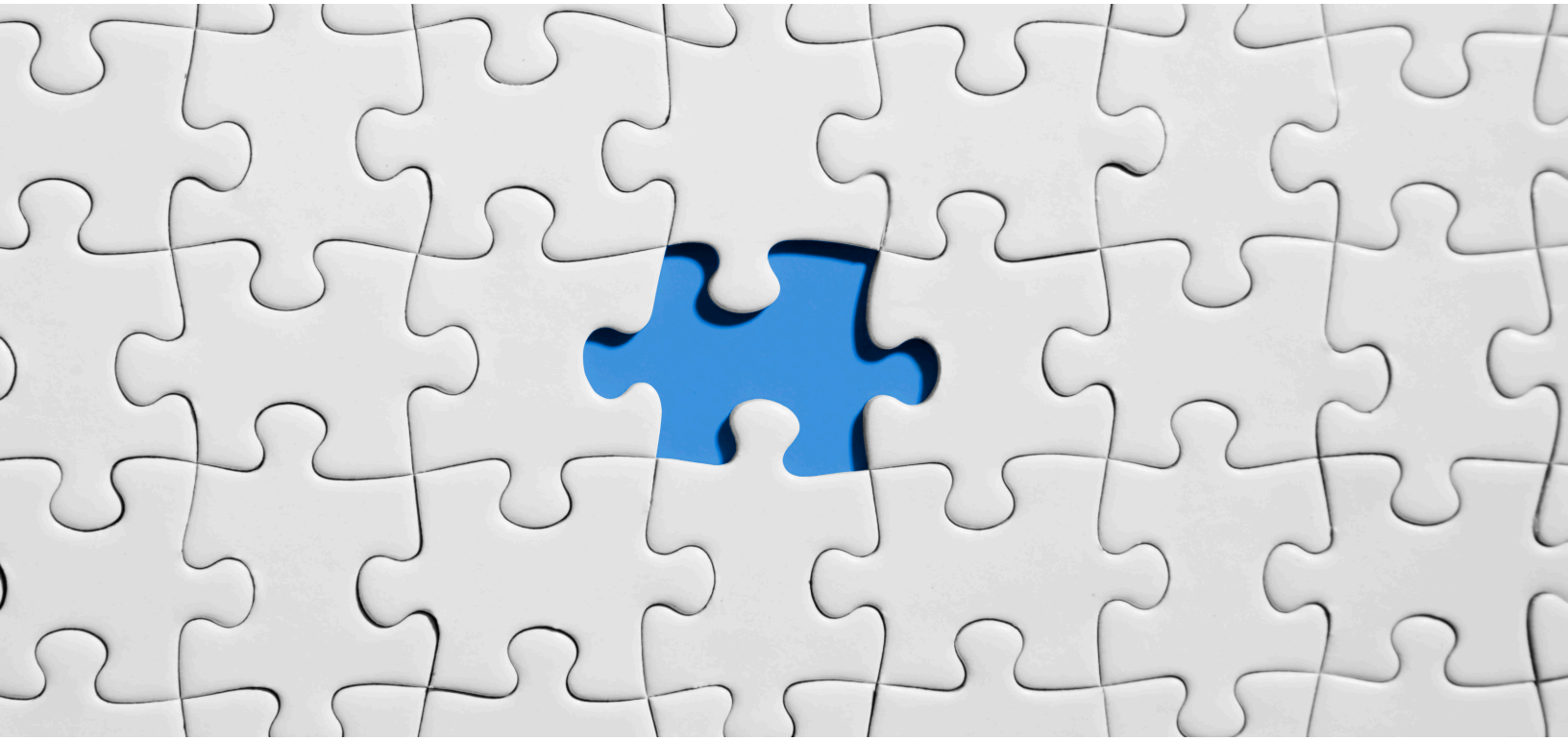
DELL Technologies

WHOLE GENOME SEQUENCING IN LIFE SCIENCE



Gangadhara N

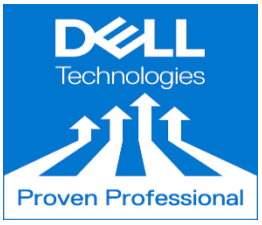
Gangadhara.n@dell.com

Lenin Ponnappa

Lenin.ponnappa@dell.com

DELL Technologies Proven Professional

Knowledge Sharing Article



The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged/Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage
- Enterprise Architect

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person Classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

Table of Contents

Disclaimer: The views, processes or methodologies published in this article are those of the authors. They do not necessarily reflect Dell Technologies' views, processes or methodologies.

