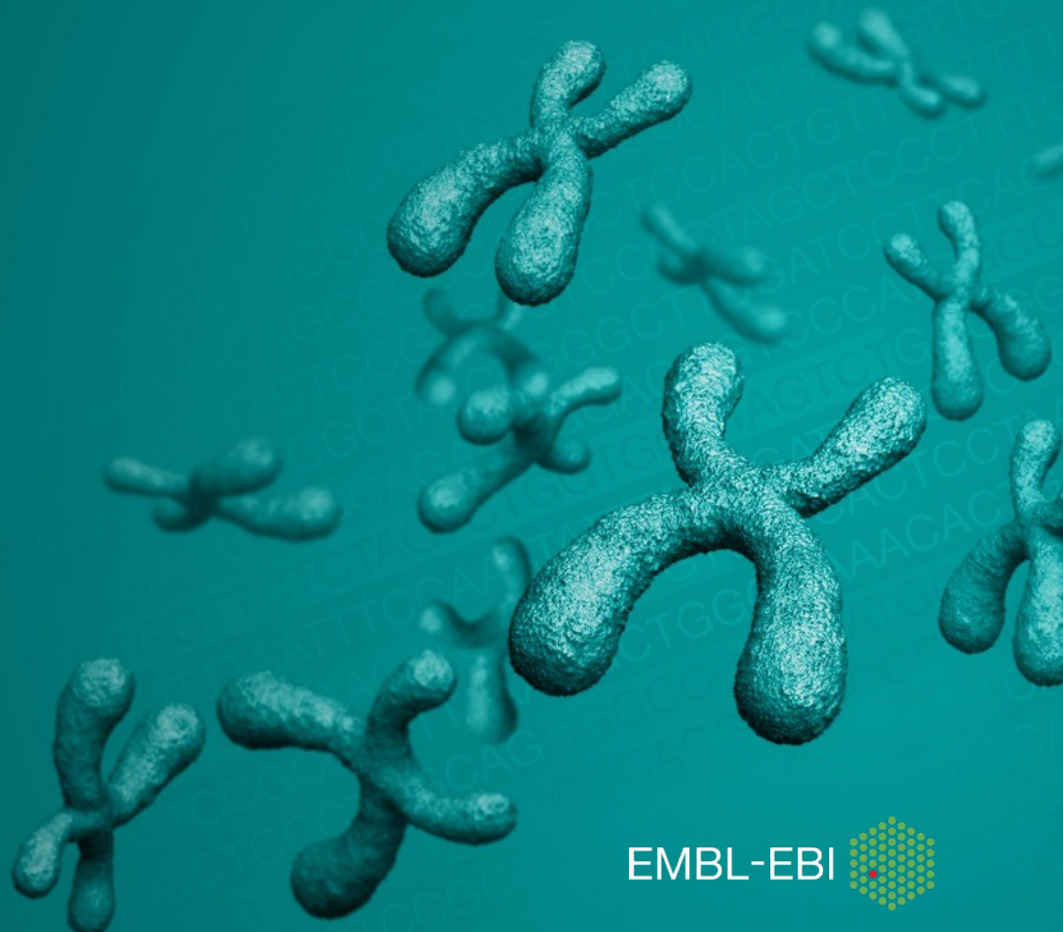
EMBL-EBI Data Transfer Use cases

Andrea Cristofori

andrea.cristofori@ebi.ac.uk

EGI Conference 2020



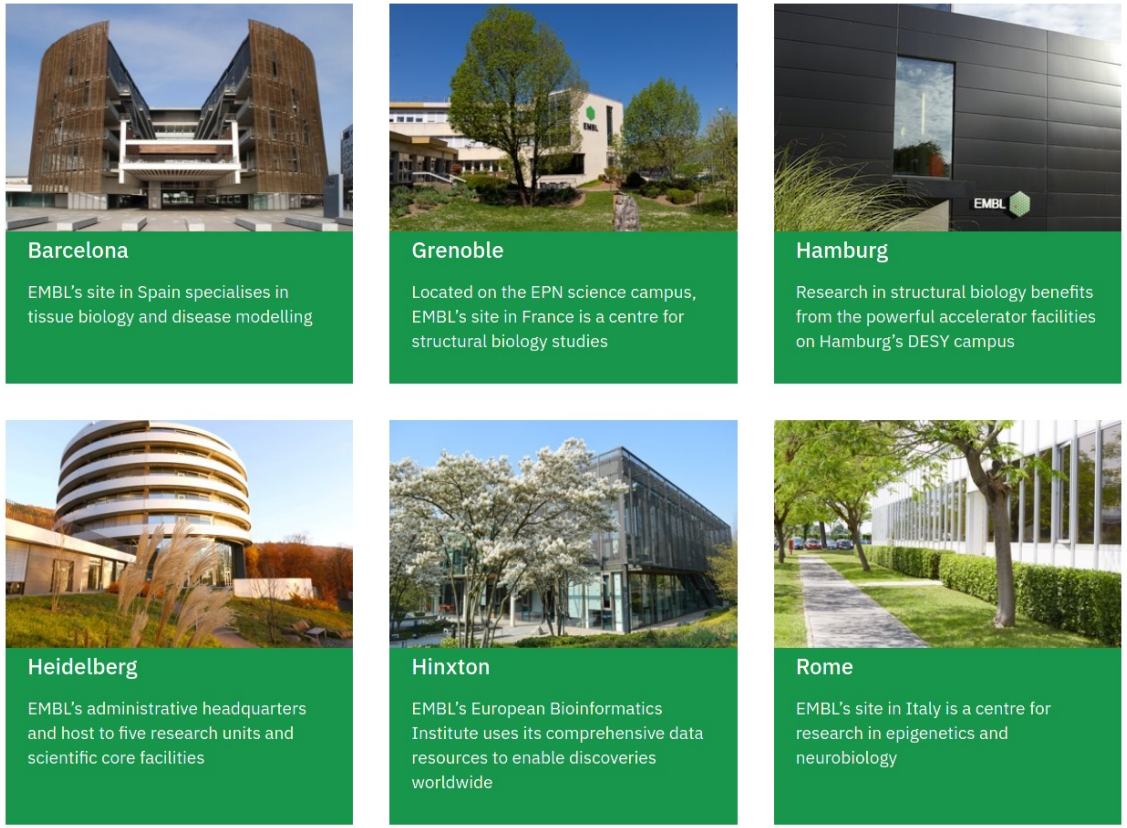
EMBL-EBI

Summary

- What is EMBL and EMBL-EBI
- What’s the amount of data we are dealing with
- How we are currently supporting the data transfer at EMBL-EBI

What is EMBL and EMBL-EBI

- EMBL: 6 sites in Europe, 27 member states
  - Heidelberg, Germany (Main Laboratory)
  - Hamburg, Germany (Structural Biology)
  - Grenoble, France (Structural Biology)
  - Barcelona, Spain (Tissue Biology, Disease Modeling)
  - Monterotondo, Italy (Epigenetics and Neurobiology)
  - Hinxton, Cambridge, UK (Bioinformatics)



