

# Unlocking the potential of big data in biology: Applications of bioinformatics

**Anna Swan**

**Senior Scientific Training Officer - Digital Learning**

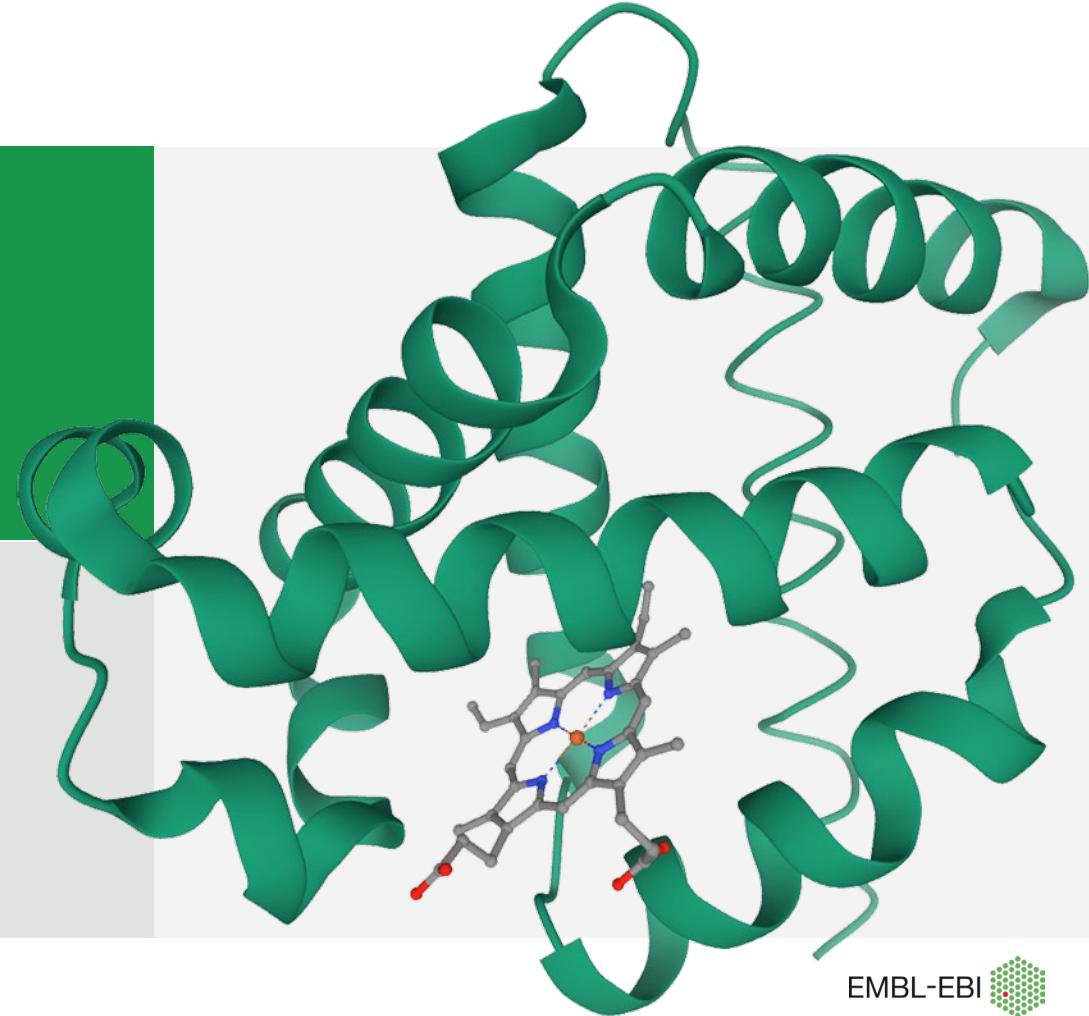
Wellcome Connecting Science - Genome Academy, 11 April 2024  
[bit.ly/GA-bioinformatics-2024](https://bit.ly/GA-bioinformatics-2024)





# Bioinformatics

- Bioinformatics is the study of biology (bio) with computers (informatics).
- Involves storing, managing and analysing huge datasets.