Introduction

Whole Genome Sequencing

Entire genome sequencing (otherwise called WGS, full genome sequencing, total genome sequencing, or whole genome sequencing) is apparently the way toward deciding the total DNA succession of a life form's genome at a solitary time. This involves sequencing the entirety of a life form's chromosomal DNA just as DNA contained in the mitochondria and, for plants, in the chloroplast. Practically speaking, genome successions that are about finished are additionally called entire genome arrangements.

Sequencing of almost a whole human genome was first cultivated in 2000 halfway using shotgun sequencing innovation. While full genome shotgun sequencing for little (4000–7000 base pair) genomes was at that point being used in 1979,[27] more extensive application profited by pairwise end sequencing, referred to casually as twofold barrel shotgun sequencing. As sequencing ventures took on longer and progressively muddled genomes, numerous gatherings started to understand that valuable data could be acquired by sequencing the two parts of the bargains of DNA. In spite of the fact that sequencing the two parts of the bargains piece and monitoring the combined information was more bulky than sequencing a solitary finish of two particular sections, the information that the two groupings were arranged in inverse ways and were about the length of a part separated from one another was significant in recreating the succession of the first objective section.

Next Generation Sequencing

Next Generation sequencing (NGS) is becoming a key methodology for exploring the sub-atomic premise of ailments. Due to its affectability and particularity, NGS has been substituting heritage advancements for investigating maladies at the atomic level. Be that as it may, NGS has confronted some exceptional difficulties because of the pace of information creation outpacing that anticipated by Moore's law [1]. The pace of upgrades in DNA sequencing is rather anticipated based on the Carlson bend—a term authored by The Economist in an article in 2006. The National Human Genome Research Institute (NHGRI) tracks DNA sequencing cost per human genome and per nucleotide base, and updates these measurements yearly. NHGRI additionally charts the forecast of Moore's law for reference. With the appearance of the Illumina NovaSeq 6000 sequencer, huge scale human genome considerations are accessible to most specialists. The taking care of and investigation of the huge measures of information delivered present considerable difficulties and raise significant issues about how to guarantee straight-forward information examination methodology when analysts share information with established researchers. The accompanying bioinformatic pipelines were utilized to make the information displayed by Thermo Fisher Scientific in specialized notes that are related with Collibri library planning units.

Advantages of Whole-Genome Sequencing

- Gives a high-goals, base-by-base perspective on the genome
- Catches both large and small variations that may be missed with focused methodologies
- Recognizes potential causative variations for additional subsequent investigations of quality articulation and guideline components
- Conveys enormous volumes of information in a short measure of time to help get together of novel genomes

An Uncompromised View of the Genome

Dissimilar to centered methodologies, i.e. exome sequencing or focused on resequencing, which investigate a constrained segment of the genome, entire genome sequencing conveys a far reaching perspective on the whole genome. It is perfect for revelation applications, such as distinguishing between causative variations and novel genome when they get together. Entire genome sequencing can identify single nucleotide variations, additions/cancellations, duplicate number changes, and enormous auxiliary variations. Because of ensuing mechanical developments, the most recent genome sequencers can perform entire genome sequencing more proficiently than any other time in recent memory.

Key Whole-Genome Sequencing Methods

- Large Whole-Genome Sequencing
- Small Whole-Genome Sequencing
- De Novo Sequencing
- Phased Sequencing

Introduction to Large Whole-Genome Sequencing

Sequencing huge genomes (> 5 Mb) can give important data to illness and populace-level investigations. Scientists regularly utilize enormous entire genome sequencing to break down tumors, explore reasons for malady, select plants and creatures for farming rearing projects, and recognize normal hereditary varieties among populaces.

