Human genetic variation (II): exploring publicly available data

DNA & RNA

Beginner

2 hours

Course overview

Genetic variation is fundamental to the evolution of all species and is what makes us individuals. Our genes have a large influence on our lives. They affect what we look like, our personalities and preferences and our susceptibility to disease. By studying genetic variation we hope to understand the molecular process that contribute to life on earth.

The study of genetic variation has been used to model human migration, understand the cause of human diseases, and to predict disease outcomes.

This is part II of our course on human genetic variation. We will and learn how to explore publicly available genetic variation data.

Prior to part II of this course

Part I of the course [9] introduces some key concepts in the field of human genetic variation including the types and possible effects of genetic variation, data formats and look common genetic variation study types. If you are new to the field we recommend that you work through part I of this course first.

The courses focus on heritable (germline) variation and will give you a taste of the resources you can use to explore genetic variation data.

Learning objectives:

- List examples of genetic variation databases
Introduction to public genetic variation data

A wealth of genetic variation data is generated within the scientific community to investigate many diverse subject areas from human disease to informing selective breeding in common bean varieties.

In addition to the continued generation of new genetic variation data, it is important for the community to have access to data that has been previously generated to aid re-evaluation and/or re-use of data in testing both new and previously established hypotheses.

What EMBL-EBI databases and resources are available for sharing, exploring and understanding genetic variation data?

**European Variation Archive** [10] (EVA): a database of genetic variation data. These datasets are submitted from the community to EVA in order to aid data sharing, and data reuse.

**Ensembl** [11]: a genome browser that provides a single point of access to annotated genomes. It includes information about genetic variants, population genetics and tools for exploring your own variant data.

**GWAS catalog** [13]: a quality controlled, manually curated database of published GWAS studies. The GWAS karyotype diagram provides an interactive way of exploring all SNP-trait associations.

**UniProt** [15]: EMBL-EBI’s resource for protein sequence and annotation data. You can use UniProt’s protein feature viewer to explore variants in relation to protein sequences, structure and function.

Standardised terminology

Each of these databases uses standardised ways of identifying and classifying variants.

For example, Sequence **Ontology** [17] (SO) provide a standard nomenclature for categorising variants based on where they fall with respect to genes and other genomic features (Figure 1). For an overview of identifiers used by different databases see the section on **variant identifiers** [18] in part I of this course.
Figure 1 A gene model with possible variant consequences.

In addition to the sequence ontology terms, an IMPACT [19] measure, agreed by Ensembl and SnpEff [20] provides a subjective classification of the severity of each class. Terms commonly used by Ensembl to describe variants are shown in Figure 2.
Figure 2 A list of commonly used sequence ontology (SO) terms for describing variants, from Ensembl. The terms are shown in order of severity (more severe to less severe) as estimated by Ensembl. View the whole list [21] of SO terms.

Known variants in databases are usually annotated with these terms, and tools such as the Ensembl VEP [22] and SnpEff [20] allow you to use them to annotate your own variants.

There are many different starting points for exploring publicly available genetic variation data. The next section features case studies that will illustrate four different ways you can access genetic variation data using a gene, variant, phenotype [23] or publication as a starting point.

Click next to view the first case study: variants in gene PKD1

Case study 1: Variants in a gene (PKD1)

In this case study we will use the human gene *polycystin 1, transient receptor potential channel interacting* (PKD1), as a starting point for exploring genetic variation data.

First, you may want to find all open access [24] genetic variants within *PKD1*. You can do this by searching for the gene symbol and filtering for coding variants in the EVA Variant browser [25].
Using the EVA variant browser to search for *PKD1* identifies 8186 variants from 17 studies (Figure 3).

![Variant Browser](image)

**Figure 3** Searching for PKD1 in the EVA variant browser ([view in EVA](#)).

The European Variation Archive ([27](#)) (EVA) is a repository for genetic variation data at EMBL-EBI. To learn more about the EVA you can watch our webinar The European Variation Archive at EMBL-EBI ([28](#)).

How can Ensembl help explore genetic variations of PKD1?

Find out next.

**Using Ensembl to explore genetic variations in the PKD1 locus**

**Finding variants**

[Open access](#) genetic variation datasets in the EVA are fed into [Ensembl](#) where they are combined with many different data types from many different resources and services. These digestible chunks of information can be accessed via web browsers, or programmatically via the [Ensembl API](#).

This means that we can use Ensembl to gain further insight into the genetic variations that are found within the human *PKD1* locus.

