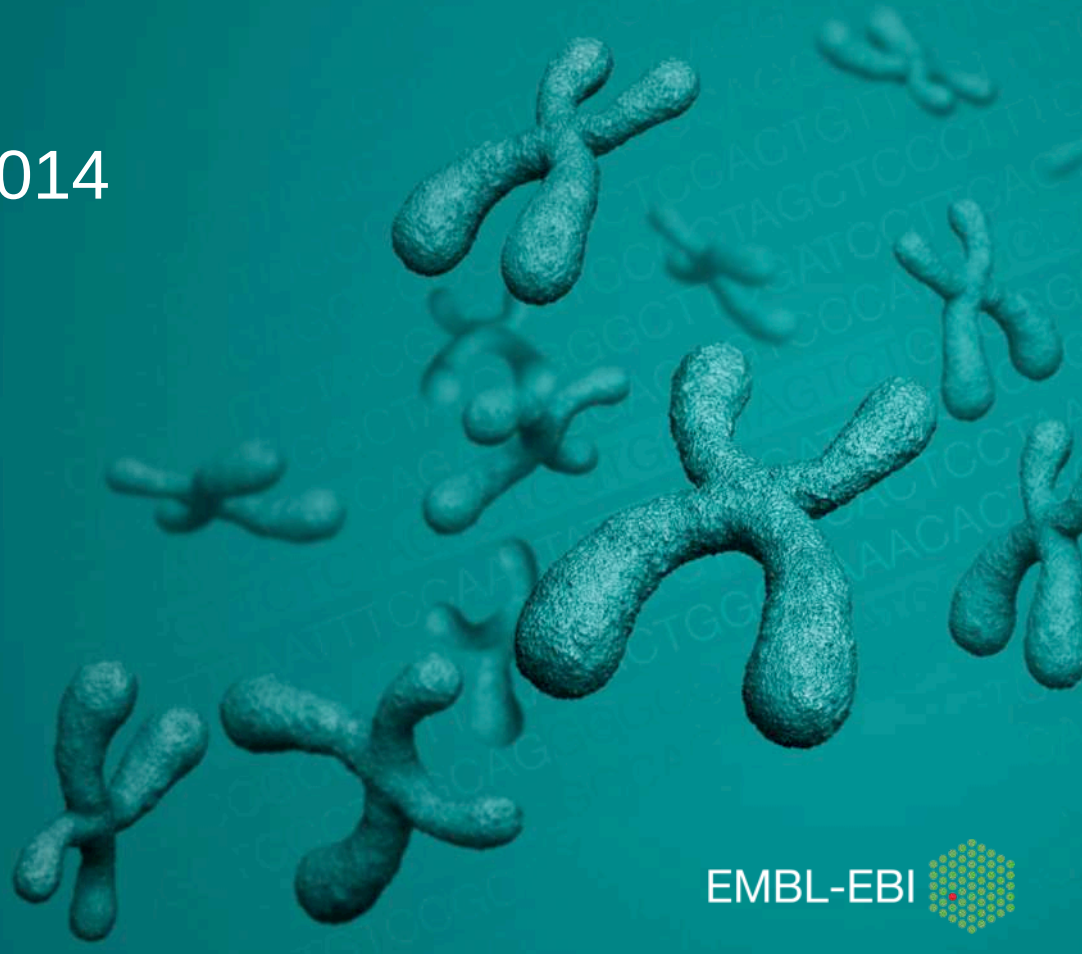
The European Bioinformatics Institute: Access to Data

RECODE workshop

Amsterdam 14 March 2014

www.ebi.ac.uk



EMBL-EBI

What is EMBL-EBI?

- Part of the European Molecular Biology Laboratory
- International, non-profit research institute
- Europe’s hub for biological data services and research
- 550 members of staff from 53 nations.

Data resources at EMBL-EBI

Genes, genomes & variation

European Nucleotide Archive
1000 Genomes
Ensembl
Ensembl Genomes
European Genome-phenome Archive
Metagenomics portal