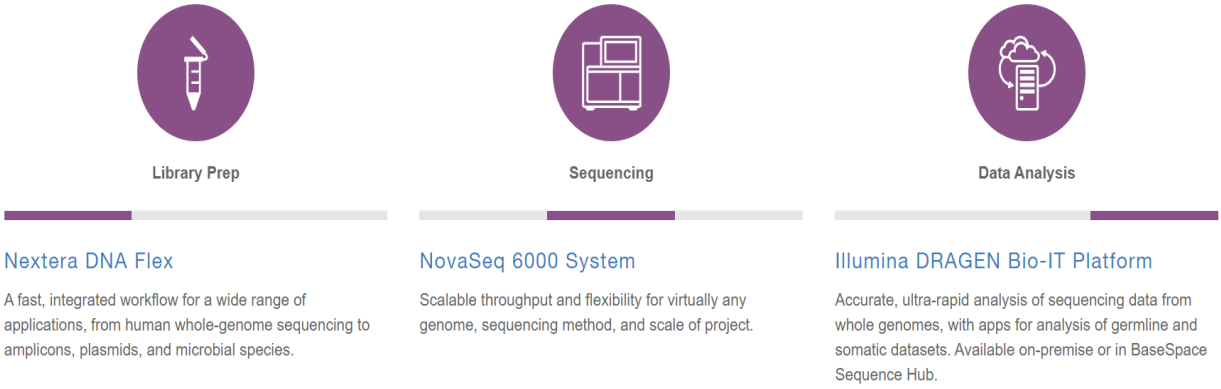
Advantages of Large Whole-Genome Sequencing

- Gives a high-goals, base-by-base perspective on the genome
- Consolidates short embeds and longer peruses to permit portrayal of any genome
- Uncovers infection-causing alleles that probably won't have been distinguished in any case
- Recognizes potential causative variations for additional follow-on investigations of quality articulation and guideline systems

A Comprehensive View of Genetic Variation

Examining the entire genome utilizing cutting edge sequencing (NGS) conveys a base-by-base perspective on all genomic modifications, including single nucleotide variations (SNV), additions and erasures, duplicate number changes, and auxiliary varieties. Matched end entire genome sequencing includes sequencing the two parts of the bargains piece, which improves the probability of arrangement to the reference and encourages location of genomic adjustments, dull successions, and quality combinations.

Recommended Workflow for Large Whole-Genome Sequencing



Comprehensive Large Genome Sequencing Workflow

Illumina sequencing by combination (SBS) science is the most generally embraced NGS innovation, producing roughly 90% of worldwide sequencing data. Notwithstanding our industry-driving information quality, Illumina offers incorporated work processes that improve the whole procedure, from library readiness to information investigation.

Below is the process flow for ech steps :

Library Preparation

- **Nextera DNA Flex Library Prep Kit**
  A quick, incorporated work process for a wide scope of utilizations, from human entire genome sequencing to amplicons, plasmids, and microbial species

- **TruSeq DNA PCR-Free Library Prep Kit**
  Library groundwork for entire genome sequencing of any species without PCR-actuated inclination

- **TruSeq Nano DNA Library Prep Kit**
  Proficient library readiness from tests with restricted accessible DNA, conveying high inclusion quality and decreased inclination

- **Nextera Mate Pair Library Prep Kit**
  Sans gel and gel-in addition to techniques for getting ready mate pair libraries for sequencing from low DNA input

- **Library Prep Kit Selector**
  Decide the best pack for your needs

Sequencing

- **NextSeq 550 System**
  Adaptable capacity to succession of the genome, exome, or transcriptome of any species, including up to 1 human genome for each run

- **HiSeq 4000 System**
  High throughput and minimal effort for generation scale genomics, sequencing up to 12 enormous genomes for each run
- **HiSeq X Series**
  $1000 human genome and ultra-high throughput for populace scale sequencing of up to 18,000 genomes for each year

- **NovaSeq 6000 System**
  Versatile throughput and adaptability for basically any genome, sequencing strategy, and size of venture

