Introduction to Module & Bioinformatics

Tim Hubbard @timjph
King's College London, King's Health Partners
Genomics England

Bioinformatics, Interpretation and Data Quality in Genome Analysis
MSc in Genomics Medicine
15th February 2016

5 days of Bioinformatics

- Day 1 - Introduction to variant analysis using NGS data and quality control
- Day 2 - Introduction to variant calling and Annotation
- Day 3 - Variant Annotation and Interpretation
- Day 4 - Researching links between genotype to clinical phenotype
- Day 5 – Additional annotation and genomic analyses

Wider overview of Bioinformatics

- …the application of computer technology to the management of biological information. Computers are used to gather, store, analyze and integrate biological information

- **Pipelines** to process experimental data
- **Repositories** to store archive
- **Resources** to organise, present, interogate
- **Analysis** to make discoveries
- **Algorithms** to make predictions

Biological Information

- **DNA** (copy of genetic material in every cell)
  - makes
- **RNA** (transcripts of expressed genes)
  - makes
- **Protein** (translation of coding regions of genes)
  - Linear sequences folds into a 3D structure
- Within **Cells**
  - Proteins interact with each other, metabolites, DNA, RNA
- Within **Organisms**
  - Cells divide, develop, interact

- **Genetics differences**
  - Change an Organism’s behavior

Biological Information – 1980s

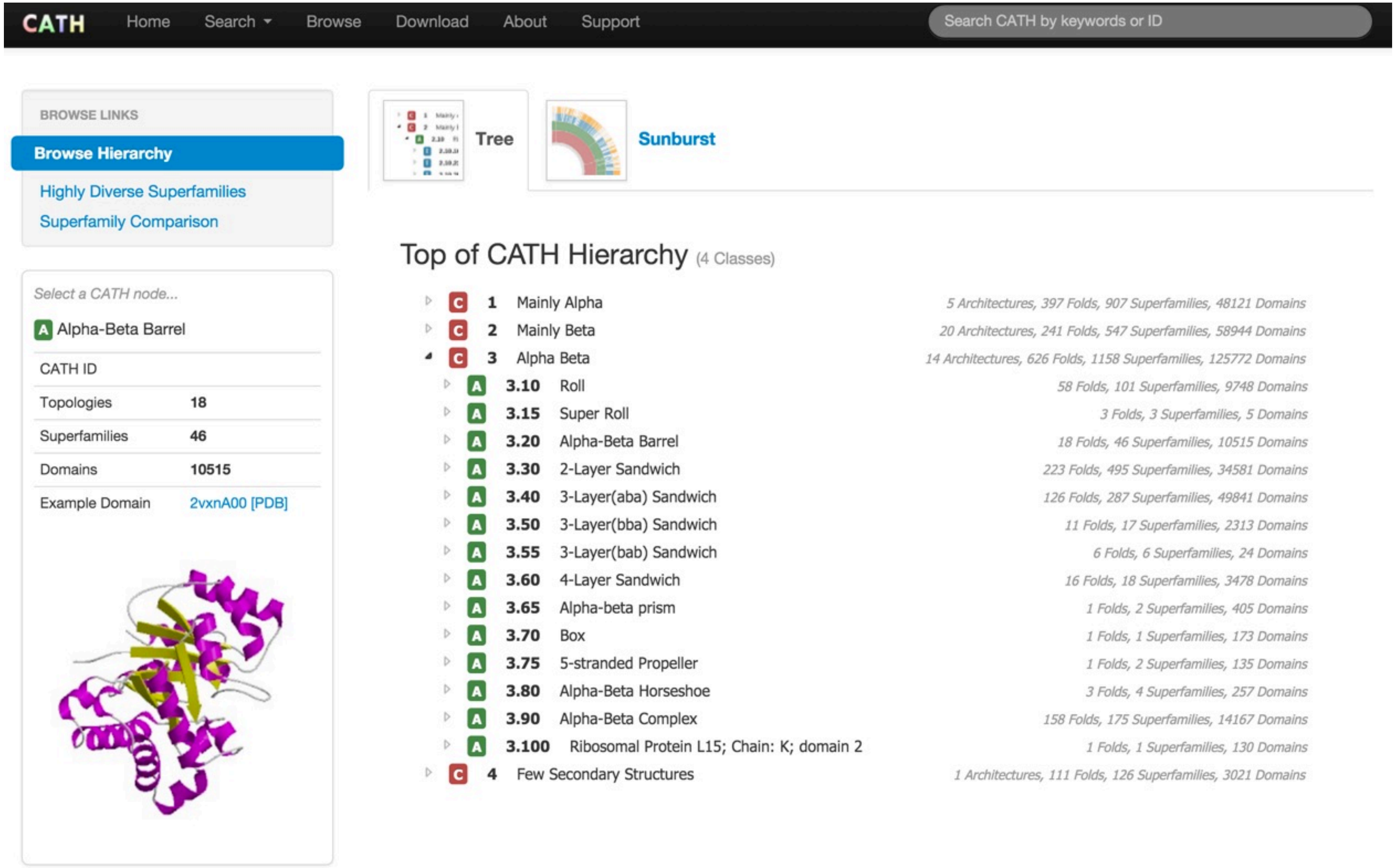
- **DNA** (copy of genetic material in every cell)
  - makes
- **RNA** (transcripts of expressed genes)
  - makes
- **Protein** (translation of coding regions of genes)
  - Linear sequences folds into a 3D structure
- Within **Cells**
  - Proteins interact with each other, metabolites, DNA, RNA
- Within **Organisms**
  - Cells divide, develop, interact

- **Genetics differences**
  - Change an Organism's behavior

Bioinformatics & Structural Biology

- **Pipelines:**
  - Xray, NMR data processing packages
- **Repositories:**
  - PDB (Protein Data Bank)
- **Resources:**
  - e.g. SCOP (Structural Classification of Proteins), CATH (Class, Architecture, Topology, Homology) databases
- **Analysis:**
  - e.g. molecular mechanisms
- **Algorithms:**
  - e.g. for Protein Structure Prediction

CATH database



Select a CATH node...