# What is EMBL-EBI?

- The home of big data in biology
- One of the six sites of the European Molecular Biology Laboratory (EMBL)
- Intergovernmental organisation



# The European Molecular Biology Laboratory



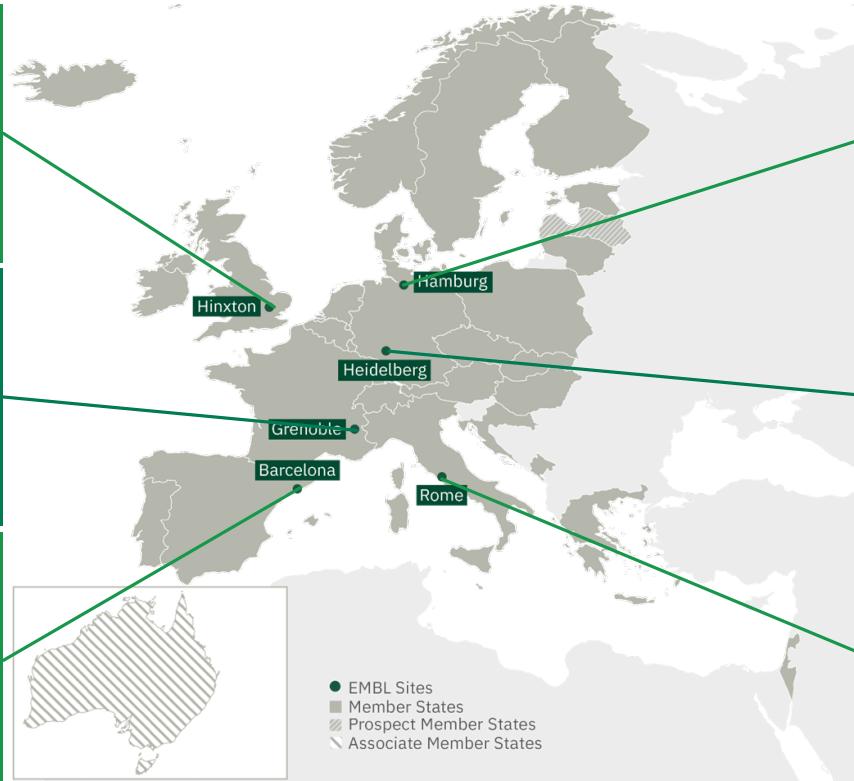
EMBL-EBI  
Bioinformatics



Grenoble  
Structural biology



Barcelona  
Tissue biology  
and disease  
modelling



Hamburg  
Structural biology



Heidelberg  
Life sciences



Rome  
Epigenetics  
and neurobiology





# What does EMBL-EBI do?



Provide data resources for the life sciences

Perform excellent research

Train the next generation of scientists

Work with the private sector

Coordinate bioinformatics in Europe

# Data resources at EMBL-EBI

The image shows two web pages side-by-side. On the left is the Ensembl Human genome browser for the GRCh38.p14 genome build. The main focus is the gene page for BRCA2 (ENSG00000139618), located at 13:32,315,086-32,400,268. The page includes a navigation menu, a summary table, and various links for protein analysis (UniProt, BLAST, Align, Peptide search, ID mapping, SPARQL) and other genomic features like Secondary Structure, Comparative Genomics, and Ontologies. On the right is the AlphaFold Protein Structure Database, developed by Google DeepMind and EMBL-EBI. The homepage features a large blue background with the text "AlphaFold Protein Structure Database" and "Developed by Google DeepMind and EMBL-EBI". It includes a search bar, examples of search terms, and a statement about its purpose: "AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research."

