World Bank – ICAR Funded
NATIONAL AGRICULTURAL HIGHER EDUCATION PROJECT
Centre for Advanced Agricultural Science and Technology (CAAST) on Genomics-Assisted Crop Improvement and Management

Training Manual

"GENOMICS OF AGRICULTURALLY IMPORTANT INSECTS"

(18th -28th September, 2019)

Division of Entomology, ICAR-IARI, New Delhi

Course Director

DR. S. SUBRAMANIAN
Principal Scientist,
Division of Entomology, ICAR-IARI New Delhi -12
e-mail: subramanian@iari.res.in

Course Coordinators

Dr. Sagar, D
Scientist,
Division of Entomology,
ICAR-IARI, Pusa campus,
New Delhi -110012
e-mail: sagar@iari.res.in

Dr. A. Kumar
Principal Scientist,
Division of Plant Pathology,
ICAR-IARI, Pusa Campus,
New Delhi-110012
Email: kumar@iari.res.in

Organised by

NAHEP- Center for Advanced Agricultural Science and Technology
ICAR-Indian Agricultural Research Institute, New Delhi
http://nahep-caast.iari.res.in/

ABOUT NAHEP-CAAST at ICAR- IARI, NEW DELHI

Centre for Advanced Agricultural Science and Technology (CAAST) is a new initiative and student centric subcomponent 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 19.99 crores for the project 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:

1. **To develop online teaching facility and online courses for enhancing the teaching and learning efficiency, and scientific communications skills**
2. **To develop and/or strengthen state-of-the art next-generation genomics and phenomics facilities for producing quality PG and Ph.D. students**
3. **To develop collaborative research programmes with institutes of international repute and industries in the area of genomics and phenomics**
4. **To enhance the skills of faculty and PG students of IARI and NARES**
5. **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.

Core-Team Members:

S.No.	Name of the Faculty	Discipline	Institute
1.	Dr. Ashok K. Singh	Genetics	ICAR-IARI
2.	Dr. Vinod	Genetics	ICAR-IARI
3.	Dr. Gopala Krishnan S	Genetics	ICAR-IARI
4.	Dr. A. Kumar	Plant Pathology	ICAR-IARI
5.	Dr. T.K. Behera	Vegetable Science	ICAR-IARI
6.	Dr. R.N. Sahoo	Agricultural Physics	ICAR-IARI
7.	Dr. Alka Singh	Agricultural Economics	ICAR-IARI
8.	Dr. A.R. Rao	Bioinformatics	ICAR-IASRI
9.	Dr. R.C. Bhattacharya	Molecular Biology & Biotechnology	ICAR-NIPB
10.	Dr. K. Annapurna	Microbiology **Nodal officer, Grievance Redressal, CAAST**	ICAR-IARI
11.	Dr. R. Roy Burman	Agricultural Extension **Nodal officer, Equity Action Plan, CAAST**	ICAR-IARI
12.	Dr. K.M. Manjaiah	Soil Science & Agri. Chemistry **Nodal officer, CAAST**	ICAR-IARI
13.	Dr. Viswanathan Chinnusamy	Plant Physiology **PI, CAAST**	ICAR-IARI

Associate Team

S.No.	Name of the Faculty	Discipline	Institute
14.	Dr. Kumar Durgesh	Genetics	ICAR-IARI
15.	Dr. Ranjith K. Ellur	Genetics	ICAR-IARI
16.	Dr. N. Saini	Genetics	ICAR-IARI
17.	Dr. D. Vijay	Seed Science & Technology	ICAR-IARI
18.	Dr. Kishor Gaikwad	Molecular Biology & Biotechnology	ICAR-NIPB
19.	Dr. Mahesh Rao	Genetics	ICAR-NIPB
20.	Dr. Veena Gupta	Economic Botany	ICAR-NBPGR
21.	Dr. Era V. Malhotra	Molecular Biology & Biotechnology	ICAR-NBPGR
22.	Dr. Sudhir Kumar	Plant Physiology	ICAR-IARI
23.	Dr. Dhandapani R	Plant Physiology	ICAR-IARI
24.	Dr. Lekshmy S	Plant Physiology	ICAR-IARI
25.	Dr. Madan Pal	Plant Physiology	ICAR-IARI
26.	Dr. Shelly Praveen	Biochemistry	ICAR-IARI
27.	Dr. Suresh Kumar	Biochemistry	ICAR-IARI
28.	Dr. Ranjeet R. Kumar	Biochemistry	ICAR-IARI
29.	Dr. S.K. Singh	Fruits & Horticultural Technology	ICAR-IARI
30.	Dr. Manish Srivastava	Fruits & Horticultural Technology	ICAR-IARI
31.	Dr. Amit Kumar Goswami	Fruits & Horticulture Technology	ICAR-IARI
32.	Dr. Srawan Singh	Vegetable Science	ICAR-IARI
33.	Dr. Gograj S Jat	Vegetable Science	ICAR-IARI
34.	D. Praveen Kumar Singh	Vegetable Science	ICAR-IARI
35.	Dr. V.K. Baranwal	Plant Pathology	ICAR-IARI
36.	Dr. (Ms.) Deeba Kamil	Plant Pathology	ICAR-IARI
37.	Dr. Vaibhav K. Singh	Plant Pathology	ICAR-IARI
38.	Dr. Uma Rao	Nematology	ICAR-IARI
39.	Dr. S. Subramanium	Entomology	ICAR-IARI
40.	Dr. M.K. Dhillon	Entomology	ICAR-IARI
41.	Dr. B. Ramakrishnan	Microbiology	ICAR-IARI
42.	Dr. V. Govindasamy	Microbiology	ICAR-IARI
43.	Dr. S.P. Datta	Soil Science & Agricultural Chemistry	ICAR-IARI
44.	Dr. R.N. Padaria	Agricultural Extension	ICAR-IARI
45.	Dr Satyapriya	Agricultural Extension	ICAR-IARI
46.	Dr. Sudeep Marwaha	Computer Application	ICAR-IASRI
47.	Dr. Seema Jaggi	Agricultural Statistics	ICAR-IASRI

48.	Dr. Anindita Datta	Agricultural Statistics	ICAR-IASRI
49.	Dr. Soumen Pal	Computer Application	ICAR-IASRI
50.	Dr. Sanjeev Kumar	Bioinformatics	ICAR-IASRI
51.	Dr. S.K. Jha	Food Science & Post Harvest Technology	ICAR-IARI
52.	Dr. Shiv Dhar Mishra	Agronomy	ICAR-IARI
53.	Dr. D.K. Singh	Agricultural Engineering	ICAR-IARI
54.	Dr. S. Naresh Kumar	Environmental Sciences; **Nodal officer, Environmental Management Framework**	ICAR-IARI

ACKNOWLEDGEMENTS

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. Joint Director (Extension), ICAR- IARI, New Delhi
11. Coordinator, School of Crop Protection, ICAR- IARI New Delhi
12. Head, Division of Entomology, ICAR- IARI, New Delhi
13. Head, Division of Plant Pathology), ICAR- IARI, New Delhi
14. Head, Division of Microbiology, ICAR- IARI, New Delhi
15. ICAR- Indian Institute of Horticultural Research, Bengaluru
16. ICAR- National Research Centre for Plant Biotechnology, New Delhi
17. ICAR- National Bureau of Agricultural Insect Resources, Bengaluru
18. ICAR- Indian Agricultural Statistical Research Institute, New Delhi
19. Nucleome Biosciences, Hyderabad

PREFACE

Molecular biological tools have redefined the contours of entomological research worldwide in the recent years. Several global research initiatives like, Manhattan Project on Entomology - I5K (an attempt to cover genomics of 5000 insects) have given new impetus to insect genomics considering the impact of insect pests in agriculture and public health which have a direct bearing on the welfare of mankind. The emerging problems of invasive pests, resurgence of sucking pest complex in various field crops, xenobiotic resistance in crop pests to insecticides and vector transmission of plant pathogens could be better tackled with a focused research thrust by strategic use of biotechnological tools. Insect molecular biological approaches offer better understanding of molecular basis of insect nutrition, host defense and behavioural physiology of pests and natural enemies. They provide strategic research support to the existing entomological research programmes to resolve conflicts in taxonomic identity of crop pests, handling xenobiotic resistance in transgenic crop systems, developing natural enemies with improved tolerance to biotic and abiotic stresses.

Insects are the largest group of animals replete with genomics databases. As on date whole genome data is available for 138 insects, transcriptomes of 116 insects, gene sets of 61 insects, 36 gene families of 60 insects, 7,544 miRNAs of 69 insects, 96,925 piRNAs from two insects, 22,536 pathways of 78 insects, 679,881 untranslated regions (UTR) of 84 insects and 160,905 coding sequences (CDS) of 74 insects. Unravelling such a big data of genomic information of inputs require sophisticated bioinformatics analytics. Molecular approaches coupled with bioinformatics analyses offer scope for gene mining the data bases for identifying novel target sites for next generation insecticides and bio rational pesticides.

Hence, it is the need of the hour to capacitate the entomological students in handling of molecular biological tools and impart exposure to insect genomics to address the ever growing insect problems in a sound manner. The training on Insect genomics of agriculturally important insects is a small beginning towards the path to Insect Genomics.

S. SUBRAMANIAN
SAGAR, D
A. KUMAR

29th July, 2019

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. ICAR- IARI has made significant contributions in the field of insect genomics and molecular biology. Pioneering efforts by Dr. N. Ramakrishnan on physical mapping of Baculovirus genomes of insect pests during 1980s laid strong foundation for insect molecular biological research in the Division of Entomology, IARI, New Delhi. Genetic engineering of baculoviruses through deletion of EGT gene, characterization of 20 hydroxy ecdyzones of insect pests, baseline studies on Bt resistance, molecular basis of xenobiotic resistance in crop pests, elucidation of Bt resistance genes in bollworm pests and molecular characterization of gut bacterial symbionts of insect pests are some of the notable research forays in the domain of Insect genomics by this Institute. Barcoding of insects are done in a big way to augment the digital database of National Pusa Collections, one of the oldest insect collections in this country.

Presently research efforts are underway on metagenomics of insect pests, RNA interference strategies for pest management and molecular characterization of insecticide/fumigant resistance in insects utilizing genomic tools. Courses like Insect Genetics and Molecular Biology are offered by the School of Post Graduate studies enable capacity building of students on cutting edge molecular biological and genomic techniques. Students are exposed to molecular biological and genomic tools to enhance their research outputs through custom designed research programmes

In order to harness the potential of genome information available on a number of Insect Genomic databases like, Fly Base, InsectBase *etc*., we need to create appropriate infrastructure facilities and human resources to face the challenges of pest management in the coming decades. With this background the Centre for Advanced Agricultural Science and Technology (CAAST) under NAHEP is organizing a 10 days Short Training course on "Genomics of Agriculturally important Insects" for the benefit of students of SAUs and Universities. I am sure that the training will impart the basics of Insect genomics to the Post Graduate students of entomology.

Dean and Joint Director (Education)
ICAR-IARI, New Delhi

Date: 29th July 2019

NAHEP-CAAST

ICAR-Indian Agricultural Research Institute, New Delhi 110 012
Genomics of Agriculturally Important Insects (18-28 September 2019)
Venue for lectures: Virology Auditorium, ICAR, IARI, New Delhi
Venue for Practicals: IBDC Unit, Division of Entomology, ICAR-IARI, New Delhi

Training Schedule

Day 1	**Wednesday, 18th September, 2019**
9.00-10.00 h	Registration
10.00-11.30 h	**Inauguration Programme:**
	Lighting of Lamp
	Welcome Address: Dr. K. M. Manjaiah, Associate Dean, P. G. School
	About NAHEP-CAAST: Dr. C. Viswanathan, Project Investigator, CAAST
	About the Training Programme: Dr. S. Subramanian, Course Director
	Students' Self introduction
	Remarks: Dr. Rashmi Aggarwal, Dean, ICAR- IARI, New Delhi
	Remarks: Dr. A. K. Singh, Joint Director (Res), ICAR- IARI, New Delhi
	Inaugural Address: Dr. A. K. Singh, Director, ICAR- IARI, New Delhi
	Vote of Thanks: Dr. Sagar, D., Course Coordinator
	Photo session
11.00-11.30h	Tea
11.30-11.45 h	**Pre Training Evaluation** Dr. S. Subramanian and Dr. D. Sagar
12.00-13.00h	Lecture 1: Insect Genomics – An overview (**Dr. S. Subramanian**)
14.15-17.00h	Practical 1: DNA isolation from insects (**Dr. Sagar, D., IARI, New Delhi**)
Day 2	**Thursday, 19th September, 2019**
9.30 -10.00h	Lab – Preparation /Observation /Mini Exercises
10.00-11.00h	Lecture 2: Genomic sequencing: An Overview **(Dr. Kishore Gaikwad, NIPB, New Delhi)**
11.00-11.15h	Tea
11.15 -13.00h	Lecture 3: Bioinformatic tools for whole genome sequencing - An introduction [Lecture cum Demonstration] (**Dr. A. Kumar, IARI, New Delhi**)
14.15-17.00h	Practical 2: RNA isolation from insects (**Dr. Sagar, D., IARI, New Delhi**)

Day 3	**Friday, 20th September, 2019**
9.30 -10.00h	Lab – Preparation /Observation /Mini Exercises
10.00-11.00h	Lecture 4: Metagenomic approaches: An overview (**Dr. K. Annapurna, IARI, New Delhi**)
11.00-11.15h	Tea
11.15-12.15 h	Lecture 5: Molecular phylogeny – Principles and practices (**Dr. Anirban Roy, IARI, New Delhi**)
12.15-13.15	Lecture 6: PCR – Real Time PCR **(Dr. A. Kumar, IARI, New Delhi)**
14.15-17.00h	Practical 3: RNA isolation and cDNA synthesis in insects (**Dr. Sagar, D., Dr. S. Subramanian and Ms. Rajna**)
Day 4	**Saturday, 21st September, 2019**
9.30 -10.00h	Lab – Preparation /Observation /Mini Exercises
10.00-11.00h	Lecture 7: Gene Silencing and Genome editing in insect pest management (**Dr. R. Asokan, IIHR, Bengaluru**)
11.00-11.15h	Tea
11.15-13.00h	Lecture 8: DNA Barcoding of insects – Principles and practices (**Dr. T. Venkatesan, NBAIR, Bengaluru**)
14.15-17.00h	Practical 4: PCR techniques for insects (**Dr. S. Subramanian and Dr. Sagar, D., IARI, New Delhi**) Practical 5: Generating barcodes for insects (**Dr. Naresh M. Meshram and Dr. P.R. Shashank, IARI, New Delhi**) Practical 6: Use of PCR for prey-predator relations (**Dr. Sachin Suroshe, IARI, New Delhi**)
	Sunday 22nd September, 2019 - HOLIDAY
Day 5	**Monday, 23rd September, 2019**
9.30 -10.00h	Lab – Preparation /Observation /Mini Exercises
10.00-11.00h	Lecture 9. House- keeping genes for expression analysis:- Insect case studies (**Dr. B. Ramcharan, NIPB, New Delhi**)
11.00-11.15h	Tea
11.15-13.00h	Lecture 10: Insect genomics – Challenges and solutions (**Dr. Dushyant Singh Baghel, Nucleome Informatics, Hyderabad**)
14.15-17.00h	Practical 7: Evaluation of RNAi constructs for insect pests (**Dr. Vinay K. Kalia**)
Day 6	**Tuesday, 24th September, 2019**
9.30 -10.00h	Lab – Preparation /Observation /Mini Exercises
10.00-11.00h	Lecture 11: Transcriptome sequencing – strategies and approaches (**Dr. Ranjeet Ranjan, IARI, New Delhi**)
11.00-11.15h	Tea

11.15-13.00h	Lecture 12: Enzyme kinetics – An over view (**Dr. Anil Dahuja, IARI, New Delhi**)
14.15-17.00h	Practical 8: Genotyping of insecticide resistance (**Dr. S. Subramanian and Dr. Suresh M. Nebapure, IARI, New Delhi**)
Day 7	**Wednesday, 25th September, 2019**
9.30 -10.00h	Lab – Preparation /Observation /Mini Exercises
10.00-11.00h	Lecture 13: Computational tools for gene annotation (**Dr. A. R. Rao, IASRI, New Delhi**)
11.00-11.15h	Tea
11.15-13.00h	Lecture 14: Quantitative PCR – principles and practices (**Dr. B. Ramakrishnan, IARI, New Delhi**)
14.15-17.00h	Practical 9: Handling of raw sequences, curation, and assembly / Gene annotation and preparation of data for accessioning (**Dr. A. Kumar, Dr. V. Govindasamy, Baskaran IARI, New Delhi**)
Day 8	**Thursday, 26th September, 2019**
9.30 -10.00h	Lab – Preparation /Observation /Mini Exercises
10.00-11.00h	Lecture 15: The Genomic Basis of Nematode Parasitology (**Dr. Uma Rao, IARI, New Delhi**)
11.00-11.15h	Tea
11.15-13.00h	Lecture 16: Molecular Markers for Entomological Research (**Dr. S. Mohankumar, TNAU**, **Coimbatore**)
14.15-17.00h	Practical 10: Molecular phylogeny (**Dr. Anirban Roy & Dr. A. Kumar, IARI, New Delhi**)
Day 9	**Friday, 27th September, 2019**
9.30 -10.00h	Lab – Preparation /Observation /Mini Exercises
10.00-11.00h	Lecture 17: RNAi approaches for pest management (**Dr. Vinay K. Kalia, IARI, New Delhi**)
11.00-11.15h	Tea
11.15-12.15h	Lecture 18: Rearing of insects for insect Genomics (**Dr. Kirti Sharma and Dr. S. Subramanian, IARI, New Delhi**)
12.15 – 13.15 h	Visit to Laboratories –Division of Entomology (**Dr. Sagar, D**.)
14.15-17.00h	Practical 11: Visit to Phenomics Facility: Dr. C. Viswanathan, IARI, New Delhi Visit to Sequencing Facility: Dr. B. Ramcharan, NIPB, New Delhi Visit to Bioinformatic Facility: Dr. A.R. Rao, ICAR-IASRI, New Delhi

Day 10	**Saturday, 28th September, 2019**
9.30 -10.00h	Lab – Preparation /Observation /Mini Exercises
10.00-11.00h	Lecture 19: Plant-Parasitic Nematode Genomics: An Update (**Dr. Vishal Somvanshi, IARI, New Delhi)**
11.00-11.15h	Tea
11.15-13.00h	**Interaction session & Post Training evaluation**
14.15-17.00h	**Valediction and Certificate distribution: CAAST Team**

CONTENTS

Chapter-1

Glossary of terminologies used in Insect Genetics, Genomics and Molecular Biology

D. Sagar, Rajna S. and S. Subramanian
Division of Entomology, ICAR-Indian Agricultural Research Institute, New Delhi
garuda344@gmail.com

A-DNA Right-handed helical form of DNA found in fibers at 75% RH in presence of sodium, potassium or cesium. The bases are tilted with regard to the helical axis and there are more base pairs per turn. The A-form may be very similar to the conformation adopted by DNA-RNA hybrids or by RNA-RNA double stranded regions.

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, DDBJ).

Acentric A chromosome, or chromosome fragment, that lacks a centromere.

Additive genes Genes that interact but do not show dominance or epistasis.

Adenine A purine and one of the nitrogenous bases found in DNA and RNA.

Agarose A polysaccharide gum obtained from agar, which is obtained from certain seaweeds, used as a gel medium in electrophoresis; used to separate DNA molecules on the basis of their molecular weight.

Allele One of two or more alternative forms of a gene at a particular locus. If more than two alleles exist, the locus is said to exhibit multiple allelism.

Allopatry Populations that are physically or geographically isolated eventually may change sufficiently through natural selection or drift that pre- or postmating isolation mechanisms develop, leading eventually to speciation.

Allozyme Allozymes are a subset of isozymes. Allozymes are variants of enzymes representing different allelic alternatives of the same locus.

Alternative splicing Gene regulation by means of alternative splicing of exons to produce different amounts of protein or even different proteins.

Amino acid One of the monomeric units that polymerize to make a protein molecule.

Aminoacyl tRNA synthetase Enzymes that catalyze the attachment of each amino acid to the appropriate transfer RNA molecule. A tRNA molecule carrying its amino acid is called a charged tRNA.

Amplification The production of additional copies of a chromosomal sequence, found as either intrachromosomal or extrachromosomal DNA.

Anchored PCR A modification of PCR that allows amplification in situations in which only one sequence is known that is suitable for a primer (rather than two).

Aneuploid A condition in which the chromosome number of an organism is not an exact multiple of the typical haploid set for the species.

Anneal The process by which the complementary base pairs in the strands of DNA combine.

Anticodon The triplet of nucleotides in a transfer RNA molecule that is complementary to and base pairs with a codon in a messenger RNA.

Antiparallel The DNA strands are parallel but point in opposite directions.

Apomorphic A character that is derived and not ancestral.

Apoptosis Programmed cell death, is a series of programmed steps that cause a cell to die via "self digestion" without rupturing and releasing intracellular contents into the surrounding environment.

Arrhenogenic A sex-determining system in which females produce male progeny only.

Asymmetric PCR Single-stranded DNA produced by providing an excess of primer for one of the two DNA strands. Single-stranded DNA produced can be sequenced directly without cloning.

ATP Adenosine triphosphate is the primary molecule for storing chemical energy in a cell.

Autoregulatory control Regulation of the synthesis of a gene product by the product itself. In some systems, excess gene product behaves as a repressor and binds to the operator of its own structural gene.

Autosomes All chromosomes except the sex chromosomes. Each diploid cell has two copies of each autosome.

B chromosomes B chromosomes are nonvital supernumerary chromosomes found in many plant and animal species, thought to be derived from one of the normal chromosomes, and are often transmitted at higher rates than expected, thus exhibiting "drive."

B-DNA A helical form of DNA formed by adding water to dehydrated A-DNA. B-DNA is the form of DNA of which Watson and Crick constructed their model in 1953. It is found in very high relative humidity. This form is thought to prevail in the living cell.

Back mutation Mutations that occur to reverse a point mutation to the original condition.

Base pair (bp) Two nucleotides that are in different strands of nucleic acid and whose bases pair by hydrogen bonding. In DNA, adenine pairs with thymine and guanine pairs with cytosine.

Bioinformatics Researchers in bioinformatics develop computer software applications that can store, compare, and analyze the very large quantities of DNA sequence data generated by the new genome technologies.

Biotechnology The manipulation of organisms to provide desirable products for human use.

BLAST Basic Local Alignment Search Tool (BLAST) is a computer program widely used to search large databases of DNA or amino acid sequences, providing sequences that have regions of similarity to the sequence(s) of interest provided by the user.

Blunt end An end of a DNA molecule, at which both strands terminate at the same nucleotide position with no extension of one of the strands.

Bootstrapping A statistical method based on repeated random sampling with replacement from an original sample to provide a collection of new estimates of a parameter, from which confidence limits can be calculated.

Cell The fundamental unit of life; each multicelled organisms is composed of cells; cells may be organized into organs that are relatively autonomous but cooperate in the functioning of the organism.

Central Dogma The Central Dogma was proposed by F. Crick in 1958. It states that the genetic information is contained in DNA, which is transcribed into RNA, which is translated into polypeptides.

Centromere A region of a chromosome to which spindle fibers attach during mitosis and meiosis.

Chaperones Protein molecules that assist with correct protein folding as the protein emerges from the cell's ribosome. Heat shock protein 70, heat shock protein 40, and chaperonins are examples.

Chelating agent A molecule capable of binding metal atoms; one example is EDTA, which binds Mg^{2+}.

Chimeric DNA Recombinant DNA containing DNA from two different species.

Chitin A water-insoluble polysaccharide that forms the exoskeletons of arthropods and crustaceans.

Chromatids Chromosome components that have duplicated during interphase become visible during the prophase stage of mitosis. Chromatids are held together at the centromere.

Chromomere A region on a chromosome consisting of densely packed chromatid fibers that produce a dark band on polytene chromosomes.

Chromosomes Units of the genome with many genes, consisting of histone proteins and a very long DNA molecule; found in the nucleus of every eukaryote.

Clade An evolutionary lineage derived from a single stem species. A branch of a cladogram.

Cladistic systematics Systematics that use only shared and derived characters as a basis of constructing classifications.

Cladogenic speciation Branching evolution of new species.

Cladogram A term used two ways by different authors. Either a dendrogram (tree) produced using the principle of parsimony, or a tree that depicts inferred historical relationships between organisms.

Clone A population of identical cells often containing identical recombinant DNA molecules. Also a group of organisms produced from one individual cell through asexual processes.

cloning vector A DNA molecule capable of replicating in a host organism; a gene is often inserted into it to construct a recombinant DNA molecule, and the vector is then used to amplify (clone) the recombinant DNA.

Cluster analysis A method of hierarchically grouping taxa or sequences on the basis of similarity or minimum distance. UPGMA is an unweighted pair group method using the arithmetic average.

Coding strand The strand of the DNA molecule that carries the biological information of a gene and that is transcribed by RNA polymerase into pre-mRNA.

Co dominant Alleles whose gene produces are both manifested in the heterozygote.

Codon A triplet of nucleotides that codes for a single amino acid.

Colony hybridization The use of *in situ* hybridization to identify bacterial colonies carrying inserted DNA that is homologous with a particular sequence (the probe).

Competent cells Bacterial cells in a state in which exogenous DNA molecules can bind and be internalized, thereby allowing transformation.

Complementary base pairing Nucleotide sequences are able to base pair; A and T are complementary

Complementary DNA (cDNA) An ss DNA that is complementary to a strand of RNA. The DNA is synthesized by an enzyme called reverse transcriptase. It is a DNA copy of the messenger RNA.

cDNA library A collection of clones containing dsDNA that is complementary to the mRNA. Such clones will lack introns and regulatory regions of eukaryotic genes.

Constitutive enzymes Enzymes that are part of the basic permanent machinery of the cell. They are formed consistently in constant amounts regardless of the metabolic state of the organism.

Constitutive heterochromatin Regions of the chromosome containing mostly highly repeated, noncoding DNA; usually near the telomeres and centromeres.

Contig Segments of DNA that partially overlap in their sequence are called contigs.

Convergent evolution The evolution of unrelated species resulting in structures with a superficial resemblance.

Copy number The number of plasmids in a cell; the number of genes, transposons, or repetitive elements in a genome.

Core DNA The DNA in the core nucleosome that is wrapped around the histone octamer. The core nucleosome is connected to others by linker DNA.

Crossing over The reciprocal exchange of polynucleotides between homologous chromosomes during meiosis.

Cytochrome The complex protein respiratory enzymes occurring within plant and animal cells in the mitochondria, where they function as electron carriers in biological oxidation.

Cytoplasm The components of the cell *not* including the nucleus.

Cytoplasmic incompatibility Reproductive incompatibility between two populations caused by factors that are present in the cytoplasm.

Cytoplasmic sex-ratio distorters Cytoplasmic genes that manipulate the sex ratio of their host to promote their own spread. Microbes (*Wolbachia*, spiroplasmas, viruses) often are transovarially and transstadially transmitted that can alter the sex ratios of insects and mites.

Cytosine A pyrimidine, one of the bases in DNA and RNA.

Cytosol The fluid portion of the cytoplasm, excluding the organelles in a cell.

Dalton A unit of mass very nearly equal to that of a hydrogen atom.

Degeneracy Refers to the genetic code and the fact that most amino acids are coded for by more than one triplet codon.

Degenerate codons Two or more codons that code for the same amino acid.

Degenerate primers Degenerate primers can be used for the PCR when a limited portion of a protein sequence is known for a gene, but the DNA sequence is not known.

Deletion The loss of a portion of the genetic material from a chromosome. The size can vary

from one nucleotide to sections containing many genes.

Denaturation Breakdown of secondary and higher levels of structure of proteins or nucleic acids by chemical or physical means.

Denatured DNA DNA that has been converted from double- to single-stranded form by a process such as heating.

Dendrogram A branched diagram that represents the evolutionary history of a group of organisms.

Deoxyribonuclease An enzyme that breaks a DNA polynucleotide by cleaving phosphodiester bonds.

Deoxyribonucleic acid (DNA) The genetic information.

Deuterotoky A form of parthenogenesis in which unfertilized eggs can develop into either males or females.

Dicentric A chromosome or chromatid with two centromeres.

Dideoxy sequencing Developed by F. Sanger and A. R. Coulson in 1975, and known as the "plus and minus" or "primed synthesis" method of DNA sequencing.

Diploid Having two copies of each chromosome.

DNA arrays DNA arrays work by hybridization of labeled RNA or DNA in solution to DNA molecules attached at specific locations on a surface.

DNA binding protein Proteins such as histones or RNA polymerase that attach to DNA as part of their function.

DNA–DNA hybridization A method for determining the degree of sequence similarity between DNA strands from two different organisms by the formation of heteroduplex molecules.

DNA ligase An enzyme that repairs single stranded discontinuities in double-stranded DNA. DNA ligases also are used in constructing recombinant DNA molecules.

DNA polymerase An enzyme that catalyzes the formation of DNA from dNTPs, using single stranded DNA as a template. Three different DNA polymerases (I, II, and III) have been isolated from *E. coli.*

DNA polymerase I The enzyme in *E. coli* that completes synthesis of individual Okazaki fragments during DNA replication.

DNA polymerase III The enzyme that primarily functions in DNA replication of *E. coli*.

DNA probe Also called a gene probe or genetic probe. Short, specific (complementary to the

desired DNA sequence), artificially produced segments of labeled DNA are used to combine with and detect the presence of a specific gene or DNA sequence within the chromosome. The presence of this labeled probe usually is detected visually.

DNA sequencing Determining the order of nucleotides in a DNA molecule.

DNase Deoxyribonuclease, an enzyme that degrades DNA.

Dominant A gene is dominant when it produces the same phenotype whether it is heterozygous or homozygous.

Double helix The base-paired structure consisting of two polynucleotides in the natural form of DNA.

Draft sequence This term has had several definitions, but generally refers to sequence that is not yet finished but is of generally high quality

ds DNA Double-stranded DNA.

Electrophoresis The separation of molecules in an electric field. Electrophoresis can be used to separate proteins or DNA molecules.

Electroporation A process used to introduce DNA into the genome of an organism.

Endonuclease An enzyme which degrades nucleic acid molecules by cleaving phosphodiester bonds internally.

Endosymbiosis Microorganisms, including bacteria, rickettsia, mycoplasmas, viruses, and yeasts, live within the cells of many eukaryotic organisms including insects.

Enhancer Sequences of DNA that can increase transcription of neighboring genes over long distances up or downstream of the gene and in either possible orientation.

Enhancer trap A method to identify genes based on their pattern of expression.

Epistatic Epistasis is the nonreciprocal interaction of nonallelic genes. A gene epistatic to another masks the expression of the second gene.

EST Abbreviation for expressed sequence tags.

Ethidium bromide A dye that binds to double stranded DNA by intercalating between the stands. DNA stained with EtBr fluoresces under UV illumination.

Euchromatin Regions of a eukaryotic chromosome that appear less condensed and stain less well with DNA-specific dyes than other segments of the chromosome.

Eukaryote An organism with cells containing a membrane-bound nucleus that reproduces by meiosis. Cells divide by mitosis. Oxidative enzymes are packaged within mitochondria.

Exogenous DNA DNA from an outside source. In genetic engineering, DNA from one organism is often inserted into another by a variety of methods.

Exon One of the coding regions of a discontinuous gene.

Exonuclease A nuclease, which degrades a nucleic acid molecule by progressive cleavage along its length, beginning at the 3' or 5' end.

Expression vector Vectors that are designed to promote the expression of gene inserts.

Extrachromosomal gene A gene not carried by the cell's chromosomes, such as mitochondrial or plasmid-borne genes.

F statistics A set of coefficients that describe how genetic variation is partitioned within and among populations and individuals, such as FST and inbreeding coefficient.

F_1 hybrid The first-generation offspring of a cross between two different strains.

Facultatively heterochromatic Chromosomal material that, unlike euchromatin, shows maximal condensation in nuclei during interphase. Constitutive heterochromatin is composed of repetitive DNA, is late to replicate, and is transcriptionally inactive. Portions of the chromosome that are normally euchromatic may become heterochromatic at a particular developmental stage (= facultative heterochromatin). An example of facultative heterochromatin is the inactivated X chromosomes in the diploid somatic cells of mammalian females.

Fingerprinting DNA fingerprinting relies on the presence of simple tandem-repetitive sequences that are present throughout the genome. The regions show length polymorphisms, but share common sequences.

Flanking sequence A segment of DNA that or follows the region of interest on the molecule.

FLP recombinase Yeast FLP recombinase is able to catalyze recombination in which a DNA segment that is flanked by direct repeats of FLP target sites (FRTs) can be excised from the chromosome. If two homologous chromosomes each bear an FRT site, mitotic recombination can occur in *Drosophila*, leading to the introduction of DNA into known, and specific, sites. FRT sites can be introduced into *Drosophila* chromosomes by P-element mediated transformation.

Foldback DNA DNA that contains palindromic sequences that can form hairpin double stranded structures when denatured DNA is allowed to renature.

Forward genetics Analysis of the phenotype or function leads to identification of interesting mutants, which might be used to analyze a particular process or clone the genes responsible for regulating this process.

Forward mutation Amutation from the wild type to the mutant.

A **back mutation** restores the wild-type phenotype.

Frameshift mutation A mutation resulting from inserting or deleting a group of nucleotides that is not a multiple of three, so that the polypeptide produced will probably have a new set of amino acids specified for downstream of the frameshift.

FST Coancestry coefficient; a measure of the relatedness of individuals.

Functional genomics Study of what traits/ functions are conferred on an organism by specific DNA sequences. Typically functional genomics occurs after the DNA sequences have been identified.

Fusion protein A hybrid protein molecule produced when a gene of interest is inserted into a vector and displaces the stop codon for a gene already present in the vector. The fusion protein begins at the amino end with a portion of the vector protein sequence and ends with the protein of interest.

G-banding Dark bands on chromosomes produced by Giemsa staining; G-bands occur in A-T rich regions of the chromosome.

Gap genes Gap gene mutants lack large areas of the normal cuticular pattern. Three wild type gap genes, *Krüppel+*, *hunchback+*, and *knirps+*, regionalize the embryo by delimiting domains of homeotic gene expression and effect position-specific regulation of the pair-rule genes.

Gating The process of shutting off a function when the value of a specific parameter attains a critical level.

Gel electrophoresis Separation of molecules on the basis of their net electrical charge and size.

Gene A segment of DNA that codes for an RNA and/or a polypeptide molecule. It includes regions preceding and following the coding region, as well as introns.

Gene amplification The production of multiple copies of a DNA segment in order to increase the rate of expression of a gene carried by the segment. The chorion genes of *Drosophila* are amplified in the ovary.

Gene boundaries Boundaries between active and inactive chromatin occur along the chromosomes. Such boundaries are established by **insulators** that act as a neutral barrier to the influence of neighboring elements.

Gene cloning Insertion of a fragment of DNA containing a gene into a cloning vector and subsequent propagation of the recombinant DNA molecule in a host organism. Recently, cloning of a DNA fragment by the polymerase chain reaction has simplified the technology.

Gene conversion A genetic process by which one sequence replaces another at an orthologous or paralogous locus, resulting in concerted evolution. May result from mismatch repair.

Gene duplication The duplication of a DNA segment coding for a gene; gene duplication produces two identical copies which may retain their original function allowing the organism to produce larger amounts of a specific protein. Alternatively, one of the gene copies may be lost by mutation and become a pseudogene, or a duplicated gene can evolve to perform a different task.

Gene expression The process by which the information carried by a gene is made available to the organism through transcription and translation.

Gene gun A method for propelling microscopic particles coated with DNA into cells, tissues, and organelles to produce stable or transient transformation.

Gene library A collection of recombinant clones derived from genomic DNA or from the cDNA transcript of an mRNA preparation. A complete genetic library is sufficiently large to have a high probability of containing every gene in the genome.

Gene regulation The mechanisms that determine the level and timing of gene expression.

Gene targeting A technique for inserting changes into a genetic locus in a desired manner. The desired locus is transferred into an embryo by microinjection where it is allowed to undergo homologous recombination into the chromosomes, replacing the original allele.

Gene transfer The movement of a gene or group of genes from a donor to a recipient organism.

Genetic code The rules that determine which triplet of nucleotides code for which amino acid during translation. There are more than 20 different amino acids and four bases (adenine, thymine, cytosine, and guanine). There are 64 potential combinations of the four bases in triplets (4×4×4). A doublet code would only be able to code for 16 (4×4) amino acids. Since only 20 amino acids exist, there is redundancy in the system so that some amino acids are coded for by two or three different triplets (codons).

Genetic distance A measure of the evolutionary divergence of different populations of a species, as indicated by the number of allelic substitutions that have occurred per locus in the two populations. The most widely used measure of genetic distance is that of Nei (1972), $D=-\ln(I)$.

Genetic diversity (**GST**) Variation in populations averaged over different loci.

Genetic engineering The deliberate modification of genes by man. Also called gene splicing, gene manipulation, recombinant DNA technology.

Genetic linkage Genes are located together on the same chromosome.

Genetic marker An allele whose phenotype is recognized and which can be used to monitor the inheritance of its gene during genetic crosses between organisms with different alleles.

Genetic sex determination system The mechanism in a species by which sex is determined. In most organisms, sex is genetically, rather than environmentally, determined.

Genome The total complement of DNA in an organism.

Genomic footprinting A technique for identifying a segment of a DNA molecule in a living cell that is bound to some protein of interest.

Genomic imprinting The process by which some genes are found to function differently when they are transmitted by the mother rather than the father, or vice versa. Mechanisms of imprinting may include methylation of the DNA. The more a gene is methylated, the less likely it is to be expressed.

Genomic library A random collection of DNA fragments from a given species inserted into a vector (plasmids, phages, cosmids). The collection must be large enough to include all the unique nucleotide sequences of the genome.

Genomics The study of genome data. The complete DNA sequences of organisms such as the human, mouse, rat, zebrafish, *D. melanogaster*, *C. elegans*, and *Arabidopsis thaliana* can provide a plethora of information on entire families of genes and whole pathways of interacting proteins. See also functional genomics, proteomics, and structural genomics.

Genotype The genetic constitution of an organism. The phenotype of the organism is its appearance or observable character.

Glycosylation A process in which a sugar or starch is linked to a protein molecule.

GMO Genetically modified organism.

Guanine A purine in one of the nucleotides in DNA and RNA.

Haploid Cells or organisms that contain a single copy of each chromosome.

Hardy–Weinberg equilibrium An equilibrium of genotypes achieved in populations of infinite size in which there is no migration, selection, or mutation after at least one generation of panmictic mating. With two alleles, *A* and *a*, of frequency *p* and *q*, the Hardy– Weinberg equilibrium frequencies of the genotypes *AA*, *Aa*, and *aa* are $p2$, $2pq$, and $q2$, respectively.

Helicase The enzyme responsible for breaking the hydrogen bonds that hold the double helix together so that replication of DNA can occur.

Helix A spiral staircase-like structure with a repeating pattern.

Helper plasmid A plasmid that is able to supply something to a defective plasmid, thus enabling the defective plasmid to function.

Heritability In the **broad sense** (hB =VG/VP), the fraction of the total phenotypic variance that remains after exclusion of the variance due to environmental effects. In the **narrow sense**, the ratio of the additive genetic variance to the total phenotypic variance (VA/VP).

Hermes A transposable element that has been engineered for transforming insects other than *Drosophila*. *Hermes* was discovered in the house fly *Musca domestica*.

Heterochromatin The regions of the chromosome that have large amounts of noncoding repetitive DNA.

Heteroduplex DNA A hybrid DNA–DNA molecule formed from tracer and driver from different individuals or species.

Heterogametic sex The sex that produces gametes containing unlike sex chromosomes. Many males are XY and thus heterogametic. Lepidopteran females are the heterogametic sex. Crossing over is often suppressed in the heterogametic sex.

Heterogeneous nuclear ribonucleoproteins (hnRNPs) Pre-mRNAs and mRNAs are associated with a set of at least 20 proteins throughout their processing in the nucleus and transport to the cytoplasm. Some of these hnRNPs contain nuclear export signals.

Heterologous DNA DNA from a species other than that being examined.

Heterologous recombination Recombination between two DNA molecules that apparently lack regions of homology.

Heteroplasmy The coexistence of more than one type of mitochondrial DNA within a cell or individual.

Heterosis Also known as hybrid vigor.

Heterozygosity Having a pair of dissimilar alleles at a locus (eg., Aa); a measure of genetic variation in a population estimated by a single locus or an average over several loci. **Heterozygote** A diploid cell or organism that contains two different alleles of a particular gene.

Highly repetitive DNA DNA made up of short sequences, from a few to hundreds of nucleotides long, which are repeated on an average of 500,000 times.

Histones Basic proteins that make up nucleosomes and have a fundamental role in chromosome structure.

Hogness box A DNA sequence 19–27 bp upstream from the start of a eukaryotic structural gene to which RNA polymerase II binds. The sequence is usually 7 bp long (TATAAAA); named in honor of D. Hogness. Often called TATA box and pronounced "tah-tah."

Holocentric Chromosomes that have diffuse centromeres.

Homeobox A conserved DNA sequence about 180 bp in size found in a number of homeotic genes involved in eukaryotic development. Homeobox genes (genes to which the homeobox is attached) are those genes that are for embryonic development.

Homeotic The replacement of one serial body part by a serially homologous body part.

Homeotic gene Genes that determine the identification and sequence of segments during embryonic development in insects. Although most genes with a homeo domain are in the homeotic class, a few are found among the segmentation genes. Homeotic genes have been described in a variety of insects other than *Drosophila*, including *Musca, Aedes, Anopheles, Blatella*, and *Tribolium.*

Homeotic mutations Mutations in which one developmental pattern is replaced by a different but homologous one. Homeotic mutations of *Drosophila* and other insects cause an organ to differentiate abnormally and form a homologous organ that is characteristic of an adjacent segment. Examples in *Drosophila* include *aristapedia* in which the antenna becomes leg like, and *bithorax* in which halteres are changed into wing like appendages.

Homoduplex DNA molecules A double-stranded DNA molecule in which the two strands come from different sources in DNA–DNA hybridization. Heteroduplex DNA will denature or melt into single strands at lower temperatures than homoduplex DNA from a single source.

Homogametic sex The sex that produces gametes with only one kind of sex chromosome. The females of many insects are XX and thus homogametic.

Homologous chromosomes Two or more identical chromosomes.

Homologous genes Two genes from different organisms and therefore of different sequence that code for the same gene product.

Homology Homology has been defined as "having a common evolutionary origin," but also is often used to mean "possessing similarity or being matched."

Homoplasy Phenomena that lead to similarities in character states for reasons other than inheritance from a common ancestor, including convergence, parallelism, and reversal.

Homozygous Diploid cells or organisms that contain two identical alleles of a particular gene.

Horizontal gene transfer The transfer of genetic information from one species to another. Mechanisms and frequency are not well understood in insects.

Hot-start PCR Hot start is a method to optimize the yield of desired PCR product and to suppress nonspecific amplification. This is done by withholding an essential component of the PCR, such as the DNA polymerase, until the reaction mixture has been heated to a temperature that inhibits nonspecific priming and primer extension. See also polymerase chain reaction (PCR).

Housekeeping genes Genes whose products are required by the cell for normal maintenance.

Humoral immunity The immune system response that consists of soluble blood serum components that fight an infection.

Hybrid dysgenesis A syndrome of genetic abnormalities that occurs when hybrids are formed between strains of *Drosophila melanogaster*, one carrying (P) and the other lacking (M) the transposable *P* element. The abnormalities include chromosomal damage, lethal and visible mutations, and sometimes sterility. Dysgenesis is caused by crossing Pmales×M females, but the reciprocal cross is not dysgenic.

Hybridization probe A labeled nucleic acid molecule used to identify complementary or homologous molecules through the formation of stable base pairs.

Hydrogen bonding A hydrogen bond is a weak electrostatic attraction between an electronegative atom (such as oxygen or nitrogen) and a hydrogen atom attached to a second electronegative atom. In effect, the hydrogen atom is shared between the two electronegative atoms.

Hypertranscription Transcription of DNA at a rate higher than normal. For many species with an XY sex-determination system, the male compensates for his single X chromosome by hypertranscribing the X chromosome. He produces a nearly equal amount of gene product compared to what is produced by females with two X chromosomes.

Imaginal discs Cells set off during embryonic development that will give rise, during the pupal stage, to adult organs.

In silico biology *In silico* biology refers to the use of computers to perform biological studies.

***In situ* hybridization** The pairing of complementary DNA and RNA strands, or the pairing of complementary DNA single strands to produce a hybrid molecule in intact chromosomes or cells.

***In vitro* packaging** The production of infectious particles by enclosing naked DNA in lambda (λ) phage packaging proteins and preheads.

Inbreeding coefficient The correlation of genes within individuals (FIT), or the correlation of genes within individuals within populations (FIS). Both FIS and FIT are measures for deviation from expected Hardy–Weinberg proportions.

Indel An insertion or deletion in a DNA sequence.

Independent assortment See Law of Independent Assortment.

Inducible enzymes Enzymes whose rate of production is increased by the presence of certain molecules.

Initiation codon AUG serves as an initiation codon when it occurs at the start of a gene; it marks the site where translation should begin. AUG also codes for methionine, so most newly synthesized polypeptides will have this amino acid at the amino terminus, although it may later be removed by posttranslational processing of the protein. AUG is the only codon for methionine, so AUGs that are not initiation codons are also found in the middle of a gene.

Insertion mutation Alteration of a DNA sequence by inserting one or more nucleotides.

Insertion sequences Insertion sequences are the simplest transposable elements, carrying no genetic information except what is needed to transpose (i.e., transposase). Usually 700–2500 bp long, denoted by the prefix IS and followed by the type number.

Insertion vectors Vectors that have a single target site at which foreign DNA is inserted.

Insulators Novel sequence elements found recently in *Drosophila* that are associated with boundaries between active and inactive genes, protecting against position. Insulators act as a neutral barrier against both positive and negative effects of the chromosomal environment.

Intercalating agent A chemical compound which is able to invade the space between adjacent base pairs of a double-stranded DNA molecule; including ethidium bromide.

Intergenic region The noncoding region between segments of DNA that code for genes.

Interphase The stage of the cell cycle when chromosomes are not visible by light microscopy. During interphase, DNA synthesis occurs.

Introgression The incorporation of genes of one species into the gene pool of another. If the ranges of two species overlap and fertile hybrids are produced, they will tend to backcross with the more abundant species.

Intron A region of eukaryotic DNA coding for RNA that is later removed during splicing; it does not contribute to the final RNA product.

Inverse PCR Inverse PCR allows amplification of an unknown DNA sequence that flanks a "core" region with a known sequence. The basic method for inverse PCR involves digesting template DNA, circularizing the digested DNA, and amplifying the flanking DNA outside the core region with the primers oriented in the opposite direction of the usual orientation. Primers for inverse PCR are synthesized in the opposite orientation and are homologous to the ends of the core region so thatDNA synthesis proceeds across the *uncharacterized* region of the circle rather than across the characterized core region.

Inversion Alteration of the sequence of a DNA molecule by removal of a segment followed by its reinsertion in the opposite orientation.

Inverted repeat Two identical nucleotide sequences repeated in opposite orientation in a DNA molecule, either adjacent to one another or some distance apart.

Ion channels The membrane passages that allow certain ions to cross the membrane.

Ionic selectivity The ability of ion channels to permit certain ions to cross the membrane, but not others.

Isozymes (isoenzymes) Multiple forms of an enzyme that differ from each other in their substrate affinity, in their activity, or in their regulatory properties. Isozymes are complex proteins of paired polypeptide subunits. They often have different isoelectric points and can be separated by electrophoresis.

Jumping genes Genes that move within the genome, usually because they are associated with transposable elements.

Junk DNA The proportion of DNA in a genome that *apparently* has no function. Also called parasitic or selfish DNA.

Kilobase A kilobase (kb) of DNA=1000 nucleotides.

Kilodalton (kDa) A unit of mass equal to 1000 daltons (Da). One dalton is nearly equal to the mass of a hydrogen atom.

Kin selection A theory put forth by W. D. Hamilton (1964) that states that an altruistic act is favored because it increases the inclusive fitness of the individual performing the social act.

Klenow fragment A portion of bacterial DNA polymerase I derived by proteolytic cleavage. It lacks the 5′-to-3′ exonuclease activity of the intact enzyme.

Lagging strand The DNA strand in the double helix which is copied in a discontinuous manner during DNA replication; short segments of DNA produced during the replication are called Okazaki fragments.

Lambda or λAdouble-stranded DNAvirus (bacteriophage) that can invade *E. coli*. Once inside the cell λ can enter a lysogenic cycle or a lytic cycle of replication, which results in death of the host cell. λ has been genetically engineered as a vector for cloning. λ is also a microliter unit of measurement, the volume contained in a cube 1 mm on a side.

Law of Independent Assortment One of Mendel's laws. The members of different pairs of factors assort independently. Different pairs of alleles assort independently into gametes during gametogenesis, if they are on different chromosomes. The subsequent pairing of male and female gametes is at random, which results in new combinations of alleles.

Law of Segregation One of Mendel's laws. The factors of a pair of characters segregate. Separation into different gametes, and thus into different progeny, of the two members of each pair of alleles possessed by the diploid parent.

Leader sequence An untranslated segment of mRNA from its 5 end to the start codon.

Leading strand The DNA strand in the double helix which is copied in a continuous fashion during DNA replication.

Lethal mutation Mutation of a gene to yield no product, or a defective gene product, resulting in the death of the organism because the gene product is essential to life.

Leucine zipper DNA binding proteins that contain four to five leucine residues separated from each other by six amino acids. The leucines on two protein molecules interdigitate and dimerize in a specific interaction with a DNA recognition sequence. Leucine zippers are involved in regulating gene expression.

Library A set of cloned DNA fragments which represent the entire genome.

Ligase DNA ligases are enzymes that catalyze the formation of a phosphodiester bond between adjacent 3′-OH and 5′-PO_4 termini in single stranded DNA. DNA ligases function in DNA repair to seal single-stranded nicks between adjacent nucleotides in a double-stranded DNA molecule.

Ligation Enzymatic joining together of nucleic acid molecules through their ends.

Likelihood methods Likelihood methods of analyzing DNA sequence data rely on genetic models and provide a basis for statistical inference. Maximum likelihood methods of tree construction assume the form of the tree and then choose the branch length to maximize the likelihood of the data given that tree. These likelihoods are then compared over different possible trees and the tree with the greatest likelihood is considered to be the best estimate.

Linkage A linkage group is a group of genes located on a single chromosome.

Linkage map A diagram of the order and relative distances between gene loci on chromosomes, based on the frequency of recombination of the linked genes in the genomes of progeny obtained from crossing parents with different genetic markers.

Linker DNA The DNA that links nucleosomes; the function of linker DNA is unresolved.

Locus The position of a gene on a chromosome; Plural: loci.

Long germ band development A pattern of development in insects such as *D. melanogaster* in which the pattern of segmentation is established by the end of blastoderm.

Lysis The process of disintegrating a cell, which involves rupturing the membranes, breaking up the cell wall and nuclear membrane.

Lysogenic During the lysogenic phase of a bacteriophage, the DNA of a virus is integrated into the chromosome of its bacterial host.

Lytic A virus in a lytic phase undergoes intracellular multiplication, and lysis of the bacterial host cell results.

Major groove The larger of the two grooves that spiral around the surface of the double helix of the DNA molecule.

Map unit In linkage maps, a 1% recombination frequency is defined as a map unit or one centimorgan. A number proportional to the frequency of recombination between two genes.

Mariner A transposable element that has been engineered for transforming insects other than *Drosophila*. *mariner* elements are widely found in arthropods and in insect parasitic nematodes, other nematodes, flatworms, hydras, humans, mouse, rat, Chinese hamster, sheep, and cows. *mariner* has been used to transform chicken, zebrafish, and a protozoan.

Marker (DNA size marker) A DNA fragment of known size used to calibrate an electrophoretic gel.

Marker (genetic) A trait that can be observed to occur (or not) in an organism. Marker genes include genes conferring resistance to antibiotics, expression of green fluorescent protein, eye color, etc.

Maternal effect gene Genes with a maternal effect are genes in the mother which have an effect on the phenotype of her progeny. Usually the result of depositing products or maternally derived mRNAs in the egg that are used or transcribed by the embryo.

Maternally inherited Characters that are transmitted primarily by cytoplasmic genetic factors (including mitochondria, viruses, some mRNAs) derived solely from the maternal parent. Also known as cytoplasmic inheritance or extranuclear heredity.

Maxam and Gilbert sequencing method A "chemical" method to sequence DNA developed in 1977 by A. M. Maxam and W. Gilbert. Single-stranded DNA derived from double stranded DNA and labeled at the 5′ end with 32P is subjected to several chemical cleavage protocols to selectively make breaks on one side of a particular base. The fragments are separated by size by electrophoresis on acrylamide gels and identified by autoradiography.

Maximum parsimony methods Taxonomic methods that focus on the character values observed and minimizing the number of changes in character state between species over the tree, making the assumption that there have been approximately constant rates of change. The changes at each node in the tree are inferred to be those that require the least number of changes to give each of the two character states of the immediate descendants.

Median melting temperature The temperature at which 50% of the double helices have denatured; the midpoint of the temperature range over which DNA is denatured.

Meiosis The sequence of events occurring during two cell divisions to convert diploid cells into haploid cells.

Meiotic drive Any mechanism that results in the unequal recovery of the two types of gametes produced by a heterozygote.

Melting of DNA Melting DNA means to denature it by heat, breaking the hydrogen bonds that hold the two strands together.

Messenger RNA (mRNA) RNA molecules which code for proteins and which are translated on the ribosomes.