A Alpha-Beta Barrel

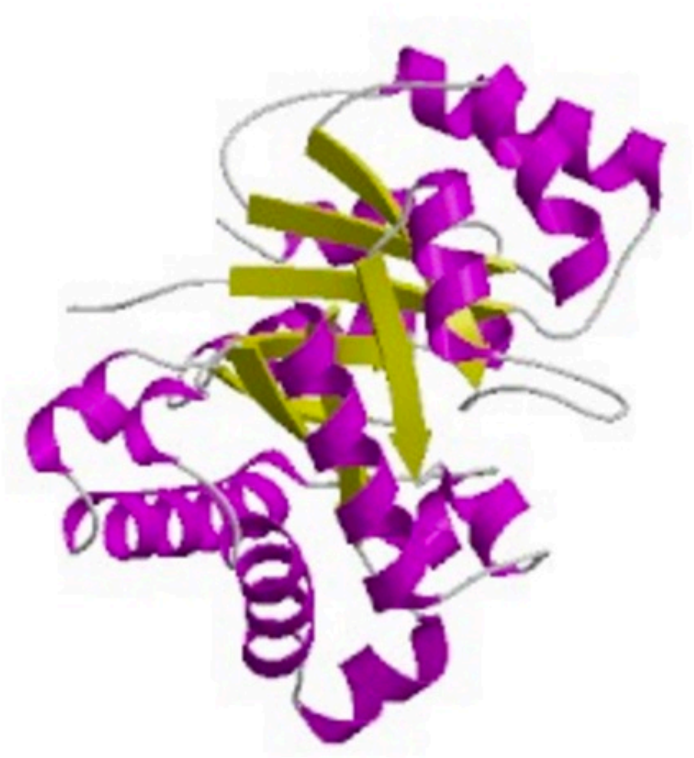
CATH ID	
Topologies	18
Superfamilies	46
Domains	10515
Example Domain	2vxnA00 [PDB]

Top of CATH Hierarchy (4 Classes)

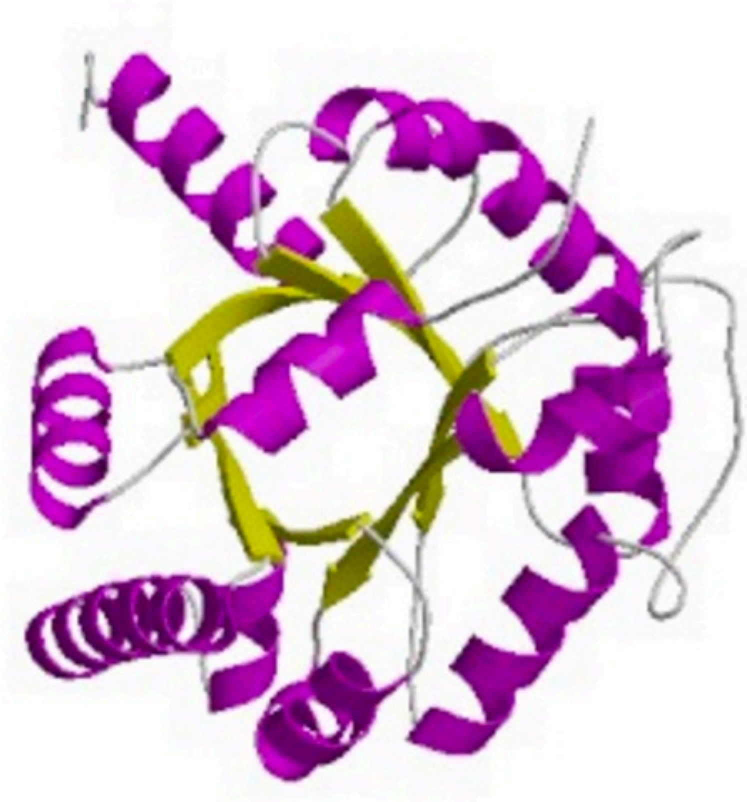
C	1	Mainly Alpha	5 Architectures, 397 Folds, 907 Superfamilies, 48121 Domains
C	2	Mainly Beta	20 Architectures, 241 Folds, 547 Superfamilies, 58944 Domains
C	3	Alpha Beta	14 Architectures, 626 Folds, 1158 Superfamilies, 125772 Domains
A	3.10	Roll	58 Folds, 101 Superfamilies, 9748 Domains
A	3.15	Super Roll	3 Folds, 3 Superfamilies, 5 Domains
A	3.20	Alpha-Beta Barrel	18 Folds, 46 Superfamilies, 10515 Domains
A	3.30	2-Layer Sandwich	223 Folds, 495 Superfamilies, 34581 Domains
A	3.40	3-Layer(aba) Sandwich	126 Folds, 287 Superfamilies, 49841 Domains
A	3.50	3-Layer(bba) Sandwich	11 Folds, 17 Superfamilies, 2313 Domains
A	3.55	3-Layer(bab) Sandwich	6 Folds, 6 Superfamilies, 24 Domains
A	3.60	4-Layer Sandwich	16 Folds, 18 Superfamilies, 3478 Domains
A	3.65	Alpha-beta prism	1 Folds, 2 Superfamilies, 405 Domains
A	3.70	Box	1 Folds, 1 Superfamilies, 173 Domains
A	3.75	5-stranded Propeller	1 Folds, 2 Superfamilies, 135 Domains
A	3.80	Alpha-Beta Horseshoe	3 Folds, 4 Superfamilies, 257 Domains
A	3.90	Alpha-Beta Complex	158 Folds, 175 Superfamilies, 14167 Domains
A	3.100	Ribosomal Protein L15; Chain: K; domain 2	1 Folds, 1 Superfamilies, 130 Domains
C	4	Few Secondary Structures	1 Architectures, 111 Folds, 126 Superfamilies, 3021 Domains

Protein 3D structures: TIM barrels

Example Domain 2vxnA00 [PDB]



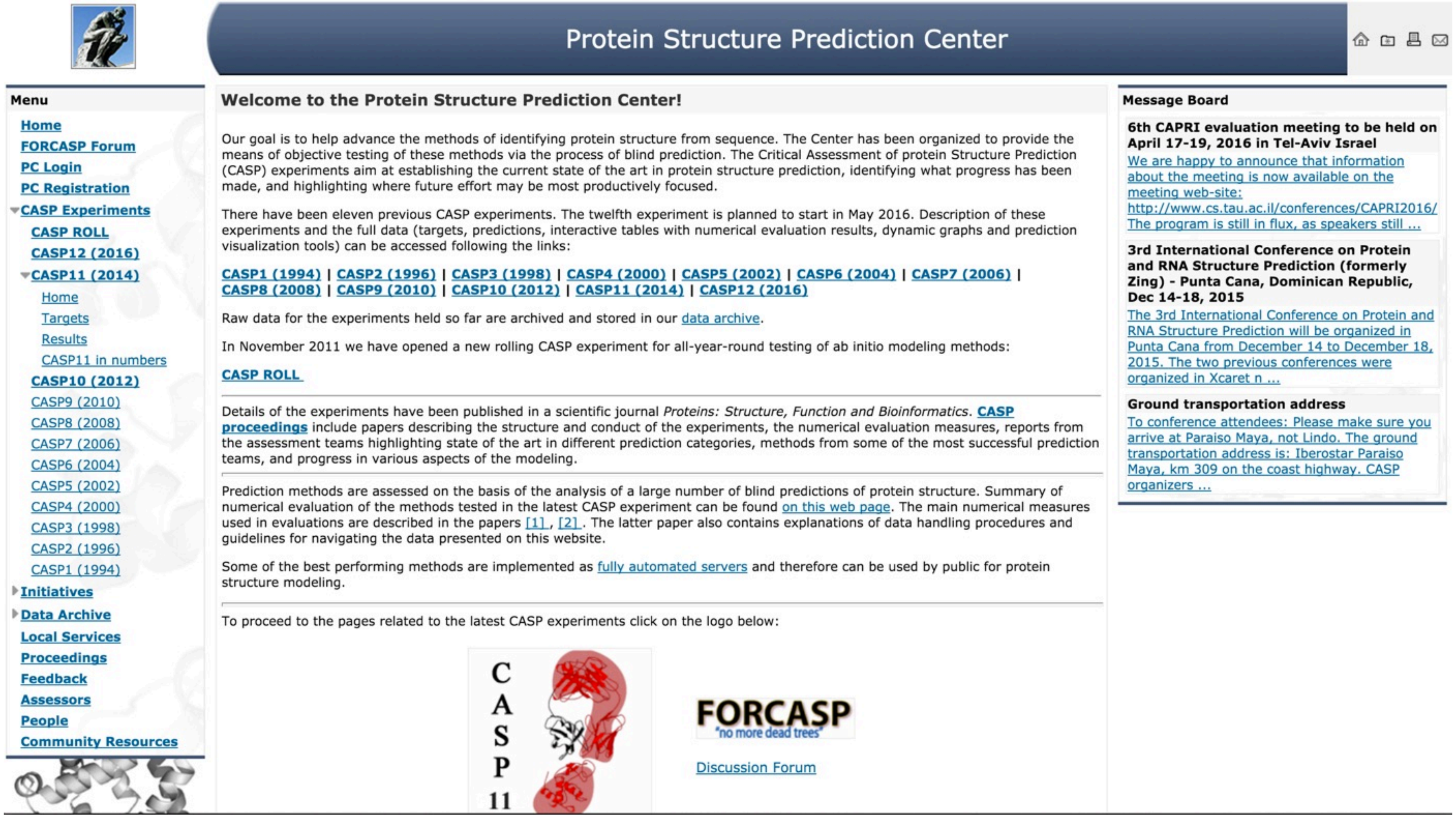
Example Domain 2vwsA00 [PDB]



