Jon K. Lærdahl,
Structural Bioinformatics

Bioinformatics for Molecular Biology

Databases & Accessing data

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Today’s Programme

- Biological databases
- Brief introduction
  - What is UNIX?
  - Why should you learn UNIX?
- Bioinformatics Core Facility
- Setting up your laptops

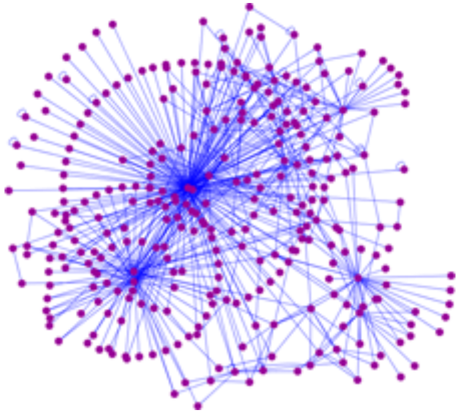
What about those of you that know Unix and Python very well?

- Very briefly on the Unix shell, file system and some commands
- UNIX basics exercise
- Tomorrow, continue on databases & working with biological sequences

Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned.

NCBI – A Science Primer

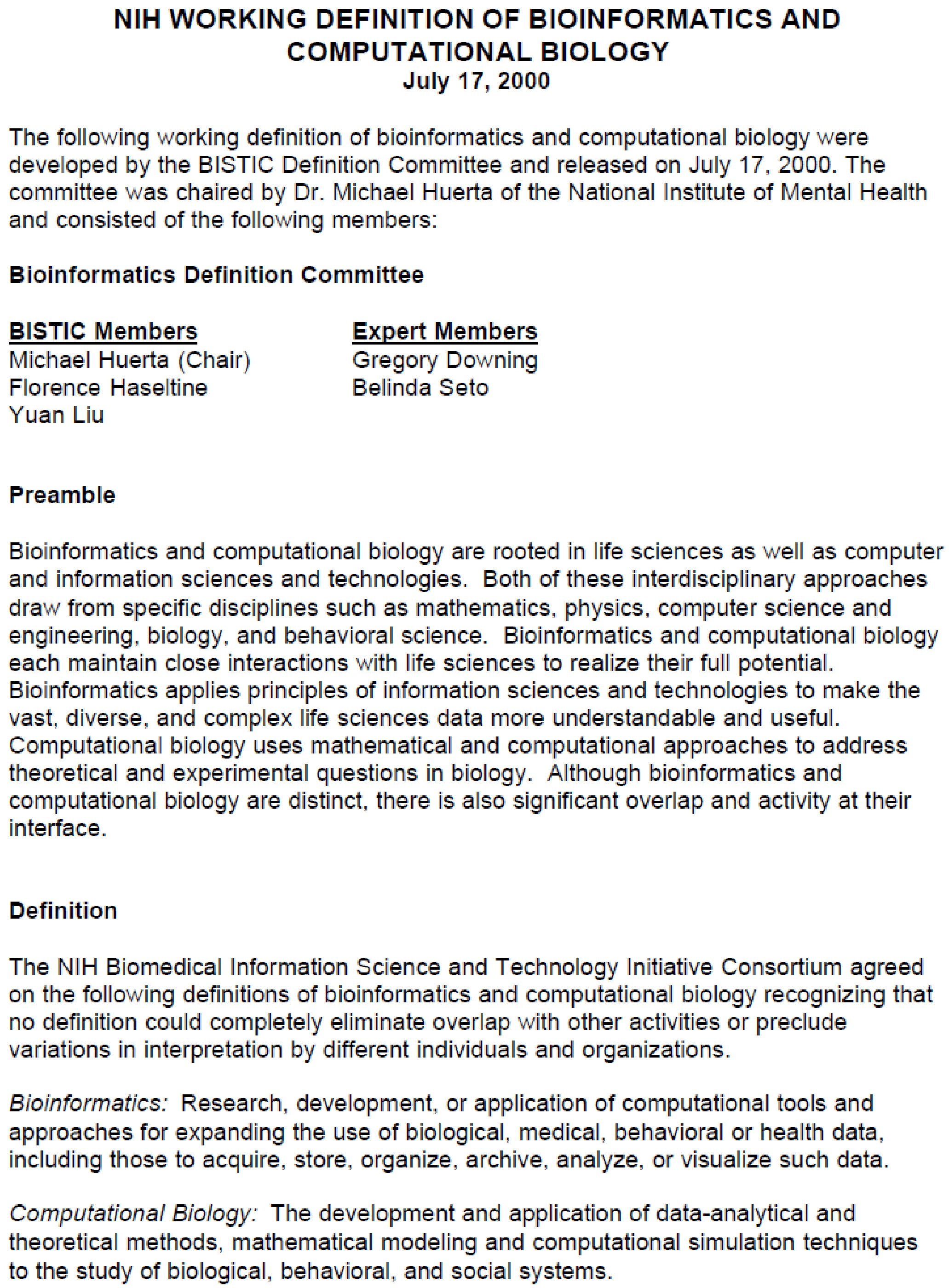
Biology in the 21st century is being transformed from a purely lab-based science to an information science as well.