Methylation In bacteria, enzymes (modification methylases) that bind to the DNA attach methyl groups to specific bases. This methylation pattern is unique to and protects the species from its own restriction endonucleases. Methylation also occurs in eukaryotes and may be involved in genomic imprinting. Genes that are methylated are less likely to be active.

M13 bacteriophage A single-stranded bacteriophage cloning vehicle, with a closed circular DNA genome of approximately 6.5 kb. M13 produces particles that contain ss DNA that is homologous to only one of the two complementary strands of the cloned DNA and therefore is particularly useful as a template for DNA sequencing.

M13 universal primer A primer derived from the M13 bacteriophage is used for sequencing reactions and has been used to identify satellite DNA sequences in many organisms.

Microsatellite DNA Pieces of the same small segment which are repeated many times.

Minor groove. The smaller of the two grooves that spiral around the surface of the DNA double helix.

Minos A transposable element that has been engineered for transforming insects other than *Drosophila*. *Minos* has a wide host range and can transform human cell lines, making it potentially useful for mutagenesis and analysis of the human genome.

Mitochondrion An organelle that occurs in the cytoplasm of all eukaryotes, capable of self-replicating.

Mitosis The sequence of events that occur during the division of a single cell into two daughter cells.

Mobile genetic element See transposable element.

Moderately repetitive DNA Nucleotide sequences that occur repeatedly in chromosomal DNA. Repetitive DNA is moderately (=middle) repetitive or highly repetitive. Highly repetitive DNA contains sequences of several nucleotides repeated millions of times. It is a component of constitutive heterochromatin. Middle-repetitive DNA consists of segments 100–500 bp long repeated 100 to 10,000 times each. This class also includes the genes transcribed into tRNAs and rRNAs.

Molecular biology A term broadly used to describe biology devoted to the molecular nature of the gene and its biochemical reactions such as transcription and translation.

Molecular clock The hypothesis that molecules evolve in direct proportion to time so that differences between molecules in two different species can be used to estimate the time elapsed since the two species last shared a common ancestor.

Molecular evolution That subdivision of the study of evolution that studies the structure and functioning of DNA at the molecular level over time.

Molecular genetics Genetic studies that focus on the molecular nature of genes and gene expression.

Molecular phylogeny An analysis of the relationships of groups of organisms as reflected by the evolutionary history detected in molecules (proteins, DNA).

Molecular systematics The detection, description, and explanation of molecular diversity within and among species.

Morphogen Molecules whose local concentration directly determines the local pattern of differentiation.

mRNA Messenger RNA.

mtDNA Mitochondrial DNA.

Monoclonal antibody A single antibody produced in quantity by cultured hybridoma cell lines.

Muller's ratchet The accumulation of deleterious mutations that can lead to extinction of a population of a sexual species.

Multigene family A group of genes that are related either in nucleotide sequence or in terms of function; they are often clustered together.

Multiple-locus, multiple-allele model A model for sex determination in Hymenoptera.

Multiplex PCR When more than one pair of primers is used in a PCR, multiple segments of target DNA can be amplified simultaneously and thus conserve template, save time, and minimize expense. See also polymerase chain reaction (PCR).

Mutagen A chemical or physical agent able to induce a mutation in a DNA molecule.

Mutant An organism expressing the effects of a mutated gene in its phenotype.

Mutation A change in the nucleotide sequence of a DNA molecule. Mutations can involve duplications, deletions, inversions, translocations, and substitutions.

Negative heterosis The inferiority of a heterozygote over that of the homozygotes with respect to one or more traits such as growth, survival, or fertility.

Neuropeptides Small molecules functioning within and without the nervous system of insects to modify behavior.

Nick A break in a single strand of a double stranded DNA molecule.

Nick translation A commonly used method of labeling DNA molecules with radioactive isotopes. DNA polymerase I is used to incorporate radiolabeled nucleotides in an *in vitro* reaction.

Nitrogenous base A purine or pyrimidine compound that forms part of the structure of a nucleotide.

Noncoding strand The polynucleotide of the DNA double helix that does not carry the genetic information, but that is the complement of the coding strand.

Nonsense mutation A mutation in a nucleotide sequence that changes a triplet coding for an amino acid into a termination codon so that a truncated polypeptide is produced which can alter the protein's activity.

Northern blotting A technique for transferring mRNAs from an agarose gel to a nitrocellulose filter paper sheet via capillary action. The RNA segment of interest is probed with a radiolabeled DNA fragment or gene.

Nuclear genome The portion of the genome contained in the nucleus of eukaryotes on chromosomes.

Nuclear pore complex A large structure forming a transport channel through the nuclear envelope.

Nucleic acid Either of the polymeric molecules DNA or RNA.

Nucleic acid hybridization The bonding of two complementary DNA strands, or one DNA and one RNA strand, to identify nucleic sequences of interest. Southern blot, Northern blot,

and plaque or colony hybridization techniques are all based on nucleic acid hybridization. All employ labeled probes to identify DNA or RNA of interest.

Nucleolus A nucleolus is an RNA-rich, spherical body associated with a specific chromosomal segment, the nucleolus organizer. The nucleolus organizer contains the ribosomal RNA genes and the nucleolus is composed of the primary products of these genes, their associated proteins, and a variety of enzymes.

Nucleosome A basic structure by which eukaryotic chromosomes are organized and compacted. Nucleosomes comprise an octamer of histone proteins with DNA coiled around them and are connected to other nucleosomes by linker DNA.

Nucleotide A compound consisting of a purine or pyrimidine base attached to a five-carbon sugar, to which a mono-, di-, or triphosphate is attached. A monomeric unit of DNA or RNA.

Nucleus The membrane-bound structure of a eukaryotic cell containing the DNA organized into chromosomes.

Null allele An allele that produces no functional product and therefore usually behaves as a recessive.

Odorant binding protein A protein that enhances the ability to smell odorants in small quantities—quantities lower than those needed to activate olfactory nerves.

Okazaki fragments Short fragments of DNA that are synthesized during replication of the lagging strand of the DNA molecule.

Oligonucleotide Short chains of single-stranded DNA or RNA nucleotides that have been synthesized by linking together a number of specific nucleotides. Used as synthetic genes or DNA probes.

Oocytes Cells produced by the ovaries that eventually become an ovum (egg cell) after meiosis.

Open reading frame (ORF) A series of codons with an initiation codon at the 5' end. Often considered synonymous with "gene" but used to describe a DNA sequence that looks like a gene but to which no function has been assigned.

Origin of replication (ORI) A base sequence in DNA that is recognized as the position at which the replication of DNA should begin. In eukaryotes, multiple origins of replication occur on each chromosome.

P element *P elements* are transposable DNA elements first found in *Drosophila melanogaster*, where they can cause hybrid dysgenesis if P containing strains are crossed with M strains lacking *P* elements. *P* elements have been engineered to serve as vectors to insert DNA into the germ line of *Drosophila* embryos.

Palindrome A DNA sequence which reads the same in both directions taking account of the two strands, i.e., 5′-AAAAATTTTTTT-3′ 3′-TTTTTTAAAAAA-5′

Paralogy Homology that arises via gene duplication.

Parasegment The visible cuticular patterns of sclerites and sutures in an insect do not represent the embryonically determined true segments. Rather, the visible "segments" are parasegments.

Parental imprinting (also genomic imprinting) The degree to which a gene expresses itself depends on which parent transmits the trait to the progeny. Imprinting may result from different patterns of DNA methylation which occur during gametogenesis in the two sexes. For such a system to maintain itself generation after generation, it would have to be reversible.

Parsimony Parsimony dictates that the minimal number of assumptions are made in an analysis.

PAS domain Protein sequence associated with signaling pathways that transmit environmental information (such as oxygen and light). Sometimes associated with protein–protein interactions.

Paternal sex ratio (PSR) The PSR condition is only carried by males of the parasitic wasp *Nasonia vitripennis* and is transmitted via sperm to fertilized eggs. After an egg is fertilized by a PSR-bearing sperm, the paternally derived chromosomes condense into a chromatin mass and subsequently are lost. The PSR chromosome itself survives, disrupting normal sex determination by changing fertilized diploid (female) eggs into haploid PSR males. PSR is the first known B chromosomeof its kind and is unusual in its ability to destroy the genome of its carrier each generation

Pathogen A virus, bacterium, parasitic protozoan, or other microorganism that causes disease by invading the body of a host; infection is not always disease because infection does not always lead to injury of the host.

PCR polymerase chain reaction

PCR-RFLP A technique that combines the PCR and RFLP analysis. Genomic DNA is amplifiedby traditional PCR. Once the DNA is amplified, it is cut with restriction enzymes, electrophoresed, and visualized by ethidium bromide staining. Because the DNA was amplified by the PCR, the DNA fragments can be visualized without having to blot and probe with a labeled probe, thus making PCRRFLP more sensitive and inexpensive than traditional RFLP analysis.

Peptide bond The chemical bond that links adjacent amino acids into a polypeptide.

Phage (bacteriophage) A virus that attacks bacteria. Frequently used as vectors for carrying foreign DNA into cells by genetic engineers.

Phagemid A phagemid is a hybrid vector molecule engineered from plasmid and M13 vectors. Phagemids provide a method for obtaining single-stranded DNA because they contain two replication origins, one a standard plasmid origin that allows production of ds DNA, and the

other from M13, which allows the synthesis of ss DNA if the host cell is superinfected with a helper phage.

Phenetic systematics Classification based on overall similarities among living organisms. All possible characters are examined and average similarities are calculated, with all characters assumed to be of equal importance.

Phenogram A branching diagram that links different taxa by estimating overall similarity based on data from characters. Characters are not evaluated as to whether they are primitive or derived.

Phenomics The study of phenotypes with knowledge of the genotypes.

Phenotype The observable characteristics of an organism that are determined by both genotype and environment.

Pheromone-binding protein Two soluble proteins are found in the lymph, a pheromone degrading esterase and a pheromone-binding protein. The pheromone-binding proteins bind species-specific pheromones and are present in very high concentrations

Phosphodiester bond The chemical bond that links adjacent nucleotides in a polynucleotide.

Phosphorylation The combination of phosphoric acid with a compound. Many proteins in eukaryotes are phosphorylated.

Phototaxis The movement of a cell or organisms toward or away from light.

Phyletic speciation The gradual transformation of one species into another without an increase in species number at any time within the lineage. Also called vertical evolution or speciation.

Phylogenetic tree A graphic representation of the evolutionary history of a group of taxa or genes.

Phylogenetics The reconstruction of the evolutionary history of a group of organisms or genes.

Phylogeny The evolutionary history of a group of taxa or genes, and their ancestors.

Physical map A map of the order of genes on a chromosome. The locations are determined by DNA sequencing, producing overlapping deletions in polytene chromosomes, or electron micrographs of heteroduplex DNAs.

Piggy Bac A transposable element that has been engineered for transforming insects other than *Drosophila*.

Plaque A clear spot on an opaque bacterial lawn in a petri dish. A plaque results after a single phage adsorbs to a bacterial cell, infects it, and lyses, releasing progeny phage. The progeny phage infect nearby bacteria and produce more phage until a clear area becomes visible to the naked eye. Each clear area contains many copies of a single phage and, if the phage is a vector containing exogenous DNA, it contains many copies of the foreign DNA.

Plaque hybridization See plaque screening.

Plaque screening Plaque screening is employed to identify, by nucleic acid hybridization with radiolabeled probes, those plaques containing specific DNA sequences.

Plasmid Circular, ds DNA molecules found in bacteria that are often used in cloning. Plasmids are independent, stable, self-replicating, and often confer resistance to antibiotics. Often used in recombinant DNA work as vectors of foreign DNA.

Pleiotropic Term used to describe a gene that affects more than one, apparently unrelated, trait.

Plesiomorphic A character used to reconstruct a phylogeny that is ancestral or primitive.

Point mutation A mutation that results from changes in a single base pair in a DNA molecule.

Pole cells The precursors of the germ cells become separated early in embryonic development in *D. melanogaster* into distinctive cells in the posterior of the egg.

Poly-A tail The processing of the 3′ end of the pre-mRNA molecule by the addition of as many as 200 adenine nucleotides, which may determine mRNA stability.

Polyacrylamide gel Polyacrylamide gels result from the polymerization of acrylamide monomers into linear chains and the linking of these chains with *N*,*N*′-methylenebisacrylamide (bis). The concentration of acrylamide and the ratio of acrylamide to bis determine the pore size of the three-dimensional network and its sieving effect on nucleic acids of different size.

Polyacrylamide gel electrophoresis (PAGE) Process by which molecules are separated based on size and charge using a polyacrylamide gel and electrical current.

Polydnaviruses The polydnaviruses are viruses with double-stranded, circular DNA genomes and found only within certain groups of parasitic Hymenoptera.

Polylinker A genetically engineered segment in a vector that allows exogenous DNA to be cloned into that region by one of two or more unique restriction sites.

Polymer A chemical compound constructed from a long chain of identical or similar units.

Polymerase chain reaction (PCR) A method for amplifying DNA by means of DNA polymerases such as *Taq* DNA polymerase. PCR fundamentally involves denaturing double stranded DNA, adding dNTPs, DNA polymerase, and primers. DNA synthesis occurs, resulting in a doubling of the number of DNA molecules defined by the primers. Additional rounds of denaturation and synthesis occur, resulting in a geometric increase in DNA molecules because each newly synthesized molecule can serve as the template for subsequent DNA amplification. Modifications of PCR primers have been developed for special purposes. The PCR is used to clone genes, produce probes, produce ssDNA for sequencing, and carry out site-directed mutagenesis. DNA sequence differences are used to identify individuals, populations, and species.

Polymorphism Two or more genetically different classes in the same interbreeding population.

Polynucleotide A polymer consisting of nucleotide units.

Polypeptide (protein) A chain of amino acids linked by peptide bonds; each protein is a gene product.

Polyploidy An increase in the number of copies of the haploid genome. Most individuals are 2n, but species are known that are polyploidy (3n, 4n, 5n, 6n) and such species are parthenogenetic because of the difficulty of maintaining normal meiosis.

Polyribosome (polysome) An mRNA molecule in the process of being translated by multiple ribosomes.

Polytene chromosomes Chromosomes in which the chromatid has duplicated up to 1000-fold without separating. Salivary gland chromosomes in *Drosophila* and other Diptera are polytene. The discrete bands of polytene chromosomes allow a physical map of genes to be constructed using light microscopy.

Position effect variegation The change in the expression of a gene when it is moved to a different region of the genome. The change in expression can be stable or variegated. Variegated position effects usually involve the suppression of wild-type gene activity when it is placed in contact with heterochromatin because of a chromosomal mutation. Under some conditions the gene may escape suppression and the final phenotype of the organism may be variegated, with patches of normal and mutant tissues.

Positive and negative selection A method for detecting and obtaining, from among many cells or organisms, those few with the desired genetic changes induced by genetic engineering. Marker genes are inserted into the organism along with the desired genes; such marker genes confer resistance to antibiotics or other chemicals and allow researchers to identify those cells/individuals that contain the newly inserted genes.

Post transcriptional processing Changes made to mRNAs, rRNAs, and tRNAs before they are finished products.

Post translational processing Changes to polypeptide chains after they have been synthesized— cleavage of specific regions to convert proenzymes to enzymes, phosphorylation, etc.

Postzygotic isolating factors Factors that help to maintain reproductive isolation between species even if mating between them does occur, such as hybrid inviability or hybrid sterility.

Pre-mRNA The unprocessed transcript of a protein-coding gene.

Prezygotic isolating factors Aspects of a species' biology that help to maintain reproductive

isolation so that mating between different species/populations does not occur, including mating discrimination or differences in habitat preferences.

Primary transcript The immediate product of transcription of a gene or group of genes which will be processed to give the mature transcript(s).

Primase The RNA polymerase that synthesizes the primer needed to initiate replication of a DNA polynucleotide during DNA replication.

Primer A short oligonucleotide that is attached to a ss DNA molecule in order to provide a site at which DNA replication can begin.

Primer-dimer artifacts Low molecular weight DNA products produced during PCR as artifacts when the reaction is carried out with high primer concentrations, too much DNA polymerase in early cycles, and small amounts of template DNA.

Prion Proteinaceous molecules found in the membranes of cells in the brains of vertebrates.
Probe A probe is a molecule labeled with radioactive isotopes or another tag that is used to identify or isolate a gene, gene product, or protein.

Prokaryote An organism whose cells lack a distinct nucleus.

Promoter A region of DNA crucial to the accuracy and rate of transcription initiation. Usually immediately upstream of the gene itself.

Proofreading A mechanism by which errors in DNA synthesis are corrected. Proofreading is carried out by a 3′ to 5′ exonuclease and increases the fidelity of the base-pairing mechanism.

Protease An enzyme that degrades proteins.

Protein The polymeric compounds made up of amino acids.

Proteoglycan A protein that is glycosylated to a variety of polysaccharide chains.

Proteome The protein complement of a cell.

Proteomics The science and process of analyzing all the proteins encoded by a genome (a proteome).

Proteasome A large protein complex in the cytoplasm of eukaryotic cells that contains proteolytic enzymes. Proteosomes break down proteins that have been tagged for destruction by the addition of ubiquitin.

Pseudogene A nucleotide sequence that is similar to a functional gene, but without accurate information so that it is not functional.

Puffing A swelling in the giant polytene chromosomes of salivary glands of many dipterans.

Pulsed field gel electrophoresis A technique for separating DNA molecules by subjecting them to alternately pulsed, perpendicularly oriented electrical fields. The technique allows separation of the yeast genome into a series of intact chromosomes on a gel. Chromosomes larger than yeast chromosomes are digested with a restriction enzyme before electrophoresis.

Purine One of the two types of nitrogenous bases that are components of nucleotides.

Pyrimidine One of the two types of nitrogenous bases that are components of nucleotides.

Q-banding Bands on chromosomes produced by quinacrine staining. The staining can only be seen under UV light and is brightest in AT-rich regions.

Quantitative genetics Analysis of the genetic influence of many genes and substantial environmental variation. It is assumed that Mendel's laws of discrete inheritance apply to complex characteristics, so that many genes, each with small effect, combine to produce observable differences among individuals in a population. Quantitative genetics determines the sum of heritable genetic influence on traits, regardless of the complexity of genetic modes of action or the number of genes involved. It does not tell us which genes are responsible for the trait.

Quantitative trait loci (QTL) Specific DNA sequences that are related to (located near to) known traits, which may be determined by multiple loci.

Radio labeling The attachment of a radioactive atom to a molecule; incorporation of 32P-dNTPs into DNA.

RAPD-PCR RAPD is derived from the term Random Amplified Polymorphic DNA. PCR using single primers of arbitrary nucleotide sequence consisting of 9 or 10 nucleotides with a 50 to 80% G+C content, and no palindromic sequences. These 10-mers can act as a primer in PCR and yield reproducible polymorphisms from random segments of genomic DNA.

Reading frame A nucleotide sequence from which translation occurs.

Real-time PCR Real-time PCR is used to quantify gene expression using a fluorescence detecting thermocycler to amplify specific sequences and measure their concentration simultaneously. See also polymerase chain reaction (PCR).

Recessive A trait or gene is recessive if it is expressed in homozygous, but not heterozygous, condition.

Reciprocal cross Crosses between individuals from two different strains (A, B), e.g., A×B and B×A.

Recombinant DNA molecule A DNA molecule created by combining DNA fragments that are not normally contiguous.

Recombinant DNA technology All the techniques involved in the construction, study, and use of recombinant DNA molecules. Often abbreviated rDNA, which can be confused with ribosomal DNA (rDNA).

Recombination A physical process that can lead to the exchange of segments of two DNA molecules and that can result in progeny from a cross between two different parents with combinations of alleles not displayed by either parent.

Redundancy Some amino acids have more than one codon. There are 64 possible combinations of four bases arranged in a triplet codon, but only about 20 amino acids.

Regulatory gene A gene that codes for a protein that is involved in the regulation of the expression of other genes.

Regulatory mutation Mutations that affect the ability to control gene expression.

Regulatory sequence A DNA sequence involved in regulating the expression of a gene (a promoter or operator).

Repetitive DNA DNA sequences that are repeated a number of times in a DNA molecule or in a genome. Some repetitive DNA is associated with heterochromatin, centromeres, and telomeres. Middle-repetitive DNA may code for ribosomal RNAs and transfer RNAs.

Replacement vectors Vectors that have a pair of insertion sites that span a DNA segment that can be exchanged with a foreign DNA fragment.

Replica plating A technique to produce identical patterns of bacterial colonies on a series of petri plates.

Replication fork The region of a dsDNA molecule that is unwound so that DNA replication can occur.

Replication origin The site(s) on a DNA molecule where unwinding of the double helix occurs so that replication can occur. There are multiple replication origins on eukaryotic chromosomes.

Reporter gene A gene used to identify or locate another gene.

Repression of gene transcription The inhibition of transcription by the binding of a repressor protein to a specific site on the DNA molecule. A repressor protein is the product of a repressor gene.

Response to selection (R) The difference in mean phenotypic value between the offspring of the selected parents and the mean phenotypic value of the entire parental generation before selection.

Restriction endonuclease An enzyme that cuts DNA only at a limited number of specific nucleotide sequences. Also called restriction enzyme.

Restriction fragment length polymorphism (RFLP) A polymorphism in an individual, population, or species defined by restriction fragments of a distinctive length. Usually caused by gain or loss of a restriction site, but could result from an insertion or deletion of DNA between two conserved restriction sites. Differences in RFLPs are visualized by gel electrophoresis.

Restriction site A specific sequence of nucleotides in a piece of dsDNA which is recognized by a restriction enzyme and which signals its cleavage.

Restriction site mapping DNA is digested with a series of different restriction endonucleases, the DNA fragments are electrophoresed, and the DNA fragments are ordered to produce a linear physical map of the locations of specific DNA sequences.

Retroelement DNA or RNA sequences that contain a gene for reverse transcriptase. There are different classes of retroelements, including retroviruses and retrons.

Retroposition The transfer of genetic information through an RNA intermediate. The genetic information carried by the DNA is transcribed into RNA, which is then reverse-transcribed into cDNA. The result is that the element is duplicated and the copy of the element is transposed.

Retro sequences Retro sequences/retro transcripts are sequences derived through the reverse transcription of RNA and subsequent integration into the genome. They lack the ability to produce reverse transcriptase.

Retrotransposon A type of transposable element that transposes by means of an RNA intermediate. At least 10 families of retrotransposons are known in *Drosophila*. Often shortened to retroposon.

Retrovirus RNA viruses that use reverse transcriptase during their life cycle. This enzyme allows the viral genome to be transcribed into DNA. The transcribed viral DNA is integrated into the genome of the host cell where it replicates in unison with the genes of the host.

Reverse genetics A particular gene is targeted for inactivation or expression in an unusual environment in order to investigate gene function. See also forward genetics.

Reverse transcriptase An enzyme that synthesizes a DNA copy from an RNA template.

Reverse transcription DNA synthesis from an RNA template, mediated by reverse transcriptase.

RFLP See restriction fragment length polymorphism (RFLP).

Ribonuclease An enzyme that degrades RNA.

Ribosomal RNA (rRNA) The RNA that acts as a structural component of ribosomes. Ribosomal RNA genes (rRNA genes) are found as tandem repeating units in the nucleolus organizer regions of eukaryotic chromosomes. Each unit is separated from the next by a nontranscribed spacer. Each unit contains three regions coding for the 28S, 18S, and 5.8S rRNAs.

Ribosome A self-assembling cellular organelle made up of proteins and RNA in which translation of mRNA occurs. Ribosomes consist of two subunits, each composed of RNA and proteins. In eukaryotes, ribosome subunits sediment as 40S and 60S particles.

Ribozyme An RNA molecule with catalytic activity. Ribozymes are known that self-splice rRNA; another ribozyme is the RNA in the large subunit of the ribosome.

Ring chromosome An aberrant chromosome with no ends.

RNA Ribonucleic acid, one of the two forms of nucleic acids.

RNA editing RNA editing involves altering the mRNA after transcription. This results in different proteins being produced from a single gene. The molecular mechanisms include single or multiple base insertions or deletions, as well as base substitutions. RNA editing occurs in both prokaryotes and eukaryotes.

RNA polymerase An enzyme capable of synthesizing an RNA copy of a DNA template.

RNA silencing (RNA interference) When double-stranded RNA (ds RNA) is injected into a cell, a defense response typically occurs in plants and animals in which the RNA is cut up into smaller chunks (about 22 nt long) and the fragments are then degraded. This process may be a defense against mobile DNA elements (TEs) which cause mutations when they insert themselves within or close to a gene. Experimentally, RNA interference can be used to silence cognate genes.

RNA surveillance A system in eukaryotic cells to degrade aberrant mRNAs.

RNA transcript An RNA copy of a gene.

S phase The portion of interphase in the cell cycle in which DNA replication occurs. The S phase occurs between the G1 and G2 phases of the interphase. Mitosis occurs after the G2 phase.

S1nuclease An enzyme that specifically degrades single-stranded DNAs or splits short single stranded segments in DNA but does not attack any double-stranded molecules. Used to convert sticky ends of duplex DNA to form blunt ends or to trim off single-stranded ends after conversion of single-stranded cDNA to the double-stranded form.

Satellite DNA Highly repeated DNA sequences with such a uniform nucleotide composition that, upon fractionalization of the genomic DNA and separation by density gradient centrifugation, they form one or more bands that are clearly different from the main band of DNA and from the smear created by other fragments of a more heterogeneous composition. The base composition of satellite DNA differs from that of the majority of DNA in a eukaryotic species, i.e., it is either A+T rich or G+C rich. Usually highly repetitive in sequence.

Secondary transposition Movement of an element after its initial insertion into the chromosome. Secondary transposition can be induced with *P* elements in *Drosophila*.

segmentation genes Genes, including the gap, pair-rule, and segment polarity classes of genes, that determine the number and polarity of the body segments during embryonic development in insects.

Selectable marker A gene that allows identification of specific cells with a desirable new genotype. Many vectors used for genetic engineering carry antibiotic resistance genes, or other genes, that allow identification of cells containing exogenous DNA.

Selection differential In artificial selection, the difference in mean phenotypic value between individuals selected as parents of the following generation and the whole population.

Selfish DNA DNA that may not provide any advantage to its carrier or host but ensures its own survival. Transposable elements are considered to be selfish DNA.

Semiconservative replication DNA replication in which each daughter double helix consists of one strand from the parent and one newly synthesized strand.

Sensory transduction Sensory cells transform and amplify the energy provided by a stimulus into an electrical signal. Sensory transduction is probably due to a change in the ionic permeability of the sensory cell membrane, which causes a depolarization of the membrane.

Sequencing The process used to obtain the sequential arrangement of nucleotides in the DNA molecule.

Sex chromosome A chromosome which is involved in sex determination.

Short germ band development A pattern of development found in some insects in which

all or most of the metameric pattern is completed after the blastoderm stage by the sequential addition of segments during elongation of the caudal region of the embryo.

Short-period interspersion pattern of genome organization This form of genome organization has single-copy DNA, 1000–2000 bp long, alternating with short (200–600 bp) and moderately long (1000–4000 bp) repetitive sequences. This pattern is found in the house fly *Musca domestica*, the Australian sheep blowfly *Lucilia cuprina*, and the wild silk moth *Antheraea pernyi*.

Shotgun cloning Genomic libraries constructed from random fragments of DNA from an organism.

Shotgun libraries Genomic libraries in which a random collection of a sufficiently large sample of cloned fragments of the DNA are present so that all the genes are represented.

Shotgun method of transformation A method for introducing foreign DNA into cells in which tiny bullets made of tungsten or other metal are coated with DNA and shot into the cell.

Silent mutation Changes in DNA that do not influence the expression or function of a gene or gene product.

Similarity A measure of the resemblance between two objects, usually on a scale of zero to one.

Single-locus, multiple-allele model A model for sex determination in Hymenoptera.

Single-strand binding proteins One of the proteins that attaches to ss DNA in the replication fork to prevent reannealing of theDNAduring DNA replication.

Site-directed mutagenesis Mutagenesis to produce a predetermined change at a specific site in a DNA molecule.

Slot blot A hybridization technique that allows multiple samples of DNA to be applied to nitrocellulose filters in specific sites (slots) using a vacuum.

Somatic cells All the eukaryotic body cells except the germ-line cells and the gametes they produce.

Southern blotting A technique developed by E. M. Southern for transferring DNA fragments isolated electrophoretically in an agarose gel to a nitrocellulose filter paper sheet by capillary action. The DNA fragment of interest is then probed with a radioactive nucleic acid probe that is complementary to the fragment of interest. The position on the filter is determined by autoradiography. The related techniques for RNA and proteins have been dubbed Northern and Western blots, respectively.

Specific activity The ratio of radioactive to nonradioactive molecules of the same kind. Probes with a high specific activity can produce a more intense signal than a probe with a low specific activity.

Spliceosome The RNA and protein particles in the nucleus that remove introns from premessenger RNA molecules. **ss DNA** Single-stranded DNA.

Stable transformation Transformation that alters the germ plasm of an organism so that the progeny transmit the trait of interest through subsequent generations.

Start codon The mRNA codon, usually AUG, at which synthesis of a polypeptide begins.

Stem cells Stem cells are able to self-renew and generate cell populations that differentiate to maintain adult tissues. There are about two stem cells in the ovary of *Drosophila* that maintain oocyte production.

Sterile insect release method (SIRM) A genetic control technique used to control or eradicate pest insects. Large numbers of mass-produced males are given nonlethal but sterilizing doses of radiation or chemical mutagens and then released. Females in natural populations mate with the sterilized males, and produce inviable progeny. After multiple releases a new generation is not produced. Used to eradicate the screwworm from North America.

Sterile male technique See sterile insect release method (SIRM).

Sticky end Single-stranded ends of DNA fragments produced by restriction enzymes; sticky ends are able to reanneal.

Stop codon One of the three mRNA codons (UAG, UAA, and UGA) that prevent further polypeptide synthesis. Also called termination codon.

Stress proteins Also called heat shock proteins. Proteins made when the cells are stressed by environmental conditions (chemicals, pathogens, heat).

Stringency Stringency, as used in hybridization reactions, refers to the conditions that can be altered to influence the ease with which a probe hybridizes to template nucleic acids.

Structural gene A gene that codes for an RNA molecule or protein other than a regulatory gene.

Structural genomics The study of protein structure based on DNA sequences.

Subclones A DNA fragment that has been cloned into one vector may be moved, or subcloned, into a second type of vector in order to perform a different procedure.

Supercoiled The coiling of a covalently closed circular duplex DNA molecule upon itself so that it crosses its own axis. A supercoil is also called a superhelix. The B form of DNA is a right-handed double helix. If the DNA duplex is wound in the same direction as that of the turns of the double helix, it is positively supercoiled. Twisting of the DNA molecule in a direction opposite to the turns of the strands in the double helix is called negative supercoiling.

Symbiont An organism living with another organism of a different species.

Sympatry Living in the same geographic location. Sympatric species have overlapping or coinciding distributions.

Synapsis The pairing of homologous chromosomes during the zygotene stage of meiosis.

Syncytium A mass of protoplasm containing many nuclei not separated by cell membranes.

Synecology The study of relationships among communities of organisms and their environment.

Syngamy The fusion of sperm and egg to form a zygote.

Synteny Synteny refers to the fact that many genes remain grouped together in the same relative positions in the genome across taxa.

Systematics The study of classification, based on evolutionary change.

tandem repeat Direct repeats in DNA codons adjacent to each other.

***Taq*DNA polymerase** A DNA polymerase that was isolated from the bacterium *Thermus aquaticus* and is tolerant of high temperatures. Used in the polymerase chain reaction.

TaqMan PCR A real-time type of PCR which uses an oligonucleotide that anneals to an internal sequence within the amplified DNA fragment.

Targeted gene replacement Replacing or modifying genes in their normal chromosomal locations has not been possible with *Drosophila* until recently. The cut-and-paste model of *P*-element transposition provided a model for inserting a gene into the double stranded gap left behind by a *P* element. The gap can be repaired, using a template provided by an extrachromosomal element that has been introduced by the investigator.

Targeted gene transfer See targeted gene replacement.

Targeted mutagenesis The ability to replace or modify DNA sequences in their normal chromosomal location.

Taxa The general term for taxonomic groups, whatever their rank. The singular form is taxon.

Taxonomy The principles and procedures according to which species are named and assigned to taxonomic groups.

tDNA-PCR Universal primers for transfer RNA can be used to generate tDNA by the PCR. The resulting fragments are visualized by gel electrophoresis and produce characteristic fingerprints for different species.

Telomerase An enzyme that adds specific nucleotides to the tips of chromosomes to form telomeres.

Telomere Telomeres are the physical ends of eukaryotic chromosomes. They protect the ends of chromosomes and confer stability. Telomeres consist of simple DNA repeats and the nonhistone proteins that bind specifically to those sequences.

Template A macromolecular mold for synthesis of another macromolecule. Duplication of the template takes two steps; a single strand of DNA serves as the template for a complementary strand of DNA or mRNA.

Termination codon One of the three codons in the standard genetic code that indicate where translation of an mRNA should stop, i.e., 5′-UAA-3′, 5′-UAG-3′, or 5′ -UGA-3′. Also called a stop codon.

Thelygenic When females produce only female progeny, as in the blowfly *Chrysomya rufifacies.*

Thelytoky Parthenogenesis in which no functional males are known; unmated females produce female progeny only, or rarely, a few males.

30-nm fiber Condensation of DNA in eukaryotic chromosomes involves formation of 30-nm fibers from supercoils of six nucleosomes per turn. The 30-nm fiber somehow is condensed further.

Tm The interpolated temperature along a DNA melting curve at which 50% of the duplex DNA formed in a DNA–DNA hybridization is double-stranded. The difference in Tm between homoduplex and heteroduplex curves is called _Tm.

Tracer DNA In DNA–DNA hybridization, single-stranded single copy DNA from one species is radioactively labeled (tracer DNA) and hybridized with unlabeled DNA (driver DNA) from the same species or from different species. DNA–DNA hybridization is used to determine the degree of sequence identity between DNAs.

Trailer segment A nontranslated sequence at the 3′ end of mRNA following the termination signal, exclusive of the poly-A tail.

Transcript An RNA copy of a gene.

Transcription The process of producing an RNA copy of a gene.

Transcriptional activator proteins Elements that stimulate transcription by binding with particular sites in the DNA.

Transcriptome The transcriptome is the profile of the genes that are expressed or transcribed from genomic DNA within a cell or tissue, with the goal of understanding cell phenotype and function. The transcriptome is dynamic and changes rapidly in response to stress or during normal cell processes such as DNA replication and cell division.

Transfection Infection of bacteria with viral nucleic acid that lacks a protein coat.

Transfer RNA (tRNA) A family of small RNA molecules (usually more than 50 types per cell) that serve as adapters for bringing amino acids to the site of protein synthesis on the ribosome.

Transformant An individual organism produced by introducing exogenous DNA.

Transformation The process of changing the genetic makeup of an organism by introducing foreign DNA. Transformation may be transient or stable (transferred to succeeding generations.)

Transgene The DNA that is inserted into the genome of a cell or organism by recombinant DNA methods.

Transgene suppression A variety of organisms, including insects, plants, and mammals, can inactivate multiple copies of inserted genes that overexpress proteins or are abnormally transcribed. Transgene silencing may be induced by methylation of the DNA or by posttranscriptional and transcriptional processes.

Transgenic organism An organism whose genome contains genetic material originally derived from an organism (not its parents) or from a different species. The transgene(s) can be transmitted to subsequent generations (stable transformation) or can be lost subsequently (unstable transformation).

Transient transformation Transient transformation involves changing the genetic makeup by introducing foreign DNA. If the genetic information is not incorporated into the germ line, the genetic changes are temporary.

Transitions Transitions are point mutations that involve changes between A and G (purines) or T and C (pyrimidines).

Translation The process by which the amino acid sequence in a polypeptide is determined by the nucleotide sequence of a messenger RNA molecule on the ribosome.

Translational regulation Gene regulation by controlling translation. Translation of mRNA can be tied to the presence of a specific molecular signal; the longevity of a mRNA molecule can be regulated; or overall protein synthesis can be regulated.

Translocation A type of mutation in which a section of a chromosome breaks off and moves to a new position in that or a different chromosome.

Transovarial transmission Transmitted to the next generation through the egg.

Transposable element An element that can move from one site to another in the genome. Transposable elements (TEs) have been divided into two classes, those that transpose with an RNA intermediate and those that transpose as DNA.

Transposase An enzyme that catalyzes transposition of a transposable element from one site to another in a DNA molecule.

Transposition The movement of genetic material from one chromosomal location to another.

Transposon A transposable element carrying several genes including at least one coding for a transposase enzyme. Many elements are flanked by inverted repeats. *Drosophila melanogaster* contains multiple copies of 50–100 different kinds of transposons.

Transposon tagging A method of cloning genes from *Drosophila* after they have been "tagged" by having the *P* element insert into them.

Transversions Transversions are point mutations that involve changes between a purine and a pyrimidine.

Triplex DNA In triplex DNA, the usual A-T and C-G base pairs of duplex DNA are present, but in addition a pyrimidine strand is bound in the major groove of the helix. DNA sequences that potentially can form triplex DNA structures appear to be common, are dispersed at multiple sites throughout the genome, and comprise up to 1% of the total genome.

Ubiquitin A protein that is present in cells of both prokaryotes and eukaryotes and is highly conserved. Ubiquitin contains 76 amino acids and plays a role in proteolysis in the proteosome. Ubiquitin-conjugating enzymes add ubiquitins to proteins carrying degradation signals. The ubiquitin is recognized by proteasomes which then cut the proteins into fragments.

Unique genes Genes present in only one copy per haploid genome, which includes most of the structural (protein-encoding) genes of eukaryotes.

Unrooted tree A phylogenetic tree in which the location of the most recent common ancestor of the taxa is unknown.

UPGMA The use of distance measurements to group taxonomic units into phenetic clusters by the Unweighted Pair-Group Method of Analysis using an arithmetic average.

Upstream Toward the 5′ end of a DNA molecule.

Uracil A pyrimidine that is one of the nitrogenous bases found in RNA.

Vector A DNA molecule capable of autonomous replication in a cell and which contains restriction enzyme cleavage sites for the insertion of foreign DNA.

Vertical gene transfer Transfer of a gene from parents to offspring. See also horizontal gene transfer.

Virus A noncellular particle that can reproduce only inside living cells; consisting only of a genetic material (either DNA or RNA) and a protein coat. Viruses are "alive" because they can reproduce, but they have no other traits of living organisms.

Vitellogenin The major yolk proteins are derived from vitellogenins, which are produced by the fat body and secreted for uptake by maturing oocytes.

Western blots Proteins are separated electrophoretically, and a specific protein is identified with a radioactively labeled antibody raised against the protein in question.

Wild type The normal form of an organism—in contrast to that of mutant individuals.

Wobble hypothesis A hypothesis to explain how one tRNA may recognize two different codons on the mRNA. Anticodons are triplets with the first two positions pairing according to base pairing rules. The third position "wobbles" and can recognize any of a variety of bases in different codons so that it can bind to either of two or more codons.

X chromosome A sex chromosome that is usually present in two copies in insect females (XX) and in one copy (unpaired) in males (XO or XY).

X-gal A lactose analogue (5-bromo-4-chloro- 3-indolyl-β-d-galactopyranoside). X-gal is cleaved by β-galactosidase into a product that is bright blue. If exogenous DNA has inserted into and disrupted the β-galactosidase gene, λ plaques will appear white or colorless. Plaques without recombinant vectors will be blue.

Y chromosome A sex chromosome that is characteristic of males in species in which the male typically has two dissimilar sex chromosomes (XY).

Z chromosome One of the sex chromosomes found in heterogametic ZW female insects.

Z-DNA A structural form of DNA in which the two strands are wound into a left-handed helix rather than a right-handed form.

Zinc finger protein Proteins with tandemly repeating segments that bind zinc atoms. Each segment contains two closely spaced cysteine molecules followed by two histidines. Each segment folds upon itself to form a fingerlike projection. The zinc atom is linked to the cysteines

and histidines at the base of each loop. The zinc fingers serve in some way to enable the proteins to bind to DNA molecules, where they regulate transcription.

Zygote A fertilized egg formed as the result of the union of the male and female gametes.

Chapter-2

Extraction of DNA from Insect tissue using CTAB method

Sakshi Gandotra, Sagar, D. and S. Subramanian
Division of Entomology, ICAR-Indian Agricultural Research Institute, New Delhi
sakshi262@gmail.com

Principle

Firstly, cell wall is disrupted physically or chemically to get a fluid containing all the cell components including DNA. The process is called cell lysis and the resulted fluid is known as lysate. During cell lysis, different chemicals and reagents are used to break down different cell components. CTAB (Cetyl Trimethyl Ammonium Bromide) is a detergent used to break open cells and solubilize its contents. The extraction process involves breaking or digestion of cell wall in order to release the cellular constituents. This is followed by disruption of the cell membrane to release the DNA into the extraction buffer. The released DNA should be protected from endogenous nucleases. EDTA is often included in the extraction buffer. EDTA acts as chelating agents and binds to Mg^{2+} an ion which acts as co-factors for DNase. The unavailability of Mg^{2+} ions leads to the deactivation of DNase activity and hence saves the DNA from degradation. DNA is pH sensitive and can be degraded on pH change. Tris acts as pH stabilizer during cell lysis process. It maintains the pH at 8. Secondly, lysate is treated with concentrated salt solution to make the broken components clumped together and leave the DNA freely floating in the solution.

The cations Na^+ or K^+ binds to negative phosphate groups of DNA and makes it more stable in aqueous solution. In the absence of Na^+ or K^+, DNA molecules repel each other and do not allow grouping of DNA molecules. Thirdly, this solution (containing lysate, detergents, surfactants, broken proteins, lipids and RNA) is centrifuged to separate the clumped debris from DNA. Lastly, DNA precipitation is done by adding ice cold alcohol plus a salt to increase the ionic strength which increases the precipitation process. Most proteins are removed by Ice cold ethanol that is added to the solution containing DNA and cell debris. Proteins get dissolved in ethanol. A pellet of DNA is obtained upon centrifugation of this solution. Supernatant is discarded except for the DNA pellet which remains stuck to the walls of the Eppendorf. Pellet is, then, suspended either in slightly alkaline solution mostly TE buffer or ultra-pure water (organic particles and dissolved gases removed) for subsequent DNA experimentation usually PCR.

Reagents Required:
Proteinase K
Lysozyme
SDS 10%
Tris-EDTA (TE) buffer (pH 8)

CTAB buffer (10%)
Chloroform: Isoamyl alcohol (24:1)
Isopropanol
Ethanol 70%
NaCl (5M)
TE buffer (10 mM)

Sample Preparation

Homogenize the insect with homogenizer in 0.85% Sodium Chloride (NaCl)) solution. Store the homogenate at 4°C and can be utilized for further experimental purpose.

Procedure:

1. Centrifuge the insect sample at 13,000 rpm for 10 minutes. Discard the supernatant.
2. Dissolve the pellet in 500 µl of Lysis buffer with gentle vortexing for 30 seconds and incubate at 37°C overnight. Lysis buffer consist of the following components:

Components	***Total volume (10ml)***
Proteinase K	80 µl
Lysozyme	100 µl
SDS 10%	500 µl
Tris-EDTA (TE) buffer (pH8)	9.32 ml

3. After completion of overnight incubation, again incubate the lysed samples at 55°C for 20 minutes.
4. Further completion of 20 minute incubation period at 55°C, add 80 µl of CTAB buffer (10%) and 100 µl of 5M NaCl solution. Mix the mixture properly by inverting the tubes and then again incubate at 65°C for 20 minutes.
5. In obtained mixture, add 700 µl of Chloroform: Isoamyl alcohol (24:1). Mix the mixture with vigorous shaking for 2 minutes and then centrifuge the tubes at 13,000 rpm for10 minutes at room temperature.
6. After centrifugation of the sample tubes, formation of two layers occurs. Collect the top phase in a clean new eppendorf collection tube and precipitate it with 360 µl of Isopropanol. Mix the collection tubes for 5 seconds.
7. Freeze the tubes at -20°C for 30 minutes. After cold freezing, again centrifuge the tubes at 13,000 rpm for 10 minutes. Discard the supernatant from the tube, pellet will be seen.
8. Add 70% ethanol for washing the pellet. After immediate addition of ethanol, centrifuge the samples at 5,000 rpm for 5 minutes. Discard the supernatant from the tubes carefully and dry the pellet at 37°C for 30 minutes.
9. After drying, resuspend the pellet in 100 µl of 10 mM TE buffer.
10. Store the DNA samples at -20°C for further experimental processes. The quality and quantity of DNA can be checked by using NanoDrop Spectrophotometer and Agarose gel electrophoresis.

Chapter-3

RNA isolation and cDNA synthesis for gene expression analysis in insects

D. Sagar, Rajna S. and S. Subramanian
Division of Entomology, ICAR-Indian Agricultural Research Institute, New Delhi
garuda344@gmail.com

RNA isolation using Triazol reagent

Principle: RNA (Ribonucleic acid) is a polymeric substance present in living cells and many viruses, consisting of a long single-stranded chain of phosphate and ribose units with the nitrogen bases adenine, guanine, cytosine, and uracil, which are bonded to the ribose sugar. RNA is used in all the steps of protein synthesis in all living cells and carries the genetic information for many viruses. The isolation of RNA with high quality is a crucial step required to perform various molecular biology experiment. TRIzol Reagent is a ready-to-use reagent used for RNA isolation from cells and tissues. TRIzol works by maintaining RNA integrity during tissue homogenization, while at the same time disrupting and breaking down cells and cell components. Addition of chloroform, after the centrifugation, separates the solution into aqueous and organic phases. RNA remains only in the aqueous phase. After transferring the aqueous phase, RNA can be recovered by precipitation with isopropyl alcohol. But the DNA and proteins can recover by sequential separation after the removal of aqueous phase. Precipitation with ethanol requires DNA from the interphase, and an additional precipitation with isopropyl alcohol requires proteins from the organic phase. Total RNA extracted by TRIzol Reagent is free from the contamination of protein and DNA. This RNA can be used in Northern blot analysis, in vitro translation, poly (A) selection, RNase protection assay, and molecular cloning.

Chemicals/Reagents required

1. Triazol
2. Chloroform
3. Isopropanol
4. Ethanol

Procedure:

1. Homogenize larvae in Triazol at 4˚C. Initially add 300 µl homogenize and then add 700 µl of Triazol (Homogenization is an important step)
2. Add 200 µl of chilled chloroform and vortex it for 15 seconds, and leave at room temperature for 15 minutes.

3. Centrifuge at 12,000 rpm at 4˚C for 15 minutes. There will be three phases visible within the tube. Transfer the aqueous phase (top) to a fresh tube, be careful not to contaminate the solution with the other phases.
4. Add 500 µl of chilled isopropanol and mix gently
5. Incubate at room temperature for 10-15 minutes
6. Centrifuge at 10,000 rpm at 4˚C for 15 minutes
7. RNA get pelleted, remove isopropnaol
8. Wash RNA pellet with 75% ethanol
9. Centrifuge at 8000 rpm for 10 min at 4˚C
10. Excess ethanol was removed by pipetting out
11. Remove ethanol by drying it on dry bath at 55˚C.
12. RNA pellet was dissolved in 70 µl nuclease free water.
13. Keep at room temp for 5 minutes and then at 55˚C for 10-15 minutes
14. Firstly the dissolved RNA kept @ 4˚C for 10-15 minutes and after 30 minutes OD was measured in nanospectrophotometer.

RNA isolation using RNAzol® RT reagent

Principle: RNAzol® RT is the most effective reagent for isolation of total RNA and small RNA from samples of human, animal, plant, bacterial and viral origin. This patented reagent provides higher yield and quality of isolated RNA than previous reagents based on the single-step method. RNAzol® RT isolates pure and undegraded RNA that is ready for RT-PCR without DNase treatment.

- No chloroform-induced phase separation is necessary to obtain pure RNA. Just add water to remove DNA, proteins, polysaccharides and other contaminants.
- The isolation procedure can be completed in less than one hour and is performed at room temperature, including all centrifugation steps.
- RNAzol® RT isolates total RNA, or large RNA and small RNA in separate fractions. The large RNA fraction contains rRNA and mRNA. The small RNA fraction contains tRNA, small RNA and microRNA down to 10 bases.
- The isolated RNA is ready for RT-PCR, qRT-PCR, microarrays, poly A+ selection, northern blotting, RNase protection assay and other molecular biology applications.
- Due to the removal of impurities, the RNA pellets are smaller and solubilize more easily than pellets obtained from previous single-step reagents.

Material

- RNAzol® RT
- Ethanol
- Diethyl pyrocarbonate (DEPC) treated water

Method

1. Homogenize 100mg tissue in RNAzol at 4˚C. Initially add 300 µl homogenize and then add 700 µl of RNAzol (Homogenization is an important step)
2. Centrifuge at 5,000 g at 4˚C for 5 minutes
3. Collect the supernatant in a new eppendroff (Collect approximately 750 µl)
4. Add 400µl of chilled DEPC treated water
5. Shake vigorously for 15 seconds/times.
6. Leave at room temp for 15 minutes incubation
7. Centrifuge at 12,000 g at 4˚C for 15 minutes
8. Collect supernatant in a new eppendroff (Collect approximately 750-800 µl)
9. Add 600 µl of chilled 75% ethanol and mix well
10. Leave at room temperature for 10 minutes
11. Centrifuge at 12,000 g at 4˚C for 10 minutes
12. RNA get pelleted
13. Wash RNA pellet with 75% ethanol
14. Spin at 5000g for 5 min at 4˚C
15. Excess ethanol was removed by pipetting out
16. RNA pellet was dissolved without drying in nuclease free water.
17. Keep at room temp for 5 minutes and then at 55˚C for 10-15 minutes
18. Firstly the dissolved RNA kept @ 4˚C for 10-15 minutes and after 30 minutes OD was measured in nanospectrophotometer and use it cDNA synthesis.

Precautions

i. Take precautions to avoid RNase contamination during isolation and handling of RNA
ii. Microcentrifuge tubes, tips, falcon tubes used in RNA isolation should be treated with 0.1% DEPC water
iii. Prepare the reagents in DEPC treated water only

cDNA synthesis

Complementary DNA (cDNA) synthesis describes the generation of complementary DNA (cDNA) from an RNA template by reverse transcription. cDNA is DNA synthesized from a single-stranded RNA (e.g., messenger RNA (mRNA) or microRNA) template in a reaction catalyzed by the enzyme reverse transcriptase. Reverse transcriptases (RTs) use an RNA template and a primer complementary to the RNA to direct the synthesis of the first strand cDNA, which can be used directly as a template for the Polymerase Chain Reaction (PCR), notably gene expression analysis using real-time PCR (qPCR). cDNA is derived from mRNA, so it contains only exons but no introns**.**

Chemicals/Reagents required

1. Template RNA

2. Oligo dT/Random hexamer/gene specific primer
3. 5x Reaction buffer
4. dNTP (10mM)
5. Ribolock RNase inhibitor
6. Reverse transcriptase.

1) S_1/1st cycle : 70°C , 5 min (12 µl)
    - Template : 3000ng/3µg
    - OligodT :1µl
    - Water : 11-x (x=template)
2) S_2/second cycle : 37°C , 5 minutes (12 µl + 7 µl = 19 µl)
    - 5x reaction buffer : 4.1
    - 10mM dNTP mix : 2.0
    - Ribonuclease RNAse inhibitor :1.1
3) S_3/third cycle : 42°C , 60 min followed by 70°C, 10 min (19 µl + 1 µl = 20 µl)
    - Reverse transcriptase: 1µl and spin it
4) Hold at 4˚c for infinite time.

Purity and concentration of RNA will be determined by Nano Drop spectrophotometer by measuring absorbance ratios of 260/280, 260/230 and integrity of RNA was determined by 1% agarose gel electrophoresis. Only samples that satisfy the quantity, purity and integrity will be used for synthesis of cDNA using cDNA synthesis kit (Thermo Scientific Revert aid first strand cDNA synthesis kit) with the following cycling program: (Template RNA-3000ng, OligodT primer-1µl and Nuclease free water (NFW) 11-x(x=template) at 70ºC for 5 minutes followed by the addition of 5x reaction buffer (4µl), 10 mM dNTPs (2µL) and RNAase inhibitor (1µl) at 37ºC for 5 minutes followed by reverse transcriptase (1 µL) at 42ºC for 60 min and 70ºC for 10 minutes and hold at 4°C for infinite time. After the RT-PCR, products will be visualized on 2% agarose gel.

Chapter-4

Polymerase Chain Reaction and Agarose Gel Electrophoresis

Rahul Kumar Chandel, Sagar, D. and S. Subramanian
Division of Entomology, ICAR-Indian Agricultural Research Institute, New Delhi
rahulchandelbiotech@gmail.com

The polymerase chain reaction (PCR) is a laboratory (invitro) technique for generating large quantities of a specified DNA. The PCR was originally developed in 1983 by the American biochemist Kary Mullis. He was awarded the Nobel Prize in Chemistry in 1993 for his pioneering work. Obviously, PCR is a cell free amplification technique for synthesizing multiple identical copies (Billions) of DNA of interest. PCR is now considered as a basic tool for the molecular biologist, as like photocopier a basic requirement in an office so PCR machine in a molecular biology laboratory.

