsystems Technology) is used for automated microbial genome annotation. The MicroScope and COMBREX databases enable the investigators to curate the data collectively. In addition to the genome annotation and analysis, information on protein subcellular localizations for bacteria and archaea are available in PSORTdb. There are many special-purpose databases such as EcoGene, GenoBase, EcoCyc and PortEco for *Escherichia coli*; SubtiWiki and SporeWeb for *Bacillus subtilis*; ProPortal for *Prochlorococcus*; CyaoBase for cyanobacteria; Rhizobase for nitrogen-fixing bacteria such as *Azoarcus*, *Azospirillum*, *Klebsiella* and *Frankia*; and Saccharomyces Genome Database for *Saccharomyces cerevisiae*.

The Pathosystems Resource Integration Center (PATRIC) has genomic data and others including transcriptomics, protein-protein interactions, sequence typing of more than 10,000 bacteria of NIAID category A to C/emerging/reemerging pathogens. Likewise, the Pathogen portal (http://pathogenportal.org) has resources on bacterial and eukaryotic pathogens. The annotation database of GeneDB deals with pathogenic bacteria, eukaryotes, and viruses. The ViralZone knowledge base is an important resource for molecular and epidemiological information. The Eukaryotic Pathogen Database Resources (EuPathDB) has individual bases of specific eukaryotic pathogens including pathogenic amoeba, *Cryptosporidium*, *Toxoplasma*, *Giardia*, *Trypanosoma* and *Plasmodium*, accessible through a common portal. Information of *Candida*, *Aspergillus*, *Cryptococcus* and other fungi are available in the FungiDB. The aligned and annotated rRNA gene sequences from bacteria, archaea, and fungi are available in the Ribosomal Database Project (RDP) while the Bacterial Diversity Metadatabase (BacDive) has information on taxonomy, physiology, sampling, and environmental conditions. The Global Catalog of Microorganisms (GCM) is another important database on microorganisms. The volume of genomic data from microorganisms, especially bacterial phyla which are predicted to be about 1500 against the currently recognized 89 will be enormous. Hence, the innovative construction and evaluation of databases is essential to assist the users to make the right choice for data mining.

23. Recipe for basic molecular biology reagents

DNA extraction and quantification

Extraction buffer (Total nucleic acid)
2% w/v CTAB
1.4 M NaCl
50 mM EDTA pH 8.0
100 mM Tris-HCl pH 8.0
0.2% β-mcercaptoethanol (add just before use)

5 M NaCl
Dissolve 292.2 g of NaCl in H_2O and make upto 1 liter. Dispense into aliquots sterilize by autoclaving.

0.5 M EDTA (pH 8.0)
Add 186.1 g of sodium ethylene diamine tetra acetate $2H_2O$ to 500 ml of H_2O. Stir vigorously on a magnetic stirrer. Adjust the pH to 8.0 with NaOH. Make upto 1 litre. Dispense into aliquots and sterilize by autoclaving.

1 M Tris-HCl
Dissolve 121.1 g of Tris hydroxylmethane anminomethane in 800 ml distilled H_2O. Adjust pH to 8.0 with concentrated hydrochloric acid. Make upto 1 liter. Dispense into aliquots. Sterilize by autoclaving.

3M Sodium acetate (pH 4.8 and 5.2)
Dissolve 408.1 g of sodium acetate $3H_2O$ in 800 ml of distilled water. Adjust the pH to 4.8 to 5.2 with glacial acetic acid. Make upto 1 litre. Dispense into aliquots. Sterilize by autoclaving.

50X TAE
Dissolve 242 g of Tris in H_2O. Add 57.1 ml of glacial acetic acid and add 100 ml of 0.5M EDTA (pH 8.0). Make upto 1 liter. Dispense into aliquots. Sterilize by autoclaving.

Ethidium Bromide (10 mg/ml)
Add 1 g of ethidium bromide to 100 ml of H_2O. Stir on a magnetic stirrer for several hours to ensure that the dye has dissolved. Wrap the container in aluminium foil or transfer to a dark bottle and store at 4ºC.

Caution: Ethidium bromide is a mutagen and toxic. Wear gloves when working with ethidium bromide solutions and a mask when weighing it out.

