Biological databases and bioinformatics resources

Prof. Thomas Keane
Research and Services Team Leader

E: tk2@ebi.ac.uk
T: @drtkeane

EMBL-EBI

European Molecular Biology Laboratory

International treaty organisation to promote molecular biology across Europe

- Found in July 1974, intergovernmental treaty of nine European countries plus Israel
- 2020: 27 member states, multiple Nobel prizes
- Six European centres of excellence for molecular biologists.
  - >60 nationalities, >1800 personnel

European Bioinformatics Institute

EMBL-EBI

Europe's center for biological data services, research and training

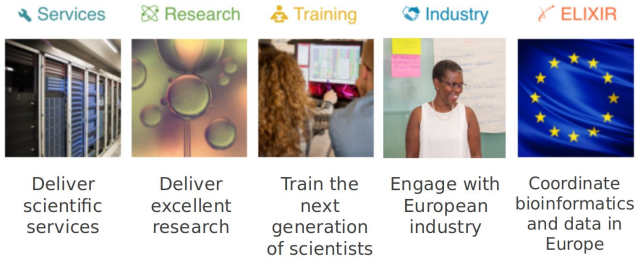
- Make the world's public biological data freely available to the scientific community
- Range of services and tools, basic research, and professional bioinformatics training

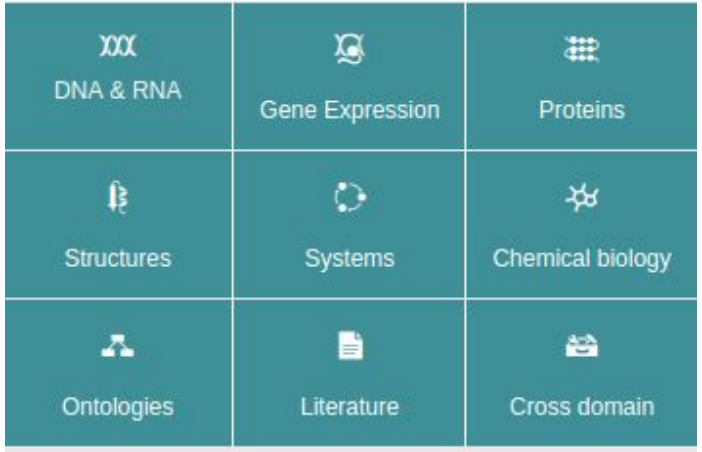
A trusted data provider for the life sciences:

- 150 Petabytes of storage
- > 40,000 CPU Cores

Part of the European Molecular Biology Laboratory

- International: >650 members of staff from 60 nations
- Home of the ELIXIR Technical hub.





