10th Anniversary Treasure Hunt

In 2017 the EMBL-EBI Training Programme is celebrating 10 amazing years of providing onsite, offsite and online training in bioinformatics.

To help celebrate this anniversary, we designed a treasure hunt to give you a taster of the free resources and (sometimes quirky) data that EMBL-EBI has on offer. This year also happened to be the UN year for “Sustainable tourism for development” and so we adopted travel as our theme.

There are 10 clues to help you navigate through the various EMBL-EBI resources (locations) to collect tourism related data.

Learning objectives:

- Cross-domain
- Beginner
- 0.5 hour
About EMBL-EBI resources

At EMBL-EBI we have a huge variety of tools, resources and databases for sharing, accessing and analysing biological data. They cover many different topics from literature and ontologies, to systems, genes, genomes, proteins, metabolites and more. Importantly, these services are interlinked which helps you explore biological systems more easily.

![EMBL-EBI's resources and databases](image)

**Figure 1** EMBL-EBI's resources and databases

You can explore all of the EMBL-EBI's tools, resources and databases on our [services page](https://www.ebi.ac.uk/training/online) [14].

In [Train online](https://www.ebi.ac.uk/training/online) [15] you will find quick tours, tutorials and webinars that introduce each of their resources and show you how to use them.

The tools and resources can be accessed online or programmatically, the latter of which is very useful when accessing large volumes of data and building analytic workflows.

EMBL-EBI web services provides a [list of tools and resources that can be accessed programmatically](https://www.ebi.ac.uk/training/online) [16] with links to the relevant help pages and documentation.
Travel has always been a part of human nature but our ancestors did not seem to travel much during the Ice Age. Studies show that prehistoric humans migrated into Europe about 45,000 years ago but only moved to the north of the continent 26,000 years later, after the retreat of the ice sheets\textsuperscript{1} \cite{17}. Recent research has shown that all individuals, who lived in Europe between ~37,000 and ~14,000 years ago are descended from a single population - the one big “family” that we have in common.

Clue

Find the paper with the title “Genetic history of Ice Age Europe” in Europe PMC \cite{18}.

What is the Open Researcher and Contributor ID (ORCID) of the co-author who works at the Toulouse university?

About Europe PMC

Europe PMC is a global open repository, providing access to biomedical literature, including articles, books, patents, and clinical guidelines. More than 32 million abstracts (27 million from PubMed \cite{19}) and 4 million full-text articles can be freely accessed through Europe PMC in a single search.

Learn more in Europe PMC: Quick tour \cite{20}.

Need a hint?

Look out for the symbol.
Nice moves (genetically speaking): Ensembl

Since the Ice Age, humans have moved all over the world. We can track the historic migration of human populations by looking at genetic variation data for modern human populations, such as that generated by the 1000 Genomes project [21][22].

The 1000 Genomes project [23] was a sequencing project to call variants in 2504 individuals from 26 populations around the world, including seven African or African-descended populations. Using Ensembl [24] you can explore all the variants identified in this project, including their frequencies and the genotypes [25] of the individuals.

Image source: History of human migration, Wikipedia [26].

Clue

Use Ensembl [24] to find the variant [27] rs555128980. There is an allele [28] in this location in which thymine (T) is substituted for a cytosine (C). Only one individual in the 1000 Genomes project has this allele. What is the sample ID for the individual with the C|T genotype [25]?

About Ensembl

Ensembl [24] provides a genome browser that acts as a single point of access for annotated genomes. Genes, genetic variants and regulatory features are all annotated onto genomes, with comparative genomics analyses carried out on genes and genomes.

Learn more in our Ensembl: Quick tour [29].

Need a hint?

Travelling west in AFR leads to GWD.
Cool destinations (polar exploration): EBI Metagenomics

The Arctic has become a popular destination for tourists who go to see the unique landscapes and animals\(^3\) [30]. Microorganisms are also an important part of this environment and contribute to global biodiversity.