Gene, protein & metabolite expression

ArrayExpress
Expression Atlas
Metabolights
PRIDE

Literature & ontologies

Europe PubMed Central
Gene Ontology
Experimental Factor Ontology

Protein sequences, families & motifs

InterPro
Pfam
UniProt

Molecular structures

Protein Data Bank in Europe
Electron Microscopy Data Bank

Chemical biology

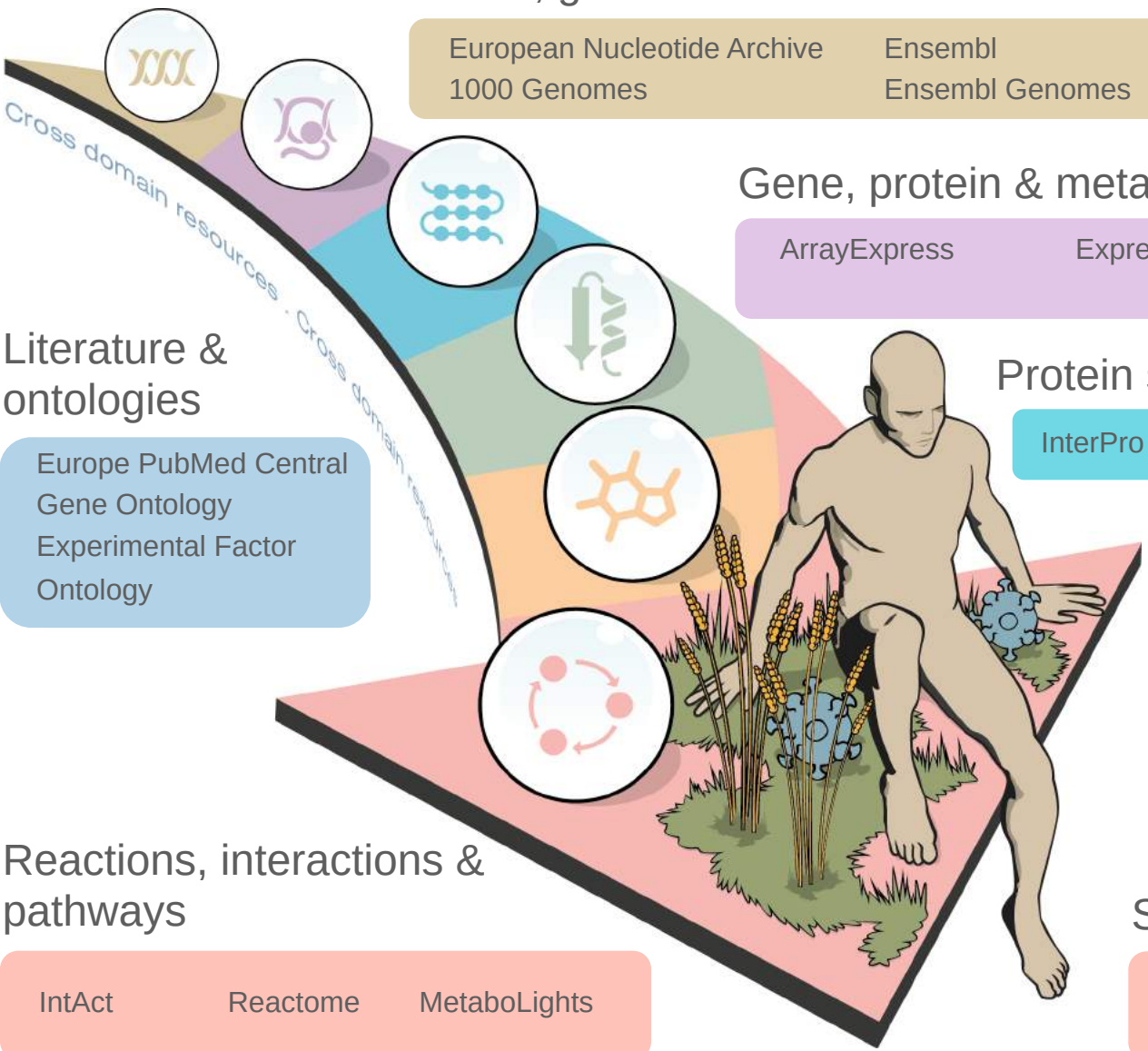
ChEMBL
ChEBI

Reactions, interactions & pathways

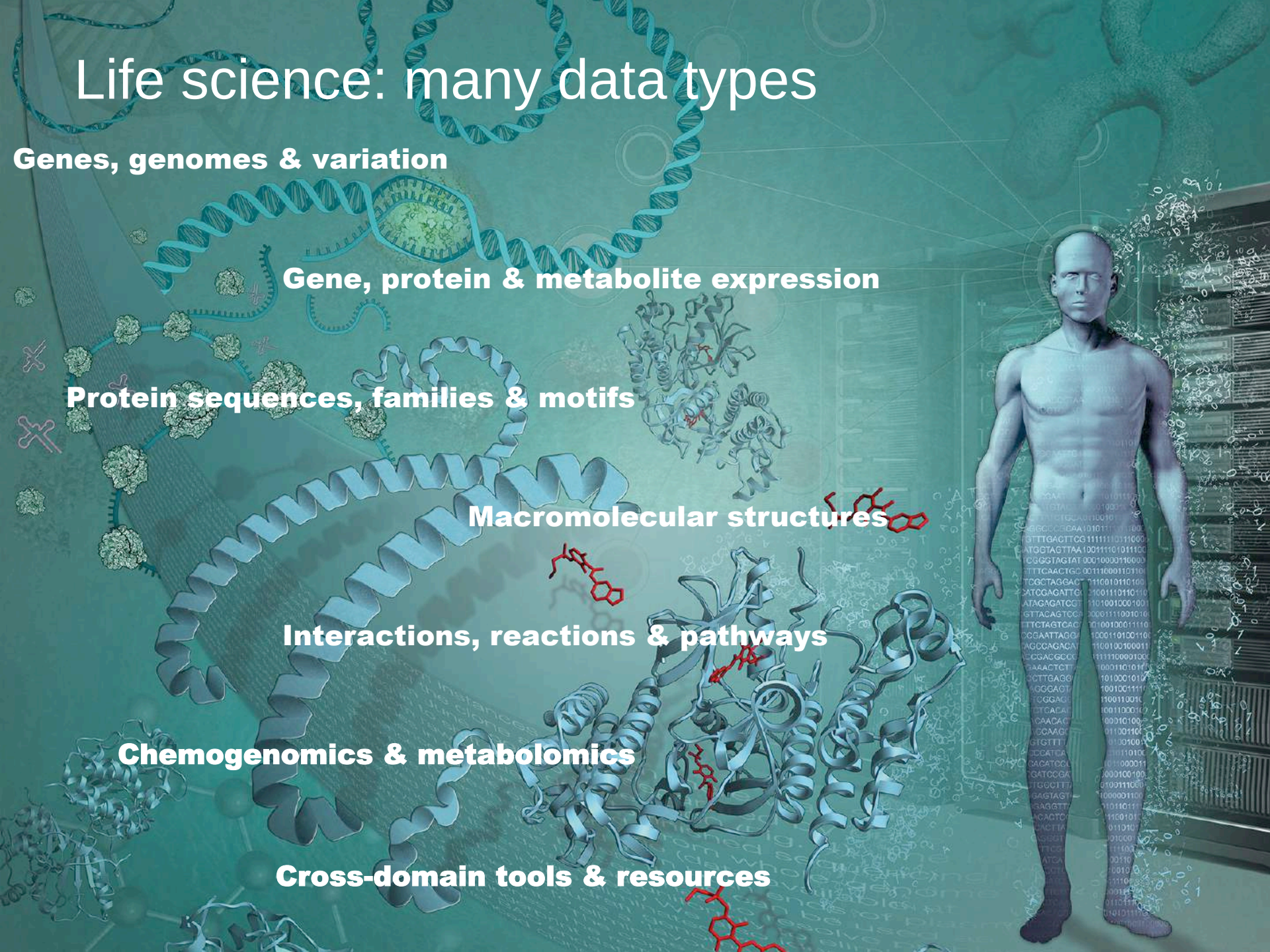
IntAct
Reactome
MetaboLights

Systems

BioModels
BioSamples
Enzyme Portal



Life science: many data types



www.ebi.ac.uk/services

EMBL-EBI | Services | Research | Training | Industry | About us

Services

Overview | A to Z | Service teams | Support

Bioinformatics services

We maintain the world's most comprehensive range of **freely available** and up-to-date molecular databases. Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our web services to access our resources programmatically.

- **DNA & RNA** — genes, genomes & variation
- **Gene expression** — RNA, protein & metabolite expression
- **Proteins** — sequences, families & motifs
- **Structures** — Molecular & cellular structures
- **Systems** — reactions, interactions & pathways
- **Chemical biology** — chemogenomics & metabolomics
- **Ontologies** — taxonomies & controlled vocabularies
- **Literature** — Scientific publications & patents
- **Other software** — cross-domain tools & resources

Popular

- Ensembl
- BLAST
- UniProt
- Literature
- PDBe
- Train online
- ArrayExpress
- Support

Bioinformatics training

Guide to resources

EMBL-EBI data sources

- Data generated by scientists
- Worldwide: many funders require that data generated with public funds be deposited in open access repositories
- EMBL-EBI receives these data
- For many data types publishers require accession numbers from public repositories before publication
- Much data automatically deposited
  - E.g. from sequencing centers