Data Resources at EMBL-EBI

Molecular Archives

BioSamples

EFO

EGA

ENA

EVA

Mouse resources

Molecular Atlas

ArrayExpress

Expression Atlas

PRIDE

Molecular & Cellular Structure

EMDB

EMPIAR

PDBe

PDBe-KB

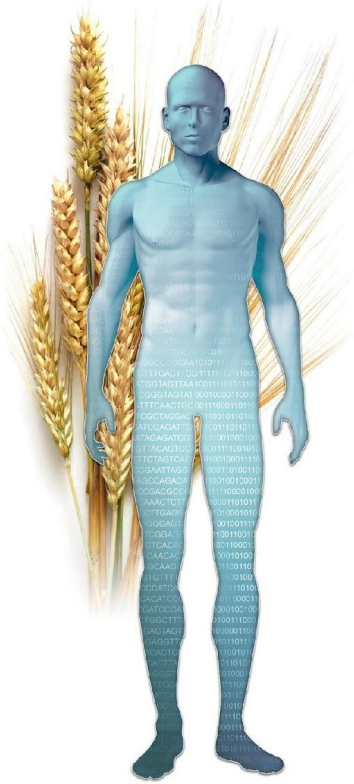
Molecular Systems

BioModels

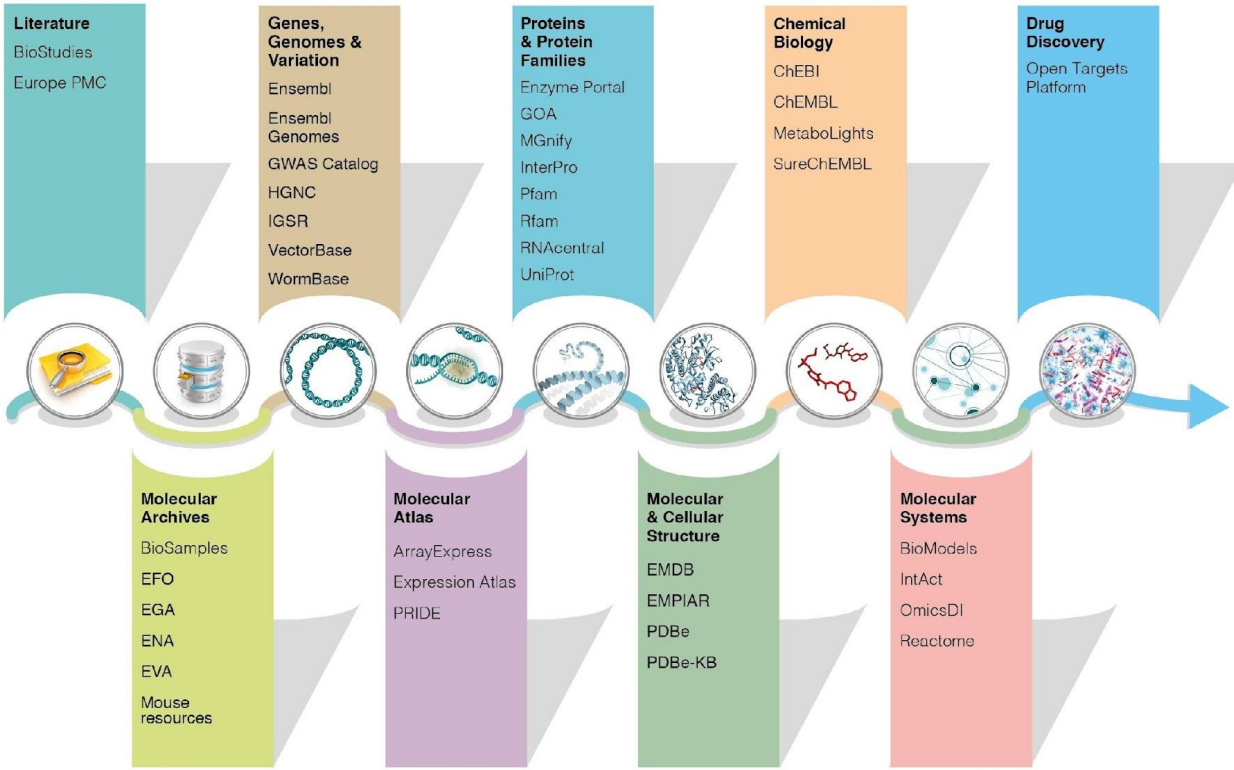
IntAct

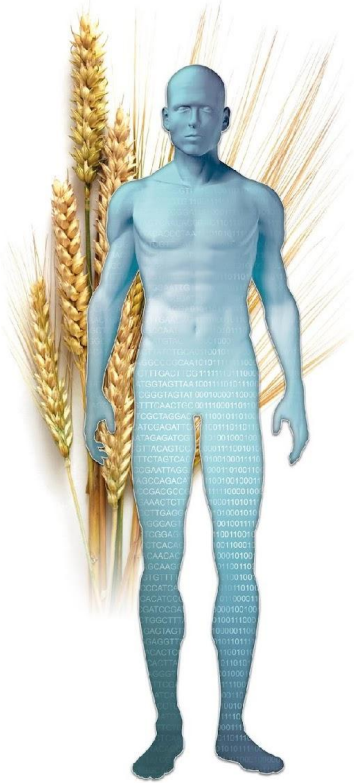
OmicsDI

Reactome



Data Resources at EMBL-EBI





COVID-19 Data Portal (2020-)

COVID-19 Data Portal

About Tools FAQ Related Resources Bulk Downloads Submit Data

Viral Sequences Host Sequences Expression Proteins Networks Cohorts More

COVID-19 Data

Accelerating research through data sharing

Search

Search

Examples: ACE2 , Severe acute respiratory syndrome 2 ...

Advanced search

Viral sequences

Raw and assembled sequence and analysis of SARS-CoV-2 and other coronaviruses.

15,933,849 records

Host sequences

Raw and assembled sequence and analysis of human and other hosts.

30,985 records

Share new data

Contact our curator teams, who will assist you with submitting your data to EMBL-EBI respositories >

Expression

Gene and protein expression data of human genes implicated in the virus infection of the host cells. Identifying cell types and genes with highest expression in SARS-CoV-2 infections.

234 records

Proteins

Curated functional and classification data on the SARS-CoV-2 protein entries and associated protein receptors.

3,856 records

Latest news

First linked datasets study

Why do we need public biological archives?

For **permanent scientific record** as evidence for scientific discoveries and support **reproducibility**

Scientists can **share data** associated with global scientific community

Finding datasets that might be relevant to your own research

Retrieve/download datasets from publications

Many **different data archives** for different data types

Why do we need public databases?

Rare disease clinical diagnosis (<5 in 10,000 of the general population)

- Patient with a very rare disease (<1 in million cases), e.g. rare blood disorder
- Clinician needs to identify other cases in the world
- Identify known genetic causes, e.g. malfunctioning gene
- **Resources: Matchmaker exchange, ClinVar, Beacon network**

Researcher studying evolution of species traits

- Evolution of traits across million of years, e.g. the eye, toxin/venom/parasite resistance, tissue regeneration
- Compare genome sequence of species with/without trait to discover genes or mechanisms
- Public databases that collate genomes, genes, regulatory data, transcription evidence, protein evidence etc.
- Species specific: Plasmodb, MGI, Gencode,
- **Genome browsers: Resources: Ensembl, UCSC Genome browser, NCBI**

Beacon Network

Ensembl

UCSC Genome Browser

ENA
European Nucleotide Archive

INSDC

Rest of world

National Center for Biotechnology Information (NCBI)

- Division of the National Library of Medicine (NLM) at the National Institutes of Health (NIH)

DNA Databank of Japan

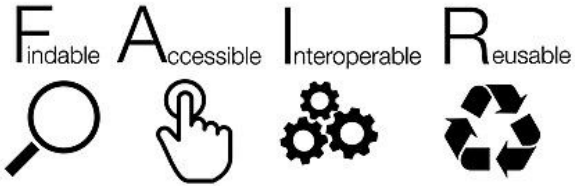
- Bioinformation and DDBJ Center provides sharing and analysis services for data from life science researches and advances science.