PCR Recipe

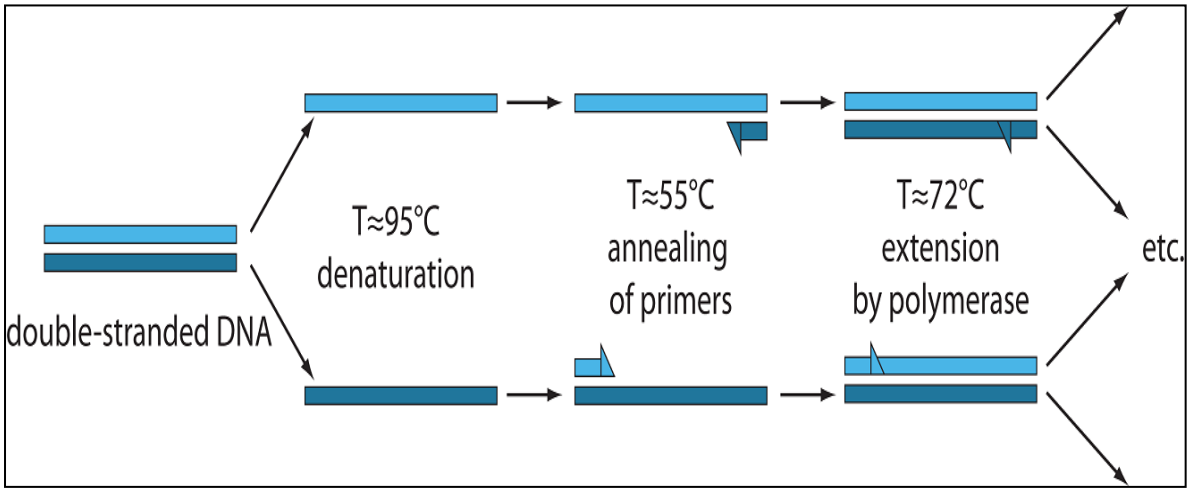
- **Template DNA ─** A target DNA 100-35000 bp in length
- **dNTPs** – Four Deoxyribonucleotides (dATP, dCTP, dGTP, dTTP)
- **Primers ─**Two primers (Synthetic oligonucleotides of 17-30 nucleotides length)
- **DNA polymerase Enzyme** – A DNA polymerase that can withstand at temperature up to 95°C (Thermostable)
- Buffer to ensure the right conditions for the reaction
- Nuclease free water to make up reaction volume

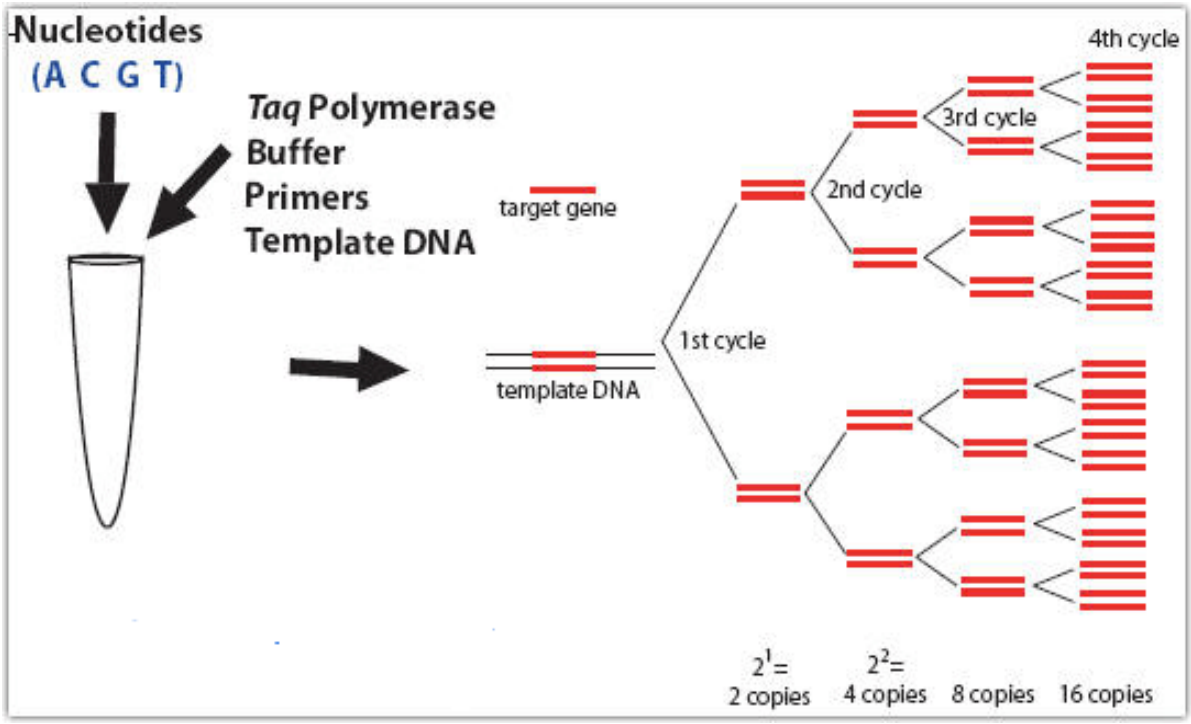
Reaction volume

- Master Mix : 12.5µl
- Forward primer : 1µl
- Reverse primer : 1µl
- Nuclease free Water : 8.5µl
- DNA Template : 2 µl
- Total volume – 25 µl

Procedure of PCR

These three steps **denaturation**, **renaturation** and **synthesis** are repeated again and again to generate multiple forms of target DNA





Analysis of PCR Products

- Gel Electrophoresis of PCR Products
- Sequencing of the PCR Products
- Cloning of PCR Products

Agarose Gel Electrophoresis

Agarose gel electrophoresis is a simple and highly effective method for separating, identifying, and purifying 0.5- to 25-kb DNA fragments. The protocol can be divided into three stages: (1)

a gel is prepared with an agarose concentration appropriate for the size of DNA fragments to be separated; (2) the DNA samples are loaded into the sample wells and the gel is run at a voltage and for a time period that will achieve optimal separation; and (3) the gel is stained or, if ethidium bromide has been incorporated into the gel and electrophoresis buffer, visualized directly upon illumination with UV light.

For the typical DNA separation experiment, however, this simple chart is sufficient for selecting a gel concentration:

Agarose Gel Concentration (%w/v)	**DNA Size Range for Optimal Separation in bp**
0.3	5,000 – 60,000
0.6	1,000 – 20,000
0.7	800 – 10,000
0.9	500 – 7,000
1.2	400 – 6,000
1.5	200 – 3,000
2.0	100 – 2,000

Usually 1 to 2% gels are used for detecting plasmids (several kb long) or their fragments (ie. from digestions). For resolving much shorter DNAs, use polyacrylamide gel electrophoresis (PAGE, see separate section). Gels with a lower percentage of agarose tend to be flimsy, so if you do use them run them at low temperature (4ºC).

Equipment

- Casting tray
- Well combs
- Voltage source
- Gel box
- UV light source
- Microwave

Reagents

- TAE
- Agarose
- Ethidum bromide (stock concentration of 10 mg/mL)

Procedure

- Measure 1 g of agarose.
- Mix agarose powder with 100 mL 1xTAE in a microwavable flask

- Microwave for 1-3 min until the agarose is completely dissolved
- Let agarose solution cool down to about 50 °C (about when you can comfortably keep your hand on the flask), about 5 mins.
- Add ethidium bromide (EtBr) to a final concentration of approximately 0.2-0.5 µg/mL (usually about 2-3 µl of lab stock solution per 100 mL gel).
- Pour the agarose into a gel tray with the well comb in place.
- Place newly poured gel at 4 °C for 10-15 mins OR let sit at room temperature for 20-30 mins, until it has completely solidified.

Loading Samples and Running an Agarose Gel:

- Add loading buffer to each of your DNA samples.
- Once solidified, place the agarose gel into the gel box (electrophoresis unit).
- Fill gel box with 1xTAE (or TBE) until the gel is covered.
- Carefully load a molecular weight ladder into the first lane of the gel.
- Carefully load your samples into the additional wells of the gel.
- Run the gel at 80-150 V until the dye line is approximately 75-80% of the way down the gel.
- Turn OFF power, disconnect the electrodes from the power source, and then carefully remove the gel from the gel box.
- Using any device that has UV light, visualize your DNA fragments.

Chapter-5

PCR Primer's Characteristics, Designing and Resuspending

Rahul Kumar Chandel, Sagar, D. and S. Subramanian
Division of Entomology, ICAR-Indian Agricultural Research Institute, New Delhi
rahulchandelbiotech@gmail.com

Good primer design is essential for successful PCR reactions. The important design considerations described below are a key to specific amplification with high yield.

Primer Length: It is generally accepted that the optimal length of PCR primers is 18-22 bp. This length is long enough for adequate specificity and short enough for primers to bind easily to the template at the annealing temperature.

Primer Melting Temperature: Primer Melting Temperature (T_m) by definition is the temperature at which one half of the DNA duplex will dissociate to become single stranded and indicates the duplex stability. Primers with melting temperatures in the range of 52-58^oC generally produce the best results. Primers with melting temperatures above 65^oC have a tendency for secondary annealing. Primer with higher GC content has a higher Tm because of more hydrogen bonds (three). A simple formula for calculation for T_m is

$T_m = 2*AT + 4 * CG$

AT = sum of A and T nucleotides and CG is sum of C and G nucleotides in the primer.

Primer Annealing Temperature (T_a): The primer melting temperature is the estimate of the DNA DNA hybrid stability and critical in determining the annealing temperature. Too high T_a will produce insufficient primer-template hybridization resulting in low PCR product yield. Too low T_a may possibly lead to non-specific products caused by a high number of base pair mismatches. Mismatch tolerance is found to have the strongest influence on PCR specificity.

$T_a = 0.3 \times T_m$ (primer) + 0.7 T_m (product) – 14.9

Where,

T_m (primer) = Melting Temperature of the primers

T_m (product) = Melting temperature of the product

GC Content: The GC content (the number of G's and C's in the primer as a percentage of the total bases) of primer should be 40-60%.

GC Clamp: The presence of G or C bases within the last five bases from the 3' end of primers (GC clamp) helps promote specific binding at the 3' end due to the stronger bonding of G and C bases. More than 3 G's or C's should be avoided in the last 5 bases at the 3' end of the primer.

Primer Secondary Structures: Presence of the primer secondary structures produced by intermolecular or intramolecular interactions can lead to poor or no yield of the product. They adversely affect primer template annealing and thus the amplification. They greatly reduce the availability of primers to the reaction.

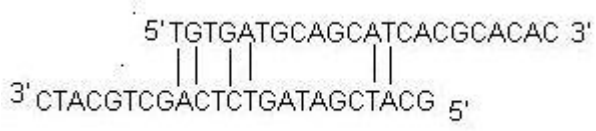
i) **Hairpins**: It is formed by intramolecular interaction within the primer and should be avoided. Optimally a 3' end hairpin with a ΔG of -2 kcal/mol and an internal hairpin with a ΔG of -3 kcal/mol is tolerated generally.

ΔG definition: The Gibbs Free Energy G is the measure of the amount of work that can be extracted from a process operating at a constant pressure. It is the measure of the spontaneity of the reaction. The stability of hairpin is commonly represented by its ΔG value, the energy required to break the secondary structure. Larger negative value for ΔG indicates stable, undesirable hairpins. Presence of hairpins at the 3' end most adversely affects the reaction.

$\Delta G = \Delta H - T\Delta S$

ii) **Self Dimer**: A primer self-dimer is formed by intermolecular interactions between the two (same sense) primers, where the primer is homologous to itself. Generally a large amount of primers are used in PCR compared to the amount of target gene. When primers form intermolecular dimers much more readily than hybridizing to target DNA, they reduce the product yield. Optimally a 3' end self dimer with a ΔG of -5 kcal/mol and an internal self dimer with a ΔG of -6 kcal/mol is tolerated generally.

iii) **Cross Dimer**: Primer cross dimers are formed by intermolecular interaction between sense and antisense primers, where they are homologous. Optimally a 3' end cross dimer with a ΔG of -5 kcal/mol and an internal cross dimer with a ΔG of -6 kcal/mol is tolerated generally.



Repeats: A repeat is a di-nucleotide occurring many times consecutively and should be avoided because they can misprime. For example: ATATATAT. A maximum number of di-nucleotide repeats acceptable in an oligo is 4 di-nucleotides.

Runs: Primers with long runs of a single base should generally be avoided as they can misprime. For example, AGCGGGGGATGGGG has runs of base 'G' of value 5 and 4. A maximum number of runs accepted is 4bp.

3' End Stability: It is the maximum ΔG value of the five bases from the 3' end. An unstable 3' end (less negative ΔG) will result in less false priming.

Avoid Template Secondary Structure: A single stranded Nucleic acid sequences is highly unstable and fold into conformations (secondary structures). The stability of these template secondary structures depends largely on their free energy and melting temperatures(T_m). Consideration of template secondary structures is important in designing primers, especially in qPCR. If primers are designed on a secondary structure which is stable even above the annealing temperatures, the primers are unable to bind to the template and the yield of PCR product is significantly affected. Hence, it is important to design primers in the regions of the templates that do not form stable secondary structures during the PCR reaction. Our products determine the secondary structures of the template and design primers avoiding them.

Avoid Cross Homology: To improve specificity of the primers it is necessary to avoid regions of homology. Primers designed for a sequence must not amplify other genes in the mixture. Commonly, primers are designed and then BLASTed to test the specificity. Our products offer a better alternative. You can avoid regions of cross homology while designing primers. You can BLAST the templates against the appropriate non-redundant database and the software will interpret the results. It will identify regions significant cross homologies in each template and avoid them during primer search.

Parameters for Primer Pair Design

Amplicon Length: The amplicon length is dictated by the experimental goals. For qPCR, the target length is closer to 100 bp and for standard PCR, it is near 500 bp. If you know the positions of each primer with respect to the template, the product is calculated as: Product length = (Position of antisense primer-Position of sense primer) + 1.

Product Position: Primer can be located near the 5' end, the 3' end or anywhere within specified length. Generally, the sequence close to the 3' end is known with greater confidence and hence preferred most frequently.

Tm of Product: Melting Temperature (T_m) is the temperature at which one half of the DNA duplex will dissociate and become single stranded. The stability of the primer-template DNA duplex can be measured by the melting temperature (T_m).

Optimum Annealing Temperature (T_a Opt): The formula of Rychlik is most respected. Our products use this formula to calculate it and thousands of our customers have reported good results using it for the annealing step of the PCR cycle. It usually results in good PCR product yield with minimum false product production.

T_a Opt = 0.3 x(T_m of primer) + 0.7 x(T_m of product) - 14.9

Where
T_m of primer is the melting temperature of the less stable primer-template pair
T_m of product is the melting temperature of the PCR product.

Primer Pair Tm Mismatch Calculation: The two primers of a primer pair should have closely matched melting temperatures for maximizing PCR product yield. The difference of 5°C or more can lead no amplification.

Primer Design using Software

A number of primer design tools are available that can assist in PCR primer design for new and experienced users alike. These tools may reduce the cost and time involved in experimentation by lowering the chances of failed experimentation.

Some primer design programs we use:

- Oligo: Life Science Software, standalone application
- GCG: Accelrys, ICBR maintains the server.
- PrimerQuest Tool-IDT
- Primer3: MIT, standalone / web application http://www-genome.wi.mit.edu/cgi-bin/primer/primer3'www.cgi
- BioTools: BioTools, Inc. ICBR distributes the license.
- Others: GeneFisher, Primer!, Web Primer, NBI oligo program, etc.
- PrimerPlex is software that can design primers for Multiplex PCR and multiplex SNP genotyping assays.

Resuspending of PCR primers and other oligos

Overview primers are often shipped and received in a lyophilized state. First create a master 100 uM stock (for each primer) and then dilute it to 10uM working stock. This reduces the number of freeze/thaw cycles that the master primer stock goes through and reduces the chances of contaminating the primary source for the primer.

Materials

- Lypholized primers
- Nuclease free water

Procedure

1) Spin down tubes. Primers should always be spun down before opening the tube for the first time. The pellet can often come dislodged during shipping and may be in the cap.

2) Prepare Master Stock, 100 µM = X nmoles lyophilized primer × 10 (molecular grade H_2O, µl). To determine the amount of water to add to the lyophilized primer simply multiply the number of nmol of primer with 10. That will be the amount of water to add to make a 100 µM primer stock. For example, if there are 38.2 nmol of primer a 100 µM primer stock is created by adding 382 µl of water. The original primer tubes are used for this 100 µM stock.

Master stock primers newly suspended in water should be allowed to settled down at room temperature for 10 minutes before they are used for working stock dilutions. Mix well before making working stock dilutions.

3) Prepare working stock of 10 µM by diluting the primer master stock in a sterile micro centrifuge tube 1:10 with sterile molecular grade water.

Chapter-6

Gene expression analysis through qPCR in insects

Rajna S., Sagar, D. and S. Subramanian
Division of Entomology, ICAR-Indian Agricultural Research Institute, New Delhi
rajnasalim@gmail.com

Introduction

Gene expression analysis can be performed in accurate, sensitive and fast manner using Reverse transcription quantitative PCR (RT-qPCR) and it is considered as the golden standard for medium throughput gene expression analysis. Real Time PCR is also known as 'fluorescence based PCR' and it quantify the nucleic acids obtained from cells or tissues. It is mostly used for two reasons, either as a primary investigative tool to determine gene expression or as a secondary tool to validate the results of DNA microarrays. Quantitative measurement of specific gene expression using qPCR is necessary for understanding basic cellular mechanisms and detecting of alteration in gene expression levels in response to specific biological stimuli.

Principle

Real time PCR combines the amplification and analysis steps of the normal PCR reaction, thus eliminating the need for post-PCR processing. Polymerase chain reaction can be divided into four phases: the linear ground phase, exponential phase, log-linear phase and plateau phase. The method is based on the principle that there is a quantitative relationship between the amount of target nucleic acid (DNA/RNA) present at the start of the assay and the amount of product amplified during its exponential phase. At the linear ground phase, (usually the first 10-15 cycles), fluorescence emission produced at each cycle has not been higher than the background. At exponential phase, the amount of fluorescence reaches a threshold where it can be detected as significantly stronger than the background fluorescence signal. The cycle in which this detection happened is known as threshold cycle (C_T). C_T value is a very important point of Real-time PCR because this value represents the amount of target gene found in the sample and it is used to calculate experimental results. If the amount of target sequence is high in the sample, the reaction reaches exponential phase more quickly and thus, the cycle (or C_T value) in which the amount of fluorescence reaches a threshold will be lower for this sample.

Materials required

Materials for RNA isolation, cDNA synthesis kit, qPCR thermal cycler plates, SYBR green, nuclease free water, specific primers *etc*.

Steps in RT PCR- For gene expression analysis

Gene expression studies using RT-qPCR has to be followed with certain steps sequentially and perfectly for a better result.

i) **Template preparation**: The RNA isolated should be highly pure and without any DNA or protein contamination. The OD values of 260/280 should be in a range 0f 2.0 to 2.2 for downstream applications. cDNA synthesis is considered as crucial step in qPCR study.

ii) **Primer synthesis**: Optimal primers are essential to ensure that only a single PCR product is amplified. Primer size should be 18-24 nucleotides long with GC content 45-55% with good GC spread. Primer melting temperatures (Tm) should be 55-60^{0}C. Primers should be designed to give product size of 100-200 bp. In order to avoid non-specific PCR products, primers should not have sequence similarity with other sequences. If oligo(dT) is used for priming in reverse transcription, primers should be located within 1000 bp of the 3' end of mRNA or better results. Some free online tools can be used for designing of primers *viz*., Primer 3, Primer blast, Real Time PCR primer sets

iii) **Housekeeping gene (HKG)/reference gene selection:** The proper housekeeping gene (HKG) is continuously expressed in all cell types and tissues. Expression level of a suitable reference gene should be stable and is not affected by the biologic and experimental condition. There is no universal housekeeping gene having invariable expression under all these circumstances. Therefore, choosing a stable housekeeping gene is crucial for the accurate interpretation of gene expression. The most frequently used housekeeping genes involved β-actin (ACTB), glyceraldehyde-3-phosphate dehydrogenase (GAPDH) elongation factor α 1 (EF α 1) and tubulin in insect related studies.

Protocol

1. Synthesized cDNA can be diluted to keep the cDNA concentration in all the samples uniform (also to remove pipetting error)
2. Reaction mixture (20 µl) of each reaction consists of
   SYBR containing master mix – 10 µl
   Forward primer – 0.2 µl
   Reverse primer -0.2 µl
   Nuclease free water – as needed to make up to 20 µl
   cDNA template (100-250 ng)- as needed to maintain the concentration
3. A programme for melting curve has to be constructed to check the contamination in reaction.
4. Finalize the thermal cycler conditions by doing semi quantitative PCR using specific primers
5. Real Time PCR thermal conditions may vary according to the experiment.
   Eg: Condition for heat shock protein genes in brown planthopper
   Initial denaturation at 95°C for 5 minutes

Denaturation 95°C for 30 seconds
Annealing 57°C for 30 seconds } 40 cycles
Final extension at 72°C for 20 seconds

6. The data obtained can be quantified using the following method

Real time Quantification

Quantification of gene expression can be measured by absolute or relative methods. Absolute quantification is an analysis method to accurately measure the copy number of a target sequence (in picograms or nanograms of DNA or RNA) in the sample, while relative quantification provides relative changes in mRNA expression levels as a ratio of the amount of initial target sequence between control and analysed samples. Relative quantification simply allows us to determine the fold changes between sample and control. Here we describe on the relative quantification of mRNA expression.

Relative quantification

Relative quantification determines the changes in steady-state mRNA level in response to different treatments (e.g., control versus experimental). During relative quantification, amounts of target and reference gene's are determined within the same sample. The housekeeping gene which helps to normalize the data for experimental error, can be co-amplified in the same tube in a multiplex assay or can be amplified in a separate tube. Normalization strategy using reference gene is a convenient method for sample to sample variation.

Data normalization using the Ct values o the target gene and the normalizer is as follows.

$$\Delta Ct_{target} = Ct_{target} - Ct_{normalizer}$$

$$\Delta Ct_{calibrator} = Ct_{calibrator} - Ct_{normalizer}$$

$$\Delta\Delta Ct = \Delta Ct_{target} - \Delta Ct_{calibrator}$$

Finally to obtain relative fold expression $\mathbf{F = 2^{-\Delta\Delta Ct}}$ (Livak and Schmittgen, 2001)

Conclusion

Real Time PCR has become one of the most reliable tools in gene expression analysis studies. The method being a rapid, accurate, sensitive, cost effective and reproducible one, this technology has become a routine and robust approach for nucleic acid based diagnostics. Choosing an appropriate reference is always been a challenge in real time PCR for obtaining biologically relevant results.

Reference

Derveaux, S., Vandesompele, J. and Hellemans, J. (2010). How to do successful gene expression analysis using real-time PCR. *Methods,* 50: 227–230.

Livak and Schmittgen (2001). Analysis of Relative Gene Expression Data Using Real-Time Quantitative PCR and the 22DD*C*T Method. *Methods,* 25: 402–408.

Yilmaz, A., Onen, H.I., Alp, E. and Menevse, S. (2014). Real-Time PCR for Gene Expression Analysis.

Chapter-7

DNA Barcodes for Insects

P.R. Shashank and Naresh M. Meshram
National Pusa Collection, Division of Entomology, ICAR-IARI, New Delhi
spathour@gmail.com

Introduction

Early attempts at molecular systematics were also termed as chemotaxonomy that made use of proteins, enzymes, carbohydrates, and other molecules that were separated and characterized using techniques such as chromatography. These have been replaced in recent times largely by DNA sequences because of advances in sequencing and computational technologies. Although earlier works focused on several DNA regions, mitochondrial DNA (mtDNA) is the most common DNA region that gain attention because it is maternally inherited and evolution of mtDNA is rapid enough to allow discrimination of species and phylogeographic groups within single species.

DNA barcoding is one of the DNA-based techniques that have been used for identification and to study molecular evolution. DNA barcoding is the use of a short standardized DNA sequence (in insects, a 658 bp fragment of the mitochondrial cytochrome c oxidase (COX I) gene) to identify and assign unknown specimens to species besides facilitating the discovery of new species. Then, the Consortium for the Barcode of Life (CBOL) was launched in May 2004 to promote DNA barcoding as a global standard for sequence-based identification of eukaryotes and now includes more than 170 organizations from 50 nations. This consortium with association of National Center for Biotechnology Information (NCBI) is in the processes of building universal barcode library. The Barcode of Life Data System (BOLD) is an open access integrated bioinformatics workbench that supports all phases of the analytical pathway from specimen collection to tightly validated barcode library. BOLD act as repository for the specimen and sequence records that form the basic data unit of all barcode studies and it creates a atmosphere where in a group of scientist can collaborate with each other.

Procedure:

Sample collection

One to three legs of sample specimens were separated carefully in such way that coxa should be included without damaging the specimen. Separated parts preserved in 95 % Ethyl Alcohol and kept in refrigerator for future DNA extraction.

Extraction and isolation of DNA

Pre DNA extraction

Takeout stored moth legs in a cavity block and remove the scale by rubbing gently with very fine Camlin brush (000'). Transfer the legs to another cavity block containing distilled water and wash. Washed legs are placed on tissue paper for drying.

DNA Extraction:

There are commercially available kits for extraction of insect genomic DNA. All the kits come with extraction protocols. However, traditional method of DNA extraction is useful for the beginners. The Cetyl trimethylammonium bromide (CTAB) procedure for isolating DNA was simplified from one proposed by Doyle and Doyle (1990). The steps are as follows:

i. All six legs of adult moths were taken in a 1.5 ml eppendorf tube and powdered well using liquid nitrogen.

ii. The ground tissue was transferred to a 50 ml sterile centrifuge tube by adding 15ml of CTAB buffer (Extraction buffer- 2%C-TAB, 1.4 M NaCl, 20 mM EDTA (Disodium) and 10mM Trisbase (pH 8)). Then 50 μL of β- mercaptoethanol was added to each tube and the extract was mixed well by inverting for two min. The tubes were incubated in a water bath maintained at 65^0C for an hour with constant stirring at an interval of 15 min.

iii. After an hour, 15 ml of chloroform isoamyl alcohol (24:1) was added to the incubated sample and mixed well by inverting. The tubes were then centrifuged at 10,000 rpm for 20 min.

iv. The aqueous upper phase was carefully transferred using one ml tips into fresh sterile centrifuge tubes. To this supernatant, 0.7 volumes (10.5 ml) of cold isopropanol were added.

v. The tubes were carefully inverted and kept for 5 min on ice. The precipitated DNA was then centrifuged at 6000 rpm for 20 min and sedimentation of DNA as a pellet was seen.

vi. Further the supernatant was decanted gently and the tubes were inverted on a clean filter paper.

vii. The pellet was washed twice by suspending in one ml of 70% ethanol for 5 to 10 min and centrifuged at 6000 rpm for two min.

viii. Ethanol was drained off slowly and the pellet was vacuum dried in a dessicator for 5 to 10 min. The pellet was then dissolved in 500 μL of TE buffer by flicking the tubes. TE buffer = 0.1mM Tris + 0.05mM EDTA

ix. To remove the RNA, 5µL of RNase (10mg/ml) was added into the DNA solution and incubated at 37ºC in a water bath for one hour.

x. Again the DNA solution was cleaned by washing with equal volume (500 µl) of phenol: choloroform: Isoamyl alcohol (25:24:1) by invert mixing several times and centrifuging at 6000 rpm for 15 min to separate the two phases.

xi. The aqueous upper phase was transferred into a clean 1.5 ml eppendorf tube and two volumes of absolute ethanol were added to precipitate the DNA.

xii. The pellet was washed with 70% ethanol twice and dissolved in 500 µl TE. The extracted DNA was quantified using both Nanodrop DNA quantifier and also by electrophoresis on 0.8% agarose gel.

DNA purity assessment

Two µl of isolated DNA were diluted to one ml with TE buffer and the absorbance at 260 and 280 nm were recorded against a buffer blank. A 260/280 nm ratio for all the samples was calculated to check the purity. DNA was quantified using the formula:

µg ds DNA/µl = (A260 * 40)/2

Further all the samples were diluted to a final concentration of 10 ng/µl.

DNA quantification on agarose gel

Agarose gel submersible electrophoresis is used to separate the DNA based on size (0.1 to 2.5 kb), in a submerged horizontal tray. The tray was chosen in such a way that its length allowed the required band resolution. Genomic DNA electrophoresed at 1% agarose concentration.

Preparation of agarose gel

i. One g of agarose was weighed and taken in a clean 250 ml conical flask. To this 100 ml of 1x TBE buffer was added. [TBE buffer (0.89M Tris base, 0.02 MEDTA, 0.89M Boric acid) pH = 8].

ii. Agarose was dissolved by heating at 100ºC for 15 min. After agarose completely melts, it was cooled to 60°C and 3.5 µl of ethidium bromide was added (10 mg/ml) and mixed.

iii. The ends of gel casting trays were sealed with cellophane tapes. Agarose was poured; comb was inserted and allowed to solidify. After solidification the tape on either side was removed and the gel was immersed in electrophoresis tank containing 1x TBE buffer. Then the comb was removed.

iv. To 5 µl of DNA samples 3µl of 6x loading buffer was added, mixed well and loaded into the well. 3µl of standard uncut Lambda DNA was used as marker. (Loading buffer or tracking dye 6x – 40% sucrose, 0.025% bromophenol blue, 0.25% xylene cyanol).

v. Electrophoresis was carried out at 50V for 2 to 3 hours until the bromophenol blue dye migrated two-thirds of the gel. The gel tray was removed and the gel was observed under UV transilluminator and documented using Gel Doc system. The quantity of the DNA was determined based on the intensity of the band relative to lambda uncut band.

Primers used

Mitochondrial cytochrome oxidase sununit I gene (mtCOI gene) was amplified using the forward primer LCO1490 and reverse primer HCO2198, developed by Folmer *et al.* (1994).

Name of the primer	**Sequence (5' to 3')**
LCO1490	5' GGT CAA CAA ATC ATA AAG ATA TTG G 3'
HCO2198	5' TAA ACT TCA GGG TGA CCA AAA AAT CA 3'

PCR amplification

The reaction mixture was set up in sterile 0.2 ml microfuge tubes. The reaction mixture volume per reaction was as follows:

Reaction Mixture	**Quantity (µl)**
10x Taq Polymerase buffer	5
2 mM dNTPs	5
MgCl2 (2mM)	1
Forward primer (5pMole/µl)	2
Reverse primer (5pMole/µl)	2
Taq Polymerase (1U)	0.3
Template DNA (20-25ng)	4
Sterile water	30.7
Total volume	**50.0**

Reaction mixture was vortexed and centrifuged. Amplifications were performed using master cycler with following temperature transitions:

Steps	**Temperature (ºC)**	**Time**
Initial denaturation	94	5 min
Denaturation	94	1 min
Annealing	52	30 sec
Elongation	72	1 min
Extension	72	5 min

Thermal cycle was programmed for 35 cycles with one cycle of initial denaturation and steps 2-4 were repeated 35 times.

Agarose gel electrophoresis of PCR product

The PCR products were resolved by electrophoresis using 3 percent agarose gel in1X Tris borate EDTA buffer for about 2 hours at 110V along with 100 bp ladder. The gel was stained with ethidium bromide (0.5 μg/ml), viewed under UV Tran-illuminator and photographed immediately for further interpretation using Gel-Doc system.

Sequencing

PCR products are purified and dissolved in 0.1' TE by gel extraction/PCR cleanup methods. Purified samples are sequenced at the specific commercial facilities.

Sequence analysis

Obtained DNA sequences were proofread by eye and aligned using MEGA version 4.1 (Tamura *et al.*, 2007).

Sequences were then checked for plausibility using BLAST with the blastn algorithm (Altschul *et al.* 1990; URL: http://blast.ncbi.nlm.nih.gov/Blast.cgi) as well as the BOLD Identification System (IDS, URL: http://www.boldsystems.org/index.php/IDS_OpenIdEngine)

The sequence details are analyzed carefully, submitted to NCBI for GenBank Accessions and subsequently uploaded to BOLD. All generated sequences, together with photographs and collection details, have been deposited at the Barcoding of Life Database (BOLD; www.boldsystems.org) under the specific project code. DNA barcoding analyses will be done through the online interface of the BoLD website. The taxon identification tree was based on the Kimura 2-parameter distance model (Kimura, 1980), with the filter set to sequences with length >100 basepairs, and all codon positions included.

Chapter-8

Molecular gut analysis of the predators

Sachin S. Suroshe* and Bhagyasree S.N.
Division of Entomology, ICAR-Indian Agricultural Research Institute, Pusa, New Delhi
**sachinsuroshe@gmail.com*

Principle

Entomophagous insects especially coccinellids play important role in control of soft bodied insects *viz*., mealybugs, aphids, scales and whitefly. Likewise, most of these predators forms a specific guild based on the preys they devours *i.e*. coccidophagous and aphidophagous guild etc. Guild refers to living beings competing for the same resources. As we are aware, parasitoids are specific, evolutionary tightly liked and coevolved thus, attack the particular host. However, predators are generalist might feed on different prey's consisting of host such as extraguild prey's (insect pest's) and intraguild prey's (competing predators of the same guild). That's why, the identification of the gut contents of predatory insects could provide information on predator-prey interactions *i.e*. whether released predator is feeding on the targeted prey (pests) or non targeted prey (pests) or competing predators. So, using a Multiplex polymerase chain reaction (PCR) we can determine what the field collected insect predators might have fed after release.

Procedure

The Predator- Prey System

- **The Cowpea aphid (*Aphis craccivora* Koch):** It is a polyphagous pest but mostly prefers cowpea and groundnut. This species is also accounts for the transmission of more than 50 plant viruses.
- **The six spotted lady bird beetle (*Cheilomenes sexmaculata* F.):** Almost found throughout India and the Oriental region. It remains active and found throughout the year. It is known to be an aphidophagous, but also feeds on psyllids, whiteflies, mealybugs, tingids, leaf and plant hoppers, mites, and early instar lepidopteran larvae.

DNA Extraction

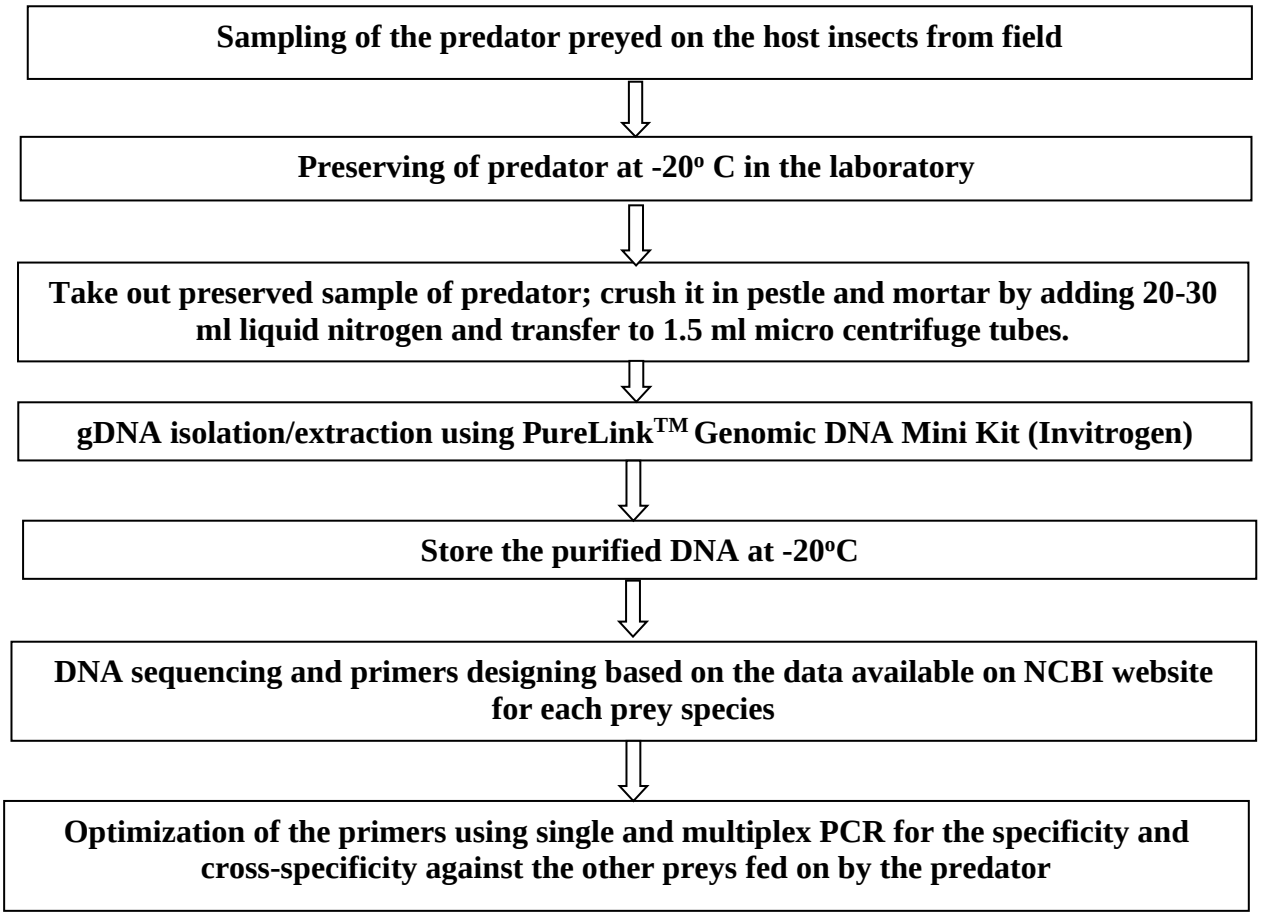
First the predator, *C. sexmaculata* is allowed to feed on the prey, *A. craccivora* in the field. Then the 5-6 predators are sampled using aspirator and kept in 70% ethanol then stored in -20 refrigerator till the DNA is extracted. DNA will be taken out from whole insect using PureLink™ Genomic DNA Mini Kit (Invitrogen).

DNA Amplification using Multiplex PCR

If you want to know the probable prey candidate (more than one) eaten by the predator, you must perform multiplex PCR. Multiplex PCR refers to the use of polymerase chain reaction to amplify several different DNA sequences simultaneously (i.e. performing many separate PCR reactions all together in one reaction). The primers which will be used is a mitochondrial gene i.e. mitochondrial cytochrome oxidase I (mtCOI). A DNA barcode is a short sequence from standardized portions of the genome (648 bp). Primers are designed using 'Primer-3' website based on the species specific mtCOI data available on NCBI website for each prey. Optimization of the primers must be carried out for its specificity and for cross-specificity against the other preys fed on by the predator.

- Different multiple primers are being used in this for amplification of the DNA in samples.
- Optimization of the primer pairs has to be done properly so that, all primer pairs could work at the same annealing temperature.

Work flow chart for Molecular gut analysis of the predators for detecting gut prey content



Chapter-9

Restriction Fragment Length Polymorphism (RFLP)

Rahul Chandel and S. Subramanian
Division of Entomology, ICAR- Indian Agricultural Research Institute, New Delhi
subramanian@iari.res.in

Restriction Fragment Length Polymorphism (RFLP) is a difference in homologous DNA sequences that can be detected by the presence of fragments of different lengths after digestion of the DNA samples in question with specific restriction endonucleases. RFLP, as a molecular marker, is specific to a single clone/restriction enzyme combination. Most RFLP markers are co-dominant and highly locus-specific. An RFLP probe is a labeled DNA sequence that hybridizes with one or more fragments of the digested DNA sample after they were separated by gel electrophoresis, thus revealing a unique blotting pattern characteristic to a specific genotype at a specific locus. Short, single- or low-copy genomic DNA or cDNA clones are typically used as RFLP probes.

Developing RFLP probes

- Total DNA is digested with a methylation-sensitive enzyme (for example, PstI), thereby enriching the library for single- or low-copy expressed sequences (PstI clones are based on the suggestion that expressed genes are not methylated).
- The digested DNA is size-fractionated on a preparative agarose gel, and fragments ranging from 500 to 2000 bp are excised, eluted and cloned into a plasmid vector (for example, pUC18).
- Digests of the plasmids are screened to check for inserts.
- Southern blots of the inserts can be probed with total sheared DNA to select clones that hybridize to single- and low-copy sequences.
- The probes are screened for RFLPs using genomic DNA of different genotypes digested with restriction endonucleases. Typically, in species with moderate to high polymorphism rates, two to four restriction endonucleases are used such as EcoRI, Mbo1

PCR-RFLP

Isolation of sufficient DNA for RFLP analysis is time consuming and labor intensive. However, PCR can be used to amplify very small amounts of DNA, usually in 2-3 hours, to the levels required for RFLP analysis. Therefore, more samples can be analyzed in a shorter time. Cleaved Amplified Polymorphic Sequence (CAPS) assay deploys PCR- RFLP to distinguish between the specific differences in gene sequences owing to the presence of SNPs in a particular locus.

Cleaved Amplified Polymorphic Sequences (CAPS)

Polymorphisms are differences in restriction fragment lengths caused by SNPs or INDELs that create or abolish restriction endonuclease recognition sites in PCR amplicons produced by locus-specific oligonucleotide primers. The CAPS assay uses amplified DNA fragments that are digested with a restriction endonuclease to display RFLP. Unique sequences primers are used to amplify a mapped DNA sequence from two related individuals (for example, from two different inbred ecotypes), A/A and B/B, and from the heterozygote A/B. The amplified fragments from A/A and B/B contain two and three RE recognition sites, respectively. In the case of the heterozygote A/B, two different PCR products will be obtained, one which is cleaved three times and one which is cleaved twice. When fractionated by agarose or acrylamide gel electrophoresis, the PCR products digested by the RE will give readily distinguishable patterns. Some bands will appear as doublets.

Advantages of CAPS

- Most CAPS markers are co-dominant and locus-specific.
- Most CAPS genotypes are easily scored and interpreted.
- CAPS markers are easily shared between laboratories.
- CAPS assay does not require the use of radioactive isotopes, and it is more amenable, therefore, to analyses in clinical settings.

Developing CAPS markers

- Sequence the gene of interest.
- Design primers to amplify 800–2,000-bp DNA fragments. Targeting introns or 3' untranslated regions should increase the chance of finding polymorphisms
- The PCR product is cloned and sequenced.
- Separately digest the amplicons with one or more restriction enzymes in the PCR amplified DNA fragments from target genotypes.
- Screen the digested amplicons for polymorphism on gels stained with ethidium bromide.

Sample Queries

Some CAPS Probes

CAPS markers are widely used in a number of molecular biological investigations such as:

- Detection of kdr genes associated with pyrethroid resistance in different insect species
- Detection of target site mutation in AchR genes or Nicotinyl receptor genes in resistant populations of insect species

Chapter-10

Genotyping of phosphine resistance in Red flour beetle *Tribolium castaneum*- A Practical Approach

Rahul Chandel, Suresh M. Nebapure and S. Subramanian
Division of Entomology, ICAR-Indian Agricultural Research Institute, New Delhi
subramanian@iari.res.in

Aim: To study the genotyping of insecticide resistance through RFLP technique in *Tribolium castaneum*

Principle

Two principal genes have been identified in association with insecticidal resistance of *T. castaneum* against phosphine gas:

- *rhp1* (Weak resistance)
- *rph2* (Strong resistance)
- Analysis of the full length length gene *rph2* reveals the presence of a number of SNP in the resistance individual in comparison to susceptible ones
- A key SNP of P45S mutation (phenyl-alanine-serine) has been detected in *T. castaneum* has been exploited for developing CAPS marker for diagnosis of phosphine resistance (Schlipalius *et al.*, 2012)

CAPS: Cleaved Amplified Polymorphic Sequence of segment of *rph2* having P45, SNP will be cleaved differently by specific restriction enzyme (*MboI*), as this restriction enzyme has unique restriction site in a region encompassing P45S mutation.

Accordingly, a PCR –RFLP technique has been developed to amplify 368 bp region of *rph2*. This amplicon of 368 bp when digested with *MboI* will cut this amplicon into 292 and 72 bp in insects showing strong resistance to phosphine. While PCR amplicons of DNA from susceptible individual remain uncut with 368 bp, as the susceptible population is devoid of the presence of this specific mutation.

Genotyping of resistance

Accordingly, the results of the PCR- RFLP of *rph2* locus will have of amplicons of of 368 for susceptible alleles (ss), while, cut fragments of 292 and 72 bp represent strongly resistance alleles (rr) homologous at rph2 locus; the heterozygous will have three fragments of sizes, 368 292 and 72bp.

Procedure

- Individual of population taken
- DNA Isolation
- PCR (Polymerase chain reaction)
- Desired amplicon *rph2* (368 bp)
- Restriction digestation (*MboI*, RE)
- Digested amplicon visualized using agarose gel electrophoresis
- Alleles frequencies will be scored.

Genomic DNA extraction and Quantification

- CTAB method (William *et al.*, 2004) and
- DNA quantified by Nanodrop

Polymerase chain reaction

PCR is done by using 2µl genomic DNA template with gene specific primers for rph2 gene in *T.castaneum*

Reaction volume

- Master Mix : 12.5µl
- Forward primer : 1µl
- Reverse primer : 1µl
- Nuclease free Water : 8.5µl
- DNA Template : 2 µl

PCR Conditions

- 95°C for 5 min
- 95°C for 1 min
- 57°C for 1 min
- 72°C for 1 min

} - 40 cycles

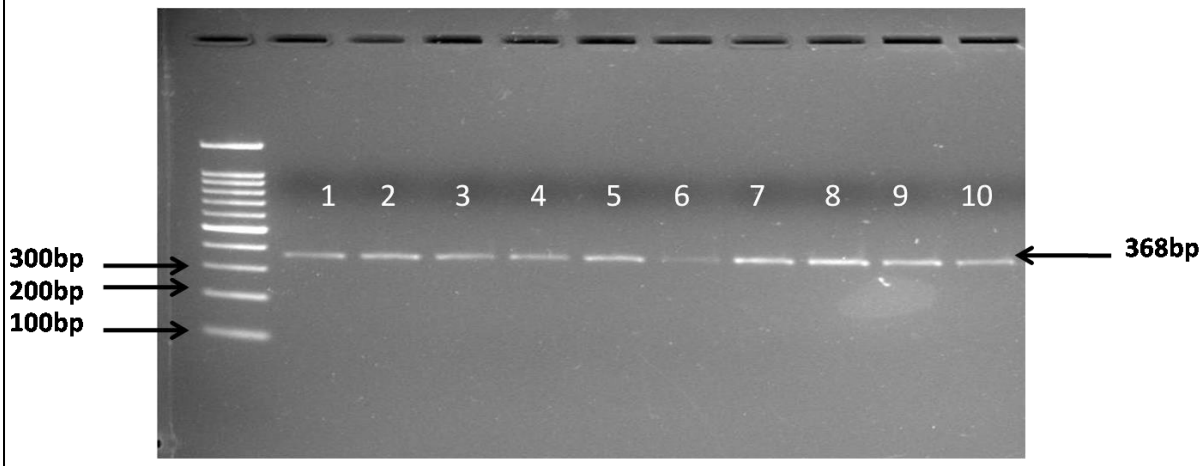
- Final extension :72°C for 10 min

Restriction digestion of PCR products

The amplified 368-bp PCR product is to be digested with *Mbo*I RE at 37 °C for overnight followed by warming up at 65 °C for 20 minutes with the following reaction volume

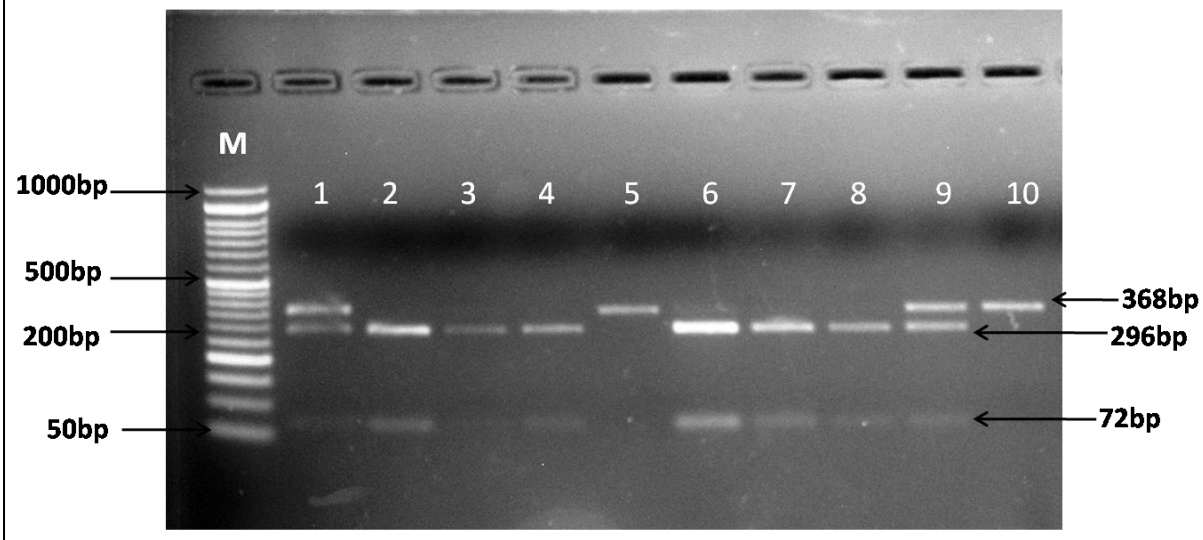
- PCR product : 8µl
- Restriction enzyme : 1µl
- 10X buffer R : 2µl
- Nuclease-free water :16µl

Agarose gel showing the PCR amplicons targeting a segment of rph2 gene



Lanes M on extremes : 100 bp DNA ladder, Lanes 1-10 on gel are PCR amplicons of rph2 gene

Agarose gel showing the PCR-RFLP – CAPS markers, Restriction digestion of PCR products with Restriction Enzyme, *MboI*



Lanes M on extremes : 50 bp DNA ladder, Lanes 1-10 on gel are digested products from susceptible SS - 368 bp; homozygous resistant (RR) - 296 bp & 72 bp; and heterozygous resistant (RS) – 368 bp,296& 72 bp

Agarose Gel Electrophoresis

The PCR product is to be visualized using 2 % agarose gel with TAE buffer. Approximate time for run in Agarose gel electrophoresis : 30 minutes.

Scoring technique for genotyping of resistance:

- Uncut single fragment (368 bp) - Susceptible (ss)
- 2 Fragments (72 and 292 bp) - Homozygous resistant (rr)
- 3 Fragments (72, 296 and 368 bp) - Heterozygous resistant (rs)

References

Chen Z, Schlipalius D, Opit G, Subramanyam B, Phillips TW (2015). Diagnostic Molecular Markers for Phosphine Resistance in U.S. Populations of *Tribolium castaneum* and *Rhyzopertha dominica*. *PLoS ONE,* 10(3): e0121343. doi:10.1371/journal.pone.0121343

Schlipalius D, Valmas N, Tuck A, Jagadeesan R, Ma L, Kaur R, *et al.* (2012). A core metabolic enzyme mediates resistance to phosphine gas. *Science*, 338(6108): 807– 810.

Chapter-11

Insect Preparation for Genomics

S. Subramanian
Division of Entomology, ICAR- Indian Agricultural Research Institute, New Delhi
subramanian@iari.res.in

A. Important considerations for insect genomics

1. Often they cannot be reared in the lab – which precludes any breeding for genome homozygocity - and instead must be collected on field trips necessitating the use of some material for species identification. Even if research colonies are available, annual and longer lifecycles can make inbreeding unrealistic
2. As many of the insect species are often physically small, such that very little DNA (nanograms) can be obtained from a single individual, necessitating pooled polymorphic individuals to make libraries. In cases with intermediate sized individuals, we prioritize a single individual for the majority of sequence, and pooled individuals for larger insert libraries, where significant material is lost in agarose gel size selection.
3. Due to the large species diversity within the arthropods, there are generally no high quality genome assemblies of phylogentically close species to aid in assembly
4. DNA preps often have to be optimized for a new insect species, as entomologists are not trained in molecular methods and standard protocols have not been determined. Different DNA extraction kits give varying results in terms of quantity and quality of DNA extracted. Hence, one needs to optimize the protocol for extraction of DNA of adequate quantity for genomic sequencing
5. The genome size varies with different groups of insect species. Although holometabola often have small (~500Mb) genomes, outside the holometabola, arthropods can have large genomes (spiders: 1.5Gb, cockroaches: 3Gb, mantis etc)
6. Hence the costs are also variable (compared to the relative stability of the 3Gb mammals).

Preparation of insects

1. The insect samples to be used for genomic sequencing are to have genome homozygocity. Homogenous populations are to be maintained from iso female lines to attain the homogeneity of the samples
2. DNA extraction methodologies: A number of methodologies viz., Kit based methods (Qiagen-QIAamp DNA Mini kit; Wizard gDNA kit (Promega), CTAB protocol (Schäffer *et al.*, 2010); Chelex-100® (Bio Rad, USA) are available for DNA extraction.

One needs to do optimization for choice of DNA extraction protocol for genomic sequencing purposes

3. In many cases input DNA quantity from a single individual can be very small, and pooled individuals increase the polymorphism and chances of a poor assembly product. But some of the above strategies require relatively small amounts of DNA (although the DNA must still not be degraded).
4. Illumina Synthetic reads require only 500ng of DNA, depending on the amount of synthetic long-read sequence desired.

DNA quantity and quality

- To estimate the concentration of DNA, the Fluorometer (Promega Quantus or Invitrogen Qubit) or Nano –Photometer can be utilized. For measurements with Fluorometer, fluorescent DNA-binding dyes are to be used.
- The purity of the DNA can be checked for DNA (A260/280) ratio.

B. DNA EXTRACTION PROTOCOL FOR INSECTS

Protocol 1: CTAB /Mercaptoethanol Method

This protocol is for isolating dna from insect tissue preferably legs, head and wings.