For example, let’s look at the Ensembl table of variants view for *PKD1* (Figure 4):
Figure 4 The Ensembl table of variants for PKD1 (view in Ensembl [30]).

**Viewing traits**

There are a total of six variants displayed for the human PKD1 gene (Figure 4). These have been filtered for pathogenic, exonic variants that exist at at allele [31] frequency of greater than 0.001.

The pathogenicity of these variants within the human PKD1 locus can be further investigated with Ensembl by viewing the associated traits in the phenotypes section [32] (Figure 5).

Figure 5 Using Ensembl to find traits associated with variants in PKD1 (view in Ensembl [32]).

Here we can see that the majority of traits associated with variants in the human PKD1 gene are held privately at the human mutation gene database. The publicly available associations are mostly to polycystic kidney disease 1.
You can learn more about the variation data in Ensembl in our webinar Variation data in Ensembl and the Ensembl VEP [33].

Next, learn how to use Uniprot to learn about protein structure and function.

**PKD1 variants, protein structure and function**

Now that we know that variants in *PKD1* are associated with a disease we can start to use EMBL-EBI protein-centric resources to understand the potential effects of variants on protein structure and function.

These resources provide detailed information on where in the protein sequence such variants lie, and whether these variants overlap with domains and/or sites, potentially affect post-translational modification residues and/or other protein structural features.

Let’s take a look at UniProt’s Feature Viewer display for the human PKD1 protein [34] (Figure 6).

![Figure 6 The UniProt Feature Viewer for PKD1 (view in UniProt [34]).](image)

The human PKD1 protein contains a number of disease reviewed variants. Not all of these variants are known to be pathogenic, however the Uniprot Sequence Viewer shows how the known variants overlap with other protein features. As you can see, there are many disease causing variants in PKD1. You can learn more about these variants and how they relate to the protein structure by clicking on them.

From UniProt you can also link out to PDB [35] where you can explore available protein structures for PKD1. We will look at this in more detail in the next case study.

As you have seen, starting with a gene of interest can unearth a wide variety of information to help you understand how variants in that gene contribute to health and disease and influence protein structure and function.
You can learn more about variation data in UniProt by watching our webinar [36].

In the next case study, we start with a variant.

**Case study 2: Search for a variant (rs334)**

This case study assumes that you have a variant identifier that you want to learn more about, for example from the results of a variant calling analysis. In this example we will be using rs334, a dbSNP identifier.

One place you might search for the identifier is Ensembl [11]. There, you can find the variant alleles and their source, along with links to further information, including population allele frequencies, sample genotypes, phenotypes that have been associated with the variant, and genes and proteins affected by the variant. rs334 has four observed alleles, where T is the reference (Figure 7).

![Ensembl screenshot](image)

**Figure 7** Search for the rs334 variant in Ensembl ([view in Ensembl](37)).

In the genes and regulation section we can see that rs334 is a missense variant in *HBB*, a haemoglobin subunit. It is associated with sickle cell anaemia and malaria resistance, and that the phenotype-associated A allele is mostly
found in African populations (you can see this in the Phenotype data and Population genetics sections). This is consistent with what we know about sickle cell anaemia; that it is caused by deformed haemoglobin protein resulting in sickled red blood cells. The same change in the protein structure also confers malarial resistance. This is advantageous to heterozygotes if they are exposed to malaria, so is most common in regions where malaria is endemic.

How can we find associated phenotypes?

Click next to find out.

Exploring phenotypes associated with a variant

To learn more about how rs334 is linked to phenotypes you can search for it in the GWAS Catalog [38]. A search for “rs334” returns the entry for that variant [12], containing all relevant data in the GWAS Catalog, such as studies, associations and traits associated with rs334 (Figure 8).

Figure 8 Studies that mention rs334 and associations between rs334 and traits can be found by searching the GWAS Catalog for rs334 (view in the GWAS Catalog [39]).
In the GWAS Catalog, rs334 is reported to be associated with urinary albumin-to-creatinine ratio and severe malaria. Summary information is provided for each association, including p-value [40], effect size (odds ratios or -coefficient), risk-allele, location and mapped gene. The Catalog includes information describing the GWAS in which this association was identified, including links to the publication, study design and ontology [17] terms to describe the phenotype and allow integration with data from other resources. Since GWAS are generally used to study complex, rather than simple Mendelian inheritance, we do not see sickle cell anaemia listed here.

To learn more about the GWAS catalog have a look at our webinar The NHGRI-EBI GWAS Catalog, a curated resource of SNP-trait associations [41].

How can we explore the effects of variants on function?

Find out next.