Clue

Hunt for the study on ‘arctic winter marine ecosystem’ in the EBI Metagenomics portal [31]. What proportion of the prokaryotic community in arctic winter marine ecosystems is made up of archaea (to the nearest whole percent)?

About EBI Metagenomics

EBI metagenomics [31] is a freely available hub for the analysis and exploration of metagenomic, metatranscriptomic, amplicon and assembly datasets. The resource provides rich functional and taxonomic analyses of user-submitted sequences, as well as analysis of publicly available metagenomic datasets held within the European Nucleotide Archive. It is one of the largest metagenomics data repositories in the world.

You can learn more about EBI metagenomics in our webinar EBI Metagenomics: analysing and exploring
Need a hint?

Treasure hunting can be taxing.

Getting high (adaptation to high-altitude): ArrayExpress

Machu Picchu, Everest Basecamp, Mont Blanc and the Cerro Torre in Argentina - beautiful, once-in-a-lifetime destinations. For most people, visiting these places will require a period of acclimatisation to avoid altitude sickness.

Some populations, for example people living in the Argentinean Andes, have enviable adaptations to their local environment that we urban people can only dream of. One of these adaptations is to low oxygen at high altitudes.

As with many other population-specific traits, it has long been suspected that adaptation to high altitudes is linked to positively-selected SNPs [33].

Clue

Seek a study in ArrayExpress [34] that investigates genetic and phenotypic variation in people living at high (>3500 m) and low (<500 m) altitudes in Argentina. What is the accession number of the nucleic acid extraction protocol used in this study?
**ArrayExpress** [34] is a database for archiving functional genomics data. In this database, you'll find data on transcriptomics (expression profiling), epigenetics (e.g. ChIP-seq, bisulphite-seq, ATAC-seq) and genomics (e.g. genotyping/SNP-typing). All data come from microarray and/or (NGS) technologies. Data are either directly submitted to ArrayExpress (often pre-published) via submission tool **Annotare** [36], or imported from a similar repository called **Gene Expression Omnibus at NCBI** [37] via an automatic pipeline.

Learn more in **ArrayExpress: discover functional genomics data quickly and easily** [38].

**Need a hint?**

Sometimes treasure is buried. Dig into the study.

**The joys of travel (food) part 1: IntAct**

Food! One of the pleasures of travel is experiencing new cultures and foods. Broad beans (or fava beans) are a staple food source across the world and turn up in many dishes $^{39}$. 

In broad beans (and many other plants) phototropins control how the plant responds to changes in light by triggering various downstream signalling pathways.

**Clue**

Find the broad bean phototropin, Vicia faba phototropin 1a (Vfphot1a) in **IntAct** [40]. What is the **UniProt** [41] accession [35] number of the protein that Vfphot1a interacts with?

**About IntAct**

**IntAct** [42] is a central, public repository where molecular interactions data can be stored and accessed. IntAct contains interaction data from literature curation or direct user submissions. Most of the data refer to protein–protein interactions, but interactions involving other types of molecules, such as small chemical compounds or nucleic acids, can also be found in IntAct. IntAct is freely available and uses an open source database system...
and analysis tools.

You can learn more in IntAct: Molecular interactions at EMBL-EBI [43].

Need a hint?

To B, or not to B. That is the question.

The joys of travel (food) part 2: ChEMBL

Mashed potatoes, baked potatoes, chips, potato noodles, gnocchi, potato bake. Like beans, potatoes are an important ingredient in food the world over from Africa, to Latin America, South East Asia, North America, Australia and Europe.

Potato blight is caused by a fungus, *Phytophthora infestans*, and is a disease that affects potatoes in the wet season. Historically, potato blight (and the failure of potato crops) has led to the starvation of hundreds of thousands of people. It is still a major disease of potatoes to this day. A lot of research goes into finding small molecules [44] with bioactivity against this fungus. Clue

Find all of the compounds with bioactivity against *Phytophthora infestans* in ChEMBL [46]. What is the CHEMBLID of the largest compound (by molecular weight)?

