

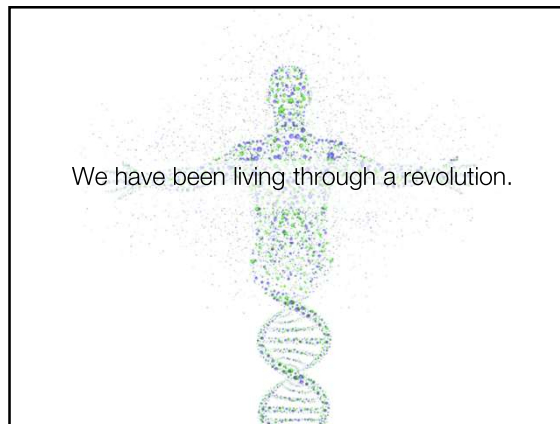
Genomics Research at EBI: Challenges in Statistical Scaling

Nick Goldman
Head of Research, EMBL-EBI
www.ebi.ac.uk



EMBL-EBI

We have been living through a revolution.



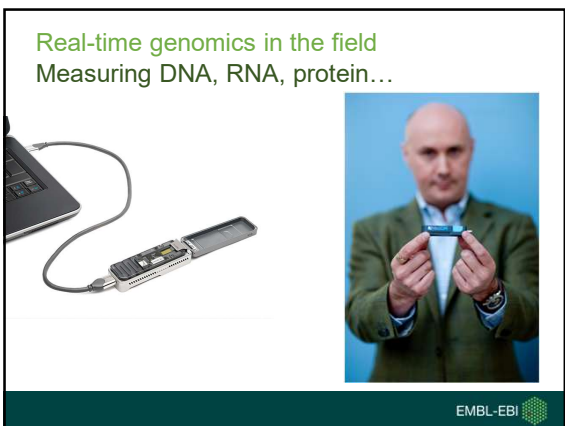
All living things are made from the same stuff:
DNA, RNA, protein...



...and agriculture, and the environment

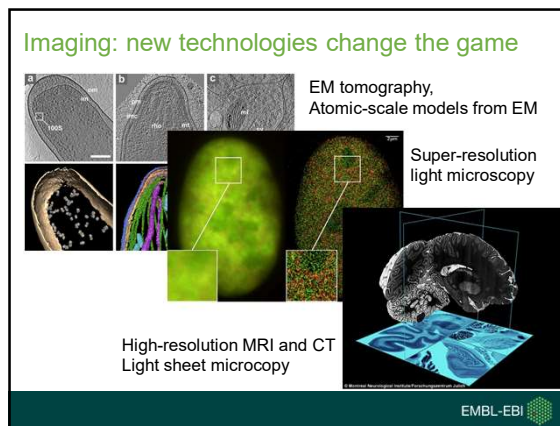


Real-time genomics in the field
Measuring DNA, RNA, protein...