Global Biodata Coalition (GBC)

- Forum for research funders to better coordinate and share approaches for the efficient management and growth of biodata resources worldwide.

FAIR Principles

FAIR principles: Provide clarity around the goals and implementation of good data management and stewardship



Guiding principles for scientific data management and stewardship

- Enhance the **reusability** of data in public biological repositories
- Specific emphasis on enhancing the ability of machines to **automatically find** and use data

Four foundational principles: Findable, Accessible, Interoperable, and Reusable

- Use-cases: 1) Human interaction 2) Automatic computer interaction, e.g. find and use data

FAIR Principles

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

beta.uniprot.org/uniprotkb/Q9Z2W8/entry

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

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins

Q9Z2W8 · GRIA4_MOUSE

Glutamate receptor 4 · Mus musculus (Mouse) · **Gene:** Gria4 (Glur4) · 902 amino acids · Evidence at protein level · **Annotation score:** 5/5

Entry | Feature viewer | Publications | External links | History

BLAST | Align | Download | Add | Add a publication | Entry feedback

Function[i]

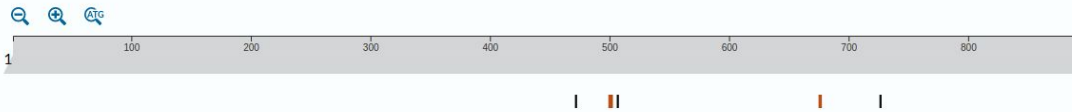
Receptor for glutamate that functions as ligand-gated ion channel in the central nervous system and plays an important role in excitatory synaptic transmission. L-glutamate acts as an excitatory neurotransmitter at many synapses in the central nervous system. Binding of the excitatory neurotransmitter L-glutamate induces a conformation change, leading to the opening of the cation channel, and thereby converts the chemical signal to an electrical impulse. The receptor then desensitize rapidly and enters a transient inactive state, characterized by the presence of bound agonist. In the presence of CACNG4 or CACNG7 or CACNG8, shows resensitization which is characterized by a delayed accumulation of current flux upon continued application of glutamate (By similarity). By Similarity

Miscellaneous

The postsynaptic actions of Glu are mediated by a variety of receptors that are named according to their selective agonists. This receptor binds AMPA (quisqualate glutamate > kainate.

Features

Showing features for region[i], binding site[i].



TYPE	ID	POSITION(S)	DESCRIPTION
-- Select --			
Region		500-502	Glutamate binding By Similarity
Region		676-677	Glutamate binding By Similarity
Binding site		472	Glutamate By Similarity
Binding site		507	Glutamate By Similarity
Binding site		727	Glutamate By Similarity

EMBL-EBI

Stable ID and URL

beta.uniprot.org/uniprotkb/Q9Z2W8/entry

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

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins

Q9Z2W8 · GRIA4_MOUSE

Glutamate receptor 4 · Mus musculus (Mouse) · **Gene:** Gria4 (Glur4) · 902 amino acids · Evidence at protein level · **Annotation score:** 5/5

Entry | Feature viewer | Publications | External links | History

BLAST | Align | Download | Add | Add a publication | Entry feedback

Function[i]

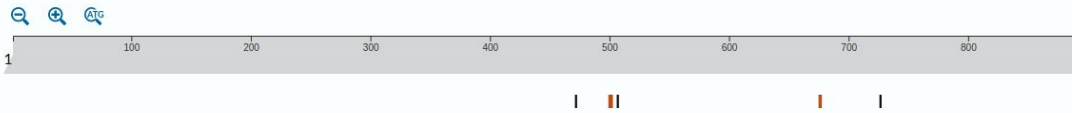
Receptor for glutamate that functions as ligand-gated ion channel in the central nervous system and plays an important role in excitatory synaptic transmission. L-glutamate acts as an excitatory neurotransmitter at many synapses in the central nervous system. Binding of the excitatory neurotransmitter L-glutamate induces a conformation change, leading to the opening of the cation channel, and thereby converts the chemical signal to an electrical impulse. The receptor then desensitizes rapidly and enters a transient inactive state, characterized by the presence of bound agonist. In the presence of CACNG4 or CACNG7 or CACNG8, shows resensitization which is characterized by a delayed accumulation of current flux upon continued application of glutamate (By similarity). By Similarity

Miscellaneous

The postsynaptic actions of Glu are mediated by a variety of receptors that are named according to their selective agonists. This receptor binds AMPA (quisqualate > glutamate > kainate.