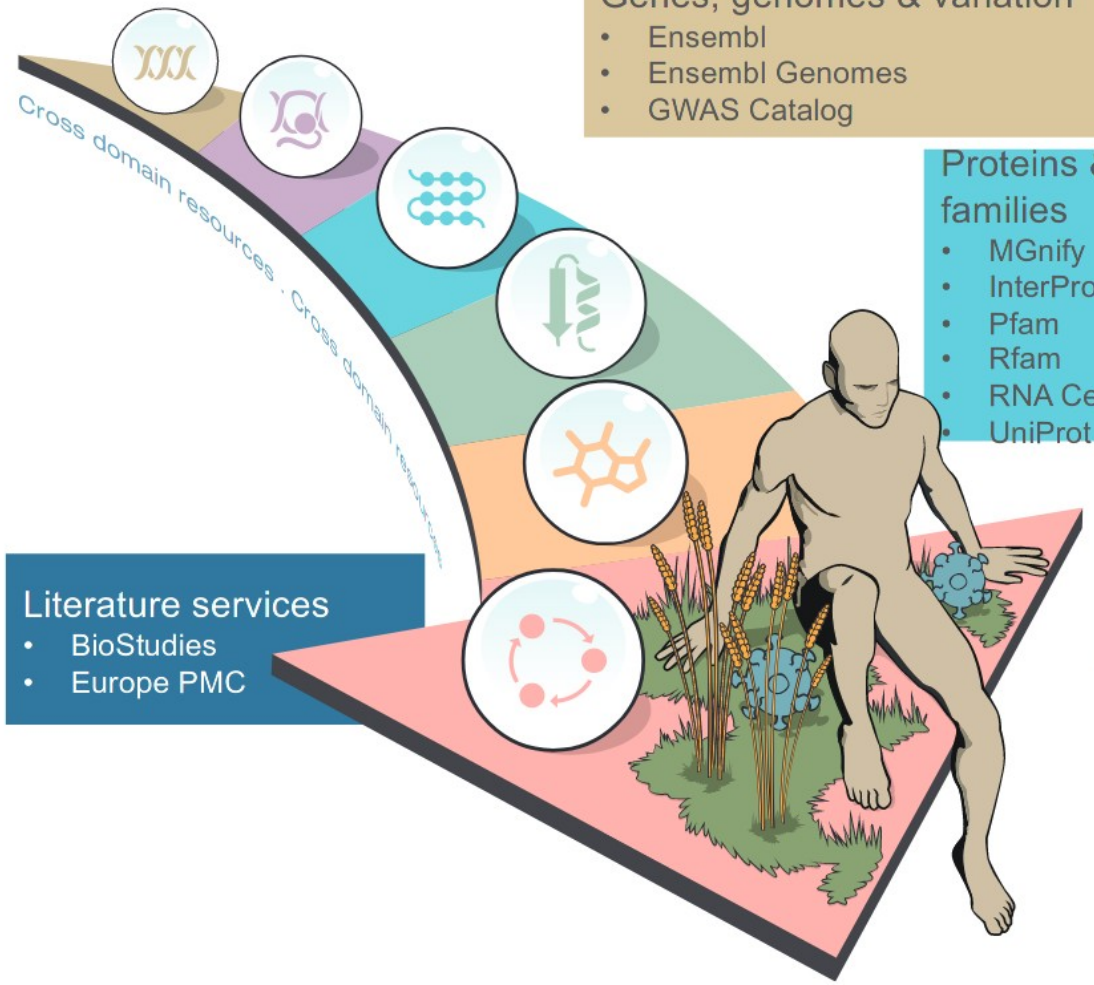
What is EMBL and EMBL-EBI

Stated mission:

- perform basic research in molecular biology
- train scientists, students and visitors at all levels
- offer vital services to scientists in the member states
- develop new instruments and methods
- actively engage in technology transfer
- and to integrate European life science research

In particular EMBL-EBI makes the world's public biological data freely available to the scientific community via a range of services and tools, perform basic research and provide professional training in bioinformatics.

Data resources at EMBL-EBI



Genes, genomes & variation
- Ensembl
- Ensembl Genomes
- GWAS Catalog

Molecular atlas
- Array Express
- Expression Atlas
- PRIDE

Proteins & protein families
- MGnify
- InterPro
- Pfam
- Rfam
- RNA Central
- UniProt

Macromolecular & cellular structure
- Protein Data Bank in Europe (PDBe)
- PDBe-KB
- Electron Microscopy Data Bank
- EMPIAR

