DNA SEQUENCING AND APPLICATION OF NGS TO STUDY AGRICULTURALLY IMPORTANT LIVESTOCK: PRESENT STATUS AND FUTURE PROSPECTS

*R.Umaya Suganthi, Manpal Sridhar

National Institute of Animal Nutrition and Physiology, Adugodi, Bangalore 560 030, Karnataka, India.

ABSTRACT

Unraveling the secret behind the sequence of DNA could be achieved by sequencing of DNA. Success in the endeavor of sequencing DNA, in part and as a whole was made possible by the amazing efforts of researchers worldwide. Sequencing of DNA started with small portions; and today, with the available Next Generation Sequencing Technologies, the whole genome of organisms could be sequenced accurately within a reasonable cost and time. Such information could be exploited to reduce infection and increase production in livestock. In this paper, the information about DNA, its structure, the efforts taken for initiating the sequencing of DNA, the latest sequencing platforms and the application of DNA sequencing in agriculturally important livestock are discussed.

Key words: DNA sequencing, next generation platforms, whole genome, livestock.

INTRODUCTION

Chromosomes are small structures, present inside the nucleus of the cells and they contain Deoxyribonucleic acid (DNA), Ribonucleic acid (RNA) and proteins. Except in some viruses, the DNA stores the genetic information in all the organisms in the earth, including viruses, bacteria, plants, birds, animals and human beings. It also transmits the genetic information from one generation to the next (from parents to progeny) and dictates the phenotype of the
organism. The composition and arrangement of DNA components are conserved among the organisms. [1]

In 1953, James Watson and Francis Crick proposed the structure of DNA. According to Watson and Crick, the DNA is in the form of double helix that measures 20Å in diameter and has two polynucleotide chains [(nucleotide: composed of purine bases- Adenine(A), Guanine (G); or pyrimidine bases- Cytosine (C), Thymine (T) and a five-carbon sugar (deoxyribose) linked by a phosphate group] that are coiled around a central axis, to form a right-handed double helix. The two chains run in opposite directions and are held together by weak thermodynamic forces. The bases of both chains are flat structures, which lie perpendicular to the axis. They are stacked on one another, 3.4Å apart and are located on the inside of the helix. The purine and pyrimidine bases in both the chains are paired to one another by hydrogen bonds. Adenine pairs with thymine by two hydrogen bonds and guanine pairs with cytosine by three hydrogen bonds. [1-3]

DNA SEQUENCING

The process of determining the order of bases in a section of DNA is called DNA sequencing. According to the Central Dogma of Molecular Genetics, a gene, composed of DNA is transcribed into ribonucleic acid (RNA) and then translated into a polypeptide and is then processed to become a protein. [1] Therefore, identifying the sequence of DNA would give the complete information about the genetic makeup of the organism. [3]

DNA REPLICATION AND DNA SEQUENCING: THE PROCESSES AND SIMILARITIES

DNA replication is a biological process in which the DNA present in the parent cell leads to the formation of two complete strands of DNA (one for each daughter cell) which are similar to the parent DNA. This process occurs in all living organisms and by this process; the genetic information is passed on from one generation to the next in an intact form. Before a cell begins to divide, specific enzymes present in cells break the hydrogen bonds between the bases of the two DNA strands and this results in the formation of two halves of the DNA. Such conformation favors the bonding of free nucleotides in the nucleus with the bases of both strands. As, A bonds only with T and C only with G, both strands form exact matches and the two are identical. [3]
DNA sequencing process is similar to DNA replication except that the process is carried out under man-made conditions. Most DNA sequencing techniques require a template, a primer, DNA polymerase enzyme and the four nucleotides. The template refers to a DNA sample with 50 million to 250 million nucleotides. The primer is a short oligonucleotide which is complementary to a region of the template and capable of being extended by addition of nucleotides. The function of DNA polymerase enzyme is to add nucleotides to the specific primer, as directed by the template strand. The techniques also represent a method by which the order of the nucleotides added to the primer can be detected. Various detection methods are available and applying the method of choice, the sequence of the DNA strand complementary to the DNA strand is determined.[4]

