1. What are 3 data formats or data tiers of ENA (European Nucleotide Archive)?

Reads, assembly, annotation

1. Provide 4 examples of genome databases and briefly describe the content and structure of these databases.
2. What is an ontology and how does a terminology differ from an ontology?

- Ontology is definied as a formal representation of a knowledge domain; it describes the attributes (items, properties) of a particular knowledge domain and creates logical links between these terms. It is represented as a DAG (Directed Acyclic Graph).

- It differs from a terminology in the sense that the terminology only defines the term or the attributes where as ontology also creates logical connections and is specific to a domain

4. What is meant by curation? Provide at least 2 examples for curated databases and describe the essential curation principles that apply to these databases.

- curation is the process of standardizing, annotating and integrating information in such a way that the integrity of the data is maintained over time; curation also reduces redundancy of the data; examples: UniprotKB, Ensembl

- Essential principles of curation involve adding annotations to the data, checking the accuracy of the data, providing unification of semantics when data comes from different sources which also reduces redundancy

5. What are the server-side programs provided by NCBI to query GEO programmatically?

Hoffman review:

1. Establish a knowledge tree on gene or protein:

Ans:

Data warehousing: bringing all the data under one roof in a single database (FIG. 5).

The first step in data warehousing is to develop a unified data model that can accommodate all the information that is contained in the various source databases.

The next step is to develop a series of software programs that will fetch the data from the source databases, transform them to match the unified data model and then load them into the warehouse.

The warehouse can then be used as a ‘one-stop shop’ for answering any of the questions that the source databases can handle, as well as those that require integrated knowledge that the individual sources do not have.

**UniRef**

UniRef provides clustered sets of sequences from UniProtKB (including additional isoforms contained within UniProtKB/Swiss-Prot records) and selected UniParc records. UniRef offers complete coverage of the sequence space at three resolutions:

The **UniRef100** database combines identical sequences from any organism into a single UniRef entry, displaying the sequence of a representative protein, the accession numbers of all the merged entries and links to the corresponding UniProtKB and UniParc records.

**UniRef90** is built by clustering UniRef100 sequences such that each cluster is composed of sequences that have at least 90% sequence identity to the longest sequence (the seed sequence) of the cluster.

**UniRef50**is built by clustering UniRef90 seed sequences that have at least 50% sequence identity to the longest sequence in the cluster.

More than 95% of the protein sequences provided by UniProtKB come from the translations of coding sequences ([CDS](https://www.ebi.ac.uk/training/online/glossary/cds)) submitted to the ENA/GenBank/DDBJ nucleotide sequence resources of the International Nucleotide Sequence Database Collaboration (INSDC).

These CDS are either generated by gene prediction programs or are experimentally proven. The translated CDS sequences are automatically transferred to the TrEMBL section of UniProtKB. The TrEMBL records can be selected for further [manual annotation](https://www.ebi.ac.uk/training/online/glossary/manual-annotation) and then integrated into the [UniProtKB/Swiss-Prot](https://www.ebi.ac.uk/training/online/glossary/uniprotkbswiss-prot) section.

In addition to translated CDS, UniProtKB protein sequences may come from:

* The [PDB](https://www.ebi.ac.uk/training/online/glossary/pdb) database;
* Sequences experimentally obtained by direct protein sequencing and submitted to [UniProt](https://www.ebi.ac.uk/training/online/glossary/uniprot);
* Sequences scanned from the literature;
* Sequences derived from gene prediction but which have not been submitted to ENA/GenBank/DDBJ. These are imported from resources such as [Ensembl](https://www.ebi.ac.uk/training/online/glossary/ensembl) and [RefSeq](https://www.ebi.ac.uk/training/online/glossary/refseq).

1. Notion of a database:
2. How to view variation data in ensemble?

The Ensembl Variation database stores areas of the genome that differ between individual genomes ("variants") and, where available, associated disease and phenotype information.  
There are different types of variants for several [species](http://www.ensembl.org/info/genome/variation/data_description.html#source):

* single nucleotide polymorphisms (SNPs)
* short nucleotide insertions and/or deletions
* longer variants classified as structural variants (including CNVs)

We predict the effects of variants on the Ensembl transcripts and regulatory features for each species. You can run the same analysis on your own data using the [Variant Effect Predictor](http://www.ensembl.org/info/docs/tools/vep/index.html).   
These data are integrated with other data sources in Ensembl, and can be accessed using the API (see links on the right handside menu) or website.

Find short variants for a gene from dbSNP, COSMIC, and [other sources](http://www.ensembl.org/info/genome/variation/sources_documentation.html) in the [Gene Variation table](http://www.ensembl.org/Homo_sapiens/Gene/Variation_Gene/Table?g=ENSG00000139618;r=13:32889611-32973805).