**Understanding the functional consequences of a variant**

Now that we know that rs334 is missense in HBB, and that it is associated with sickle cell anaemia we can start to probe the protein structure to understand the molecular mechanisms underlying this association.

**Searching Uniprot for variant identifiers**

As we saw in the first case study, you can explore proteins and their variants using UniProt [15]. As with genes or proteins you can search UniProt using the variant identifier (rs334).

There is one protein associated with this variant: human hemoglobin subunit B. By looking at the Pathology and Biotech section [42] for this protein we can see that the variant associated with sickle cell anaemia is p.Glu7Val and causes a change from a charged amino acid to a hydrophobic aliphatic amino acid. This is annotated as E -> V at position 7 (Figure 9).

![Figure 9](image)

The UniProt Pathology and Biotech section for human hemoglobin subunit B shows which variants are associated with sickle cell anaemia (view in UniProt [42]).

**Using Uniprot's feature viewer**

If we take a look at the feature viewer [43] for this protein you can see that this variant does not occur within the region of the essential heme binding residues, but does occur in an alpha-helix within a small cluster of charged residues (Figure 10).
In the final section of this case study, view the protein’s 3D structure.

**Viewing haemoglobin variants in protein structures**

Next we can use the Protein Data Bank archive [45] (PDBe) to find the three dimensional structure of haemoglobin and understand how the position of this change relates to the the 3D structure of the protein.
There are a number of different ways to search PDBe [46]. In this case we used the Uniprot [47] ID for human hemoglobin beta (P68871) as the search term and refined the results using the word ‘sickle’ to find hemoglobin structures solved for sickle cell phenotypes.

The view below shows the macromolecules tab [48] in the search results for this Uniprot ID, which gives a single macromolecule result, specifically the protein described in that Uniprot entry (Figure 11).

**Figure 11** The macromolecules tab in PDBe showing results for hemoglobin structures solved for sickle cell phenotypes. This is an interactive subsection of the original page which can be view in PDBe [48].

Comparing 2D and 3D structures

Suppose you want to explore the 2D and 3D structures of the hemoglobin subunit beta to understand how genetic variants influence the protein structure. You can do this by clicking on the name of this macromolecule, i.e. Hemoglobin subunit beta. This will take you through to the macromolecules page for that specific structure [49] (Figure 12). Here you are presented with sequence views for that particular protein, as well as a 2D topology graphic and a 3D structure viewer. The first section in the sequence view (Molecule) shows any sequence annotations for the protein in this structure, with these highlighted in orange (1D sequence annotation). Hovering over the orange bar for this example will display the change of residue 6 from glutamate (E) to valine (V) as is the case for sickle cell hemoglobin variants.

**Figure 12** The macromolecules page for haemoglobin subunit beta. This is an interactive subsection of the original page which can be viewed in PDBe [49].

If you hover over residue 6 in the schematic of the Topology 2D diagram, you will find that this specific amino acid is highlighted in yellow on the surface of the hemoglobin molecule in the 3D structure view.

The change of this amino acid in the sickle cell variant is from a hydrophilic residue (glutamate) to a hydrophobic residue (valine). This change generates a ‘sticky patch’ on the surface of the protein because the ‘water loving’ amino acid has been swapped for a ‘water hating’ one. This causes the association of multiple hemoglobin complexes, via this hydrophobic valine residue. This consequently leads to the aggregation of hemoglobin molecules into fibres, therefore producing cells with the sickle phenotype that is observed for this variant [50]. From this example you can see that by looking at the structure and understanding the type of variation involved, you can begin to draw functional conclusions about the consequences of variation.

In the next case study, we search for a phenotype.

**Case study 3: Search for a phenotype (non-melanoma skin cancer)**

With a known phenotype, you may wish to find studies focusing on it as well as variants and genes that are associated with the phenotype. In this case study we look at how you can find studies, variants and genes associated with the phenotype “non-melanoma skin cancer”.

To find studies looking at the phenotype, as well as associated genes and variants you can search for “non-melanoma skin cancer” in the GWAS Catalog [38] (Figure 13).
Figure 13 Search results for non-melanoma skin cancer in the GWAS Catalog. By searching for a phenotype you can find links to variants and studies where these links were found (view in GWAS Catalog [51]).

There are several studies in GWAS that focus on this trait. They identify many SNPs associated with non-melanoma skin cancer or more specific sub-types, including three in $MC1R$.

Another place to find information about phenotypes is Ensembl (Figure 14). Since Ensembl is not restricted by method type, it includes phenotype associations from GWAS, as well as sources such as ClinVar [52], OMIM [53], Orphanet [54] and COSMIC [55], which may be associated with either the variant or the gene.