About ChEMBL

ChEMBL is a database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data). The data is abstracted and curated from the primary scientific literature, and cover a significant fraction of the SAR and discovery of modern drugs. We attempt to normalise the bioactivities.
into a uniform set of end-points and units where possible, and also to tag the links between a molecular target and a published assay with a set of varying confidence levels. Additional data on clinical progress of compounds is being integrated into ChEMBL at the current time.

You can learn more about ChEMBL in our ChEMBL walkthrough: webinar [47].

**Need a hint?**

*Phytophthora infestans* is the of your voyage.

Look to the north-east corner. Hunt for ‘Display Bioactivities’ in

**Travel challenges (jet-lag): Complex Portal**

Our heroes got off the plane feeling completely jet-lagged. They knew this was because their transcription factor complexes which regulate the circadian rhythm had not yet adapted to their new time zone.

**Clue**

Use the Complex portal [48] to explore human transcription complexes that are involved in circadian rhythm. What is the UniProt [41] accession [35] number of the cleverly named protein that is common to all these transcription complexes?
About the Complex portal

The Complex Portal [48] is a manually curated, encyclopedic resource of macromolecular complexes from a number of key model organisms, entered into the IntAct molecular interaction database. Data include protein-only complexes, as well as protein—small molecule and protein—nucleic acid complexes.

Learn more in the Complex Portal: Quick tour [49]

Need a hint?

Looking for complexes involved in transcription? The hint is in the name.

Hurry, the clock is ticking!

Perils of travel (Zika virus and global health): PDBe and EMDB

Global travel is not just for people. Microbes and pathogens can also travel internationally leading to global health challenges such as Zika virus.

The Zika virus has been linked to microcephaly and other neurological damage in foetuses. It is spreading so rapidly that the World Health Organization has declared it to be an emergency of international concern [50]. Before 2016, there were no structures of Zika virus protein in the PDB [51], however there are now over 50, highlighting the emphasis on research into this disease.

Clue
Search the [Protein DataBank](https://www.ebi.ac.uk/pdbe) in Europe (PDBe) for Zika virus and find the entry that was published in Science and uses [electron microscopy](https://www.ebi.ac.uk/pdbe) as the experimental method.

What is the accession number for its related structure in the Electron Microscopy DataBank ([EMDB](https://www.ebi.ac.uk/pdbe))?

**About PDBe**

The [Protein Data Bank in Europe](https://www.ebi.ac.uk/pdbe) (PDBe) is a member of the Worldwide Protein Databank (wwPDB), the organisation tasked with the collection, organisation and dissemination of data about biological macromolecular structures. PDBe aims to ‘bring structure to biology’ and provides a comprehensive range of tools and services to make the extensive archive of structural and functional information in the PDB more accessible to the wider biological community. Learn more about PDBe in [PDBe:Quick tour](https://www.ebi.ac.uk/pdbe).

Did you know: you can find out more about the polyprotein encoded by the Zika virus’ single stranded RNA genome in [UniProt](https://www.uniprot.org) and [InterPro](https://www.ebi.ac.uk/interpro).

**Need a hint?**

When searching for treasure you might have to filter out the rubble from the rubies.

---

**Plants in sppaaaacccceee (future travel): Expression Atlas**

What might the future of travel hold? Will we venture into space and how will we sustain ourselves on these journeys?
Science is already trying to figure out how organisms perceive, respond and adapt to spaceflight conditions. Some studies have used the International Space Station to test how plants change their gene expression under spaceflight conditions.

You can explore the data from these and other functional genomics experiments in Expression Atlas.

Clue (to be revised)

Find an experiment in Expression Atlas that uses arrays to study the response of Arabidopsis thaliana roots, hypocotyls and leaves to spaceflight.