Algorithms: Protein structure Prediction

- Why is there a need for a prediction algorithm?
  - Vastly more protein sequences than structures
  - Xray/NMR slow, difficult, expensive
  - Ideally algorithm would predict consequence of mutation
- Strategies
  - Show sequence of unknown structure related to sequence of known structure and ‘model’ from it
  - Predict directly using knowledge of atomic interactions (physics) and simulation (*ab initio*)
  - Predict many models from ‘lego’ components from database of known structures
- History
  - 1970s–1994 Delusion
  - 1994 First CASP Competition; Steady progress since, but still not ‘solved’

Algorithms: Protein structure Prediction http://predictioncenter.org/



Welcome to the Protein Structure Prediction Center!

Our goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. The Critical Assessment of protein Structure Prediction (CASP) experiments aim at establishing the current state of the art in protein structure prediction, identifying what progress has been made, and highlighting where future effort may be most productively focused.

There have been eleven previous CASP experiments. The twelfth experiment is planned to start in May 2016. Description of these experiments and the full data (targets, predictions, interactive tables with numerical evaluation results, dynamic graphs and prediction visualization tools) can be accessed following the links:

CASP1 (1994) | CASP2 (1996) | CASP3 (1998) | CASP4 (2000) | CASP5 (2002) | CASP6 (2004) | CASP7 (2006) | CASP8 (2008) | CASP9 (2010) | CASP10 (2012) | CASP11 (2014) | CASP12 (2016)

Raw data for the experiments held so far are archived and stored in our data archive.

In November 2011 we have opened a new rolling CASP experiment for all-year-round testing of ab initio modeling methods:

CASP ROLL

Details of the experiments have been published in a scientific journal *Proteins: Structure, Function and Bioinformatics*. CASP proceedings include papers describing the structure and conduct of the experiments, the numerical evaluation measures, reports from the assessment teams highlighting state of the art in different prediction categories, methods from some of the most successful prediction teams, and progress in various aspects of the modeling.

Prediction methods are assessed on the basis of the analysis of a large number of blind predictions of protein structure. Summary of numerical evaluation of the methods tested in the latest CASP experiment can be found on this web page. The main numerical measures used in evaluations are described in the papers [1] , [2] . The latter paper also contains explanations of data handling procedures and guidelines for navigating the data presented on this website.

Some of the best performing methods are implemented as fully automated servers and therefore can be used by public for protein structure modeling.

To proceed to the pages related to the latest CASP experiments click on the logo below:

Message Board

6th CAPRI evaluation meeting to be held on April 17-19, 2016 in Tel-Aviv Israel

We are happy to announce that information about the meeting is now available on the meeting web-site: http://www.cs.tau.ac.il/conferences/CAPRI2016/ The program is still in flux, as speakers still ...

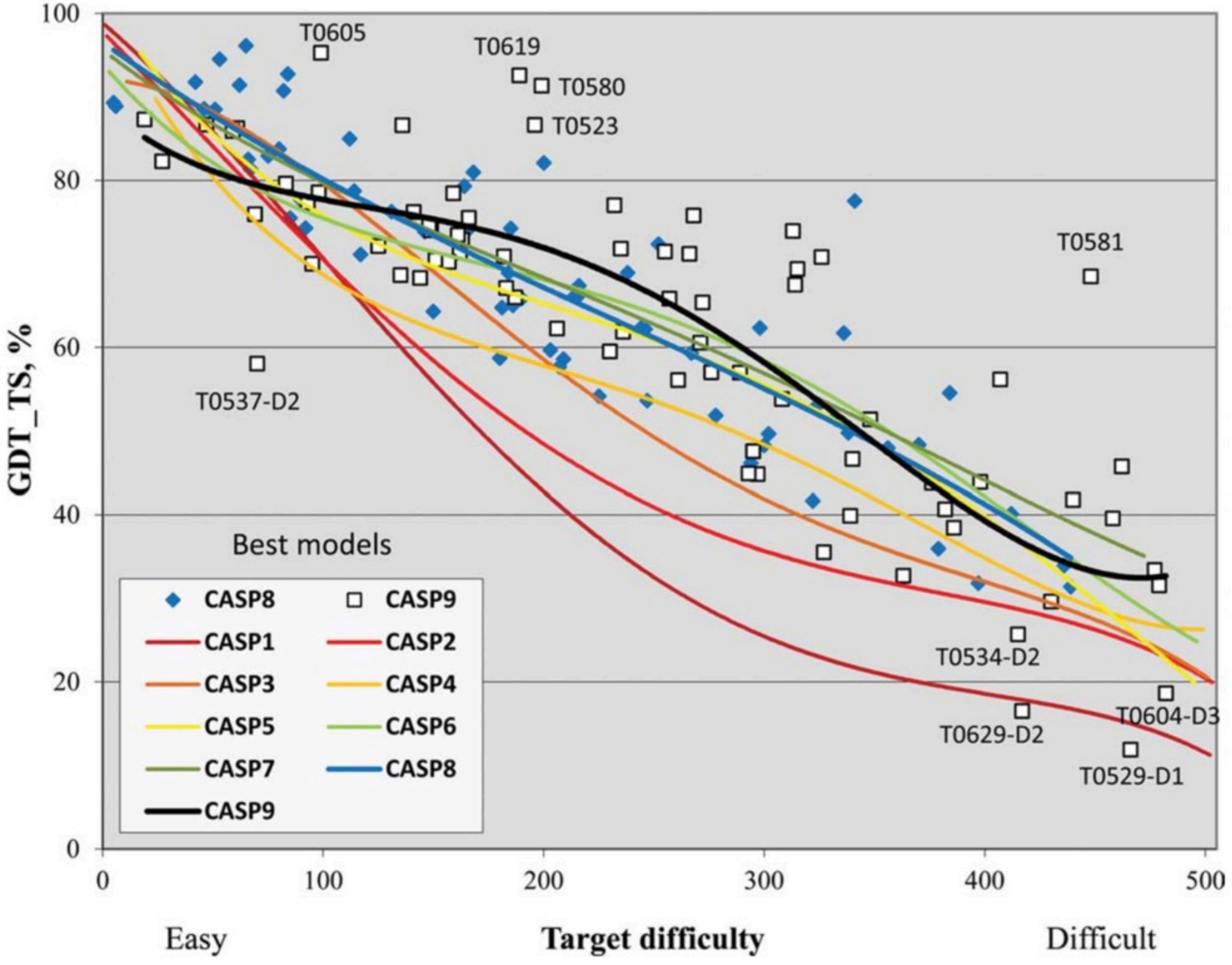
3rd International Conference on Protein and RNA Structure Prediction (formerly Zing) - Punta Cana, Dominican Republic, Dec 14-18, 2015