Figure 14 Searching for a phenotype in Ensembl yields links to variants and genes associated with the phenotype (view in Ensembl [56]).

For detailed descriptions of phenotypes, you can go to OMIM [57]. As well as clinical descriptions, OMIM also
describes the biological and molecular functions of genes associated with the phenotypes along with references.

In case study 2 we saw how you can use Ensembl, UniProt [47] and PDBe [58] to explore a gene or variant in more detail. Open Targets [59] is another resource that you can use. It was originally designed for validating potential drug targets but is also useful for getting an overview of the data and features associated with a phenotype or gene.

As we saw in the GWAS Catalog, SNPs in MC1R are associated with non-melanoma skin cancer. If we look at this gene in Open Targets we can see that it is also associated with hair colour. Indeed people with red hair and freckles often have certain SNPs in MC1R (Figure 15).

**Figure 15** Searching for MC1R in open targets reveals that it is associated with hair colour. This is an interactive subsection, the full site can be viewed in Open Targets [60].

In the last case study, we search for data related to a specific publication.

**Case study 4: Starting with the literature**

Sometimes you might start with a specific publication. Perhaps you want to access the data that is described so that you can include it in your own analyses.

In this case study we are going to use the Nature paper [61] from the Wellcome Trust Case Control Consortium (WTCCC) as the starting point and investigate how we can access and analyse the data. The WTCCC is a consortium that was put together to help understand human variation using high-throughput technology.

You can find the paper in Europe PMC [62] by searching for keywords (e.g. author names, phenotypes of interest, etc.) or known literature identifiers (e.g. PMID [63]:17554300) (Figure 16). Searching for the paper in Europe PMC is a useful way to get started as the search results include direct links to information about the genes, proteins, variants etc. that are mentioned in the paper.
Figure 16 Europe PMC record of the WTCCC Nature publication from 2007 (view in Europe PMC [64]).

To learn more about Europe PMC watch our webinar Europe PMC: get the most from literature searches [65].

Access the data, both raw and processed, next

Accessing raw and processed data

Finding the raw data

In the paper it is stated that the participants gave their informed consent for the raw data to be shared provided that it is held securely in a controlled-access resource. The European Genome-phenome Archive [66] (EGA) is the EMBL-EBI’s controlled-access data resource. This means that if you want to access data held at the EGA you must apply and be approved [67]. This is a relatively straightforward process and access is usually granted to qualified investigators for appropriate use.

There are 33 studies at EGA that were submitted by WTCCC; a screenshot of these studies is shown in Figure 17.
Figure 17 The WTCCC data is held at the controlled-access EGA across 33 studies. Users must apply for access to the data owners in order to download these datasets (view in EGA [68]).

Once you have been granted access to the data you can download the raw data for use in your own studies.

Accessing processed data

In addition to accessing the raw WTCCC data, we are also able to access the processed data. Because the paper used Genome-Wide Association Study (GWAS) methods we can find the results in the GWAS Catalog by searching with the article’s PubMed ID (Figure 18).
Figure 18 The processed data from the WTCCC publication are available at the GWAS Catalog (view in GWAS Catalog [69]).

The data from the WTCC GWAS can be downloaded and reused in meta-analyses. For example, Zeggini et al combined these data with two other GWAS studies to uncover SNPs associated with type II diabetes [70].

It is also possible to browse specific variants using the GWAS Catalog as a starting point, as we saw in case study 2 [71].

To learn more about a particular variant we can link from the GWAS Catalog to Ensembl to further analyse the associated information. In turn, you can probe the effect of specific variants on protein structure and function using UniProt and PDBe as we did in case study 2, [71].

In the final page of this course, discover other resources to help you learn more.

Summary

In parts I and II of this course we have introduced some key topics in the field of human genetic variation. In this part of the course we’ve learnt about some of the public resources available for exploring genetic variation data. In the case studies we used these resources to answer specific biological questions and showed you how you can use genes, variants, phenotypes or the literature as a starting point in your research.

What’s next?

Perhaps you would like to try searching for your favourite gene or variant using the resources featured in this course. You can use the case studies as a guide to help find your way through the different resources.

Many of the resources mentioned in this course (for example Ensembl, EVA, UniProt, GWAS Catalog and PDBe) are developed and maintained by the EMBL-EBI. If you would like to learn more about these resources please have a look at our free online courses in Train online.

As a starting point we recommend:
Where to find help

In addition to our Train online courses, every EMBL-EBI resource provides detailed help and documentation pages on their website. You can also contact their help desk to ask a specific question.