Location: 13:32,315,086-32,400,268

Gene: BRCA2 ENSG00000139618

Gene-based displays

Summary

- Splice variants
- Transcript comparison
- Gene alleles

Sequence

- Secondary Structure
- Comparative Genomics
- Genomic alignments
- Gene tree
- Gene gain/loss tree
- Orthologues
- Paralogues

Ontologies

- GO: Biological process
- GO: Molecular function
- GO: Cellular component
- Phenotypes

Genetic Variation

- Variant table
- Variant image
- Structural variants

Gene expression

Protein

- Molecular interactions
- Regulation
- External references
- Supporting evidence

ID History

- Gene history

Configure this page

Custom tracks

Export data

Share this page

Bookmark this page

Genes (Comprehensive set from GENCODE)

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search Help

A2VEY9 · APP\_DROME

Protein<sup>1</sup> Palmitoyltransferase app

Gene<sup>1</sup> app

Status<sup>1</sup> UniProtKB reviewed

Organism<sup>1</sup> Drosophila melanogaster

Home About FAQs Downloads API

AlphaFold Protein Structure Database

Entry Variant viewer Feature viewer

BLAST Align Download Add

Function<sup>1</sup> Palmitoylates Dihis which is required for...

Catalytic activity<sup>1</sup> Rhea 36683 hexadecanoyl-CoA EC 2.3.1.225

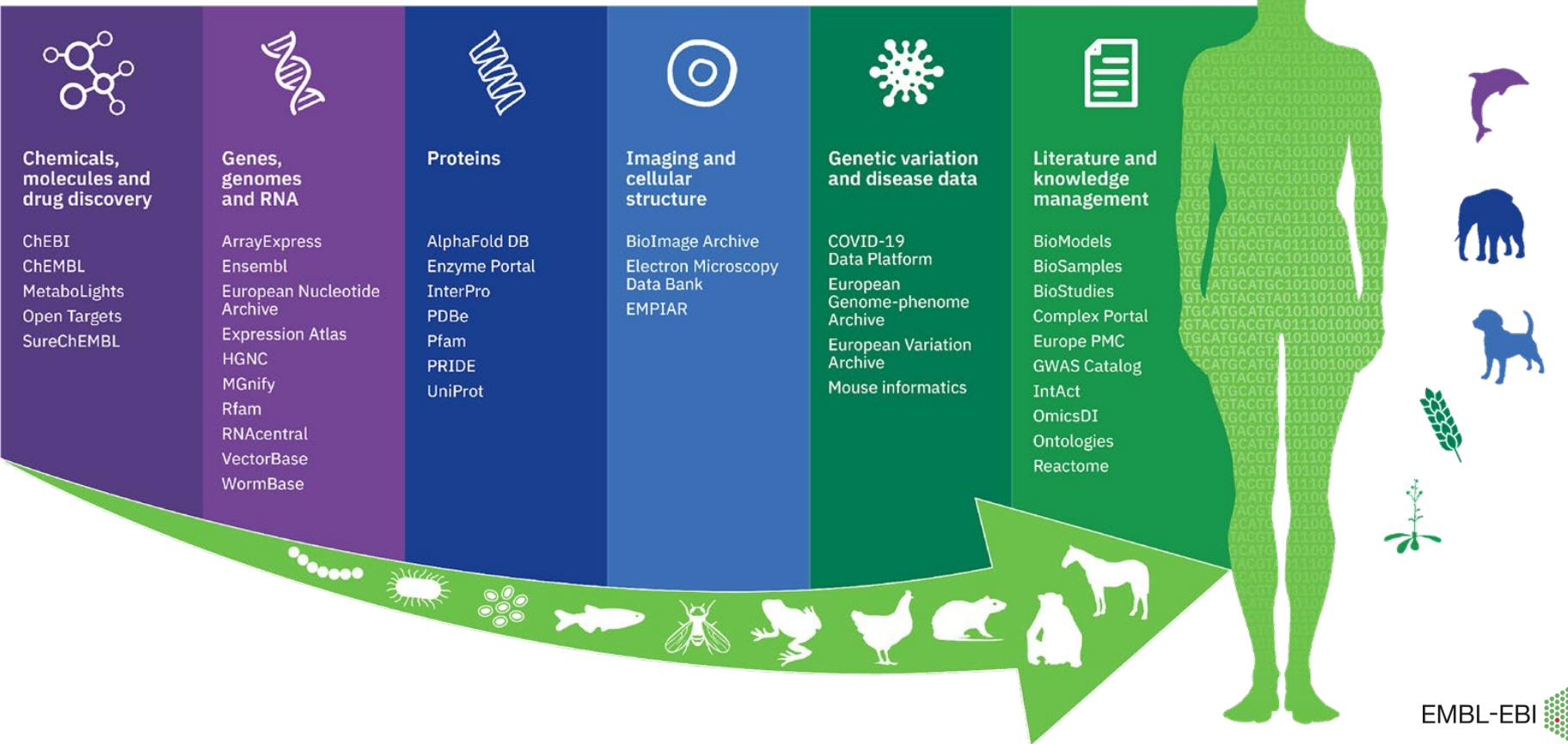
Search for protein, gene, UniProt accession or organism or sequence search Examples: MENFQKVEKIGEGTYGV... Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli

See search help Go to online course

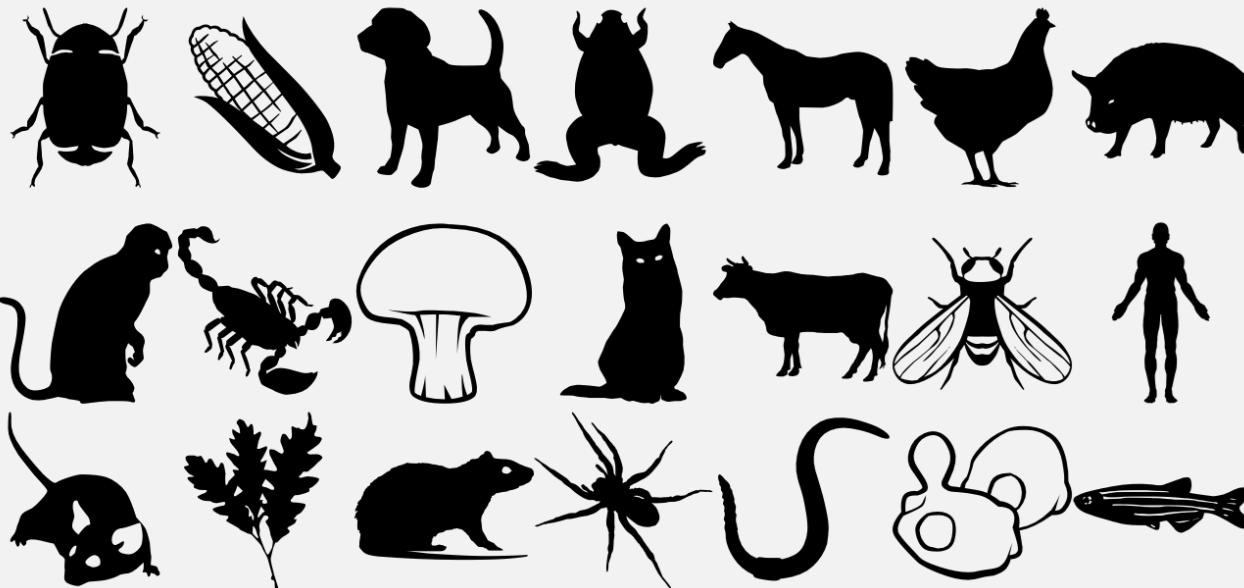
AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research.



# Data resources at EMBL-EBI



# Data for many species





# Searching for a gene, protein or chemical

The image shows two side-by-side screenshots of the EMBL-EBI website. On the left is the main EMBL-EBI homepage, featuring a teal header with the text "EMBL's European Bioinformatics Institute" and "EMBL-EBI". Below this is the tagline "Unleashing the potential of big data in biology". A search bar contains the placeholder "Find a gene, protein or chemical" with a blue outline, and a "Search" button next to it. Below the search bar are links for "All" and "Example searches: blast keratin bft1 | About EBI Search". At the bottom of the homepage are buttons for "Find data resources" and "Submit data". On the right is the "EBI Search" interface, which also has a teal header with the text "EMBL's European Bioinformatics Institute" and "EBI Search". Below this is the tagline "Access all EMBL-EBI resources". The search bar contains the query "mstn" and a "Search" button. Below the search bar are links for "Advanced search" and "Examples: VAV\_HUMAN, tp53, Sulston...". The main content area displays the search results for "mstn", showing 41 results out of 8,548 in All results. It includes a "Filter your results" sidebar with categories like Source, Gene and protein summaries, and Samples & ontologies. The "Gene and protein summaries" section lists "Myostatin" (MSTN) from Human (Homo sapiens) and "Myostatin" (Mstn) from House Mouse (Mus musculus). The "Samples & ontologies" section lists "Reporter vector pGL3-MSTN-3.8kb". The footer of the page includes a link to "https://www.ebi.ac.uk/" and the EMBL-EBI logo.