EMBL-EBI

Imaging: new technologies change the game



EM tomography,
Atomic-scale models from EM

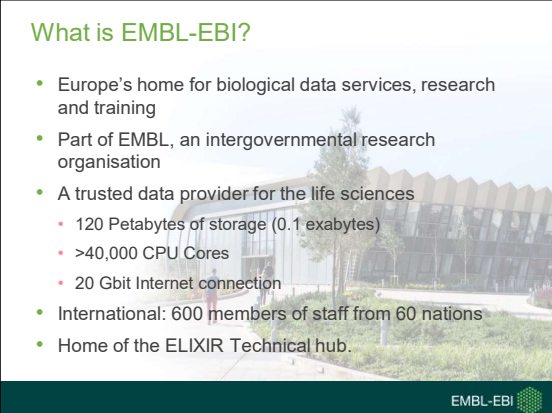
Super-resolution
light microscopy

High-resolution MRI and CT
Light sheet microscopy

EMBL-EBI

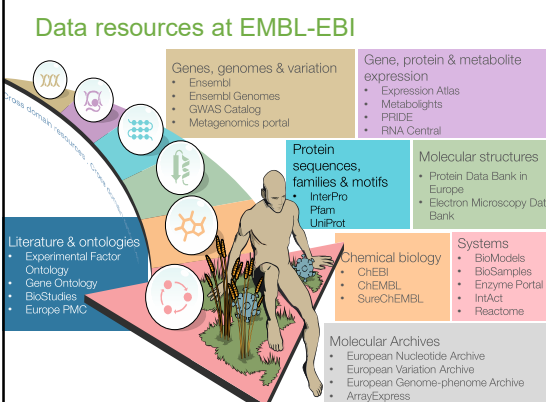
What is EMBL-EBI?

- Europe's home for biological data services, research and training
- Part of EMBL, an intergovernmental research organisation
- A trusted data provider for the life sciences
 - 120 Petabytes of storage (0.1 exabytes)
 - >40,000 CPU Cores
 - 20 Gbit Internet connection
- International: 600 members of staff from 60 nations
- Home of the ELIXIR Technical hub.



EMBL-EBI

Data resources at EMBL-EBI



- Genes, genomes & variation**
 - Ensembl
 - Ensembl Genomes
 - GWAS Catalog
 - Metagenomics portal
- Gene, protein & metabolite expression**
 - Expression Atlas
 - Metaboblights
 - PRIDE
 - RNA Central
- Protein sequences, families & motifs**
 - InterPro
 - Plan
 - UniProt
- Molecular structures**
 - Protein Data Bank in Europe
 - Electron Microscopy Data Bank
- Literature & ontologies**
 - Experimental Factor Ontology
 - Gene Ontology
 - BioStudies
 - Europe PMC
- Chemical biology**
 - ChEMBL
 - ChEMBL
 - SureChEMBL
- Systems**
 - BioModels
 - BioSamples
 - Enzyme Portal
 - InrAct
 - Reactome
- Molecular Archives**
 - European Nucleotide Archive
 - European Variation Archive
 - European Genome-phenome Archive
 - ArrayExpress

What services do we provide?



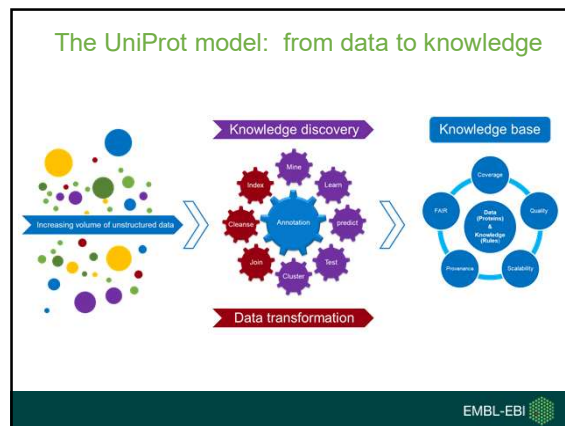
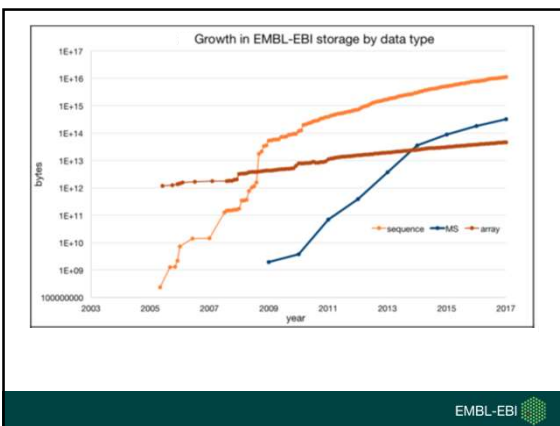
- Labs around the world send us their data and we...
- Archive it
- Classify it
- Share it with other data providers
- Analyse, add value and integrate it
- ...provide tools to help researchers use it
- A collaborative enterprise