CTAB EXTRACTION BUFFER – (For 50ml)

- 1M Tris-HCl 5ml
- 0.5 M EDTA 2ml
- 5M NaCl 17.5ml
- 10% CTAB 10ml (5 gm in 50 ml)
- Distilled Water 15.5 ml to make up the volume.
- Set the pH to 7.5-8.0. (Autoclave the whole content before use)

Methodology

1. Dry the alcohol dipped samples, transfer it to microcentrifuge tubes.
2. Add 60 0C pre-warmed CTAB (600ul) +3ul of Beta-Mercaptoethanol+ 10ul of 20% SDS
3. After crushing add 3ul 0f Proteinase-K in each tube.
4. Vortex each tube for 5 minutes vigorously.
5. Keep it overnight incubation, and vortex it after the interval of 2-3 hours if possible, after incubation, cool the samples to room temperature, and centrifuge at 14,000 rpm for 10 min
6. Add equal volume of Phenol: chloroform: Isoamyl alcohol (25:24:1) and mixed it for 5 minutes (vortex).
7. After mixing, centrifuged at 14,000 rpm for 10 minutes.
8. Then, take supernatant fresh microfuge tube and discard the pellet.
9. Add Chloroform: isoamyl alcohol (24:1), 600ul in each tube and mix it (vortex).

10. Centrifuge at 12,000 rpm for 10 minutes.
11. Take the supernatant in fresh microfuge tube and discard the debris.
12. Add chilled isopropanol (400 ul) and mixed slowly until white flakes appear.
13. Then keep it at -35 or-50^{o}C in deep freeze for 1hour.
14. Brought the sample at room temperature.
15. Centrifuge at 10,000 rpm for 10 minutes.
16. Decant the supernatant and add 70% chilled ethanol (400ul) +ammonium acetate (100 ul) to the pellet for washing.
17. Centrifuge at 10,000 rpm for 10 minutes.
18. Decant the supernatant carefully; add 400ul of absolute alcohol and Centrifuge at 10,000 rpm for 10 minutes.
19. Decant the supernatant and dry the pellet at room temperature.
20. Dissolve it in nuclease free water.

Protocol 2: Rapid, Cost efficient genomic DNA isolation protocol

In molecular ecology and entomology, there is a constant need to develop a rapid and cost-efficient genomic DNA extraction protocol for large numbers of samples for downstream systematic analytical survey to establish genetic and adaptive diversity within natural populations. Though several commercial genomic DNA extraction kits have been developed, they are generally either expensive or not readily available, especially for researchers in developing and under-developed countries around the world. Developing a simple time- and cost-efficient protocol for DNA extraction is not only essential for molecular studies in many parts of the world where the price of the kits is unaffordable, but also highly desirable when a large number of samples are to be processed. In insects, polyphenol bound proteins, created by phenol-oxidases present in the cuticle, are some of the major contaminating compounds in the extracted DNA samples. The mechanisms by which these phenolic compounds inhibit the DNA-polymerases are still not well understood, although it has been speculated that polyphenols can bind to the DNA itself resulting in PCR inhibition.

A simple genomic DNA extraction protocol for insect tissues that is time and cost-efficient, free of PCR inhibiting contaminants, and not reliant on toxic reagents such as phenol/ chloroform. to combine the above protocol with currently available in-house sequencing technologies.

Reagents and equipment

- DNA extraction buffer: 1% SDS (from 10% (w/v) SDS stock); 0.5 M NaCl (from 5 M NaCl stock solution). Working solution of the extraction buffer with 1% SDS and 0.5 M NaCL was freshly prepared from stock solutions of 10% (w/v) SDS and 5 M NaCl solution before the DNA extraction
- Isopropanol
- 70% (v/v) Ethanol
- Table-top microcentrifuge

Insect collection and DNA extraction

Insects collected from outdoors/ lab /field populations can be used for DNA extraction. Insect tissue, from the thoracic region, of approximately 10–20 mg is to be macerated at room temperature (RT) in a sterile Eppendorf tube using a hand-operated homogenizer without buffer. 400 µl of DNA extraction buffer is added to the homogenate and thoroughly mixed using a vortex for 20 s. The suspension is then spun (13,000 rpm, 1 min, RT) using a table-top microcentrifuge. The supernatant is transferred into a new Eppendorf tube, and 400 µl of isopropanol is added, mixed gently by inversion, and then spun (13,000 rpm, 1 min, RT) as above. The supernatant is to be discarded, and the DNA pellet is washed with 500 µl 70% (v/v) ethanol and spun down (13,000 rpm, 1 min, RT) using a table-top microcentrifuge. The ethanol was thereafter discarded. The excess ethanol was blotted from the pellet by inverting it on a clean paper-towel, and then the pellet was air-dried to remove any residual ethanol. The pellet (DNA) was dissolved in 100 µl ddH2O and stored at 4˚C for immediate use or at -20 to -80˚C for long-term storage.

Remarks

The uniqueness of this protocol resides in its simplicity and its environmentally-friendly aspect making it appealing to scientists. Using this protocol, the DNA extraction can be accomplished within 5 min by a non-specialist. Students using this method for the first time require minimal supervision. The protocol is reliable and can therefore be used for DNA extraction in animal cells. It has the advantage of requiring no use of liquid nitrogen, ice, or the addition of hazardous reagents. It is therefore suitable not only for large scale sample extraction but also for high quality DNA extraction in laboratories with relatively limited equipment or funds.

Protocol 3. Chelex Technique

i. Insect specimen are to be separated and softened by deionized water, and transfer to a clean autoclaved 1.5ml micro-centrifuge tube.
ii. Then add 20 µl of deionized water in tube and grind specimen with pipette tube and convert it into suspension, add 100 µl autoclaved 1× PBS/1% saponin solution and mix it by vortexing and incubate it at room temperature for 20 min and centrifuge for 2 min at 20,000 × g, discord the supernatant.
iii. Then again re-suspend remaining solution (pellet) in 100 µl 1× PBS, and centrifuge again at 20,000 × g for 2 min.
iv. In next step remove supernatant, re-suspend pellet by 5 sec vortexing in 25 µl of 20% w/v Chelex-100 resin suspension in deionized water and 75 µl sterile deionized water, and make hole in lid of tube by 23 G hypodermic needle. Boil this suspension on water bath for 10 min and then centrifuge at 20,000 × g for 1 min, and isolate the DNA solution].
v. The Chelex(®) 100 resin approach can be used to extract DNA from insect samples. Through this technique even from a single egg, substantial amount of mitochondrial

DNA can be extracted, amplified and sequence as barcode region for identification of species.

Remarks

This technique required three simple reagents and 37 min time for complete extraction. The Chelex method is simple and sustainable approach.

Protocol 4: Chelax method for 96 well plates

Material needed (for a 96-well PCR plate)

- 14.4 mL of a home-made 10% solution of Biotechnology Grade Chelex 100® (Bio Rad. USA) resin in purified water, obtained by mixing 14.4 mL of purified water with 1.4 g of Chelex
- 960 lL of Proteinase K (>600 mAU⁄mL; Qiagen))
- forceps, filter paper, spirit lamp, magnetic agitator,
- Thermocycler

Methodology

i. Carefully remove 1–2 legs from a specimen with clean forceps and place them on clean filter paper. This will remove most of the ethanol.
ii. Pick up the leg(s) and place them at the bottom of the first well of a 96-well PCR plate. Care is required as static electricity may move the leg: the leg must be at the bottom of the well. If static electricity makes tissue handling too tricky, it is possible to fill-in each well of the plate well with 5 lL of 95% ethanol prior to this step. The step 5 should then be lengthened to the point when all ethanol is evaporated.
iii. Dip the tips of the forceps in 95% ethanol and burn the ethanol in a flame. This will destroy all remaining tissue and DNA; the forceps are now ready for the next sample.
iv. Repeat steps 1–3 with all samples until each of the 96 wells contains 1–2 legs.
v. Incubate at room temperature for 30 min to ensure that all remaining ethanol has evaporated.
vi. Place a bottle containing a 10% Chelex® 100 solution on a magnetic agitator and mix until homogeneous.
vii. Pipette 10 µl of Proteinase K into each well, ensuring that the legs are submerged in the Proteinase K.
viii. Pipette 150 µl of the 10% Chelex® 100 Resin solution in each well.
ix. Seal the plate and incubate it at 55 ˚C overnight (NB: most thermocyclers have an 'Incubate' function).
x. Store the plate in deep freezer at – 20® C.

Remarks

When working on arthropods, it is often desirable to extract DNA from the leg muscle, as this part of an arthropod will usually contain fewer pigments and potential PCR inhibitors. A further drawback is that Chelex extractions perform poorly on dry, old museum arthropod samples. It is then suitable only for fresh and well-preserved material.

Chapter-12

DNA Sequencing, Data handling, Curation, Assembly and Submission of sequences to GenBank

S. Subramanian*, V. Govindasamy and A. Kumar
ICAR-Indian Agricultural Research Institute, New Delhi
**subramanian@iari.res.in*

Computer databases are an increasingly necessary tool for organizing the vast amounts of biological data currently available and for making it easier for researchers to locate relevant information. In 1979, the Los Alamos Sequence Database was established as a repository for biological sequences. In 1982, this database was renamed GenBank and, later the same year, moved to the newly instituted National Center for Biotechnology Information (NCBI), where it lives today. By the end of 1983, more than 2,000 sequences were stored in GenBank, with a total of just under 1 million base pairs (Cooper & Patterson, 2008).

At about the same time, a joint effort between NCBI, the European Molecular Biology Laboratory (EMBL), and the DNA Databank of Japan (DDBJ) created the International Nucleotide Sequence Database Collaboration (INSDC) to collect and disseminate the burgeoning amount of nucleotide and amino acid sequence data that was becoming available. Since then, the INSDC databases have grown to contain over 95 billion base pairs, reflecting an exponential growth rate in which the amount of stored data has doubled every 18 months (Figure 1). The advent of next-generation sequencing technologies, metagenomics, genome-wide association studies (GWAS), and endeavors such as the 1000 Genomes Project will only increase the tremendous volume and complexity of this and other sequence data collections (Siva, 2008).

Sequence Data Repositories

As previously mentioned, the INSDC is a collaboration of NCBI's GenBank in the U.S., EMBL in Europe, and the DDBJ in Japan. Each of these databases accepts direct submissions of biological sequences from individual researchers, from sequencing projects, and from patent applications from around the world. Sequences are entered into the database and given a unique identification or accession number. These submitted entries are stored in a "library" of records, and each entry is "owned" by—and can only be updated by—its submitter. The data integrated in these entries include the submitter's name, the originating organism, the definition, the actual sequence, related references, and more. (Examples of IDs and and entries are accessible through these links.) The submitted entries are then shared across the three repositories on a daily basis, and releases of the data are made regularly. This has been a boon to the research community, facilitating the sharing of sequence data and allowing the advancement of research.

These sequence repositories have become the universal, comprehensive, and authoritative resources for the exponentially growing amount of sequence data currently available to researchers.

DNA Sequencing

The sequencing reaction is a key technique that forms the basis for Genomic data base. Sanger sequencing technique is primarily used for DNA sequencing. This technique is based on the complementary base-pairing property of DNA. When a single-strand DNA fragment is isolated and places with primers, DNA polymerase, and the four types of deoxyribonucleoside triphosphate (dNTP), a new DNA strand complementary to the existing one will be synthesized. In the DNA sequencing reaction, dideoxyribonucleoside triphosphate (ddNTP) is added besides the above components, and the four types of ddNTPs are bound to four different fluorescent dyes. The synthesis of a new strand will stop when a ddNTP instead ofa dNTP is added. Therefore, with abundant template single-strand DNA fragments, we'll be able to get a set of complementary DNA segments of all different lengths, each one stopped by a colored ddNTP. Under electrophoresis, these segments of different lengths will run at different speeds, with the shortest segments running the fastest and the longest segments running the slowest. By scanning the color of all segments ordered by their length, we'll be able to read the nucleotide at each position of the complementary sequence and therefore read the original template sequence. This technique is implemented in the first generation of sequencing machines.

Next Generation Sequencing

The new technology can read huge amount of shorter DNA sequences at much higher efficiency. The DNA fragments are first cut into short fragments and ligated with some adaptor sequences. Next, *in-vitro* amplification is performed to generate an array of million PCR colonies or "polonies." Each polony which is physically isolated from the others contains many copies of a single DNA fragment.

BLAST

In bioinformatics, BLAST (basic local alignment search tool is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA and/or RNA sequences. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold.

BLAST, which *The New York Times* called *the Google of biological research* is one of the most widely used bioinformatics programs for sequence searching. It addresses a fundamental problem in bioinformatics research. Before BLAST, FASTA was developed by David J. Lipman and William R. Pearson in 1985. Subsequently, Altschul, along with Warren Gish, Webb Miller, Eugene Myers, and David J. Lipman at the National Institutes of Health designed

the BLAST algorithm, which was published in the *Journal of Molecular Biology* in 1990 and cited over 75,000 times.

BLAST is also often used as part of other algorithms that require approximate sequence matching. BLAST is available on the web on the NCBI website. The input sequences are prepared in FASTA or Genbank format. To run the software, BLAST requires a query sequence to search for, and a sequence to search against (also called the target sequence) or a sequence database containing multiple such sequences. BLAST will find sub-sequences in the database which are similar to sub sequences in the query.

The BLAST program can either be downloaded and run as a command-line utility "blastall" or accessed for free over the web. The BLAST web server, hosted by the NCBI, allows anyone with a web browser to perform similarity searches against constantly updated databases of proteins and DNA that include most of the newly sequenced organisms.

There are a number of BLAST programmes available for specific analysis.

i. Nucleotide-nucleotide BLAST (blastn): This program, given a DNA query, returns the most similar DNA sequences from the DNA database that the user specifies.
ii. Protein-protein BLAST (blastp) : This program, given a protein query, returns the most similar protein sequences from the protein database that the user specifies.
iii. Position-Specific Iterative BLAST (PSI-BLAST) (blastpgp)
    a. This program is used to find distant relatives of a protein. First, a list of all closely related proteins is created. These proteins are combined into a general "profile" sequence, which summarises significant features present in these sequences.
    b. By including related proteins in the search, PSI-BLAST is much more sensitive in picking up distant evolutionary relationships than a standard protein-protein BLAST.
iv. Large numbers of query sequences (megablast)
    a. When comparing large numbers of input sequences via the command-line BLAST, "megablast" is much faster than running BLAST multiple times.

However, BLASTn and BLASTp are the most commonly used among the several BLAST programs,

Gen Bank

GenBank (https://www.ncbi.nlm.nih.gov/genbank/submit/) is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and

notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release.

How to submit data to GenBank

The most important source of new data for GenBank ® is direct submissions from scientists. GenBank depends on its contributors to help keep the database as comprehensive, current, and accurate as possible. NCBI provides timely and accurate processing and biological review of new entries and updates to existing entries, and is ready to assist authors who have new data to submit.

Receiving an Accession Number for your Manuscript

Most journals require DNA and amino acid sequences that are cited in articles be submitted to a public sequence repository (DDBJ/ENA/Genbank - INSDC) as part of the publication process. Data exchange between DDBJ, ENA and GenBank occurs daily so it is only necessary to submit the sequence to one database, whichever one is most convenient, without regard for where the sequence may be published. Sequence data submitted in advance of publication can be kept confidential if requested. GenBank will provide accession numbers for submitted sequences, usually within two working days. This accession number serves as an identifier for your submitted your data, and allows the community to retrieve the sequence upon reading the journal article. The accession number should be included in your manuscript, preferably in a footnote on the first page of the article, or as required by individual journal procedures.

What is data submission?

Data reporting is the process of collecting and submitting data which gives rise to accurate analyses of the facts on the ground; inaccurate data reporting can lead to vastly uninformed decision-making based on erroneous evidence.

There are several options for submitting data to GenBank:

- **BankIt**, a WWW-based submission tool with wizards to guide the submission process
- **tbl2asn**, a command-line program, automates the creation of sequence records for submission to GenBank using many of the same functions as Sequin. It is used primarily for submission of complete genomes and large batches of sequences and is available by FTP for use on MAC, PC and Unix platforms.
- **Submission Portal**, a unified system for multiple submission types. Currently only ribosomal RNA (rRNA), rRNA-ITS, Influenza or Norovirus sequences can be submitted with the GenBank component of this tool. This will be expanded in the future to include other types of GenBank submissions. Genome and Transcriptome Assemblies can be submitted through the Genomes and TSA portals, respectively.

- **Sequin**, NCBI's stand-alone submission tool to propagate features from one record to another is available by FTP for use on for MAC, PC, and UNIX platforms. NCBI is phasing out support of the Sequin submission tool.

Further References/ Online Resources:

1. Webinar: A Submitter's Guide to GenBank, Part 1. The webinar was presented December 17, 2014 and outlines using BankIt, a web-based submission tool at NCBI, to submit sequence data to the GenBank® database. Part 2 is scheduled for Jan. 7, 2015.

For more information, see: http://www.ncbi.nlm.nih.gov/education.;
https://www.youtube.com/watch?v=OZxxsRm0pP4

2. How to submit a sequence to GenBank:
https://www.slideshare.net/minhazahmed21/how-to-submit-a-sequence-in

Chapter-13

Insect Metagenomics: Principles and Practices

Sakshi Gandotra, Sagar, D. and S. Subramanian
Division of Entomology, ICAR-Indian Agricultural Research Institute, New Delhi
sakshi262@gmail.com

Metagenomics is a molecular technique used to examine DNA acquired from environmental samples, in order to study the community of microorganisms present, without the necessity of obtaining pure cultures. Metagenomic approaches are often applied in two ways: targeted metagenomics or shotgun metagenomics. In targeted metagenomics, the diversity of a single gene (16S/18S rRNA) is probed to identify the full complememt sequences of that particular gene in an environment. This approach is often employed to investigate both the phylogenetic diversity and relative abundance of the particular gene in a sample. In shotgun metagenomics, the total genomic complement of an environmental community is probed through genomic sequencing. In this approach, environmental DNA is extracted and fragmented to prepare sequencing libraries which are further sequenced to determine the total genomic content of the sample.

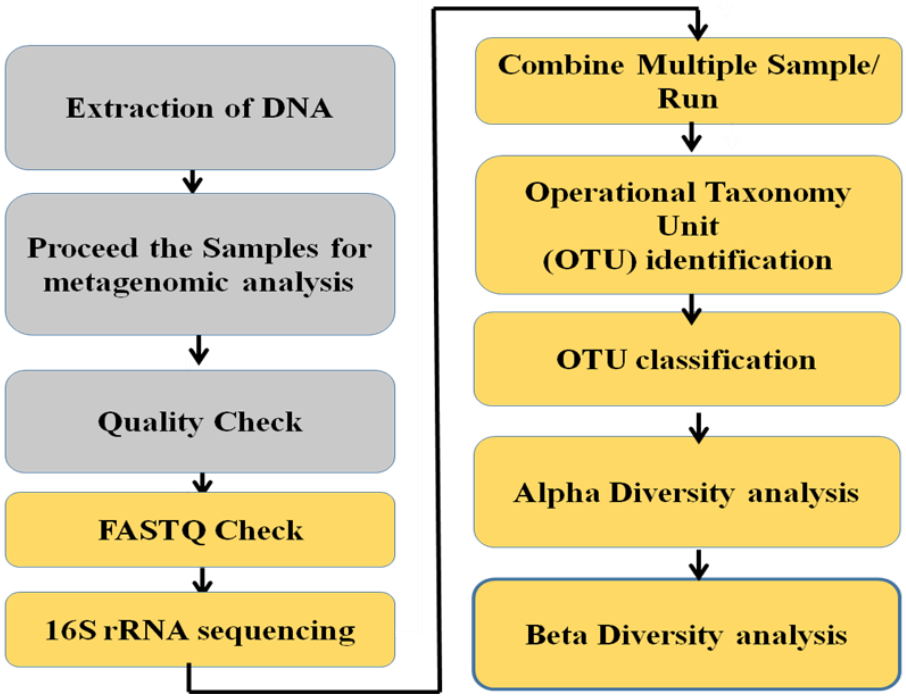
Insect Metagenomics

Insect metagenomics technique provides further knowledge on the functions, structure and diversity of entire bacterial communities prevailing or inhabiting the insects. Metagenomic approaches using 16S rRNA gene yields a better and more inclusive representation community of bacteria and resulted in a spectacular development in understanding of the microbes harboring the insect gut. Next Generation Sequencing (NGS) is one of the most useful and high-throughput sequencing technology that has revolutionized the study of metagenomic and molecular biology. With this high throughput technology it is possible to sequence RNA and DNA faster and less expensive than other first generation sequencing methods such as Sanger sequencing.

Approaches to Metagenomic Analysis of Insect Gut

For metagenomic approach the DNA from insect gut is extracted and the gene of interest is PCR amplified using primers designed to amplify the greatest diversity of sequences for the gene of interest that is 16S rRNA. These PCR products then can be send for sequencing. Further, these amplified genes are sequenced using NGS (Illumina Miseq platform), which results in the thousands of small subunit rRNA reads per sample and can probe hundreds of samples simultaneously metagenomic libraries are screened for novel physiological, metabolic, and genetic features. Although time-consuming and labor-intensive, metagenomic is the most powerful environmental approach that offers possibilities to discover novel genes

and novel biomolecules through the expression of genes from uncultivated and unknown bacteria in recipient host cell.



Design of the experiment followed for metagenomic analysis

Data Processing

Reads are processed for bioinformatics analysis pipeline for different regions of 16S rRNA gene. The Fastq quality check of the sample read is performed. The quality of the bases of the sequences is checked and their composition and GC content is determined. A propriety wet-lab approach is followed for sequencing 16S rRNA V3region of bacteria. The paired-end sequences read contain some fraction of the V3 region, spacer, and conserved region. As a first step, the spacer and conserved region from paired-end reads are trimmed. To identify high-quality V3 region sequences; Spacer sequence filter, Mismatch filters, Conserved region filter, and Read quality filters are used. The reads are combined and clustered into Operational Taxonomy Units (OTUs). The reads from filtered OTUs are processed using QIIME (Quantitative Insights Into Microbial Ecology) program to build a representative sequence for each OTU. The representative sequence is aligned to the Greengenes core set reference databases using PyNAST (Python Nearest Alignment Space Termination program) and the taxonomic classification is assigned to the OTUs. The diversity indices (Shannon, Chao1) are calculated using the QIIME software. The UPGMA (Unweighted Pair Group Method with Arithmetic Mean) phylogenetic trees with Jacknife tests are constructed by calculating the distance matrix by using Weighted and Unweighted UniFrac approaches.

Chapter-14

Terminologies and concept of sequence analysis

Anirban Roy
Advanced Centre for Plant Virology, Division of Plant Pathology, ICAR-Indian Agricultural Research Institute, New Delhi

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. The development of new rapid, inexpensive next generation high-throughput technologies sequencing over the last 10 years or so is changing the ways we think about the application of sequences to plant biology.

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

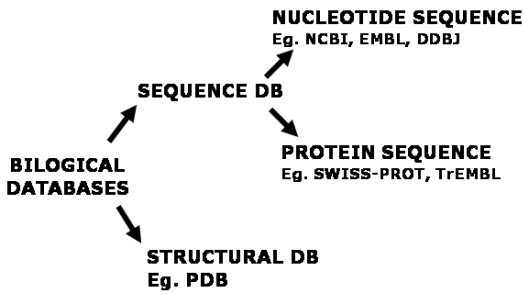
Biological Database:

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 can not 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 can not say there is homology

Homologous Gene Super family

A) Orthologous Gene

Same sequence and Same function but found in different taxa. E.g. - DNA Polymarase of Goat, DNA Polymarase 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 an 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

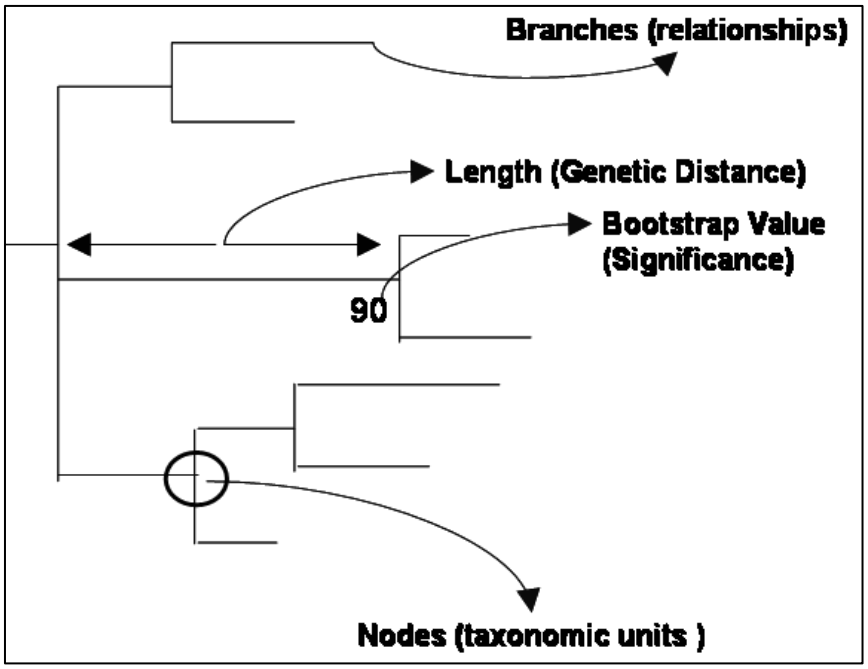
- Sequences that share similarity with query sequence
- Helps to retrieve those sequences

What We Do Not Know From BLAST

- Can not quantify the sequence similarity
- Can not 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

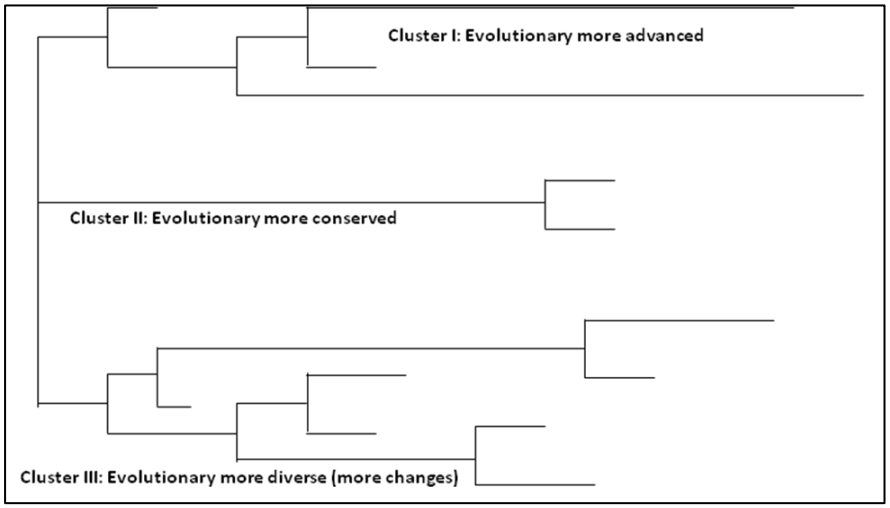


The branching pattern of a tree is called the TOPOLOGY
Representation of relationship through LINE : DENDROGRAM

Types of dendrogram

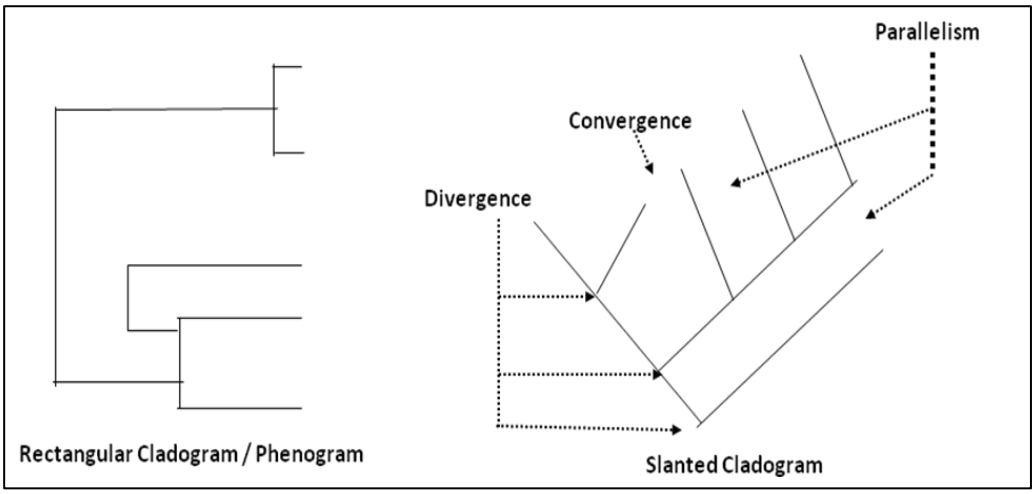
Phylogram

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.



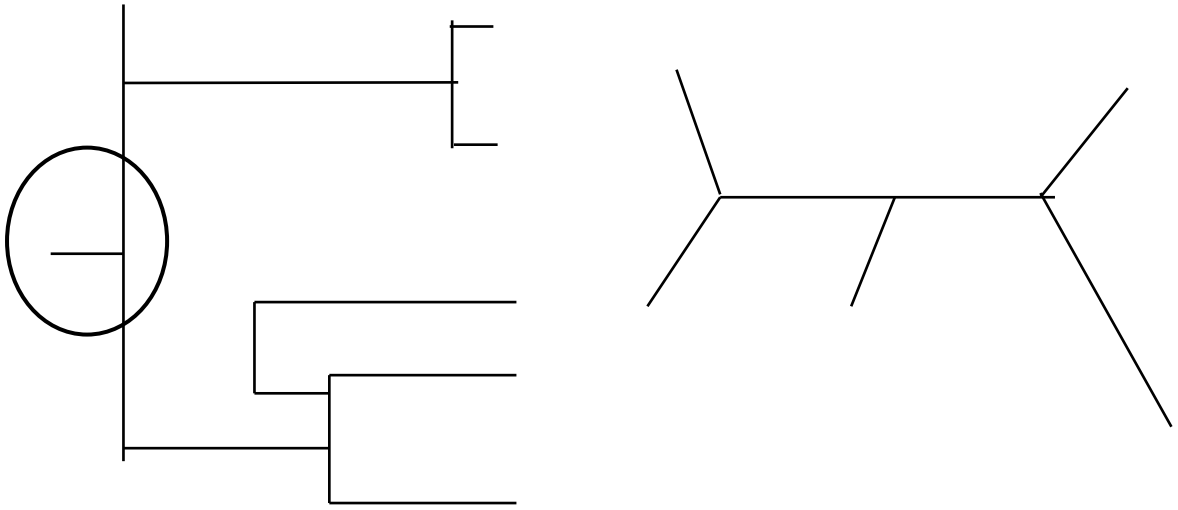
Cladogram

- A dendrogram depicting the hypothesised 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



Rooted **Unrooted**

Methods for constructing phylogenetic tree:

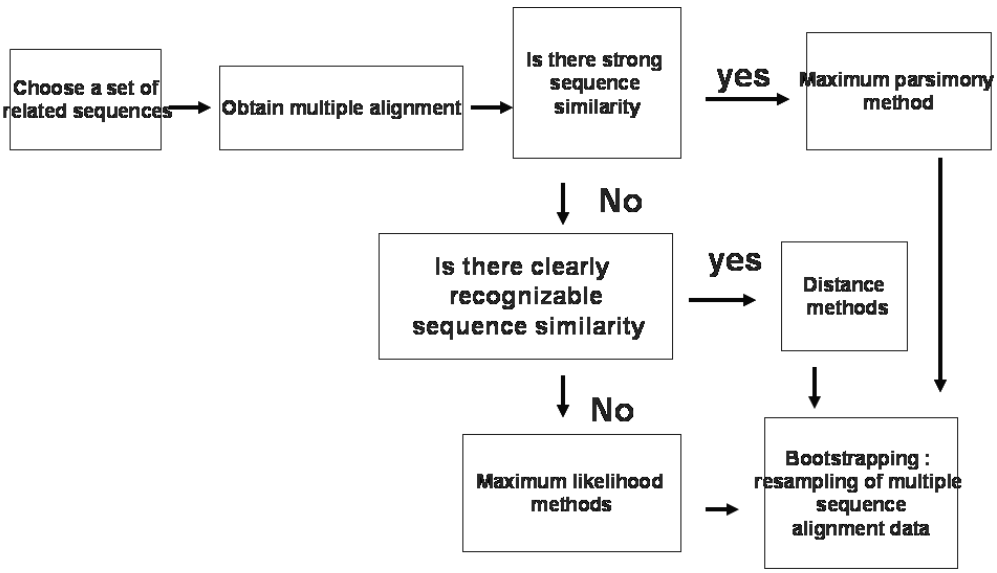
- Character based and distance based method for tree development
- Nebourhood joining tree - distance based method
- Parsimony tree – character based tree. Use when sequences are quite similar, e.g. – strains of different viruses. Use small numbers of sequences for parsimony analysis.

Bootstraping

- 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:



Basic sequence analysis steps

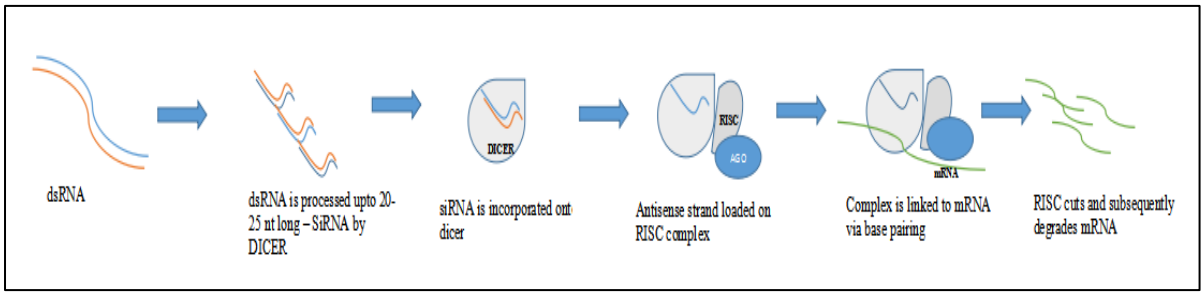
1. After obtaining a sequence purify it from the contamination of vector sequence. Use online service like VecScreen (www.ncbi.nlm.nih.gov/tools/vecscreen/) for the purpose. After an initial idea from VecScreen, carefully see the border region between vector and insert using Bioedit Sequence Alignment 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) or discontiguous 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 concensus phylogenetic tree.

Chapter-15

Evaluation of RNAi constructs by feeding assay against insects

Vinay Kalia
Division of Entomology, ICAR- Indian Agricultural Research Institute, New Delhi
vkalia@iari.res.in

RNAi is a post transcriptional mechanism of silencing gene function by inserting short homologous sequence of messenger RNA (mRNA) to prevent translation of proteins. RNAi is a gene silencing mechanism triggered by double-stranded RNA (dsRNA). In the first *step*, the trigger dsRNA is processed by Dicer (species-specific RNase-III like enzymes) into short interfering *RNA* (siRNA). In the second *step*, siRNAs are loaded into the effector complex *RNA*-induced silencing complex (RISC) which is associated with an Argonaut protein (AGO). The Argonaut protein uses the antisense/guide RNA to associate with target mRNA and then slicing of the target occurs.



An important aspect of the RNA interference pathway is the transport of RNAi information. In insects RNAi can be divided in cell-autonomous [RNA silencing effect takes place within cells where dsRNA is expressed or introduced] and non-cell-autonomous RNAi, the interfering effect takes place in tissues/cells different from the location of application or production of the dsRNA. There are two different kinds of non-cell-autonomous RNAi i.e., environmental RNAi [Signal is picked up by cells from immediate environment such as gut or haemocoel] and systemic RNAi response [when silencing signal spreads to neighbouring cells from epicentre of cells].

For the efficient application of RNAi in insect control, we have to focus on non-cell-autonomous RNAi. *The insect will have to internalize the dsRNA of a target gene through feeding.* In order to silence the target gene, this dsRNA must be taken up from the gut lumen into the gut cells demonstrating environmental RNAi. The midgut is designed to absorb nutrients from the gut lumen with its large absorption area created by the microvilli. These characteristics make the tissue very interesting as a potential dsRNA uptake location. If the target gene is expressed in a tissue outside of the gut, the silencing signal will also have to spread via cells and tissues, which is systemic RNAi.

In this class we will evaluate the efficacy of synthesised dsRNA against *Helicoerpa armigera* by feeding bioassays

Principle: Bioassay refers to the procedure for determination of the relation between a physiologically active agent and the effect, which it produces in a living organism. According to Finney (1952), the term biological assay means the potency of any stimulus, physical, chemical or biological, physiological or psychological by means of the reactions which it produces in living matter. The principle of bioassay is to compare the response of treated insect to those of untreated insect under the same conditions. The response may be based upon mortality, growth inhibition, antifeedant activity etc. Bioassay is affected by experimental conditions therefore success of bioassay depends upon both biotic (type of test insect as well as stage, age, size and sex of insect) and abiotic factors (temperature, humidity, amount of illumination, amount and type of food).

Materials Required

1. ds RNA construct; 2. Negative control dsRNA construct; 3. neonates of *H. armigera;* 4. Diet; 5. Microapplicator; 6. Plastic containers; 7. Tissue role; 8. Spatula; 9 nuclease free water and Pestle mortar

Method

1. Prepare different concentrations *viz.*, 0.5, 1.0 and 5.0 µg of dsRNA/g of diet using stock solution suitably diluted to volume of 0.1 ml nuclease free water to mix uniformly in 10 gm of diet.
2. As given above do similarly for –ve control dsRNA construct. In control use nuclease free water only
3. After mixing, divide the diet in three parts and transferred to small plastic containers (5×2 cm). Each container serve as one replicate, with three replications per concentration.
4. Release ten neonates on the treated diet (3g diet) per replication and fed for four days after that they were separated and transferred individually on fresh untreated diet.
5. Record observations on mortality and phenotypic effect after every 24 h up to the adult emergence.
6. As given in step 1 to 4 repeat and kept the treated as well as control sample separately.
7. Collect samples for expression analysis from second fraction after every 24 hours till 96 hours from treatment and control.
8. Wash collected larvae with 70% alcohol followed by nuclease free water and kept in RNAse free tube in -80^{o}C in RNAse later™ for further use expression analysis by Real Time PCR.

Datasheet for Bioassay

1. Name of dsRNA construct (gene) ----------
2. No. of concentration/dose ----------
3. List of concentrations ----------
4. Control (nuclease free water) ----------
5. Negative control (unrelated dsRNA) ----------
6. Insect name ----------
7. Stage of insect ----------
8. No. of insect / concentration ----------
9. Total No. of insects including control ----------
10. Location of insect source if laboratory; ----------
11. Method of Bioassay:
    Diet incorporation/Diet overlay ----------
12. Conditions of bioassay:
    Temperature ----------
    Humidity ----------
    Light:Dark period ----------
13. Duration of bioassay ----------

14. Observation Record:

S.No.	Conc. of Toxin	Mortality/total No. of insects				Total Mortality
		24h	48h	72h	96h	
1.	Control					
2.						
3.						
4.						
5.						
6.						
7.						
8.						
9.						
10.						

15. Analyze the mortality data by probit analysis; Calculate LC_{50} and LT_{50} using MLP (Ross, 1977)
16. Observation on growth inhibition : Take weight of 10-15 larva: control and treatment

References

Finney, D.J. (1952). *Probit analysis*. Cambridge University Press. 2nd Ed.

Ross G.E.S. (1977). *Maximum likelihood programme*. The numerical algorithms Group, Rothamsted Experiment Station, Harpenden, UK.

Chapter-16

Estimation of detoxification enzymes associated with insecticide resistance in insects

Rajna S, Sagar. D and S. Subramanian
Division of Entomology, ICAR-Indian Agricultural Research Institute, New Delhi
email: rajnasalim@gmail.com

Introduction

The indiscriminative and nonjudicious use of insecticides in field conditions is causing a high risk of development of resistance to insecticides in many of the field pests. The development of resistance towards insecticide is majorly due to three basic mechanisms *viz*., decreased penetration of insecticides into insect system, enhanced detoxification of insecticides, and target site insensitivity. Among these, detoxifying enzymes play an important role in the development of resistance. Three enzymes *viz*., esterase, cytochrome P450 monoxygenase and glutathione S transerase are playing major role in detoxifying the insecticides. In this chapter, we discuss the procedure for detection of these detoxification enzymes using microplate reader.

Materials required

Microplate reader, micropestle, microcentrifuge tubes, hand held homogenizer, sodium and potassium phosphate buffers with different molarity and pH as in protocol, Triton X 100, substrates and co-substrates for the assays as described in the protocol

Procedure

Protocols for three different assays are described here.

Estimation of esterase activity: Esterases are generally known as hydrolases which detoxifies insecticides through hydrolysis, yielding acid and alcohol. Here, the esterase activity is measured by microplate assay using α–naphthyl acetate as the substrate, which is converted into α naphthol by esterase and the product is measured at 450nm (Guo *et al.*, 2013).

Steps

1. Homogenization of the sample in 250 µl ice cold sodium phosphate buffer (0.1 M, pH 7.5), containing 0.1% (w/v) triton X-100 using a handheld homogenizer.
2. The samples are centrifuged at 10,000g at 4^{o}C for 10 min and the supernatant is collected in a fresh tube and use as the enzyme source.
3. *Preparation of substrate solution*: The substrate solution (1 mM α-naphthyl acetate) is prepared in 0.2M, pH 6.0 sodium phosphate buffer containing fast blue RR salt (15 mg/ 25 ml)

4. *Enzyme reaction mixture*: The enzyme assay is carried out by adding 25 µl sodium phosphate buffer (0.1 M, pH 6.0) in microplate well followed by 50 µl enzyme source and 200 µl substrate solution.
5. Microplate wells with buffer and substrate solution alone is considered as blank
6. Optical density values can be measured at 25^0C for 10 min, giving 1 minute interval at 450nm.
7. Estimation of enzyme activity using standard curve of α naphthol.

Estimation of Cytochrome P450 monoxygenase activity: In this method, the cytochrome P450 monoxygnease activity is estimated as general oxidase level, which is an indirect measure of cytochrome P450 using heme peroxidation. Here, heme peroxidation, along with the substrate 3,3',5,5'-Tetramethylbenzidine (TMBZ), Hydrogen peroxide is used as co-substrate. With the presence of H_2O_2, the microsomal oxidases utilize the TMBZ and develop two oxidized TMBZ molecules (Brogdon *et al.*, 1997; Penilla *et al.*, 2007).

Steps

1. Homogenization of the sample in 250 µl ice cold 250 µl ice-cold potassium phosphate buffer (0.625 M, pH 7.2), containing 0.1% (w/v) Triton X-100 using a handheld homogenizer.
2. The samples are centrifuged at 10,000g at 4^0C for 10 min and the supernatant is collected in a fresh tube and can be used as enzyme source.
3. *Substrate solution*: The TMBZ substrate solution is prepared by dissolving 2 mg of TMBZ tablet in 2.5 ml of methanol and 7.5 ml of 0.25 M sodium acetate buffer (pH 5.0).
4. *Enzyme reaction mixture*: The reaction mixture in the microplate consisted of 80 µl of 0.625 M potassium phosphate buffer (pH 7.2), 20 µl of enzyme source, 200 µl TMBZ solution, 25 µl of 3% H_2O_2
5. Reaction mixture without enzyme source is considered as the blank.
6. Absorbance measurement after 5 min incubation at room temperature at 650 nm in microplate reader.
7. A standard curve for heme peroxidase activity was prepared using different concentrations of cytochrome C from bovine heart.
8. The cytochrome P450 (general oxidase) activity obtained from plate reading was expressed as equivalent units (EU) of cytochrome P450 per milligram of protein using the standard curve of cytochrome C.

Estimation of GST activity: The GST assay protocol is based on the GST-catalyzed reaction between GSH (Glutathione reduced) and the GST substrate CDNB (1 cholro 2,4 dinitrobenzene). The GST-catalyzed formation of GS-CDNB produces a dinitrophenyl thioether which can be detected by spectrophotometer at 340 nm Habig *et al.* (1974).

Steps

1. Homogenization using 400µl of 0.2M sodium phosphate buffer, pH 7.5 (1mMof EDTA, 0.1mM of DTT, 1mM of PTU and 1 mM of PMSF and 0.1% glycerol).
2. The samples are centrifuged for 10 min at 10,000 g at 4^0C and the resulting supernatant will be taken as enzyme source.
3. *Substrate preparation*: The CDMB and GSH substrate solutions are prepared in 0.2M sodium phosphate buffer (pH 6.5) with concentrations 0.4 mM and 4 mM respectively (Guo *et al.*, 2013., Nauch and Nauen 2005).
4. *Reaction mixture*: The reaction mixture will be prepared in microplate wells by adding 100 µl enzyme source, 100 µl CDMB and 100 µl GSH.
5. The absorbance is measured after 1 min incubation at room temperature for 10 minutes giving 1 minute interval at 340 nm and 25^0C in microplate reader.
6. The wells containing all reaction components, except enzyme source is used as blank.
7. A change in absorbance per minute is converted into µmol CDNB conjugated/ min using the molar extinction coefficient ($^{\varepsilon}$molar) of the resulting 2, 4−dinitrophenyl−glutathione: $^{\varepsilon}_{340nm}$ = 9.6 mM^{-1} cm^{-1} (Habig *et al.*, 1974).

$$\text{CDNB-GS in µmol/min} = \frac{\text{ABS (increase in absorbance after time period)} \times \text{Volume of total solution used} \times \text{Conversion factor}}{\text{Extinction coefficient of 2, 4−dinitrophenyl−glutatione} \times \text{Time of run}}$$

Protein estimation: The specific activity of the enzyme is expressed as µmoles or nanomoles of product/min/ mg protein. Protein content of the enzyme samples is determined following Bradford method using bovine serum albumin (BSA) as the standard.

Conclusion

The knowledge on the biochemical mechanisms on resistance can provide insight into the molecular mechanisms underlying and also to develop new molecules with novel modes of action which will be helpful in overcoming insecticide resistance.

References

Brogdon, W.G., McAllister, J.C. and Vulule, J. (1997). Heme peroxidase activity measured in single mosquitoes identifies individuals expressing an elevated oxidase for insecticide resistance, *J. Am. Mosq. Control Assoc*., 13: 233–237.

Habig, W.H., Pabst, M.J., Fleischner, G., Gatmaitan, Z., Arias, I.M. and Jakoby, W.B. (1974). The identity of glutathione S–transferase B with ligand in a major binding protein of liver, *Proc. Natl. Acad. Sci.*, 71: 3879– 3882.

Penilla, R.P., Rodriguez, A.D., Hemingway, J., Trejo, A., Lopez, A.D. and Rodríguez, M.H. (2007). Cytochrome P450 based resistance mechanism and pyrethroid resistance in the field *Anopheles albimanus* resistance management trial, *Pestic. Biochem. Physiol*., 89: 111–117.

Chapter-17

Gene Silencing and Genome Editing in Insect Pest Management

R. Asokan
Division of Biotechnology, ICAR-Indian Institute of Horticultural Research, Bengaluru–560089, Karnataka
asokaniihr@gmail.com

The noble cause of feeding 125-crore hungry stomachs is the Herculean task faced by the Indian agriculture today and the population is slated to increase by leaps and bounds in the ensuing years. This calls for more land to be brought under cultivation and also increasing productivity of crops. While there is no scope to further increase the area under cultivation, increasing the productivity is only the plausible, viable option to meet the demand. It has been estimated India's population will have roughly 200-crore by 2040 and we have to go a long way from the current production level of 800 million tons of major crops. Insecticides account for and. excessive reliance on chemical insecticides (61% of the total pesticides usage) has already resulted in many control failures due to accelerated development of resistance and in contamination of soil, water and adversely affected non-target organisms. Therefore there is an urgent need to look for an effective, ecofriendly alternative for insect pest management, where a new and novel approach called RNA interference (RNAi), called *gene silencing* and recently, *genome editing* poised to play a vital role. Employing this, it has now become possible to control insect pests by silencing or editing some of the vital genes that play an important role in insect-host plant interaction, growth & development, flight, reproduction etc by delivering cognate double stranded RNA (dsRNA) either as spray or through transgenic plant or through *gene drive*. These approaches are not generic but has to be species specific.

RNA interference

RNA interference (RNAi) is a generic term where the silencing is brought about by small interfering RNA (siRNA) derived from endogenous or exogenous double stranded RNA (dsRNA) or micro RNA (miRNA). RNAi is a nucleotide based defense mechanism for maintaining genome integrity and is more elaborate in the basal genomes. Initially RNAi has been employed as a laboratory tool for identification of gene functions which later became an important tool in various spheres of scientific research that include insect pest management. One important observation by Fire & Mello, 1998 in the free-living nematode, *C. elegans* that exogenously applied double stranded RNA (dsRNA) could also elicit RNAi response, caused the paradigm shift in the realization potential of RNAi. RNAi occurs at two levels viz. transcriptional gene silencing (TGS) which occurs in the nucleus and as post transcriptional gene silencing (PTGS) which occurs in the cytoplasm. The later involves degradation of target messenger RNA (mRNA) through the production of small interfering RNAs (siRNAs) from

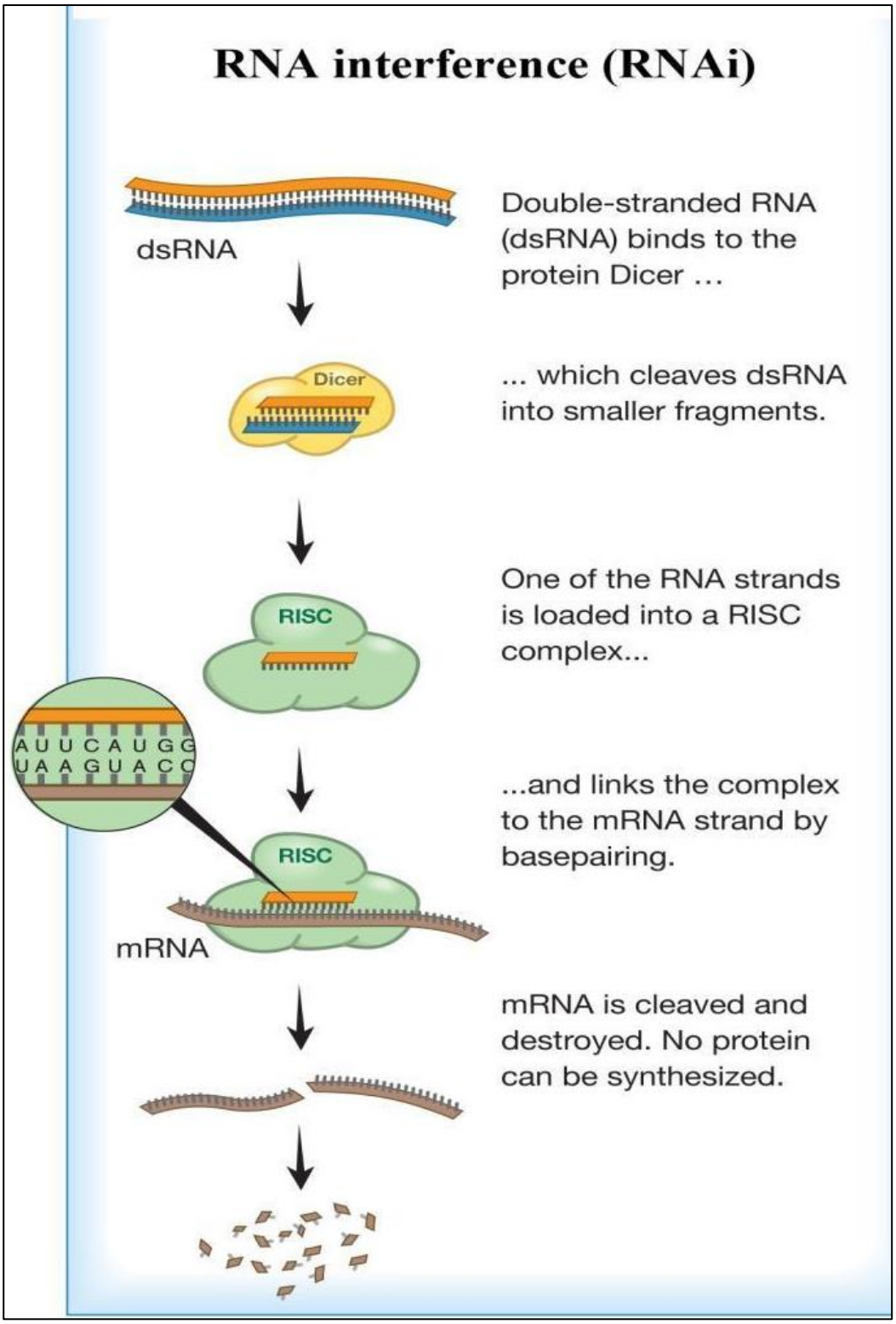
the dsRNA, which is cleaved by dsRNA-specific endonucleases referred to as dicers (RNaseIII). The siRNAs are 21 bp dsRNA fragments carrying two base extensions at the 3' end of each strand; one strand of the siRNA is assembled into an RNA-induced silencing complex (RISC) in conjunction with the argonaute multi-domain protein, which contains an RNaseH-like domain responsible for target degradation. The process is closely related to post-transcriptional gene regulation by micro RNAs (miRNAs), where the end-result is inhibition of translation initiation, and shares many of the same components. In plants and nematodes, RNAi can have systemic effects on gene expression, so that gene knockout spreads throughout the organism and persists over development. The basis of this effect is thought to lie in the presence of an RNA-dependent RNA polymerase (RdRP) that is able to interact with the RISC complex and generate new dsRNA based on the partially degraded target template by using the hybridized siRNA strands as primers. The synthesized dsRNA is then acted on by the dicer enzymes to generate new siRNAs (secondary siRNAs), thus acting as an amplification step. In this way, once a dsRNA is introduced into a cell, its effect can persist over development; in addition, the dsRNAs can be exported to neighbouring cells and thus spread the gene knockout effect through the organism.