Data policies

- (Almost) all data and all software are freely available for unrestricted use worldwide
  - Private for limited time until publication
  - Many data access methods provided
    - Web based
    - ftp
    - Programmatic access: REST, SOAP
    - RDF/OWL for some data
- Services are also freely available for unrestricted use worldwide

Terms of use

- Sample terms/licences/policies
  - http://www.ebi.ac.uk/about/terms-of-use
  - http://www.uniprot.org/help/license
  - http://www.ensembl.org/info/about/legal/code_licence.html
  - https://www.ebi.ac.uk/arrayexpress/help/data_availability.html
  - http://www.wwpdb.org/wwpdb_charter.html
  - http://www.ebi.ac.uk/biomodels-main/termsofuse
  - PRIDE: http://code.google.com/p/ebi-pride/
  - https://www.ebi.ac.uk/chebi/aboutChebiForward.do

Software licences

EMBL-EBI produces much software

- Freely distributable provided credit and copyright are maintained

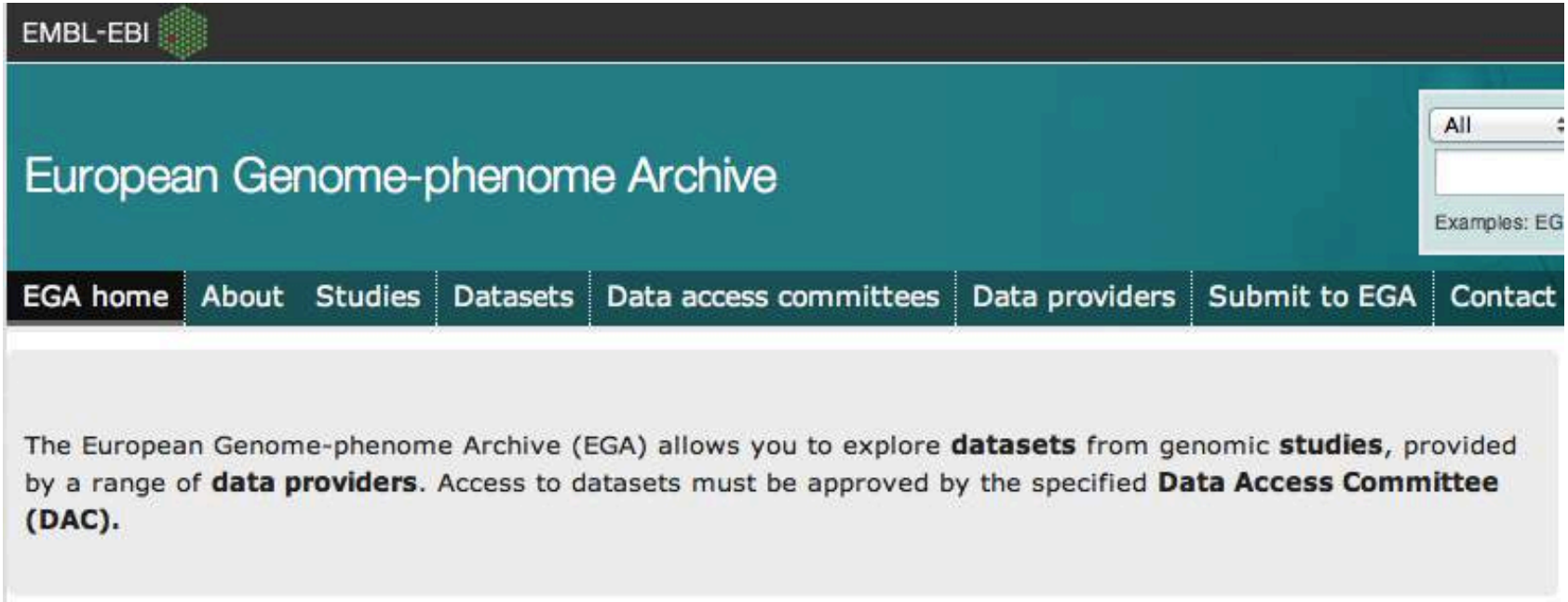
Often an Apache license:

```
Copyright (c) EMBL European Bioinformatics Institute, Hinxton, Cambridge,
UK. (http://www.ebi.ac.uk) The InterProScan software itself is provided
under the Apache License, Version 2.0
(http://www.apache.org/licenses/LICENSE-2.0.html). Third party components
(e.g. member database binaries and models) are subject to separate
licensing - please see the individual member database websites for
details.
```