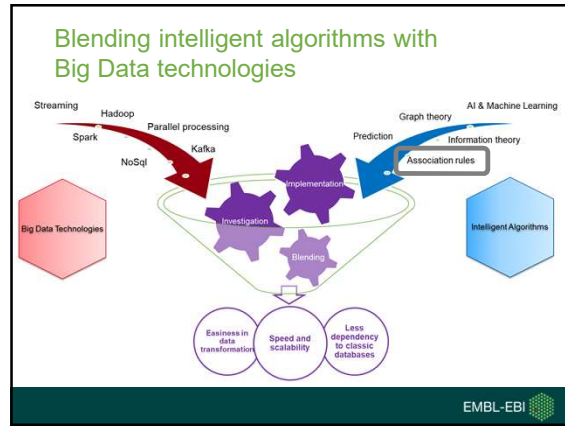
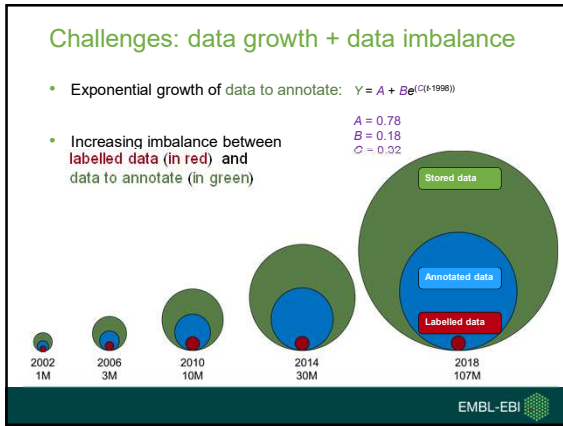
EMBL-EBI

Principles of FAIR data

- Findable**
 - identifiers, metadata, indexed
- Accessible**
 - retrievable by identifier using open, free protocol
- Interoperable**
 - formal, accessible language using FAIR vocabularies
- Re-usable**
 - clear, accessible license, provenance, community standards

EMBL-EBI





Association-Rule-Based Annotator (ARBA)

- A **multi-class** predictor to annotate protein entries
- Theoretical paper: *International Journal on Artificial Intelligence Tools* (2014)
- Generates **representative** rules
- Extracts association rules (exhaustive knowledge)
- Filter them and build concise models (representative and concise knowledge)
- Provides techniques and structures to evaluate the prediction
- Cross-validation, confusion matrix, ROC curve, experiment, visualisation

Example 5.1. Given a set of rules $R = \{R_1, R_2, R_3, R_4, R_5, R_6, R_7, R_8, R_9, R_{10}\}$, the following solution is found. Let R' be the set of rules such that $R' \subseteq R$.

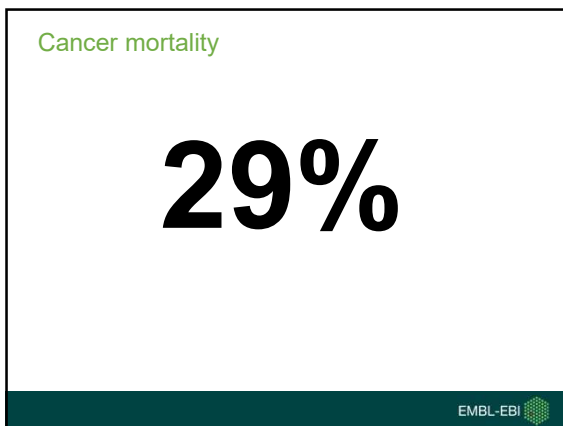
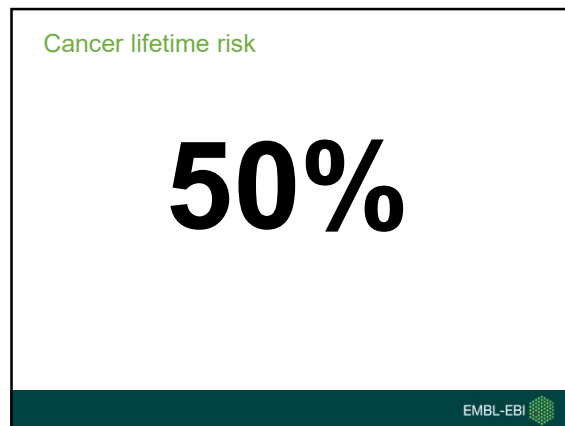
(1) First we show that there is no R' such that $R' \subseteq R$ and $R' \neq R$. Suppose $R' \neq R$. Then $R' \subseteq R$ and $R' \neq R$ implies that there is at least one rule $R_i \in R$ such that $R_i \notin R'$. This is a contradiction because R' is supposed to be a solution.