Features

Showing features for region[i], binding site[i].



TYPE	ID	POSITION(S)	DESCRIPTION
-- Select --			
Region		500-502	Glutamate binding By Similarity
Region		676-677	Glutamate binding By Similarity
Binding site		472	Glutamate By Similarity
Binding site		507	Glutamate By Similarity
Binding site		727	Glutamate By Similarity

EMBL-EBI

Stable ID and URL

Links to other resources

beta.uniprot.org/uniprotkb/Q9Z2W8/entry

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

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins

Q9Z2W8 · GRIA4_MOUSE

Glutamate receptor 4 · Mus musculus (Mouse) · **Gene:** Gria4 (Glur4) · 902 amino acids · Evidence at protein level · **Annotation score:** 5/5

Entry | Feature viewer | Publications | External links | History

BLAST | Align | Download | Add | Add a publication | Entry feedback

Function[i]

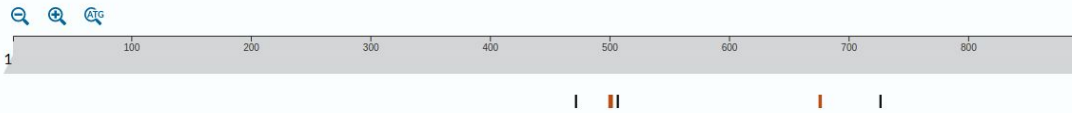
Receptor for glutamate that functions as ligand-gated ion channel in the central nervous system and plays an important role in excitatory synaptic transmission. L-glutamate acts as an excitatory neurotransmitter at many synapses in the central nervous system. Binding of the excitatory neurotransmitter L-glutamate induces a conformation change, leading to the opening of the cation channel, and thereby converts the chemical signal to an electrical impulse. The receptor then desensitizes rapidly and enters a transient inactive state, characterized by the presence of bound agonist. In the presence of CACNG4 or CACNG7 or CACNG8, shows resensitization which is characterized by a delayed accumulation of current flux upon continued application of glutamate (By similarity). By Similarity

Miscellaneous

The postsynaptic actions of Glu are mediated by a variety of receptors that are named according to their selective agonists. This receptor binds AMPA (quisqualate glutamate > kainate.

Features

Showing features for region[i], binding site[i].



TYPE	ID	POSITION(S)	DESCRIPTION
-- Select --			
Region		500-502	Glutamate binding By Similarity
Region		676-677	Glutamate binding By Similarity
Binding site		472	Glutamate By Similarity
Binding site		507	Glutamate By Similarity
Binding site		727	Glutamate By Similarity

EMBL-EBI

Stable ID and URL

Links to other resources

Rich metadata

beta.uniprot.org/uniprotkb/Q9Z2W8/entry

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

- Function
- Names & Taxonomy
- Subcellular Location
- Phenotypes
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence
- Similar Proteins

Q9Z2W8 · GRIA4_MOUSE

Glutamate receptor 4 · Mus musculus (Mouse) · **Gene:** Gria4 (Glur4) · 902 amino acids · Evidence at protein level · **Annotation score:** 5/5

Entry | Feature viewer | Publications | External links | History

BLAST | Align | Download | Add | Add a publication | Entry feedback

Function

Receptor for glutamate that functions as ligand-gated ion channel in the central nervous system and plays an important role in excitatory synaptic transmission. L-glutamate acts as an excitatory neurotransmitter at many synapses in the central nervous system. Binding of the excitatory neurotransmitter L-glutamate induces a conformation change, leading to the opening of the cation channel, and thereby converts the chemical signal to an electrical impulse. The receptor then desensitizes rapidly and enters a transient inactive state, characterized by the presence of bound agonist. In the presence of CACNG4 or CACNG7 or CACNG8, shows resensitization which is characterized by a delayed accumulation of current flux upon continued application of glutamate (By similarity). By Similarity

Miscellaneous

The postsynaptic actions of Glu are mediated by a variety of receptors that are named according to their selective agonists. This receptor binds AMPA (quisqualate glutamate > kainate.

Features

Showing features for region, binding site.

1 100 200 300 400 500 600 700 800

TYPE	ID	POSITION(S)	DESCRIPTION
-- Select --			
Region		500-502	Glutamate binding By Similarity
Region		676-677	Glutamate binding By Similarity
Binding site		472	Glutamate By Similarity
Binding site		507	Glutamate By Similarity
Binding site		727	Glutamate By Similarity