The 3rd International Conference on Protein and RNA Structure Prediction will be organized in Punta Cana from December 14 to December 18, 2015. The two previous conferences were organized in Xcaret n ...

Ground transportation address

To conference attendees: Please make sure you arrive at Paraiso Maya, not Lindo. The ground transportation address is: Iberostar Paraiso Maya, km 309 on the coast highway. CASP organizers ...

CASP progress: 1994-2010



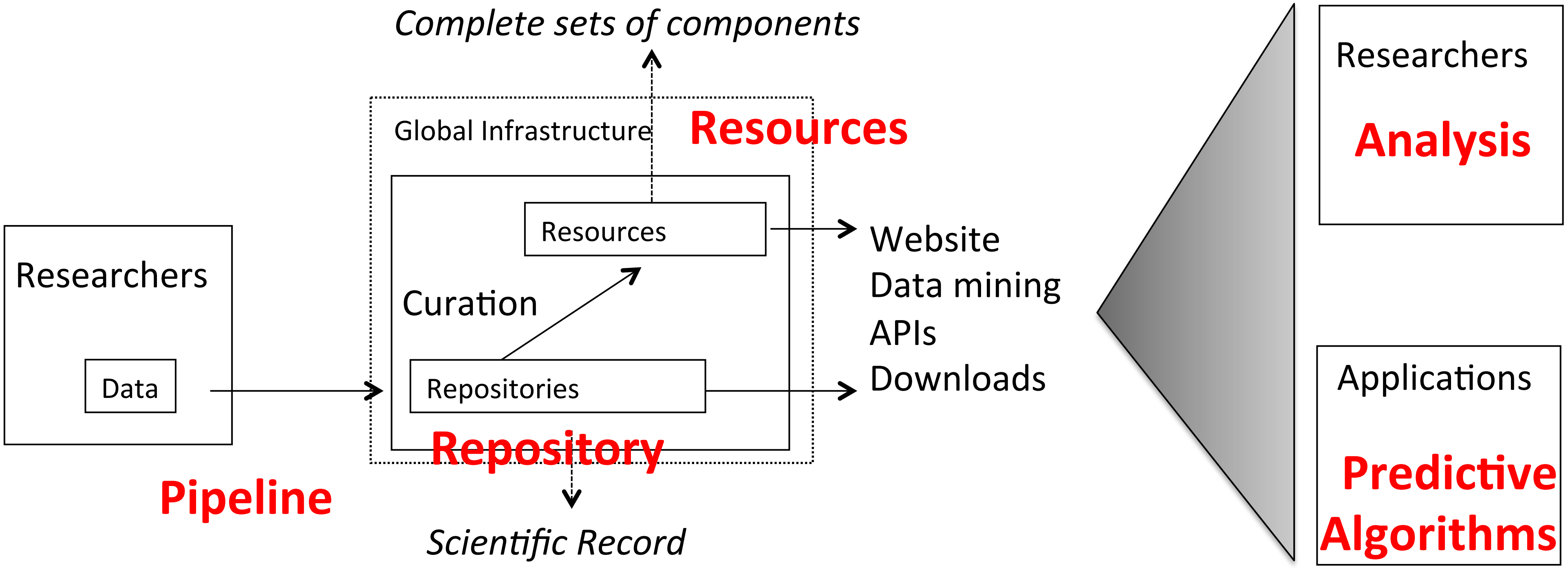
Bioinformatics & Genomics

- **Pipelines:**
  - assemble whole genome from fragments of DNA
- **Repositories:**
  - Genbank/ENA/DDBJ
- **Resources:**
  - e.g. Ensembl
- **Analysis**:
  - e.g. genes, evolution, conservation etc.
- **Algorithms:**
  - e.g. gene prediction

Algorithms: Gene prediction

- Why is there a need for a prediction algorithm?
  - Vastly more genome sequences than annotated genomes
  - Ideally algorithm would predict consequence of mutation
- Strategies
  - Collect transcriptome data; map back onto genome to annotate
  - Look for conserved regions between related genomes
  - Predict directly using knowledge of regulatory regions and simulation (*ab initio*)
- History
  - Ab initio algorithms performed well on single gene regions, but once large regions of genome sequenced, showed to perform poorly for vertebrates (~1998)
  - Transcriptomics data increasing easy to collect and algorithms relying on it because gold standard (Ensembl) + Curation for better accuracy (GENCODE)
  - Can still find new genes missing from transcriptome collection