Alternatively, you can get in touch via our website using this form [74]. Be sure to include your email address and pick the most appropriate topic so that we can direct your question to the people best able to answer it.

References


Your feedback

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

Contributors

David Armstrong [1]
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 Emery [3]
EMBL-EBI
Scientific Training Officer - Training

Laura completed her undergraduate degree in Genetics (BSc first class hons) at the University of Nottingham. After dabbling in experimental work on a summer placement at the Wellcome Trust Sanger Institute, she found joy in using computer-based analyses to explore biological questions. Laura completed her PhD in Evolutionary Genetics at the University of Edinburgh’s Institute of Evolutionary Biology. There she used a bioinformatics approach to investigate patterns of codon usage in Archaea while under the supervision of Prof Paul Sharp FRS.

She also worked as a postdoc at the University of Manchester, where she explored molecular co-evolution among interacting proteins. Having obtained qualified teacher status (QTS), Laura joined EMBL-EBI as a Scientific Training Officer in 2012 and is responsible for the development and delivery of training courses.
Jackie MacArthur [4]
EMBL-EBI
Project Leader (GWAS Catalog) - Cunningham team: Variation Annotation

Andrew Nightingale [5]
EMBL-EBI
Bioinformatician - Martin team: Protein Function development

Andrew has been a Bioinformatician with the Protein Function Development team at EMBL-EBI since 2012, and has a background in biochemistry and bioinformatics. Before joining EMBL-EBI, Andrew worked for a biotechnology company specialising in structure-based drug design by NMR. Andrew is responsible for enhancing the links between UniProt [47] and the genomics and translational biology communities to provide clinical and/or translational context to UniProtKB [77].

Emily Perry [6]
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 [7]

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.

Gary Saunders [8]

EMBL-EBI
Genetic Variation Scientific Curator - Paschall team: Variation

Gary Saunders is an EBI curator of the European Variation Archive and related resources: the Database of Genomics Variants archive and the European Genome-phenome Archive. It is Gary's responsibility to manage the data within these resource(s) to ensure accuracy, clarity and discoverability. Previous to this position, Gary was a curator of the GENCODE project, which provides the gene set for the Ensembl genome browser.

Gary moved into curation following the completion of his PhD at the University of Glasgow, where he employed a variety of phylogenomic and bioinformatic methods to investigate drug resistance in nematode parasites of human and livestock importance. ORCID iD: 0000-0002-7468-0008

Source URL: https://www.ebi.ac.uk/training/online/course/human-genetic-variation-ii-exploring-publicly-available-data

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/laura.emery
[4] https://www.ebi.ac.uk/training/online/trainers/jalm_12464
Human genetic variation (II): exploring publicly available data
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

[49] http://www.ebi.ac.uk/pdbe/entry/pdb/5e83/protein/2
[50] https://www.ebi.ac.uk/training/online/course/human-genetic-variation-ii-exploring-publicly-available-data/references/#genvar2_ref_1
[51] https://www.ebi.ac.uk/gwas/efotraits/EFO_0009260
[53] https://www.omim.org/
[55] http://cancer.sanger.ac.uk/cosmic
[56] http://www.ensembl.org/Homo_sapiens/Phenotype/Locations?ph=25242
[57] https://www.omim.org
[58] https://www.ebi.ac.uk/training/online/glossary/pdbe
[59] https://www.targetvalidation.org/
[60] https://www.targetvalidation.org/target/ENSG00000258839/associations
[61] http://europemc.org/articles/PMC2719288
[63] https://www.ebi.ac.uk/training/online/glossary/pmid
[65] https://www.ebi.ac.uk/training/online/course/europe-pmc-get-most-literature-searches
[67] https://www.ebi.ac.uk/ega/about/access
[68] https://www.ebi.ac.uk/ega/datap provid/EGA00000000001
[69] https://www.ebi.ac.uk/gwas/search?query=17554300
[70] https://www.ebi.ac.uk/training/online/course/human-genetic-variation-ii-exploring-publicly-available-data/references/genvar2_ref_2
[71] https://www.ebi.ac.uk/training/online/course/human-genetic-variation-ii-exploring-publicly-available-data/case-studies/case-study-2-search
[72] https://www.ebi.ac.uk/training/online/course/ensembl-browser-webinar-series-2016
[73] https://www.ebi.ac.uk/training/online/course/pdbe-searching-biological-macromolecular-structure
[74] http://www.ebi.ac.uk/support/
[77] https://www.ebi.ac.uk/training/online/glossary/uniprotkb