Cloning and Transformation

100 mM $CaCl_2$
Add 1.47 g of $CaCl_2.2H_2O$ in H_2O. Make upto 100 ml. Sterilize by autoclaving. Store at -20°C.
100mM PIPES (piperazine-1,2-bis[2-ethanesulfonic acid]**)**
Dissolve 1.51g PIPES in 50 ml of DD water. Store at -20°C.

LA Medium (Luria agar medium)

950 ml of deionized water, add 10 g of bactotryptone, 5 g of yeast extract and 10 g of NaCl and 15 g of agar. Boil to dissolve content, adjust pH to 6.8 with 5 N NaOH and make volume upto 1 litre with deionized water. Sterilize by autoclaving.

Luria Bertani broth

It is same as above but prepared without agar.

X-gal (5-bromo-4chloro-3-indolyl-β-D-galactopyranoside)

Dissolve X-gal in dimethyl formamide to make a 20 mg/ml solution and store at –20ºC.

IPTG (Isopropyl thio-β-galactoside)

Dissolve 2.0 g of IPTG in 8 ml of distilled H_2O. Make upto 10 ml/ Filter sterilize through a 0.22 micron disposable filter. Dispense into aliquots. Store at –20ºC.

Antibiotics

Dissolve 5 mg of ampicillin in 10 ml of sterile double distilled water. Filter-sterilize and dispense into 200 µl aliquots and store at –20ºC.

Media and antibiotic stocks for fungal/bacterial transformation

<u>**Stock solution**</u>

K-buffer (pH-7.0)

K_2HPO_4- 200g/l
KH_2PO_4- 145g/l
Autoclaved and stored at 4°C

MN solution

MgSO4.7H_2O- 30g/l
NaCl-15g/l
Autoclaved and stored at 4°C

20% Glucose- Dissolve 20g glucose in 100ml of autoclaved molecular biology grade water, filter sterilized and stored at 4°c.

20% $(NH4)_2SO_4$: Dissolve 20g of Ammonium sulphate in 100ml of autoclaved molecular biology grade water and stored at 4°c.

1% $CaCl_2.4H_2O$- 20mg of Calcium chloride tetra hydrate was dissolved in 2ml of autoclaved MOLECULAR BIOLOGY GRADE WATER, filter-sterilize and stored at 4C.

0.1% $FeSO_4.7H_2O$- 2 mg of ferric sulphate hepta hydrate was dissolved in 2 ml of autoclaved MOLECULAR BIOLOGY GRADE WATER, filter sterilize and store at 4C.

0.4 M MES (pH- 5.3)
MES-7.808g
MOLECULAR BIOLOGY GRADE WATER-100ml
Autoclaved and stored at 4°C.

1M Glucose- glucose 18g, MOLECULAR BIOLOGY GRADE WATER- 100ml, filter sterilize stored at 4° C.

0.2 M Acetosyringone- Acetosyringone- 39.24 mg, DMSO- 1 ml, Filter sterilize and stored at -20°C

Preparation of induction media

K- buffer- 10ml/l
MN Sol- 20ml/l
20% $(NH_4)_2$ SO_4-2.5ml/l
Glucose-3.6g/l
1% $CaCl_2.2H_2O$- 1 ml/l
0.1% $FeSO_4.7H_2O$- 1ml/l
0.4 MES-100 ml/l
50% glycerol- 10 ml/l
Autoclaved and cooled to approximately 55° C. Then add 200µl acetosyringone to it.

Antibiotic stocks

Cefotaxime (125mg/ml): Dissolve 125mg of cefotaxime in 1 ml of autoclaved MOLECULAR BIOLOGY GRADE WATER and filter sterilize the solution prior to use, store it at -20°C.

Kanamycin (50mg/ml): Dissolve 50 mg of kanamycin monosulphate in 1 ml of autoclaved MOLECULAR BIOLOGY GRADE WATER , filter-sterilize the solution prior to use, store it at -20°C.

Ampicillin (50mg/ml): Dissolve 50mg of ampicillin in 1 ml of autoclaved MOLECULAR BIOLOGY GRADE WATER, filter sterilize the solution prior to use and store the stock at -20°C.

Rifampicin (10mg/ml): Dissolve 10mg of rifampicin in 1 ml of methanol and store the stock at -20°C.

Chloramphenicol (100mg/ml): Dissolve 100 mg of chloramphenicol in 1 ml of ethanol, filter sterilize and stored at -20°C.

