= European Nucleotide Archive =

The European Nucleotide Archive (ENA) is a repository providing free and unrestricted access to annotated DNA and RNA sequences . It also stores complementary information such as experimental procedures , details of sequence assembly and other metadata related to sequencing projects . The archive is composed of three main databases : the Sequence Read Archive , the Trace Archive and the EMBL Nucleotide Sequence Database (also known as EMBL @-@ bank) . The ENA is produced and maintained by the European Bioinformatics Institute and is a member of the International Nucleotide Sequence Database Collaboration (INSDC) along with the DNA Data Bank of Japan and GenBank .

The ENA has grown out of the EMBL Data Library which was released in 1982 as the first internationally supported resource for nucleotide sequence data . As of early 2012 , the ENA and other INSDC member databases each contained complete genomes of 5 @,@ 682 organisms and sequence data for almost 700 @,@ 000 . Moreover , the volume of data is increasing exponentially with a doubling time of approximately 10 months .

= = History = =

The European Nucleotide Archive originated from separate databases , the earliest of which was the EMBL Data Library , established in October 1980 at the European Molecular Biology Laboratory (EMBL) , Heidelberg . The first release of this database was made in April 1982 and contained a total of 568 separate entries consisting of around 500 @,@ 000 base pairs . In 1984 , referring to the EMBL Data Library , Kneale and Kennard remarked that " it was clear some years ago that a large computerized database of sequences would be essential for research in Molecular Biology " .

Despite the primary distribution method at the time being via magnetic tape , by 1987 , the EMBL Data Library was being used by an estimated 10 @,@ 000 scientists internationally . The same year , the EMBL File Server was introduced to serve database records over BITNET , EARN and the early Internet . In May 1988 the journal Nucleic Acids Research introduced a policy stating that " manuscripts submitted to [Nucleic Acids Research] and containing or discussing sequence data must be accompanied by evidence that the data have been deposited with the EMBL Data Library . "

During the 1990s the EMBL Data Library was renamed the EMBL Nucleotide Sequence Database and was formally relocated to the European Bioinformatics Institute (EBI) from Heidelberg . In 2003 , the Nucleotide Sequence Database was extended with the addition of the Sequence Version Archive (SVA) , which maintains records of all current and previous entries in the database . A year later in June 2004 , limits on the maximum sequence length for each record (then 350 kilobases) were removed , allowing entire genome sequences to be stored as a single database entry .

Following the uptake of Sanger sequencing , the Wellcome Trust Sanger Institute (then known as The Sanger Centre) had begun cataloguing sequence reads along with quality information in a database called The Trace Archive . The Trace Archive grew substantially with the commercialisation of high @-@ throughput parallel sequencing technologies by companies such as Roche and Illumina . In 2008 , the EBI combined the Trace Archive , EMBL Nucleotide Sequence Database (now also known as EMBL @-@ Bank) and a newly developed Sequence (or Short) Read Archive (SRA) to make up the ENA , aimed at providing a comprehensive nucleotide sequence archive . As a member of the International Nucleotide Sequence Database Collaboration , the ENA exchanges data submissions each day with both the DNA Data Bank of Japan and GenBank .

= = EMBL Nucleotide Sequence Database = =

The EMBL Nucleotide Sequence Database (also known as EMBL @-@ Bank) is the section of the ENA which contains high @-@ level genome assembly details , as well as assembled sequences and their functional annotation . EMBL @-@ Bank is contributed to by direct submission from genome consortia and smaller research groups as well as by the retrieval of sequence data associated with patent applications .

As of release 114 (December 2012) , the EMBL Nucleotide Sequence Database contains approximately 5 × 1011 nucleotides with an uncompressed filesize of 1 @.@ 6 terabytes .

= = = Data classes = = =

The EMBL Nucleotide Sequence Database supports a variety of data derived from different sources including , but not limited to :

Expressed sequence tags with their associated sample data .

Nucleotide sequence being generated from whole genome sequencing projects at varying stages of assembly , including complete contigs and annotated , fully assembled sequence .

Data relating to transcriptomics , such as complementary DNA , with optional annotation .

Novel or extended annotations of existing coding sequences , for example new sequence versions with corrected start or stop codons .

= = = EMBL @-@ Bank format = = =

The EMBL Nucleotide Sequence Database uses a flat file plaintext format to represent and store data which is typically referred to as EMBL @-@ Bank format . EMBL @-@ Bank format uses a different syntax to the records in DDBJ and GenBank , though each format uses certain standardised nomenclature , such as taxonomies as defined by the NCBI Taxon database . Each line of an EMBL @-@ format file beings with a two @-@ letter code , such as AC to label the accession number and KW for a list of keywords relevant to the record ; each record ends with / / .

= = Sequence Read Archive = =

The ENA operates an instance of the Sequence Read Archive (SRA) , an archival repository of sequence reads and analyses which are intended for public release . Originally called the Short Read Archive , the name was changed in anticipation of future sequencing technologies being able to produce longer sequence reads . Currently , the archive accepts sequence reads generated by next @-@ generation sequencing platforms such as the Illumina Genome Analyzer and ABI SOLiD as well as some corresponding analyses and alignments . The SRA operates under the guidance of the International Nucleotide Sequence Database Collaboration (INSDC) and is the fastest @-@ growing repository in the ENA .

In 2010 the Sequence Read Archive made up approximately 95 % of the base pair data available through the ENA , encompassing over 500 @,@ 000 @,@ 000 @,@ 000 sequence reads made up of over 60 trillion (6 × 1013) base pairs . Almost half of this data was deposited in relation to the 1000 Genomes Project wherein the researchers published their sequence data to the SRA in real @-@ time . In total , as of September 2010 , 65 % of the Sequence Read Archive was human genomic sequence , with another 16 % relating to human metagenome sequence reads .

The preferred data format for files submitted to the SRA is the BAM format , which is capable of storing both aligned and unaligned reads . Internally the SRA relies on the NCBI SRA Toolkit , used at all three INSDC member databases , to provide flexible data compression , API access and conversion to other formats such as FASTQ .

= = Data access = =

The data contained in the ENA can be accessed manually or programmatically via REST URL through the ENA browser . Initially limited to the Sequence Read Archive , the ENA browser now also provides access to the Trace Archive and EMBL @-@ Bank , allowing file retrieval in a range of formats including XML , HTML , FASTA and FASTQ . Individual records can be accessed using their accession numbers and other text queries are enabled through the EB @-@ eye search engine .

Additionally , sequence similarity @-@ based searches implemented using De Bruijn graphs offer another method of retrieving records from the ENA .

The ENA is accessible via the EBI SOAP and REST APIs , which also offer access to other databases hosted at the EBI , such as Ensembl and InterPro .

= = Storage = =

The European Nucleotide Archive handles large volumes of data which pose a significant storage challenge . As of 2012 , the ENA 's storage requirements continue to grow exponentially , with a doubling time of approximately 10 months . To manage this increase , the ENA selectively discards less @-@ valuable sequencing platform data and implements advanced compression strategies . The CRAM reference @-@ based compression toolkit was developed to help reduce ENA storage requirements .

= = Funding = =

Currently the ENA is funded jointly by the European Molecular Biology Laboratory , the European Commission and the Wellcome Trust . The emerging ELIXIR framework , coordinated by EBI director Janet Thornton , aims to secure a sustainable European funding infrastructure to support the continued availability of life science databases such as the ENA .