- **Sequencing Platform Comparison Tool**
  Think about sequencing stages and distinguish the best framework for your lab and applications

- **Sequencing Reagents and Flow Cells**
  Discover packs that incorporate sequencing reagents, stream cells, and additionally cushions custom-made to each Illumina sequencing framework

Data Analysis and Storage

- **Illumina DRAGEN Bio-IT Platform**
  The Illumina DRAGEN (Dynamic Read Analysis for GENomics) Bio-IT Platform gives ultra-quick optional examination of NGS information

- **BaseSpace Sequence Hub**
  The Illumina genomics registering condition for NGS information examination and the board

- **BaseSpace Whole-Genome Sequencing App**
  Rapidly extricates natural data from entire genome successions, utilizing Isaac arrangement and variation calling

- **BaseSpace Tumor-Normal App**
  Empowers analysts to recognize physical variations from a tumor and coordinated typical example pair

- **Integrative Genomics Viewer**
  Partners single quality or rundown of qualities with comment information for pathways, ailments, tissues, and little atoms

- **BaseSpace Correlation Engine**
  A developing library of curated genomic information to help scientists in recognizing ailment instruments, tranquilize targets, and biomarkers

- **BaseSpace Variant Interpreter**
  Use driving comment databases and an amazing sifting interface to distinguish infection related variations quickly

Introduction to Small Whole-Genome Sequencing

Small genome sequencing (≤ 5 Mb) includes sequencing the whole genome of a bacterium, infection, or other organism, and afterward contrasting the succession with a known reference. Sequencing little microbial genomes can be valuable for nourishment testing in general wellbeing, irresistible malady observation, atomic study of disease transmission contemplates, and ecological metagenomics.

Advantages of Small Genome Sequencing:

- Allows investigation of all genes from single organism culture
- Sequences thousands of organisms in parallel
- Provides comprehensive analysis of the microbial or viral genome
- Aids discovery of new biomarkers within a microbial or viral sample by providing distinct gene information from homologous chromosomes, supporting haplotyping

Fast, Culture-Free Microbial Analysis

In contrast to conventional methodologies, little genome sequencing contemplates utilizing cutting edge sequencing (NGS) that don't depend on work serious cloning steps. NGS likewise empowers researchers to arrangeme several life forms at the same time by means of multiplexing. NGS can recognize low-recurrence variations, genomic improvements, and other hereditary changes that may be missed or are too exorbitant to even consider identifying utilizing different strategies. For little genomes, DNA libraries can be arranged, sequenced, and investigated in as little as 2 days.

Recommended Workflow for Small Whole-Genome Sequencing

Comprehensive Small Whole-Genome Sequencing Workflow

Illumina sequencing by amalgamation (SBS) science is the most generally received NGS innovation, creating around 90% of worldwide sequencing data. Notwithstanding our industry-driving information quality, Illumina offers coordinated work processes that streamline small genome sequencing, from library arrangement to information investigation.

Below is the process flow for each step:

Library Preparation

- **Nextera DNA Flex Library Prep Kit**
  A fast, integrated workflow for a wide range of applications, from human whole-genome sequencing to amplicons, plasmids, and microbial species

- **Nextera XT DNA Library Prep Kit**
  Fast and easy library preparation workflow for small genomes, PCR amplicons, and plasmids

- **Nextera Mate Pair Library Prep Kit**
  Gel-free and gel-plus methods for preparing mate pair libraries for sequencing from low DNA input

Sequencing

- **iSeq 100 System**
  Affordable, fast, and accessible sequencing power for targeted or small genome sequencing in any lab

- **MiniSeq System**
  Access cost-effective sequencing, even for low numbers of samples. Sequence up to 50 small genomes per run

- **MiniSeq System**
  Speed and simplicity for targeted and small genome sequencing applications, processing up to 384 small genomes per run

- **Platform Comparison Tool**
  Compare sequencing platforms and identify the best system for your lab and applications

- **Sequencing Reagents**
  Find kits that include sequencing reagents, flow cells, and/or buffers tailored to each Illumina sequencing system

Data Analysis and Storage

- **Local Run Manager**
  An on-premises software solution for creating sequencing runs, monitoring run status, and analyzing data

