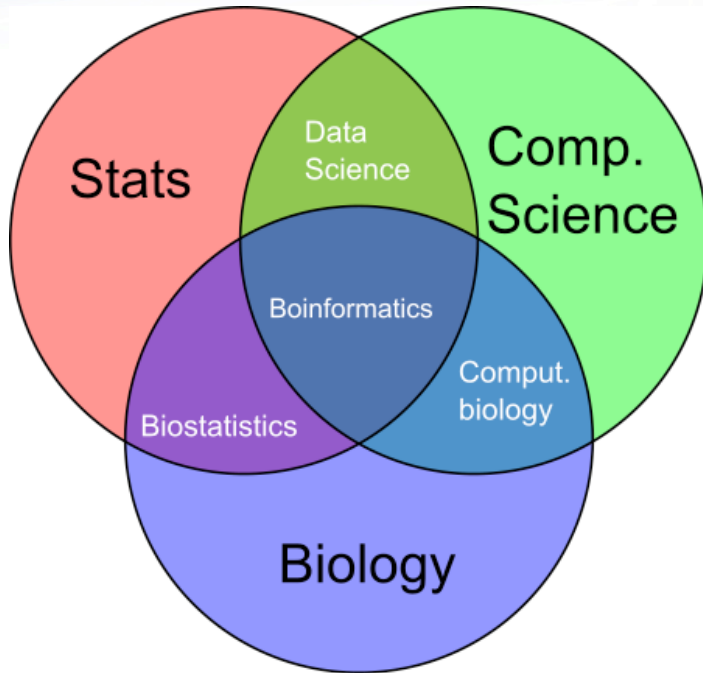


# IMBB 2016

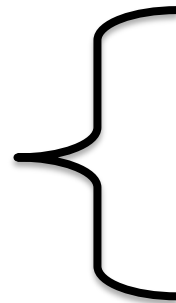
## Online Bioinformatics resources

Dedan Githae, Bioinformatics  
d.githae@cgiar.org

# Bioinformatics resources

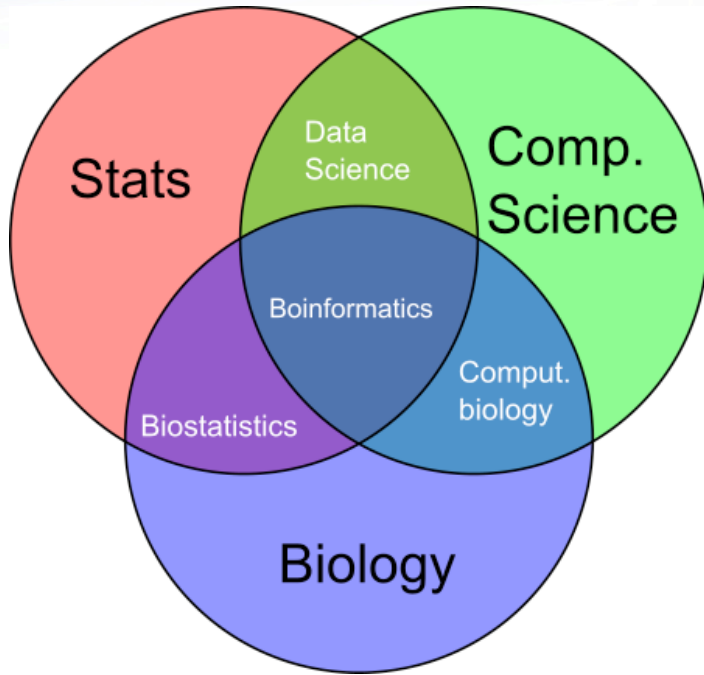


**BIOINFORMATICS**

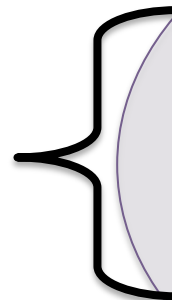


**1. Data (Locally or otherwise)**

**2. Tools (softwares and databases) to handle Data  
(see 1.)**



## BIOINFORMATICS



**1. Data**

**2. Tools (softwares and databases) to handle Data (see 1.)**

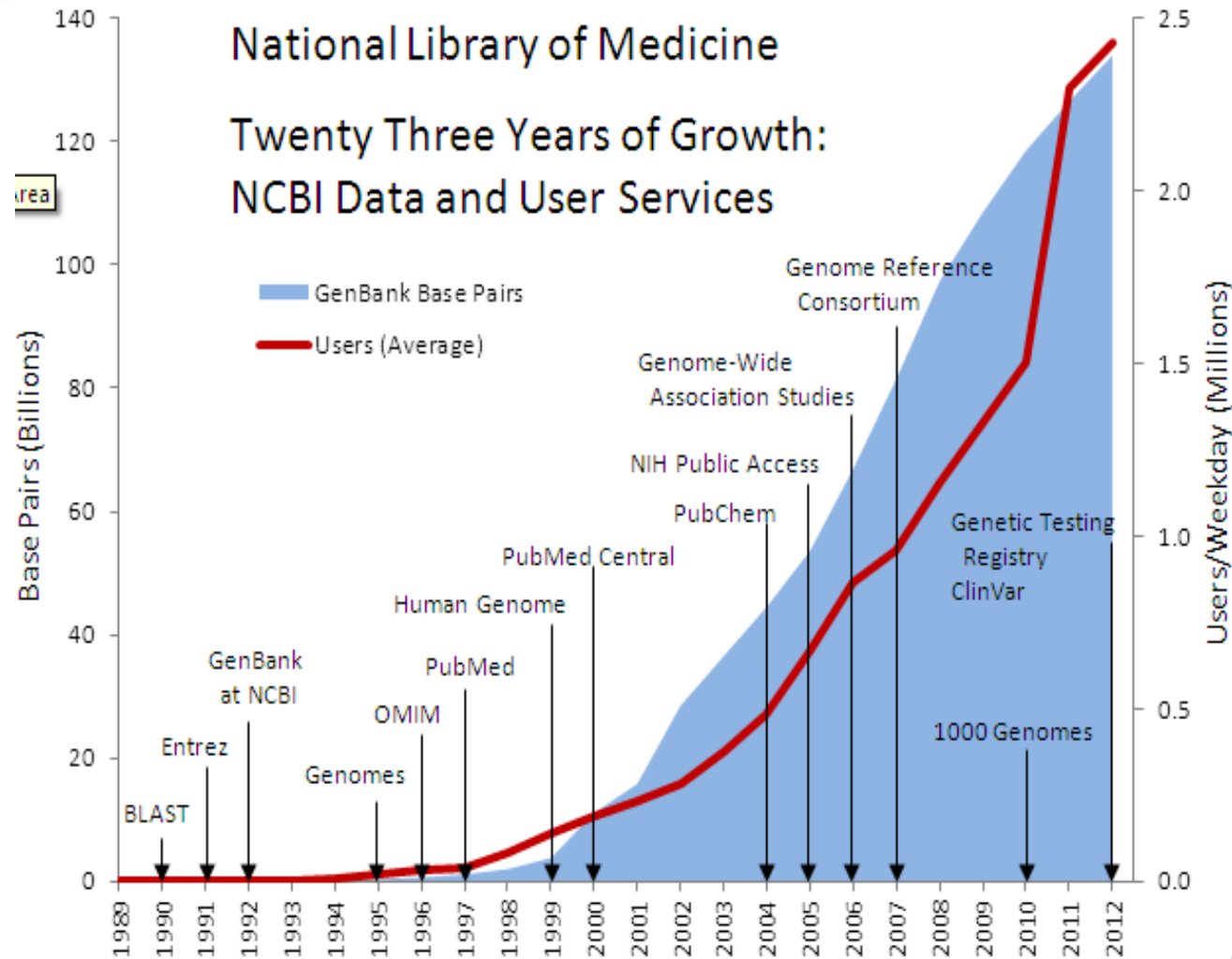


You need both  
Data and right tools  
(resources) to do  
Amazing research

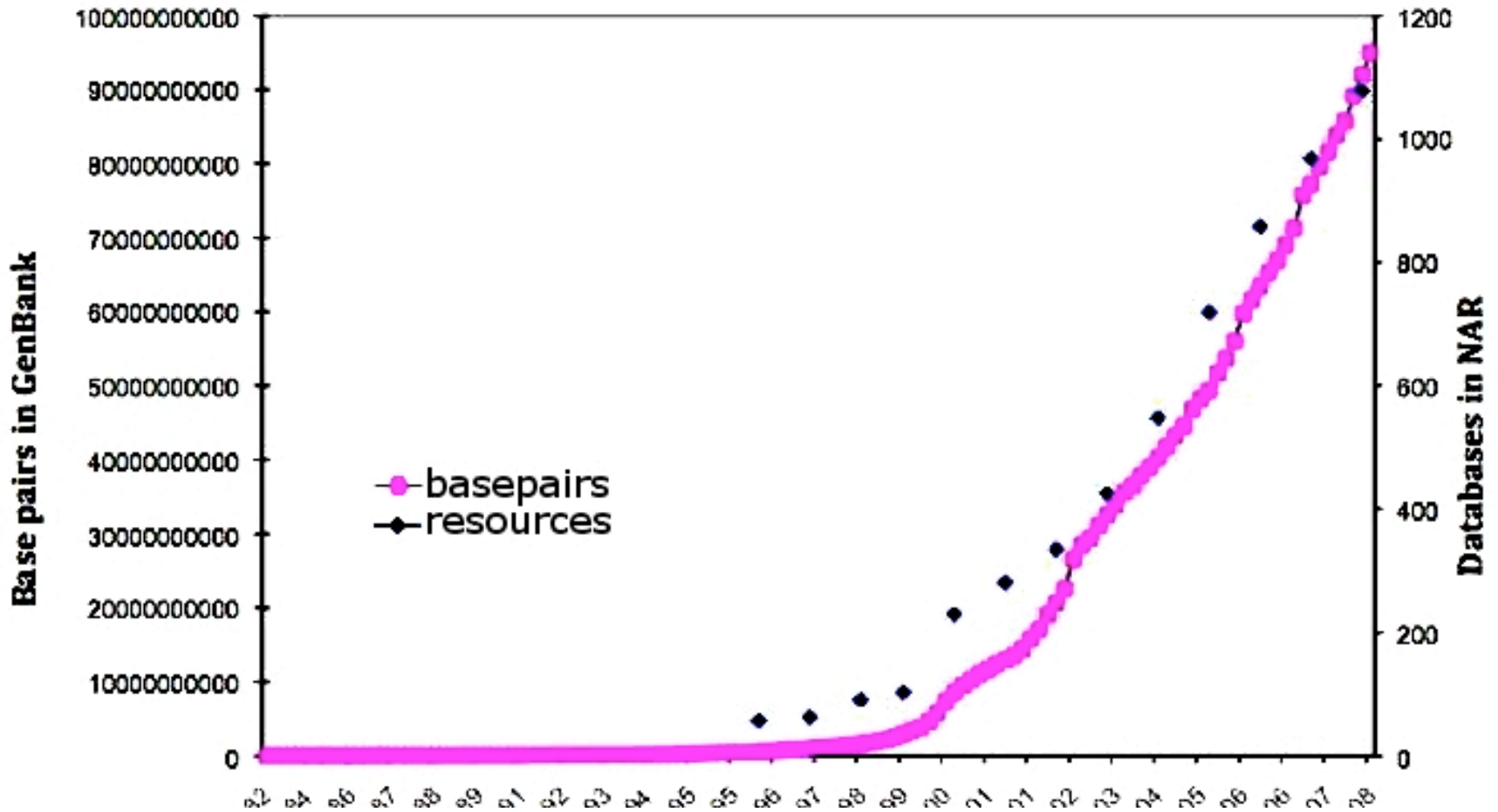
# 1. Data

- What data is available?
  - How much data is out there related to mine?
  - Where do I get the data?
  - What data is relevant to mine?
  - How much has been done on my topic?
- ...etc

# (Un)fortunately..



# (Un)fortunately..



Citation: Lathe, W., Williams, J., Mangan, M. & Karolchik, D. (2008)  
Genomic Data Resources: Challenges and Promises. Nature Education 1(3):2

# 1. Data: modern technology

- Lower cost
- Improved efficiency
  - The sequencers have GREATLY increased the rate at which data is being produced.