Use Expression Atlas to explore the gene-set enrichment results for this experiment.

What is the GO accession number for the ontology with the lowest observed/expected ratio?

About Expression Atlas

Expression Atlas is a resource for finding information about gene and protein expression across different species and biological conditions such as tissues, cell types, developmental stages and diseases. Expression Atlas contains thousands of selected datasets that have been manually curated, annotated to ontologies, checked for high quality and processed using standardised pipelines.

Need a hint?

The "plot" thickens. Go towards the light.

Did you know: you can also use ArrayExpress to access experiments in Expression Atlas?
For the love of travel friends and survival skills: EBI Search

Throughout the year we are featuring a series of ‘Top tens’, in which we share some interesting, historical and surprising facts about the EBI training programme.

So far we’ve featured 10 milestones, 10 things that make our training unique, and 10 things that our trainers need on the road.

Clue

Have a look through our top 10s [63]. Find the animal that one of our trainers has as a travel companion. Use EBI Search [64] to find information about this particular type of feathery friend and a paper on their hidden keys to survival.

What is the PubMed [19] ID (PMID [65]) of this paper?

About EBI-search

EBI-search is a quick and easy way of getting an overview of the data and resources that the EBI has on a particular topic. It can be accessed online from the EMBL-EBI [64] homepage or programmatically using our RESTful webservices interface.

EBI search gives you access to the latest and most up to date nucleotide and protein sequences, structures ranging from chemicals to macro-molecular complexes, gene-expression experiments, binary level molecular interactions, reaction maps, pathway models, functional classifications, biological ontologies, and comprehensive literature libraries.

For more information on EBI search see the help and documentation page [66].

Need a hint?
Check your answers

Did you solve all the clues?

Check your answers against ours (correct as of July 2017):

1. Historic travel: Europe PMC
   Answer: 0000-0002-9159-8030

2. Nice moves (genetically speaking): Ensembl
   Answer: HG02568 from the Gambia

3. Cool destinations (polar exploration): EBI Metagenomics
   Answer: ~10%

4. Getting high (adapting to high-altitude): ArrayExpress
   Answer: P-GSE55175-4

5. The joys of travel (food) part 1: IntAct
   Answer: Q4W5W8

6. The joys of travel (food) part 2: ChEMBL
   Answer: CHEMBL526909
7. Travel challenges (jet-lag): Complex Portal
Answer: O15516

8. Perils of travel (Zika virus and global health): PDBe and EMDB
Answer: EMD-8116

Answer: GO:0005794 or GO:0043565

10. For want of travel friends and survival skills: EBI Search
Answer: PMID:26490794

Contributors

David Armstrong
EMBL-EBI
Scientific Database Curator - Kleywegt team: Protein Data Bank in Europe Scientific Database Curator - Velankar team: PDBe content and integration
Melissa Burke [2]

EMBL-EBI
Scientific Training Officer (e-learning): Training Team

Melissa is the Scientific Training Officer (e-learning) for the Training Team at the EMBL-EBI. She joined the Training Team in July 2016 after having worked as a Scientific Curator for ArrayExpress/Expression Atlas at the EMBL-EBI. She has a PhD in Molecular Parasitology and has worked internationally as a postdoctoral researcher specialising in the functional genomics of infectious diseases.

Laura Huerta [3]

EMBL-EBI
Senior Scientific Curator - Petryszak team: Gene Expression

Laura works as a Senior Scientific Curator to improve the content of functional genomics data in Expression Atlas. She is involved in developing and implementing metadata standards with a particular interest on data integration through ontology annotation. She manages training activities, delivering courses on functional genomics resources at EMBL-EBI and worldwide. She also interacts with software developers to improve user experience of Expression Atlas website. Laura joined EMBL-EBI in 2015 after receiving her PhD in Molecular Biology from the Polytechnic University of Valencia, focused on studying plant development regulated by hormones through the generation and analysis of transcriptomics data.