http://code.google.com/p/interproscan/wiki/InterProScan5HowToRun

European Genome-phenome Archive

- Holds EMBL-EBI data that are not freely accessible
- www.ebi.ac.uk/ega



EGA data

- EMBL-EBI does not hold medical records
- Does hold human data (clinical or research), de-identified and consented for research
- Original submitters retain control of the data in their own data sets
  - Consent terms differ for different data sets
- Access to data granted by original submitter
  - Data access application through EMBL-EBI
- Data accesses by direct download or by ftp/aspera
  - Distributed using individual encryption key
  - Some large datasets difficult to download

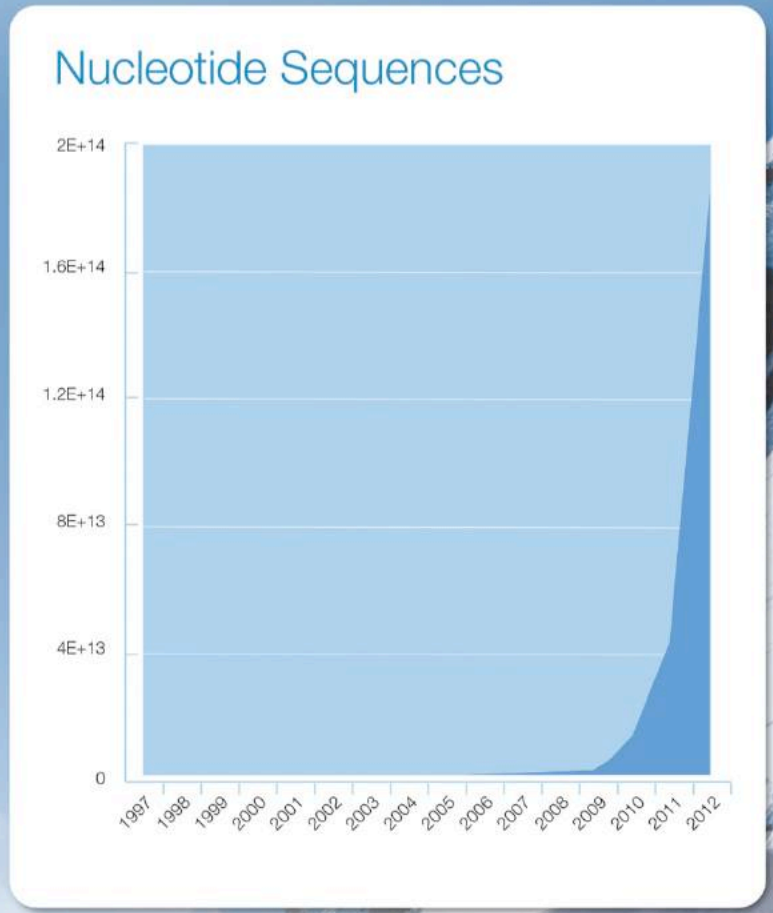
Clinical data: other access models

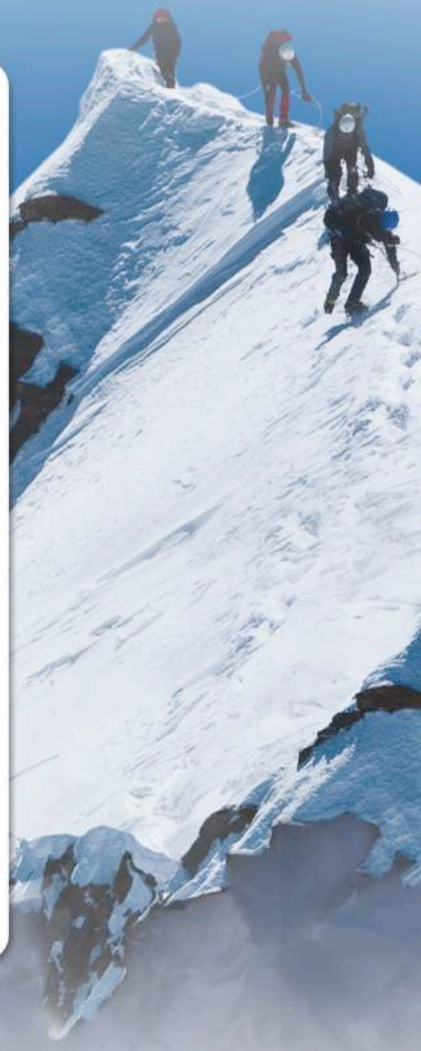
- Data submitted to U.S. NIH
  - Control passes to NIH
  - NIH data access committee grants/denies access
  - E.g. the Cancer Genome Atlas
    - http://cancergenome.nih.gov

Misuse of data?