EMBL's European Bioinformatics Institute  
**EMBL-EBI**  
Unleashing the potential of big data in biology

Find a gene, protein or chemical All

Example searches: blast keratin bft1 | About EBI Search

Find data resources → Submit data →

Latest news →

Funding announcement

UKRI funding supports EMBL-EBI infrastructure upgrade in response to increasing data demand

30 Nov 2023

The Global Biodiversity enabling biodiversity worldwide

29 Nov 2023

EBI Search  
Access all EMBL-EBI resources

mstn

Search Examples: VAV\_HUMAN, tp53, Sulston... Advanced search

Search results for **mstn**

Showing 41 results out of 8,548 in All results

Give us feedback on these results

Filter your results

Source

All results (8,548)

- Genomes & metagenomes (369)
- Nucleotide sequences (3,205)
- Protein sequences (1,674)
- Macromolecular structures (667)
- Bioactive molecules (2)
- Gene expression (181)
- Diseases (24)
- Molecular interactions (12)
- Gene-Disease Associations (592)
- Reactions & pathways (12)
- Protein families (45)
- Protein expression data (2)
- Enzymes (1)
- Literature (1,475)
- Samples & ontologies (287)

Gene and protein summaries (2 results • includes expression, structures, literature...)

Myostatin  
MSTN (ENSG00000138379)  
Human (Homo sapiens)

Myostatin  
Mstn (ENSMUSG00000026100)  
House Mouse (Mus musculus)

Samples & ontologies (287 results)

Source: Taxonomy (ID: 1219472)

Reporter vector pGL3-**MSTN-3.8kb**

<https://www.ebi.ac.uk/>



# Data resources at EMBL-EBI

107 million requests to our data resources on an average day

- 1 Scientists generate data, make discoveries
- 2 Deposit with EMBL-EBI on publication
- 3 We archive and share data with global collaborators and all scientists



- 4 We classify, enrich, combine and analyse
- 5 We distribute both raw and “value added” data resources
- 6 Scientists design new experiments on basis of shared global knowledge

# What is open data?

- Open data can be freely used, re-used and redistributed by anyone.
- When research data is open others can use it to ask new questions and get new insights.
- Open data saves repeating experiments
- Open data drives new discoveries.
- EMBL-EBI data resources are open data.





# We don't wear lab coats

Biologists, physicists  
mathematicians

Software engineers

Biocurators



Bioinformaticians

Data wranglers

And more!

