# Genomics: An introduction to EMBL-EBI resources

Giulietta M. Spudich [1]

- DNA & RNA
- Beginner
- 0.5 hour

This course will introduce the field of genomics and provide an overview of the different resources at the EBI which are related to this field. The course is aimed at graduates in the life sciences or related subjects, with an interest in using bioinformatics resources to solve problems related to genome biology.

Learning objectives:

- Know what genomics is and the genomics resources available at the EBI
- Know what Ensembl and Ensembl Genomes are
- Know what the European Nucleotide Archive (ENA) is
- Know what the European Genome–phenome Archive (EGA) is
- Know what the difference between Ensembl and Ensembl Genomes and when to use them

## What is genomics?

Genomics is the study of whole genomes of organisms, and incorporates elements from genetics. Genomics uses a combination of recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyse the structure and function of genomes. It differs from 'classical genetics' in that it considers an organism's full complement of hereditary material, rather than one gene or one gene product at a time. Moreover, genomics focuses on interactions between loci and alleles [2] within the genome and other interactions such as epistasis [3], pleiotropy [4] and heterosis [5](Figure 1.1). Genomics harnesses the availability of complete DNA sequences for entire organisms and was made possible by both the pioneering work of Fred Sanger [6] and the more recent next-generation sequencing technology.

Fred Sanger's group established techniques of sequencing, genome mapping, data storage, and bioinformatic analyses in the 1970s and 1980s. This work paved the way for the human genome project in the 1990s (1 [7]), an enormous feat of global collaboration that culminated in the publication of the complete human genome sequence [8] in 2003. Today, next-generation sequencetechnologies have led to spectacular improvements in the speed [8], capacity and affordability of genome sequencing. Moreover, advances in bioinformatics [9] have enabled hundreds of life-science databases and projects that provide support for scientific research. Information stored and organised in these databases can easily be searched, compared and analysed. We will explore some key genomics resources in the following sections of this course.