THE JOURNEY OF DNA SEQUENCING

It was only in the 1960s that methods were developed to sequence the nucleotides in nucleic acids. Initially the amino acid sequence of proteins was determined from which the nucleotide sequence was determined in the DNA or RNA and this technique was called reverse genetics and this did not give expected results. The similarity in chemical properties of DNA molecules, high chain length of DNA, and presence of only four bases and absence of base specific DNAases made sequencing of DNA a big task.[5]

A lot of effort has been put by scientists across the world for successful sequencing of DNA. The various events and success stories on DNA sequencing can be discussed as: before and after automation of DNA sequencing. The milestones that led to successful emanation of DNA sequencing are summarized in Table 1.

Table 1. The milestones in DNA Sequencing before automation [5]

<table>
<thead>
<tr>
<th>Name</th>
<th>Researcher involved with year</th>
<th>Milestones/Discoveries and their significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sinsheimer</td>
<td>1959</td>
<td>The first and foremost requirement for DNA sequencing is the availability of DNA molecule in pure form. The genome of bacteriophage φX174 was (first DNA molecule) purified to homogeneity [6]</td>
</tr>
<tr>
<td>Wu and Kaiser</td>
<td>1968</td>
<td>The partial sequence of phage lambda DNA was reported. Use of oligonucleotide primers in DNA sequencing reactions was introduced. However, the method could be applied only to short</td>
</tr>
<tr>
<td>Year</td>
<td>Event Description</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>1970</td>
<td>Sequencing the whole DNA as such is a huge task and this problem was overcome by the discovery of type 2 restriction enzymes by Hamilton Smith and Coworkers. These restriction enzymes can recognize and cleave (cut) DNA at specific short nucleotide sequences, typically 4-6 base pairs in length and thus provided a general method for cutting large DNA molecule into a number of smaller pieces.</td>
<td></td>
</tr>
<tr>
<td>1975</td>
<td>Developed a method called plus-minus, and the complete sequence of the φX DNA genome was sequenced. In this method, DNA is synthesized but is made to terminate at a given base and the products were separated by gel electrophoresis technique. The gels were then placed in contact with X-ray film for a suitable time and the film was developed (autoradiography). When the film was developed, molecules differing by a single nucleotide in length could be resolved as discrete bands. The plus-minus method could deduce a sequence of ~50 bases in a single experiment but it could not be applied for longer sequences.</td>
<td></td>
</tr>
<tr>
<td>1977</td>
<td>Chemical method was developed. In this method, DNA is cleaved at specific bases. The fragments thus obtained were resolved by a technique called Sodium Dodecyl Sulphate Polyacrylamide gel Electrophoresis technique and bands were produced for every sequence position and this method was widely adopted.</td>
<td></td>
</tr>
<tr>
<td>1977</td>
<td>Dideoxy method was developed. This method used single-stranded DNA template and DNA synthesis was driven by a primer but with chain-terminating nucleotide analogs. Dideoxy method of sequencing was used to read 100 base pairs to 400 base pairs length of DNA.</td>
<td></td>
</tr>
<tr>
<td>1981</td>
<td>Complete sequence of 16,569-base pair human mitochondrial genome was published.</td>
<td></td>
</tr>
<tr>
<td>1984</td>
<td>Complete nucleotide sequence of B95-8 strain of Epstein-Barr virus was established.</td>
<td></td>
</tr>
<tr>
<td>1982</td>
<td>Gene Bank was created to provide genetic sequences.</td>
<td></td>
</tr>
</tbody>
</table>
AUTOMATION OF DNA SEQUENCING
The automation of DNA sequencing was first explained by Leroy Hood in collaboration with Applied Biosystems (ABI) in the year 1986. In this method, sequencing data could be collected directly to a computer without autoradiography of the sequencing gel and develop programs to automatically interpret the data to produce an actual sequence. [5]