# What skills are needed for bioinformatics?

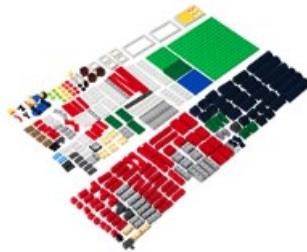
1 Data Collection



2 Data Preparation



3 Data Visualization



4 Data Analysis



5 Data Storytelling



Organisational

Problem-solving

Creativity and Experimentation

Communication



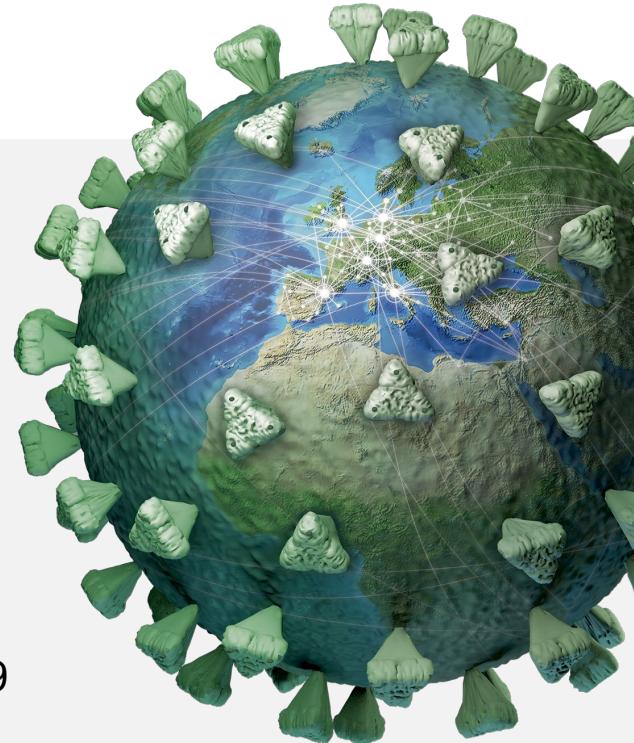
# Pandemic preparedness

Data science was essential in the COVID-19 pandemic.

EMBL-EBI supported the pandemic response:

- Set up the COVID-19 Data Portal to access SARS-CoV-2 molecular and genomic data from all over the world
- Supported countries to set up data sharing platforms
- Revealed insights on new ‘variants’
- Analysed molecular causes of different immune responses
- Identified existing drugs that could be used to treat COVID-19

EMBL-EBI and collaborators are helping to improve European pandemic preparedness.



# Sustainable food production

Bioinformatics helps to feed a growing population in a changing climate.

- Plant genomics – identify which species will be most tolerant to drought and pests while still providing optimum nutrition
- Pests and pollinators – genomics can inform strategies for dealing with pests while protecting pollinators
- Precision breeding – linking genes to traits, farmers and breeders can make food production more sustainable





# Biodiversity conservation

Bioinformatics helps us

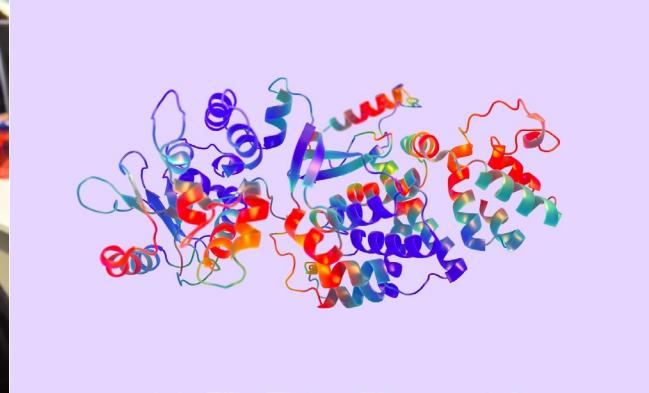
- understand and protect biodiversity
- develop clean technologies to reduce environmental pollution

## Darwin Tree of Life project

- Sequence 77,000 eukaryotic species in Britain and Ireland



# A growing field



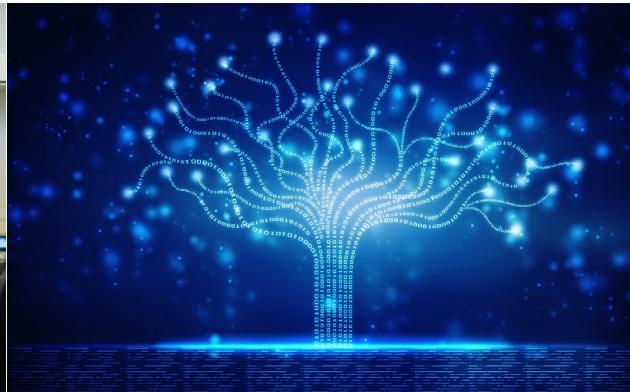
**100 million**  
requests to our websites  
on an average day

**2 million scientists**  
access our websites  
every month

**7 out of 10 users say**  
“not having access to EMBL-EBI data resources would have  
a major impact  
on my work”



# Want to learn more and develop your skills?



[EMBL's TeachingBASE](#)

[Tree of Life project tutorial](#)

[EMBL-EBI tutorials -  
Bioinformatics for the terrified](#)



# Thank you

<https://www.ebi.ac.uk/>