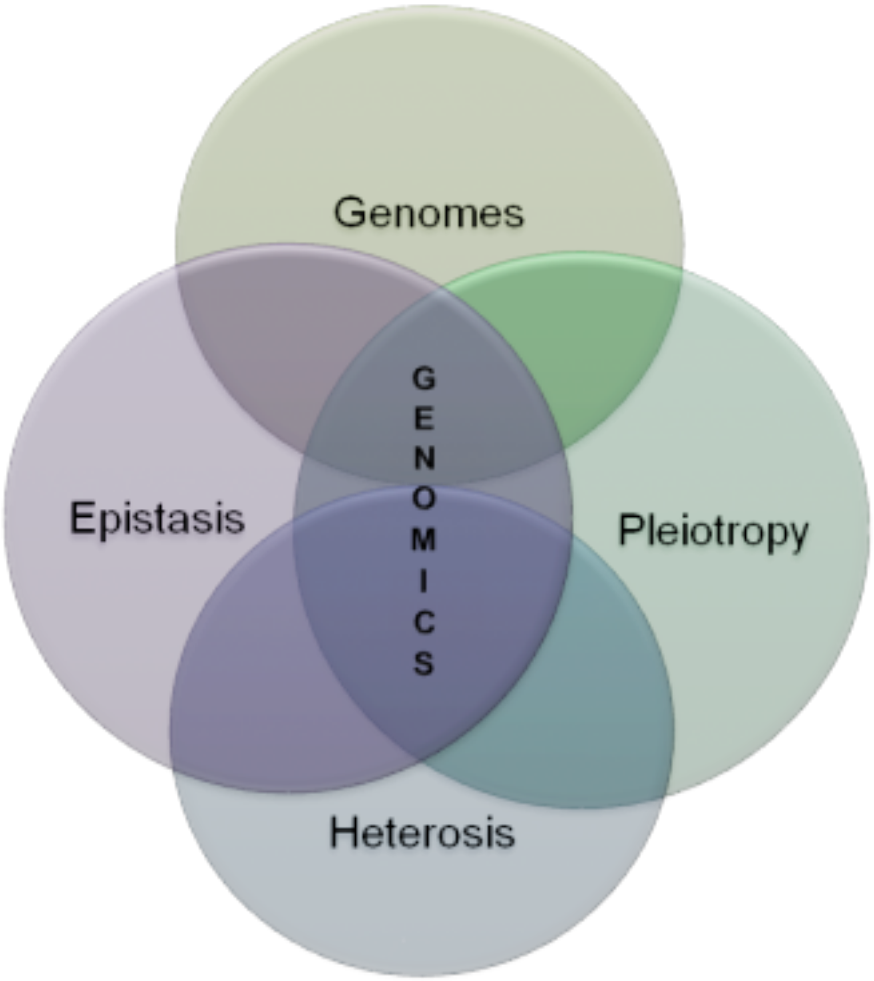


**Figure 1.1** Genomics studies the genomes of whole organisms and other intragenomic interactions.

## Genomics resources at the EBI

This section will introduce you to the genomics resources at the EBI (Figure 1.2).

The resources include:

- Ensembl **[10] and** Ensembl Genomes **[11]**
  Ensembl [12] provides a genome browser with information such as gene sequence, splice variants, sequence polymorphisms, comparative analysis and sequences implicated in gene regulation [13] for mostly vertebrate genomes.

  Ensembl Genomes [14] extends the Ensembl system to invertebrate metazoa [15], protists, bacteria, plants and fungi.

- **European Nucleotide Archive (ENA)** [16] provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation [17]. ENA consists of three main databases: the Sequence Read Archive [18] (SRA), the Trace Archive and EMBL-Bank [19].

- **European Genome-phenome Archive (EGA)** [20] is a service for permanent archiving and sharing of all types of genetic and phenotypic data resulting from biomedical research projects.