EMBL-EBI

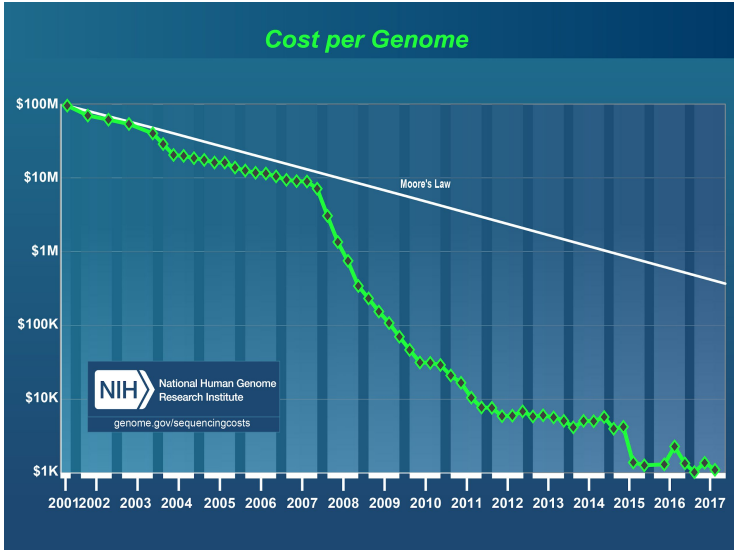
DNA Sequencing

DNA sequencing is the process of working out the order of the bases (A, C, G and T) in a strand of DNA

- Single gene, 000's bp, Bacterial genome, ~2-5Mbp, Human genome ~3Gbp

Evolution of DNA Sequencing

- 1970's Sanger sequencing method
- 1990's Capillary sequencing
- 2000's Second generation sequencing
- Now: Third generation sequencing



DNA Sequencing

Sequencing technologies can only sequence short stretches of DNA

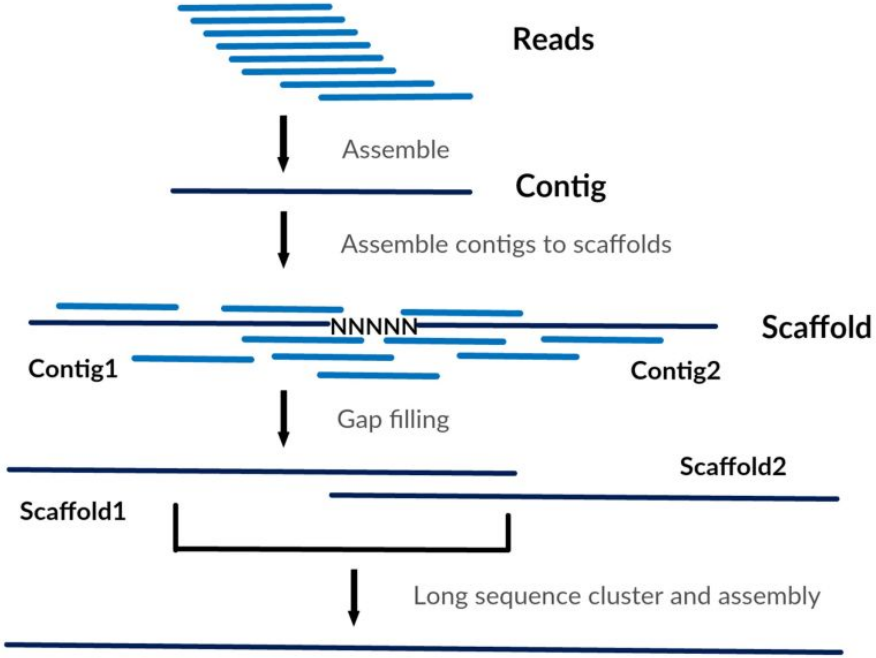
Given many (millions or billions) of reads, produce a linear (or perhaps circular) genome

Clone based

- Select clones using markers
- Sequence clones separately

Whole-genome shotgun

- Fragment whole-genome and sequence
- De novo assembly



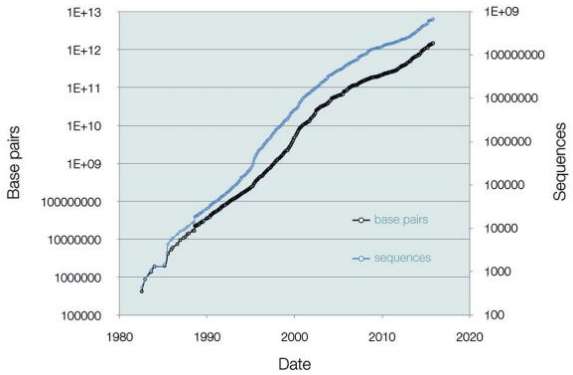
Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis

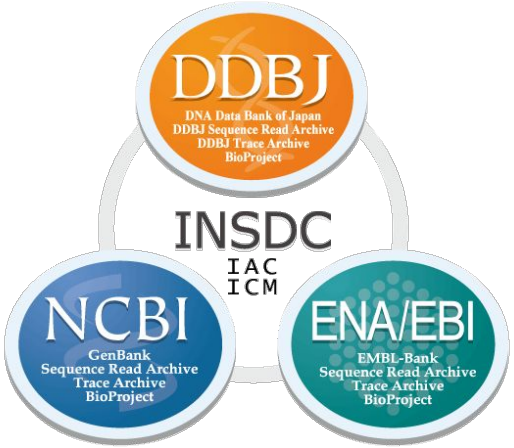
Yan Guo [a,*], Yulin Dai [a], Hui Yu [a], Shilin Zhao [a], David C. Samuels [b], Yu Shyr [c,*]