Maria Levchenko [4]

EMBL-EBI
Europe PMC Community Manager - McEntyre team: Literature services

Maria is the Community Manager for Europe PMC [67], a literature repository and partner in PubMed [19] Central International. In her role Maria engages with all stakeholders of Europe PMC to communicate their feedback to the team of developers so that they can deliver the best tools for scholarly publishing and communication. She has a PhD in Biochemistry from the The University of Göttingen.
Alex Mitchell [5]

EMBL-EBI
Curation co-ordinator, InterPro and EBI Metagenomics

Alex Mitchell is co-ordinator for the InterPro and EBI Metagenomics databases at EMBL-EBI. He obtained his DPhil in pharmacology from the University of Oxford, and was previously employed as a molecular biologist at the Institute of Psychiatry. He moved to the University of Manchester to work on protein sequence analysis and functional classification, before joining EMBL-EBI in 2011.

Sarah Morgan [6]

EMBL-EBI
Training Coordinator

Sarah has a BSc and MSc in Biomedical Sciences, a PhD from Cranfield University and an MA in Healthcare Law and Ethics from the University of Manchester. Over a 10 year academic career her research has focused on tumour biomarker characterisation and cell-surface interaction. Beyond research she took an active role in the provision of postgraduate taught courses for the University as Lecturer in Molecular Medicine, leading the development and direction of a number of MSc Courses including the MSc in Translational Medicine, the MSc in Molecular Medicine, and more recently acting as director for the MSc Programmes in Advanced Biosciences. She has lectured on a variety of biomedical techniques and topics, including ethics and governance of research.
Grace Mugumbate [7]
EMBL-EBI
Scientific Training Officer

Prudence Mutowo [8]
EMBL-EBI
Bioactivity Data Manager - Leach team: ChEMBL

Sandra Orchard [69]
EMBL-EBI
Proteomics Services Team Coordinator

Sandra Orchard leads the Molecular Interaction Team, responsible for developing resources enabling the network analysis of large-scale datasets as well as supplying basic interaction data and information about protein complexes. She is responsible for the production and maintenance of the IntAct Molecular Interaction database and the Complex Portal. She has previously contributed to the annotation of the UniProtKB, InterPro and GOA databases. She also applies her experience to provide hands-on training in several resources, including UniProtKB, InterPro, IntAct and Reactome.
**Emily Perry**

EMBL-EBI
Ensembl Outreach Project Leader

Emily is the Outreach Project Leader for Ensembl: she is responsible for the team that teaches workshops, creates training materials and help pages, manages social media, answers helpdesk queries and aids development of new tools for the resource. Emily started at EMBL-EBI as an Ensembl Outreach Officer in September 2012 and became the Project Leader in March 2015. Before working at EMBL-EBI, Emily did her PhD in molecular biology at the MRC Human Genetics Unit in Edinburgh, then worked for the University of Edinburgh's SCI-FUN group, touring Scottish secondary schools with an interactive science roadshow.

**Sangya Pundir**

EMBL-EBI
User Experience Manager - Martin team: UniProt development

Sangya Pundir is a User Experience (UX) Manager in EMBL-EBI's UniProt team, where she established a user-centred process for the redesign of the world's leading protein resource. To make UniProt easy for researchers to explore, Sangya conducts usability testing and information-gathering methods such as card sorting, contextual studies and workshops. Before she came to EMBL-EBI, Sangya worked at a healthcare consultancy, designing bespoke management systems. She holds an MSc in Biotechnology, Bioprocessing and Business Management from the University of Warwick.
Lorna Richardson [12]

EMBL-EBI
Scientific Curator

Lorna is a scientific curator in the sequence families team at EMBL-EBI. She is responsible for InterPro, Pfam and Genome Properties curation, as well as outreach and training. Prior to joining the sequence families team at the EBI, she worked with the eMouseAtlas project at the University of Edinburgh, curating 3D anatomy and gene expression data. She has been working in the field of biological database curation since 2001.