(2) Next, we show that $R' = R$ is the only solution. Suppose $R' \subseteq R$ and $R' \neq R$. Then $R' \subseteq R$ and $R' \neq R$ implies that there is at least one rule $R_i \in R$ such that $R_i \notin R'$. This is a contradiction because R' is supposed to be a solution.

(3) Finally, we show that $R' = R$ is the only solution. Suppose $R' \subseteq R$ and $R' \neq R$. Then $R' \subseteq R$ and $R' \neq R$ implies that there is at least one rule $R_i \in R$ such that $R_i \notin R'$. This is a contradiction because R' is supposed to be a solution.

Category	Value
ARBA	0.95
ARBA + ARBA	0.95
ARBA + ARBA + ARBA	0.95
ARBA + ARBA + ARBA + ARBA	0.95
ARBA + ARBA + ARBA + ARBA + ARBA	0.95

EMBL-EBI

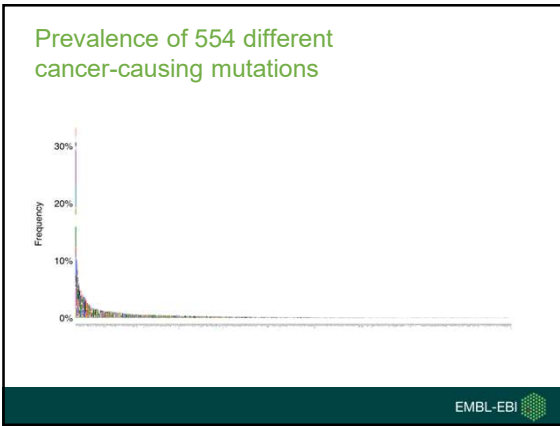


Cancer is a genetic disease

Structural abnormalities

Point mutations

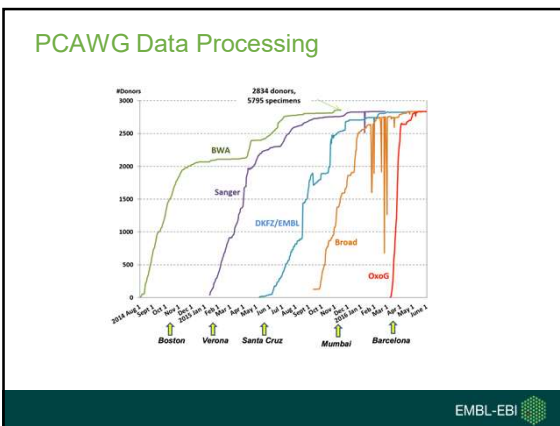
EMBL-EBI



Pan-Cancer Analysis of Whole Genomes

- 2,778 tumour types
- 39 cancer types
- 300 scientists
- 650 TB of data
- 3 mutation-calling pipelines
 - 46,197,106 point mutations
 - 288,416 structural variants
- 32 papers on bioRxiv.org