Data, Databases & Bioinformatics



Complete sets of components

Global Infrastructure

Resources

Researchers

Analysis

Researchers

Data

Resources

Curation

Repositories

Website
Data mining
APIs
Downloads

Repository

Pipeline

Scientific Record

Applications

Predictive Algorithms

Submission

Reuse

Policies
Systems

Sustainable
funding

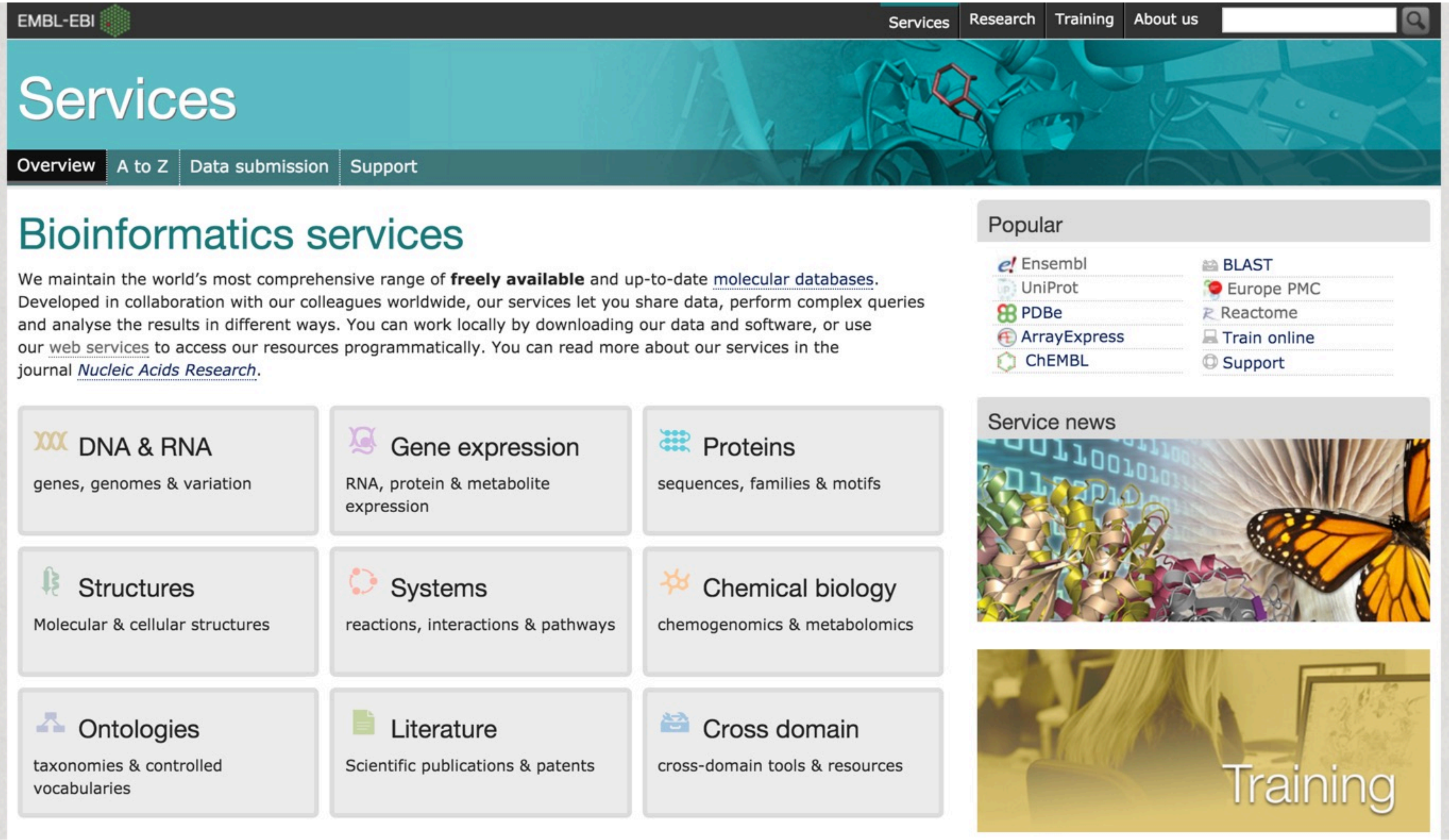
Discoverability
Easy of use (+access)

Organisation, Scale, History

- Centralised repositories for raw data
  - one data type, one repository
  - mandatory submission linked to publication
- Infrastructure to organise raw data for access
  - human genome presented to user as whole chromosomes instead of thousands of fragments
- Curated databases of biological objects
  - supported by evidence from raw data repositories

- First repositories >40 years old
- 1,000s of full time staff supporting infrastructure distributed worldwide

Repositories in EU: EBI http://www.ebi.ac.uk/services



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. You can read more about our services in the journal *Nucleic Acids Research*.

DNA & RNA
genes, genomes & variation

Gene expression
RNA, protein & metabolite expression

Proteins
sequences, families & motifs

Structures
Molecular & cellular structures

Systems
reactions, interactions & pathways

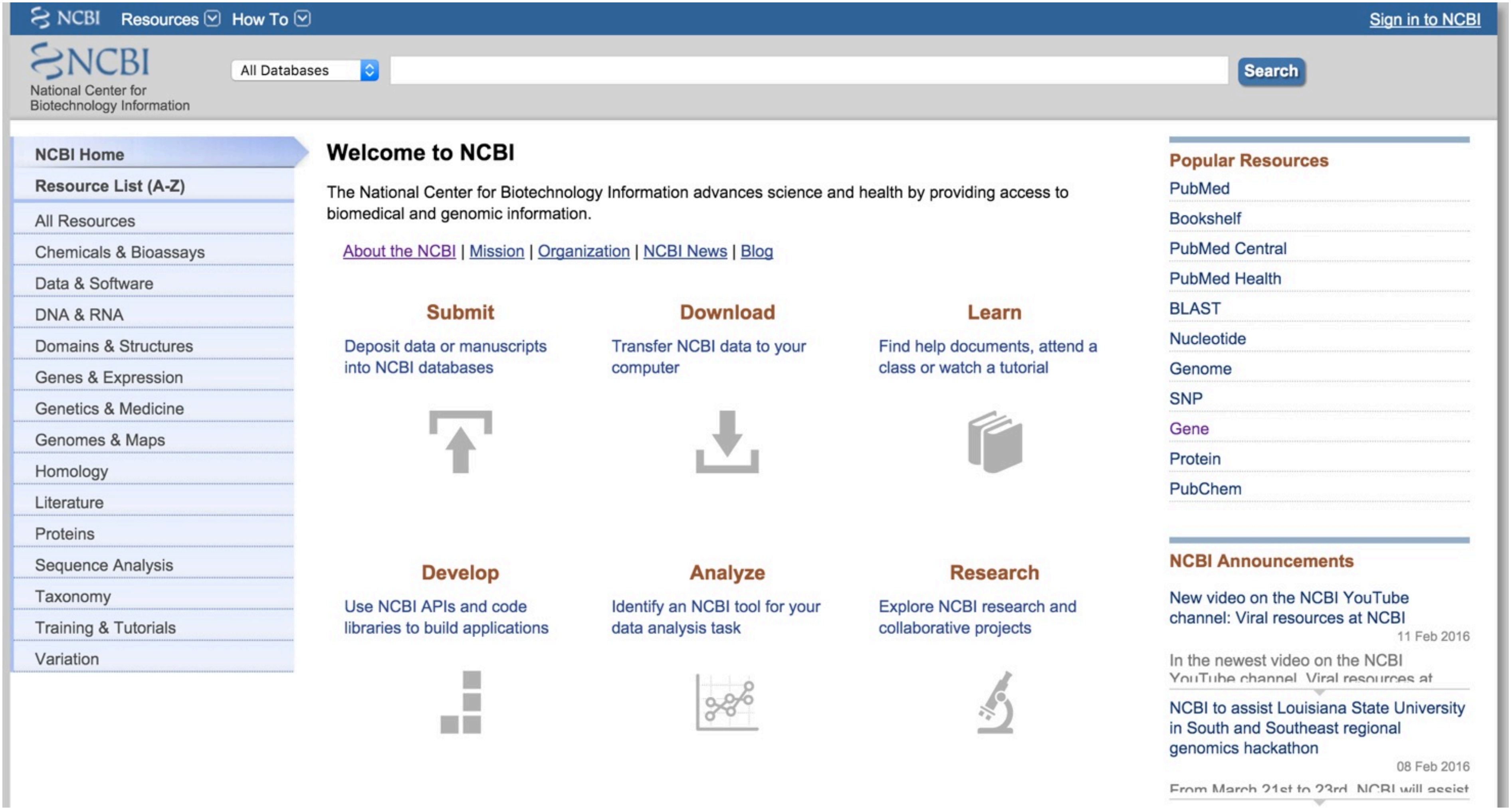
Chemical biology
chemogenomics & metabolomics