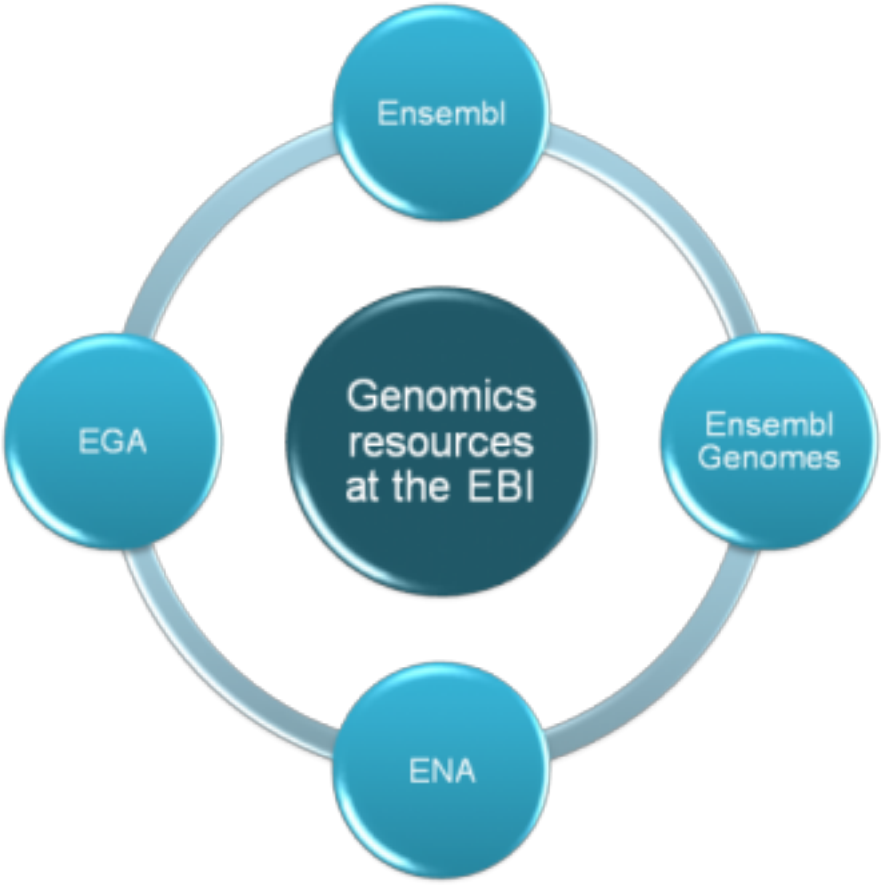


**Figure 1.2** The radial cycle graph shows what genomics resources are available at the EBI.

## Ensembl and Ensembl Genomes

The Ensembl project [21], developed jointly by the EBI [22] and the Wellcome Trust Sanger Institute [23], has been used for the annotation [17], analysis and display of vertebrate genomes since 2000 (2 [7]). Its aim is to enable genomic science by providing high quality, integrated annotation on chordate [24] and selected eukaryotic genomes within a consistent and accessible infrastructure (3 [7]). All supported species include comprehensive, evidence-based gene annotations and a selected set of genomes includes additional data focused on variation, comparative, evolutionary, functional and regulatory annotation. The most advanced resources are provided for key species including human, mouse, rat and zebrafish, reflecting the popularity and importance of these species in biomedical research. As of Ensembl release [25] 64 (September 2011), 58 species are supported.

Since 2009, the Ensembl site has been complemented by the creation of five new sites, for bacteria, protists, fungi, plants and invertebrate metazoa [15], in the Ensembl Genomes project [14] developed at the EBI (4 [7]). Both Ensembl and Ensembl Genomes are accessible through their web browsers, the BioMart tool [26] and Perl APIs [27].

You can use Ensembl if you want to find out about the sequence of a gene, homologues in other species, variations across populations, or explore chromosomal regions around the gene of interest (Figure 1.3).

You can find more information in 'Where do the data come from? [28]' section in Ensembl: browsing chordate genomes.

**Figure 1.3** Ensembl and Ensembl genomes resources provide a wealth of information on genomes of different species used in biomedical research.

## European Nucleotide Archive (ENA)

The European Nucleotide Archive (ENA) [29] provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation [17] (5 [7]). The ENA consists of three main databases (Figure 1.4): the Sequence Read Archive [18] (SRA), the Trace Archive and EMBL-Bank [19]. It comprises three parts: ENA-Annotation, ENA-Assembly and ENA-Reads. ENA-Annotation contains detailed functional annotation, for example of individual, well characterised coding sequences. ENA Assembly is designed for efficient storage of sequence assemblies. Finally, ENA-Reads is optimised for the efficient storage of sequence trace information (6 [7]).

Data arrive at ENA from a variety of sources. These include submissions of raw data [30], assembled sequences and annotation from small-scale sequencing projects or data coming from the major European sequencing centres and through the exchange with partners in the International Nucleotide Sequence Database Collaboration [31] (INSDC).

You should use ENA to retrieve information on nucleotide sequences of interest, if you want to:

- perform a comprehensive sequence-similarity search,
- if you want raw data from electrophoresis-based sequencing machines (held in the European Trace Archive),
- raw data from next-generation (array-based) sequencing platforms .

Moreover, if you have sequenced a gene or transcript, you can submit your data to EMBL-Bank [32], that constitutes Europe's primary nucleotide sequence resource and is part of the ENA.