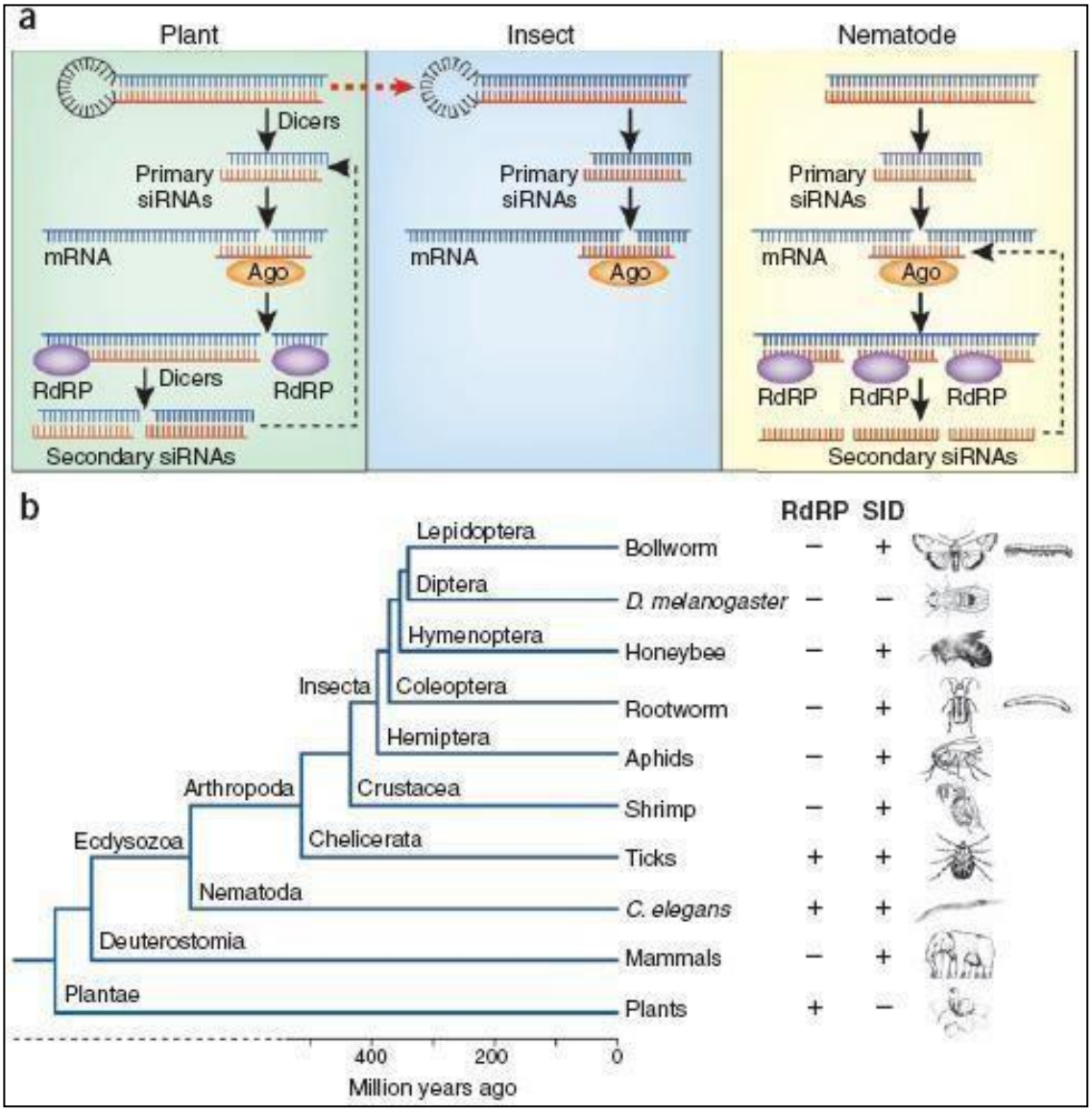
RNA interference was first observed in petunias, when Napoli *et al.* (1990) discovered that introduction of a pigment-producing gene under control of a powerful promoter suppressed expression of both the introduced gene and the homologous endogenous gene, a phenomenon they called co-suppression. Co-suppression was subsequently found to occur in many species of plants and fungi and to occur at the post-transcriptional level. Such post-transcriptional gene silencing (PTGS) was shown to be mediated by a diffusible, trans-acting molecule in both *Neurospora* and plants.

The first miRNAs to be discovered, lin-4 and let-7, were identified through loss-of-function mutations affecting control of postembryonic development in *C. elegans*. Other miRNAs were soon identified in *C. elegans*, *Drosophila*, and mouse by combinations of forward genetics, cDNA cloning, bioinformatics, and reverse genetics. By 2002, miRNAs were firmly established as a large class of conserved regulatory molecules in animals and plants. miRNAs have been shown to play a role in developmental timing, cell death, cell proliferation, and oncogenesis. This class of small RNA may represent 2–3% of the total number of genes in humans, and estimates of miRNA target-binding sites indicate that miRNAs may play a role in regulating as many as 30% of mammalian genes. Dicer, was shown to process both long dsRNA into siRNAs and cytoplasmic miRNA precursors (pre-miRNAs) into mature miRNAs. Other components of the RNAi and miRNA silencing pathways have been shown to be closely related, most importantly Argonaute, which is a key component of the RNA-induced Silencing Complex (RISC) and miRNAs function in a silencing complex that is similar, if not identical, to RISC to regulate expression of target genes either through cleavage of mRNA or translational repression: if the miRNA exhibits perfect complementarity to its target mRNA, the mRNA is cleaved (typically in plants), if there is only partial complementary, translational repression occurs (typically in animals). The further discovery of endogenous small RNAs, distinct from miRNAs, that function in transcriptional silencing and genome stability has

driven the adoption of the more general terms, RNA silencing and small RNAs to describe the collection of related silencing pathways and their RNA guides.

Figure 1: General mechanism of RNAi



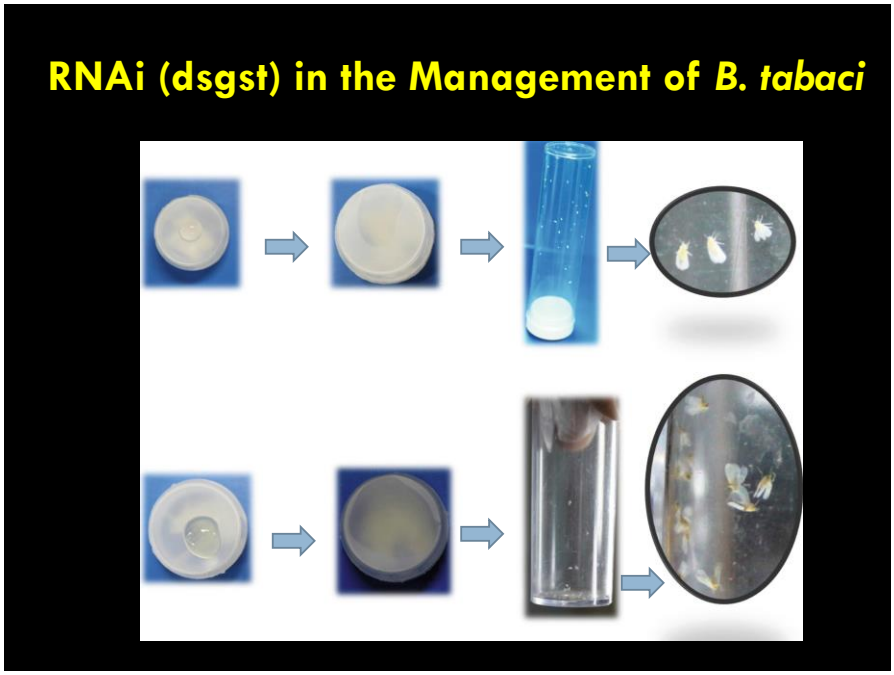


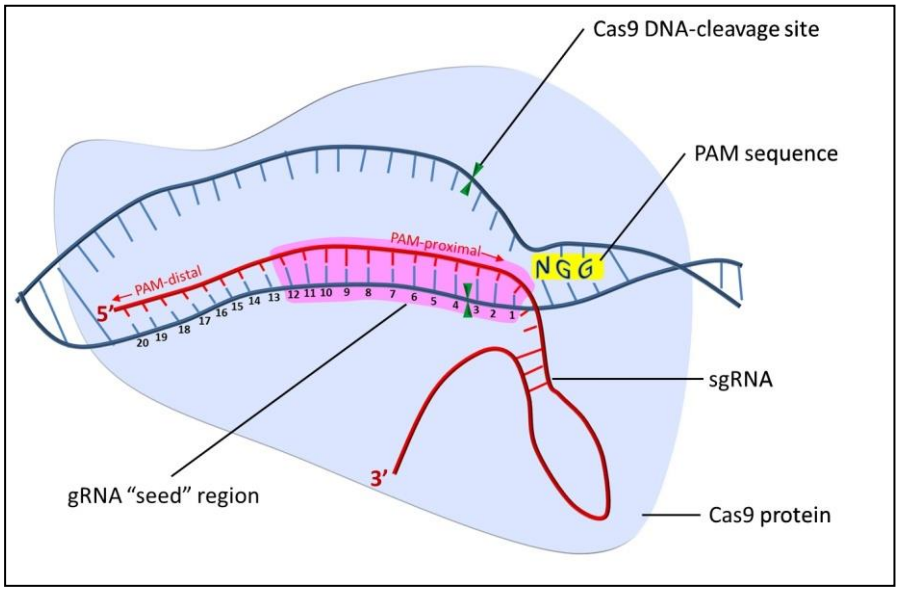
(Source: Gordon and Water house, 2007

Applications of RNAi in insect pest management

Most of the studies on RNAi in entomology were on establishing the function of genes that are involved in various metabolic processes. But many researchers envisaged the potential of RNAi for field level insect pest management, but success was not forthcoming due to many reasons like poor understanding on the mechanism of RNAi in different orders of insects, amplification and systemic spread of the silencing signal in the treated insects, amenability of different orders of insects for the RNAi approach etc. In the year 2007 two research groups from China and USA experimentally validated the utility of RNAi in insect pest management. The choice of the gene for RNAi mediated silencing could be many and generally falls into two broad categories viz. silencing of genes that results in quick control e.g. genes involved in insect-

plant interactions, digestion, moulting etc and silencing of genes that results in long term management e.g. genes involved in pheromone biosynthesis, pheromone reception, migration, flight, diapause etc. Mao *et al.* (2007) demonstrated that it is possible for a no chemical management of the cotton boll worm, *Helicovera armigera* by transgenic cotton mediated silencing of cytochrome P450. Similarly Baum *et al.* (2007) proved that RNAi is approach is feasible in the management of the coleopteran pest the corn root worm (*Diabrotica virgifera virgifera*). Likewise different sets of genes could be targeted depending upon the requirement. Similarly there are many examples of application of dsRNA in insect pest management and some of them are on the sweet potato whitefly, *Bemisia tabaci*; melon aphid, *Aphis gossypii*, diamondback moth, *Plutella xylostella*.





Genome editing

In the year 2015 yet another important development happened in Biology which paved the way to edit the genomes at will. This revolutionary technology is known as CRISPR/Cas system originally discovered in bacteria. Naturally, the bacteria defend against invading DNAs with an adaptive immune strategy using a class of RNA-guided nucleases. The RNAs are generated from a Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) in bacterial genomes. The CRISPR-associated (Cas) nucleases are then guided to foreign DNA targets by these CRISPR RNAs (crRNAs) by the recognition of protospacer adjacent motifs (PAMs) in the foreign DNA sequence. After complementary DNA-RNA base paring, the Cas nuclease cleaves the target DNA by forming site-specific double-stranded breaks (DSB). Therefore, the defensive CRISPR/Cas system provides researchers with a promising tool for specific genome editing, as in vivo DSB will be repaired mainly through non-homologous end joining (NHEJ) or homology-directed repair. (HDR) pathways, introducing insertions or deletions. The most explored CRISPR/Cas system for genome editing to date employs the *Streptococcus pyogenes* Cas9 nuclease, which can recognize the PAM sequence NGG (N is A, T, C or G). Regarding its application in insect research, CRISPR/Cas9 has fallen behind its exploding popularity in mammalian studies. However, CRISPR/Cas9 applications in insects have been reported increasingly, mainly in *Drosophila melanogaster, Bombyx mori* and *Aedes aegypti*. The ease of retargeting Cas9 by simply designing a short guide RNA sequence combining the crRNA and transactivating crRNA has enabled large-scale genome experiments to determine the gene functions. This technology is not only becoming more and more popular in functional genomics studies, but it is also being exploited as a tool for the potential control of pest insects and vector-borne diseases. Before the CRISPR/Cas9 system was discovered, zinc finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) technologies were used for genome modification; both technologies can be used to design a DNA-binding domain that can effectively recognize and modify virtually any sequence, and both technologies have been widely applied in various fields. ZFNs and TALENs, however, require the use of a variety of nucleases, and the off-target effects of nucleases can lead to cellular toxicity. In addition, methods using ZFNs and TALENs are complex and labor-intensive. Compared with RNA interference (RNAi) technology, CRISPR/Cas9 generates changes at the genomic level that are stable and heritable, and the mutant gene can be transmitted to the next generation, while gene silencing by RNA.

Conclusion

The recent demonstration of the potential of RNAi and genome editing in insect pest management has opened a new avenue which will fuel a futuristic approach where application of chemical insecticides will no longer be needed. Identification of suitable gene targets and genome targets and delivery method will usher a new ecofriendly approach that is safer to non-target organisms including humans.

Chapter-18

Enzyme kinetics – an indispensable tool for understanding metabolic pathways

Sandeep Kumar, Minnu Sasi and Anil Dahuja
Division of Biochemistry, ICAR-IARI, New Delhi-110012

Introduction

A catalyst is an agent that speeds up the rates of reactions without themselves undergoing any permanent change. In biological systems, macromolecules called the enzymes act as catalysts, which work by lowering the activation energy of the reaction. The enzymes are highly specific not only for their substrates, the substances that are acted upon by enzyme, but also for the type of reaction they catalyze. This property of the enzymes allows them to distinguish between different substances - even the optical isomers of a compound. Each given cell in an organism contains large number of enzymes because every reaction taking place in a cell is catalyzed by a particular enzyme. In the absence of enzymes most of the reactions of cellular metabolism would come to standstill. It is thus difficult to imagine life on this planet without enzymes.

Urease was the first enzyme to be purified and crystallized by James B. Sumner in the year 1926, which was found to be a protein later on. This led to the development of a concept that all enzymes are proteins; however, later on it was found that some RNA molecules, known as Ribozymes, do possess catalytic activity. The enzymes can be monomeric or oligomeric depending upon whether they contain one or more polypeptides in their structure. Many enzymes require a non-protein component for carrying out the catalysis. This non-protein component is known as coenzyme while the combination of proteinaceous enzyme with coenzyme is known as holoenzyme. The holenzyme without coenzyme is known as Apoenzyme. A coenzyme e.g. Nicotinamide Adenine Dinucleotide (NAD) can associate to different apoenzymes and act on different substrates. The metal ions are essential for the activity of some enzymes and can be involved in the catalytic process through their ability to attract or donate electrons. They may also contribute to maintain the tertiary and quaternary structures of the enzyme molecules or they may help in binding with the substrate by coordination links.

Sometimes, several different enzymes with complementary actions form organized complexes known as "Multienzyme Complexes". In these complexes various enzymes are arranged in such a way that the product of the reaction catalyzed by the first enzyme is received as substrate by the second enzyme and from this to subsequent enzymes. The sequence of the action of enzymes is determined by their spatial arrangement within the complex. Such kind of arrangement of enzymes improves the efficiency of the system as a whole, as intermediates are

never released in the medium and so the chances of their modification or being lost are the least. The respiratory chain present in the inner mitochondrial membrane is the perfect example of a Multienzyme Complex. The other category of complex enzyme existing in the cell is multifunctional enzymes, in which several different catalytic sites are present on a single polypeptide chain. An example of this type of enzyme is fatty acid synthase.

In general enzymes are synthesized in the cell cytoplasm and are then either exported to place of action within the cell or secreted outside the cell and fulfills their mission there itself. Most of the enzymes are intracellular and are localized in different intracellular compartments, where they perform their specific functions.

Enzyme catalyzed reactions

During the enzyme (E)-catalyzed conversion of substrate (S) into product (P), the enzyme and substrate forms a complex ES, which then dissociates in to enzyme and product:

$$E + S \leftrightarrow ES \rightarrow E + P$$

To form the ES complex, the substrate is attached to a defined place on the enzyme known as active site. The active site is constituted by a binding and a catalytic site both of which are having a highly specific three dimensional structure. The active site is composed of a small number of precisely distributed amino acids, whose side chains play an essential role in enzyme catalysis. The amino acid residues that participate in the formation of the active site are sometimes located far away from each other in the polypeptide chain but these are brought close to each other during the folding and twisting of the polypeptide chain. The binding of the substrate on the binding site occurs in such a way that the groups in the substrate which need to be modified during the reaction are positioned exactly at the catalytic site. Once the ES complex is formed, the enzyme undergoes a conformational change to form a transition state, from which it can easily be converted into product(s).

Some enzymatic reactions involve two or more different substrates. In these cases, the active site of the enzyme provides a favorable habitat to each substrate to interact with the other in perfect orientation, promoting the formation of the transition state, reducing the activation energy, and increasing the reaction rate. Many of us wonder if active site of the enzyme is self-sufficient to carry out the reaction, then why macromolecules (Protein and RNA) and not the smaller molecules have been selected to function as enzymes during the course of evolution. The most satisfactory explanation to this question would be that it is very difficult for small molecules to acquire a three-dimensional configuration required to accommodate a given substrate and create an accurate environment for its requisite modification for conversion into product.

Two hypotheses have been put forward so far and successfully accepted to explain the mechanism of action of enzymes viz. "lock and key" and "induced fit hypothesis, while the former demands a perfect structural complementarity between enzyme and substrate for the

formation of ES complex and consider enzyme to be having a rigid and static structure but the latter regards the enzyme structure to be more flexible allowing it maneuver its confirmation when it comes in contact with the substrate, adapting to it and orienting essential residues to obtain the optimal conformation to form the ES complex.

Enzyme kinetics- measurement of enzyme activity

The activity of an enzyme can be determined in a medium by measuring the amount of product formed or the amount of substrate consumed by the enzyme in a given time. The assay method could be continuous or discontinuous. In discontinuous method, the rate of the enzymatic reaction could be monitored by removing samples from the reaction mixture at known times after addition of enzyme, stopping the reaction quickly and measuring the amount of product formed. The continuous method involves the continuous measurement of some property (absorbance, fluorescence, refractive index, chemiluminescence etc.) that changes during the course of the reaction. If there is no change in any of the property, then reaction can be coupled to some reaction where change in property is observed.

However, following precautions should be taken to obtain the reliable estimate of the enzyme activity in both the above mentioned methods:

- The substrate, buffers etc. should be of as high purity as possible, since contaminants may affect the activity of the enzyme.
- It must be ascertained that enzyme preparation, does not contain any compound that interferes with the assay.
- Enzyme should be stable during the time taken for assay.
- The pH, temperature should remain constant during the assay of enzyme.

The enzyme activity depends on the total amount of enzyme present in the sample and is not influenced by changes produced in the reaction mixture by the action of enzyme on substrate provided the activity is measured at initial velocity.

The initial velocity refers to the time in the enzymatic reaction in which the amount of substrate used by the enzyme is still negligible compared to the total substrate existing in the mixture. It is accepted that the enzyme activity is determined at initial velocity before the substrate consumption has reached 20% of the total originally present in the sample; however it is preferable to set initial velocity at a lower limit (about 5%).

The amount of enzyme is usually specified in International Units, which is defined as the amount of enzyme that catalyzes the conversion of one micromole of substrate per minute under defined conditions of pH, temperature, and pressure. The specific activity indicates the relative purity of the enzyme preparation. It correlates enzymatic activity, not to the volume, but to the total protein present in the sample. It is defines as the units of enzymes per milligram of protein in the sample. The other parameter which is very important in this context is turnover number (K_{cat}) i.e. the number of substrate molecules converted into product by an enzyme

molecule when fully saturated with substrate in a unit time. The number of enzyme molecules can be easily calculated from amount of purified enzyme, if its molecular weight is known.

Factors Affecting Enzyme activity

Enzyme concentration: Enzyme activity is directly proportional to the enzyme concentration. However, this relationship holds true only if initial velocity of the enzyme is determined using saturating amounts of substrate and reaction is not prolonged beyond the time required to consume more than 10% of the originally present substrate.

Substrate Concentration: To study the effect of substrate concentration on enzyme activity, the enzyme activity is determined at a single enzyme concentration but under varying substrate concentrations while maintaining all other reaction conditions constant. Then a plot as shown in figure 1 is obtained, which clearly indicates that relationship between substrate concentration and enzyme activity is hyperbola. The hyperbolic curve is a characteristic of single substrate reactions. However, the hyperbolic curve can also be obtained for two-substrate reactions, if concentration of one of the substrate is taken in excess. The Figure 1 also indicates that after a certain level of the substrate concentration, there is no more increase in the enzyme activity, howsoever more substrate one may add in the reaction mixture; the reaction behaves as zero order reaction. The curve at this point becomes horizontal and the enzyme is said to have achieved its maximum velocity (V_{max}) at this point. At this point, the enzyme becomes saturated with substrate and virtually all enzyme molecules are present in the enzyme-substrate (ES) complex form. Theoretically V_{max} is possible only at the infinite concentration of the substrate and hence difficult to predict. To establish a precise relationship between initial velocity and substrate concentration, Michaelis and Menten defined a constant called K_m (Michaelis constant). The Km corresponds to substrate concentration at which rate of the reaction is half the maximum. The Michaelis and Menten described the hyperbolic relationship between enzyme and substrate by the following equation:

$$v = \frac{Vmax\ [S]}{Km + [S]}$$

where v corresponds to the initial rate at substrate concentration [S].

For the determination of K_m and V_{max} a double reciprocal plot, popularly known as Lineweaver Burk Plot, as shown in Figure 2, is obtained in which inverse of initial velocity (1/v) is plotted against inverse of substrate concentration (1/[S]). There are other powerful graphical methods available to calculate V_{max} and K_m but Lineweaver Burk Plot is still the most commonly used one. The Km for different enzymes is different but for a particular enzyme its value is characteristics of that enzyme and varies for each of its substrates when determined under the same conditions of temperature and pH. The K_m value of most enzymes is taken as measure of its affinity with the substrate and both are inversely related. In general, lower the km value higher is its affinity for the substrate.

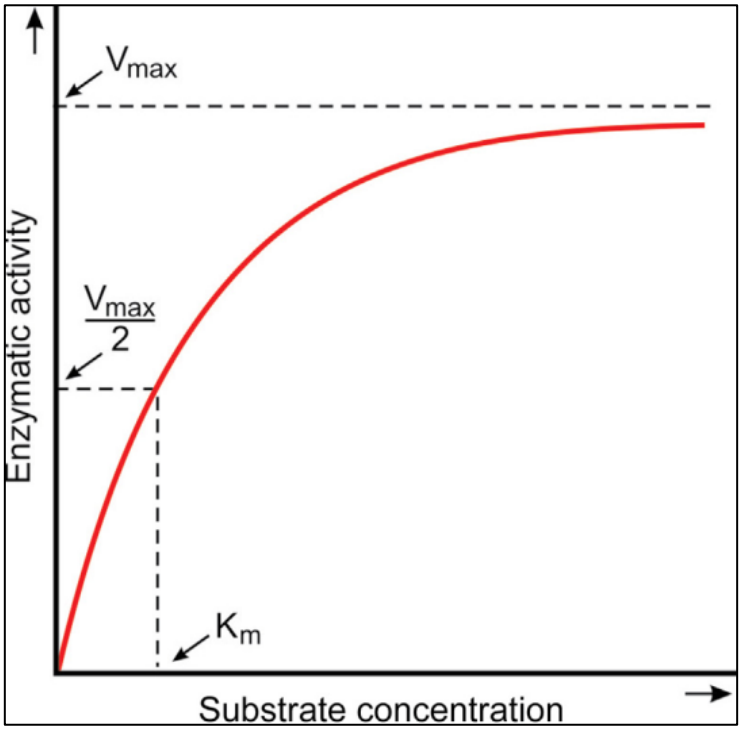


Figure 1: Effect of substrate concentration on enzyme activity

When an enzyme acts on several substrates, the K_m value is usually different for each of them. The substrate with lower Km value is usually considered the natural or physiological substrate for that enzyme. The parameters K_{cat}, K_m and V_{max} are of great practical significance. For example, the ratio of K_{cat}/ K_m can be utilized to compare catalytic efficiency of the enzymes; higher the ratio more is the catalytic efficiency. The role of an enzyme in metabolism can also be judged by relating the Km value to the prevailing concentration of the substrate.

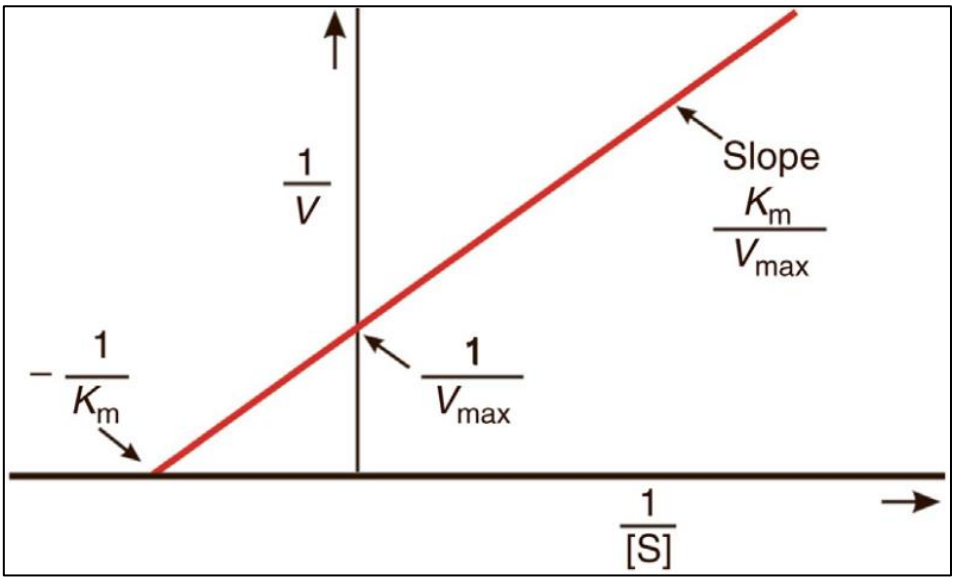


Figure 2: Lineweaver-Burk Plot for the determination of K_m and V_{max}

Temperature: The rate of a chemical reaction increases with the rise in temperature, as a result of increase in the kinetic energy of the system. Although enzyme activity increase with temperature, a maximum value is reached; this corresponds to the optimal temperature for the catalytic activity of a given enzyme. Above this temperature, enzyme activity rapidly drops.

pH: The pH changes in the reaction medium affect the state of ionization of functional groups on both the enzyme and substrate molecules, thereby influencing the formation of ES complex. The optimal pH is the one at which the state of dissocation of the essential groups is most appropriate for interaction of enzyme and substrate to form ES complex. For most enzymes, optimum activity is between pH 6 and 8. Below or above these values, the reaction rate drops more or less rapidly because of the denaturation of enzyme followed by its inactivation. However, there are certain exceptions as well; Gastric Pepsin has acidic pH optimum (~1.5) while alkaline phosphatase exhibits optimal activity at pH of 9.5.

Use of enzyme inhibitors for the modulation of their activity

Enzyme inhibition is a well-known tool to block many important biochemical and physiological processes, resulting in the elucidation of new metabolic pathways as well as in in-depth knowledge of many kinetic mechanisms of enzyme-driven reactions. For this reason, the development of potent inhibitors is an important area of research in the pharmaceutical and agrochemical fields among others. The enzyme inhibitors are chemical agents that inhibit the catalytic activity of enzymes. The reduction n enzyme activity can be reversible or irreversible.

Irreversible inhibitors: These inhibitors produce permanent changes in the enzyme molecule causing definitive alteration of its catalytic capacity. Examples are organophosphates that produce irreversible inhibition of acetylcholinesterase-an important enzyme of insect nervous system.

Reversible inhibitors: There are three types of reversible inhibitors: competitive, non-competitive, and uncompetitive.

Most of the *competitive Inhibitors* have structural similarity with the substrate and bind to the free enzyme only. They increase the value of the Km but do not modify the V_{max} of the enzyme. This type of inhibition can be reversed by increasing the substrate concentration. At higher substrate concentration, the inhibitor bound to the enzyme gets displaced. *Noncompetitive inhibitors* bind to free enzyme as well as ES complex. They bind to enzyme at a site different from the active site and they decrease the V_{max} without altering the K_m. Also, the binding of the substrate to the enzyme is not affected by the presence of the inhibitor and vice versa, hence there is formation of a ternary complex between enzyme-substrate and Inhibitor (ESI), which is a dead end complex. Inhibition of enzyme containing sulfhydryl groups by metal ions is a very good example of non-competitive inhibition. *Uncompetitive inhibitors* bind to the ES complex only and form an inactive ESI complex. Such inhibitors decrease both the K_m and V_{max}

Concluding Remarks

The enzyme kinetics is a great tool for understanding the roles and regulation of various enzymes involved in catalyzing various reactions in different metabolic pathway, which in turn can help us to devise strategies for enhancing the nutritional quality of food crops and/or augmenting their tolerance against biotic and abiotic stresses for improved crop yields.

References

Bowden A.C. (2012). *Fundamentals of Enzyme Kinetics*. Fourth Edition. Wiley Blackwell

Dixon M., Webb E.C., Thorme C.J.R. and Tipton K.F. (1979). *Enzymes.* 3rd Ed. Longman.

Maragoni A.G. (2003). *Enzyme Kinetics- A Modern Approach*. John Wiley.

Palmer T. 2001. E*nzymes: Biochemistry, Biotechnology and Clinical Chemistry, 5th* Ed. Harwood Publ.

Price N.C. & Stevens L. 2003. *Fundamentals of Enzymology.* Oxford University Press.

Wilson K. & Walker J. (Eds.) 2000. *Principles & Techniques of Practical Biochemistry*. 5th Ed. Cambridge Univ. Press.

Bergmeyer HU. 1983. *Methods of Enzymatic Analysis.* Verlag Chemie, Academic Press.

Chapter-19

DNA Barcoding and Its Application in Identification of Species

T. Venkatesan
ICAR-National Bureau of Agricultural Insect Resources, Bengaluru-560024, Karnataka
tvenkat12@gmail.com

Insects are the most abundant of all life on earth. The estimated world totals of described, living species in the 29 orders of the class Insecta amount to 1,004,898 (Adler and Foottit, 2009). The figure of 4 million has been accepted as being the most commonly cited figure based on recent publications for the total number of species. India, with 2% of global space, is among the top ten mega diversity nations in the world in terms of insect diversity, with about 7.10% of the world insect fauna. Ghorpade (2010) provided an estimate of 54,346 described species of insects in 27 orders from India, with nearly as many species yet to be discovered. Some other estimates by the Zoological Survey of India and other sources are much higher, ranging from 62000 to 80000 described species. According to Mayr & Ashlock (1991), it took nearly 200 years for taxonomists to describe 1.7 million species which is only 10 per cent of the total number of species estimated. In this context identification of insects has been a monumental task where it calls for availability of more number of specialists and funding. But with the dwindling interest in taxonomy and fund availability, classification and identification of various life forms particularly insects has been major challenge to the scientific community. With the advent of molecular biology and molecular tools identification of life forms including insects has become quick, precise and easy. Development of species-specific markers enables even a non-specialist to identify insects to species level.

Difficulties in morphological identification

Further, we need about 15000 taxonomists working for centuries to complete the task of classifying the remaining 90 per cent of the unidentified organisms. Economic development and increased international commerce are leading to higher extinction rates and introduction of invasive pest species. Therefore there is a need for faster species identification and information about their biodiversity for conserving them before they vanish from the face of Earth. Contributions of morphological taxonomy is enormous but has also some draw backs like

The task of routine species identification has several limitations including incorrect identification due to both phenotypic plasticity and genetic variability in the characters employed for species recognition and presence of morphologically cryptic taxa in some plant groups. Further, unambiguous identification of species requires the availability of complete plants at reproductive stages, which are generally not available throughout the year. Moreover, the use of morphological taxonomic keys often demands a high level of expertise that misdiagnoses are common. Thus the limitations in morphology-based identification systems

and the dwindling pool of taxonomists urgently require a new robust approach for taxon recognition. Hence there is a need for an adjunct tool that facilitates rapid identification of species where molecular identification popularly called "DNA barcoding" becomes handy.

Advantages of DNA barcoding

A DNA barcode is a short sequence from standardized portions of the genome (a 648 bp of mtCOI). DNA barcoding is technically a simple and rapid approach, in which a small DNA fragment is amplified by PCR from total genomic DNA and PCR product is directly sequenced. The species identification is done by comparing the query sequence with the reference database of DNA barcode library. The current work worldwide is targeted to generate such reference barcode library.

DNA barcoding can be a tremendously useful and exciting new tool in our arsenal of species identification methods, provided taxonomists are supported to discover and describe species that are subsequently identified by the use of DNA barcodes (Wheeler, 2008). Insect biodiversity is a valuable and vulnerable genetic resource. Abrupt changes in climate may endanger the survival of vulnerable species and trigger the loss of unique gene pool in the populations. Reliable identification of species diversity and species inter-relationships within a genus is of paramount importance for bioprospecting of species for alien-gene transfer and mining of genes for medicinally active biomolecules. Microgenomic identification systems, which permit species discrimination through the analysis of a small segment of genome, represent one extremely promising approach for the genetic identification of biological diversity at species level. DNA barcoding is extremely useful for unambiguous identification of biological specimens and more efficiently managing species diversity in Gene Banks. DNA barcoding is being done for several organisms including insects and other arthropods under various initiatives such as the Barcode of Life.

The main advantage of DNA barcoding is the rapid acquisition of molecular data. Mitochondria are energy-producing organelles, found in nearly every cell in nearly every plant and animal species. The mitochondrial genome in particular has turned out to be exceedingly useful in tracing evolutionary history, as it is present in all eukaryotic organisms, evolves rapidly as compared to nuclear DNA, and does not undergo meiosis and recombination, processes that scramble the evolutionary lineages of nuclear genes. The generation of molecular data from the CO1 region was based on accepted DNA bar-coding principles .i.e. barcoding protocols developed by the Barcoding of Life (iBoL) Initiative. A first International conference in this direction was organized at Munich in the year 2002 by German Science Association (DFG). In this regard, Paul Hebert of University of Guelph, Canada developed the use of one mitochondrial gene, mitochondrial cytochrome oxidase I (mtCOI) as Universal identification marker for identification of animal species which include insects.

The following are the set of Universal primers for the identification of species by Folmer *et al.* (1994).

Universal Primers

LCO 1490 5'-GGT CAA ATC ATA AAG ATA TTG G-3'
LCO 2198 5'-TAA ACT TCA GGG TGA CCA AAA AAT CA-3'

For identifying species a sequence length of 617 nucleotides is recommended and 669 nucleotides for phylum analysis. The sequence generated is aligned and checked similarity with the help of NCBI nucleotide database. The barcode is then generated by uploading the sequence with NCBI accession number.

While morphological data are usually time consuming and needs specialists. DNA barcoding techniques are uniform, practical method of species identification of insects and can be used for the identification of all developmental stages of insects, their food webs, biotypes and this may not be possible with morphology based taxonomy. Fragments or damaged specimens identity can be determined using DNA barcode (Pons, 2006). The COI gene has proved to be suitable for species identification in a large range of animal taxa, including butterflies and moths (Hebert *et al.* 2004a; Janzen *et al.*, 2005; Burns *et al.*, 2008); mayflies (Ball *et al.*, 2005), spiders (Greenhouse *et al.*, 2005), mosquitoes (Kumar *et al.*, 2007) and wasps (Smith *et al.*, 2008). In USA, 25,000 DNA barcodes have been generated for insects belonging to Hymenoptera, Lepidoptera, Hemiptera, Diptera and Trichoptera. In Canada, 30,000 DNA barcodes have been developed for various groups of insects in Lepidoptera, Hymenoptera, etc. There are DNA barcodes available for mosquitos, honeybees, fruit flies, ants (www.boldsystems.org).

DNA-based species identification will speed up analysis of known species and reveal cryptic species within species by population genetic analysis. DNA bar-coding can play an important role in studying the arrival of invasive species. DNA bar-coding can pinpoint the geographic source of an invading species and measure the distances over which pest species can travel. DNA barcoding can be advantageous for monitoring illegal trade in animal byproducts. DNA barcoding may lead to discover new species by sampling biodiversity hotspots, unexplored regions. The COI gene has proved to be suitable for species identification in a large range of animal taxa, including butterflies and moths (Hebert *et al.* 2004a, Janzen *et al.* 2005, Hajibabaei *et al* 2006b, Burns *et al.* 2008); mayflies (Ball *et al* 2005), spiders (Greenhouse *et al.*, 2005), mosquitoes (Kumar *et al.*, 2007) and wasps (Smith *et al.*, 2008). Development of automatable DNA chip-based approaches & protocols will be very useful to identify and quantify species.

Overview of DNA barcoding

- 2003-Hebert *et al.*, proposed the technique of using COI gene (648 bp).
- 2004-Barcode of life project-initiated by CBOL (Canadian Barcode of Life) to promote DNA barcode as a global std
- 2010- IBOL (International Barcode of Life)- 26 countries aims for automated identification system based on DNA barcode library of all eukaryotes (5 million

specimens of 500 000 sps.). Also aims for new protocols, informatics, equipment, extn methods etc).

- CBOL & IBOL started campaigns fish (Fish-BOL); birds, mammals, marine life and insects.
- Further, Europe (ECBOL), Norway (NorBOL), Mexico (MexBOL) and Japan (JBOLI) started projects as part of IBOL

Work done at NBAIR

Protocols for DNA barcoding of insects were standardized and characterization of cytochrome oxidase-I gene (COI) was done for parasitoids, predators and invasive insect pest and DNA-barcode has been generated for the same (Table 1). Nearly 700 DNA barcodes of parasitoids, predators and invasive pests been generated. A barcoding of species of *Trichogramma* will form a very important molecular aid for resolving the taxonomic problem in the identification of these important egg parasitoids. Currently several biocontrol lab in India both public and private insectaries are invariably mass producing and field releasing the *Trichogramma* of more than one species and very often there is a problem of mislabeling and identification. Further DNA barcoding technique was used to identify and to see the variation in *Glyptapanteles* sp., *Microplitis* sp.,m *Bemisia tabaci* and *Trichogramma* spp. The DNA barcoding of *Brotispa longissima*, an alien invasive pest is a great significant for the rapid identification of the pest as soon as the invasive pest enters in India. Very recently, the DNA barcoding helped to confer the identification of the population of *Acerophagus papayae* an important parasitoid of invasive pest against papaya mealybug *Paracoccus marginatus*, imported from Puerto Rico and fortuitously introduced and observed in Pune.

Other Applications

- Identification of extinct species
- EPA-Identify insects in rivers & streams as critical indicators.
- Quarantine of fruit flies.
- Illegal trade of endangered or potected insects
- Monitoring disease vectors-Mosquitos
- FDA and CBOL- working on DNA barcodes for economically important fishes and can detect economic fraud especially for trading tuna fish and bush meat.
- Partial (780 bp) mitochondrial cytochrome c oxidae subunit and near complete nuclear 18 S rDNA (1780 bp) sequences were directly compared to assess their relative usefulness as markers for species identification and phylogenic analysis of coccidian parasites (phylum Apicomplexa) i.e. Eimeria spp. Infecting chickens.
- Used for identification of forensically important blowflies.
- Can be used to identify earhworms
- Can be used to identify bird species

Development of Linkages

International Scenario in DNA Barcode

The International Nucleotide Sequence Database Collaborative is a partnership among GenBank in the U.S., the Nucleotide Sequence Database of the European Molecular Biology Lab in Germany, and the DNA Data Bank of Japan. They have agreed to CBOL's data standards for barcode records. Barcode of Life Database (BOLD) was created and is maintained by University of Guelph in Ontario. It offers researchers a way to collect, manage, and analyze DNA barcode data. The Data Analysis: Specimens are identified by finding the closest matching reference record in the database. CBOL has convened a Data Analysis Working Group to improve the ways that DNA barcode data can be analyzed, displayed, and used.

An automated DNA-based system will free taxonomists from routine identifications, allowing them to direct their efforts to new collections, descriptions and assessments of taxonomic relationships. In 2003, Paul D.N. Hebert from the University of Guelph, Ontario, Canada, proposed the compilation of a public library of DNA barcodes that would be linked to named specimens. This library would "provide a new master key for identifying species, whose power will rise with increased taxon coverage and with faster, cheaper sequencing". The goal of a DNA barcoding library is the construction of an enormous, online, freely available sequence database. Participants in the DNA barcode initiative come in many configurations, including consortia, databases, networks, labs, and projects that range in size from local to global.

The largest consortia are

- The International Barcode of Life (iBOL) Project is a Canadian-led research alliance which spans 26 countries and brings together hundreds of leading scientists in the task of collecting specimens, obtaining their DNA barcode records and building an informatics platform to store and share this information for use in species identification and discovery. By 2015, iBOL participants will gather DNA barcode records for five million specimens representing 500,000 species, delivering a highly effective identification system for species commonly encountered by humanity and laying the foundation for subsequent progress. iBOL's principal funding partners within Canada are the Canada Foundation for Innovation, the Ontario Ministry of Research and Innovation, the Government of Canada through Genome Canada in collaboration with the Ontario Genomics Institute, the Natural Sciences and Engineering Research Council of Canada, and the International Development Research Centre.
- CBOL, the Consortium for the Barcode of Life, promotes barcoding through conferences, outreach activities, working groups and workshops, but does not generate any barcode data. CBOL is the designated lead organization for iBOL's Working Group for Outreach and Collaborations. CBOL is based at the Smithsonian Institution's National Museum of Natural History in Washington, DC.

- ECBOL, the European Consortium for the Barcode of Life, was established as part of the research infrastructure efforts of EDIT, the European Distributed Institute of Taxonomy.

Databases

There are two central DNA barcode databases. BOLD, the Barcode of Life Data Systems at the University of Guelph, is a public workbench for barcoding projects. Researchers can assemble, test, and analyze their data records in BOLD before uploading them to GenBank, EMBL, and DDBJ which comprise the International Nucleotide Sequence Database Collaboration. They are the permanent public repositories for barcode data records.

DNA Barcoding Software

The Barcode of Life Data Systems (**BOLD)** is freely available to any researcher online with registration. The results are displayed in tables showing the most closely related species and related taxa. The BOLD is an online workbench that aids collection, management, analysis, and use of DNA barcodes. It is an official informatics for the

Barcode of life project (Ratansingham and Hebert, 2007), developed by the Canadian Centre for DNA Barcoding (CCDB). BOLD is open to the public. It consists of three components (MAS, IDS, and ECS) that address the needs of various groups in the barcoding community.

- BOLD-MAS (Management and analysis) provides a repository for barcode records coupled with analytical tools. It serves as an online workbench for the DNA barcode community.
- BOLD-IDS (identification engine) provides a species identification tool that accepts DNA sequences from the barcode region and returns a taxonomic assignment to species level when possible.
- BOLD-ECS (external connectivity) provides web developers and bioinformaticians the ability to build tools and workflows that can be integrated with the BOLD framework. BOLD-ECS supplies REST services that allow access to public sequence and specimen data.

National Scenario in DNA Barcode

So far 106236 barcodes have been generated (Pers. Comm. from BOLD database) in the world and 2600 barcodes have been generated in India.

- DNA Barcoding of Butterflies and skippers from Western Ghats, India & True Bug species University of Pune & Nation Centre for Cell Sciences, Pune, funded by DBT.
- DNA Barcoding of Mosquitoes of India: Vector Control Research Centre, Indian council of Medical Research, Puducherry, India.

- Identification of Satyrine butterflies of Peninsular India through DNA Barcodes (IISc & KFRI) funded by DBT.
- DNA Barcoding of sucking pests: IIHR, Bangalore

Integrated approach

Some taxonomists are concerned that DNA barcoding will compete with traditional taxonomic studies. However, it is emphasized that DNA barcoding is inseparably linked to taxonomy. The integration of various types of data such as morphological, ecological, physiological and molecular data including DNA barcodes will improve species discovery and description processes. Recently, multilocus DNA-barcoding approach (ITS region and 18S RNA) are progressively emerging and is now commonly accepted (particularly in cases where COI is not species specific). Hence it was proposed the combination of morphological and molecular characters, which has the advantage of bridging the gap between the classical taxonomy and molecular taxonomy and the DNA barcoding approach. This integrated approach based on the use of several different markers to carry out combined releases has been suggested to deal with taxonomic problems, such as species boundaries (Dayarat 2005, Will *et al* 2005, Roe & Sperling 2007).

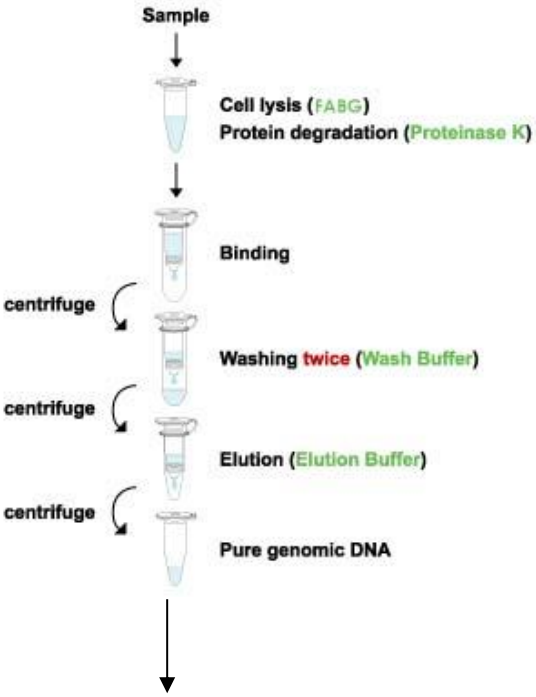
Limitations of the Barcode: Gap areas

- Maternal inheritance: mtDNA genes are maternally inherited which sometimes may result in interspecific hybridization or endosymbiont infections that generate transfer of mitochondrial genes outside the species (Hurst & Jiggins 2005, Dasmahapatra & Mallet 2006); the occurrence of indirect selection on mitochondrial DNA arising from male-killing microorganisms and cytoplasmic incompatibility incompatibility inducing symbionts (eg. Wolbachia) (Johnson & Hurst, 1996; Funk *et al.,* 2000; Whitworth *et al.,* 2007).
- Possible presence of nuclear copies of COI-in the nuclear genome (Nuclear mitochondrial DNA's-NUMT's) (Williams & Knowlton, 2001).
- Different rates of genomic evolution of COI genes are also limitations, since they are not equal for all the organisms.
- The COI based identification sometimes fails to distinguish closely related animal species, underlining the requirement of the other mitochondrial or nuclear regions (Sevilla *et al.*, 2007).
- Lack of monophyly of several groups studied. About 23% of animal species are polyphletic if their mtDNA are accurate indicating that using a mtDNA to assign a species name to an animal will be ambiguous or erroneous in 23% of the time.
- For groups of species that diverged in the recent past, marker alone will not be enough to clearly determine their taxonomic position or resolve specific limits, no matter the species concept used.

