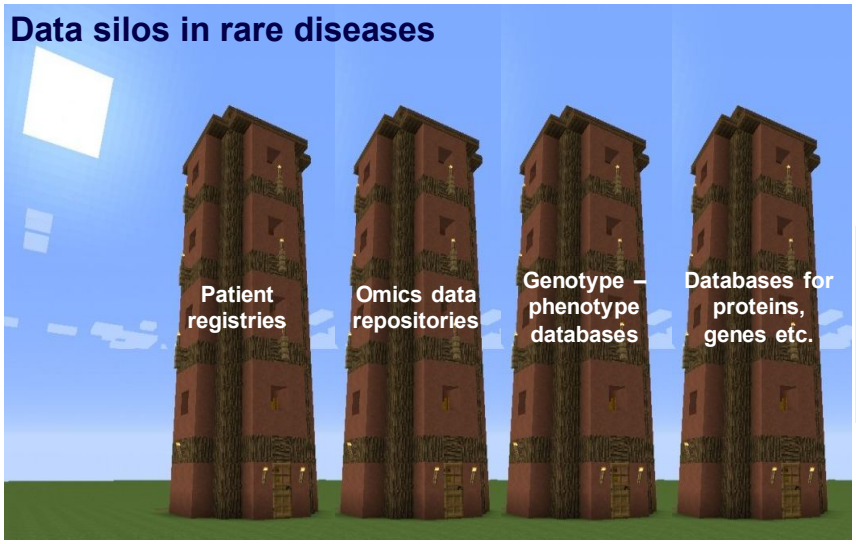
Molecular data for rare diseases: The added value of data interoperability

Friederike Ehrhart[1,2], Annika Jacobsen[3], Salvador Capella[4], Nasim Bahram Sangani[1,2], Eric Smeets[2], Marco Roos[3], Chris T Evelo[1,2], Leopold MG Curfs[2]

[1] Department for Bioinformatics – BiGCaT, Maastricht University, The Netherlands; [2] Rett Expertise Centre – Governor Kremers Centre (GKC), Maastricht University Medical Centre, The Netherlands [3] Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. [4] Barcelona Supercomputing Centre, Barcelona University, Barcelona, Spain

What we have



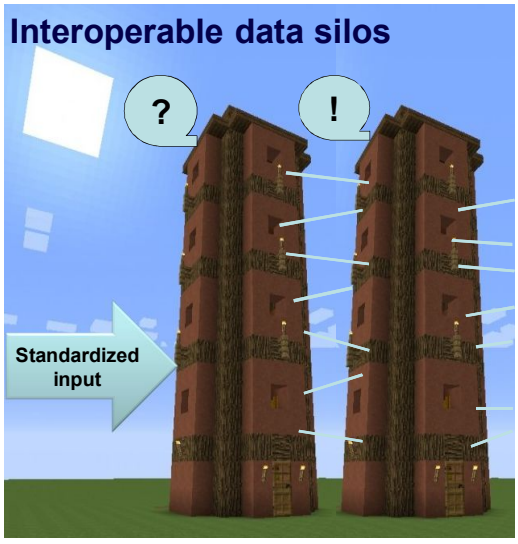
Patient registries
- Patient data - limited access
- Offline collections at clinicians or clinical researchers
- Interoperability status often not good
- No standard formats – very different form and input

Omics data repositories
- Anonymous
- Complete genomes/ transcriptomes etc.
- Large bulky datasets
- Different formats available for fast call/retrieval of data

Genotype-Phenotype databases
- Anonymous genetic variants linked to a phenotype (usually a disease) or labelled disease causing (or benign)
- Different standards for variant and phenotype description
- Often linked to anonymous patient ID

Databases for single molecular entities
- Possibly linked to original patient/source
- Cross links with other similar databases (e.g. bridgeDB) often available

What we want

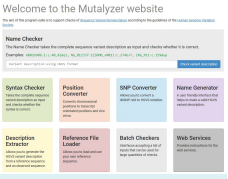


What are the problems?

The way rare disease data is captured and put in databases and database set up:
- Different input information, different minimum information standards
- Broad variety of formats – those which are currently used and historic formats (e.g. HGVS, RS)
- Consequent application of these formats
- Different reference sequences or reference SNPs
- Low degree of interoperability of phenotypic data (use of ontologies), disease information or pathogenicity scores
- Re-use or data accession permission often hidden

Possible solutions

- Instead of trying to enforce the use of one format improvement of tools which can translate between them e.g. Mutalyzer
- Manual curation is still needed (for about 10% of data)
- Implementation and enforcement of FAIR principles
- Improve Meta-database approaches (beacon) (Fairsharing.org)



https://mutalyzer.nl/

What would be the benefit?

Patient registries
- Correlation of "Natural history" e.g. life expectancy, longitudinal health profiles, with genetic data and other omics data profiles

Omics data repositories
- Comparative genomics/ transcriptomics etc.
- Identification of pathogenic and benign variants and correlation with phenotype information
- Genetic background – identification of modifier genes

Genotype-phenotype databases
- Correlation studies variation/phenotype
- Sorting pathogenic from benign enabling diagnosis
- Supporting genetic counselling

Databases for single molecular entities
- In depth understanding of pathogenic mechanisms
- Protein/gene function
- Adding detailed information to patient phenotypes

This work has been funded by ELIXIR and Stichting Terre.

Correspondence to:
friederike.ehrhart@maastrichtuniversity.nl
ORCID: 0000-0002-7770-620X
http://www.bigcat.unimaas.nl

Department Bioinformatics - BiGCaT
T +3143 3882913
F +3143 3881996

Maastricht University
P.O. Box 616
6200 MD Maastricht, The Netherlands