# Genomics

**Genomics** is a field that studies the entire collection of an organism’s DNA or genome. It involves sequencing, analyzing, and comparing the information contained within genomes. Since sequencing has become much less expensive and more efficient, vast amounts of genomic information is now available about a wide variety of organisms, but particularly microbes, with their smaller genome size. In fact, the biggest bottleneck currently is not the lack of information but the lack of computing power to process the information!

## Sequencing

**Sequencing**, or determining the base order of an organism’s DNA or RNA, is often one of the first steps to finding out detailed information about an organism. A bacterial genome can range from 130 kilobase pairs (kbp) to over 14 Megabase pairs (Mbp), while a viral genome ranges from 0.859 to 2473 kbp. For comparison, the human genome contains about 3 billion base pairs.

### Shotgun sequencing

**Shotgun sequencing** initially involves construction of a **genomic library**, where the genome is broken into randomly sized fragments that are inserted into vectors to produce a library of clones. The fragments are sequenced and then analyzed by a computer, which searches for overlapping regions to form a longer stretch of sequence. Eventually all the sequences are aligned to give the complete genome sequence. Errors are reduced because many of the clones contain identical or near identical sequences, resulting in good “coverage” of the genome.



Shotgun Sequencing. By Commins, J., Toft, C., Fares, M. A. [[CC BY-SA 2.5](http://creativecommons.org/licenses/by-sa/2.5)], [via Wikimedia Commons](https://commons.wikimedia.org/wiki/File%3AWhole_genome_shotgun_sequencing_versus_Hierarchical_shotgun_sequencing.png)

### Second generation DNA sequencing

**Second-generation DNA sequencing** uses **massively parallel methods**, where multiple samples are sequenced side-by-side. DNA fragments of a few hundred bases each are amplified by PCR and then attached to small bead, so that each bead carries several copies of the same section of DNA. The beads are put into a plate containing more than a million wells, each with one bead, and the DNA fragments are sequenced.

### Third- and fourth-generation DNA sequencing

Third-generation DNA sequencing involves the sequencing of single molecules of DNA. Fourth-generation DNA sequencing, also known as “post light sequencing,” utilizes methods other than optical detection for sequencing.

## Bioinformatics

After sequencing, it is time to make sense of the information. The field of **bioinformatics**combines many fields together (i.e. biology, computer science, statistics) to use the power of computers to analyze information contained in the genomic sequence. Locating specific genes within a genome is referred to as **genome annotation**.

### Open Reading Frames (ORFS)

An **open reading frame**or **ORF** denotes a possible protein-coding gene. For double-stranded DNA, there are six reading frames to be analyzed, since the DNA is read in sets of three bases at a time and there are two strands of DNA. An ORF typically has at least 100 codons before a stop codon, with 3’ terminator sequences. A **functional ORF** is one that is actually used by the organism to encode a protein. Computers are used to search the DNA sequence looking for ORFs, with those presumed to encode protein further analyzed by a bioinformaticist.

It is often helpful for the sequence to be compared against a database of sequences coding for known proteins. [GenBank](http://www.ncbi.nlm.nih.gov/Genbank/) is a database of over 200 billion base pairs of sequences that scientists can access, to try and find matches to the sequence of interest.  The database search tool **BLAST** (**basic local alignment search tool**) has programs for comparing both nucleotide sequences and amino acid sequences, providing a ranking of results in order of decreasing similarity.

BLAST Results.

## Comparative Genomics

Once the sequences of organisms have been obtained, meaningful information can be gathered using **comparative genomics**. For this genomes are assessed for information regarding size, organization, and gene content.

Comparison of the genome of microbial strains has given scientists a better picture regarding the genes that organisms pick up. A group of multiple strains share a **core genome**, genes coding for essential cellular functions that they all have in common. The **pan genome** represents all the genes found in all the members of species, so provides a good idea of the diversity of a group. Most of these “extra” genes are probably picked up by horizontal gene transfer.

Comparative genomics also shows that many genes are derived as a result of gene duplication. Genes within a single organism that likely came about because of gene duplication are referred to as **paralogs**. In many cases one of the genes might be altered to take on a new function.  It is also possible for gene duplication to be found in different organisms, as a result of acquiring the original gene from a common ancestor. These genes are called **orthologs**.

## Functional Genomics

The sequence of a genome and the location of genes provide part of the picture, but in order to fully understand an organism we need an idea of what the cell is doing with its genes. In other words, what happens when the genes are expressed? This is where **functional genomics** comes in – placing the genomic information in context.

The first step in gene expression is transcription or the manufacture of RNA. **Transcriptome**refers to the entire complement of RNA that a cell can make from its genome, while **proteome** refers to all the proteins encoded by an organisms’ genome, in the final step of gene expression.

### Microarrays

**Microarrays**or **gene chips** are solid supports upon which multiple spots of DNA are placed, in a grid-like fashion. Each spot of DNA represents a single gene or ORF. Known fragments of nucleic acid are labeled and used as **probes**, with a signal produced if binding occurs. Microarrays can be used to determine what genes might be turned on or off under particular conditions, such as comparing the growth of a bacterial pathogen inside the host versus outside of the host.

### Proteomics

The study of the proteins of an organism (or the proteome) is referred to as **proteomics**. Much of the interest focuses on **functional proteomics**, which examines the functions of the cellular proteins and the ways in which they interact with one another.

One common technique used in the study of proteins is **two-dimensional gel electrophoresis**, which first separates proteins based on their isoelectric points. This is accomplished by using a pH gradient, which separates the proteins based on their amino acid content. The separated proteins are then run through a polyacrylamide gel, providing the second dimension as proteins are separated by size.

**Structural proteomics** focuses on the three-dimensional structure of proteins, which is often determined by protein modeling, using computer algorithms to predict the most likely folding of the protein based on amino acid information and known protein patterns.

### Metabolomics

**Metabolomics** strives to identify the complete set of metabolic intermediates produced by an organism. This can be extremely complicated, since many metabolites are used by cells in multiple pathways.

### Metagenomics

**Metagenomics** or **environmental genomics** refers to the extraction of pooled DNA directly from a specific environment, without the initial isolation and identification of organisms within that environment. Since many microbial species are difficult to culture in the laboratory, studying the **metagenome** of an environment allows scientists to consider all organisms that might be present. Taxa can even be identified in the absence of organism isolation using nucleic acid sequences alone, where the taxon is known as **phylotype**.

### *Key Words*

genomics, sequencing, shotgun sequencing, genomic library, second generation DNA sequencing, massively parallel methods, third- and fourth-generation DNA sequencing, bioinformatics, genome annotation, open reading frame/ORF, functional ORF, GenBank, BLAST/basic local alignment search tool, comparative genomics, core genome, pan genome, paralog, ortholog, functional genomics, transcriptome, proteome, microarray/gene chips, probe, proteomics, functional proteomics, two-dimensional gel electrophoresis, structural proteomics, metabolomics, metagenomics/environmental genomics, metagenome, phylotype.

### *Study Questions*

1. What does the field of genomics encompass?
2. What is shotgun sequencing and how does this allow for the complete sequencing of an organism’s genome?
3. What are the basic differences among 2nd, 3rd, and 4th generation sequencing?
4. What is an open reading frame and how can scientists use it to determine information about a genome and its products?
5. How does functional genomics differ from comparative genomics? What are the tools used in functional genomics and what information can be obtained from each?

## LICENSE

[General Microbiology](https://open.oregonstate.education/generalmicrobiology) by Linda Bruslind is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/), except where otherwise noted.