Chemistry services
- ChEBI
- ChEMBL
- MetaboLights
- SureChEMBL

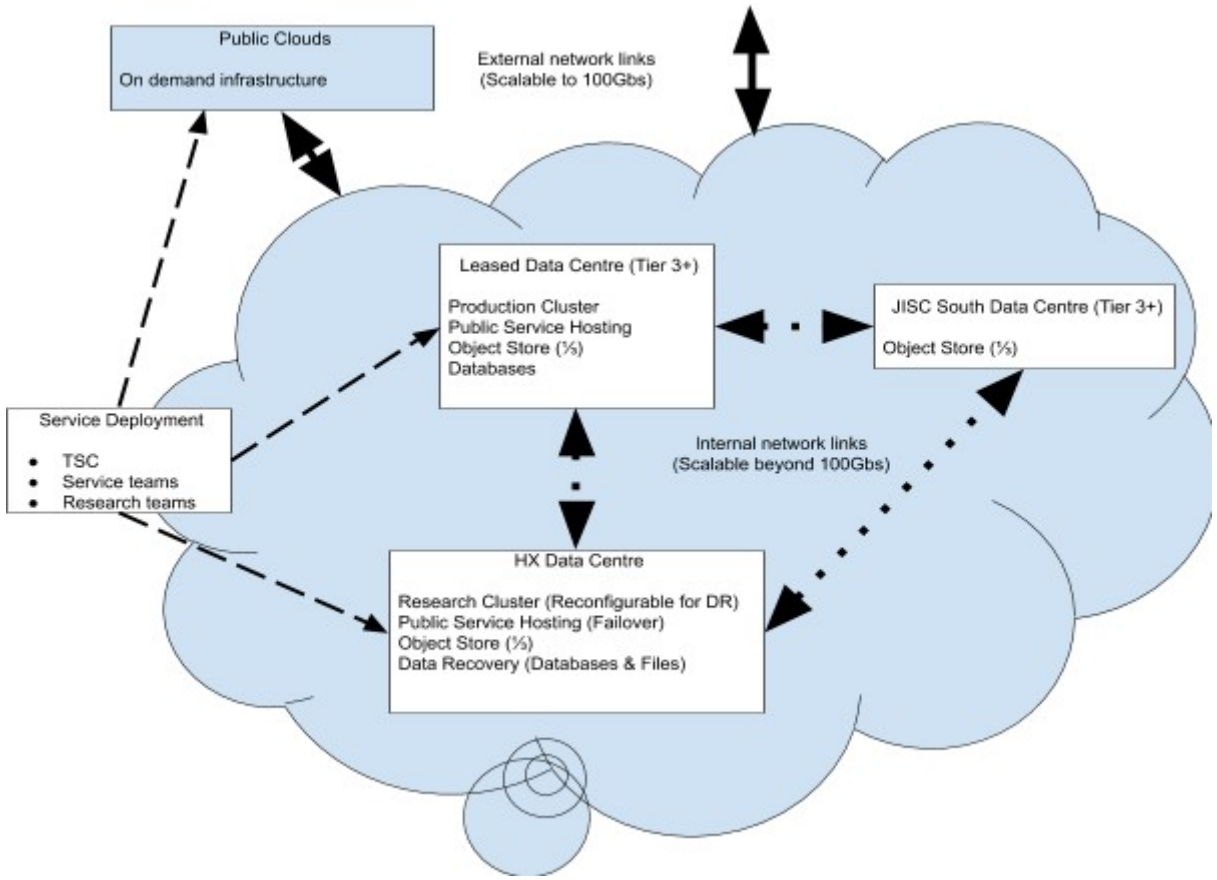
Molecular systems
- BioModels
- IntAct
- OmicsDI
- Reactome

Literature services
- BioStudies
- Europe PMC

Molecular archives
- European Nucleotide Archive
- European Variation Archive
- European Genome-phenome Archive
- Experimental Factor Ontology
- BioSamples
- Mouse Resources

How we are currently supporting the data transfer at EMBL-EBI

- Raw Storage (241PB):
  - Object Store: 103PB
  - NAS: 81PB
  - HPC Storage: 27PB
- Tape: for disaster recovery of the Object Store
- Out (3.5-4PB/month):
  - Globus Connect Server
  - Aspera
  - HTTP/FTP/rsync



EMBL-EBI

How we are currently supporting the data transfer at EMBL-EBI

- 2 geographically separated data centres used for data transfer, data store and processing + 1 as data store only
  - 100Gbs Internal networking capacity
  - Max 100Gbs External networking capacity (can be increased when to that value if requested)
- Service Resilience:
  - NAS storage: replicated on 2 data centres
  - FIRE: in house developed S3 compatible object store, geodispersed on the 3 data centre

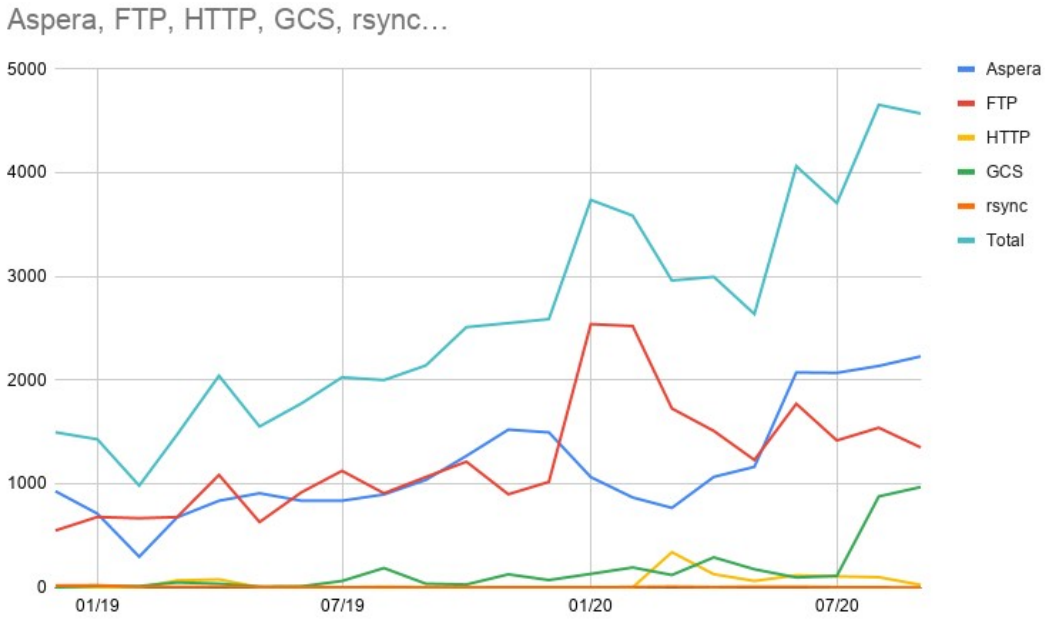
How we are currently supporting the data transfer at EMBL-EBI