Do you know?

The first genome was sequenced by Frederick Sanger. He sequenced the complete genome (5368 bp) of a bacteriophage Φ-X174 by dideoxy chain termination method in 1977

24. Useful online resources for genomics research

Bacteriology related web-resources

1. http://www.atlasplantpathogenicbacteria.it/
2. http://www.isppweb.org/names_bacterial.asp
3. http://www.bacterialgenomics.org/
4. http://www.brinkman.mbb.sfu.ca/
5. http://bamics3.cmbi.ru.nl/
6. https://genome.jgi.doe.gov/
7. http://it.umich.edu/projects/sitemaker/
8. https://wwwnc.cdc.gov/eid/
9. http://microbes.ucsc.edu/

MLST web resources

1. http://genome.ppws.vt.edu/cgi-bin/MLST/home.pl
2. http://pubmlst.org/
3. http://www.mlst.net/
4. http://www.pasteur.fr/recherche/genopole/PF8/mlst/
5. http://mlstdb.hku.hk:14206/MLST_index.html
6. http://genome.ppws.vt.edu/cgi-bin/MLST/docs/MLSTMLSA.pl

Culture Collections and Genome information

1. http://www.jcvi.org
2. http://www.bacterio.cict.fr/
3. http://www.dsmz.de/
4. https://www.uv.es/cect
5. https://www.lgcstandards.com/IN/en

Bioinformatic web resources

1. http://tolweb.org/tree/phylogeny.html
2. http://rdp.cme.msu.edu/index.jsp
3. http://lowelab.ucsc.edu/GtRNAdb/
4. http://mamit-trna.u-strasbg.fr/
5. *http://*www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi
6. http://www.ridom.de/doc/Ridom_16S_rDNA_tour.pdf
7. http://www.tolweb.org/tree/phylogeny.html
8. http://www.genomesonline.org/index2.htm
9. http://insilico.ehu.es/

Genome databases

1. http://www.ncbi.nlm.nih.gov/BLAST/
2. http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?mode=Root
3. http://www.ncbi.nlm.nih.gov/taxonomy
4. http://www.ncbi.nlm.nih.gov/genomeprj
5. http://www.ncbi.nlm.nih.gov/unigene

6. http://www.ddbj.nig.ac.jp/welcome-e.html
7. http://www.ebi.ac.uk/genomes
8. https://genome.jgi.doe.gov
9. http://www.phytopathdb.org/

Molecular Biology Protocols and cloning vectors

1. http:// http://molbiol.ru
2. http://www.ndsu.edu/pubweb/~mcclean/plsc431/cloning/clone3.htm
3. https://www.addgene.org/protocols/
4. https://www.neb.com//media/nebus/files/brochures/cloning_tech_guide.pdf
5. https://www.jove.com/science-education/5074/molecular-cloning

Institutions involved in genome research

1. https://www.broadinstitute.org
2. http://guides.lib.berkeley.edu/Plant-and-Microbial-Biology
3. http://www.nipgr.res.in/home/home.php
4. http://www.jcvi.org/cms/home/
5. https://www.csiro.au/

Whole Genome database

1. http://mbgd.genome.ad.jp/
2. http://fungi.ensembl.org/Puccinia_graminis/Info/Index
3. http://fungi.ensembl.org/Saccharomyces_cerevisiae/Info/Index
4. http://fungi.ensembl.org/Aspergillus_nidulans/Info/Index
5. http://alternaria.vbi.vt.edu/index.html
6. https://fungi.ensembl.org/Magnaporthe_oryzae/Info/Index

Tools used in Omics

1. https://omictools.com/

Metagenomic data analysis

1. http://metagenomics.anl.gov/
2. https://www.ebi.ac.uk/metagenomics/
3. https://omictools.com/metagenomics-category

Practical Manual

Genomics of Plant Pathogens and Agriculturally Important Microbes

December 19th to 31th, 2018

Course Director

Dr. C. Viswanathan

Course Coordinators

Dr. A. Kumar
Dr. K. Annapurna

Published by

NAHEP- Center for Advanced Agricultural Science and Technology
©Indian Council of Agricultural Research

Organized by
Center for Advanced Agricultural Science and Technology (CAAST)

प्रो. एम एस स्वामीनाथन पुस्तकालय
Prof. M S SWAMINATHAN LIBRARY