- **BWA Whole-Genome Sequencing App**
  Processes whole-genome sequencing data using BWA alignment and GATK variant calling

- **BaseSpace Whole-Genome Sequencing App**
  Quickly extracts biological information from whole-genome sequences, using Isaac alignment and variant calling

- **SPAdes Genome Assembler**
  De novo assembler suitable for single-cell and isolate genomes

- **BaseSpace Sequence Hub**
  The Illumina genomics computing environment for NGS data analysis and management

Introduction to De Novo Sequencing

De novo sequencing refers to sequencing a novel genome where there is no reference sequence available for alignment. Sequence reads are assembled as contigs, and the coverage quality of de novo sequence data depends on the size and continuity of the contigs (i.e. the number of gaps in the data). Next-generation sequencing (NGS) enables faster, more accurate characterization of any species compared to traditional methods, such as Sanger sequencing. Illumina offers mate pair sequencing and long-read technology to complement shorter reads for comprehensive, accurate characterization of any species.

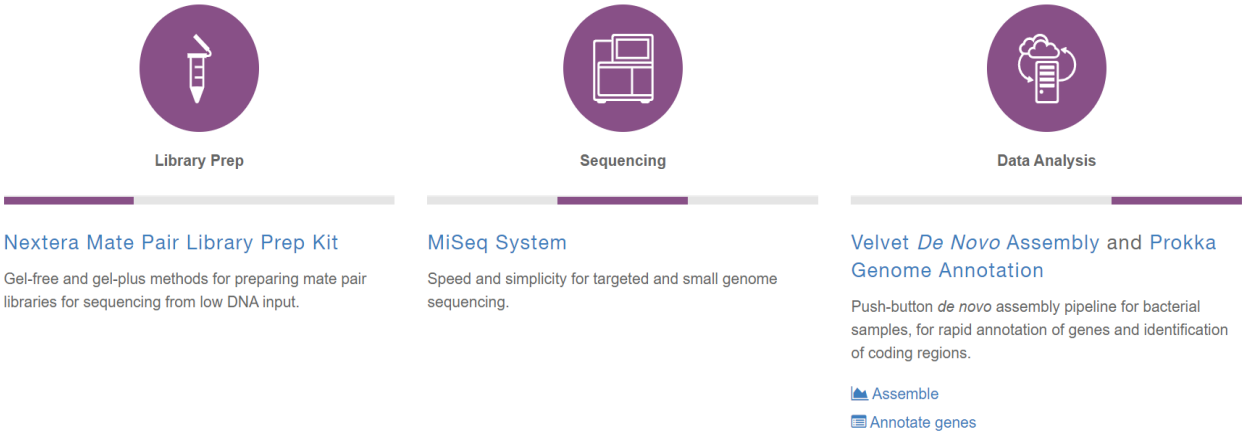
Advantages of De Novo Sequencing

- Generates accurate reference sequences, even for complex or polyploid genomes
- Provides useful information for mapping genomes of novel organisms or finishing genomes of known organisms
- Clarifies highly similar or repetitive regions for accurate de novo assembly
- Identifies structural variants and complex rearrangements, such as deletions, inversions, or translocations

Accurate Genome Assembly

When sequencing a genome for the first time, use a combined approach for higher-quality assemblies. Combining short-insert, paired-end and long-insert, mate pair sequences is the ideal way to maximize coverage. The short reads, sequenced at higher depths, can fill in gaps not covered by the long inserts. This combination enables detection of the widest range of structural variant types and is essential for accurate identification of complex rearrangements. Synthetic long reads can also aid assembly by providing long contigs that are "stitched" together from shorter reads to maintain accuracy

Recommended Workflow for De Novo Bacterial Sequencing



Comprehensive De Novo Genome Sequencing Workflow

Illumina sequencing by amalgamation (SBS) science is the most generally received NGS innovation, creating around 90% of worldwide sequencing data. Notwithstanding our industry-driving information quality, Illumina offers coordinated work processes that streamline de novo sequencing, from library arrangement to information investigation.