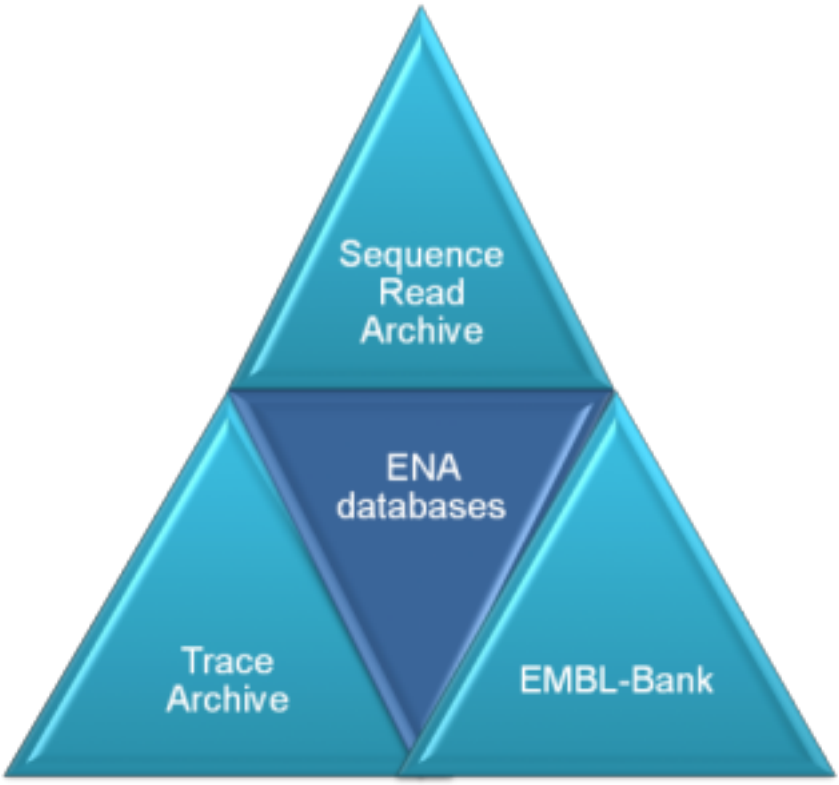


**Figure 1.4** Three main databases contribute to the European Nucleotide Archive as shown in the figure.

## European Genome-phenome Archive (EGA)

The European Genome-phenome Archive (EGA) [33] provides a service for archiving, processing and disseminating all types of potentially identifiable genetic and phenotypic human data at EMBL-EBI. The EGA stores processed and raw data [30] from many types of experiments including case control studies, cancer sequencing and population studies (7 [7]). Available data types include single nucleotide polymorphism [34] (SNP [34]) and copy number variation (CNV) genotypes [35], whole genome sequence and phenotype [36] data. Each data type is stored at the EGA using methods designed to ensure that the storage and distribution is done in accordance with the consent and confidentiality agreements that the research participants agreed to at the time of entry into the study. Strict protocols govern how information is managed, stored and distributed by the EGA project.

You should use the EGA if you want to find genetic and phenotypic data resulting from biomedical research projects. EGA contains a large number of whole genome association studies. Access to these studies, which contain information on identifiable individuals, is controlled by a dataset access committee (DAC) and you need to obtain an account [37] to gain access to these data. Finally, you can submit [38] to the EGA to archive data on subjects who have consented to the use of their individual genetic data for biomedical research, but not for unlimited public data release (Figure 1.5).


**Figure 1.5** The EGA contains exclusive data collected from individuals whose consent agreements authorize data release only for specific research use.

## Summary

- **Genomics** is the study of the entire genome of organisms, in contrast to genetics, in which single genes are studied;
- **The genomics resources at the EBI** include Ensembl [10], Ensembl Genomes [11], ENA and EGA [39];
- **Ensembl and Ensembl Genomes** provide databases with information on genomes from different species. The difference between the two is that Ensembl collects information on vertebrate genomes; Ensembl Genomes extends these information to invertebrate metazoa [15], protists, bacteria, plants and fungi;
- **The** European Nucleotide Archive **[40] (ENA)** provides a comprehensive, accessible and publicly available repository for nucleotide sequence data;
- **The European Genome-phenome Archive (EGA)** is a permanent repository for all types of potentially identifiable genetic and phenotypic data from biomedical research projects. The EGA contains data collected from individuals who have given consent for its use in research, but not for open public distribution.

Want to know more, why not watch our recent 'Resources for Genomics' webinar below.