Amy Tang [13]

EMBL-EBI
Functional Genomics Curation and Training Project Leader

Amy studied Human Genetics at the University of Nottingham for her bachelor degree. She did her PhD at the Brockdorff's lab [70], then at MRC Clinical Sciences Centre / Imperial College London, working on the epigenetic control of X-chromosome inactivation in mouse early development. The project was split 50:50 into wet-lab and bioinformatics, this is when she first got interested in bioinformatics. After a short postdoc in the same lab, she returned to study at Imperial for a MSc in Bioinformatics and Systems Biology, which helped her move laterally from wet-lab to bioinformatics.

Amy now leads the curation efforts for ArrayExpress [34] and Expression Atlas [61] databases in the Gene Expression Team [71], promoting reproducible research by maintaining curation quality, and also by driving the development of the user-friendly ArrayExpress data submission tool Annotare [72]. To allow users make the most out of the databases and to educate researchers on best practices for reproducible research, she also manages all training activities for the Team, delivering most of the hands-on courses on functional genomics resources at EMBL-EBI and worldwide. Through listening to Annotare submitters’ needs and gathering feedback at training courses, she also acts as users’ advocate and works with software developers to improve user experience of ArrayExpress and Expression Atlas websites.

Before joining EMBL-EBI, Amy was a senior bioinformatician in the Ensembl [24] GeneBuild team, curating gene models for key model organisms such as mouse and rat.

Your feedback

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


Source URL: https://www.ebi.ac.uk/training/online/course/10th-anniversary-treasure-hunt

Links
[1] https://www.ebi.ac.uk/training/online/trainers/davida_1
[2] https://www.ebi.ac.uk/training/online/trainers/mburke
[3] https://www.ebi.ac.uk/training/online/trainers/lauhuema
[4] https://www.ebi.ac.uk/training/online/trainers/levchmar_10147
[5] https://www.ebi.ac.uk/training/online/trainers/mitchell
[6] https://www.ebi.ac.uk/training/online/trainers/sarahm
[7] https://www.ebi.ac.uk/training/online/trainers/grace_11792
[8] https://www.ebi.ac.uk/training/online/trainers/prudence_11864
[9] https://www.ebi.ac.uk/training/online/trainers/orchard%40ebi.ac.uk
[10] https://www.ebi.ac.uk/training/online/trainers/emily
[11] https://www.ebi.ac.uk/training/online/trainers/sangya.pundir
[12] https://www.ebi.ac.uk/training/online/trainers/lornar_8729
[13] https://www.ebi.ac.uk/training/online/trainers/amytyang
[14] http://www.ebi.ac.uk/services
[15] http://www.ebi.ac.uk/training/online
[19] https://www.ebi.ac.uk/training/online/glossary/pubmed
[20] https://www.ebi.ac.uk/training/online/course/europe-pmc-quick-tour-0
[21] https://www.ebi.ac.uk/training/online/glossary/1000-genomes-project
[22] http://europepmc.org/articles/PMC4750478
[23] http://www.internationalgenome.org/about/
[25] https://www.ebi.ac.uk/training/online/glossary/genotype
[27] https://www.ebi.ac.uk/training/online/glossary/variant
[28] https://www.ebi.ac.uk/training/online/glossary/allele
[29] https://www.ebi.ac.uk/training/online/course/ensembl-quick-tour-0
[31] https://www.ebi.ac.uk/metagenomics/
[32] https://www.ebi.ac.uk/training/online/course/ebi-metagenomics-analysing-and-exploring-metagenomics-data
[34] http://www.ebi.ac.uk/arrayexpress
[35] https://www.ebi.ac.uk/training/online/glossary/accession
[36] https://www.ebi.ac.uk/fg/annotare