Ontologies
taxonomies & controlled vocabularies

Literature
Scientific publications & patents

Cross domain
cross-domain tools & resources

Repositories in USA: NCBI http://www.ncbi.nlm.nih.gov/



NCBI Home

Resource List (A-Z)

- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News | Blog

Submit

Deposit data or manuscripts into NCBI databases

Download

Transfer NCBI data to your computer

Learn

Find help documents, attend a class or watch a tutorial

Develop

Use NCBI APIs and code libraries to build applications

Analyze

Identify an NCBI tool for your data analysis task

Research

Explore NCBI research and collaborative projects

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- PubMed Health
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI Announcements

New video on the NCBI YouTube channel: Viral resources at NCBI

11 Feb 2016

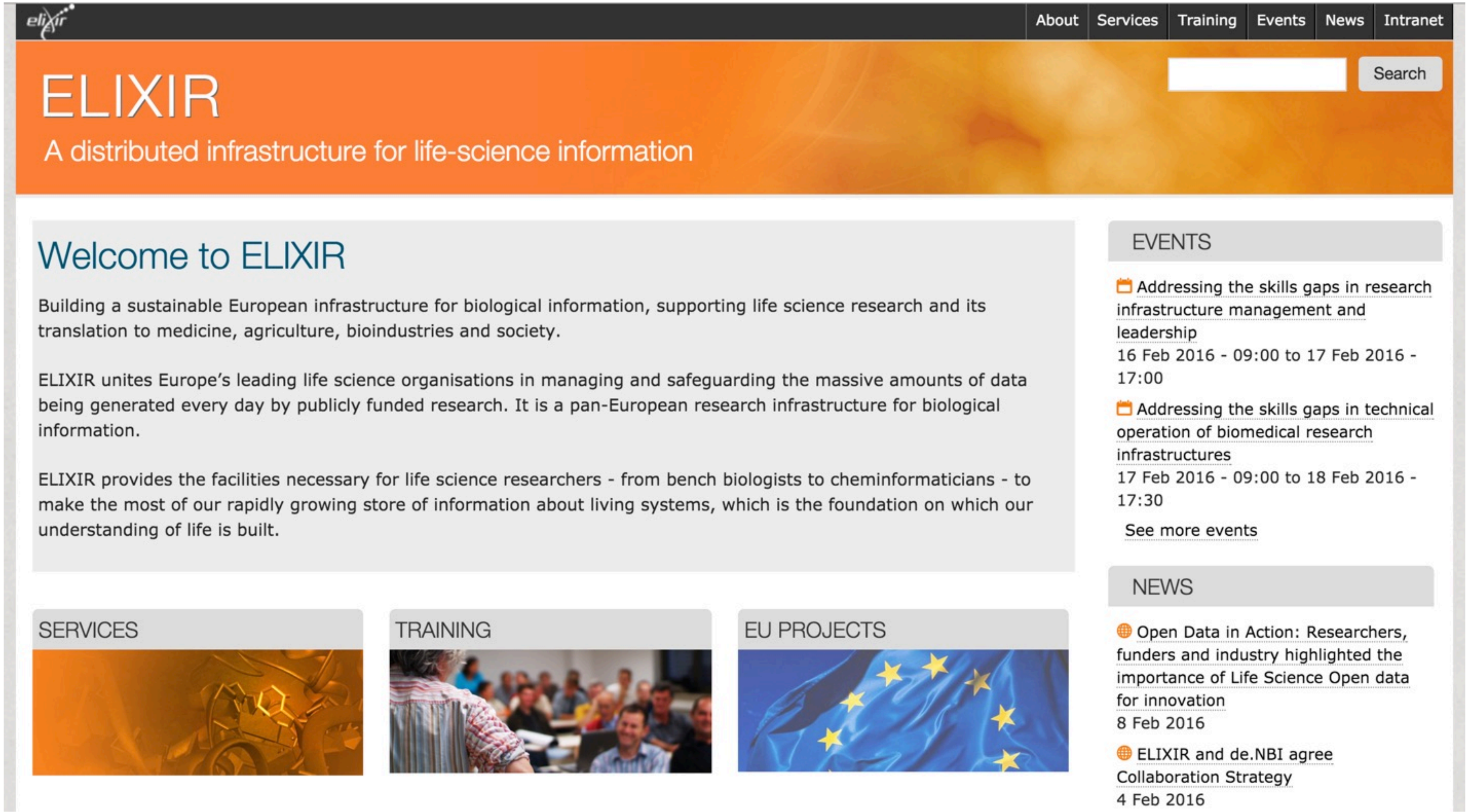
In the newest video on the NCBI YouTube channel, Viral resources at

NCBI to assist Louisiana State University in South and Southeast regional genomics hackathon

08 Feb 2016

From March 21st to 23rd, NCBI will assist

Scale up in EU: ELIXIR
https://www.elixir-europe.org/



ELIXIR

A distributed infrastructure for life-science information

Welcome to ELIXIR

Building a sustainable European infrastructure for biological information, supporting life science research and its translation to medicine, agriculture, bioindustries and society.

ELIXIR unites Europe's leading life science organisations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. It is a pan-European research infrastructure for biological information.

ELIXIR provides the facilities necessary for life science researchers - from bench biologists to cheminformaticians - to make the most of our rapidly growing store of information about living systems, which is the foundation on which our understanding of life is built.