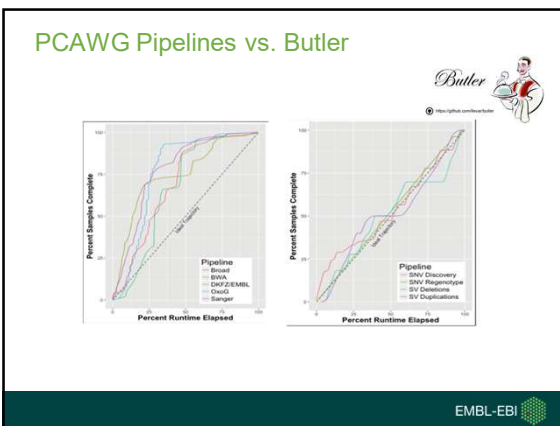
EMBL-EBI



Lessons Learned

- Infrastructure fails
 - Disk failures
 - Network events
- Bioinformatics tools fail
 - Bad data
 - Software defects
- End user must be self-reliant in diagnosis and resolution of issues
- Manual investigation and resolution of failure is a major contributor to project costs and timeline slip

EMBL-EBI



Big data (multi-omics) association genetics: statistical challenges and opportunities

- Challenge: large-scale multiple testing problem
 - need to consider potentially millions of loci
 - account for confounding
 - need appropriate corrections (e.g. false discovery rate)
 - scalability to large cohorts
- Opportunity: large datasets allow testing of modelling assumptions/fit better models
 - inference of confounding structures
 - not possible before large-scale hypothesis testing/large datasets
 - more power from large datasets
 - modelling rich and high-dimensional molecular phenotypes

EMBL-EBI

Scalable linear mixed models to account for known and unknown confounding factors

population 1
population 2

phenotype

$y = X\beta + u + \psi$

phenotype SNP relatedness noise

relatedness

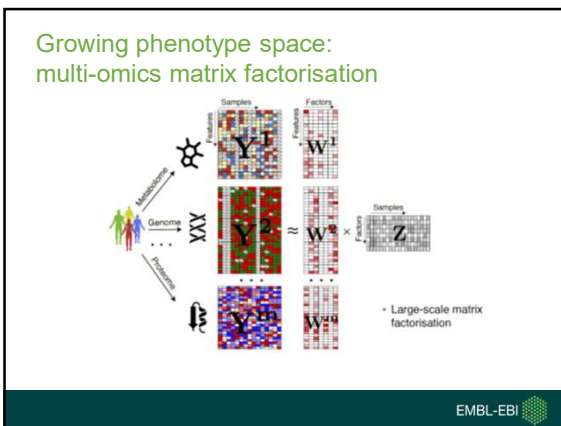
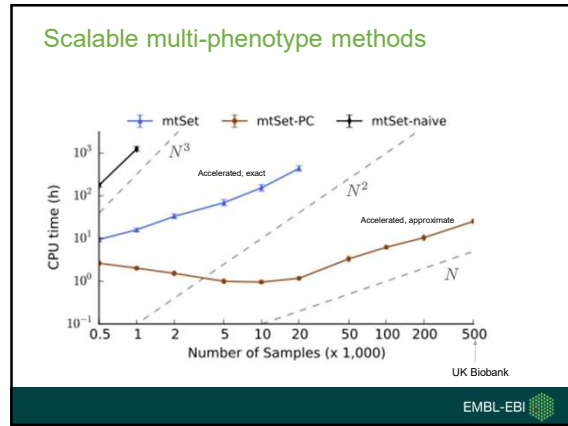
Individuals

Typical confounding

- Population structure
- Environment/batch

$u \sim \mathcal{N}(0, \Sigma)$

EMBL-EBI



Acknowledgements

Cancer genomics

- Moritz Gerstung
- Sergei Iakhnin (EMBL-HD)

Metagenomics

- Rob Finn

Association studies

- Oliver Stegle
- Pablo Casale

UniProt database

- Maria Martin
- Rabie Saidi

EMBL-EBI

Thank you!

Follow on twitter: @emlebi

EMBL-EBI