First Generation Sequencing Technologies
Once the sequencing methods were established, every effort was taken to move sequencing into an automated mode and the AB1 370A DNA sequencer was introduced and the automated Sanger method was developed. This was considered as a ‘first generation” technology. Using the automated sequencing, 337 new human genes were identified. The genomes of Haemophilus influenza, Mycoplasma genitalium, Eschericia coli were completely sequenced by automated sequencers. In 1996, the first commercial DNA sequencer was introduced by AB1 (AB1 Prism 3700 capillary sequences) and used for sequencing Drosophila melanogaster. In late 1996, complete sequence of yeast Saccharomyces cerevisiae (12.0kb) was sequenced and it was the first eukaryotic organism to be sequenced. Later the genome of the Bacillus subtilis was sequenced. This was then followed by sequencing of C. elagans in 1998 and it was the first animal genome to be sequenced.[5]

Next Generation Sequencing Technologies (NGS)
The requirement for delivery of fast, inexpensive and accurate genome information led to the development of newer methods of automated sequencing called next-generation sequencing (NGS) technologies. Basically, the methods involved in NGS technologies comprise the preparation of template; sequencing and data analysis. NGS produces enormous data at reasonable price. Using NGS technologies, the whole genome could be sequenced within a reasonable time. So these technologies could be used to sequence and study the genomes of hundreds of species of agronomic, evolutionary and ecological importance and as well as biomedical interest.[17]

Second generation sequencing technologies
In the recent years various firms have come out with their NGS platforms that use specific protocols for sequencing and they represent the Second Generation Sequencing Technologies. Such platforms includes: Roche 454 pyrosequencing, reversible terminator
sequencing by Illumina, sequencing by ligation of ABI/SoLiD, and single-molecule sequencing by Helicos.

Roche/454’S Genome Sequencer employs pyrosequencing and the run time is faster. The technique exploits emulsion PCR to achieve clonal amplification of target sequence. The sequencing machine contains many picoliter-volume wells with each one having a single bead and sequencing enzymes. Pyrosequencing uses luciferase to generate light for detection of the individual nucleotide incorporated into the nascent DNA.\[17\] Illumina’s Genome Analyzer performs massively parallel sequencing functions using a reversible terminator-based method that enables detection of single bases as they are incorporated into growing DNA strands. The supported Oligonucleotide Ligation and Detection system (SOLiD3) developed by Applied Biosystems is a highly accurate (greater than 99.94% basecalling accuracy) and massively parallel platform that enables completion of large-scale sequencing in a cost effective manner. This platform employs ligation reaction for sequencing using a repertoire of all possible oligonucleotides of a fixed length that are labeled according to the sequence position. The Polonator G.007 platform developed by Life Technologies is a least expensive platform that involves sequencing by ligation (joining of DNA strands) and is used in resequencing of bacterial genome. Helicos Genetic Analysis System, developed by Helicos Biosciences is the first commercial single molecule sequencer.\[18-20\] Here, the DNA fragments along with added polyA tail adapters are attached to the flow cell surface. After attachment, extension-based sequencing is carried out with cyclic washes of the flow cell with fluorescently labeled nucleotides, similar to Sanger sequencing. The second generation sequencing platforms were successfully used for various research applications. However, they did suffer from drawbacks including high cost of the instruments that varied between 0.5 to 1.0 million US dollars, complexities in sample preparation and chemistry (fluorescent labeling and enzyme substrate reaction) and optics and instrumentation.\[17\]