SERVICES

TRAINING

EU PROJECTS

EVENTS

Addressing the skills gaps in research infrastructure management and leadership
16 Feb 2016 - 09:00 to 17 Feb 2016 - 17:00

Addressing the skills gaps in technical operation of biomedical research infrastructures
17 Feb 2016 - 09:00 to 18 Feb 2016 - 17:30

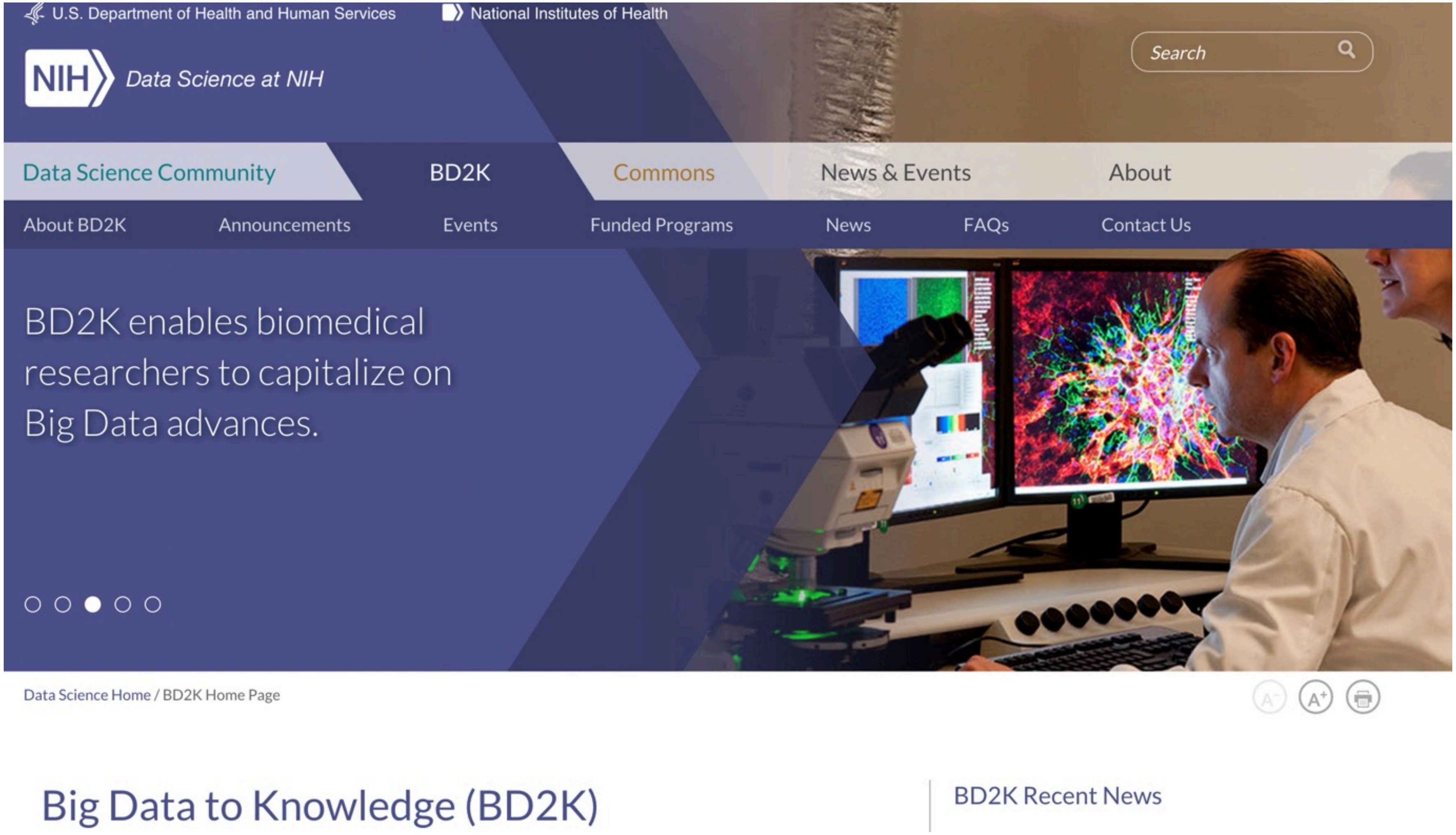
See more events

NEWS

Open Data in Action: Researchers, funders and industry highlighted the importance of Life Science Open data for innovation
8 Feb 2016

ELIXIR and de.NBI agree Collaboration Strategy
4 Feb 2016

Scale up in USA: BD2K
https://datascience.nih.gov/bd2k



U.S. Department of Health and Human Services

National Institutes of Health

NIH Data Science at NIH

BD2K enables biomedical researchers to capitalize on Big Data advances.

Data Science Home / BD2K Home Page

Big Data to Knowledge (BD2K)

BD2K Recent News

Biological research is a grand project

- Build complete models of biological systems

- Future application to human medicine
  - Disease redefinition
  - Improved drug development
  - Personalised medicine

Bioinformatics & Clinical Data

- **Pipelines:**
  - extract, clean, anonymise data from Electronic Health Records (EHRs)
- **Repositories:** NONE (privacy)
- **Resources:**
  - e.g. HSCIC (Health and Social Care Information Center), CPRD (Clinical Practice Research Datalink), PHE (Public Health England), NIHR HIC (Health Informatics Collaborative), GEL (Genomics England)
- **Algorithms:**
  - e.g. Text extraction

Bioinformatics & Genomic Medicine

- **Pipelines:**
  - align individual genome to reference and call *variants*
- **Repositories:**
  - e.g. dbSNP (variants), [But not for clinical – privacy]
- **Resources:**
  - e.g. Decipher, Ensembl
- **Analysis**:
  - e.g. New targets for drug development
- **Algorithms:**
  - e.g. Prediction of disease causing variants

Bioinformatics algorithm assessments

- Protein Structure
  - CASP – Critical Assessment of Structure Prediction (since 1994, CASP11 in 2014)
- Gene prediction
  - GASP, RGASP – Gene prediction and RNAseq assessments
- Variant effect prediction
  - CLARITY Challenge – 2012
    - http://genes.childrenshospital.org/
  - CAGI – 2010, 2011, 2013
    - https://genomeinterpretation.org/

Biological Information – in theory

- **DNA** (copy of genetic material in every cell)

  ↓ *predict*

- **RNA** (transcripts of expressed genes)

  ↓ *predict*

- **Protein** (translation of coding regions of genes)

  ↓ *predict*

- Interactions that make up **Cell**

  ↓ *predict*

- Interactions that make **Organisms**

- **Genetics differences**

  ↓ *predict consequences for disease*

Biological Information – in reality

- **DNA** (copy of genetic material in every cell)

  ↓ *predict*

- **RNA** (transcripts of expressed genes) ← *Expression, other omics data*

  ↓ *predict*

- **Protein** (translation of coding regions of genes)