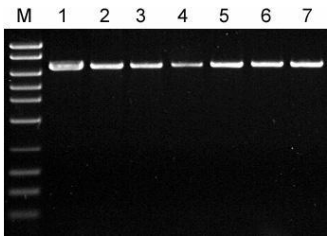
- DNA barcoding raises analytical and statistical issues. Only few studies have compared algorithms for species assignments and comparisons between the approaches are needed.
- Identification constraints in BOLD commonly arise when the unknown specimens come from the currently under-described part of biodiversity

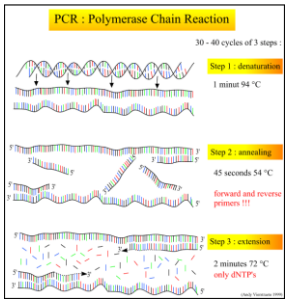
DNA Barcode Protocol

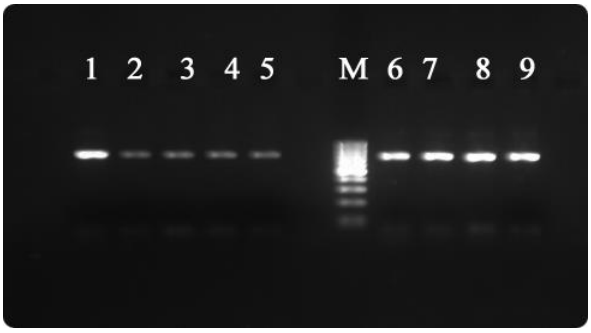
Extraction



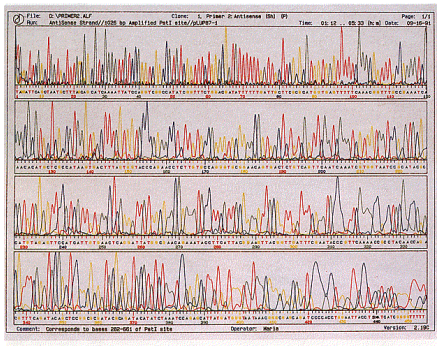
Agarose Gel Electrophoresis







DNA Sequencing



TTTGGATTTTGATCAGGCATAGTAGGAACATCACTCAGGGTCTTAATTCGAACTGAACTT
GGAAACCCTGGATCA
TTAATTGGAAATGATCAAATTTATAATGTCATTGTAACTGCCCACGCATTCATTATAATTT
TCTTCATAGTTATG
CCAACCTTAATTGGAGGATTCGGAAACTGATTAGTGCCCCTGATAATTGGAGCACCTGAT
ATAGCCTTCCCCCGG
CTAAATAATATAAGATTCTGACTTCTTCCCCCTTCTTTATCCTTACTTCTCCTTAGAAGAA
TGAGAGAAAGAGGA
GCTGGAACCGGTTGAACAGTATATCCCCCACTCTCAGCCAATATTGCCCATAGAGGAGCT
TCAGTAGACCTCGCA
ATTTTTAGGCTACACTTGGCAGGAATTTCATCAATTTTAGGGGCCGTAAATTTCATTTCAA
CTATTATAAATATA
CGACCGTCTGGTATAGACCTAGATAAAACCCCGCTATTCCCCTGAGCAGTAATCATTACT
GCTATTCTTCTTTTA
TTATCACTTCCTGTTTTAGCCGGAGCTATTACTATACTTCTAACAGATCGAAACATTAACA
CTTCATTTTTTGAC
CCAGCAGGAGGAGGTGATCCTATTTTATACCAACATCTATTCTGATTTTTTGGTCACCCTT
GAAGTTTAA

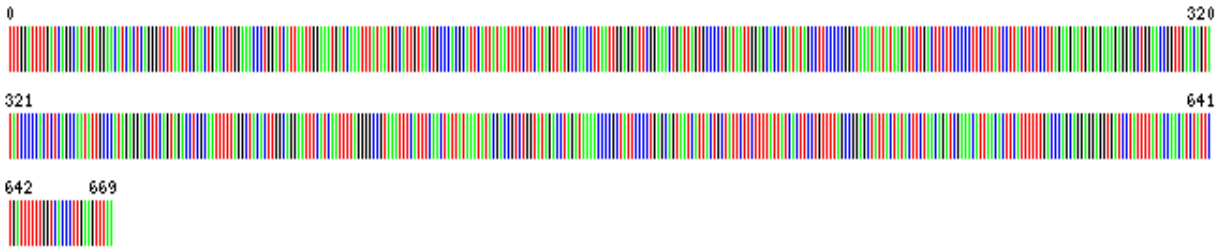
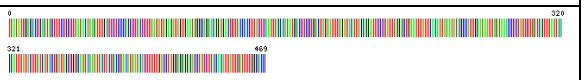
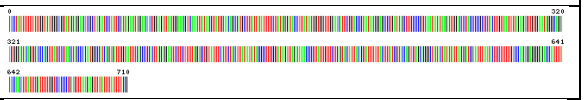
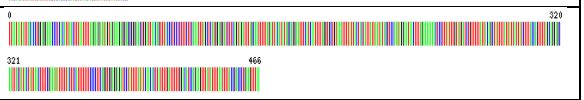
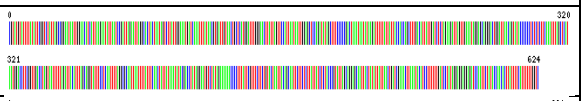
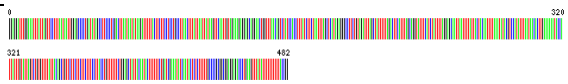
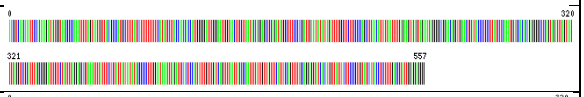
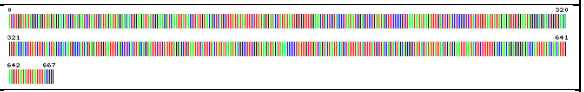
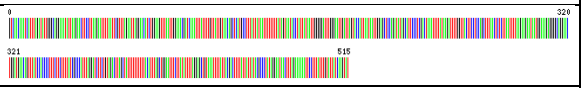
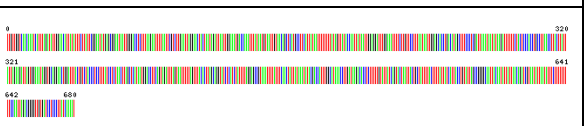
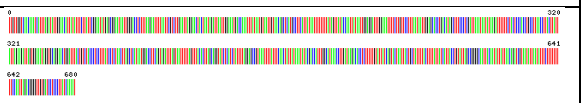
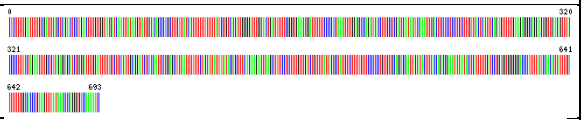
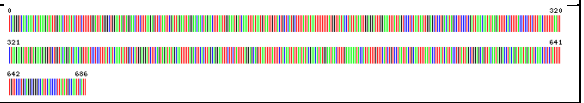
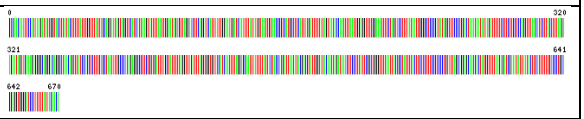
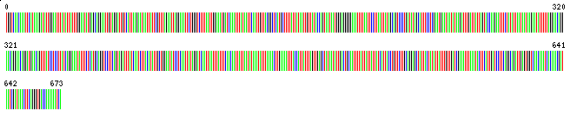
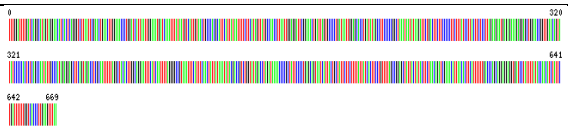
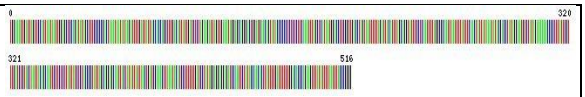
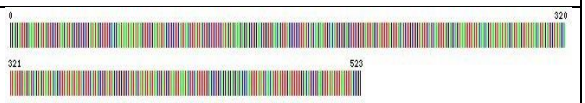
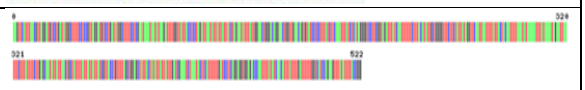
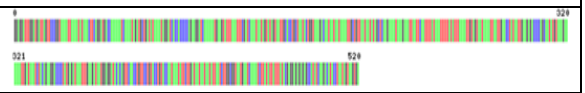
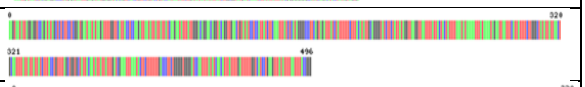
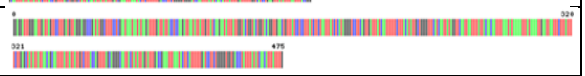
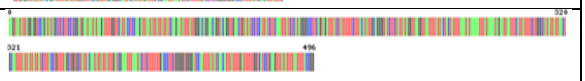
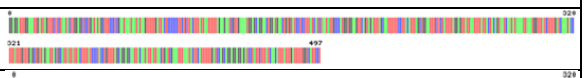
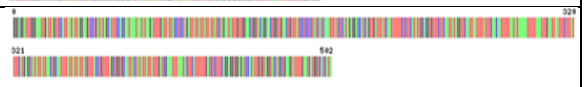
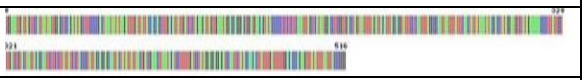
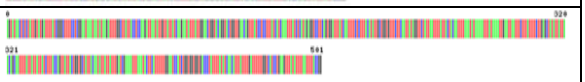
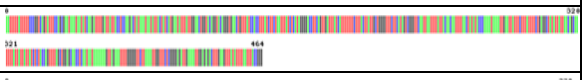
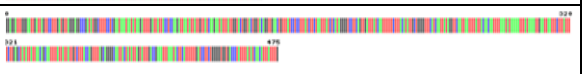
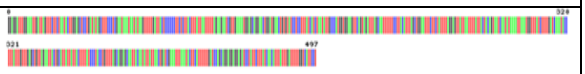
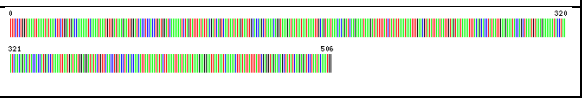


Table 1: Some DNA barcodes of insects generated at NBAII, Bangalore

Sl. No.	Name of the species	Barcode ID	Genbank Acc.no.	DNA barcode
1	*Coccinella transversalis*	AINCC00910	HQ658149	
2	*Harmonia axyridis*	AINCC008-10	HQ658148	
3	*Chilocorus nigrita*	AINCC006-10	HQ270157	
4	*Cheilomenes sexmaculata*	AINCC005-10	HQ270156	
5	*Cryptolaemus montrouzieri*	AINCC003-10	HQ270154	
6	*Illeis cincta*	AINCC002-10	HQ270153	
7	*Brumoides suturalis*	AINCC001-10	HQ694829	
8	*Curinus coeruleus*	AINCC011-11	JF323039	
9	*Coccinella septempunctata*	AINCC012-11	*JF323040*	
10	*Hyperaspis maindroni*	AINCC013-11	*JF323041*	
11	*Rodolia amabilis*	AINCC014-11	*JF323042*	
12	*Scymnus latemaculatus*	AINCC015-11	*JF323043*	
13.	*Henosepilachna vigintioctopunctata*	AINCC010-11	*HQ270155*	
14	*Acerophagus pappayae*	ACERO001-10	*HQ231257*	
15	*Brontispa longissima*	BRLON001-10	*HM446251*	

16	*Cheilomenes sexmaculata*	NBAII001-10	*FJ154102.*	
17	*Chrysoperla zastrowi sillemi*	CZASG001-10	*GU817334.*	
18	*Trichogramma japonicum*	TRINB007-10	*GU975843*	
19	*T. pretiosum*	TRINB008-10	*GU975846*	
20	*T. semblidis*	TRINB009-10	*GU975847*	
21	*T. cacoeciae*	TRINB010-10	*GU975838*	
22	*Trichogrammato idea bactrae*	TRINBO11-10	*GU975840*	
23	*Tr. fulva*	TRINBO12-10	*GU975839*	
24	*Tr. robusta*	TRINBO13-10	*GU975837*	
25	*T. achaeae*	TRINAC	*GU975841*	
26	*T. brassicae*	TRIN B002-10	*GU975842*	
27	*T. chilotreae*	TRINB003-10	*GU975844*	
28	*T. chilonis*	TRINB003-10	*GU975845*	
29	*T. dendrolimi*	*TRINB005-10*	*GU975835*	
30	*T. flandersi*	*TRINB006-10*	*GU975843*	
31	*Cotesia flavipes*	*CFBLR001-10*	*GQ853456*	

References

Alcaide, M., Rico, C., Ruiz, S., Soriguer, R. and Mun˜oz, J. (2009). Disentangling Vector-Borne Transmission Networks: A Universal DNA Barcoding Method to Identify Vertebrate Hosts from Arthropod Bloodmeals. *PLoS ONE*, 4(9): e7092.doi:10.1371/journal.pone.0007092

Folmer, O., Black, M., Howh, W., Lutz, R. and Vrijenhoek, R. (1994). DNA primer for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, 3: 294–299.

Greenstone, M.H., Rowley, D.L., Heimbach, U., Lundgren, J.G., Pfannenstiel, R.S. and Rehner, S.A. (2005). Barcoding of generalist predators by polymerase chain reaction: Carabids and spiders. *Molecular Ecology*, 14: 3247–3266.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H. and Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly, *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences, USA*, 101: 14812–14817.

Kumar, N.P., Rajavel, A.R., Natarajan, R. and Jambulingam, P. (2007). DNA barcodes can distinguish species of Indian mosquitoes (Diptera: Culicidae). *Journal of Medical Entomology*, 44: 1–7.

Miller, J.S., Brower, A.V.Z. and Desalle, R. (1997). Phylogeny of the neotropical moth tribe Josini (Notodontidae: Dioptinae): Comparing and combining evidence from DNA sequences and morphology. *Biological Journal of the Linnean Society*, 60: 297–316.

Mitchell J. Eaton E Greta L. Meyers. (2010). Barcoding bushmeat: molecular identification of Central African and South American harvested vertebrates. *Conserv Genet*., 11: 1389–1404.

Ratnasingham, S. and Hebert, P.D.N. (2007). BOLD: The Barcode of Life Data System (www. Barcodinglife.org). *MolecularEcology Notes*, 7: 355–364.

Smith, M.A., Rodriguez, J., Whitefield, J., Deans, A., Janzen, D.H., Hallwachs, W. and Hebert, P.D.N. (2008). Extraordinary diversity of parasitoid wasps exposed by iterative integration of natural history, DNA barcoding, morphology and collections. *Proceedings of the National Academy of SciencesUSA*, 105: 12359–12364.

Utsugi, J., Toshihide, K. and Motomi, I. (2011). Current progress in DNA barcoding and future implications for entomology. *Entomological Science*, 14: 107-124.

Chapter-20

Computational Tools for Gene Annotation

A.R. Rao
Indian Agricultural Statistics Research Institute, New Delhi – 110 012

Introduction

Very efficient programs for searching a text for a combination of words are available on many computers. The same methods can be used for searching for patterns in biological sequences, but often they fail. This is because biological 'spelling' is much sloppier than English spelling: proteins with the same function from two different organisms are almost certainly spelled differently, that is, the two amino acid sequences differ. Similarly in DNA many interesting signals vary greatly even within the same genome. Hidden Markov Models (HMMs) are very well suited for many tasks in molecular biology, although they have been mostly developed for speech recognition since the early 1970s.

What is a Hidden Markov Model?

An HMM is similar to Markov chain, but is more general and flexible, and allows to model phenomena that can not be explained well with a regular Markov chain model. It is a discrete-time Markov model with some extra features. The main advantage is that when a state is visited by the Markov chain, the state "emits" a letter from a fixed-independent alphabet. Letters are emitted via a time independent, but usually state-dependent, probability distribution over the alphabet. When the HMM runs there is, first, a sequence of states visited, which are denoted by $q_1, q_2, q_3, \ldots,$ and second, a sequence of emitted symbols, denoted by $\phi_1, \phi_2, \phi_3, \ldots$. Generation of symbols can be visualized as a two step process as follows:

Initial → emission → transition → emission → transition → emission → …
q_1 ϕ_1 to q_2 ϕ_2 to q_3 ϕ_3

Denoting entire sequence of q_i's by $\boldsymbol{Q}$ and the entire sequence of ϕ_i's by $\boldsymbol{\phi}$, one can write "the observed sequence $\boldsymbol{\phi} = \phi_1, \phi_2, \phi_3, \ldots$" and "the state sequence $\boldsymbol{Q} = q_1, q_2, q_3, \ldots$". Quite often, the sequence $\boldsymbol{\phi}$ is known but not the sequence $\boldsymbol{Q}$. In such a case the sequence $\boldsymbol{Q}$ is called "hidden". An important feature of HMMs is that one can efficiently answer several questions about $\boldsymbol{\phi}$ and $\boldsymbol{Q}$.

An HMM consists of the following five components:

1. A set of N states $S_1, S_2, \ldots, S_N$.
2. An alphabet of M distinct observation symbols $A = \{a_1, a_2, \ldots, a_M\}$.
3. The transition probability matrix $P = (p_{ij})$, where

P_{ij} = Prob ($q_{t+1} = Sj \mid q_t = Si$).

4. The emission probabilities: For each state *Si* and a in *A*,
   b_i (*a*) = Prob (*Si* emits a symbol *a*)
   The probabilities b_i (*a*) form the elements in an $N \times M$ matrix $B = (b_i (a))$.
5. An initial distribution vector $\pi = (\pi_i)$, where π_i = Prob ($q_1 = S_i$).

The components 1 and 2 describe the structure of the model, and 3-5 describe the parameters. It is convenient to let $\lambda = (P, B, \pi)$. There are three main calculations that are frequently required in HMM theory. Given some observed output sequence $\boldsymbol{\phi} = \phi_1, \phi_2, \phi_3, \ldots$, these are:

(i) Given the parameters λ, efficiently calculate
Prob ($\boldsymbol{\phi} \mid \lambda$).
That is efficiently calculate the probability of some given sequence of observed outputs.

(ii) Efficiently calculate the hidden sequence $\boldsymbol{Q} = q_1, q_2, q_3, \ldots, q_T$ of states that is most likely to have occurred, given $\boldsymbol{\phi}$. That is, calculate
argmax Prob (Q | $\boldsymbol{\phi}$).
Q

(iii) Assuming a fixed topology of the model, find the parameters = (P, B, π) that maximize Prob ($\boldsymbol{\phi} \mid \lambda$).

Details on the above said algorithms (i) to (iii) are given in Ewens and Grant (2001).

Applications of HMMs

The most popular use of the HMM in molecular biology is as a 'probabilistic pro-file' of a protein family, which is called a profile HMM. From a family of proteins (or DNA) a profile HMM can be made for searching a database for other members of the family. Profile HMM treats gaps in a systematic way. HMMs are particularly well suited for problems with a simple 'grammatical structure,' such as gene finding. In gene finding several signals must be recognized and combined into a prediction of exons and introns, and the prediction must conform to various rules to make it a reasonable gene prediction. An HMM can combine recognition of the signals, and it can be made such that the predictions always follow the rules of a gene.

From regular expressions to HMMs

Regular expressions are used to characterize protein families, which is the basis for the PROSITE database (Bairoch *et al.*, 1997). Using regular expressions is a very elegant and efficient way to search for some protein families, but difficult for other. As already mentioned in the introduction, the difficulties arise because protein spelling is much more free than English spelling. Therefore the regular expressions sometimes need to be very broad and complex. Let us imagine a DNA motif like this:

```
A  C     A     -     -     -     A     T     G
T  C     A     A     C     T     A     T     C
A  C     A     C     -     -     A     G     C
A  G     A     -     -     -     A     T     C
A  C     C     G     -     -     A     T     C
-------------------------------------------------------------------------
T  G     C     T     -     -     A     G     G
-------------------------------------------------------------------------
A  C     A     C     -     -     A     T     C
-------------------------------------------------------------------------
```

A regular expression for this is
[AT] [CG] [AC] [ACGT]* A [TG] [GC],

meaning that the first position is A or T, the second C or G, and so forth. The term '[ACGT]*' means that any of the four letters can occur any number of times. The problem with the above regular expression is that it does not in any way distinguish between the highly implausible sequences

T G C T - - A G G

which has the exceptional character in each position, and the consensus sequence with the most plausible character in each position (the dashes are just for aligning these sequences with the previous ones).

A C A C - - A T C

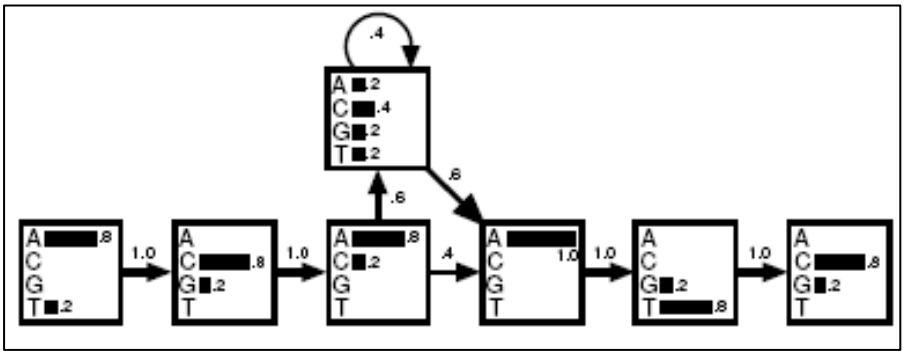


Figure 1: A hidden Markov model derived from the text above. The transition probabilities are shown with arrows. In each state the histogram shows the probabilities of the four nucleotides

It is possible to make the regular expression more discriminative by splitting it into several different ones, but it easily becomes messy. The alternative is to score sequences by how well they fit the alignment. To score a sequence, there is a probability of 4/5= 0. 8 for an A in the first position and 1/5=0.2 for a T, because out of 5 letters 4 are As and one is a T. Similarly in the second position the probability of C is 4/5 and of G 1/5 and so forth. After the third position

in the alignment, 3 out of 5 sequences have 'insertions' of varying lengths, so the probability of making an insertion is 3/5 and thus 2/5 for not making one. To keep track of these numbers a diagram can be drawn with probabilities as in Figure 1.

The above mentioned figure indicates a hidden Markov model. A box in the drawing is called a state, and there is a state for each term in the regular expression. All the probabilities are found by counting in the multiple alignments how many times each event occur, just as described above. The only part that might seem tricky is the 'insertion', which is represented by the state above the other states. The probability of each letter is found by counting all occurrences of the four nucleotides in this region of the alignment. The total counts are one A, two Cs, one G, and one T, yielding probabilities 1/5, 2/5, 1/5 and 1/5 respectively. After sequences 2, 3 and 5 have made one insertion each, there are two more insertions (from sequence 2) and the total number of transitions back to the main line of states is 3. Therefore there are 5 transitions in total from the insert state, and the probability of making a transition to itself is 2/5 and the probability of making one to the next state is 3/5.

It is now easy to score the consensus sequence ACACATC. The probability of the first A is 4/5. This is multiplied by the probability of the transition from the first state to the second, which is 1. Continuing this, the total probability of the consensus is

$$P(\text{ACACATC}) = .8 \times 1 \times .8 \times 1 \times .8 \times .6 \times .4 \times .6 \times 1 \times 1 \times .8 \times 1 \times .8 = 4.7 \times 10^{-2}$$

Making the same calculation for the exceptional sequence yields only $0\_0023 \times 10^{-2}$ which is roughly 2000 times smaller than for the consensus. This way one can achieve the goal of getting a score for each sequence; a measure of how well a sequence fits the motif.

Table 1: Probabilities and log-odds scores for the 5 sequences in the alignment and for the consensus sequence and the 'exceptional' sequence

	Sequence	**P x 100**	**Log odds**
Consensus	A C A C — — A T C	4.7	6.7
Original sequences	A C A — — — A T G	3.3	4.9
	T C A A C T A T C	0.0075	3.0
	A C A C — — A G C	1.2	5.3
	A G A — — — A T C	3.3	4.9
	A C C G — — A T C	0.59	4.6
Exceptional	T G C T — — A G G	0.0023	-0.97

Table 1 shows the calculated probabilities of other four original sequences in the alignment. The probability depends very strongly on the length of the sequence. Therefore the probability itself is not the most convenient number to use as a score, and the log-odds score shown in the last column of the table is usually better. It is the logarithm of the probability of the sequence divided by the probability according to a null model. The null model is one that treats the sequences as random strings of nucleotides, so the probability of a sequence of length L is 0.25^L. Then the log-odds score is

$$\text{Log-odds for sequence S} = \log \frac{P(S)}{0.25^L} = \log P(S) - L \log 0.25.$$

One can also use other null models instead. Often one would use the over-all nucleotide frequencies in the organism studied instead of just 0.25. For instance, the calculation of the log-odds of the consensus sequence is

Log-odds (ACACATC) = 1.16 + 0 + 1.16 + 0 + 1.16 – 0.51 + 0.47 – 0.51 + 1.39 + 0 + 1.16 + 0 + 1.16 = 6.64

If the alignment had no gaps or insertions we would get rid of the insert state, and then all the probabilities associated with the arrows (the transition probabilities) would be 1 and might as well be ignored completely. Then the HMM works exactly as a weight matrix of log-odds scores, which is commonly used.

Profile HMMs

A profile HMM is a certain type of HMM with a structure that in a natural way allows position dependent gap penalties. A profile HMM can be obtained from a multiple alignment and can be used for searching a database for other members of the family in the alignment very much like standard profiles (Gribskov *et al.*, 1987). The structure of the model is shown in Figure 2. The squares are called the match states, the diamonds the *insert* states, and the circles the delete states. The edges not shown have transition probability zero. State *Begin* is the *start* state, so that the process always starts in state *Begin*. A transition never moves to the left, so that as time progress the current state gradually moves to the right, eventually ending in match state *End*, the *end* state. When this state is reached the process ends. A match or delete state is never visited more than once. The bottom line of states is called the main states, because they model the columns of the alignment. In these states the probability distribution is just the frequency of the amino acids or nucleotides as in the above model of the

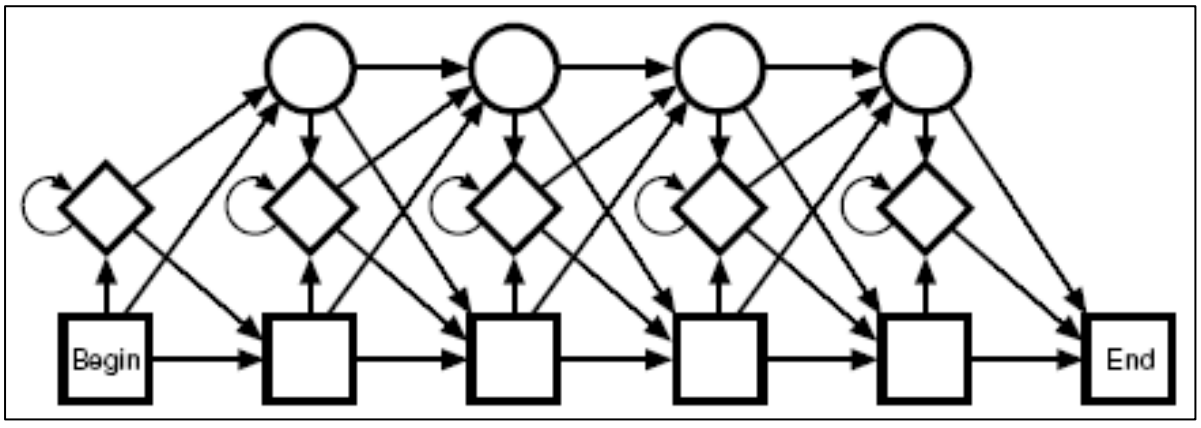


Figure 2: The structure of the profile HMM

DNA motif. The second row of diamond shaped states is called insert states and is used to model highly variable regions in the alignment. They function exactly like the top state in Figure 1, although one might choose to use a fixed distribution of residues, *e.g.* the overall distribution of amino acids, instead of calculating the distribution as in the example above. The top line of circular states is called delete states. These are a different type of state, called a silent or null state. They do not match any residues, and they are there merely to make it possible to jump over one or more columns in the alignment, *i.e.*, to model the situation when just a few of the sequences have a '–' in the multiple alignment at a position.

As an example consider a multiple alignment as shown in Figure 3. A region of this alignment (columns 9 and 10) has been chosen to be an 'insertion,' because an alignment of this region is highly uncertain. The rest of the alignment is the columns that will correspond to main states in the model. For each non-insert column we make a main state and set the probabilities equal to the amino acid frequencies.

G G W W R G d y . g g k k q L W F P S N Y V
I G W L N G y n e t t g e r G D F P G T Y V
P N W W E G q l . . n n r r G I F P S N Y V
D E W W Q A r r . . d e q i G I V P S K - -
G E W W K A q s . . t g q e G F I P F N F V
G D W W L A r s . . s g q t G Y I P S N Y V
G D W W D A e l . . k g r r G K V P S N Y L
G D W W E A r s l s s g h r G Y V P S N Y V
G D W W Y A r s l i t n s e G Y I P S T Y V
G E W W K A r s l a t r k e G Y I P S N Y V
G D W W L A r s l v t g r e G Y V P S N F V
G E W W K A k s l s s k r e G F I P S N Y V
G E W C E A q t . k n g q . G W V P S N Y I
S D W W R V v n l t t r q e G L I P L N F V
L P W W R A r d . k n g q e G Y I P S N Y I
R D W W E F r s k t v y t p G Y Y E S G Y V
E H W W K V k d . a l g n v G Y I P S N Y V
I H W W R V q d . r n g h e G Y V P S S Y L
K D W W K V e v . . n d r q G F V P A A Y V
V G W M P G l n e r t r q r G D F P G T Y V
P D W W E G e l . . n g q r G V F P A S Y V
E N W W N G e i . . g n r k G I F P A T Y V
E E W L E G e c . . k g k v G I F P K V F V
G G W W K G d y . g t r i q Q Y F P S N Y V
D G W W R G s y . . n g q v G W F P S N Y V
Q G W W R G e i . . y g r v G W F P A N Y V
G R W W K A r r . a n g e t G I I P S N Y V
G G W T Q G e l . k s g q k G W A P T N Y L
G D W W E A r s n . t g e n G Y I P S N Y V
N D W W T G r t . . n g k e G I F P A N Y V

Figure 3: An alignment of 30 short amino acid sequences. Shaded area represents most conserved region and is the main states in the HMM. The unshaded area represents insert states

To estimate the transition probabilities we count how many sequences use the various transitions, just like the transition probabilities were calculated in the DNA motif example. The model is shown in Figure 4. There are two transitions from a main state to a delete state shown with dashed lines in the figure, that from *begin* to the first delete state and from main state 12 to delete state 13. Both of these correspond to dashes in the alignment. In both cases only one sequence has gaps, so the probability of these delete transitions is 1/30. The fourth sequence continues deletion to the end, so the probability of going from delete 13 to 14 is 1 and from delete 14 to the end is also 1.

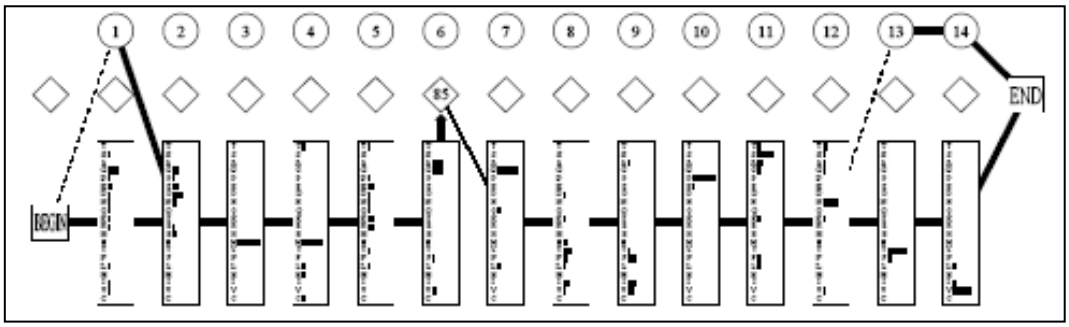


Figure 4: A profile HMM made from the alignment shown in Figure 3.

Searching a database

It is discussed earlier how to calculate the probability of a sequence in the alignment by multiplying all the probabilities (or adding the log-odds scores) in the model along the *path* followed by that particular sequence. However, this path is usually not known for other sequences which are not part of the original alignment, and the next problem is how to score such a sequence. Obviously, if one can find a path through the model where the new sequence fits well in some sense, then one can score the sequence as before. All it needs is to 'align' the sequence to the model. It resembles very much the pairwise alignment problem, where two sequences are aligned so that they are most similar, and indeed the same type of dynamic programming algorithm can be used.

For a particular sequence, an alignment to the model (or a path) is an assignment of states to each residue in the sequence. There are many such alignments for a given sequence. For instance an alignment might be as follows. Let us label the amino acids in a protein as A1, A2, A3, *etc.* Similarly we can label the HMM states as M1, M2, M3, *etc.* for match states, I1, I2, I3 for insert states, and so on. Then an alignment could have A1 match state M1, A2 and A3 match I1, A4 match M2, A5 match M6 (after passing through three delete states), and so on. For each such path we can calculate the probability of the sequence or the log-odds score, and thus we can find the *best* alignment, *i.e.*, the one with the largest probability. Although there are an enormous number of possible alignments it can be done efficiently by the above mentioned dynamic programming algorithm, which is called the Viterbi algorithm. The algorithm also gives the probability of the sequence for that alignment, and thus a score is

obtained. The log-odds score found in this manner can be used to search databases for members of the same family.

Model estimation

As presented so far, one may view the profile HMMs as a generalization of weight matrices to incorporate insertions and deletions in a natural way. There is however one interesting feature of HMMs, which has not been addressed yet. It is possible to estimate the model, i.e. determine all the probability parameters of it, from unaligned sequences. Furthermore, a multiple alignment of the sequences is produced in the process. Like many other multiple alignment methods this is done in an iterative manner. One starts out with a model with more or less random probabilities, or if a reasonable alignment of some of the sequences is available, a model is constructed from this alignment. Then, when all the sequences are aligned to the model, we can use the alignment to improve the probabilities in the model. These new probabilities may then lead to a slightly different alignment. If they do, we then repeat the process and improve the probabilities again. The process is repeated until the alignment does not change. The alignment of the sequences to the final model yields a multiple alignment. Although this estimation process sounds easy, there are many problems to consider to actually making it work well. One problem is choosing the appropriate model length, which determines the number of inserts in the final alignment. Another severe problem is that the iterative procedure can converge to suboptimal solutions. It is not guaranteed that it finds the optimal multiple alignments, i.e. the most probable one.

HMMs for gene finding

One ability of HMMs, which is not really utilized in profile HMMs, is the ability to model grammar. Many problems in biological sequence analysis have a grammatical structure, and eukaryotic gene structure is one such example. If one can consider exons and introns as the 'words' in a language, the sentences are of the form exon-intron-exon-intron...intron-exon. The 'sentences' can never end with an intron, at least if the genes are complete, and an exon can never follow an exon without an intron in between. Obviously this grammar is greatly simplified, because there are several other constraints on gene structure, such as the constraint that the exons have to fit together to give a valid coding region after splicing. Formal language theory applied to biological problems is not a new invention. In particular David Searls (1992) has promoted this idea and used it for gene finding [Dong and Searls, 1994], but many other gene finders use it implicitly. Formally the HMM can only represent the simplest of grammars, which is called a regular grammar, but that turns out to be good enough for the gene finding problem, and many other problems. Krogh (1998) outlined an approach to find genes with the weight on the principles rather than on the details.

Signal sensors

One may apply an HMM similar to the ones already described directly to many of the signals in a gene structure. In Figure 5 an alignment is shown of some sequences around acceptor sites

from human DNA. It has 19 columns and an HMM with 19 states (no insert or delete states) can be made directly from it. Since the alignment is gap-less, the HMM is equivalent to a weight matrix.

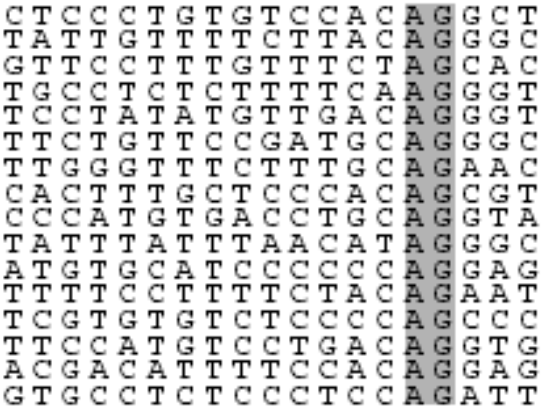


Figure 5: Examples of human acceptor sites (the splice site 5' to the exon). Except in rare cases, the intron ends with AG, which has been highlighted. Included in these sequences are 16 bases upstream of the splice site and 3 bases downstream into the exon.

There is one problem: in DNA there are fairly strong dinuclotide preferences. A model like the one described treats the nucleotides as independent, so dinucleotide preferences cannot be captured. This is easily fixed by having 16 probability parameters in each state instead of 4. In column two we first count all occurrences of the four nucleotides given that there is an A in the first column and normalize these four counts, so they become probabilities. This is the conditional probability that a certain nucleotide appears in position two, given that the previous one was A. The same is done for all the instances of C in column 1 and similarly for G and T. This gives a total of 16 probabilities to be used in state two of the HMM. Similarly it can be extended to all the other states. To calculate the probability of a sequence, say ACTGTC, we just multiply the conditional probabilities

$$P\,(ACTGTC\ldots) = p_1\,(A) \times p_2(C/A) \times p_3\,(T/C) \times p_4\,(G/T) \times p_5\,(T/G) \times p_6(C/T) \times \ldots$$

Here p1 is the probability of the four nucleotides in state 1, p2 (x/y) is the conditional probability in state 2 of nucleotide x given that the previous nucleotide was y, and so forth. A state with conditional probabilities is called a first order state, because it captures the first order correlations between neighboring nucleotides. It is easy to expand to higher order. A second order state has probabilities conditioned on the two previous nucleotides in the sequence, i.e., probabilities of the form p (x/y, z). Small HMMs like this are constructed in exactly the same way for other signals: donor splice sites, the regions around the start codons, and the regions around the stop codons.

Coding regions

The codon structure is the most important feature of coding regions. Bases in triplets can be modeled with three states as shown in Fig. 6. The figure also shows how this model of coding regions can be used in a simple model of an unspliced gene that starts with a start codon (ATG), then consists of some number of codons, and ends with a stop codon.

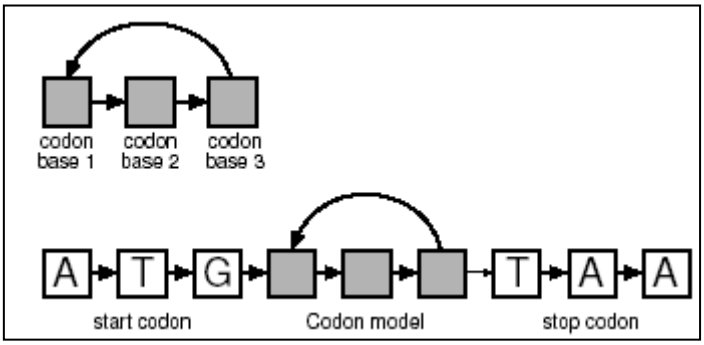


Figure 6: Top: A model of coding regions

Since a codon is three bases long, the last state of the codon model must be at least of order two to correctly capture the codon statistics. The 64 probabilities in such a state are estimated by counting the number of each codon in a set of known coding regions. These numbers are then normalized properly. For example the probabilities derived from the counts of CAA, CAC, CAG, and CAT are

P (A/CA) = c (CAA)/[c (CAA)+c (CAC)+c (CAG)+c (CAT)
P(C/CA) = c (CAC)/[c (CAA)+c (CAC)+c (CAG)+c (CAT)
P (G/CA) = c (CAG)/[c (CAA)+c (CAC)+c (CAG)+c (CAT)
P (T/CA) = c (CAT)/[c (CAA)+c (CAC)+c (CAG)+c (CAT)

where c (xyz) is the count of codon *xyz*. One of the characteristics of coding regions is the lack of stop codons. That is automatically taken care of, because p (A/TA), p (G/TA) and p (A/TG), corresponding to the three stop codons TAA, TAG and TGA, will automatically become zero. For modeling codon statistics it is natural to use an ordinary (zeroth order) state as the first state of the codon model and a first order state for the second. However, there are actually also dependencies between neighboring codons, and therefore one may want even higher order states.

Softwares and websites on HMMs

There are two program packages available free of charge to the academic community. One, developed by Sean Eddy, is called hmmer (pronounced 'hammer'), and can be obtained from his web-site (http://genome.wustl.edu/eddy/hmm.html). The other one, called SAM (http://www.cse.ucsc.edu/research/compbio/sam.html), was developed by Anders Krogh and the group at UC Santa Cruz, and it is now being maintained and further developed under the command of Richard Hughey. The gene finder sketched above is called HMMgene. The current version of HMM gene is available at the web site http://www.cbs.dtu.dk/services/HMMgene/.

The first HMM based gene finder is probably EcoParse developed for *E. coli* (Krogh *et al.*, 1994). VEIL (Henderson *et al.*, 1997) is a recent HMM based gene finder for human genes. The main difference from HMMgene is that it does not use high order states (neither does EcoParse), which makes good modeling of coding regions harder. Two recent methods use so-called generalized HMMs. Genie (Kulp *et al.*, 1996; Reese *et al.*, 1997; Kulp *et al.*, 1997) combines neural networks into an HMM-like model, whereas GENSCAN [Burge and Karlin (1997)] is more similar to HMMgene, but uses a different model type for splice site. Also, the generalized HMM can explicitly use exon length distributions, which is not possible in a standard HMM. Web pointers to gene finding can be found at http://www.cbs.dtu.dk/krogh/genefinding.html. Other applications of HMMs related to gene finding are: detection of short protein coding regions and analysis of translation initiation sites in Cyanobacterium (Yada and Hirosawa, 1996; Yada *et al.*, 1997), characterization of prokaryotic and eukaryotic promoters (Pedersen *et al.*, 1996), and recognition of branch points (Tolstrup *et al.*, 1997). Apart from the areas mentioned here, HMMs have been used for prediction of protein secondary structure (Asai *et al.*, 1993), modeling an oscillatory pattern in nucleosides (Baldi *et al.*, 1996), modeling site dependence of evolutionary rates (Felsenstein and Churchill, 1996), and for including evolutionary information in protein secondary structure prediction (Goldman *et al.*, 1996).

References

Asai, K., S. Hayamizu and K. Handa, 1993. *Computer Applications in the Biosciences*, 9:141–146.

Bairoch, A., P. Bucher and K. Hofmann, 1997. *Nucleic Acids Research, 25:217*–221.

Baldi, P., S. Brunak, Y. Chauvin, and A. Krogh, 1996. *Journal of Molecular Biology* 263:503–510.

Burge, C. and S. Karlin, 1997. *Journal of Molecular Biology, 268:78*–94.

Dong, S. and D. B. Searls, 1994 *Genomics, 23:540*–551.

Ewens, W.J. and G. R. Grant, 2001. Statistical methods in Bioinformatics – An Introduction. *Springer-Verlag New York, Inc.*

Felsenstein, J. and G. A. Churchill, 1996. *Molecular Biological Evolution* 13.

Goldman, N., J. L. Thorne and D. T. Jones, 1996. *Journal of Molecular Biology*, 263:196–208.

Gribskov, M., A. D. McLachlan and D. Eisenberg, 1987. *Proc. of the Nat. Acad. of Sciences of the U.S.A., 84:4355*–4358.

Henderson, J., S. Salzberg and K. H. Fasman, 1997. Finding genes in DNA with a hidden Markov model Journal of Computational Biology.

Krogh, A., 1998. An introduction to Hidden Markov Models for biological sequences. In *Computational methods in molecular biology*, edited by S.L. Salzberg, D.B. Searls and S. Kasif, pp 45-63. Elsevier.

Krogh, A., I. S. Mian and D. Haussler, 1994. *Nucleic Acids Research, 22:4768*–4778.

Kulp, D., D. Haussler, M. G. Reese and F. H. Eeckman, 1996. A generalized hidden Markov model for the recognition of human genes in DNA In States, D., Agarwal, P., Gaasterland, T., Hunter, L., and Smith, R. (Eds.), *Proc. Conf. on Intelligent Systems in Molecular Biology* pp. 134–142 Menlo Park, CA. AAAI Press.

Kulp, D., D. Haussler, M. G. Reese and F. H. Eeckman, 1997. Integrating database homology in a probabilistic gene structure model In Altman, R. B., Dunker, A. K., Hunter, L., and Klein, T. E. (Eds.), *Proceedings of the Pacific Symposium on Biocomputing* New York. World Scientific.

Pedersen, A. G., P. Baldi, S. Brunak and Y. Chauvin, 1996. Characterization of prokaryotic and eukaryotic promoters using hidden Markov models In *Proc. of Fourth Int. Conf. on Intelligent Systems for Molecular Biology* pp. 182–191 Menlo Park, CA. AAAI Press.

Reese, M. G., F. H. Eeckman, D. Kulp and D. Haussler, 1997. Improved splice site detection in Genie In Waterman, M. (Ed.), *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB)* New York. ACM Press.

Searls, D. B., 1992. *American Scientist, 80:579*–591.

Tolstrup, N., P. Rouz´e and S. Brunak, 1997. *Nucleic Acids Research, 25:3159*–3164.

Yada, T. and M. Hirosawa, 1996. *DNA Res., 3:355*–361.

Yada, T., T. Sazuka and M. Hirosawa, 1997. *DNA Res., 4:1*–7.

Chapter-21

Genome Sequencing: An Overview

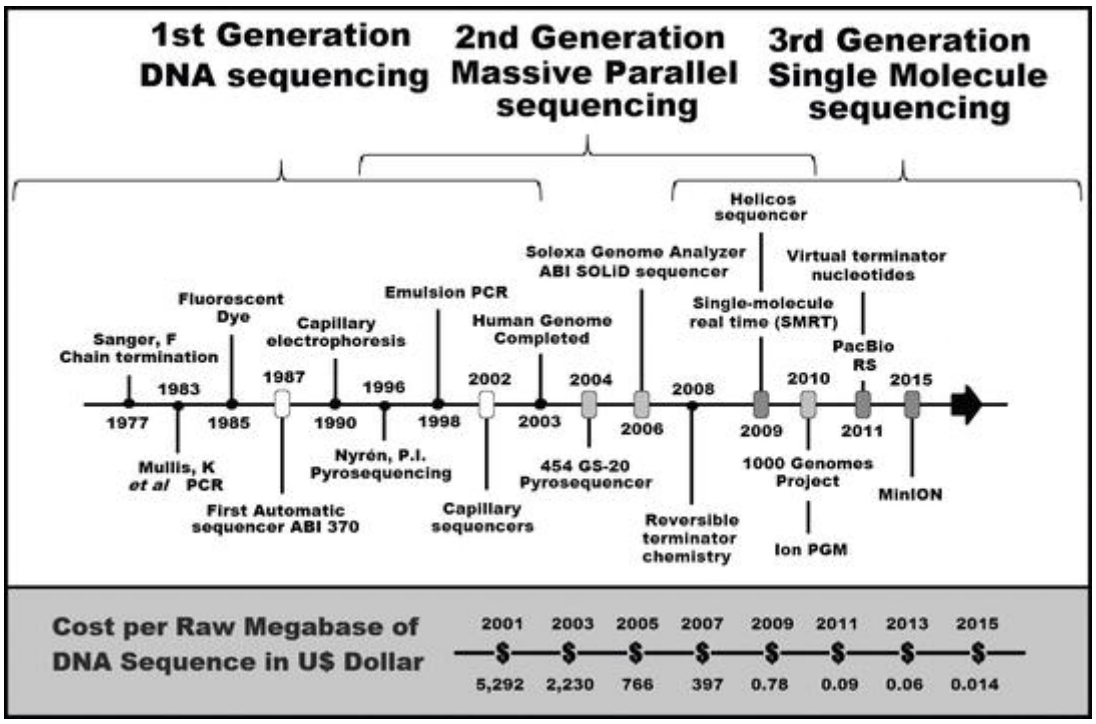
Kishor Gaikwad
National Institute for Plant Biotechnology, New Delhi-110012
kish2012@gmail.com

A genome is a collection of the entirecellular DNA in a manner that allows for proper packaging into chromosomes and specific expression allowing a cell to pass on its genetic information to the next generation. The genomes are present in the nucleus and mitochondria and additionally in plantschloroplast. 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 areas on the genome and for that matter the entire genome to understand its complexity and structure. This is the world of structural and functional genomics and biologists are decoding the alphabets of life with an intent of understanding its complexity and harnessing the benefits.

Genome sequencing was considered an uphill task and an expensive proposition in early days. In the 1960s and 70s it was understood that the whole genomes of big organisms were in proportion to their size. But these theories proved to be in contrast to the experimental evidences as it was shown that the genome of a big mammal was smaller than that of the lily plant. Even the prokaryotes seemed to have too much DNA in their genome than they could handle or package in the constraints of their single cell. It looked as if most of the DNA in the genome was redundant and was initially referred to as the junk DNA occupying space and energy of that organism. Then came the era of rapid advances in the understanding of the genetic code and how it worked in the pro and eukaryotes and one of the first sequence information was generated. In 1977 two methods for sequencing DNA were introduced. One method, referred to as Maxam-Gilbert sequencing, named after the two scientists at Harvard University who developed the technique, uses different chemicals to break radioactively labeled DNA at specific base positions. The other approach, developed by Frederick Sanger in England and called the chain termination method (also called the Sanger method), uses a DNA synthesis reaction with special forms of the four nucleotides that, when added to a DNA chain, stop (terminate) further chain growth. Thus a couple of hundreds reads could be completed in a day and was considered a major turnaround for gene sequencing. At this point, genome sequencing was still not thought of even as a distant possibility.

However all changed in the 90s with advent of PCR and high capacity cloning vectors. New vectors that could take insert DNA of sizes upto 1000kb (1Mb) were developed. These were the Bacterial artificial chromosomes (BACs) and the Yeast artificial chromosomes (YACs). Thus large insert libraries were developed and sequencing genomes became a reality. With advances in the sequence chemistry new non isotopic fluorescent dyes were developed and the era of automated capillary electrophoresis dawned. At the same time the computer industry

was going through its own revolution in the form of newer and faster than before machines and easy to use softwares. The synchronization or simple coincidence of these two technologies led to development of new machines. Thus it is possible to sequence the *E. coli* genome of around 4.5 Mb in less than the time required to watch a movie or even newer and faster machines that can finish human genome sequencing in flat 3 days otherwise which required around a decade. We have traveled a long way from the first sequence of *H. influenza* in 1995 to complete sequence of humans in 2001 and rice in 2002.



(DOI: 10.5772/intechopen.69337)

Figure 1: A historical snapshot of DNA sequencing platforms

The most important landmark in genome sequencing was in the latter part of the last decade when the next generation sequencers were launched by 454 and Solexa (Fig 1& 2). This revolutionized the way by which genome sequencing could be achieved. All the earlier projects including humans, Arabidopsis and rice were completed by a particular approach, known as the BAC by BAC approach. This took more time and was a high cost and labour intensive experiment. These approaches also required the availability of high density molecular maps which is available only in few plants. NGS technology changed all that and whole genome sequencing (WGS) became a routine feature where just few microgram of genomic DNA was enough to get a draft genome assembly. The improvements in NGS technology has been rapid and targeted more towards sequencing each base. One could now reach deep into the cell to sequence that one elusive transcript which normally would escape detection by any other methods. Thus the longer reads of a 454 machines and deeper coverage provided by Illumina / Solid/Ion Torrent systems merged together with the Sanger backbone, became a regular approach to sequence and assemble complex genomes.Suddenly the genebank became populated with assemblies of all types of eukaryotes and plants in particular.These included

draft genomes, organelle genomes, deep transcript profiles, small RNA profiles, ethylated C etc.Just when the biologists were grappling with the huge amount of genome and transcriptome data, along came the chemistry of single molecule sequencing. Notably, two systems known as Pacific Biosciences and Oxford nanopore were launched and now provide the luxury of longer reads ranging upto average 10 kb or more. With advent of newer chemistries like HiC, and 10X Genomics, it has now become relatively easier to obtain bigger read lengths. Combined with Optical mapping system like BioNano, developing a genome sequence to chr level assembly is relatively easier.

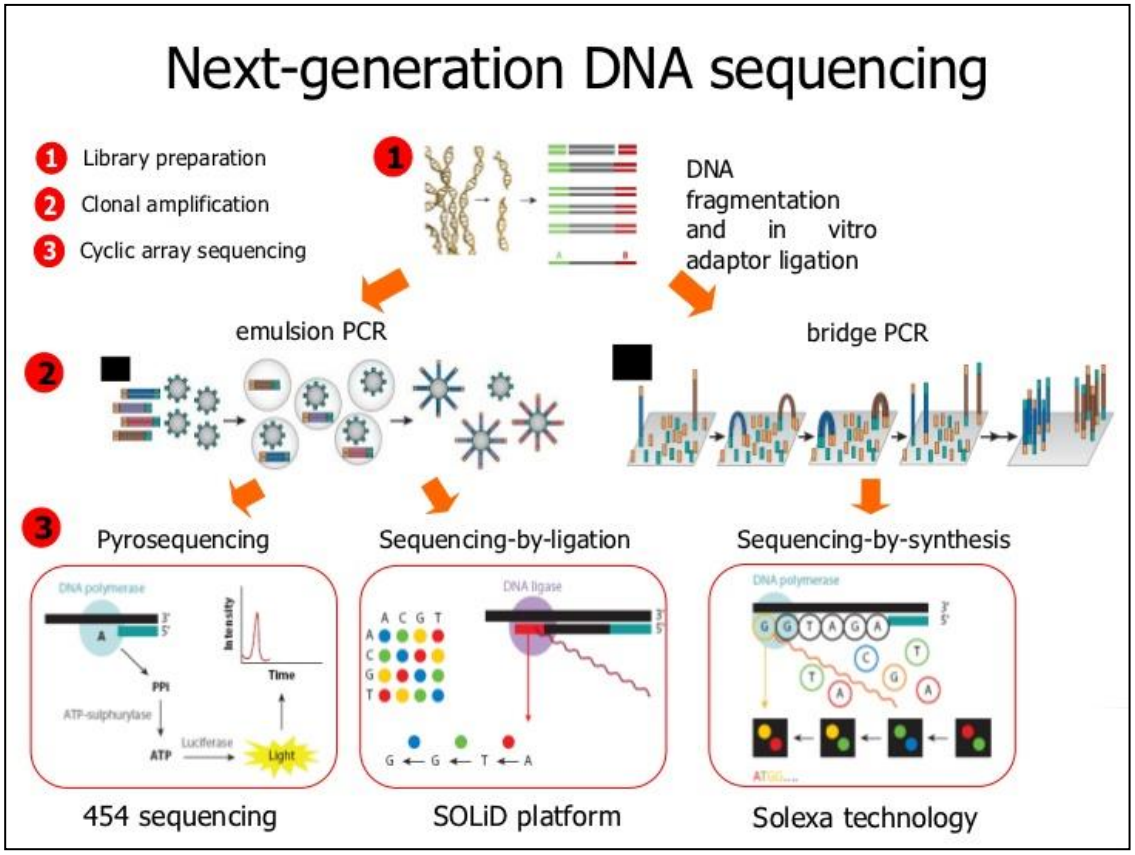


Figure 2: A overview of early Next Generation Sequencing technologies

Thus genome sequencing today has become easy atleast in a sense that all genomes can be decoded and analyzed completely and cost is no longer a hurdle. Over the decade the cost has come down to manageable proportions, from millions of dollars for a genome like humans and rice to few hundred for a complete genome today excluding the cost of the platform and data analysis. All the above technologies however come with some disadvantage. From a bioinformatics point of view, assembling each genome is challenging due to various factors. Assembling a genome sequenced by BAC by BAC approach and Sanger methods is much easier and accurate. But assembling small reads generated by Illumina and Ion Torrent systems is not that easy. Thus preparing a hybrid assembly of all different chemistries seems to be the only way out and mapping each base becomes an uphill task. Similarly annotation of such trancriptomes is not easy due to presence of spliced transcripts and MiRNAs. Thus the

challenges of data analysis now are a bigger worry than the actual sequencing itself. No single pipeline exists that can cater to all the complex eukaryotic genomes; each assembly will require probably newer and faster algorithms.

This requires constant interaction between biology and bioinformatics and development of species targeted interfaces and databases. Today ambitious projects like the 1000 human's genome project or the 3000 rice genome project have been completed. The 1000 genome human project is expected to map every SNP present in the genome and use the association for disease or particular traits. The 3000 rice genome project aims to capture all the diversity in the rice germplasm for utilization in trait improvement.

Thus resequencing of genome assumes greater significance as this effort will lead to identification of useful associations between genes and phenotype. As evident from the graph below, NGS has strongly contributed to growth of eukaryotic genomes in recent times and it keeps getting bigger and bigger every day (Figure 3).

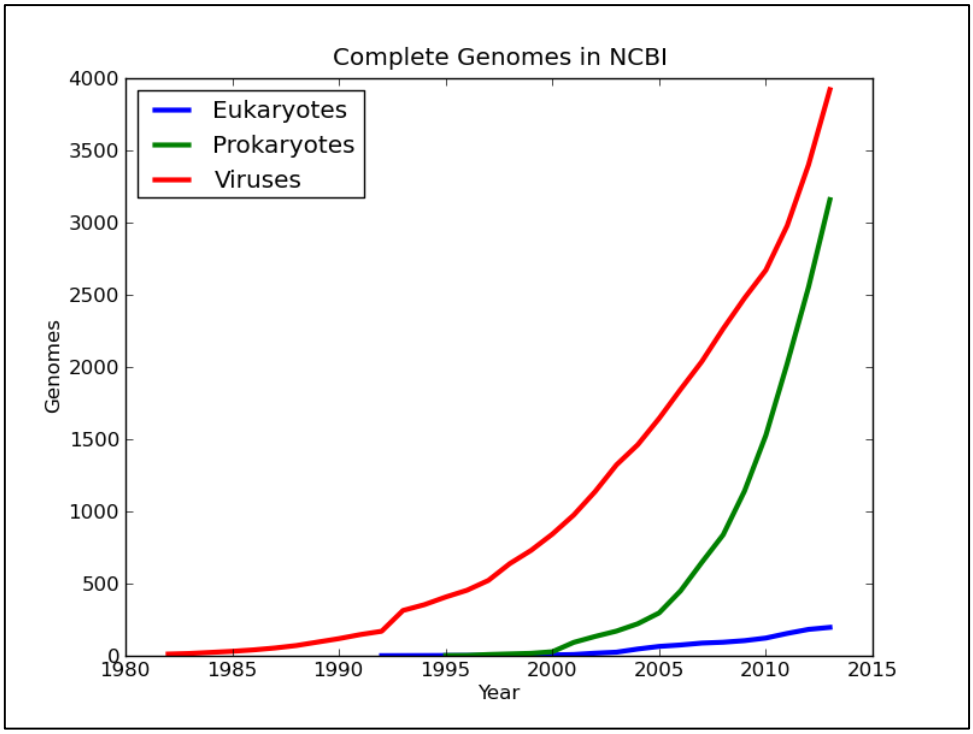


Figure 3: Growth in genome sequencing

One can only imagine the amount of data these studies will generate and help us understand the organisms and also harness the knowledge generated from such projects. Huge amount of biodiversity in important organisms provides us with immense potential for crop improvement that remained hidden for long. Genome sequencing of entire germplasm of plants, fungi, insects, nematoides, viruses have the potential to unravel new genes that could assist in providingfood and nutritional security to the masses.