DNA Databases

International Nucleotide Sequence Database Collaboration (INSDC)

- **Three global partners** that capture, preserve and provide comprehensive public-domain **nucleotide sequence** information
- Establishes standards, formats and protocols for submission and access of nucleotide data (reads, genome assembly, gene annotation)
  - e.g. Feature Table Definitions document, the INSDC country list and conventions in the description of experimental support for annotated features





DNA Databases - INSDC

Data Type	DDBJ	EMBL-EBI	NCBI
NGS reads	Sequence Read Archive	European Nucleotide Archive (ENA)	Sequence Read Archive
Capillary reads	Trace Archive		Trace Archive
Annotated sequences	DDBJ		GenBank
Samples	BioSamples	BioSamples	BioSamples
Studies	BioProject	BioProject	BioProject

European Nucleotide Archive (ENA)

Globally comprehensive scientific record of open nucleotide data

- Established in early 80's
- Platform for the management, sharing, integration and dissemination of sequence data

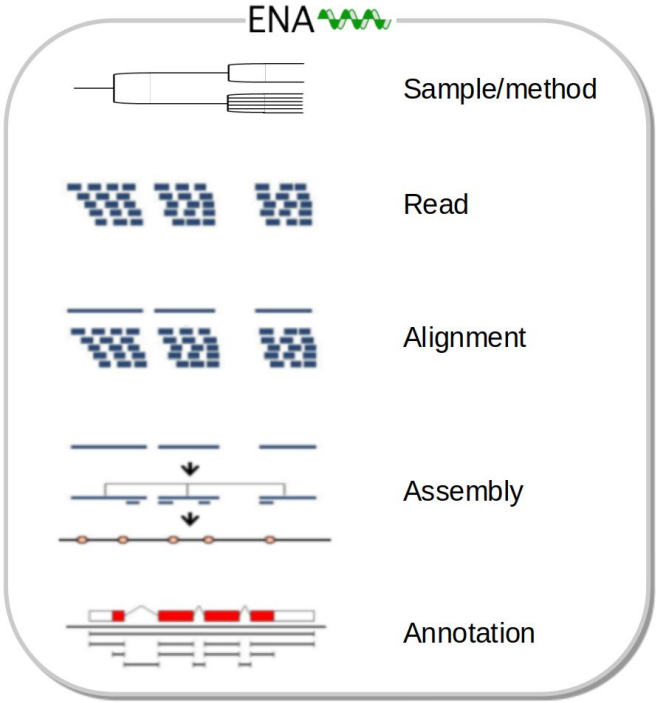
Permanent identifiers/accessions for all objects

- Samples, studies, experiments, data objects

Big data

- 1.3 petabase pairs across >1 million taxa, 2,000-5,000 active data providers, global consumer user base

Services: Submission, discovery and retrieval software, tools and services



http://www.ebi.ac.uk/ena/

Project: PRJEB402

This project will rely on samples collected during the scientific expedition Tara-Oceans (2009-2012). By March 2012, the schooner Tara, equipped with innovative systems for sampling of 11 organismal size-ranges covering entire planktonic communities from viruses to animals, has collected standardized genetic (total DNA/RNA), morphological, and physico-chemical (contextual) samples from 153 sites across the world oceans, locations carefully selected with input from near-real-time remote sensing and in-situ hydrographic criteria. Overall, a total of ~50,000 biological samples and ~13,000 contextual measures from 3 depths will be analysed. The metagenomics component of the project consists of size fractionated plankton samples that are submitted to barcoding and shotgun sequencing, as well as isolated single-cell amplified protists and single-organisms isolated metazoans that are sequenced as reference genomes.

Organism:	marine metagenome
Secondary Study Accession:	ERP004109
Study Title:	Tara-oceans samples barcoding and shotgun sequencing
Center Name:	Genoscope CEA
ENA-FIRST-PUBLIC:	2012-08-22
ENA-LAST-UPDATE:	2021-07-23

Show More

Component Projects

Accession	Title/Description
PRJEB1787	**Shotgun Sequencing of Tara Oceans DNA samples corresponding to size fractions for prokaryotes.** Seawater was filtered from different depths to retain small cell sizes (Bacteria Organisms). The DNA... See more
PRJEB1788	**Shotgun Sequencing of Tara Oceans DNA samples corresponding to size fractions for large DNA viruses** Seawater was filtered from different depths to retain small cell sizes. The DNA was extracted and su... See more
PRJEB36282	**Amplicon sequencing of Tara Oceans DNA samples corresponding to size fractions for prokaryotes.** Analysis of 16S DNA in Tara Oceans Prokaryotes size fractions through amplicon sequencing: Seawater ... See more

Genetic Variation

SNPs/SNVs . . . Single Nucleotide Polymorphism/Variation

```
ACGTTTAGCAT
ACGTTCAGCAT
```