  ↓ *predict*

- Interactions that make up **Cell** ← *iPS cells, organoids*

  ↓ *predict* ← *Tests, sensors*

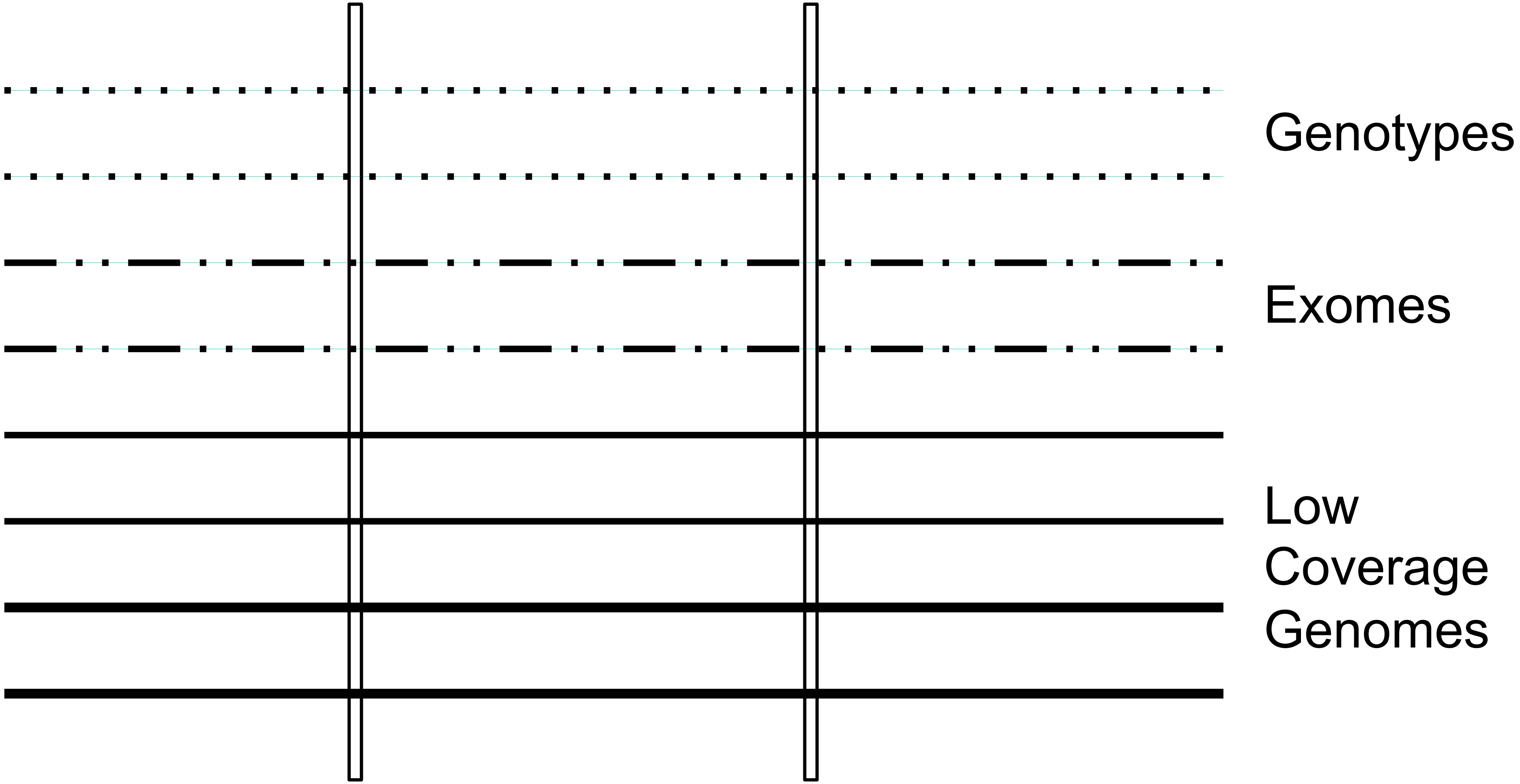
- Interactions that make **Organisms** ← *Observation, self reporting*

- **Genetics differences**

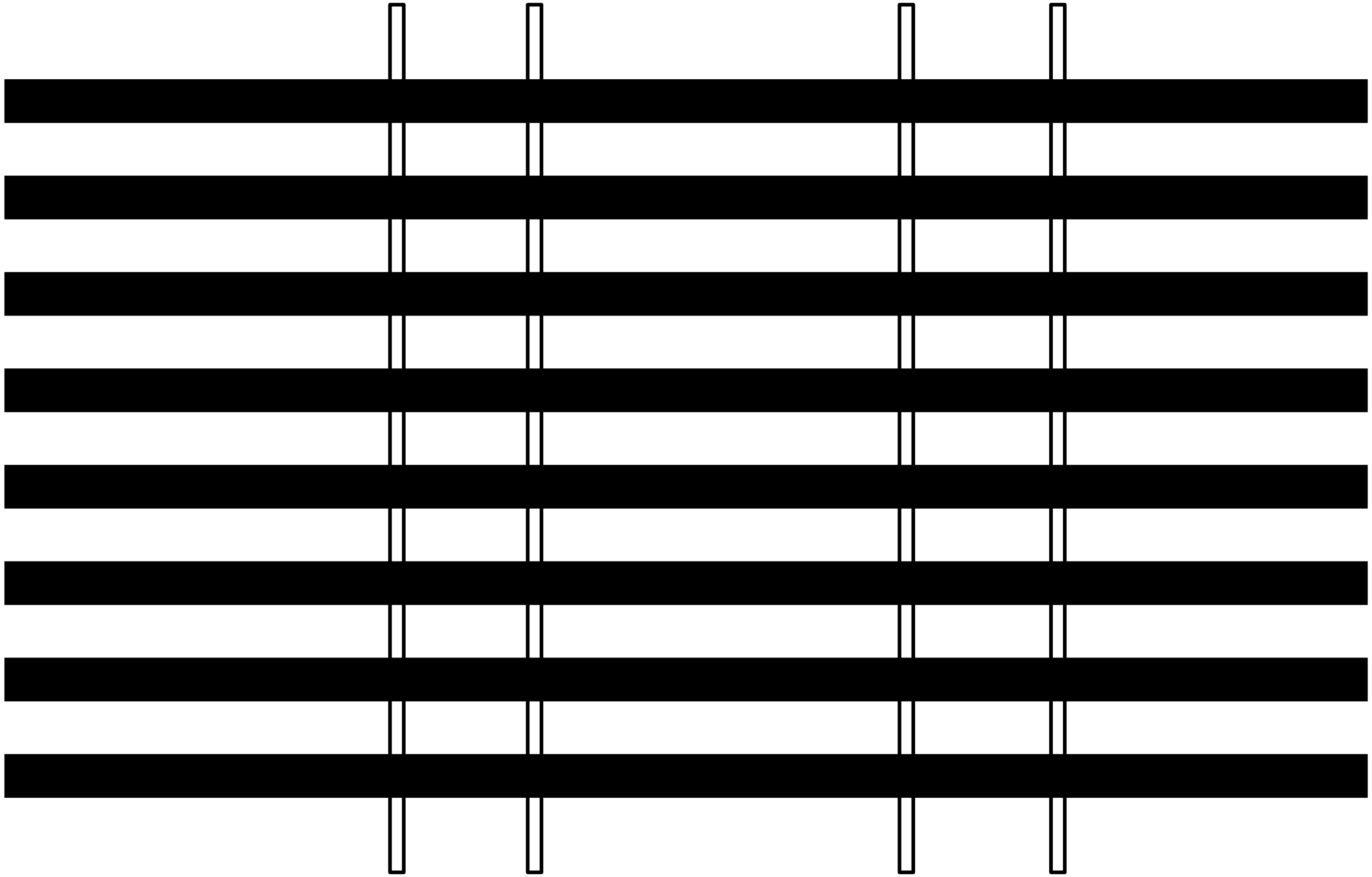
  ↓ *predict consequences for disease*

High throughput data collection cost effective way to extend understanding

From Genome Wide Association Studies (GWAS) to Whole Genome Analysis (WGA)



…to Whole Genome Analysis of Individual Genomes



Clinical
Coverage
Genome