  - The sequencers have GREATLY increased the amount of biological data produced at a FASTER rate.
    - Both supersede the rate at which data can be interpreted.

Computing power  
+  
access to information (data)  
+  
right software (tools)



increase rate at answering  
research questions  
+  
new discoveries

# 1. Data(bases)

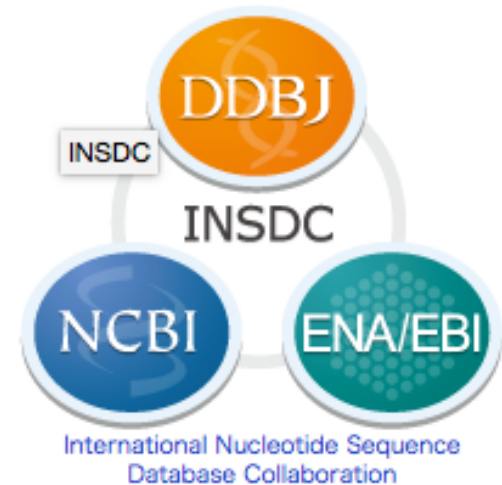
**Database:** a structured set of data held in a computer, especially one that is accessible in various ways.

## Sequence Databases

### Primary (DNA):

Consist of data derived experimentally, e.g. nucleotide sequences and 3D structures. The three, cooperate to make publicly available sequences available

- GenBank (USA), <http://www.ncbi.nlm.nih.gov/nucleotide/>
- European Nucleotide Archive (Europe) <http://www.ebi.ac.uk/ena> and
- DNA Database of Japan. <http://www.ddbj.nig.ac.jp/>



### Secondary databases:

Data is derived from analysis or treatment of primary data, such secondary structures, hydrophobicity plots, and domains are stored



# Protein sequence databases

- **Uniprot** (Universal protein resource): Database of protein sequences, and functional information.
  - *Swiss-prot (548,208): Manually annotated and reviewed*
  - *TrEMBL (46,714,516): Automatically annotated and not reviewed*
- **PIR**: Protein information resource

– <http://www.uniprot.org/>

– <http://pir.georgetown.edu/>

The screenshot displays the UniProt website interface. At the top, there is a search bar with 'UniProtKB' selected and an 'Advanced' search option. Below the search bar, there are navigation links for 'BLAST', 'Align', and 'Upload lists', along with 'Help' and 'Contact' links. The main content area features a grid of database links: UniProtKB (Swiss-Prot: Manually annotated and reviewed; TrEMBL: Automatically annotated and not reviewed), UniRef (Sequence clusters), UniParc (Sequence archive), and Proteomes. A 'Supporting data' section includes links for Literature citations, Taxonomy, Subcellular locations, Cross-ref. databases, Diseases, and Keywords. On the right, there is a 'News' section with social media icons and two news items: 'An old unwanted guest being shown the door | Cross-references for isoform sequences: Ensembl Genomes UniProt release 2014\_04' and 'Minority report | Cross-references for isoform sequences UniProt release 2014\_03'. At the bottom, there are sections for 'Getting started' (Text search, BLAST), 'UniProt data' (Download latest release, Statistics), and 'Protein spotlight' (On The Garden Pea).

