

DTU





**DTU Health Technology
Bioinformatics**

Association analysis

*Gisle Vestergaard
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gisves@dtu.dk*

Cohort based studies

GWAS: Genome-wide association study

MWAS: Metagenomice-wide association study



Fig. by NHS:
<https://www.genomiceducation.hee.nhs.uk>

How researchers compare genomic information to identify genetic alterations

GWAS

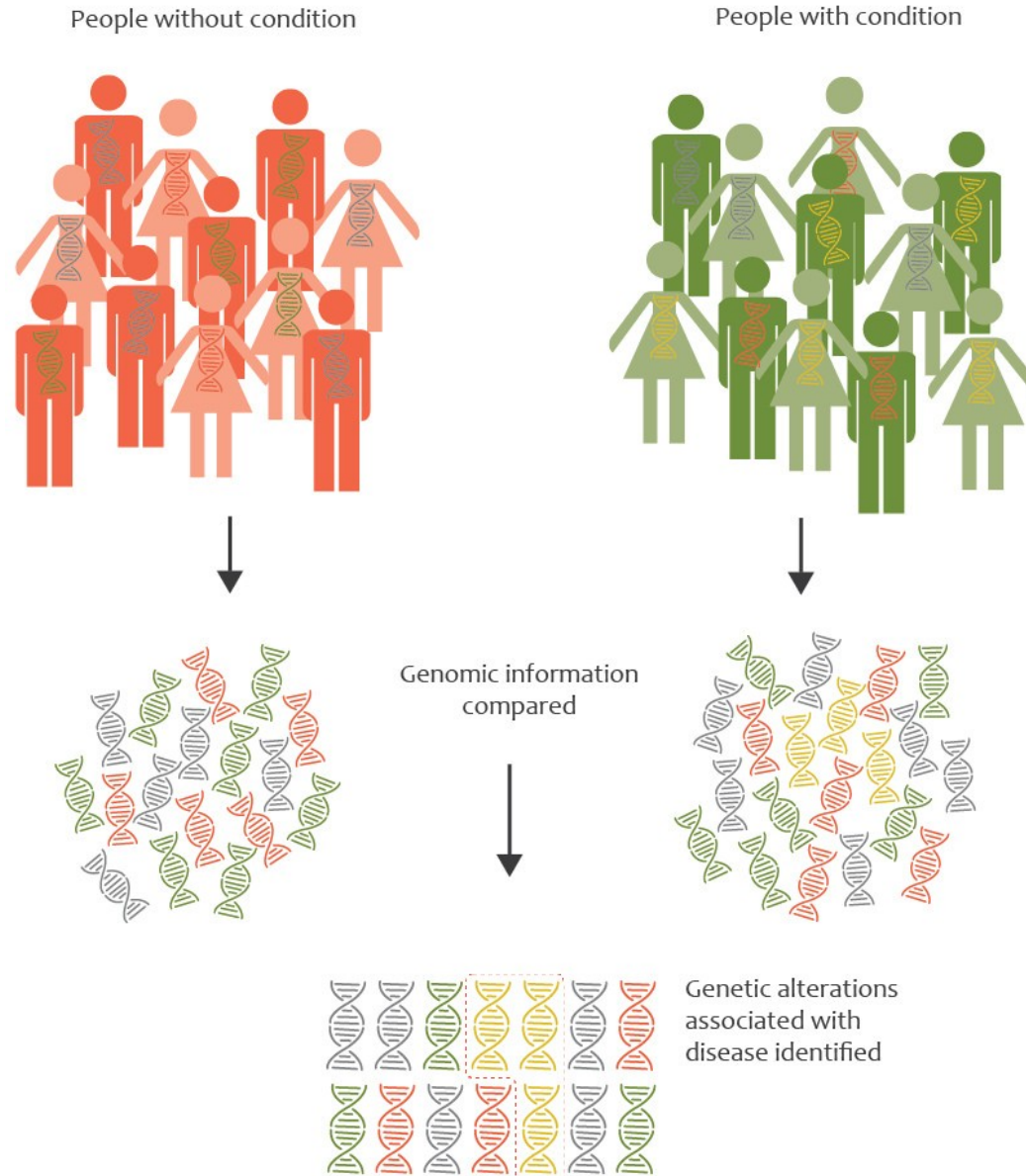
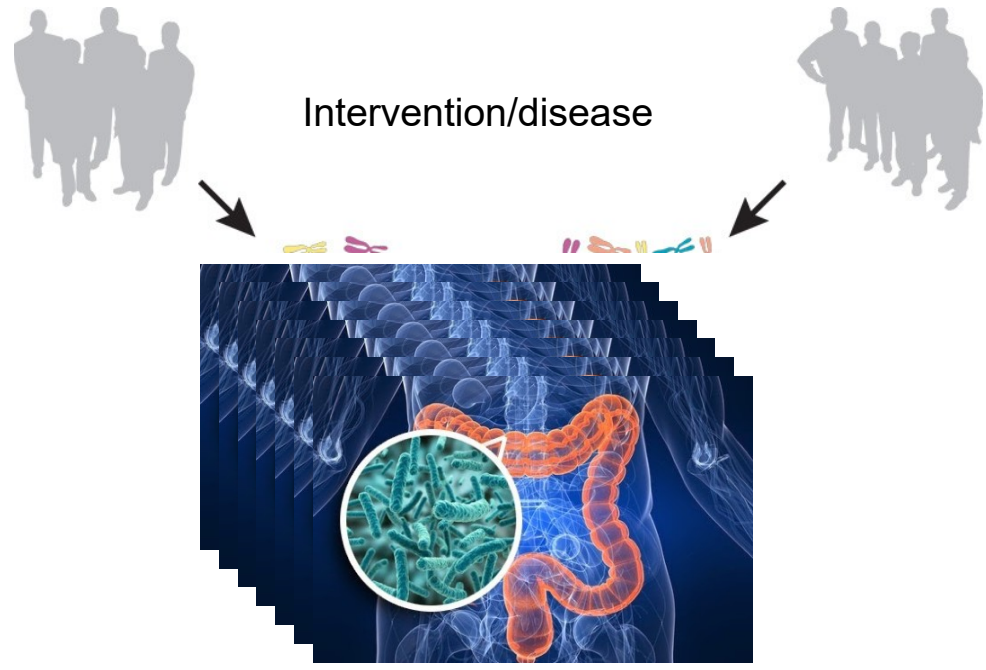



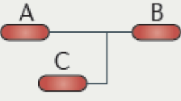


Fig. by NHS:
<https://www.genomicseducation.hee.nhs.uk>

MWAS



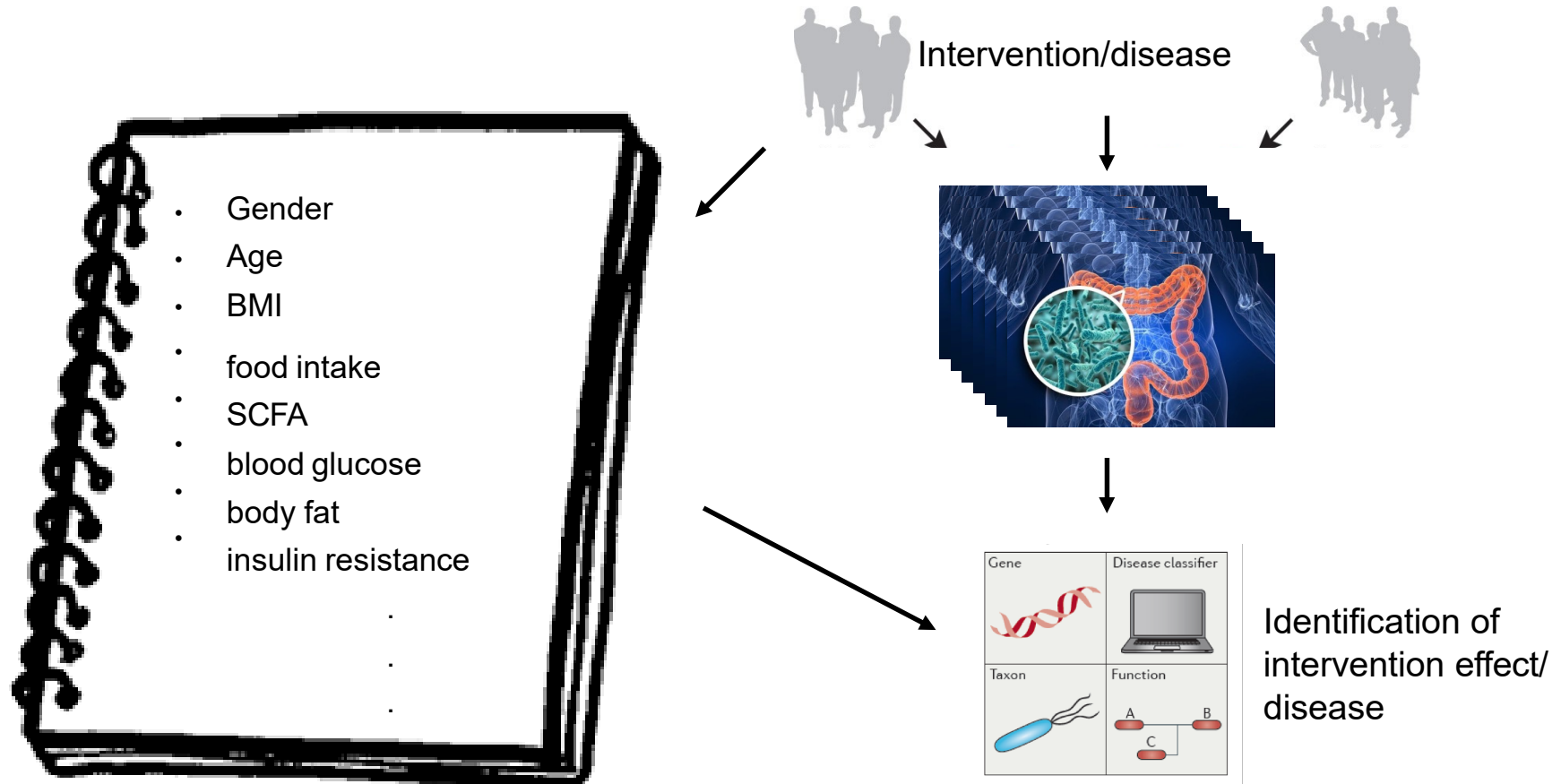
Gene 	Disease classifier 
Taxon 	Function 

Identification of intervention/disease

MWAS

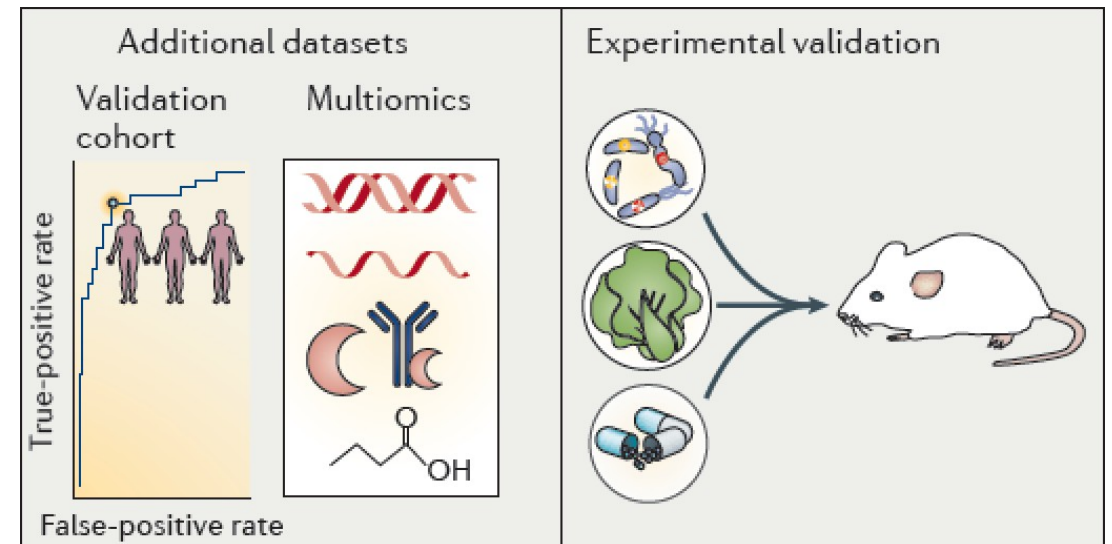
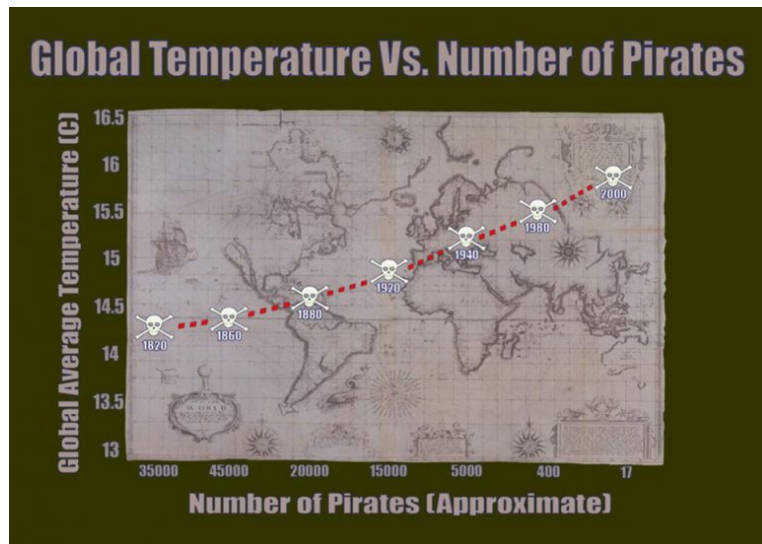
“The associations that can be identified by MWAS are not limited to the identification of taxa that are more or less abundant, as is the case with taxonomic approaches, but additionally include the **identification of microbial functions** that are enriched or depleted.” *Wang and Jia (2016)*

Metadata association



Remember

- **Correlation does not imply causation**
- Validation
 - Literature
 - Follow-up studies
 - Otheromics: metatranscriptomics, metaproteomics, metabolomics
 - Longitudinal studies
 - Experimental evidence such as animal models



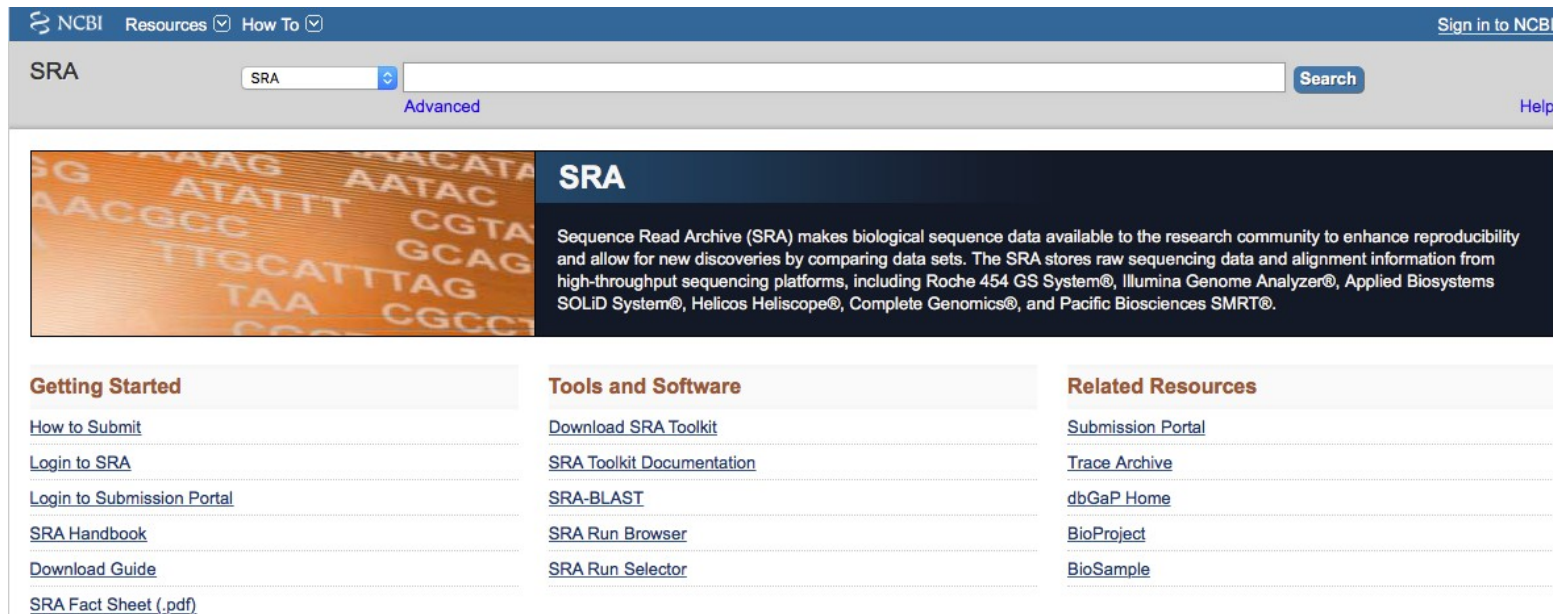
Good to know: Functions and Public Databases

- Finding data that can help answer your question



SRA

- SRA:
 - short read archive (part of NCBI)
 - <https://www.ncbi.nlm.nih.gov/sra/>



The screenshot shows the NCBI SRA website homepage. At the top, there is a navigation bar with "NCBI Resources" and "How To" menus, and a "Sign in to NCBI" link. Below this is a search bar with "SRA" entered and a "Search" button. A "Help" link is also visible. The main content area features a large banner with a background image of DNA sequence data and the text: "SRA Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®." Below the banner are three columns of links: "Getting Started" (How to Submit, Login to SRA, Login to Submission Portal, SRA Handbook, Download Guide, SRA Fact Sheet (.pdf)), "Tools and Software" (Download SRA Toolkit, SRA Toolkit Documentation, SRA-BLAST, SRA Run Browser, SRA Run Selector), and "Related Resources" (Submission Portal, Trace Archive, dbGaP Home, BioProject, BioSample).

Searching SRA

<https://www.ncbi.nlm.nih.gov/sra/advanced>

SRA Advanced Search Builder

Builder

Organism

Using one of the available fields to narrow the search by "Organism"

Hide index list

Index: available values for the field "Organism" and number of records

Previous 200

Next 200

Refresh index

Add new builder line

Show index list

Choosing boolean operator

AND

Accession

All Fields

Author

BioProject

BioSample

Filter

Modification Date

Organism

Properties

Publication Date

Text Word

Title

#2

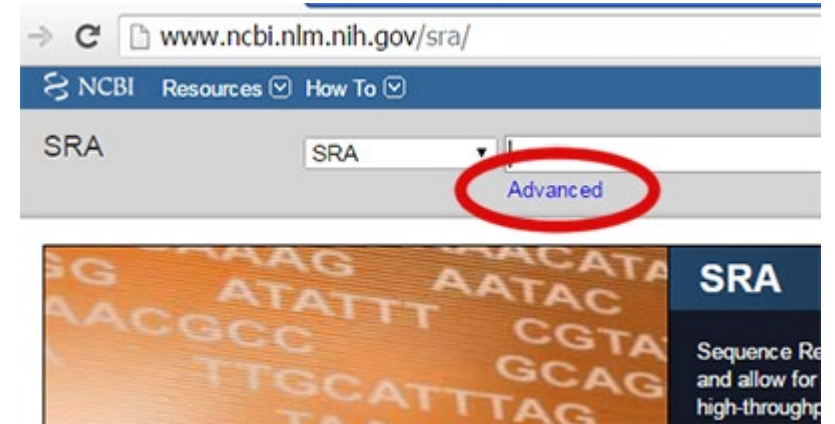
Using search # from history for new query

All available fields where Filter and Properties can be especially useful

Download history

Clear history





Search	Add to builder	Query	Items found	Time
#2	Add	Search (mus musculus[Organism]) AND tumor cell line[Text Word]	52	13:50:47



ENA

- ENA:
 - European Nucleotide Archive
 - <https://www.ebi.ac.uk/metagenomics/> (metagenomic datasets)

The screenshot shows the EBI Metagenomics website. At the top left is the EMBL-EBI logo. The main header features the 'EBI Metagenomics' logo and a search bar with a 'Search' button. A navigation menu below the header includes 'Home', 'Search', 'Submit data', 'Projects', 'Samples', 'Comparison tool', 'About', and 'Contact'. On the right side of the navigation menu, it says 'Not logged in' and 'Login'. The main content area has a dark blue background with the text 'Submit, analyse, visualize and compare your data.' and a prominent 'SUBMIT DATA' button. Below this, there are four columns of statistics:

 64649 data sets	 49902 amplicons 103 assemblies 688 metabarcoding 12897 metagenomes 1059 metatranscriptomes	 Public 61572 runs 43960 samples 735 projects	 Private 2966 runs 2703 samples 131 projects
---	--	---	--

Searching ENA

- ENA:
 - European Nucleotide Archive
 - <https://www.ebi.ac.uk/metagenomics/> (metagenomic datasets)

By selected biomes



Soil (379)



Host-associated
human (83)



Human digestive
system (64)



Engineered (55)



Host-associated
mammals (54)



Marine (46)



Host-associated
plant (44)



Forest soil (26)



Freshwater (21)



Grassland (19)

Functional annotations

GO: Gene ontology

- defines concepts/classes used to describe gene function, and relationships between these concepts

KEGG: Kyoto Encyclopedia of Genes and Genomes

- database resource for understanding high-level functions and utilities of the biological system

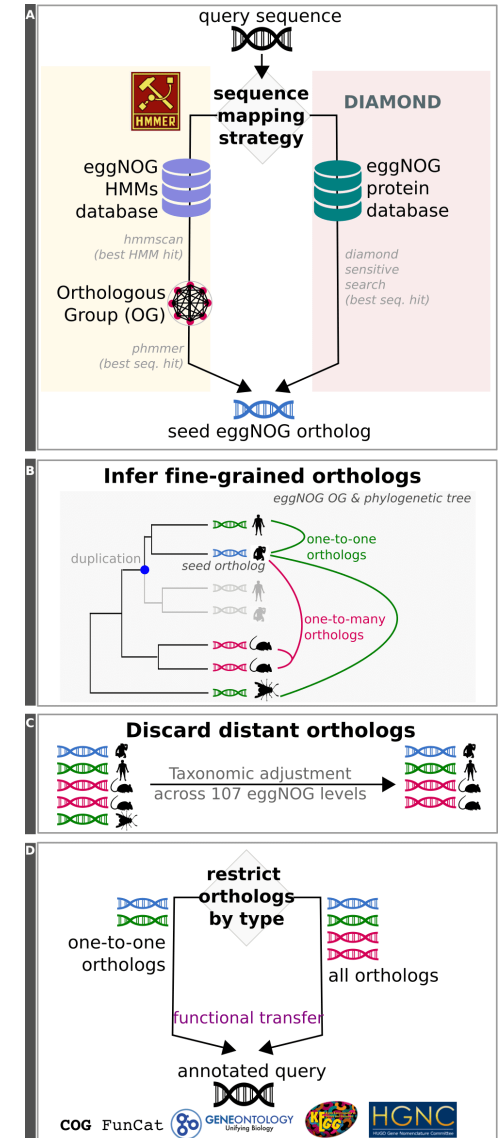
eggNOG

- A database of orthologous groups and functional annotation
- Cross-references to other databases

EggNOG-mapper

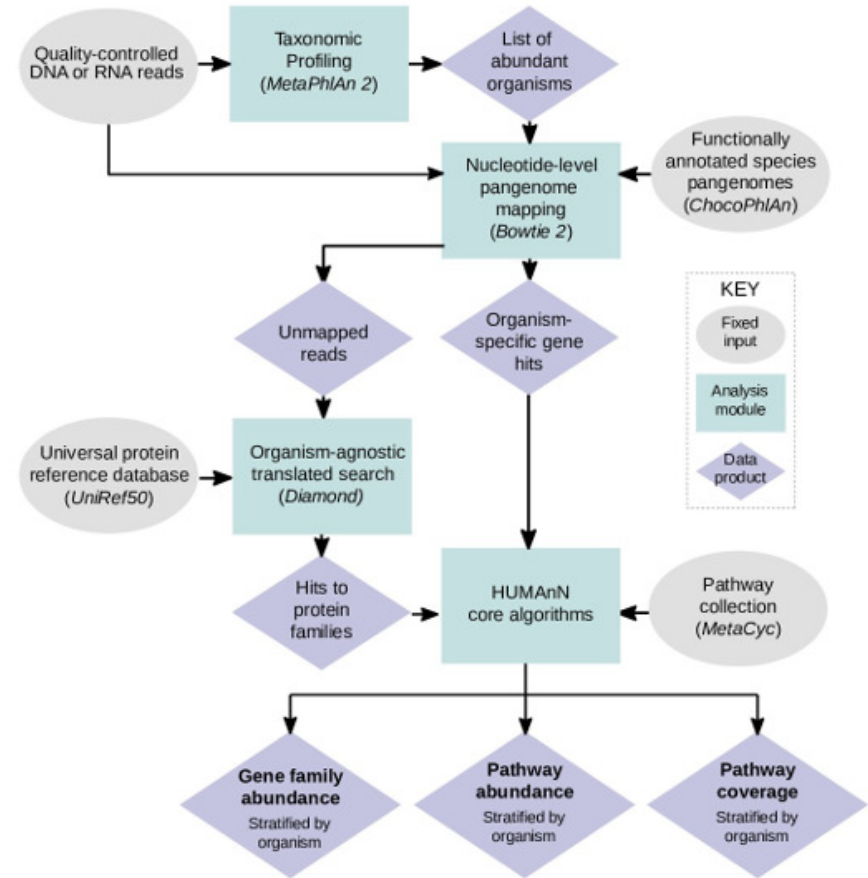
1. query_name
2. seed eggNOG ortholog
3. seed ortholog evaluate
4. seed ortholog score
5. Predicted taxonomic group
6. Predicted protein name
7. Gene Ontology terms
8. EC number
9. KEGG_ko
10. KEGG_Pathway
11. KEGG_Module
12. KEGG_Reaction
13. KEGG_rclass
14. BRITE
15. KEGG_TC
16. CAZy
17. BiGG Reaction
18. tax_scope: eggNOG taxonomic level used for annotation
19. eggNOG OGs
20. bestOG (deprecated, use smallest from eggnog OGs)
21. COG Functional Category
22. eggNOG free text description

- Functional annotation of reads or ORFs
- Cross-references to other databases
- Complex tabulated output



HUMAnN2

- Very easy to install and use



**How to hands on:
Association analysis exercise**