References

Pop *et. al.* Trends Genet. 2008 :249:142-149

Hamilton and Buell,Plant J. 2012 Apr;70(1):177-90

Todd P *et.al.* Plant Genome, 2013: 6(2):1-7

Shendure *et.al Nature*, 550, pages 345–353 (2017)

Heather *et.al* Genomics (2016) 107:1-8

Levy and Myers, Annual Review of Genomics

Chapter-22

Metagenomics: An overview

K. Annapurna, Rajeev Kaushik and V. Govindasamy
Division of Microbiology, ICAR-Indian Agricultural Research Institute, New Delhi-110012
Email: annapurna96@yahoo.co.in

Microbes have important roles to play in various ecosystems, however, many remain to be characterized in detail due to unavailability of their complex nutritional requirements. Metagenomic tools are commonly used to investigate such complex microbial communities, sampled directly from the environment, without culturing or isolating a single organism. The term metagenomics was coined in 1998 (Handelsman *et al.*, 1998). Various metagenomic approaches help in understanding various aspects of total microbial community in a sample, and allows one to characterize them functionally and phylogenetically (Langille *et al.*, 2013; Vieites *et al.*, 2008). The amplification of specific targeted hypervariable regions (V1-V9) of 16S rRNA, 18S rRNA, ribosomal ITS, NifH, among others, by PCR before sequencing permit diversity analysis (Morgan and Huttenhower, 2012). Findings of metagenomic studies have revealed complex microbial interactions in environmental samples and it was observed that a) the relative abundance of the most abundant microorganisms of a particular group is not necessarily associated with the importance of that group in the functioning of the community; b) the most abundant organisms may not always play the most critical role in a community, while organisms constituting only 0.1% of the community (e.g., nitrogen fixers) can have very important functions (Dinsdale *et al.*, 2008).

Types of Metagenomics Studies

A) The most commonly used methods of microbial identification using high throughput metagenome sequencing data are

1) Whole metagenomic shotgun sequencing
2) Amplicon based method which includes 16S ribosomal RNA for bacteria, internal transcribed spacer (ITS) and 18S region for fungi and eukaryotes, respectively

Shotgun Metagenomics: Shotgun metagenomic analysis has the ability to identify the majority of the organisms (culturable and unculturable bacteria) in the environmental sample. A community biodiversity profile is created, which is further associated with functional composition analysis of organism lineages (Tringe *et al.*, 2005). Shotgun metagenomic studies can be divided into two types:

a) Sequence-based screens, which describe the microbial diversity and genomes of a particular environmental sample
b) Functional screens, which identify the functional gene products, but do not determine from what species the genetic material originated.

Before initiating a whole metagenomic study, an understanding of the potential microbial diversity and the relative abundance of species in the environmental sample is very important. For example, the metagenome of soil samples will consist of a more complex microbial community, than human skin. Hence, for proper coverage, more data must be generated in case of soil than for human skin. A higher sequencing depth also allows the detection of rare taxa. This makes shotgun metagenomic sequencing much more expensive than 16S sequencing, in order to achieve the coverage and depth needed for species identification.

Amplicon based sequencing: 16S sequencing is a widely used technique that relies on the variable regions (V1-V9) of the bacterial 16S rRNA gene to make community-wide taxonomic assignments. It is also used for microbial diversity analysis and has been used for various environmental samples, such as soil and gut microflora of animals and humans. Some degree of divergence is allowed during the sequence similarity assessment stage of the analysis; typically, nearly identical sequences (497%) are clustered into Operational Taxonomical Units (OTU). The limitation of this method is that, if any two organisms have the same 16S rRNA gene sequence, they may be classified as the same species in a 16S analysis, even if they are from different species. Because 16S analysis is based on the 16S rRNA gene, with OTUs defined as taxa, it is generally not possible to distinguish strains, nor, in some cases, closely related species. The OTUs are analyzed at each taxonomic level, but are less precise at the species level.Amplicon based 18S/ITS is one of the basic components of fungal cells and comprises both conserved and hypervariable regions. The internal transcribed spacer region, ITS, is located between the 18S and 5.8S rRNA genes and has a high degree of sequence variation. The 18S rRNA is mainly used for high resolution taxonomic studies of fungi, while the ITS region is widely used for analyzing fungal diversity in environmental samples. Taxonomic studies of fungi are often based on the nuclear ribosomal gene cluster, which includes the 18S or small subunit (SSU), 5.8S subunit, and 28S or large subunit (LSU) rRNA genes. ITS1 and ITS2 have been found to be the most suitable markers for fungal phylogenetic analysis due to their variable sequences, conserved primers and multicopy nature. Various pipelines, such as QIIME, MG-RAST and Mothur are used to perform taxonomic and functional analysis. In addition to rRNA genes, other amplicon-based studies are performed in order to focus on specific functions such as nitrogen fixation activity, and diversity analysis of nitrogenase reductase (nifH) genes.

B) Meta-transcriptomics and Meta-proteomics: RNA-Seq analysis of microbial communities in a complex ecosystem is known as Meta-transcriptomics (Zhang *et al.*, 2017). Co-expression of gene clusters and transcript abundance, followed by functional annotation, can be studied in environmental samples (Oyserman *et al.*, 2016). The quantitation of mRNA and pathway expression can be carried out using meta-transcriptomics. The challenge associated with meta-transcriptomic approaches is to get high quality RNA from environmental samples; given this, it is an efficient approach to elucidate gene expression, and to discover novel genes in the microbial community. The study of the proteome expressed in the microbial community is known as Meta-proteomics. This has been used to investigate microbial activities along with complex metabolic

pathways in soil ecosystems. Community metaproteomics are emerging as complementary approaches to metagenomics and can provide large-scale characterization of proteins in the microbiota. Meta-genomics, along with meta-transcriptomics and meta-proteomics, provide insights into functional dynamics, prediction of the *in situ* microbial responses/activities, and the production capabilities of microbial communities.

Applications of metagenomics:

a) Metagenomics has a wide range of applications from clinical to environmental samples, from food safety to industrial waste, and also has the ability to identify pathogens.
b) Metagenomics provides information about the diversity of organisms in environmental samples and has provided insights in industrial research.
c) Functional metagenomics has been used for identification of several biocatalysts, which are available in the market. Novel cellulases with improved enzymatic characteristics have been identified.
d) The use of metagenomics, metatranscriptomics and metaproteomics approaches enhance enzyme discovery and can be used to efficiently screen for highly active enzymes.
e) Recent studies have stated that metagenomics can be used as a bioremediation tool; in comparison with other approaches of bioremediation, metagenomics gave better degrading ratios. Metagenomics study help in identifying different widespread microorganism and their respective functions in polluted environment; these microorganisms are the best tools in nature to degrade toxic pollutants.
f) Metagenomics has been used in clinical diagnostics
g) Viral metagenomics has the power to identify the root cause of novel epidemic diseases.
h) Metagenomics is also used in medical or forensic investigations, and to solve challenges in the field of medicine, agriculture and ecology.
i) Major applications of metaproteomics have included investigation of acid mine drainage biofilms, activated sludge, soil, human gut microbiota, and other environmental samples.

References

Handelsman, J., Rondon, M.R., Brady, S.F., Clardy, J., Goodman, R.M. 1998. Molecular biological access to the chemistry of unknown soil microbes: A new frontier for natural products. Chemistry & Biology 5(10), R245–R249.

Langille, M.G., Zaneveld, J., Caporaso, J.G.2013. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology 31 (9), 814–821

Vieites, J.M., Guazzaroni, M.E., Beloqui, A., Golyshin, P.N., Ferrer, M. 2008. Metagenomics approaches in systems microbiology. FEMS Microbiology Reviews 33 (1), 236–255.

Morgan, X.C., Huttenhower, C. 2012. Human microbiome analysis. PLOS Computational Biology 8 (12), e1002808

Dinsdale, E.A., Edwards, R.A., Hall, D. 2008. Functional metagenomic profiling of nine biomes. Nature 452 (7187), 629

Tringe, S.G., Von Mering, C., Kobayashi, A. 2005. Comparative metagenomics of microbial communities. Science 308 (5721), 554–557

Zhang, Y., Sun, J., Mu, H., Lun, J.C., Qiu, J.W. 2017. Molecular pathology of skeletal growth anomalies in the brain coral *Platygyracarnosa*: A meta-transcriptomic analysis. Marine Pollution Bulletin.

Oyserman, B.O., Noguera, D.R., del Rio, T.G., Tringe, S.G., McMahon, K.D. 2016. Metatranscriptomic insights on gene expression and regulatory controls in Candidatus Accumulibacterphosphatis. The ISME Journal 10 (4), 810.

Chapter-23

Molecular markers for Entomological Research

S. Mohankumar
Centre for Plant Molecular Biology and Biotechnology, Tamil Nadu Agricultural University, Coimbatore, Tamil Nadu
Email: smktnau@gmail.com

Introduction

Insect pests constitute the most diverse group of organisms on earth with more than one million described species and millions more either awaiting for description or simply discovered (Grimaldi and Engel, 2005). Molecular tools come handy in undersatanding the diversity of the insect species. The information of population structure and genetic diversity at inter and intra specific level is crucial in order to develop an effective integrated pest management strategy. Analysis of population genetic structure, i.e. the distribution of genetic variation within and among populations, is a key aspect to understand insect pest population dynamics in agricultural scenarios and understanding population structures provides the most fundamental information for reliable identification of species and design of management strategies. Moreover, understanding the genetics of pest invasion may help to identify the origin, the number of introductions and the spread of the infestation of a pest in an area. Hence, in the present paper, studies on inter and intra species variation in insect pests of crops, natural enemies and pollinators using molecular tools are discussed.

Molecular markers in genetic diversity of insect pests

Many different types of molecular markers have been developed and can be used for this purpose (Avise, 2004).In insects, DNA based markers are widely used to obtain the basic information of organisms at molecular levels and that would help to measure the genetic diversity and gene flow between species, identifies haplotypes and lineages or predicts migration and colonization history of insect pests (Bosio *et al.* 2005). In addition, the molecular marker based data provide the information to differentiate sympatric species from allopotaric species and parapatric species (Margonari *et al.*, 2004). The details of gene flow and genetic variations within and between insect species, measured from marker data, are critical to establish meaningful explanation for population structure and dynamics (Mendelson & Shaw 2005). DNA based molecular markers are used to infer phylogeny and biogeography of insect populations and to understand modes of evolution and evolutionary trajectories (Prasad *et al.*, 2005). Also, diagnostic molecular markers, based on linkage to certain traits or genes, are used for diagnostic purposes of individual insects (Ullmann *et al.*, 2003). Greater level of polymorphism could be obtained by using DNA markers than by using other markers (i.e. protein) (Richardson *et al.*, 1986). This is because of mutations in introns or even in the codons

of a gene can potentially provide more variation at the DNA level than at the protein level. Moreover, DNA samples are more stable than proteins and are unchanged for detection at all time and tissue of the organism unlike proteins. Over the last past 25 years, DNA makers have made a significant contribution to rapid rise of molecular studies of genetic relatedness, phylogeny, population dynamics or gene and genome mapping in insects.

Current trends of application of DNA marker techniques in diverse field of insect studies demonstrated that mitochondrial (mtDNA*COI*) and nuclear DNA based markers (Simple sequence repeat (SSR), Inter simple sequence repeat (ISSR), random amplified polymorphic DNA (RAPD), and amplified fragment length polymorphism (AFLP) have contributed significant progresses towards the genetic basis of insect diversity. In addition to the above mentioned markers, other approaches including transposon display, sequence-specific amplification polymorphism (S-SAP), repeat-associated polymerase chain reaction (PCR) markers have been identified as alternate marker systems in insect studies. Besides, whole genome microarray and single nucleotide polymorphism (SNP) assays are also becoming more popular to screen genome-wide polymorphisms in fast and cost effective manner.

DNA Barcoding and Meta-barcoding Approach

Traditionally, taxa have been identified using morphological characters *viz*., size, shape, colour, and anatomical structures. However, this can result in sometimes intractable problems especially for small insects, members of cryptic and polymorphic species and immature stages (Smith *et al.*, 2006). DNA barcoding represents the complementary tool for identifying the cryptic and previously overlooked species. Mitochondrial enzyme Cytochrome c Oxidase I (*COI* – 658 bp) coding DNA has been identified as an effective marker for the development of DNA barcode in insect pests because of their high frequency of mutations.

DNA barcoding of Neotropical skipper butterfly *Astraptesfulgerator*discovered the occurrence of ten cryptic species, which were earlier considered as single species for more than a century (Hebert*et al.*, 2004). Rebijith*et al.* (2012) studied the usefulness of *cytochrome c oxidaseI (CO-I)* gene for the discrimination of mirid species in India *viz. Helopeltis antonii*, *H. thievora*, *H. bradyi* and *Pachypeltis maesarum*in their various life stages. The results showed 1.0 per cent intraspecific divergence for all the four species examined, whereas the inter-specific distances were in the range of 7 to 13 per cent and demonstrated that the DNA barcode certainly helps the identification of mirids in India and will stand as a decisive tool in formulating integrated pest management (IPM) strategy, quick identification of invasive and cryptic species, haplotypes and biotypes.

The red lined geometrid *Crypsiphonaocultaria* (Donovan), shows external morphological differences, but was previously considered a single species. The *COI* barcode was used to test the hypothesis of two separated species (Ounap and Viidalepp, 2009), and the results showed that a phylogeny based approach allowed the delimitation of *C. ocultaria*and the new species *C. tasmanica*, though distance-based delimitation is problematic due to substantial overlap in

intra and interspecific genetic distances. Oba *et al.* (2015) studied the identification and biodiversity of Japanese click beetle (Coleoptera; Elateridae) using DNA barcoding approach.

DNA barcoding approaches also play a major role in discrimination of cryptic species. Cryptic species are the species which do not interbreed but whose morphology is similar, making them very difficult or impossible to differentiate using morphological criteria alone. The DNA barcoding approaches have been used to discriminate the cryptic species in diverse groups of insects. Revelation of cryptic species by barcode data has been documented in several insect species including sphingid moths (Vaglia *et al.,* 2008), leaf mining micro moths (Nieukerken *et al.,* 2012), aphids (Carletto *et al.,* 2009), white flies (Ashafq *et al.,* 2014) and thrips (Iftikhar *et al.,* 2016).

Understanding the interactions between insect pests and its natural enemies is essential for effective biological control of insect pests and biodiversity protection. With this, DNA barcoding approach has been successfully used in studies of host and parasitoid interactions. However, this approach does not allow the simultaneous detection of hosts and parasitoids. The recent advent of high-throughput sequencing such as DNA meta-barcoding approach have been used to resolve many ecological interactions (Hajibabaei, 2012). Meta-barcoding is the use of Next Generation Sequencing (NGS) for the automated identification of multiple species from a single bulk sample containing entire organisms or from a single environmental sample containing degraded DNA (Taberlet*et al.,* 2012b). Sigut*et al.* (2017) studied the host parasitoid interactions in Lepidoptera and Hymenoptera (Saw flies) (5 hosts and 14 parasitoids) using DNA meta-barcoding.

Population structure and insect pest invasions

Invasive species may rapidly spread in the new areas due to superior competitive ability and adaptation to new environment (Sax *et al.*, 2005) which may affect the local fauna and flora and potentially cause serious damage to agriculture (Pimentel *et al.,* 2001). In this connection, the information on geographical origins, introduction routes and biology in native regions of such invasive species is crucial for identifying the means of transport, preventing reintroduction and development of effective management and eradication techniques (Estoup and Guillemaud, 2010). Among the molecular markers, the microsatellite markers and the universal mtDNA based markers are highly used to measure the invasion of insect pests in different geographical regions.

Hosokawa *et al.* (2014) using mitochondrial DNA sequences of the introduced and East Asian native *Megacopta* populations identified as a well-supported clade consisting of the introduced populations and *M. punctatissima* populations, which strongly suggests that the invading *M. cribraria* populations are derived from a *M. punctatissima* population in the Kyushu region of Japan. Fraimout *et al.* (2015) reported the development of polymorphic microsatellite markers in the invasive spotted wing drosophila, *Drosophilla Suzuki* designed from recent genetic resources and their cross amplifications in closely related *Drosophilla* species of the *suzuki* sub group. *Bactrocera correcta* is one of the most destructive pests of several horticultural crops

in tropical and subtropical region of the world and the recent genetic diversity and population structure analysis of *B. correcta* with *mtCOI* and 12 microsatellite markers demonstrated low genetic diversity across Asian populations and also reported that the *Bactrocera* species may have originated in India (Qin *et al.*, 2016).

Luo and Agnarssan (2017) studied the global mtDNA genetic structure and invasion history of a major pest of citrus, *Diaphorinacitri* using cytochrome oxidase I (*COI*) gene and the analysis of molecular variation showed that route involved the expansion of lineage B from southern Asia into North America via West Asia. The second, the expansion of some lineage A individuals from Southeast Asia into East Asia, and the third involved both lineages from Southeast Asia spreading westward into Africa and subsequently into South America.

Next Generation Sequencing (NGS) and SNP Markers

Next generation sequencing tools or high-throughput sequencing (Illumina (Solexa) sequencing platforms such as Roche 454 sequencing. Ion torrent: Proton / PGM sequencing) have been widely used in the evolutionary studies (Dijk*et al.*, 2014). these studies have increased the level of understanding of insect genome in the area of mitogenomics and phylogenomics and provided valuable insights into classification and evolution of insects.

The study of complete genome of mitochondria of organism is called Mitogenomics. Complete genomes are now available for numerous arthropod species and entire mitochondrial genomes have been sequenced for nearly 500 insect species (Cameron, 2014). The studies on Mitogenomics of Hymenoptera have demonstrated many interesting features including gene arrangements are conserved in the basal Hymenoptera, i.e., the grade Symphyta, whereas frequent gene rearrangements are observed in the derived clade, Apocrita (Dowton *et al.*, 2002) with approximately equal amounts of gene shuffling, inversion, and translocation (Dowton and Austin, 1999).

Gill et *et al.* (2014) used bulk de novo mitogenome assembly from pooled total DNA of weevil to study the higher level phylogenetic relationships in the weevils (Coleoptera: Curculionoidea) and produced 92 assembled mitogenomes through single Illumina MiSeq run and the results demonstrated a separate origin of wood-boring behavior by the subfamilies Scolytinae, Platypodinae, and Cossoninae.

Phylogenomics is the combination of phylogenetics with genome data, has emerged as a powerful method to study the evolution of species and systematics. It provides opportunities for comparative genomics and large-scale multigene phylogenies of diverse lineages of insects. Phylo genomic investigations help us better understand systematic and evolutionary relationships of insect species that play important roles as herbivores, predators, detritivores, pollinators, or disease vectors (Behura, 2015). Misof*et al.* (2014) used Phylogenomic approach and reported the timing and pattern of insect evolution. They have dated the origin of insects to the Early Ordovician (~ 479 million years ago (Ma)), of insect flight to the Early Devonian

(~406 Ma), of major extant lineages to the Mississippian (~345 Ma), and the major diversification of holometabolous insects to the Early Cretaceous.

Marker Assisted Selection for pest resistance

Extensive research, including the generation of many genetic maps in different crops, has demonstrated that the majority of the amplified restriction fragments correspond to unique loci in the genome. The preferable approaches are all based on screening a limited number of samples with a relatively large number of primer pairs.

The classical approach for the identification of loci involved in complex polygenic traits consists in the screening of a large number of individuals from a segregating population with a set of markers that are evenly distributed throughout the genome. Subsequently, statistical analysis is performed to identify regions in the genome that are involved in the trait. The laborious nature of this approach makes it unrealistic to screen sufficiently large populations to precisely locate the quantitative trait loci (QTL). As a consequence, the QTL cannot be localized precisely on the map and closely linked markers cannot be obtained, thereby preventing the broad scale application of indirect selection for quantitative traits.

Marker assisted backcross breeding is now becoming a standard application in modern plant breeding. In selection for high recurrent parent genomic content, the DNA fingerprints are used to calculate the per cent recurrent parent genome in each backcross individual, thereby taking the genome representation of the markers into account. When negative characteristics are linked with the trait that needs to be introgressed, molecular markers can be used to select for recombinants in the region. After phenotypic testing of these recombinants, individuals may be selected in which the region responsible for the linkage drag has been removed from the locus of interest. Yencho *et al.* (2000) reviewed how molecular markers can be used to increase the understanding of the mechanisms of plant resistance to insects and develop insect resistant crops. Also, genes controlling resistance to different races or biotypes of a pest or pathogen, or genes contributing to agronomic or seed quality traits can be pyramided together to maximize the benefit of MAS through simultaneous introgression (Dwivedi *et al.,* 2007). The use of molecular markers is very reliable method to diagnose the integration of the two genes in the plants.

Examples of Marker assisted breeding for pest resistance

Crop	**Pest**
Rice	Brown plant hopper, Gall midge
Wheat	Hessian fly
Maize	Corn earworm, South western corn borer and sugarcane borer
Sorghum	Sorghum midge
Cotton	*H. armigera*
Chick pea	Pod borer

Cowpea	Aphid, Bruchid
Green gram	Bruchid
Soybean	Corn ear worm, Soybean aphid
Pea	Russian wheat aphid
Bean	Common bean mosaic virus, Bean pod weevil (*Apiongodmani* Wagner) in common bean.
Potato	Leafhopper
Tomato	*H. armigera*,

References

Ashafq, M., Hebert, P.D.N. Mirsa, M.S. Khan, A.M. Mansoor, S. and Shah, G.S. 2014. DNA barcoding of *Bemisiatabaci* complex (Hemiptera; Aleyrodidae) reveals southerly expansion of the dominant whitefly species on cotton in Pakistan. *PLoS ONE*, 9: e104485.

Behura, S.K. 2015. Insect phylogenomics. *Insect Mol Biol.*, 24(4): 403–411.

Cameron, S.L. 2014. Insect mitochondrial genomics: implications for evolution and phylogeny. *Annu. Rev. Entomol*., 59: 95–117.

Carletto, J., Blin, A., and Vanterberghe – Masutti, F. 2009. DNA based discrimination between the sibling species *Aphis gossypii* Glover and *Aphis frangulae*Kattenbach. *SystEntomol.* 24:307-314.

Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Termes, C. 2014. Ten years of next-generation sequencing technology. *Trends Genet*. 30, 418–426.

Estoup, A., and Guillemaud, T. 2010. Reconstructing routes of invasion using genetic data: why, how and so what? *Mol. Ecol.*, 19: 4113–4130

Fraimout, A., Loiseau, A., Price, D. K., Xuereb, A., Martin, J.-F., Vitalis, R., Felous, S., Debat, V., &Estoup, A. (2015). New set of microsatellite markers for the spotted-wing Drosophila suzukii (Diptera :Drosophilidae) : A promising molecular tool for inferring the invasion history of this major insect pest. *European Journal of Entomology*, *112*(4), 855-859.

Gillett, C.P., Crampton-Platt, A., Timmermans, M.J., Joredl, B.H., Emerson, B.C and Vogler, A.P. 2014. Bulk de novo mitogenome assembly from pooled total DNA elucidates the phylogeny of weevils (Coleoptera: Curculionoidea). *Mol. Biol. Evol*., 31: 2223–2237.

Hajibabaei M. The golden age of DNA metasystematics. Trends in Genetics. 2012;28:535–537.

Hebert, P.D.N., Cywinska, A., Ball, S.L., and Dewaard, J.R. 2003. Biological identifications through DNA barcodes. *Proc. R. Soc. Series B.* 270: 313–321.

Hebert, P.D.N., Penton, E.H., Burns, J.M., Janzen, D.H., Hallwachs, W. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptesfulgerator. *Proc. Natl. Acad. Sci. U.S.A.* 101, 14812-14817.

Hosokawa, T., Nikon, N. and Fukatsu, T. 2014. Fine scale geographical origin of an insect pest invading North America. *PLoS ONE,* 9(2):e89107.

Iftikhar, R., Ashfaq, M., Rasool, A., and Hebert. P.D.N. 2016. DNA Barcode analysis of thrips (Thysanoptera) diversity in Pakistan reveals cryptic species complex, *PLoS ONE*, 11(1):e01460114.

Junqueira A. C., Azeredo-Espin A. M., Paulo D. F., Marinho M. A., Tomsho L. P., Drautz-Moses, D.I. Purbojati, R.W. Ratan A and Schuster, S.C. 2016. Large-scale mitogenomics enables insights

into Schizophora (Diptera) radiation and population diversity. *Sci. Rep.*, 6: 21762. DOI: 10.1038/srep21762

Luo, Y. and Agnarsson, I. 2017. Global mtDNA genetic structure and hypothesized invasion history of a major pest of citrus *Diaphorinacitri*(Hempitera; Liviidae). *Ecol. Evol*., 1 – 10.

Nieukerken, E.V., Doorneweerd, C., Stokvis, F.R., Groenenberg, D.S.J. 2012. DNA barcoding of the leaf mining moth sub genus Ectodemia S. Str. (Lepidoptera; Noctuidae) with COI and EF1-a two are better than one in recognizing cryptic species. *Contrib Zool*., 81: 1 – 24.

Oba Y, Ôhira H, Murase Y, Moriyama A, Kumazawa Y (2015) DNA Barcoding of Japanese Click Beetles (Coleoptera, Elateridae). PLoS ONE 10(1): e0116612.

Õunap E, Viidalepp J (2009) Description of *Crypsiphonatasmanica* sp. nov. (Lepidoptera: Geometridae:Geometrinae), with notes on limitations in using DNA barcodes for delimiting species. Aust J Entomol 48: 113-124.

Pimentel, D., McNair, S., Janecka, J., Wightman, J., Simmonds, C., *et al*. 2001. Economic and environmental threats of alien plant, animal, and microbe invasions. *AgricEcosyst Environ*, 84: 1–20.

Qin, Y., Buahom, N., Krosch, M.N. Du, Y., Wu, Y., Malacrida, A.R., Deng, Y.L., Liu, J.Q., Jinag, X.L., and Li, Z.H. 2016. Genetic diversity and population structure in Bactroceracorrecta inferred from mtDNA and microsatellite markers. *Sci. Rep*., 6, 38476; doi: 10.1038/ srep38476

Rebijith, K. B., Asokan, R., Krishna Kumar, N. K., Srikumar, K. K., Ramamurthy, V. V., and Shivarama Bhat, P. 2012. DNA barcoding and development of species-specific markers for the identification of tea mosquito bugs (Miridae: Heteroptera) in India. *Environ. Entomol.* 41(5): 1239–1245.

Sax DF, Kinlan BP. Smith KF. A conceptual framework for comparing species assemblages in native and exotic habitats. *Oikos*, 108: 457–464.

Schwentner, M., Combosch, D.J., Nelson, J.P. and Giribet, G. 2017. A phylogenomic solution to the origin of insects by resolving Crustacean-Hexapod relationships. *Curr. Biol*., 12(12): 1818–1824.

Šigut M, Kostovčík M, Šigutová H, Hulcr J, Drozd P, and Hrcek, J. (2017) Performance of DNA metabarcoding, standard barcoding, and morphological approach in the identification of host–parasitoid interactions. *PLoS ONE*,12(12): e0187803.

Smith, M.A., Woodley, N.E., Janzen, D.H., Hallwachs, W., Hebert, P.D.N. 2006. DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera :Tachinidae). Proc. Natl. Acad. Sci. U.S.A. 103, 3657-3662.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.*,21: 2045–2050.

Vaglia, T., Haxaire, J., Kitching, I.J., Meusnier, I., and Rougerie, R. 2008. Morphology and DNA barcoding reveal three cryptic species within the *Xylophanesneoptolemeus* and loelia species groups (Lepidoptera; Sphingidae). *Zootaxa*, 1923:18-36.

Chapter-24

Quantitative PCR-Principles and Practices

B. Ramakrishnan
Division of Microbiology, ICAR-Indian Agricultural Research Institute, New Delhi–110012
ramakrishnanbala@yahoo.com

The quantitative PCR (qPCR) is the keystone method for absolute or relative quantification of DNA or RNA in the molecular investigations in biological sciences, medicine, and molecular diagnostics. In the PubMed database, there are about 183,105 and 172,765 citations until today for qPCR and real-time PCR, respectively. Higuchi *et al.* (1992) provided the The first demonstration of real-time analysis of PCR.The company "Idaho Technology" that manufactured a rapid-cycle PCR added flow cytometry optics and led to the "LightCycler" real-time instruments (Lyon and Wittwer 2009; Wittwer 2017). The commercial LightCycler systems (Idaho Technology in 1997 and Roche in 1998) pioneered the melting analysis, the use of generic indicator SYBR Green I, and rapid temperature cycling for making immediate data presentation to real-time PCR. Ririe *et al.* (1997) introduced the melting analysis as a way for characterizing PCR products after the last cycle of PCR.Applied Biosystems introduced the first real-time thermocyclers commercially in 1997 (Wilhelm and Pingoud 2003). Instruments with competing capabilities of high-throughput and automated technology with lower turnaround times are currently available in the market.

In the PCR reactions, the amount of starting target sequence has a quantitative relationship with the amount of PCR accumulated at each cycle. The target sequence will double approximately during each cycle in an optimized reaction. In the qPCR, the amount of amplified product is linked to the fluorescence intensity of the fluorescent reporter molecule and the fluorescent signals can be measured at the endpoint (endpoint semi-quantitative PCR) or during the exponential phase of amplification (real-time PCR). The presence of inhibitors, pyrophosphate accumulation or limitations of reagents makes the end pointquantitation of PCR products inaccurate. The exponential phase, rather than the plateau- or other phases of the PCR reaction is useful to determine the starting quantity of target sequence.

The labelling of amplified DNA can be achieved by different fluorescent dyes, also known as qPCR chemistries. The fluorescent reporter molecules are generally of two classes: double-stranded DNA-binding dyes and dye-labelled probes. In each cycle of PCR, the fluorescence intensity is measured, which is proportional to the increases in the amplicon concentrations. The fluorescence intensity is measured during the whole PCR process. The resulting plots of fluorescence intensity vs cycle number are prepared with the background fluorescence at a common starting point (baseline correction). The cycle at which the fluorescence intensity

exceeds a detection threshold, the Ct (threshold cycle) correlates with the number of target sequences present. Higher the initial number of target sequences in the sample, faster the fluorescence intensity will increase during the PCR reaction.

The cycle in which fluorescence intensity can be detected is termed as quantitation cycle (Cq), and lower Cq values mean higher initial copies of target sequences. The threshold cycle (Ct) or the quantitation cycle (Cq) are determined by the standard curve dilution series, also known as "absolute" quantification. The standard curve approach is employed when the objective of the experiment is to measure the exact number of target sequences. The templates for the standard curve quantification can include genomic DNA, cDNA, total RNA, or plasmids containing the cloned gene of interest. The log values of the initial template copy number of each dilution are plotted against the Ct generated to prepare the standard curve, and the linear regression line is calculated. The Ct values of unknown samples can be compared to the standard curve for quantification of initial copy numbers. The standard curve dilution series need to consider the entire range of concentrations that will be measured in the unknown samples. The linearity of the regression line, denoted by R^2 or Pearson Correlation Coefficient, should be close to 1. The standard templates (DNA or RNA) are to be quantified using the standard methods (UV spectrophotometry or nucleic acid binding dyes) in replicates, with no template controls (NTC).

The 'relative quantification' approach is employed to examine 'gene expression' by measuring the relative concentration of the gene of interest in unknown samples compared to a calibrator. The calibrator, which is a baseline for the expression of a target gene serves as a benchmark to which other samples can be compared. The differences in Ct values as fold-changes between the unknown sample and the calibrator are expressed to state whether the gene of interest is up- or down-regulated. In addition to comparing the expression of target gene in control versus experimental sample, normalization using a reference gene whose expression is constant in both the control and experimental samples are desired. The actual amplification efficiencies of target and normalizer need to be established using the multiple standard curve approach to verify the reproducibility of efficiency measurements.

The qPCR chemistries include the non-specific DNA binding dyes such as SYBR Green I and EvaGreen. These non-binding dyes are easy to use, reasonably priced, and have relatively low fluorescence when free in solution. The fluorescence intensity increases by more than 1000-fold, and proportionately to the double DNA concentration. The inherent disadvantage of non-specific DNA binding dyes is their non-specificity. Hence, the primer dimer formation due to non-specific binding of primers needs to be avoided. The presence of non-specific signal can be detected by the melt-curve analysis. The amplified sequences can be characterized in the melting curve analysis using their apparent melting temperature (T_m), which is a function of product length and base composition. Following the PCR amplification, the amplified sequences can be slowly melted while fluorescence from the dye is monitored. The amplified

DNA melts with increasing temperature and the fluorescence density gets decreased. In the PCR amplification where the amplified sequences consist of molecules of homogeneous length and base composition, a single thermal transition is detected. Otherwise, multiple thermal transitions in the fluorescence intensity will be detected. The specific and non-specific amplicons based on the melting temperature (T_m) of the reaction end-products can be differentiated from the fluorescence versus temperature curve (also known as the dissociation curve).

The probe-based chemistries provide a higher level of detection specificity. The internal, fluorescent probe will not fluoresce, remain quenched in the absence of a specific target sequence. When the probe hybridizes to the target sequence, the fluorescence of reporter dye can be detected. TaqMan probes which are the third oligonucleotide in the PCR reaction and have the fluorescent reported dye (FAM) attached to the 5' end and a quencher (TAMRA) or a dark quencher (Black Hole Quencher) at the 3' end. The fluorescence from the reporter dye is quenched as long as both the reporter and quencher are in close proximity. TaqMan probes use the FRET (Fluorescence Resonance Energy Transfer) quenching mechanism. The probe designing is done to anneal to one strand of the target sequence, just slightly downstream of one of the primers. Taq polymerase which has 5'-3' nuclease activity encounters the probe, displaces and degrades the 5' end, and releases the reporter dye free into solution. The separation of reporter dye and quencher will help to detect the reporter dye. TaqMan chemistry (specialized probes such as MGB or LNA probes) is used in a multiplex reaction where a separate probe is designed for each target sequence (allele) and each probe labelled with different fluorophores (e.g., FAM and HEX).

The qPCRmethods with detection of deviations in the amplification efficiency of individual reactions are more precise than end-point determinations. The most challenging steps of qPCR analysis are the primer and probe designing, and the fluorescence detection chemistry. The techniques of qPCR and reverse transcription (RT)-qPCR have become very popular now and are used in biomedical and agricultural research, and biotechnological applications. To improve the quality, reproducibility, and interpretability of qPCR data, the 'minimum information for the publication of quantitative real-time PCR experiments' (MIQE) needs to be considered in all the applications, more so for genotyping and diagnosis (Taylor *et al.* 2010; Bustin and Nolan 2017). The applications of qPCR are growing rapidly, which indicates that qPCR will continue to remain one of the sought-after technique in medicine, molecular life sciences and agricultural sciences.

References

Bustin S, Nolan T (2017) Talking the talk, but not walking the walk: RT-PCR as a paradigm for the lack of reproducibility in molecular research. Eur J Clin Invest 47:756-774.

Higuchi R, Dollinger G, Walsh PS, Griffith R. (1992) Simultaneous amplification and detection of specific DNA sequences. Biotechnology (N Y) 10:413–417.

Lyon E, Wittwer CT (2009) LightCycler technology in molecular diagnostics. J Mol Diagn 2009; 11:93–101.

Ririe KM, Rasmussen RP, Wittwer CT (1997) Product differentiation by analysis of DNA melting curves during the polymerase chain reaction. Anal Biochem245:154–60.

Taylor S, Wakem M, Dijkman G, Alsarraj M, Nguyen M (2010) A practical approach to RT-qPCR publishing data according to the MIQE guidelines. Methods 50:S1-S5.

Wilhelm J, Pingoud A (2003) Real-Time polymerase chain reaction. ChemBioChem 4: 1120-1128.

Wittwer CT (2017) Democratizing the real-time PCR. 63:924-925.

Chapter-25

RNAi in Insect Pest Management

Vinay Kalia
Division of Entomology, ICAR-Indian Agricultural Research Institute, New Delhi-110012
vkalia@iari.res.in; vkalia_1998@yahoo.com

Every living creature in this universe is engaged in a constant struggle for food and shelter, of them some are our competitors like insect pests. Insect pests cost billions of dollars in the form of crop losses. In India, insects alone are reported to cause crop losses of US $ 1.5 billion annually. Sprayed insecticides are the most widespread tactic for insect control. These are associated with environmental hazards, may cause pest resurgence and develop resistance in target pests. More than 500 species of insects and related arthropods have evolved resistance to one or more insecticides globally. Resistance is a critical area of pest control and is indeed a global phenomenon. Resistance to chemical pesticides has been documented for over six decades. With the advent of GM technology transgenic crops have emerged as a dynamic tool for the management of insect pests based on practical success shown by *Bacillus thuringiensis* (Bt) toxin in the protection of wide categories of crops. However, the success of this technology is threatened by development of resistance to Bt toxins in sprays and transgenic crops, just like synthetic insecticides. Therefore, there is a need to look for another sustainable and environment friendly approach to manage insect pests and one such approach could be ribonucleic acid interference (RNAi) technology.

The main antiviral immune system of insects is the post transcriptional gene silencing mechanism known as RNAi. This, post transcriptional mechanism of silencing gene function by inserting short homologous sequence of messenger RNA (mRNA) to prevent translation of proteins. For decades, RNAi was known to occur in plants (as *post transcriptional gene silencing*) and fungi (as *quelling*) but was reported for the first time in animals (*Caenorhabditis elegans*). In 2003, Science named it as the "Breakthrough of the Year" and Fortune magazine hailed it as "Biotech's Billion Dollar Breakthrough". Andrew Fire and Craig Mello unveiled the underlying mechanism of this phenomenon in *C. elegans* for which they were awarded Nobel Prize. Fire *et al.* (1998) systematically clarified that double-stranded RNA (dsRNA) was more effective than either sense or antisense RNA in silencing genes, and they called dsRNA-induced silencing phenomenon as RNAi. This pathway needs interfering molecules and special type of enzyme called dicer (RNase III) enzyme. The interfering molecules are small interfering RNAs (siRNAs) and microRNAs (miRNAs). These interfering molecules are then loaded onto a "RNA induced silencing complex'' (RISC) which is associated with an Argonaute protein (AGO). RISC helps the interfering molecules to find their target where the cleavage of target mRNA based on homology binding can occur (Fig.1). Since it's discovery RNAi has been used successfully as a powerful reverse genetic tool to study the function of

various genes associated with insecticidal target protein and thereby controlling various agriculturally important insect pests.

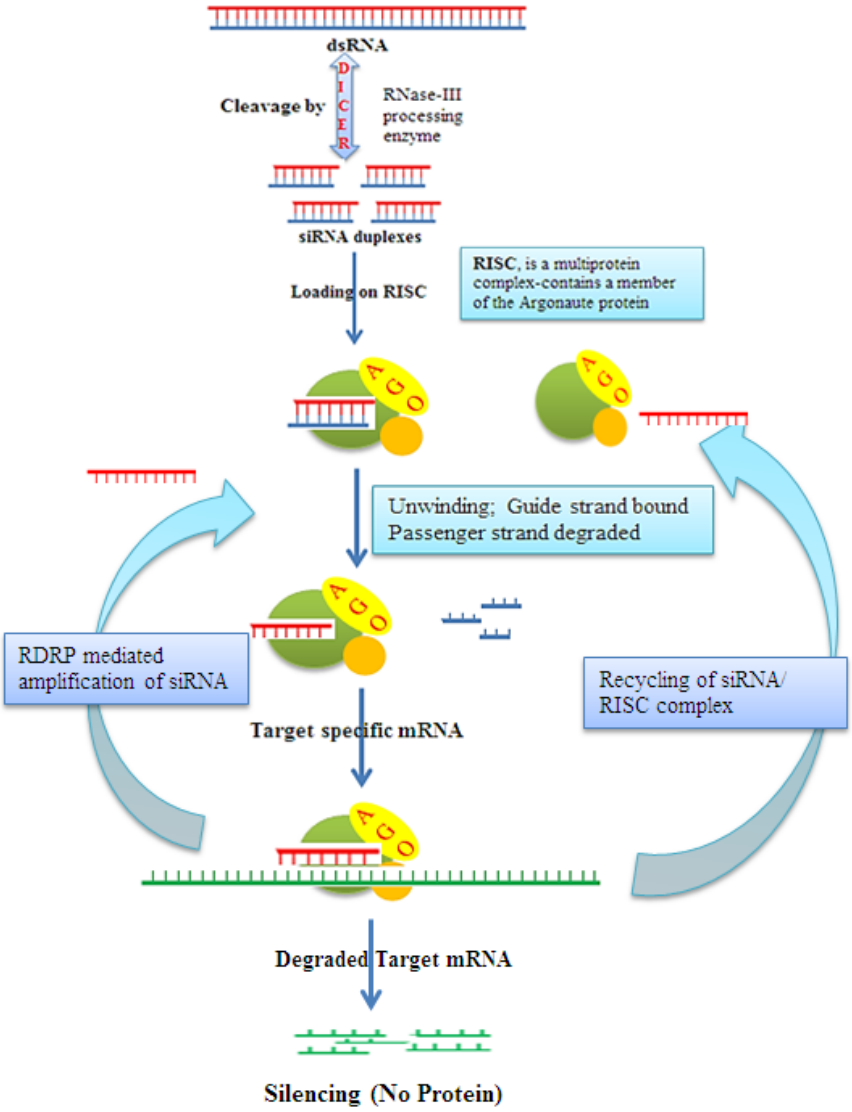


Figure 1: RNA interference mechanism

RNAi has progressed much in the last decade and became a powerful and efficient genetic tool that has greater utility. Thus it has also become a choice of method for controlling insect pest problem. It is a tailor-made insecticide, which is highly species specific. However, the efficiency of this mechanism varies depending upon several factors such as the test insect, stage of insect, candidate gene selection, delivery system adopted and stability of the trigger molecule. Apart from the successful implication in diverse areas, there are certain drawbacks of this technology such as 'off-target' effects, lack of sensitivity of various insect species. Further research would relieve these limitations and support the manifestation of this genetic tool with much more reliability.

RNAi mechanism in insect

In insect an antiviral immune defense system involved the RNAi pathway. It is activated when dsRNA molecule, as products of viral replication are recognized in the cytoplasm and processed into siRNA by RNase type III enzyme Dicer-2. Cleavage of siRNA generated from viral dsRNA targets is further used by an Argonaute 2 (AGO2) protein associated with RISC (multi protein complex) which involves the siRNA guide strand (Figure 1). Inventively, this RNAi mechanism also be triggered by an artificial administration of gene-specific long dsRNA. This dsRNA treatment can result in functional knockdown effects that in fact can be considered as auto immune defect since it is an antiviral immune defense mechanism of insects.

The RNA precursor molecules from the RNAi pathways in insects are identified as small RNAs. These are of three type viz. small interfering RNAs (siRNAs; 20-25 nucleotides) microRNAs (miRNAs; 21-24 nucleotides) and the PIWI-interacting RNAs (piRNAs; 24-30 nucleotides). Both miRNAs and siRNAs share a common RNase-III processing enzyme, Dicer. While, piRNAs are independent of Dicer activity. In the siRNA pathway, the dsRNA are processed by Dicer into siRNA duplexes. But the miRNAs are generated from endogenous transcripts in nucleus as pre miRNAs. It is then processed by Enzyme Drosha and finally transported to cytoplasm. Both these RNAs contain ribonucleoprotein particles (RNPs). The siRNAs contain RISC "RNA-induced silencing complex" and a miRNA contain miRNPs. Every RISC or miRNPs contains a member of the Argonaute (AGO) protein family. The AGO protein uses the guide RNA to associate with target RNAs and then slicing of the target mRNA occurs (Fig.1). In most cases a single Dicer is responsible for both the siRNA and miRNA pathways. While in *Drosophila melanogaster* has two prologues, one is Dicer1 which process miRNAs and other Dicer2 which process siRNAs. In Drosophila, AGO-1 is involved in the miRNA pathway and AGO-2 in the siRNA pathway. Although in *Tribolium castaneum* only one type of AGO protein has been identified in miRNA (i.e., Tc Ago1) and two classes of AGO protein in siRNA (viz. Tc-AGO 2a & Tc-AGO 2b). Whereas in *Bombyx mori* three RNAi pathways has been identified (Kolliopoulou *et al.*, 2014). As such RNAi has become the most widely used reverse genetics research tool in insect and have great potential to contribute to novel strategies for species specific control of insect pests and to overcome viral infections in disease vectoring and beneficial insects.

Transport of RNAi information

There are two types of RNAi response observed in insects *viz.*, cell-autonomous and non-cell-autonomous RNAi. In the case of *cell-autonomous RNAi*, the silencing process is limited to the cell in which the dsRNA is expressed or introduced and comprises the RNAi process within individual cells. In case of *non-cell-autonomous RNAi* as the name depicts, the interfering effect takes place in tissues/cells different from the site of application or production of the dsRNA. There are two different kinds of non-cell-autonomous RNAi i.e., environmental RNAi and systemic RNAi. *Environmental RNAi* describes all processes in which dsRNA is taken up by a cell from the environment such as gut or haemocoel. This process can also be observed in

unicellular organisms. In *systemic RNAi*, silencing signal spreads to neighbouring cells from epicentre of cell. Systemic RNAi can only take place in multicellular organisms because it includes processes in which a silencing signal is transported from one cell to another or from one tissue type to another.

For the efficient application of RNAi in insect control, non-cell-autonomous RNAi has to be considered. For applied aspects, insect have to internalize the dsRNA of a target gene through feeding. In order to silence the target gene, this dsRNA must be taken up from the gut lumen into the gut cells, representing environmental RNAi. The insect midgut consists of a single layer of columnar cells with microvilli, endocrine cells, and stem cells at the base, grouped in the so-called nidi. The mid gut is designed to absorb nutrients from the gut lumen with its large absorption area created by the microvilli, with many channels and endocytosis device. These features make gut as a potential dsRNA uptake location. If the target gene is expressed in a tissue outside of the gut, the silencing signal will also have to spread via cells and tissues, which is systemic RNAi. In some insects *viz.*, light brown apple moth, *Epiphyas postvittana* , cricket *Gryllus bimaculatus*, Western Corn Rootworm *Diabrotica virgifera virgifera*, diamondback moth, *Plutella xylostella*, *Tribolium castaneum* and *Schistocerca gregaria* repoted that administration of dsRNA can result in the generation of RNAi response throughout the entire body (systemic RNAi). However, the mechanism of short distance as well as long distance intercellular transfer of systemic RNAi signal is still to elucidate.

Some reports indicated that the cellular uptake of dsRNA in insects occurs via scavenger receptor-mediated endocytosis both in cultured cells and in vivo (referred to as environmental RNAi). It is also known that the uptake of dsRNA depends on it's length. It occurs efficiently for long dsRNA molecules of around 200–500 base pairs (bp) and even up to 1000 bp. However, for shorter constructs such as siRNAs, this efficiency decreases. Moreover, it has been shown that lipophorins can adhere to dsRNA fragments in the insect hemolymph, suggesting a possible role of these proteins in either protection, transport, or both, throughout the body (Wynant *et al.*, 2014a). In addition, two main findings have been reported for *Drosophila melanogaster* i.e., viral infection of cultured Drosophila cells increased the formation of nanotube-like structures through which short-distance transport of dsRNA & RISC components can occur and flies use hemocyte-derived exosome-like vesicles to systemically spread an antiviral siRNA signal in the hemolymph.

Application of RNAi in Entomology

The research application of RNAi in entomology has elucidated the functions of several genes. RNAi has been used to study various mechanisms related to insect development reproduction, behaviour and other complex biosynthetic pathways. RNAi technique have been exploited for various species of orders like Coleoptera, Lepidoptera, Diptera, Hemiptera, Orthoptera, Blattodea and Hymenoptera (Vogel *et al.*, 2019). The silencing efficiency varied among stages of insect, even insect to insect with in the same order as well as among different orders. The significance and compilation of various categories of RNAi experiments in entomology are

detailed in many articles. RNAi also finds relevance in other aspects of insect science. It is quite beneficial in maintaining the beneficial insects and saving them from various parasites and pathogens. RNAi has also been useful in elucidating the importance of various immunological pathways and host-parasite relationships. However, the RNAi systemic spreading mechanism is not conserved across organisms, and its elucidation is an essential step in developing an efficient method to control agricultural pests by RNAi technology. The research on RNAi for the control of insect pests has made significant growth in recent years. The availability of genomic sequences of insects has further widened the horizons for the testing of this technology against various insect groups.

Implications of RNAi in insect pest management

The success of RNAi on target insects, is depend upon several factors such as the concentration of dsRNA, the nucleotide sequence, the length of the dsRNA fragment, mode of delivery and the life stage of the target insects. The midgut is the most attractive target for RNAi exploitation in insects and the midgut epithelial tissue is the primary site of exposure to dsRNA. The stability of dsRNA molecule after ingestion and the efficiency of the silencing process is determined by the gut pH and nucleases. Delivery of dsRNA is a major challenge in RNAi-based plant protection method. After identifying the target gene, choosing a convenient strategy to deliver the dsRNA into the insect body is very important. Microinjection is a good strategy for functional genomic studies but this method is not suitable to control insect pests in the field. Sprayable RNAi-based products are in the process of development and are expected to be on the market shortly. These products can be divided into the four categories namely direct control agents, resistance repressors, developmental disruptors, and growth enhancers. Spraying the dsRNA might be useful to control some pest population in the field, but not all like piercing-sucking pests feeding on phloem sap, or stem borer pests feeding in the plant stems. So, Delivery of dsRNA to sucking pests and borer pests could be achieved by supplying dsRNA through *root absorption or trunk injection* into plant vessels, where these insects can naturally acquire dsRNA through sucking or chewing. Bedsides transgenic planta have been made to express dsRNA, cautiously selected to silence essential genes in target pest.

More than a decade back Baum *et al.* has given a proof of concept by developing transgenic corn crop to express dsRNA against the V-ATPase A transcript of the Western corn rootworm (WCR) *Diabrotica virgifera virgifera*. Feeding of WCR on this modified plant resulted in larval stunting and in the premature death of the insect. Moreover, dsRNA act as a crop protectant as feeding damage to the transgenic corn was greatly reduced (Baum *et al.*, 2007). Similarly Mao *et al.* (2007) showed that plant-mediated expression of dsRNA targeting the cytochrome P450 monooxygenase gene (CYP6AE14) could increase the toxic effects of gossypol, a cotton metabolite that is otherwise tolerated by the cotton bollworm. Silencing of CYP6AE14 led to delayed larval growth when gossypol was supplemented in the diet. These studies are good example of the application potential of dsRNA-mediated plant protection. Although, dsRNA-expressing transgenic plants reduce the crop damage, but effective killing of the pest population has not yet been attained. As RNAi is a budding technology within the

field of agriculture, so it will take certain time to make it's place. However, first RNAi-based insecticides for the control of WCR have already been approved by the United States Environmental Protection Agency (EPA). Monsanto and Dow Agro sciences developed a RNAi insecticide known as SmartStax ProR. This plant-incorporated protectant (PIP) will employ a pyramid strategy: several different Bt-proteins, as well as dsRNA targeting the WCRSnf7 gene, will be expressed in this plant (Head *et al.*, 2017). This combined strategy is designed to lead to the instant death of the insect, while also reducing the chances that insects will develop resistance against this PIP (Head *et al.*, 2017). Similarly, Ni *et al.* (2017) reported pyramiding RNAi and Bt counters insect resistance to Bt crops. They developed two types of transgenic cotton plants producing dsRNA from the *Helicoverpa armigera* designed to interfere with its metabolism of juvenile hormone (JH). They focused on suppression of JH acid methyltransferase (JHAMT), which is crucial for JH synthesis, and JH-binding protein (JHBP), which transports JH to organs. Both types of RNAi cotton were effective against Bt-resistant insects. Bt cotton and RNAi acted independently against the susceptible strain.

As stringent and lengthy regulatory rules from protection agencies may hinder transgenic crop releases, so it's better to go for non-transformative RNAi strategies. Limitations of inadequate RNAi efficiency must be solved by protecting dsRNA by packaging and through the use of delivery systems such as bacterial, viruses, nanoparticles, liposomes etc. A list of various delivery systems presented in Table 1 (Vogel *et al.*, 2019).

Bacteria and Viruses

The delivery of dsRNA using bacteria has many advantages when compared with plant-mediated dsRNA delivery or *in vitro* synthesized dsRNA delivery. The application of bacteria-expressed dsRNA is cost effective when compared with *in vitro* synthesized dsRNA. Moreover, large-scale production of bacteria, which express dsRNA for use as pesticide, could turn into a reality soon. Persistent and large-scale delivery of dsRNA is required to kill an insect pest and also to reduce the resistance development. The bacteria-expressed dsRNA pesticides can be sprayed on crops at any time because of the ease of producing large amounts of bacteria-expressing dsRNAs. Initially, the recombinant *Escherichia coli* was engineered for dsRNAs production. Recently, the use of symbiontic bacteria has been shown to be a promising delivery strategy as well. Symbiont-mediated RNAi is an interesting strategy in which the relationship between culturable symbiotic gut bacteria and the hosts can be exploited in order to constitutively produce dsRNA to induce RNAi in the host. The symbiont-mediated RNAi is a versatile technology to study the gene function and also a bio pesticide to control the pest population.

Recently, plant viruses have been investigated as tools to trigger RNAi in plants. Generally, plants respond to infections caused by viruses through the siRNAi pathway (vsRNAs). Therefore, if an insect-specific RNAi inducer sequence is introduced into an engineered plant virus, this will produce siRNAs in the plant that are specific for insect targets. The RNAi effect can be induced when insects feed on plants containing engineered virus to produced specific

siRNAs. All plant-infecting viruses move inside the plant systematically through the phloem. For that reason, the recombinant plant viruses might target the phloem-feeding insect pests.