# More...

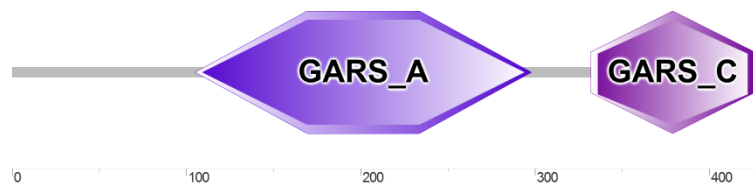
- **PROSITE**: Patterns of amino acids- For example N-glycosylation site motif takes the form:

**N{P}[ST]{P}**

To mean: Asn, followed by anything but Pro, followed by either Ser or Thr, followed by anything but Pro

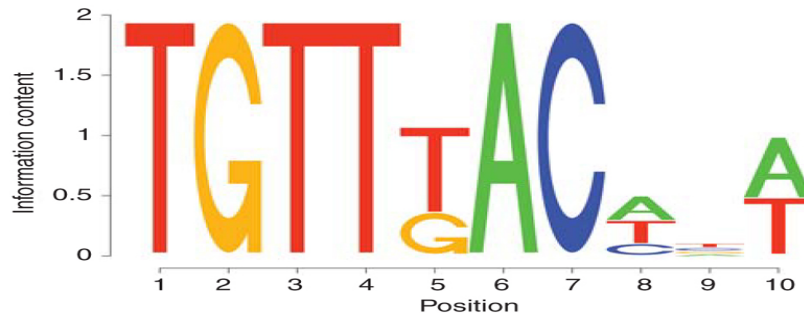
- **PRODOM**: Database of Protein domain families:  
<http://prodom.prabi.fr/prodom/current/html/home.php>

Other Databases include: **SMART; PROSITE; NCBI; CATH**



# 1. Data(bases)

- **PFAM:** Protein families database of alignments and HMM  
<http://pfam.xfam.org/>
- **PRINTS:** Protein “fingerprints” i.e. conserved motifs to characterise protein family
  - <http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php>

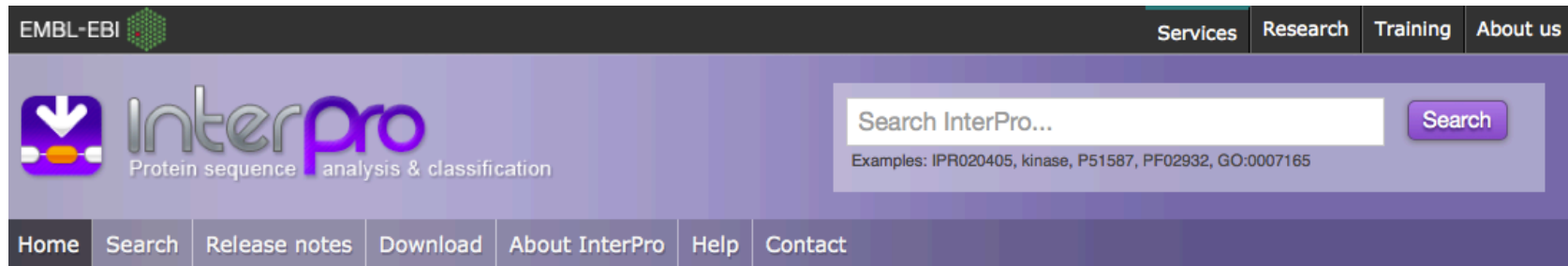


- **SignalP:** Predicts signal peptide prediction including cleavage site prediction  
<http://www.cbs.dtu.dk/services/SignalP/>

# 1. Data(bases)

- **Interpro**: provides functional analysis of proteins by classifying them into families and predicting domains and important sites.

<http://www.ebi.ac.uk/interpro/>



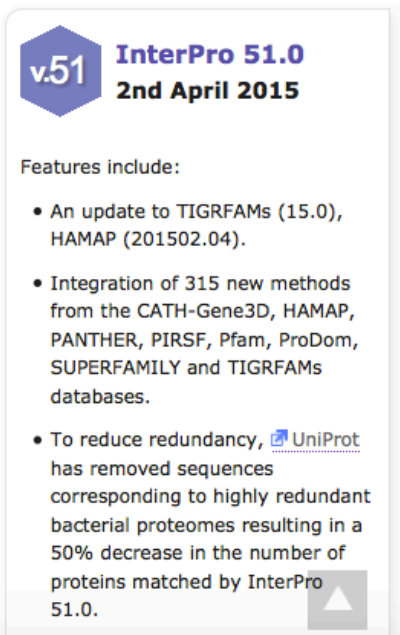
The screenshot shows the top navigation bar of the InterPro website. On the left is the EMBL-EBI logo. On the right are links for Services, Research, Training, and About us. Below this is the InterPro logo with the tagline 'Protein sequence analysis & classification'. To the right of the logo is a search bar with the placeholder text 'Search InterPro...' and a 'Search' button. Below the search bar are example search terms: 'Examples: IPR020405, kinase, P51587, PF02932, GO:0007165'. At the bottom of the header is a horizontal menu with links for Home, Search, Release notes, Download, About InterPro, Help, and Contact.