MNPs . . . Multi-Nucleotide Polymorphism

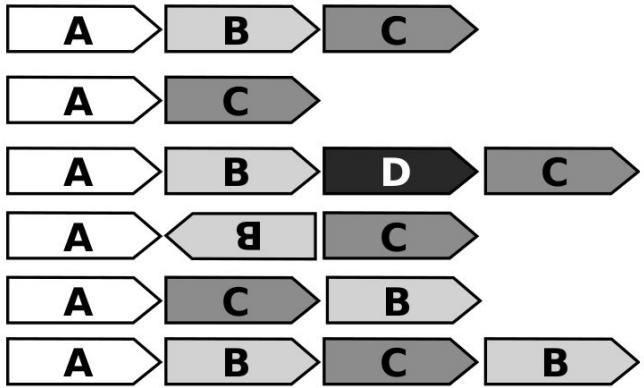
```
ACGTCCAGCAT
ACGTTTAGCAT
```

Indels . . . short insertions and deletions

```
ACGTTTAGCA-TT
ACGTT-AGCAGTT
```

SVs . . . Structural Variation

Deletion
Insertion
Inversion
Translocation
Duplication



European Variation Archive (EVA)

Short variants (SNPs/indels)

- dbSNP at NCBI since 1998
- EVA at EBI since 2014

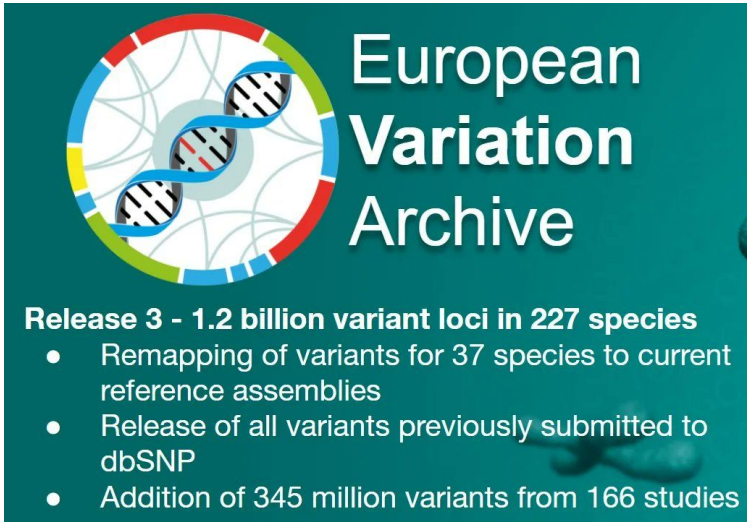
dbSNP
Short Genetic Variations

Structural Variants

- dbVar at NCBI since 2010
- DGVa + EVA at EBI

Data requirements

- Can be openly shared
- Described in Variant Call Format (VCF) files
- Samples genotypes and/or allele frequencies
- Reference sequence registered at INSDC



EVA Data Submissions

European Variation Archive

Home | Submit Data | Study Browser | Variant Browser | GA4GH | API | dbSNP Import Progress | Help | Feedback

EVA / SUBMIT

Submit

Please read our Data Requirements and the Key stages of submission below. All data valid for EVA submission shall be made available via the Study Browser and will be browsable using both the Variant Browser and the EVA API. Variant Effect Predictor annotations shall be available for variants mapped to genome assemblies that are known to Ensembl.

Data submitted to the EVA is brokered to our collaborating databases at NCBI, dbSNP and dbVar. It is therefore unnecessary to submit data to multiple resources.

Data requirements

EVA accepts **all** types of **precise** genetic variants, in **any species** providing the following requirements are met:

1. Data is described in **valid VCF file(s)**. This can be tested prior to submission using the EVA VCF Validation Suite found here. For help with converting variation data to VCF, please see our help pages.
2. Data includes **sample genotypes** and/or **allele frequencies**
3. The reference sequence used is INSDC registered, or will be at point of submission. A "reference" can be any of the following, but not restricted to:
   - Assembly, e.g. GCA_000002285.2
   - Transcriptome/Transcript, e.g. GCJV01000000, KY286086
   - Gene sequence, e.g. X76482

   PLEASE NOTE: Sequence identifiers in VCF must match those in the reference FASTA file.
4. If consent was gathered for any **individual human genotype data** then a consent statement must be completed prior to submission.

Variant accessions (ss# and rs#) and study accessions will only be provided for data which satisfies all data requirements. More details on whether your data is suitable can be found here.

Alternative resources for data not accepted by EVA

- Submit structural variations that cannot be expressed in VCF(s) to DGVa.
- Submit variations with sensitive clinical data to EGA.
- Submit variations with clinically relevant genetic variant data, i.e. data that relates genetic variation(s) with clinical significance values (e.g. pathogenic, benign, etc.), to the ClinVar archive at NCBI.