Nanoparticle-mediated RNAi

Nanoparticles are polyplex-based delivery systems, consisting of either natural or synthetic polymer subunits. In order to increase stability and uptake efficiency, dsRNA can also be incorporated into a nanoparticle. The most utilized nanoparticles are chitosan-derived. Chitosan is a non-toxic, biodegradable molecule that can be obtained by deacetylation of chitin, one of the most abundant biopolymers in nature that is particularly known for its structural function in the exoskeleton of arthropods. dsRNAs were entrapped by the polymer chitosan with electrostatic forces to form a chitosan/dsRNA nanoparticle, which were delivered into the insect by ingestion. Chitosan nanoparticles are designed by self-accumulation of polycations with dsRNA via electrostatic forces among positive and negative charges of the amino groups in the chitosan and phosphate groups on the backbone of the nucleic acid, respectively. This method is suitable with long dsRNA and siRNA.

TABLE 1 | Overview of delivery systems used for the successful delivery of dsRNA in several economically important insect orders.

Insect order	Category	Delivery system	Species	Target gene[a]	Reference
Lepidoptera	Micro-organism	Bacteria	*Spodoptera exigua*	*Chitin synthase A (SeCHSA)*	Tian et al., 2009
			Spodoptera exigua	*Chymotrypsin 2 (SeCHY2)*	Vatanparast and Kim, 2017
			Helicoverpa armigera	*Ultraspiracle protein (USP)*	Yang and Han, 2014
			Sesamia nonagrioides	*Juvenile hormone esterase (SnJHE)*	Kontogiannatos et al., 2013
	Viral	BmNPV	*Sesamia nonagrioides*	*Juvenile hormone esterase (SnJHE)*	Kontogiannatos et al., 2013
		AcMNPV	*Heliothis virescens*	*Juvenile hormone esterase (HvJHE)*	Hajós et al., 1999
		Sindbis Virus	*Bombyx mori*	*Broad-Complex (Br-C)*	Uhlirova et al., 2003
	Nanoparticle	FNP	*Ostrinia furnacalis*	*Chitinase-like gene CHT10*	He et al., 2013
		Guanylated polymers	*Spodoptera exigua*	*Chitin synthase B*	Christiaens et al., 2018
Coleoptera	Micro-organism	Bacteria	*Leptinotarsa decemlineata*	*β-actin (actin), Protein transport protein sec23 (Sec23), Coatomer subunit beta (COPβ)*	Zhu et al., 2011
	Proteinaceous	PTD-DRBD	*Anthonomus grandis*	*Chitin synthase II (AgChSII)*	Gillet et al., 2017
Hemiptera	Micro-organism	Bacterial symbiont – *Rhodococcus rhodnii*	*Rhodnius prolixus*	*Nitrophin 1 (NP1), Nitrophin 2 (NP2), Vitellogenin (Vg)*	Whitten et al., 2016
		Bacterial symbiont – *BFo2*	*Frankliniella occidentalis*	*α-Tubulin (Tub)*	Whitten et al., 2016
Diptera	Micro-organism	Yeast symbiont – *Saccharomyces cerevisiae*	*Drosophila suzukii*	*γ-Tubulin 23C (γTub23C))*	Murphy et al., 2016
		Chlamydomonas reinhardtii	*Anopheles stephensi*	*3-hydroxykynurenine transaminase (3-HKT)*	Kumar et al., 2013
		Pichia pastoris	*Aedes aegypti*	*Juvenile hormone acid methyl transferase (AeaJHAMT)*	Van Ekert et al., 2014
	Nanoparticles	*Chitosan*	*Aedes aegypti*	*Semaphorin-1a (sema1a)*	Mysore et al., 2013
			Aedes aegypti	*Single-minded (Sim)*	Mysore et al., 2014
			Aedes aegypti	*Vestigial gene (vg)*	Kumar et al., 2016
			Anopheles gambiae	*Chitin synthase 1 (AgCHS1), Chitin synthase 2 (AgCHS2)*	Zhang et al., 2010
	Liposomes	Lipofectamine 2000, Cellfectin, Transfectin, BMRIE-C	*Drosophila melanogaster*	*γ-Tubulin (γ-Tub)*	Whyard et al., 2009
			Drosophila sechellia	*γ-Tubulin (γ-Tub)*	Whyard et al., 2009
			Drosophila yakuba	*γ-Tubulin (γ-Tub)*	Whyard et al., 2009
			Drosophila pseudoobscura	*γ-Tubulin (γ-Tub)*	Whyard et al., 2009
		Lipofectamine 2000	*Drosophila suzukii*	*Alpha-coatomer protein (alpha COP), Ribosomal protein S13 (RPS13), Vacuolar H[+]-ATPase E subunit (Vha26)*	Taning et al., 2016
		Effectene	*Aedes aegypti*	*Inositol-requiring enzyme 1 (Ire-1), X-box binding protein-1 (Xbp-1), Caspase-1 (Cas-1), SREBP cleavage-activating protein (Scap), site-2 protease (S2P)*	Bedoya-Pérez et al., 2013
			Aedes aegypti	*Mitogen-activated protein kinase p38*	Cancino-Rodezno et al., 2010

[a] *All target genes listed gave a more efficient knockdown than similar treatment with naked dsRNA.*

Liposomes

Liposomes are composed of natural lipids and they are non-toxic and easily biodegradable. Liposomes form naturally when transfection agents are brought into an aqueous environment. During this process, the positively charged lipids envelop the negatively charged nucleic acid material, forming compact lipid bilayer particles similar to the phospholipid bilayer of the cell membrane. Cell entry of the liposome-encapsulated dsRNA is then achieved through lipofection.

Delivery of dsRNA using microorganism (bacteria and viruses), polymer nanoparticles, liposomes could increase efficacy in attaining a potent RNAi response. The choice of the delivery method of course depend on the conditions, on the target insect and on the reason for impaired RNAi-efficiency. For example, liposomes and polymers could be used where a limited cellular uptake is causing the insect to be irresponsive to RNAi. When stability of dsRNA in the insect body is the main issue, polymer- or liposome nanoparticles and bacteria could be used. Insect virus-mediated delivery could be a solution for cellular uptake, degradation and in cases where the insect is difficult to reach, since the dsRNA would be immediately produced inside the insect cells infected with the viruses.

Future Perspectives

RNAi has an immense potential to become a successful approach for insect pest management. However, several research and ethical issues need to be addressed before this technology can be applied on a commercial level. RNAi technology has been extended to a large number of insect species from various orders. Many differences in components and mechanisms among insect orders and between insects and other organisms still need to be worked out. Progress made so far in RNAi technology provides an ample evidence for RNAi to become a successful alternate for insect pest management. As far as research is concerned the production of dsRNA can be achieved by using molecular biology kits available commercially. However, the cost/µg of dsRNA synthesis is commercially not viable. Since large amounts of dsRNA are required, more cost-efficient methods for mass production has to be developed. Bacteria expressed dsRNA is considered as one of the most cost-effective methods, and biotech companies are investing in this production method to produce large quantities. Several agricultural companies are working toward improvements of low-cost production of ready-to-spray RNAi products, rather than to create genetically modified crops that cost millions and take years. Moreover, many regulatory hurdles from governmental agencies as well as non-acceptance of public of these genetically modified crops. Exploration of RNAi to use in insect pest management has taken a leap, and one can easily believe that it has a potential to become the powerful pest management tool. Despite various doubts associated with this technology, the days are not too far when one would see RNAi would stand alongside Bt technology in insect pest management programs.

**Reproduced from Vogel et al., 2019

References

Baum J. A., Bogaert T., Clinton W., Heck G. R., Feldmann P., Ilagan O., *et al.* (2007). Control of coleopteran insect pests through RNA interference. Nat. Biotechnol. 25, 1322–1326.

Bolognesi R., Ramaseshadri P., Anderson J., Bachman P., Clinton W., Flannagan R., *et al*. (2012). Characterizing the mechanism of action of double-stranded RNA activity against western corn rootworm (Diabrotica virgifera virgifera LeConte). PLoS ONE 7:e47534.

Head, G. P., Carroll, M. W., Evans, S. P., Rule, D. M., Willse, A. R., Clark, T. L., *et al.* (2017). Evaluation of SmartStax and SmartStax PROmaize against western corn rootworm and northern corn rootworm: efficacy and resistance management. Pest Manag. Sci. 73, 1883–1899. doi: 10.1002/ps.4554.

Huvenne, H., & Smagghe, G. (2010). Mechanism of dsRNA uptake in insects and potential of RNAi in pest control: a review. Journal of Insect Physiology, 56, 227–235.

Katoch, R, Amit Sethi, Thakur, N., Murdock, L. L. (2013). RNAi for insect control : Current perspectives and future challenges. Applied Biochemistry Biotechnology 171, 847-873.

Kolliopoulou A, Swevers L. (2014). Recent progress in RNAi research in Lepidoptera: intracellular machinery antiviral immune response and prospects for insect pest control. Curr Opin Insect Sci. 6:28–34. DOI: http://dx.doi.org/10.1016/j.cois.2014.09.019

Mao, Y. B., Cai,W. J., Wang, J. W., Hong, G. J., Tao, X. Y., Wang, L. J., *et al.* (2007). Silencing a cotton bollworm P450 monooxygenase gene by plant-mediated RNAi impairs larval tolerance of gossypol. Nature Biotechnology, 25, 1307–1313.

Ni M, Ma W, Wang X, Gao M, Dai Y, Wei X, *et al.* (2017). Next-generation transgenic cotton: pyramiding RNAi and *Bt* counters insect resistance. Plant Biotechnology Journal, 15, 1–10.

Vogel E, Santos D, Mingels L, Verdonckt T-W and Broeck JV (2019) RNA Interference in Insects:Protecting Beneficials and Controlling Pests. Front. Physiol. 9:1912. doi: 10.3389/fphys.2018.01912

Chapter-26

The Genomic Basis of Nematode Parasitology: as a Host & a Pathogen

Uma Rao*, Victor Phani and Vishal S. Somvanshi
Division of Nematology, ICAR-Indian Agricultural Research Institute, New Delhi, India
**Email: umarao@iari.res.in, umanema@gmail.com*

Globally plant parasitic nematodes (PPNs) are responsible for considerable yield losses amounting to an estimated $157 billion and national losses due to plant parasitic nematodes is estimated as Rs. 21 billion annually. The Present management practices like field sanitation, host resistance, biological control, chemical control etc. are not adequate. Each of these approaches has their own advantages and disadvantages. Recognizing the limitations of current nematode management practices, there is a pressing need to develop environmentally suitable and sustainable new generation management approaches tailor-made for controlling various PPNs, particularly for the most damaging species of root knot and cyst nematodes. One such approach is by using the genomic information of an organism for exploring genes involved in vital pathways of nematode life and disease cycles that could serve as potential targets for designing transgenic plants with required field tolerance or development of novel nematicidal molecule / transgenic crops.

Parasitism is a natural phenomenon, where an organism (parasite) lives in or on another organism (host) and obtains nourishment. Being a diversified animal, nematodes provide apt opportunity to study host-parasite interaction with flexibility of using them from both ends. Over few decades, knowledge of nematode-host interplay increased exponentially with better clarification using genomic approaches. Recent efforts in application of Genome sequencing, transcriptomics, proteomics, metabolomics and epigenetics in both free living nematodes- *Caenorhabditis elegans, C. briggsae* and several plant, animal and human parasitic nematodes has helped unraveling the nematode biology, initiation and progression of parasitism, and host immune or defense responses. This has resulted in availability of large amount of genomics data that made it possible to identify genes controlling several critical pathways / functions etc via comparative genomics. The in-depth mastery helped us to unfold the parasitism gene function(s), immune evasion mechanisms, identification of drug targets, novel drug designing, drug resistance and epidemiological intricacies. Genomic studies involving animal- and human-parasitic nematodes are dealt under "Helminthology" and "Nematologists" are predominantly engaged with plant-parasitic (PPNs) and free-living nematodes.

Genomic data of nematodes is very useful for deciphering the gene function using both by *in silico* analysis and also experimental procedures. One such experimental strategy is by employing RNA interference. With the recent advent of gene expression control via small interfering RNA (siRNA) and micro RNA (miRNA) molecules, RNAi based transgenics is

becoming the trend to suppress the menace of plant parasitic nematodes (PPNs). Induction of RNAi by delivering double-stranded RNA (dsRNA) has been very successful in the model non-parasitic nematode, *C. elegans*, while in PPNs, dsRNA delivery was accomplished by soaking the nematodes with dsRNA solution mixed with the neurotransmitters like resorcinol, octopamine, serotonin etc. Using *in vitro* dsRNA delivery approaches, down regulation of various housekeeping genes led to reduced parasitic ability, delayed egg hatching, impaired motility, and ability to locate and invade roots, demonstrated in root-knot, cyst, lesion, pine wilt and burrowing nematodes. The success of the *in vitro* dsRNA ingestion and down-regulation of the target genes inspired the *in planta* delivery of dsRNA to the feeding nematodes. The most convincing success of *in planta* delivery of dsRNA to feeding nematodes resulting in knocking down of several key nematode genes came from root-knot nematodes. Hence limitations of existing nematode management practices could be overcome by employing RNAi based approach for curbing population build up. Peptide based transgenics produce functional proteins which could have off target effects on non target organisms but RNAi based transgenics is superior to that as it does not produce any functional proteins and targets organism in sequence specific manner. Although RNAi based transgenics are still in preliminary stage but it offers novel management strategy for the future.

In this endeavor, our laboratory has identified and functionally validated using RNAi several gene targets in *M. incognita, M. graminicola* and *Heterodera avenae* for their role in infection, development and reproduction. This ongoing work in our laboratory have shown that several novel parasitism genes in PPNs help in host recognition (*odr*s), invasion through root tissues (hydrolases and effectors), development and maintenance of feeding sites (esophageal secretion-genes) etc. Additionally, housekeeping genes of neuropeptidergic function (*flp*s, *nlp*s, cholinergic-neuropeptides) and lipid metabolism (*sbp*) have also been identified in root-knot and cyst nematodes; and few of them have been successfully employed for management via transgenic approaches. Furthermore, studies involving *C. elegans* and its bacterial and fungal parasites helped to understand the microbial pathogenesis process. The transcriptomic analyses of root-knot nematode, *M. incognita* hyperparasitized by *Pasteuria*, helped to identify bacterial adhesion regulatory genes (*far*, *SeBP*, mucin) and its pathogenic progression. This knowledge can be utilized to identify the nematode kill-switches used by microbes for drug designing, and can be extrapolated to understand microbial pathogenesis in other organisms, including humans. All the 'omics' technologies are nowadays comfortably accessible furnishing enormous data. We therefore need to think carefully about how the resources will enhance our perception of parasitism to develop new therapeutic tools.

Chapter-27

Plant-Parasitic Nematode Genomics: An Update

Vishal Singh Somvanshi and Uma Rao
Division of Nematology, ICAR-Indian Agricultural Research Institute, New Delhi 110012
Email: vssomvanshi@iari.res.in; umarao@iari.res.in

Nematodes are one of the largest groups of metazoans. They are ubiquitous in nature and are found in almost all possible ecosystems. It is estimated that 1 – 10 million species of nematodes could be present in the environment, but only~27000 species have been described. Although popular as animal, humanor plant parasites, majority of nematodes are beneficial for the environment and play critical role in nutrient recycling. The life cycle of nematodescomprise of an embryonicstage, followed by four larval and adult stages. In some nematode species, an alternative 3rd larval stage known as dauer stage, which isable of survive adverse environmental conditions is also found. In parasitic nematodes, some or all stages can be parasitic. The life span of nematodes may vary greatly from a few days to several years. Phylum Nematoda has traditionally been classified into two classes- Adenophoreaand Secernentea. However, the molecular analysis of phylogenetic evolution in Phylum Nematoda revealed that Adenophorea was ancestral, and Secernentea arose from Adenophorea. As per the new analysis, Phylum Nematoda has been classified into three major lineages, Chromadoria, Enoplia and Dorylaimia, distributed into 5 major clades. It was also suggested that animal- and plant-parasitism evolved at least four and three different times, respectively, during the course of nematode evolution.

Plant-parasitic nematodes (PPNs) cause an estimated annual global crop yield loss of about 10%, amounting to US $173 billion.The globally identified top ten most dangerous plant-parasitic nematode are:(1) root-knot nematodes (*Meloidogyne* spp.); (2) cyst nematodes (*Heterodera*and *Globodera* spp.); (3) root lesion nematodes (*Pratylenchus* spp.); (4) the burrowing nematode *Radopholussimilis*; (5) *Ditylenchusdipsaci*; (6) the pine wilt nematode *Bursaphelenchusxylophilus*; (7) the reniform nematode *Rotylenchulusreniformis*; (8) *Xiphinema index* (the only virus vector nematode to make the list); (9) *Nacobbusaberrans*; and (10) *Aphelenchoidesbesseyi*.In India, a recent estimate indicated that plant-parasitic nematodes cause an approximate yield loss of ₹102,039.79 million (US$1577 million) to various crops. The top plant nematode threats to Indian agriculture are *Meloidogyne* spp., *Rotylenchulusreniformis*, *Globodera* and *Heterodera*spp., *Pratylenchus*spp., *Aphelenchoidesbesseyi*, *Ditylenchusangustus*, *Radopholussimilis*, *Tylenchulussemipenetrans*, *Hirschmanniella* spp., *Helicotylenchusmulticinctus*, and *Tylenchorhynchusbrevilineatus*. Also, two extremely dangerous nematodes not present in India are pinewood nematode (*Bursaphelenchusxylophilus*) and red ring nematode of coconut (*B. cocophilus*).

Despite the fact that PPNs are an enormous threat to the crops, there is an acute scarcity of management options for these parasites. Most of the chemical pesticides used for PPN

management are general biocides or insecticides which have either been banned or are being phased out because of their ill effects on the environment. At present, only four chemicals, fluopyram, fluensulfone, fluazaindolizine and tioxazafenare registered as 'nematicides' for the management of PPN. The situation is not better for animal/human parasitic nematodes, where only two classes of drugs, i.e. benzimidazoles (BZ), and nicotinic acetylcholine receptor (nAChR) agonists are being used to fight the nematode menace. Hence, there is an urgent need for the discovery of newer options and strategies for nematode management. However, the discovery of a new chemical molecule and developing it into a ready-to-use finished product is a highly expensive and time-consuming process. Genomics provide a viable option for development of new nematode management options.

Thegenome of nematode *Caenorhabditis elegans* was sequenced in 1998. It is the only nematode for which the genome sequence is complete. Since then genomes of 131 nematodes have been sequenced and published on WormBase, the primary nematode genome repository. In 2019 the *C. elegans* genome was re-sequenced using long read sequencing approach and newer genes were discovered. Nematode genomes show several interesting features. Nematode genomes are compact andrange from 19.67 Mb to 265 Mb in size. In spite of such a major size difference, the number of genes in Nematodes is almost comparable to that of humans. On an average, the ematode genomes also show high gene density. The nematodesshow a large number of genes for which no homologues could be recognized outside the same nematode genus, indicating high rates of gene gain and loss through horizontal gene transfer. In addition, the nematode genes are organised in operons. Comparative genomic analysis has revealed that nematode genomes do not show any genomic featurewhich could be used to differentiate parasitic nematodes from free-living nematodes. Several novel genes are also known to arise in Phylum Nematoda.

A list of all the sequenced plant parasitic nematodes, the technologies used for sequencing and their genome statistics is provided in Table 1.

Table 1: A comparison of publishedplant-parasitic nematode genomes (Modifoed from Somvanshi *et al.*, 2018)

S. No.	Nematode	Assembly Size	No. of Scaffolds	N50 value (kbp)	CEGMA score (Complete/ Partial %) or Complete%	Year
1	*Meloidogyne incognita*	86.1	2995	62.5	77/80.6	2008
2	*Meloidogyne hapla*	53.0	3452	37.6	94.8/96.8	2008
3	*Bursaphelenchusxylophilus*	74.6	5527	949.8	97.6/98.4	2011
4	*Meloidogyne floridensis*	96.7	58696	3.7	58.1/77.4	2014
5	*Globoderapallida*	124.6	6873	122	74.19/80.65	2014
6	*Pratylenchuscoffeae*	19.7	5821	10	NA	2015
7	*Globoderarostochiensis*	95.9	4377	89	93.55/95.56	2016
8	*Meloidogyne enterolobii*	162.4	46090	9.2	81 /NA	2017

	L30					
9	*Meloidogyne floridensis* SJF1	74.9	9134	13.2	84	2017
10	*Meloidogyne incognita* W1	122.1	33735	16.4	83	2017
11	*Globoderaellingtonae*	119.1	2248	360	92/96	2017
12	*Meloidogyne javanica*VW4	142.6	34394	14.2	90	2017
13	*Meloidogyne arenaria* HarA	163.7	46509	10.5	91	2017
14	*Ditylenchus destructor*	112	1761	570	91	2016
15	*Meloidogyne incognita*	183.5	12091	38.6	97	2017
16	*Meloidogyne javanica*	256.3	31341	10.4	96	2017
17	*Meloidogyne arenaria*	235.5	26196	16.4	95	2017
18	*Meloidogyne graminicola*	38.18	4304	20.4	84.27/90.73	2018

Recent advancements in PPN genomics and transcriptomics has made it possible to identify promising molecular targets and metabolic choke points that may be exploited for their management through transgenic approaches or target-based drug discovery.Although new drugs or chemicals designed using genomic information are still awaited, the transgenic crop-based approaches, such as host delivered RNAi has been very successfully used for nematode management.It is expected that increased use of genomic information would give a major boost to find smarter options for animal- and plant-parasitic nematode managemen.

References

Abad, P., Gouzy, J., Aury, J.M., Castagnone-Sereno, P., Danchin, E.G., Deleury, E., Perfus-Barbeoch, L., Anthouard, V., Artiguenave, F., Blok, V.C., Caillaud, M.C., Coutinho, P.M., Dasilva, C., De Luca, F., Deau, F., Esquibet, M., Flutre, T., Goldstone, J.V., Hamamouch, N., Hewezi, T., Jaillon, O., Jubin, C., Leonetti, P., Magliano, M., Maier T.R., Markov, G.V., *et al.* 2008. Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. Nature Biotechnology 26:909-915.

Dieterich, C., Sommer, R.J. 2009. How to become a parasite: lessons from the genomes of nematodes.Trends in Genetics. 25:203–9

Kikuchi, T., Eves-van den Akker, S., and Jones, J. T. 2017. Genome evolution of plant-parasitic nematodes. Annual Review of Phytopathology 55:333-354.

Lunt, D.H., Kumar, S., Koutsovoulos, G., and Blaxter, M.L. 2014. The complex hybrid origins of the root knot nematodes revealed through comparative genomics. PeerJ 2:e356.

Mitreva, M., Smant, G.,Helder, J. 2009. Role of horizontal gene transfer in the evolution of plant parasitism among nematodes. Methods in Molecular Biology. 532:517–35

Opperman, C.H., Bird, D.M., Williamson, V.M., Rokhsar, D.S., Burke, M., Cohn, J., Cromer, J., Diener, S., Gajan, J., Graham, S., Houfek, T.D., Liu, Q., Mitros, T., Schaff, J., Schaffer, R., Scholl, E., Sosinski, B.R., Thomas, V.P., and Windham, E. 2008. Sequence and genetic map of *Meloidogyne hapla*: A compact nematode genome for plant parasitism. Proceedings of the National Academy of Sciences of the United States of America 105:14802-14807.

Sommer, R.J. and Streit, A. 2011. Comparative Genetics and Genomics of Nematodes:Genome Structure, Development, and Lifestyle. Annual Review of Genetics. 45:1–20

Chapter-28

Transcriptomic Approaches for Elucidating the Genes Network Associated with Source and Sink in Wheat

Ranjeet R. Kumar and Shelly Praveen
Division of Biochemistry, Indian Agricultural Research Institute, New Delhi-110 012
ranjeetranjaniari@gmail.com

Plant omics and new biotechnologies such as massively parallel sequencing and microarray analysis were preferred tools to identify and characterize the differentially expressed genes (DEGs) under stress in different modal species (Gong *et al.* 2014), but Next-Generation Sequencing (NGS) is now the most preferred technique used for transcriptome study. RNA sequencing (RNA-Seq) is revolutionizing the study of the transcriptome. A highly sensitive and accurate tool for measuring expression across the transcriptome, it is providing visibility to previously undetected changes occurring in disease states, in response to therapeutics, under different environmental conditions and across a broad range of other study designs. RNA-Seq allows researchers to detect both known and novel features in a single assay, enabling the detection of transcript isoforms, gene fusions, single nucleotide variants, allele-specific gene expression and other features without the limitation of prior knowledge. The transcriptome data generated through NGS provide very useful information on the regulatory pathway networks operating in plants under different conditions (Rensink and Buell, 2005). NGS data help to identify DEGs, along with their functional annotation under different stresses, which has been used in the past to elucidate the different mechanisms as well as pathways associated with stress-tolerance (Mochida *et al.*, 2006). RNA-Sequencing (RNA-Seq) is now the preferred technology used for global genome identification and de novo transcriptome of various organisms (Mousavi *et al.*, 2014). In this technology, small stretches of cDNAs are sequenced at a very high coverage and assembled using different programs to reconstruct the contigs (Figure 1). It has been successfully used for the relative expression study, editing 5' and 3' ends of annotated genes and functional gene identification with their respective exons and introns (Nagalakshmi *et al.*, 2010). In agriculturally important crops, RNA-Seq has been mostly used for the identification of novel and conserved stress-responsive and pathway-associated genes involve in tolerance and nutrient responsive regulation (Kugler *et al.*, 2013). Other than this, RNA-Seq offers numerous advantages over gene expression arrays like more sensitive and accurate measurement of gene expression, helps in identification of both known and novel genes, can be used in any crop plant species. Various crops such as chickpea (Molina *et al.*, 2011), rice (Mizuno *et al.*, 2010), sorghum (Dugas *et al.*, 2011), soybean (Hao *et al.*, 2011), and parsley (Li *et al.*, 2014) have been subjected to high-throughput NGS for identification of abiotic stress-related genes. RNA-Seq is virtually changing the area of sequencing and transcriptome profiling (Wang *et al.*, 2009). Recently, de novo assembly has been used in some of the agriculturally important crops like chilli pepper (Liu *et al.*, 2013),

coconut (Fan *et al.*, 2013), rice (Zhang *et al.*, 2013), and sugarcane (Cardoso-Silva *et al.*, 2014) for identification of novel genes. Earlier, Roche GS-FLX 454 was the most widely used platform for de novo transcriptome sequencing, due to its long read length in different organisms, for example, ginseng (Sun *et al.*, 2010), A. thaliana (Wall *et al.*, 2009), and maize (Vega-Arreguin *et al.*, 2009). The Illumina transcriptome was mainly used for the sequencing of organisms whose genome was sequenced (Li *et al.*, 2010). It was confirmed in due course of time that the relatively short reads can be effectively assembled by the Illumina transcriptome, or whole genome de novo sequencing and assembly with the advantage of paired-end sequencing (Maher *et al.*, 2009).

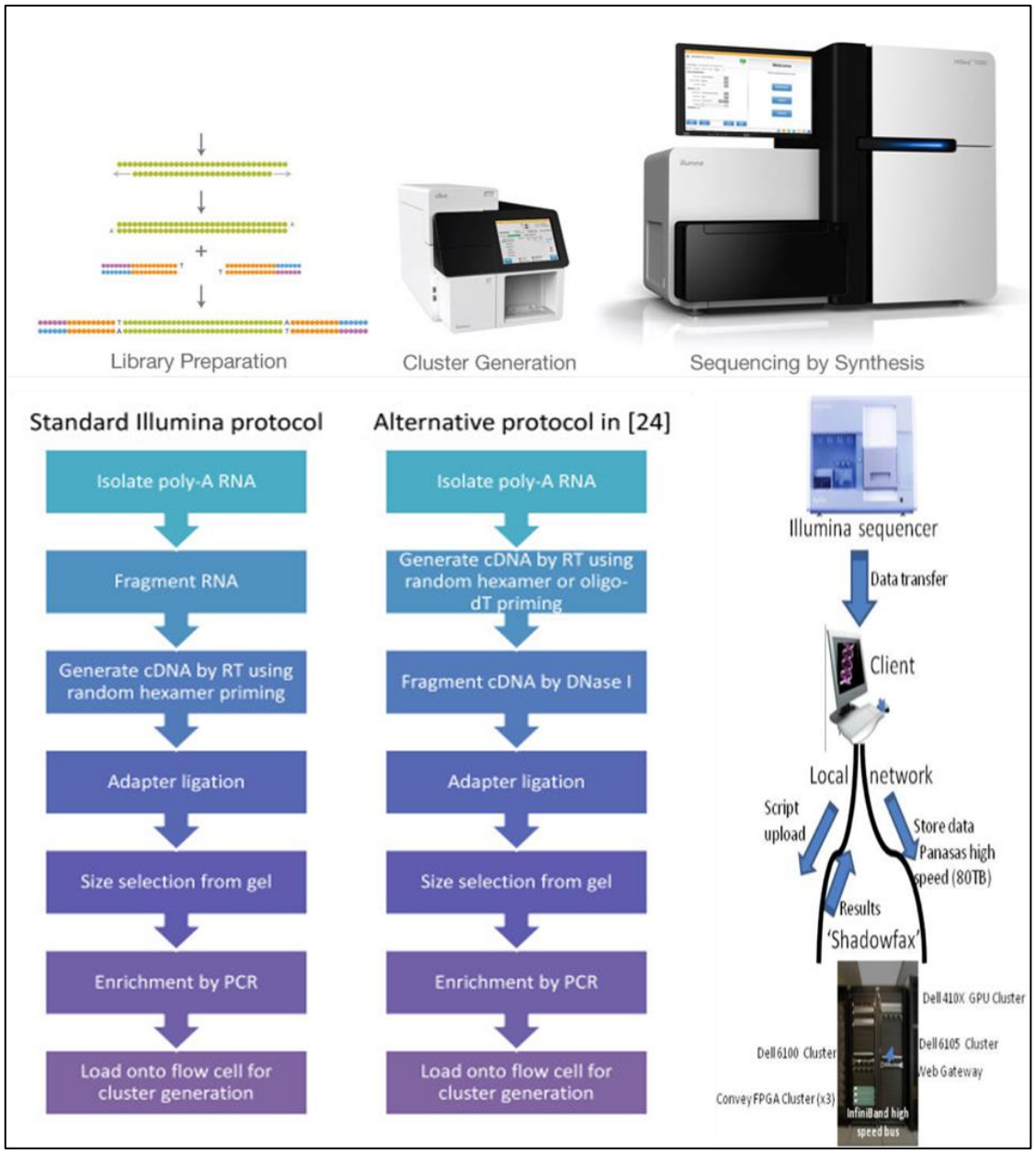


Figure 1: Schematic representation of protocol used for the Next-Generation Sequencing using Illumina HiSeq.

Starch

Starch is the most significant form of carbon reserve in plantsin terms of the amount made, the universality of its distributionamong different plant species, and its commercial importance. It consists of different glucose polymers arranged into a threedimensional,semi-crystalline structure-the starch granule. Thebiosynthesis of starch involves not only the production of thecomposite glucans but also their arrangement into an organizedform within the starch granule. The formation of the starchgranule can be viewed as a simple model for the formationof ordered three-dimensional polysaccharide structuresin plants. Understanding the biochemical basis for the assemblyof the granule could provide a conceptual basis forunderstanding other higher order biosynthetic systems suchas cellulose biosynthesis. For example, one emerging concept is that structurewithin the granule itself may determine or influence the wayin which starch polymers are synthesized. Starch is synthesized in leaves during the day from photosynthetically fixed carbon and is mobilized at night. It is alsosynthesized transiently in other organs, such as meristemsand root cap cells, but its major site of accumulation is in storageorgans, including seeds, fruits, tubers, and storage roots. Almost all structural studies have used starch from storageorgans because it is readily available and commercially important;we therefore focus on starch biosynthesis in storageorgans. However, where aspects of transient biosynthesis areclearly different from long-term reserve synthesis, referenceis made to biosynthesis in non-storage tissues. Starch is synthesized in plastids, which in storage organscommitted primarily to starch production are called amyloplasts (Figure 2). These develop directly from proplastids and have littleinterna1 lamellar structure. Starch may also be synthesized inplastids that have other specialized functions, such as chloroplasts (photosynthetic carbon fixation), plastids of oilseed (fatty acid biosynthesis), and chromoplasts of roots such ascarrot (carotenoid biosynthesis).

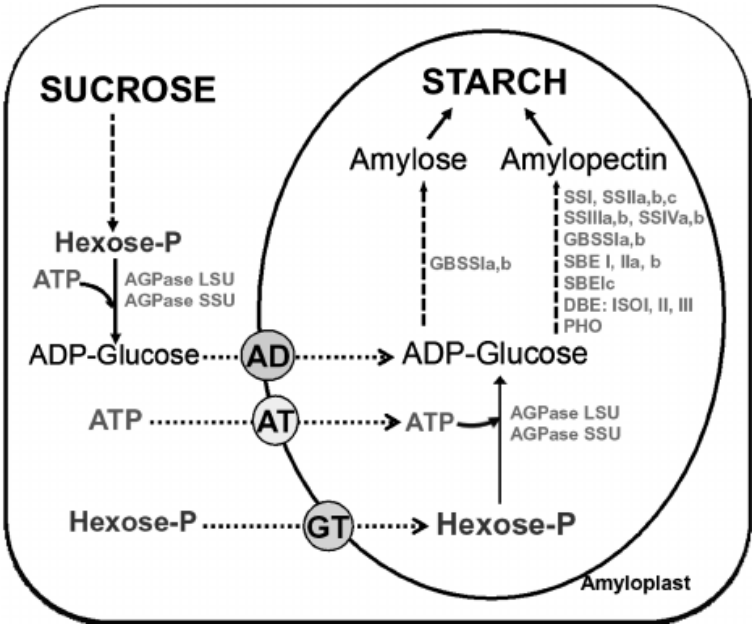


Figure 2: Schematic representation of starch biosynthesis pathway operating in plants

The Biochemistry of Starch Biosynthesis

The biosynthetic steps required for starch biosynthesis arerelatively simple, involving three committed enzymes: ADPglucosepyrophosphorylase (ADPGPPase; EC 2.7.7.23), starchsynthase (SS; EC 2.4.1.21), and starch branching enzyme (SBE;EC 2.4.1.28). Amylose and amylopectin are synthesized from ADPglucose, which is synthesized from glucose-1-phosphate andATP in a reaction that is catalyzed by ADPGPPase and thatliberates pyrophosphate. This enzyme is active within theplastid, which means that its substrates, glucose-1-phosphateand ATP, must also be present in the plastid. In chloroplasts, ATP may be derived from photosynthesis, but in non-photosyntheticplastids, it must be specifically imported from the cytosol, probably by an ADP/ATP translocator. The glucose-1-phosphate canbe supplied by the reductive pentose phosphate pathway inchloroplasts via phosphoglucoisomerase and phosphoglucomutase. In non-photosynthetictissues, it may be imported directly from the cytosol or synthesized in the plastid from glucose6-phosphatevia the action of a plastidialphosphoglucomutase. The pyrophosphate produced by ADPGPPase is removedby inorganic alkaline pyrophosphatase, which is probably confinedto plastids in both photosynthetic and nonphotosynthetictissues. The removal of this plastidial pyrophosphate effectivelydisplaces the equilibrium of the ADPGPPase reactionin favor of ADPglucose synthesis (Figure 3).

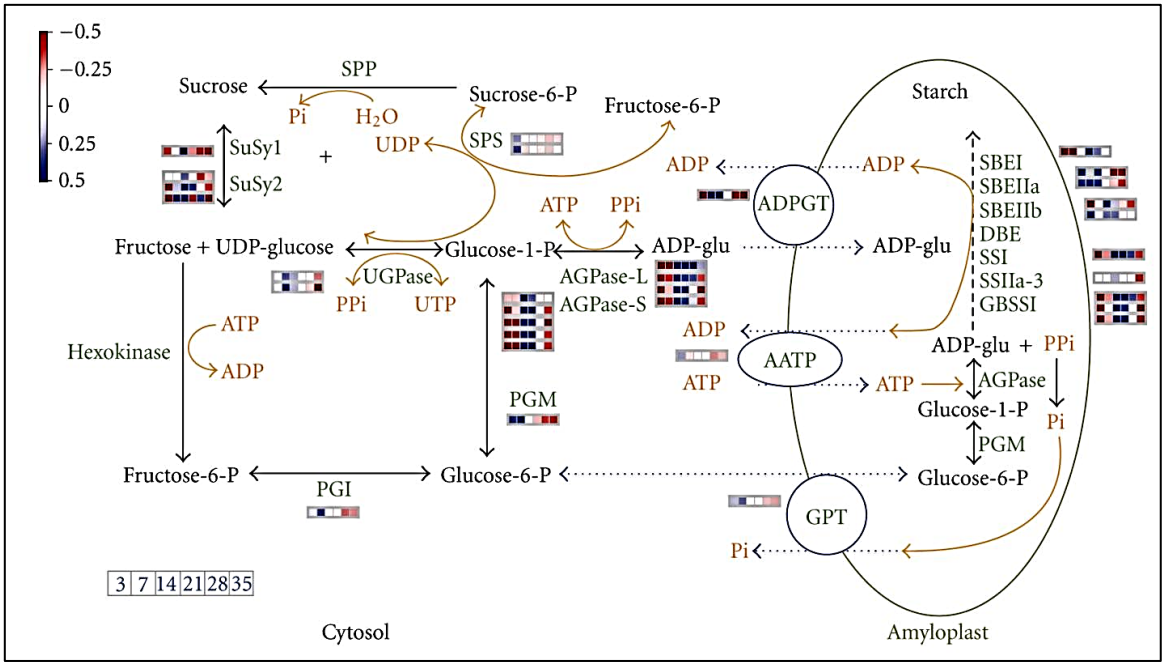


Figure 3: Schematic of the metabolic flux and expression of genes involved in sucrose to starch pathway in developing wheat seed. Upper box legend indicates level of gene expression and lower box legend indicates the time or developmental stage as days post-anthesis.

Further, SS catalyzes the synthesisof α (1-4) linkage between the non-reducing end ofa preexisting glucan chain and the glucosyl moiety of ADPglucose, causing the release of ADP. SSs can use both amyloseand amylopectin as substrates in vitro. How the initial primersfor the synthesis of glucan chains are produced in vivo is notknown. The α (1-6) branches in starch polymers are made by SBE, which hydrolyzes an (l-4) linkage within a chain and then catalyzesthe formation of an α (1-6) linkage between the reducingend of the "cut" glucan chain and another glucose residue, probably one from the hydrolyzedchain. SBEs show somespecificity for the length of the α (1-4) glucan chain that theywill use as a substrate.

Remodelling Starch Biosynthesis pathway

The genes associated with starch biosynthesis pathway has not been fully identified and characterised. With the advent of technology like NGS and gel-free proteomics, now it becomes easy to identify the respective transcripts and their proteins associated with SBP. Even, the transcriptional regulation of the genes coding for these enzymes has not yet been fully explored. Transcriptional regulation may be a more important mechanism for long-term control of genes expression especially during caryopsis development (Table 1).

Table 1: Identification of novel transcripts and SNIPs based marker associated with fructose, starch and sucrose metabolism pathways in wheat using de novo transcriptomic approach.

Differentially Expressed Genes Associated with Starch Biosynthesis

Sample	Fructose Pathway	Starch Metabolism	Sucrose Metabolism
HD85T Vs. HD29T	11	15	8
HD85C Vs. HD29T	17	16	5
HD29C Vs. HD29T	11	12	8
HD85C Vs. HD29C	7	9	7
HD85T Vs. HD29C	12	21	8
HD85C Vs. HD85T	20	38	10

Novel Genes Associated with Source and Sink in wheat under HS

Sample	Fructose Pathway	Starch Metabolism	Sucrose Metabolism
HD29C1	62	60	51
HD29C2	63	69	35
HD29T1	42	48	35
HD29T2	46	34	28
HD85C1	80	41	34
HD85C2	93	44	33
HD85T1	78	58	33
HD85T2	94	69	36

Sample Name	No. Of Predicted SNP
HD29C	17171
HD29T	24719
HD85C	18936
HD85T	44959

Single-Nucleotide Polymorphism (SNIPS) Identified in wheat under HS

Posttranslational regulation including phosphorylation, interaction with 14-3-3 regulatory proteins and posttranslational redox activation, appear to be essential regulatory mechanisms controlling starch biosynthesis by providing a rapid response to short-term environmental changes. The pathway in terms of synthesis and regulation has not been extensively studied. In our lab, we have executed whole transcriptome sequencing of contrasting wheat cvs. HD2985, HD2329, Raj3765 and BT-Schomburgkin a tissue specific manner at different stages of growth and development and under differential stress treatment. The transcriptome data generated from the developing endosperm tissue of contrasting wheat cvs. was analysed and characterised using different bioinformatics software's. We identified ~87, 100, and 46 novel transcripts associated with fructose, starch and sucrose metabolism pathway which was further validated in our lab. We identified 12 putative soluble starch synthase genes using de novo transcriptomic approach and cloned five of them for further functional validation (Table 1). Similarly, we identified 4 AGPase, 2 SBE and 2 GBSS genes from contrasting wheat cvs. based on the information generated using transcriptomic approach. We also identified ~100000 SNIPs lying in differentially expressed genes associated with starch metabolism (Table 1). These are the potential resources to be utilized for the breeding program in order to develop a 'climate-smart' crop.

References

Dugas D V, Monaco MK, Olson A, et al (2011) Functional annotation of the transcriptome of Sorghum bicolor in response to osmotic stress and abscisic acid. BMC Genomics 12:514. doi: 10.1186/1471-2164-12-514

Kumar RR, Goswami S, Sharma SK, et al (2015) Harnessing Next Generation Sequencing in Climate Change: RNA-Seq Analysis of Heat Stress-Responsive Genes in Wheat (*Triticum aestivum* L.). Omi A J Integr Biol 19:632–647.

Kumar RR, Pathak H, Sharma SK, et al (2014) Novel and conserved heat-responsive microRNAs in wheat (*Triticum aestivum* L.). Funct Integr Genomics. doi: 10.1007/s10142-014-0421-0

Maher CA, Palanisamy N, Brenner JC, et al (2009) Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci U S A 106:12353–8.

Nagalakshmi U, Waern K, Snyder M (2010) RNA-seq: A method for comprehensive transcriptome analysis. Curr. Protoc. Mol. Biol.

Chapter-29

Real time PCR Technique

A. Kumar*, Asharani Patel, Neelam Sheoran, Kuleshwar Prasad Sahu, Mukesh Kumar
Division of Plant Pathology, ICAR-Indian Agricultural Research Institute, New Delhi
**Email: kumar@iari.res.in*

Introduction

Over last several years, the development of novel chemistries and instrumentation platforms enabling detection of PCR products on real time basis has lead to wide spread adoption of real-time PCR as the method of choice for quantitative changes in gene expression. This is called "real-time PCR" because it allows us to actually view the increase in the amount of DNA as it is amplified. Real-time polymerase chain reaction, also known as quantitative real time polymerase chain reaction (qPCR) is used to amplify and simultaneously quantify a targeted DNA molecule. It enables both detection and quantification (as absolute number of copies or relative amount when normalized to DNA input or additional normalizing genes) of a specific sequence in a DNA sample. Quantitative real-time PCR (qPCR) is a sensitive technique for the detection and quantitation of specific DNA sequence in the environmental samples.

Principle

In Real time PCR, using sequence specific primers, the relative copies of a particular DNA or RNA sequence can be determined. We use the term relative since this technique tends to be used to compare relative copy numbers between tissues, organisms, or different genes relative to a specific housekeeping gene. 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.

Primer designing

RT-PCR primers specific to gene are to be designed using online IDT Primer-Quest software available at http://eu.idtdna.com or Primer 3 plus software available at http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi with the following parameters: optimal length, 20-25 base pairs; GC content, 50-55%; melting temperature, 52-60°C; amplicon length, 120 to 200 base pairs; maximum self-complementarity at the 3′ end - five nucleotides, and absence of stable hairpins & dimers. Primer specificity and quality

parameters were checked with the help of Oligo-Analyzer (https://www.idtdna.com/calc/analyzer).

Template preparation

A critical aspect of performing real time PCR is to begin with a template that is of high purity. The PCR template DNA is to be prepared critically using the commercial available kits to avoid inhibitors which could potentially interfere with cyclic reactions. The concentration of DNA should be about 5-30 nanograms.

Dyes & Fluorescence detection chemistry in qPCR

Probe based Quantitative PCR

Probe based qPCR relies on the sequence–specific detection of a desired PCR product. Unlike SYBR based qPCR methods that detect all double–stranded DNA, probe based qPCR utilizes a fluorescent–labeled target-specific probe resulting in increased specificity and sensitivity. Additionally, a variety of fluorescent dyes are available so that multiple primers can be used to simultaneously amplify many sequences. This chemistry is ideal for high throughput. Ready mixes contain all necessary components for qPCR you simply add the fluorescent detection chemistry, primers and template.

SYBR® Green based Quantitative PCR

SYBR Green I, a commonly used fluorescent DNA binding dye, binds all double-stranded DNA and detection is monitored by measuring the increase in fluorescence throughout the cycle. SYBR Green I has an excitation and emission maxima of 494 nm and 521 nm, respectively. Specificity of Sigma's SYBR based qPCR detection is greatly enhanced by the incorporation of a hot–start mediated taq polymerase, JumpStart Taq.

Two major applications for qPCR in plant pathology is explained below

i. Gene expression analysis using qPCR

Objectives: To estimate changes in gene expression or transcriptional changes

Quantitative PCR combines PCR amplification and detection into a single step. With qPCR, fluorescent dyes are used to label PCR products during thermal cycling. Real-time PCR instruments measure the accumulation of fluorescent signal during the exponential phase of the reaction for fast, precise quantification of PCR products and objective data analysis. With the help of qPCR we can analyze the changes in gene expression in a given sample (treated sample) relative to another reference sample such as an untreated control sample. In relative quantification, one can analyze changes in gene expression in a given sample relative to another reference sample (such as an untreated control sample).

Gene expression analysis using qPCR involves the following steps

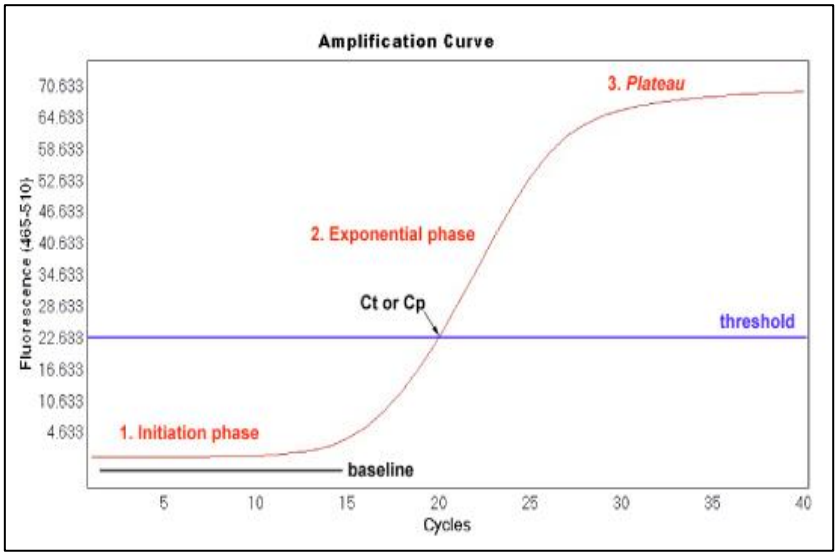
1. Sample Preparation

- Isolate total RNA
- Treat this isolated total RNA with DNAse I to avoid contamination with genomic DNA.
- Reverse transcribe this RNA and use the synthesized cDNA as a template for real-time quantitative PCR for gene expression analysis.
- Alternatively, nowadays one step qPCR reaction can be performed wherein cDNA synthesis and subsequent amplification can be done in one reaction.

2. qRT-PCR Amplification

- For PCR amplification, prepare reaction mixture containing cDNA template, gene-specific forward and reverse primers (Design gene-specific RT-PCR primers using primer 3 plus software) and SYBR green mix.
- For normalizing expression levels, use a constitutively expressed gene such as housekeeping gene for example 18S rRNA, GAPDH, β Actin etc
- Amplify the genes in a Real Time PCR machine. DNA is amplified using an initial denaturation at 95ºC for 3 min, followed by 35 cycles of 95ºC for 15s, annealing for 15s and extension 72ºC for 15s . Reaction is completed with a final extension step of 10 min at 72ºC.
- Agarose gel (2.0-2.5%) electrophoresis of the qPCR products can be performed to confirm that the individual qPCR products correspond to a single homogeneous cDNA fragment of expected size.

3. Data Analysis



Amplification plots represent the accumulation of product over the duration of the real-time PCR experiment consists of the following components

1. **Baseline:** During initial cycles of PCR, there is little change in fluorescence signal. An increase in fluorescence above the baseline indicates detection accumulated PCR product.
2. **Threshold line:** Point at which a reaction reaches a fluorescent intensity above background. It is set in the exponential phase of the amplification for the most accurate reading.
3. **Cycle Threshold, CT:** The cycle at which the sample reaches threshold level. CT value of 40 or more means no amplification and cannot be included in the calculations.
    - After visualizing the amplification curve, import the data into Real Time analysis software for further analysis.
    - The relative expression of genes is calculated using comparative Ct method which involves:
    - Comparing Ct values of the samples with a control or calibrator such as a non-treated sample.
    - The Ct values of both the calibrator and the samples are normalized to an endogenous housekeeping gene.
    - This gives ΔCt value of control and the sample.
    - The comparative Ct method is also known as 2-ΔΔCt method, where ΔΔCt = ΔCt, sample - ΔCt, reference Fold change = Efficiency-ΔΔCt or 2-ΔΔCt (which gives relative gene expression)

Transcript or copy number quantitation using qPCR

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. In absolute quantification using the standard curve method, one can quantitate unknowns based on a known quantity. First of all one has to create a standard curve that will be used to compare unknowns to the standard curve and extrapolate a value.

References

Saiki R. K; Gelfand D. H; Stoffel S; Scharf S. J; Higuchi R; Horn G. T; Mullis K. B; Erlich HA. Primer directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science, 1988; 29, 239(4839):487-91.

Pfaffl, MW; Tichopad, A; Prgomet, C; Neuvians, TP (2004). "Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: Best Keeper—Excel-based

tool using pair-wise correlations". Biotechnol Lett. 26 (6): 509–515. doi:10.1023/b:bile.0000019559.84305.47

Stephen A. Bustin; Vladimir Benes; Jeremy A. Garson; Jan Hellemans; Jim Huggett; Mikael Kubista; Reinhold Mueller; Tania Nolan; Michael W. Pfaffl; Gregory L. Shipley; Jo Vandesompele & Carl T. Wittwer. (Apr 2009). "The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments". *Clin. Chem*. 55 (4): 611–22. doi:10.1373/clinchem.2008.112797

Boggy G, Woolf PJ (2010). Ravasi T (ed.). "A Mechanistic Model of PCR for Accurate Quantification of Quantitative PCR Data". *PLOS ONE*. 5 (8): e12355.

Higuchi, R.; Dollinger, G.; Walsh, P.S.; Griffith, R. (1992). "Simultaneous amplification and detection of specific DNA-sequences". *Bio-Technology*. 10 (4): 413–417. doi:10.1038/nbt0492-413.

Kubista, M; Andrade, JM; Bengtsson, M; Forootan, A; Jonak, J; Lind, K; Sindelka, R; Sjoback, R; Sjogreen, B; Strombom, L; Stahlberg, A; Zoric, N (2006). "The real-time polymerase chain reaction". *Mol. Aspects Med*. 27 (2–3): 95–125. doi:10.1016/j.mam.2005.12.007

Logan J, Edwards K, Saunders N (editors) (2009). *Real-Time PCR: Current Technology and Applications*. Caister Academic Press. ISBN 978-1-904455-39-4.

Chapter-30

Recipe for molecular biology reagents

1. Lysis buffer for DNA Extraction
Components Total volume (10ml)
Proteinase K 80 µl
Lysozyme 100 µl
SDS 10% 500 µl
Tris-EDTA (TE) buffer (pH8) 9.32 ml

2. Composition of Tris- EDTA (TE) buffer
Components Total Volume (25ml)
1 M Tris Cl 25 µl
0.5 M EDTA (pH 8.0) 5 µl
Milli Q Water 24.97 ml

3. Preparation of 1 M Tris Cl (pH 8.0)
Dissolve 121g Tris base (MW-121.14) in 800 ml water. Adjust the pH to 8.0 with concentrated HCl (approx. 45ml). Add bidest to make the volume to 1 litre.

4. Preparation of EDTA (0.5 M) (pH 8.0)
Dissolve 186.1g EDTA ($Na_2EDTA\text{-}2H_2O$) (MW-372.24) in 700 ml of water by adjusting the pH to 8.0 with 10 M NaOH (approx. 45 ml), add make the volume to 1 litre.

5. Preparation of CTAB/NaCl solution (10% CTAB in 0.7 M Nacl)
Dissolve 4.1g NaCl in 80 ml and slowly add 10 g of Cetyl Trimethyl Ammonium Bromide (CTAB) while heating and stirring. If necessary, heat to 65°C to dissolve. Adjust final volume to 100 ml

6. Preparation of TAE Buffer
One liter 50X stock of TAE
Tris-base: 242 g
Acetate (100% acetic acid): 57.1 ml
EDTA: 100 ml 0.5M sodium EDTA
Add dH_2O up to one litre.
To make 1x TAE from 50X TAE stock, dilute 20ml of stock into 980 ml of DI water

7. Preparation of 5 M NaCl

Dissolve 292.2 g of NaCl in H2O and make upto 1 liter. Dispense into aliquots sterilize by autoclaving.

8. 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.

9. 3M Sodium acetate (pH 4.8 and 5.2)

Dissolve 408.1 g of sodium acetate 3H2O 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

10. Preparation protocol of 0.1% DEPC treated water

Wrap the reagent bottle with aluminum foil, add 0.1 ml DEPC (Diethyl pyrocarbonate) to 100 ml of the autoclaved distilled water, shake it overnight and autoclave it.