## InterPro: protein sequence analysis & classification

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. We combine protein signatures from a number of member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool. [Read more about InterPro](#)



The screenshot shows the search interface of the InterPro website. At the top is a tab labeled 'Analyse your protein sequence'. Below this is a large, empty text input field. At the bottom of the input field are three buttons: 'Search', 'Clear', and 'Example protein sequence'.



The screenshot shows a news announcement for InterPro 51.0. It features a blue hexagonal icon with 'v.51' inside. To the right of the icon is the text 'InterPro 51.0' and '2nd April 2015'. Below this is the heading 'Features include:' followed by a bulleted list of updates. The list includes an update to TIGRFAMs (15.0), HAMAP (201502.04), integration of 315 new methods from various databases, and a reduction in redundancy by removing highly redundant bacterial proteomes, resulting in a 50% decrease in the number of proteins matched by InterPro 51.0. A small upward-pointing arrow is visible at the bottom right of the announcement box.

**v.51 InterPro 51.0**  
**2nd April 2015**

Features include:

- An update to TIGRFAMs (15.0), HAMAP (201502.04).
- Integration of 315 new methods from the CATH-Gene3D, HAMAP, PANTHER, PIRSF, Pfam, ProDom, SUPERFAMILY and TIGRFAMs databases.
- To reduce redundancy, [UniProt](#) has removed sequences corresponding to highly redundant bacterial proteomes resulting in a 50% decrease in the number of proteins matched by InterPro 51.0.

## **Metabolic pathways**

BRENDA: comprehensive enzyme information

<http://www.brenda-enzymes.org/>

KEGG pathway DB (Kyoto Encyclopaedia of Genes and Genomes)


<http://www.genome.jp/kegg/>

Reactome

<http://www.reactome.org/>

# Genomic databases

## Hymenoptera Genome Database



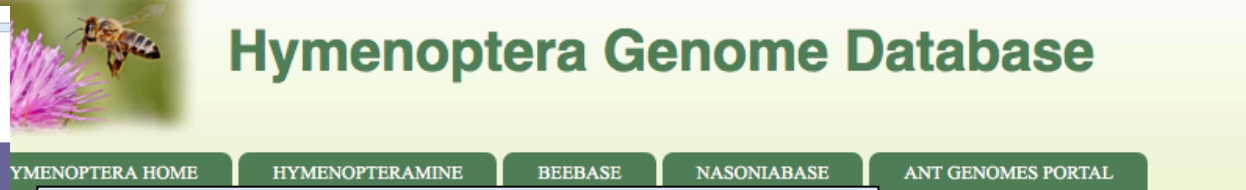
www.yeastgenome.org

SGD **Saccharomyces** GENOME DATABASE

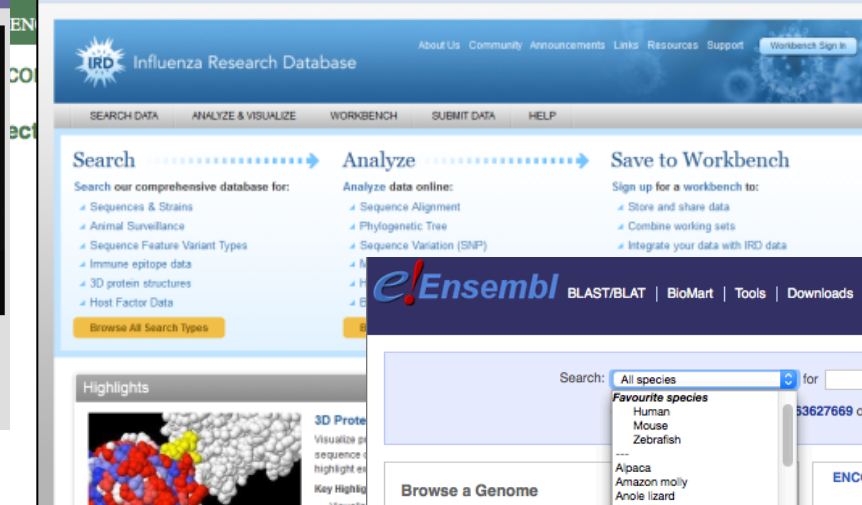
SGD Analyze Sequence Function Literature

**Localization of active Ras in a wild type strain**  
Image courtesy of S. Colombo and E. Martegani, University Milano Bicocca

**About SGD**  
The *Saccharomyces* Genome Database (SGD) provides comprehensive integrated biological information for the budding yeast *Saccharomyces*



HYMENOPTERA HOME HYMENOPTERAMINE BEEBASE NASONIABASE ANT GENOMES PORTAL



IRD Influenza Research Database

About Us Community Announcements Links Resources Support Workbench Sign In

SEARCH DATA ANALYZE & VISUALIZE WORKBENCH SUBMIT DATA HELP

**Search** **Analyze** **Save to Workbench**

Search our comprehensive database for:

- Sequences & Strains
- Animal Surveillance
- Sequence Feature Variant Types
- Immune epitope data
- 3D protein structures
- Host Factor Data

Analyze data online:

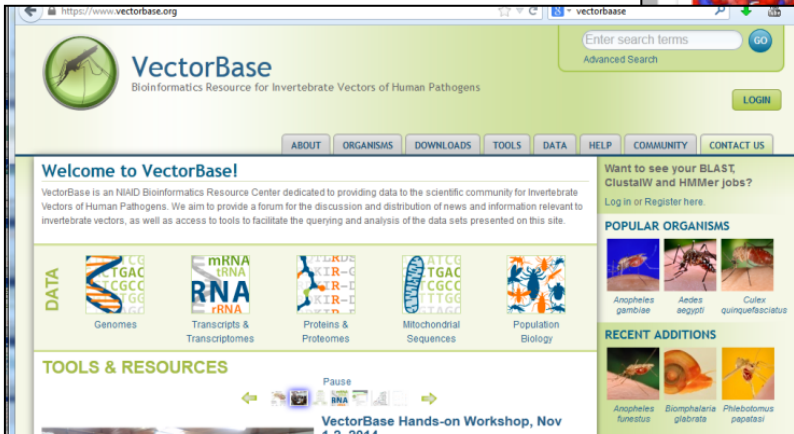
- Sequence Alignment
- Phylogenetic Tree
- Sequence Variation (SNP)

Sign up for a workbench to:

- Store and share data
- Combine working sets
- Integrate your data with IRD data

Highlights

3D Prote



https://www.vectorbase.org

VectorBase  
Bioinformatics Resource for Invertebrate Vectors of Human Pathogens

Enter search terms GO

Advanced Search

LOG IN

ABOUT ORGANISMS DOWNLOADS TOOLS DATA HELP COMMUNITY CONTACT US

**Welcome to VectorBase!**

VectorBase is an NIAID Bioinformatics Resource Center dedicated to providing data to the scientific community for Invertebrate Vectors of Human Pathogens. We aim to provide a forum for the discussion and distribution of news and information relevant to invertebrate vectors, as well as access to tools to facilitate the querying and analysis of the data sets presented on this site.

**DATA**

- Genomes
- mRNA RNA Transcripts & Transcriptomes
- Proteins & Proteomes
- Mitochondrial Sequences
- Population Biology

**TOOLS & RESOURCES**

Want to see your BLAST, ClustalW and HMMER jobs? Log in or Register here.

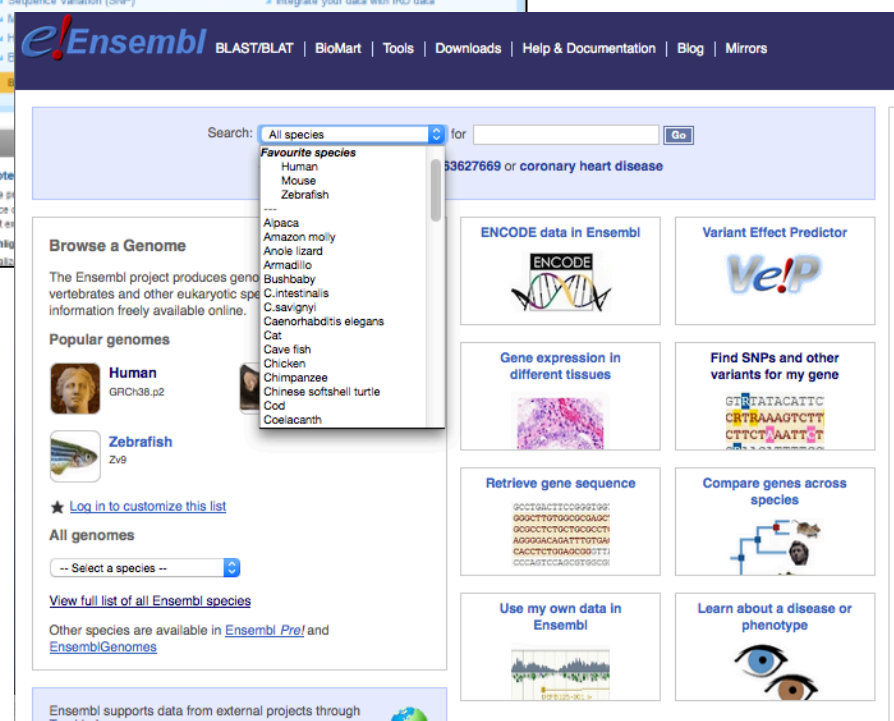
**POPULAR ORGANISMS**

- Anopheles gambiae
- Aedes aegypti
- Culex quinquefasciatus

**RECENT ADDITIONS**

- Anopheles funestus
- Biomphalaria gabriata
- Phlebotomus papatasi

VectorBase Hands-on Workshop, Nov 1-2, 2014



Ensembl BLAST/BLAT BioMart Tools Downloads Help & Documentation Blog Mirrors

Search: All species for 33627669 or coronary heart disease Go

**Favourite species**

- Human
- Mouse
- Zebrafish
- Alpaca
- Amazon molly
- Anole lizard
- Armadillo
- Bushbaby
- C.intestinalis
- C.savignyi
- Caenorhabditis elegans
- Cat
- Cave fish
- Chicken
- Chimpanzee
- Chinese softshell turtle
- Cod
- Coelacanth

**Browse a Genome**

The Ensembl project produces genomic data for vertebrates and other eukaryotic species. Information is freely available online.