Below is the process flow for each step:

Library Preparation

- **Nextera Mate Pair Library Prep Kit**
  Gel-free and gel-plus methods for preparing mate pair libraries for sequencing from low DNA input

- **10x Genomics Chromium Genome Library Prep Kit**
  Whole genome prep that provides variant calling and phasing for sequencing on Illumina platforms from low DNA input

- **Dovetail Service**
  De novo assembly for a large range of genomes using the proprietary Dovetail Chicago TM method on Illumina platforms from multiple DNA inputs

- **Long-Read Sequencing Technology**
  Highly accurate assembly of DNA fragments into long reads for whole-genome sequencing or genome phasing

- **NRGene Service**
  De novo assembly of complex genomes for Ag researchers to deliver long, phased sequences and accurate assembly results on Illumina platforms

Sequencing

- **MiSeq System**
  Speed and simplicity for targeted and small genome sequencing

- **NextSeq 550 System**
  Flexible desktop sequencer supporting multiple applications, sequencing 1 Nextera Mate Pair sample per run

- **HiSeq 4000 System**
  High throughput and low cost for production-scale genomics

- **NovaSeq 6000 System**
  Scalable throughput and flexibility for virtually any genome, sequencing method, and scale of project

- **Platform Comparison Tool**
  Compare sequencing platforms and identify the best system for your lab and applications

- **Sequencing Reagents**
  Find kits that include sequencing reagents, flow cells, and/or buffers tailored to each Illumina sequencing system

Data Analysis and Storage

- **Nextera Mate Pair**
  De novo assembler suitable for single-cell and isolate genomes

- **BaseSpace Velvet De Novo Assembly App**
  De novo assembly of bacterial genomes using the Velvet assembler, with a focus on Nextera Mate Pair data

- **BaseSpace Whole-Genome Sequencing App Visualization**
  Quickly extracts biological information from whole-genome sequences, using Isaac alignment and variant calling

- **SPAdes Genome Assembler**
  De novo assembler suitable for single-cell and isolate genomes

- **BaseSpace Sequence Hub**
  The Illumina genomics computing environment for NGS data analysis and management

Introduction to Phased Sequencing

Generally, entire genome sequencing produced a solitary agreement arrangement without recognizing variations on homologous chromosomes. Staged sequencing, or genome staging, addresses this restriction by recognizing alleles on maternal and fatherly chromosomes. This data is frequently significant for understanding quality articulation designs for hereditary sickness examination.

Benefits of Phased Sequencing

Next generation sequencing (NGS) empowers entire genome staging without depending on trio examination or factual induction. By distinguishing haplotype data, staged sequencing considers complex characteristics, which are regularly affected by cooperations among various qualities and alleles. Staging can likewise give significant data to explore hereditary ailments, as interruptions to alleles in cis or trans positions on a chromosome can cause hereditary issues.

Phasing can help researchers:

- Analyze compound heterozygotes
- Measure allele-specific expression
- Identify variant linkage

Human Whole-Genome Phasing Solutions

10x Genomics Partnership

The 10x Genomics Chromium Genome Library Prep Kit provides long-range information on a genome-wide scale, including variant calling, phasing, and extensive characterization of genomic structure using NGS technology

BaseSpace Sequence Hub

The Illumina genomics computing environment, offering specialized apps for NGS data analysis and management

Conclusion

In this way, the different and various methodologies of Whole Genome Sequencing is explained with their advantages. These advances imply that life forms that had recently been considered excessively mind boggling or too costly to even think about warranting entire genome sequencing would now be able to be considered by analysts in their own research facilities. As stated by Dr. Christiaan Henkel: “We can finally start looking at massive genomes which we have previously had to ignore because we couldn’t sequence them.”

References

- Mak, A. C. et al. Genome-Wide structural variation detection by genome mapping on nanochannel arrays. Genetics 202(1):351–62 (2016).
- Fleischmann, R.D. et al. (1995). Wholegenome random sequencing and assembly of Haemophilus influenzae. Science 269(5223): 496–512.
- Whole-Genome Sequencing - Elaine R. Mardis, in Genomic and Personalized Medicine (Second Edition)