**Third generation sequencing technologies**

The demand for a low cost, handy, time saving sequencing technology that could work with a novel chemistry impacted the market players to come out with third generation sequencing platforms. This includes Pacific Biosciences’ real-time single molecule sequencing (PacBioRS), Complete Genomics’ combined proanchor hybridization and ligation (cPAL) and Ion Torrent of Life Technologies, Inc.
PacBioRS is a real-time single molecule-single polymerase sequencing platform which has a zero-mode wave guided nanostructures of 100nm holes. Inside these nanostructures, DNA polymerase performs sequencing by synthesis from sequentially introduced fluorophore labeled nucleotides.\cite{21-23} On the other hand, the proanchor hybridization and ligation (cPAL) sequencing employs rolling circle amplification of small DNA sequences into structures called nanoballs. Later unchained sequencing by ligation is used to reveal the nucleotide sequence. However, the major drawback with this system is the difficulty in mapping the short sequencing reads with the reference genomic database. Different from the other third generation platforms, the personal genomic machine (PGM) of Ion Torrent Technology has an array of semiconductor chips that sense minor changes in pH and detect nucleotide incorporation events by the release of a hydrogen ion from natural nucleotides. It is a bench top machine that does not require any special enzymes or labeled nucleotides for its functioning, thus making it a more advantageous one.\cite{24} Besides, the Ion Torrent PGM (Life Technologies), two other benchtop machines are available now. They are 454 GS Junior (Roche), MiSeq (Illumina) and are efficient in sequencing *Escherichia coli* O104:H4 genome.\cite{25}

**WHOLE GENOME SEQUENCES OF AGRICULTURALLY IMPORTANT LIVESTOCK**

Chicken is a main source of animal protein in both rural and urban areas throughout the world. Chickens were reported to be domesticated by Harappans of the Indus valley during the 2500-2100 BC and subsequently dispersed to various parts of the world. The red jungle fowl is the nearest ancestor to the domestic chicken. Until the middle of the last century most chicken breeds were raised in fairly small flocks for meat and eggs. The metamorphosis of poultry from a backyard venture to a multimillion dollar industry took place in the last eighty years. Today, the Indian poultry industry is booming and emerging as the world's second largest market with a growth of 12-15% year on year. Above and beyond a nutrient packet, the chicken (*Gallus gallus*) is an important model organism and serves as the main laboratory model for the ~9,600 avian species that exist today.\cite{26} In addition to being the organism of importance in terms of nutritional, economical and research fronts, the chicken also holds credit as the first non mammalian organism and agricultural animal whose genome was sequenced completely. The physical map\cite{27} and draft sequence of the 1.2 Gb chicken genome were completed at the Washington University Genome Sequencing Center. The first draft sequence of the Red Jungle fowl (USD 001) was assembled and the chicken genome is
reported to have 20,000-23,000 genes. The assembly was generated from 6.6 fold coverage in whole-genome shotgun reads which is a combination of plasmid, fosmid and bacterial artificial chromosome (BAC)-end read pairs.[28] The chicken genome has a haploid content of 1.2 x 10⁹ base pairs (bp) of DNA; approximately 40% that of either mouse or human and thus, the chicken genome is about one-third the size of the human genome.[29] In addition to a pair of sex chromosomes, chickens have 38 pairs of autosomes: 5 macro-, 5 intermediate, and 28 microchromosomes. A total of 571 ncRNA (non coding RNA) genes, from over 20 distinct gene families, were identified within the chicken genome assembly using bioinformatic approaches.[30,31] A total of between 20,000 and 23,000 protein-coding genes in chicken has been predicted by Ensembl.[28]

Besides chicken, cattle are agriculturally important as they provide meat and milk. They convert fibre into products that are consumed by man. Knowledge of the genome and genes of various functions in cattle would help to improve the nutrition of mankind. Cattle genome sequence has been established using bacterial artificial chromosome (BAC) and whole-genome shotgun (WGS) sequences.[32] The results revealed the presence of 22,000 genes in cattle. Further it was also observed that the number and organization of genes related to reproduction, immunity, lactation, and digestion in cattle were different with that of other species namely rat, human mouse (placental mammals, opossum (marsupial); and platypus (monotreme). The scientific evidence obtained is expected to help in genetic improvement in beef and dairy industries. Similarly, the genome sequence of a female domestic Duroc pig (Sus scrofa) has been published recently.[33] Similar to chickens, pigs have been domesticated by humans and the genomes have been modified right from the domestication to the present modern breeding practices. The report indicates that in pigs, the genes associated with immune response and olfaction exhibited fast evolution.