Wikipedia:

Bioinformatics is a branch of biological science which deals with the study of methods for storing, retrieving and analyzing biological data, such as nucleic acid (DNA/RNA) and protein sequence, structure, function, pathways and genetic interactions. It generates new knowledge that is useful in such fields as drug design and development of new software tools to create that knowledge. Bioinformatics also deals with algorithms, databases and information systems, web technologies, artificial intelligence and soft computing, information and computation theory, structural biology, software engineering, data mining, image processing, modeling and simulation, discrete mathematics, control and system theory, circuit theory, and statistics.

Bigger than biology?



NIH WORKING DEFINITION OF BIOINFORMATICS AND COMPUTATIONAL BIOLOGY

July 17, 2000

The following working definition of bioinformatics and computational biology were developed by the BISTIC Definition Committee and released on July 17, 2000. The committee was chaired by Dr. Michael Huerta of the National Institute of Mental Health and consisted of the following members:

Bioinformatics Definition Committee

BISTIC Members	**Expert Members**
Michael Huerta (Chair)	Gregory Downing
Florence Haseltine	Belinda Seto
Yuan Liu	

Preamble

Bioinformatics and computational biology are rooted in life sciences as well as computer and information sciences and technologies. Both of these interdisciplinary approaches draw from specific disciplines such as mathematics, physics, computer science and engineering, biology, and behavioral science. Bioinformatics and computational biology each maintain close interactions with life sciences to realize their full potential. Bioinformatics applies principles of information sciences and technologies to make the vast, diverse, and complex life sciences data more understandable and useful. Computational biology uses mathematical and computational approaches to address theoretical and experimental questions in biology. Although bioinformatics and computational biology are distinct, there is also significant overlap and activity at their interface.

Definition

The NIH Biomedical Information Science and Technology Initiative Consortium agreed on the following definitions of bioinformatics and computational biology recognizing that no definition could completely eliminate overlap with other activities or preclude variations in interpretation by different individuals and organizations.

Bioinformatics: Research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, archive, analyze, or visualize such data.

Computational Biology: The development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological, behavioral, and social systems.

Certainly not exactly clear distinction between bioinformatics and the rest of science

CLS (Computational Life Science)

If you want to do state-of-the art research in biology or molecular medicine in 2014 you need bioinformatics/CLS/informatics competence!!

