bionivid

The Genome Education
Nurturing Careers

The bionivid Science Blog

UNDERSTANDING NGS FILE FORMATS

MARCH - 2025 - IV

BIONIVID TECHNOLOGY PRIVATE LIMITED

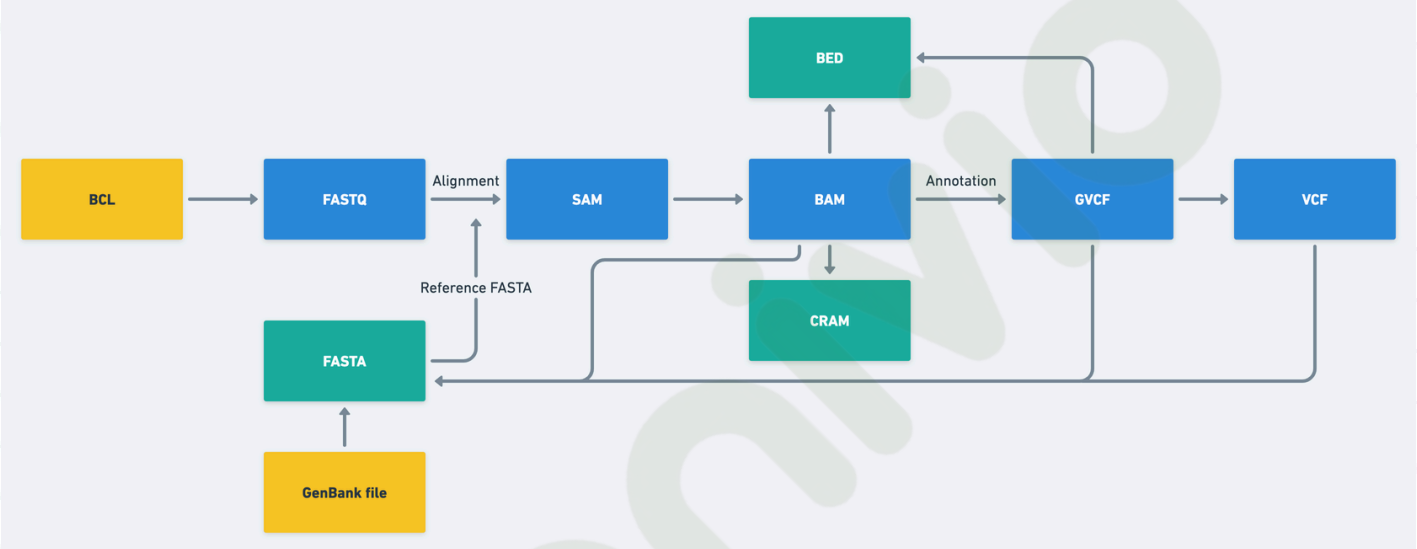
BENGALURU, KARNATAKA, INDIA.

info@bionivid.com | sales@bionivid.com

+91 95356 19191 | +91 96064 56771

www.bionivid.in

Next-Generation Sequencing (NGS) technologies generate extensive data, necessitating a variety of file formats to store and manage sequencing information efficiently. These formats support different stages of sequencing workflows, from raw read storage to alignment and annotation.



Why So Many Formats?

NGS file formats exist to optimize data management at each stage of sequencing analysis. Different formats cater to:

- Efficient storage and processing requirements
- Specialized data types

  (raw sequences, alignments, annotations, etc.)
- Compatibility with specific bioinformatics tools and pipelines

NGS file formats are categorized into raw sequence files, coordinate files, parameter files, annotation files, and metadata files, each serving a distinct purpose.

Raw Sequence Data Formats

- **FASTA / FNA (FASTA Nucleotide Format)**: A widely used format for storing nucleotide or protein sequences, identified by a unique header line starting with '>'.
- **FASTQ**: The most commonly used format, containing nucleotide sequences and Phred quality scores in ASCII format.
- **BCL (Base Call Format)**: Generated by Illumina sequencers and converted to FASTQ through demultiplexing.
- **uBAM (Unaligned Binary Alignment Map)**: Used by platforms like PacBio for storing raw reads before alignment.
- **SFF (Standard Flowgram Format)**: Used in 454 sequencing for storing raw reads and quality scores.

Quality Scoring and Base Calling Formats

- **Phred Scores:** Used in FASTQ files to indicate base-calling confidence.
- **QUAL Files:** Store base quality scores separately.
- **CSFASTA (Color Space FASTA):** Used in SOLiD sequencing, encoding sequences with colors instead of nucleotides.
- **PRB (Probability Score Format):** Used in Illumina sequencing to store base-call probabilities.

Read Alignment Formats

- **SAM (Sequence Alignment/Map):** A text-based format used to store aligned sequencing reads.
- **BAM (Binary Alignment/Map):** A compressed binary version of SAM, enabling faster analysis.
- **CIGAR Strings:** Used within SAM/BAM files to represent sequence alignment details, including matches, insertions, deletions, and skipped regions.
- **QSEQ:** A tab-delimited file format used by Illumina, containing raw sequencing reads before conversion to FASTQ.
- **SCARF (Solexa Compact ASCII Read Format):** Used in older Solexa sequencing technologies.

Variant and Structural Data Formats

- **VCF (Variant Call Format):** Stores detected genetic variants, including SNPs and structural variations.
- **BED (Browser Extensible Data):** Represents genomic regions without sequence data, optimizing computational efficiency.
- **GFF/GTF (General Feature Format/General Transfer Format):** Store gene annotations and feature information related to genomic sequences.

Data Storage and Public Repositories

- **SRA (Sequence Read Archive):** Standardized by NCBI, EBI, and DDBJ for storing raw sequencing reads.
- **Index Files (.bai, .tbi, .fai):** Facilitate quick retrieval of sequences within large datasets.
- **CSV/TSV:** Simple tabular formats for metadata and structured sequencing information.
- **HDF:** A hierarchical data format used in PacBio and Oxford Nanopore sequencing for efficient storage and retrieval.

Multiplexing and Barcode Identification

To maximize sequencing efficiency, NGS runs often pool multiple samples using unique DNA barcodes:

- **Manifest Files:** Specify sample barcodes to aid demultiplexing.
- **Dual Indexing:** Reduces misidentification errors in multiplexed sequencing.

Conclusion

Understanding NGS file formats is crucial for managing sequencing data efficiently. Each format plays a unique role in processing and analysis, ensuring accurate results and seamless bioinformatics workflows. Mastery of these formats enables researchers to make the most of high-throughput sequencing technologies, advancing discoveries in genomics and transcriptomics.

bionivid

BIONIVID TECHNOLOGY PRIVATE LIMITED

BENGALURU, KARNATAKA, INDIA.

info@bionivid.com | sales@bionivid.com

+91 95356 19191 | +91 96064 56771

www.bionivid.in

The Genome Education

Nurturing Careers