## Quiz: Genomics

| | |
|---|---|
| **Questions:** | *5* |
| **Attempts allowed:** | *Unlimited* |
| **Available:** | *Always* |
| **Pass rate:** | *75 %* |

**Backwards navigation:** *Allowed*

# Your feedback

Please tell us what you thought about this webinar. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

# Learn more

## Find out more

- yourgenome.org [41] is a website intended to help people understand genetics and genomic science and the implications for everyone.
- NCBI Genome [42] The National Center for Biotechnology Information provides access to biomedical and genomic information.
- Genome New Network [43] The Genome News Network (GNN) produces an online magazine that covers important developments in genomics research around the world.
- Ensembl [12] The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.
- Ensembl Genomes [44] Powered by Ensembl, Ensembl Genomes allows you to browse genomes of metazoa [15] not represented in Ensembl, protists, bacteria, plants and fungi.
- The European Genome-phenome Archive (EGA) [20] The EGA repository allows you to explore datasets from numerous genotype [35] experiments, supplied by a range of data providers.
- European Nucleotide Archive [16] ENA provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation [17].

### Books

- Introduction to genomics by Arthur Lesk. Oxford University press
- Genome: The autobiography of a species in 23 Chapters by Matt Ridley. Harper Collins publisher
- Genomes by Terence A Brown. Oxford: Wiley-Liss 2nd edition

## Recommended courses

On-line

Ensembl:Browsing the chordate genomes [45]

EMBL-EBI Hands on Training Programme Courses:

http://www.ebi.ac.uk/training/course/hands-training-ebi-next-generation-sequencing-workshop [46]

# References

1. Bentley,D.R., Balasubramanian,S., Swerdlow,H.P., Smith,G.P., Milton,J., Brown,C.G., Hall,K.P., Evers,D.J., Barnes,C.L., Bignell,H.R. et al. (2008) Accurate whole human genome sequencing using reversible terminator chemistry. [47] Nature, 456, 53–59.

2. Flicek,P., Aken,B.L., Ballester,B., Beal,K., Bragin,E., Brent,S., Chen,Y., Clapham,P., Fairley,S. et al. (2010) Ensembl's 10th year. [48] Nucleic Acids Res.

3. International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. [49] Nature, 432, 695–716.

4. Kersey,P.J., Lawson,D., Birney,E., Derwent,P.S., Haimel,M., Herrero,J., Keenan,S., Kerhornou,A., Koscielny,G., KaÅNhaÅN ri,A. et al. Ensembl Genomes: Extending Ensembl across the taxonomic space. [50] Nucleic Acids Res.

5. Li,W., McWilliam,H., Richart de la Torre,A., Grodowski,A., Benediktovich,I., Goujon,M., Nauche,S. and Lopez,R. (2010) Non redundant patent sequence databases with value-added annotations at two-levels. [51] Nucleic Acids Res.

6. Leinonen,R., Akhtar,R., Bonfield,J., Bower,L., Corbett,M., Cheng,Y., Demiralp,F., Faruque,N., Goodgame,N., Gibson,R. et al. (2010) Improvements to services at the European Nucleotide Archive. [52] Nucleic Acids Res.

7. Capture and release. [53] Nature Genetics 43,921 (2011)

# Contributors

**Annalisa Mupo**

EMBL-EBI

Visitor, Functional Genomics [54]

Annalisa Mupo is a visiting scientist at the European Bioinformatics Institute in Cambridge. She earned her PhD in Molecular Medicine - Human Genetics from the European School of Molecular Medicine (SEMM) [55] in 2008 working on a genetic disease named DiGeorge Syndrome.

**Giulietta Spudich**

EMBL-EBI

Outreach project leader, PANDA

Giulietta is the outreach project leader for the Protein and Nucleotide Database group (PANDA) at the EBI, and has been working on Ensembl [56] workshops, help and documentation since 2006. She has a PhD from the Department of Molecular and Cell biology at the University of California, Berkeley, and completed her postdoctoral research in biochemical studies of Myosin VI at the MRC-LMB in Cambridge, UK. She strives to support wet-lab biologists in the fantastic bioinformatics resources at the EBI.