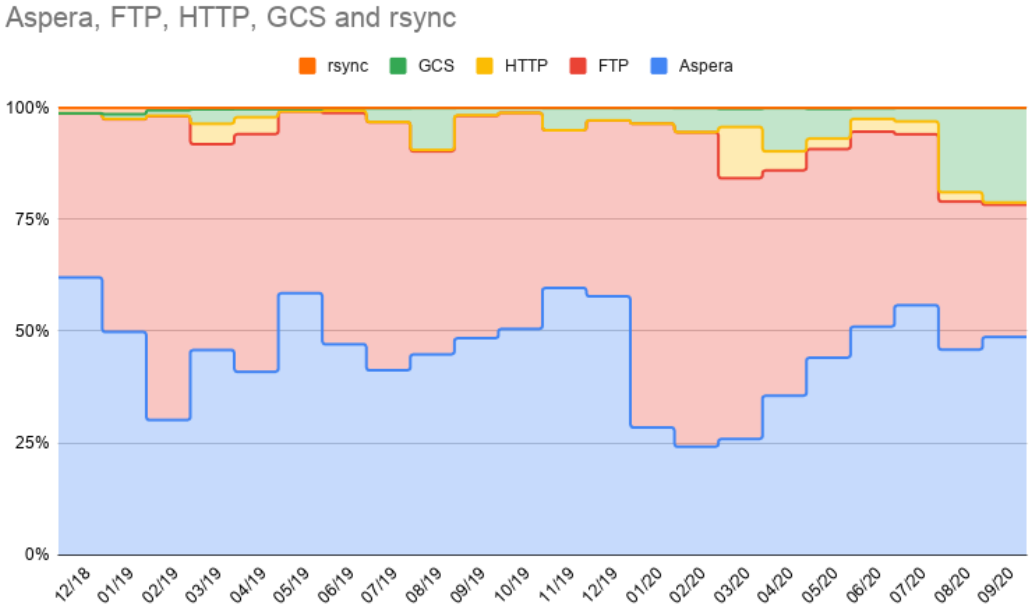
- In the last 2 years (VMs + containers when possible):
  - 90% of the FTP/HTTP/rsync infrastructure has been migrated from physical machines to VM + containers giving more flexibility
  - 100% of the Globus Connect Server has been migrated to VMs and new network configuration is underway
  - We are evaluating the migration of Aspera to Vms
- All VMs are being configured with 2 network interfaces:
  - One for external traffic
  - One for access to internal resources

How we are currently supporting the data transfer at EMBL-EBI

- How public data is accessed:
  - Public data are replicated on the NAS storage and made available through the same endpoint
  - Data on the object store is available on both data centre (access to this data generate some extra traffic between data centre as the files are geodispersed)
  - For FTP/HTTP/rsync and in part Aspera this is done through Amazon Route 53 for balancing and failover
  - For Globus Connect Server, Globus service itself balance the load across the available DTN on the two datacentre (for now is not possible to give a different “weight” to different nodes)
- How private data is accessed:
  - Private data is generally the result of submission from collaborator and is not present in one of the replicated storage
  - Dedicated storage area are available in each data centre and different endpoint are used on each one

What’s the amount of data we are dealing with





EMBL-EBI

How we are currently supporting the data transfer at EMBL-EBI

- Consideration:
  - Recent experience with Globus Connect Server shows very good performance among other advantages (e.g.: if more resources are needed we can easily provide additional VMs)
  - We need to keep compatibility and support all other services for the forseeable future (rsync)
  - Aspera is historically popular however the licensing scheme might require more investment in future as we see the
- For the future the link from EBI to the external world should allow increase on the traffic for the next few years when, with the current trends, an upgrade might be necessary (currently ~20Gb/s are continuosly used with period of time at ~40Gb/s)

Questions?