**Popular genomes**

- Human** GRCh38.p2
- Zebrafish** Zv9

★ Log in to customize this list

**All genomes**

-- Select a species --

[View full list of all Ensembl species](#)

Other species are available in [Ensembl Pre/](#) and [EnsemblGenomes](#)

Ensembl supports data from external projects through

**ENCODE data in Ensembl**

**Variant Effect Predictor**

**Gene expression in different tissues**

**Find SNPs and other variants for my gene**

**Retrieve gene sequence**

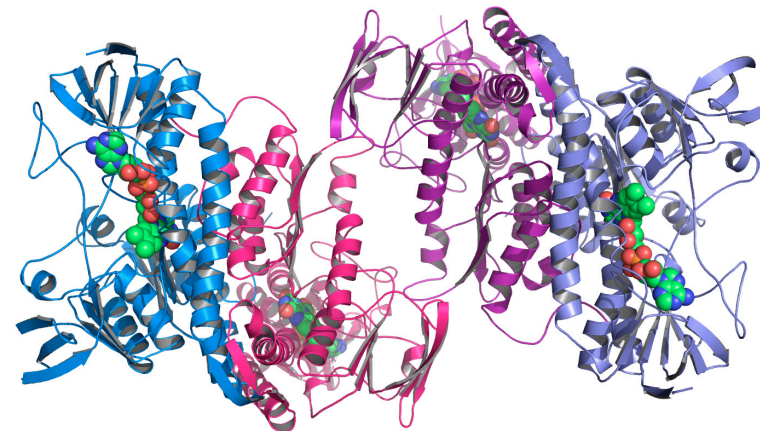
**Compare genes across species**

**Use my own data in Ensembl**

**Learn about a disease or phenotype**

# Protein structure databases

- **Protein Databank (PDB)** consists of experimentally validated protein structure e.g. x-ray crystallography, NMR.
- **ModBase**: A database of annotated comparative protein structure models (Modelled proteins)
- **SCOP**: Structural classification of Proteins Depending on  $\alpha$  ;  $\beta$  ;  $\alpha + \beta$  ; membrane & cell surface proteins; small proteins; coiled coil proteins, etc.
- **CATH**: hierarchical domain classification of protein structures in the Protein Data Bank (Class | Architecture | Topology | Homologous super-families)



# Software tools

- **Journals**

e.g. *Bioinformatics*, *Nucleic Acids Research*, *Journal of Molecular Biology*, *Protein science* publish papers on cutting edge developments and innovations in computational biology methods

- Most **biological databases** have software resource listings- e.g. Sequence searching, visualisation resources (genome / alignment / genome level).

- **Web servers:** “Simple” web implementation of the softwares. Clear inputs, outputs, parameters, graphical data representation and downloadable results. [www.ebi.ac.uk](http://www.ebi.ac.uk)

Examples?

## Mobyle @Pasteur

The screenshot shows the Mobyle @Pasteur website interface. At the top, there is a search bar with a "Search" button and a "[more]" link. Below the search bar, the content is organized into three main sections: "Programs", "Workflows", and "Tutorials".

- Programs** (blue header):
  - alignment
  - assembly
  - database
  - display
  - genetics
  - hmm
  - information
  - nucleic
  - phylogeny
  - protein
  - sequence
  - structure
- Workflows** (blue header):
  - alignment
  - database
  - phylogeny
  - blast\_to\_multialign
  - hmm\_build\_search
  - mafft-cons-tree
  - protein\_distance\_phylogeny
- Tutorials** (red header):
  - data formats
  - BMPS\_tutorial
  - registration
  - setpbystep



# Where to get information

- ◆ **Journal Website:** Almost every major journal provides a web access to abstracts is usually free, even when the content is subscription.
- ◆ **E-journals:** Some electronic journals are online-only journals; some are online versions of printed journals, and some consist of the online equivalent of a printed journal, but with additional online-only (sometimes video and interactive media) material.

The screenshot shows the BMC Bioinformatics website. At the top left is the BMC Bioinformatics logo with an Impact Factor of 3.02. A search bar is located at the top right. Below the logo are navigation tabs: Home, Articles, Authors, Reviewers, About this journal, and My BMC Bioinformatics. The main content area is titled 'Articles' and features a filter bar with options for 'All articles', 'Sections', 'Most popular', 'Archive', 'Supplements', and 'Article collections'. Below the filter bar are dropdown menus for 'Show', 'All article types', 'All sections', and 'Supplements', along with 'Vol.', 'Art. No.', and 'Jump' buttons. A pagination bar shows 'Page 1 of 186' and 'Articles per page: 25 | 50 | 100'. Three article listings are visible, each with a title, authors, journal name, volume, issue, and date, and links for 'Abstract' and 'Provisional PDF'.