APPLICATION OF NGS TO STUDY AGRICULTURALLY IMPORTANT LIVESTOCK: PRESENT STATUS AND FUTURE PROSPECTS
Development of NGS technologies has induced a spurt in genome sequencing and is being exploited to unravel the science involved in the variations in livestock characteristics and few such studies have been carried out in turkey, chicken and cattle.

NGS and Turkey
Globally, turkey (Melearis gallopavo) is Mainly reared for its meat. It is the agriculturally most
important avian species next to chicken.\textsuperscript{[34,35]} The complete genome of turkey has been sequenced\textsuperscript{[36]} and is the domestic avain genome to be sequenced. Sequencing was done by the next-generation sequencing platforms, Roche 454 and Illumina GAII. The assembly has been created entirely on the basis of data obtained from NGS platforms except that BAC end sequences were used for aligning the chromosomes. The DNA of Nici (Nicholas Inbred), a female turkey has been used for sequencing. As expressed by the authors, Nici is from a subline (sib-mating for nine generations) originally derived from a commercially significant breeding line but with a genome that is extensively heterozygous.\textsuperscript{[36]} The genome assembly thus generated consists of approximately 1.1 gigabases of which 917 megabytes of sequences were aligned to specific chromosomes. Nearly 16,000 genes were identified and 15,093 were recognized a protein coding and 611 as non-coding RNA genes. Single nucleotide variants were identified (920,126) with transitions, to be more common than transversions.

Phylogenetic analyses indicate that the last common ancestor to the turkey and chicken lived between 20 and 40 million years ago.\textsuperscript{[37,38]} Comparision of the genome assembly of turkey with that of other birds (comparative genomics), chicken and zebrafinch and with mammalian species revealed genes specific for Turkey and avian lineage\textsuperscript{[39,40]} It was observed that 881 gene families were specific to turkeys and 271 were specific to birds. Certain genes, for example, the homologs for innate immune gene families (egs. chemokines, pattern recognition receptors) were observed in turkey, chicken, zebra finch and mammals. In contrary, all the three birds lacked the genes that encode tumor necrosis factor superfamily members TNFSF1 and TNFSF3 the lymphotoxins that controls lymph node formation in mammals and such difference is related to the absence of lymph nodes in birds.\textsuperscript{[41,42]}

**NGS and chicken**

In an organism like chicken, identification of single nucleotide polymorphisms (SNPs) would be of immense value in breeding studies. Approximately 2.8 million SNPs exists between the Red Jungle fowl (RJF #256 from the inbred line UCD001) and the three chicken breeds that includes a female layer (White Leghorn); a male broilers (Cornish); and a female Silkie and Roche 454 Titanium and FLX sequencing platforms were used in such studies.\textsuperscript{[43]}