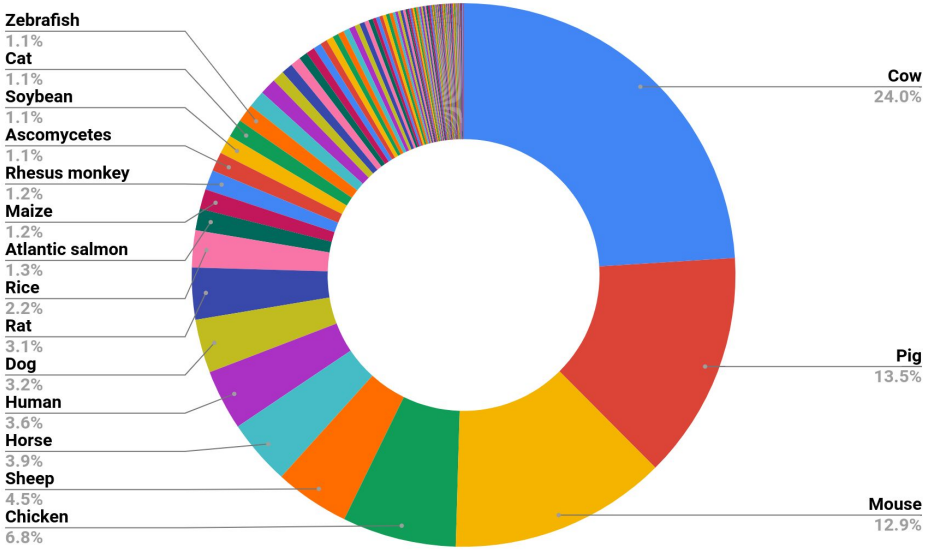
Key stages of EVA submissions

Prepare

- Prepare valid VCF file(s), which can be validated prior to submission using the EVA VCF validation suite.
- Complete a metadata template describing the samples and analyses in your study. Please provide as much metadata as possible since this information is extremely useful for downstream analysis and is directly related to the frequency at which datasets

EVA Browser

EVA Studies



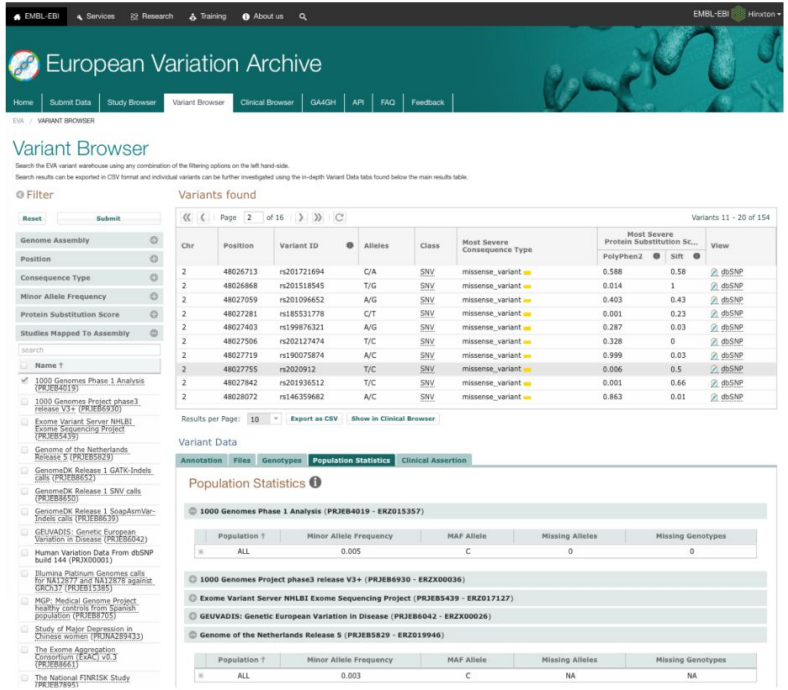


Figure 2: The EVA Variant Browser highlighting a missense variant in the human MSH6 gene on chromosome 2. The allele frequency of this variant is shown in two of the studies archived at the EVA in the bottom panel and options for data filtering and discovery are displayed on the left.

Many Data Sources

http://www.ensembl.org/info/genome/variation/sources_documentation.html

Secondary Resources

Collate data from several primary resources (DNA, RNA, protein evidence)

Resources that act as portals for a particular:

- Group of species (e.g. mammals, fish, bacteria, parasites)
- Model organism (e.g. mouse, human, yeast, fly)
- Research community (e.g. farm animals)

Genome browsers

- Not species specific, but may prioritise
- Ensembl, UCSC Genome Browser, ZENBU genome browser, PlantGDB

Species specific databases

- Plasmodb, Mouse Genome Informatics (MGI), Typdb, Saccharomyces Genome Database (SGD), FlyBase

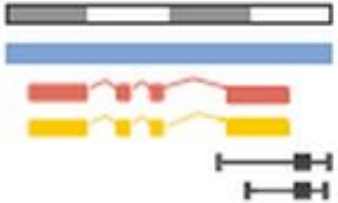
Ensembl Genome Browser

Ensembl project aggregates, processes, integrates and redistributes genomic datasets

- Began with initial release of human genome
- Expanded into animals, plants, fungi, bacteria

Genome browser enables comparative genomic, clinical genetics,

- Genomic sequence + automated gene annotation
- Genetic variation - Small and large scale sequence variation with phenotype associations
- Comparative Genomics - Whole genome alignments, gene trees
- Regulation - Potential promoters and enhancers, DNA methylation



FAIR Principles

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards