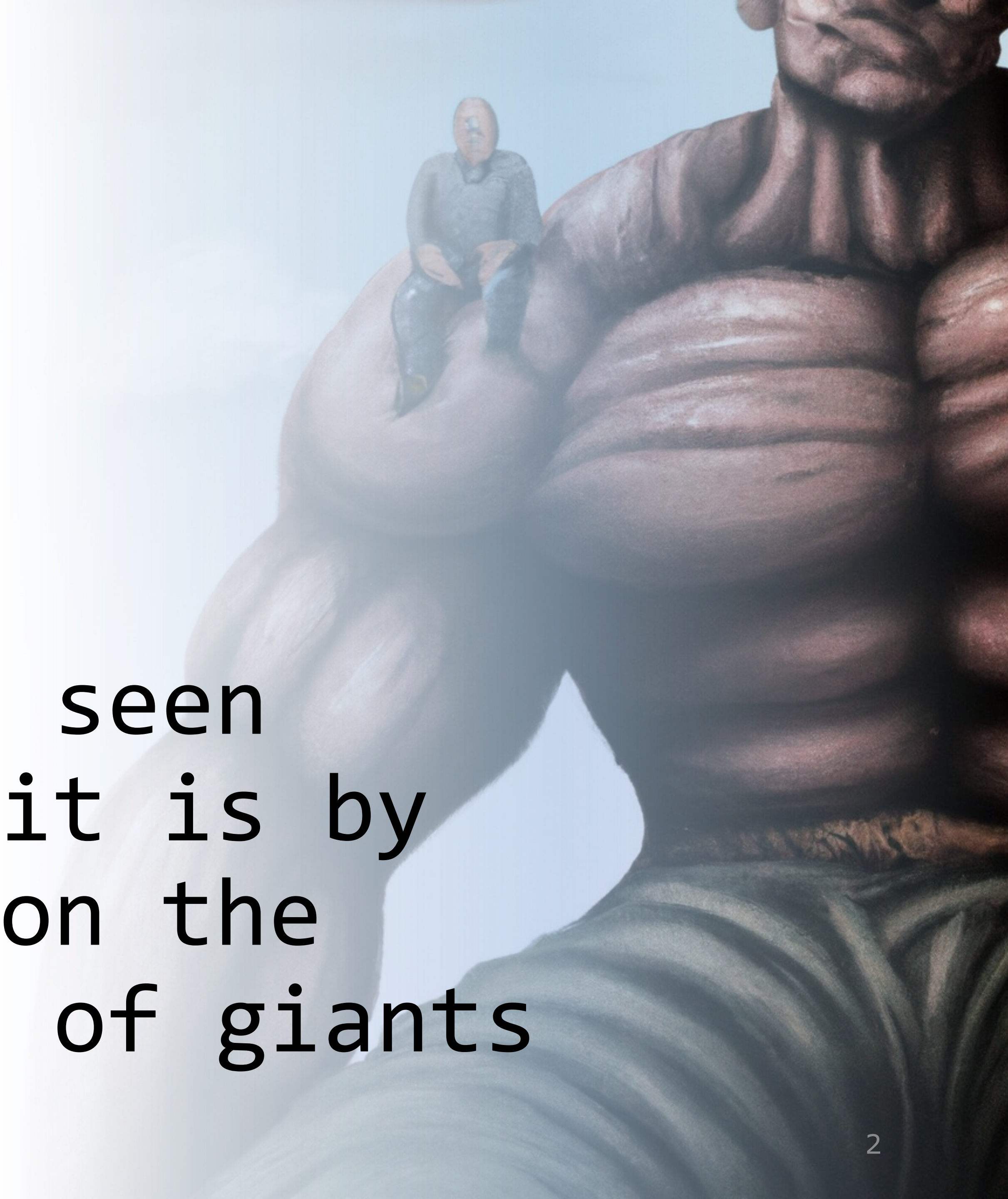
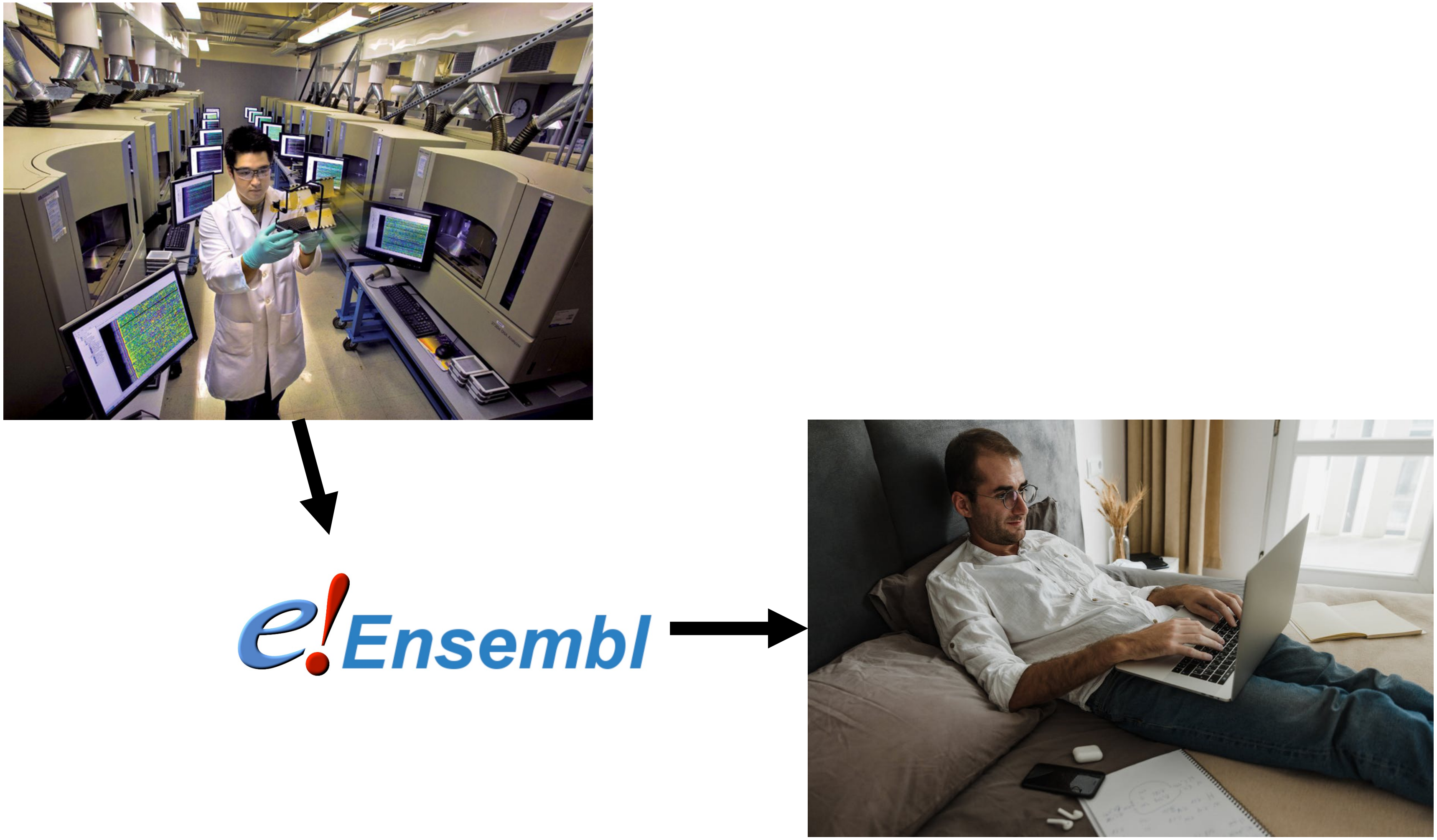
Introduction to raw sequence databases

Geert van Geest
Interfaculty Bioinformatics Unit, UniBe
Training group, SIB



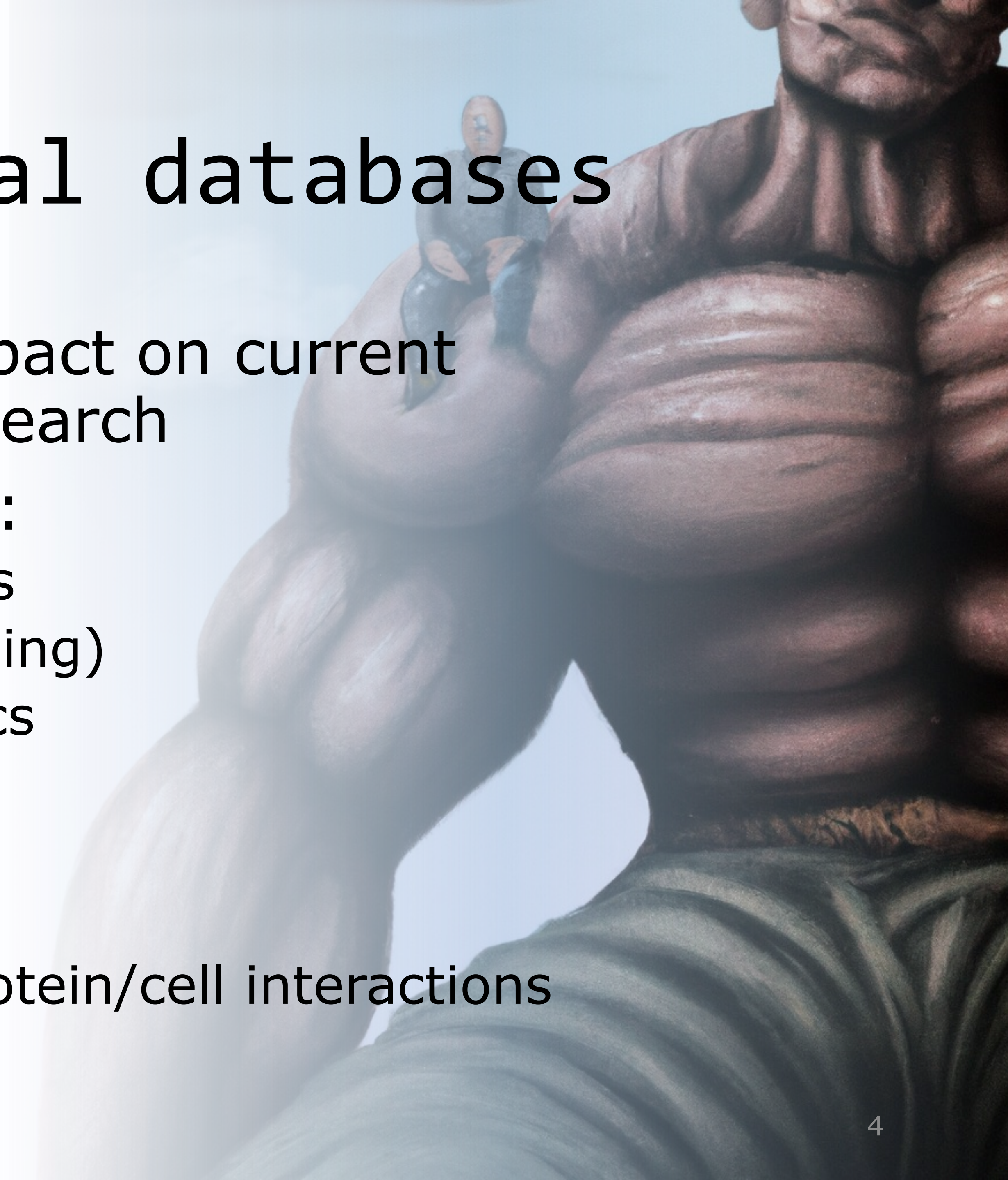
If I have seen further, it is by standing on the shoulders of giants

Current giant-shoulder-standing



Biological databases

- Immense impact on current biological research
- Databases in:
  - Nucleic acids
  - Protein (folding)
  - Metabolomics
  - Taxonomy
  - Imaging
  - Cell lines
  - Molecule/protein/cell interactions
  - …



Question

Biological sequence databases

- **Proteins**: UniProtKB/Swiss-Prot, InterPro
- **Genomes + annotations**: Ensembl, ENA, GenBank/RefSeq, UCSC, ENCODE
- **Raw sequencing data**: INSDC Sequence read archives of ENA, NCBI and DDBJ

What is a biological database?

- Organizes and standardizes biological information
- (Curated) addition and modification
- Quick searches
- Access by the community through APIs

FAIR principles

- **F**indable, **A**ccessible, **I**nteroperable, **R**eusable
- To ensure transparency, reproducibility, and reusability
- Enables reuse by:
  - People – same data, other questions
  - Machines - database connections, meta-analyses etc.
- Storage in biological databases typically makes data FAIR

https://www.go-fair.org/fair-principles/

INSDC

International Nucleotide Sequence Database Collaboration

Data type	DDBJ	EMBL-EBI	NCBI
Next Generation reads	Sequence Read Archive	European Nucleotide Archive	Sequence Read Archive
Assembled Sequences	DDBJ		GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

DDBJ
DNA Data Bank of Japan

ENA
European Nucleotide Archive

NIH NLM
National Center for Biotechnology Information

Other raw sequencing data portals

- Genome Sequence Archive (Chinese)
- Human controlled access:
  - European Genome Phenome Archive (EGA)
  - JGA – Japan
  - dbGaP – US
- Expression data: ArrayExpress, GEO
- Metagenomics: MGnify

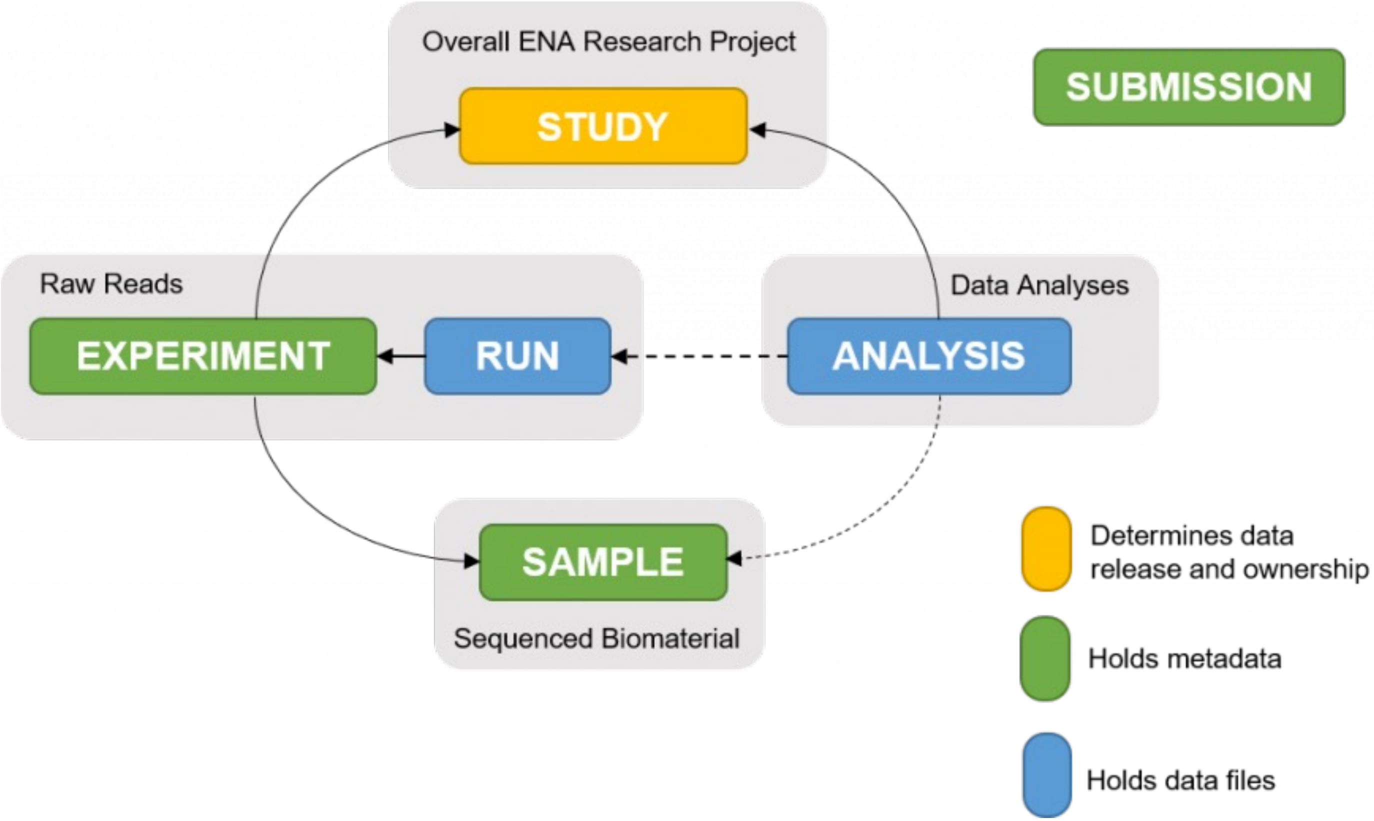
Other portals

- Use the data submission wizard: https://www.ebi.ac.uk/submission/

Databases are interconnected

- INSDC databases are interconnected
- Ensembl uses ENA, RefSeq and UniProtKB for its annotations
- ArrayExpress and GEO submit to their respective SRA
- RefSeq is based on INSDC
- dbGaP and EGA are interconnected
- …

ENA database



Searching ENA

- Raw reads, sequences, assemblies
- Based on any metadata column, like:
  - Organism
  - Geographic origin
  - Sequencing method
  - ..

Downloading sequence data

- Through the browser
- File Transfer Protocol (FTP):
  - Command line: wget or curl
  - FileZilla
- SRA tools
- https://nf-co.re/fetchngs