The screenshot shows the article page for 'Direct and site-specific quantification of RNA 2'-O-methylation by PCR with an engineered DNA polymerase' in the journal 'Nucleic Acids Research'. The journal title is at the top in a large font. Below it are navigation links: 'ABOUT THIS JOURNAL', 'CONTACT THIS JOURNAL', 'SUBSCRIPTIONS', 'CURRENT ISSUE', 'ARCHIVE', and 'SEARCH'. The article title is prominently displayed in the center. Below the title are the authors 'Joos Aschenbrenner and Andreas Marx' and their affiliation: 'Department of Chemistry, Konstanz Research School Chemical Biology, University of Konstanz, Universitätsstraße 10, D-78457 Konstanz, Germany'. There are buttons for 'Submit a manuscript', 'Register', 'Sign up for article alerts', and 'Contact us', along with a 'Follow us on Twitter' link. The article abstract is visible, starting with 'Methylation of the 2'-hydroxyl-group of ribonucleotides is found in all major classes of RNA in eukaryotes...'. On the right side, there are links for 'Previous | Next Article' and 'Table of Contents', and a section for 'This Article' with details like 'Nucl. Acids Res. (05 May 2016) 44 (8): 3495-3502' and 'First published online: March 25, 2016'. There are also links for 'Full Text (HTML) Free', 'Full Text (PDF) Free', and 'SUPPLEMENTARY DATA'. At the bottom, there are 'Classifications' and 'Services' sections.



Search

## Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE SEARCH

Institution: ILRI Sign In as Personal Subscriber

Oxford Journals > Science & Mathematics > Nucleic Acids Research > Volume 44 Issue 8 > Pp. 3495-3502.

Home Articles Authors Reviewers About this journal My BMC Bioinformatics

### Articles

All articles Sections Most popular Archive Supplements Article collections

Show -- All article types -- -- All sections --  Supplements

Page 1 of 186

1 2 3

Display/download options

### Direct and site-specific quantification of RNA 2'-O-methylation by PCR with an engineered DNA polymerase

« Previous | Next Article »  
Table of Contents

#### This Article

Nucl. Acids Res. (05 May 2016) 44(8): 3495-3502.  
doi: 10.1093/nar/gkw200  
First published online: March 25, 2016

This article is Open Access  
Abstract **Free**

» Full Text (HTML) **Free**  
Full Text (PDF) **Free**  
SUPPLEMENTARY DATA

All Versions of this Article:  
gkw200v1  
44/8/3495 **most recent**

#### Classifications

Chemical Biology and Nucleic Acid Chemistry

#### Services

Article metrics  
Alert me when cited

Joos Aschenbrenner and Andreas Marx\*

Department of Chemistry, Konstanz Research School Chemical Biology, University of Konstanz, Universitätsstraße 10, D-78457 Konstanz, Germany

\*To whom correspondence should be addressed. Tel: +49 7531 885139; Fax: +49 7531 885140; Email: andreas.marx@uni-konstanz.de

Received February 17, 2016.  
Revision received March 11, 2016.  
Accepted March 14, 2016.

#### Abstract

Methylation of the 2'-hydroxyl-group of ribonucleotides is found in all major classes of RNA in eukaryotes and is one of the most abundant posttranscriptional modifications of stable RNAs. In spite of intense studies, the multiple functions of RNA 2'-O-methylation are still not understood. One major obstacle in the field are the technical demanding detection methods, which are typically laborious and do not always deliver

Automated pipeline for RT-PCR primer design, targeted at exon-junction sites.

# Information

- ◆ **Servers** (eg NCBI Pubmed; Google scholar; SCOPUS) A search engine to search references and abstracts on life sciences and biomedical topics in multiple databases

NCBI Resources How To Sign in to NCBI

PubMed.gov PubMed  Search

US National Library of Medicine National Institutes of Health Advanced

**PubMed**

PubMed comprises more than 23 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites.

**PubMed Commons**

PubMed's new commenting system

More

## Using PubMed

[PubMed Quick Start Guide](#)

[Full Text Articles](#)

[PubMed FAQs](#)

[PubMed Tutorials](#)

[New and Noteworthy](#)

## PubMed Tools

[PubMed Mobile](#)

[Single Citation Matcher](#)

[Batch Citation Matcher](#)

[Clinical Queries](#)

[Topic-Specific Queries](#)

## More Resources

[MeSH Database](#)

[Journals in NCBI Databases](#)

[Clinical Trials](#)

[E-Utilities](#)

[LinkOut](#)

You are here: NCBI > Literature > PubMed

[Write to the Help Desk](#)

## GETTING STARTED

[NCBI Education](#)

[NCBI Help Manual](#)

[NCBI Handbook](#)

## RESOURCES

[Chemicals & Bioassays](#)

[Data & Software](#)

[DNA & RNA](#)

## POPULAR

[PubMed](#)

[Bookshelf](#)

[PubMed Central](#)

## FEATURED

[Genetic Testing Registry](#)

[PubMed Health](#)

[GenBank](#)

## NCBI INFORMATION

[About NCBI](#)

[Research at NCBI](#)

[NCBI News](#)

# In a nutshell

Lots of data available...

More data being produced .....

A ton of software out there..

And new, better computational algorithms being produced...

PHASE TWO: INTERPRETATION

SLEDMAN The Star Ledger





*Thank You*

**Dedan Githae**  
*Bioinformatician*  
*d.githae@cgiar.org*

***Online Bioinformatics resources***  
*<http://hub.africabiosciences.org>*



**biosciences**  
eastern and central **africa**



**NC STATE**  
**UNIVERSITY**

**ILRI**

INTERNATIONAL  
LIVESTOCK RESEARCH  
INSTITUTE