

Plant genetics: extract DNA and explore the challenge of gene sequencing

Bioinformatics

Researchers today depend on access to large datasets of many different types, including genes, proteins, and the behaviour of small molecules.

Breakthrough methods, such as DNA sequencing, have changed the face of research in such a short time that they are considered to be disruptive technologies: they are completely changing how biologists work. These days, all biologists need to be able to use databases and data-analysis tools, and they need to have some level of coding knowledge.

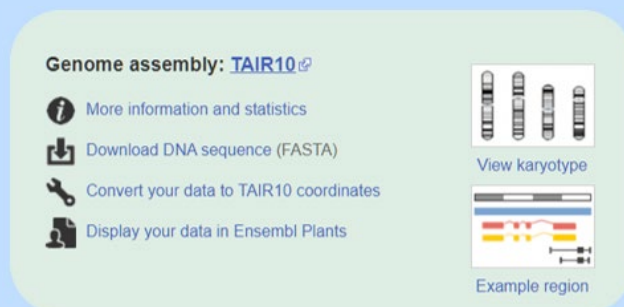
Life-science experiments are generating a flood of data every day, which is good news for researchers but poses practical challenges. The amount of data produced is doubling twice as quickly as computer storage and processing power, and this rate is increasing.

Bioinformatics makes it possible to collect, store, and add value to this data, so that researchers in many fields can find and analyze them efficiently. EMBL-EBI is one of few places in the world that has the capacity and expertise to fulfil this important task.

Ensembl

The Ensembl project was started in 1999 to annotate the human genome and make all data publicly and freely available via the web. Many more genomes have since been added to Ensembl, and the range of available data has also expanded to include comparative genomics, variation, and regulatory data.

In 2009, the Ensembl Genomes project was launched, with specific web portals for plant, fungal, invertebrate metazoan, bacterial, and protist genomes. These portals provide the evolutionary context in which genes can be understood, as well as coverage of all major nonvertebrate experimental organisms, species of agricultural importance, pathogens, and vectors.



Genome assembly: [TAIR10](#)

- More information and statistics
- Download DNA sequence (FASTA)
- Convert your data to TAIR10 coordinates
- Display your data in Ensembl Plants

View karyotype

Example region

Some example functions on the EnsemblPlants database. Scientists can look up published genes for plants they are studying.

Image: [EnsemblPlants/EMBL](#)

Where does the data come from, who uses it, and what for?

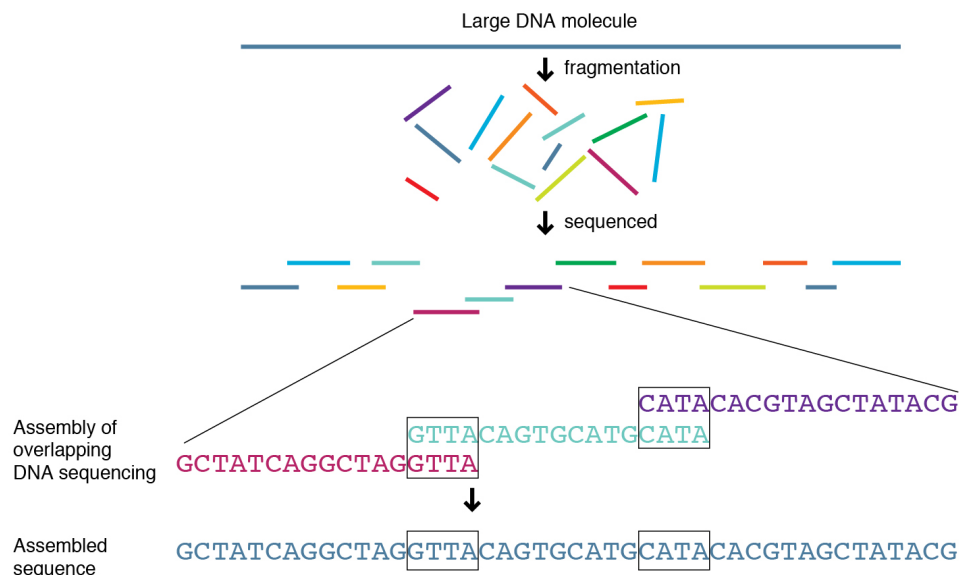
Data cycle steps:

1. Scientists around the world generate data and make discoveries.
2. Scientists publish their findings and deposit data with EMBL-EBI (just like new books are made available to all through libraries).
3. EMBL-EBI experts archive and make the data available to all scientists free of charge.
4. EMBL-EBI experts also add value to the data by classifying, enriching, and analyzing it.
5. EMBL-EBI distributes/makes available both raw data and added-value data.
6. Researchers then use the data to develop new experiments or ask other research questions and make new discoveries.

Sequencing a genome

Why can't we currently sequence a genome from start to finish?

Once DNA is extracted and sequenced, the sequence needs to be reassembled. We can't currently sequence a genome from start to finish – it has to be broken up into smaller fragments. These small fragments then need to be arranged in the correct order before scientists can start analyzing the genome.



A scheme of sequencing, showing how the sequenced fragments need to be assembled.
Image: [National Human Genome Research Institute](#)



Genomes can be very large: the human genome is 3 gigabases (3 000 000 000 nucleotides) in length. Until recently, the sequencing technologies with which most genomes were sequenced only produced reads that were a few hundred bases long. Newer 'long-read' technologies can now produce sequence reads in the order of 10 000s of bases long, which is an enormous improvement and simplifies the genome-assembly problem a great deal, but it is still far short of chromosome lengths for many organisms.

Sequencing is also not yet 100% accurate, so small errors can occur. To account for this, each base is sequenced multiple times – this means that at the end of sequencing there will be lots of pieces of DNA sequence, which need to be assembled in order. It's like putting together a jigsaw puzzle when you have only part of the picture (if you have existing DNA sequences for comparison). If you don't have existing sequences to compare them with, it's like doing a jigsaw with no picture, and you have to look for sections of sequence overlap to get them in the right order – this stage is called assembly.