**Source URL:** https://www.ebi.ac.uk/training/online/course/genomics-introduction-ebi-resources

**Links**
[1] https://www.ebi.ac.uk/training/online/trainers/giulietta
[2] https://www.ebi.ac.uk/training/online/glossary/allele
[3] https://www.ebi.ac.uk/training/online/glossary/epistasis
[4] https://www.ebi.ac.uk/training/online/glossary/pleiotropy
[5] https://www.ebi.ac.uk/training/online/glossary/heterosis
[6] http://www.sanger.ac.uk/about/people/biographies/fsanger.html
[7] https://www.ebi.ac.uk/training/online/course/genomics-introduction-ebi-resources/references
[8] http://www.yourgenome.org/downloads/animations.shtml
[9] http://nar.oxfordjournals.org/content/39/suppl_2/W3.full
[10] https://www.ebi.ac.uk/training/online/glossary/ensembl
[11] https://www.ebi.ac.uk/training/online/glossary/ensembl-genomes
[12] http://www.ensembl.org/index.html
[13] https://www.ebi.ac.uk/training/online/glossary/gene-regulation
[14] http://www.ensemblgenomes.org
[15] https://www.ebi.ac.uk/training/online/glossary/metazoa
[16] http://www.ebi.ac.uk/ena/
[17] https://www.ebi.ac.uk/training/online/glossary/annotation
[18] https://www.ebi.ac.uk/training/online/glossary/sequence-read-archive
[19] https://www.ebi.ac.uk/training/online/glossary/embl-bank
[20] http://www.ebi.ac.uk/ega/
[21] http://www.ensembl.org
[22] https://www.ebi.ac.uk/training/online/../../
[23] http://www.sanger.ac.uk/
[24] https://www.ebi.ac.uk/training/online/glossary/chordate
[25] https://www.ebi.ac.uk/training/online/glossary/ensembl-release
[26] http://www.ensembl.org/biomart/martview
[27] http://www.ensembl.org/info/docs/api/index.html
[28] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/what-ensembl
[29] http://www.ebi.ac.uk/ena/home
[30] https://www.ebi.ac.uk/training/online/glossary/raw-data
[31] http://www.insdc.org/
[32] http://www.ebi.ac.uk/embl/

[33] http://www.ebi.ac.uk/ega/home
[34] https://www.ebi.ac.uk/training/online/glossary/single-nucleotide-polymorphism
[35] https://www.ebi.ac.uk/training/online/glossary/genotype
[36] https://www.ebi.ac.uk/training/online/glossary/phenotype
[37] https://www.ebi.ac.uk/ega/node/67
[38] https://www.ebi.ac.uk/ega/submission
[39] https://www.ebi.ac.uk/training/online/glossary/ega
[40] https://www.ebi.ac.uk/training/online/glossary/european-nucleotide-archive
[41] http://www.yourgenome.org/dgg/
[42] http://www.ncbi.nlm.nih.gov/sites/genome
[43] http://www.genomenewsnetwork.org/
[44] http://www.ensemblgenomes.org/
[45] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes
[46] http://www.ebi.ac.uk/training/course/hands-training-ebi-next-generation-sequencing-workshop
[47] http://europepmc.org/abstract/MED/18987734
[48] http://europepmc.org/abstract/MED/19906699
[49] http://europepmc.org/abstract/MED/15592404
[50] http://europepmc.org/abstract/MED/19884133
[51] http://europepmc.org/abstract/MED/19884134
[52] http://europepmc.org/abstract/MED/19906712
[53] http://europepmc.org/abstract/MED/21956386
[54] https://www.ebi.ac.uk/training/online/glossary/functional-genomics
[55] http://www.semm.it/
[56] https://www.ebi.ac.uk/training/online/glossary/term/66