**NGS and cattle**

Livestock species are subjected to selection for economically important quantitative traits and the information on genomic features that control such molecular aspects is limited. Throughout the world dairy population has been subjected to more than 50 year of intense
selection for milk production. Population genetic theory predicts that when alleles are under natural or artificial selection, their frequencies will change over time.\textsuperscript{[44]} So a study by Larkin et al.\textsuperscript{[45]} reconstructed the haplotypes of two influential ancestors in the pedigrees of the contemporary Holstein-Friesian (HF) population to identify chromosome segments and polymorphisms in dairy cattle that have been subjected to recent selection. The group selected the genomes of two of the most influential dairy sires in history, Walkway Chief Mark and his father Pawnee Farm Arlinda Chief. The two sires were selected based on the fact that each of these bulls accounts for 7\% of the genomes of the current North American Holstein cow populations. The DNA of Mark, Chief and 92 Mark sons obtained from US Dairy Bull DNA Repository were subjected to Emulsion-based clonal amplification and sequencing in a 454 Genome Sequencer FLXTitanium system (454 Life Sciences).\textsuperscript{[46,47]} The authors aligned 20.5 Gbp (approximately 7.3x coverage) and 37.9 Gbp (approximately 13.5x coverage) of the Chief and Mark genomic sequences respectively. They also detected more than 2.5 million high-quality autosomal SNPs with approximately 1.2 million and 757,000 high-quality SNPs in Mark and Chief respectively. From the polymorphisms in 49 chromosomal segments, a number of genes that are linked to economically important traits were identified to be subjected to selection in the studied HF cattle. The haplotypes of Chief reconstructed by resequencing indicated that Chief was heterozygous for two (Single Nucleotide Polymorphisms) SNPs in located in the 5’UTR on the gene that endoces ADP ribosylation factor-like 4A (ARL4A) on BTA4 chromosome. This information is important as one of these SNPs at BTA4 position 21,573,888 was associated with milk-production traits in the North American Holstein population.\textsuperscript{[48]} The study also revealed a strong selection for the gene (SCARB2 on BTA6) that encodes a receptor for the groups A, B and C of enterovirus 71 (EV71) that causes hand, foot and mouth disease in humans. This virus is again related to the picornavirus that causes foot and mouth disease in cattle and this gene has 118 SNPs heterozygous in Chief. The other interesting finding of this study was the strong selection observed in the gene that encodes plasminogen, the deficiency of which leads to milk stasis and premature mammary gland involution in mice.\textsuperscript{[49]}

**FUTURE PROSPECTS OF NGS IN AGRICULTURALLY IMPORTANT LIVESTOCK**

By taking advantage of NGS technologies genes involved in various functions, altered by their environment, important to their production and/or are involved in immune or physiological processes could be identified. Integration of DNA sequencing into the drug
discovery process will allow the identification of genes responsible for the susceptibility particular animal/ species to specific infections and as well as identifying diagnostic and/or theranostic markers. Additionally, genes responsible for the resistance of a particular animal/ species to specific infections could also be determined. The NGS technology could also be used to obtain the complete genome of the microbial flora of livestock. This in turn would help the animal scientists to develop vaccines and new therapeutics.\textsuperscript{50} Furthermore, sequencing will also enable a more sophisticated choice of drugs based on an animal’s or infectious organism’s genetic background (Fig.1).

The next generation sequencing technologies, thus offer the combination of speed, cost and accuracy demanded to meet the growing need for the study of genetic makeup of emerging infectious microbes, drug discovery and development processes and identification of genetic background of economic traits of livestock which ultimately would help in the improvement of livestock resources.

**CONCLUSION**

The effort of man to unveil the secrets about the structure, organization, maintenance of specific traits and susceptibility or resistance of living beings to specific diseases led to the identification of methods and technologies to study about DNA. The most advanced of such technologies is DNA sequencing. Humans started rearing livestock when their life pattern
shifted from hunter gatherers to settlers and till today they form a main source of nutrition and income throughout the world. It is no doubt that DNA sequencing is a powerful technique for livestock researchers. Presently, research efforts are being made throughout the world to exploit NGS technologies in veterinary field; but needs to be strengthened. Such research efforts are expected to focus on identification of individual genes controlling/affecting economic traits (milk, meat, egg production) and disease resistance in agricultural animals, which will ultimately benefit the farmers.

Conflict of interest: There is no conflict of interest.

REFERENCES


34. FAO: FAOSTAT: Production Database. [http://faostat.fao.org/]