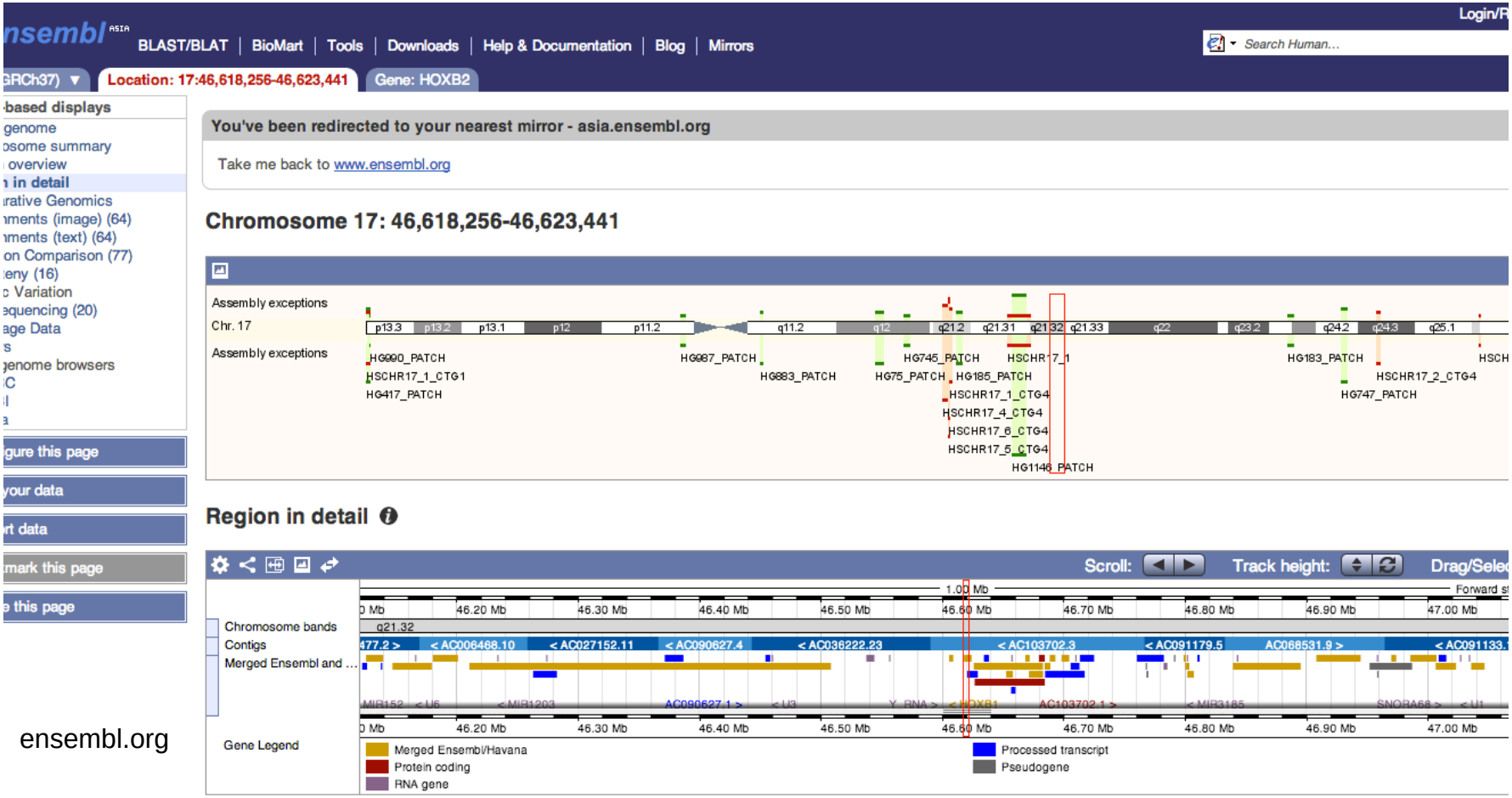
- Molecular data difficult to misuse
- Benefits of public access to data type held by EMBL-EBI far outweigh any potential drawbacks

Bigger and bigger data





Questions?



ensembl.org

EMBL-EBI