Some examples

Vol 467 | 2 September 2010 | doi:10.1038/nature09322

Genome-wide measurement of RNA secondary structure in yeast

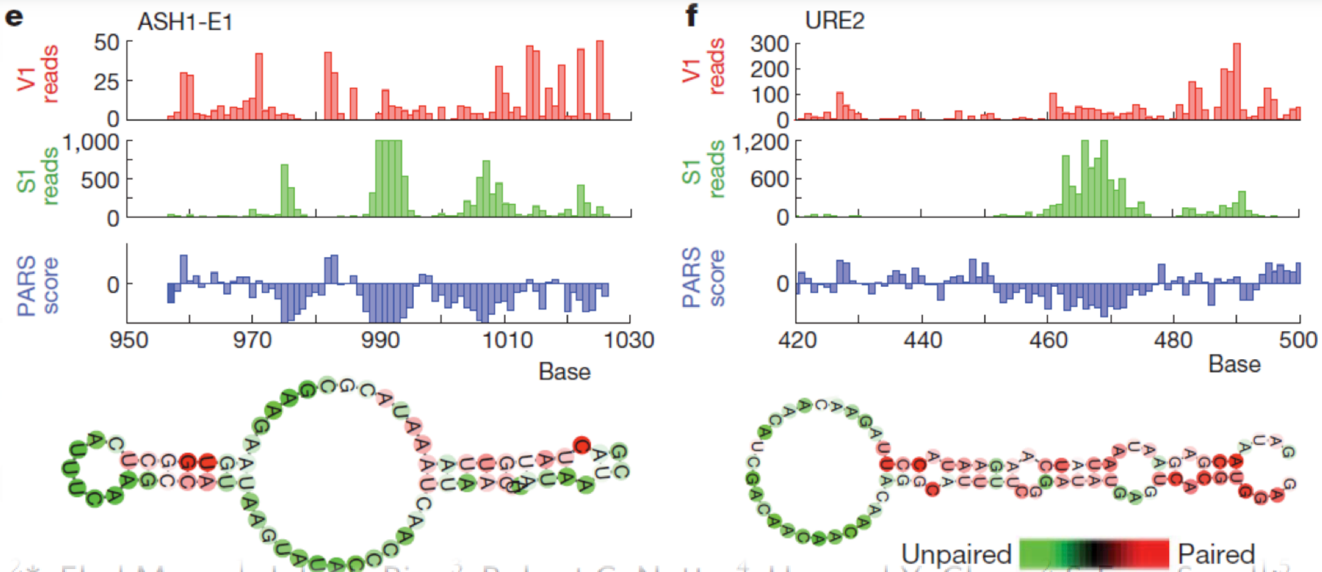
Michael Kertesz[1]*†, Yue Wan[2]*, Elad Mazor[1], John L. Rinn[3], Robert C. Nutter[4], Howard Y. Chang[2] & Eran Segal[1,5]

The structures of RNA molecules are often important for their function and regulation[1–6], yet there are no experimental techniques for genome-scale measurement of RNA structure. Here we describe a novel strategy termed parallel analysis of RNA structure (PARS), which is based on deep sequencing fragments of RNAs that were treated with structure-specific enzymes, thus providing simultaneous *in vitro* profiling of the secondary structure of thousands of RNA species at single nucleotide resolution. We apply PARS to profile the secondary structure of the messenger RNAs (mRNAs) of the budding yeast *Saccharomyces cerevisiae* and obtain structural profiles for over 3,000 distinct transcripts. Analysis of these profiles reveals several RNA structural properties of yeast transcripts, including the existence of more secondary structure over coding regions compared with untranslated regions, a three-nucleotide periodicity of secondary structure across coding regions and an anti-correlation between the efficiency with which an mRNA is translated and the structure over its translation start site. PARS is readily applicable to other organisms and to profiling RNA structure in diverse conditions, thus enabling studies of the dynamics of secondary structure at a genomic scale.

that typically have 5′ hydroxyl (Supplementary Fig. 3). Thus each observed cleavage site provides evidence that the cut nucleotide was in a double-stranded (for V1-treated samples) or single-stranded (for S1-treated samples) conformation. As a quantitative measure at nucleotide resolution representing the degree to which a nucleotide was in a double- or single-stranded conformation, we took the log ratio between the number of sequence reads obtained for each nucleotide in the V1 and S1 experiments. A higher (lower) log ratio, or PARS score, thus denotes a higher (lower) probability for a nucleotide to be in a double-stranded conformation.

We performed four independent V1 experiments and three independent S1 experiments, which were highly reproducible across replicates (correlation = 0.60–0.93, Supplementary Table 1), resulting in over 85 million sequence reads that map to the yeast genome, of which approximately 97% mapped to annotated transcripts (Supplementary Table 2). At an average nucleotide coverage above 1.0, we obtained structural information for over 3,000 yeast transcripts (Supplementary Table 3 and Supplementary Fig. 4a), covering in total over 4.2 million transcribed bases, which is approximately 100-fold more than all published RNA footprints to date.

Vol 467|2 September 2010|doi:10.103



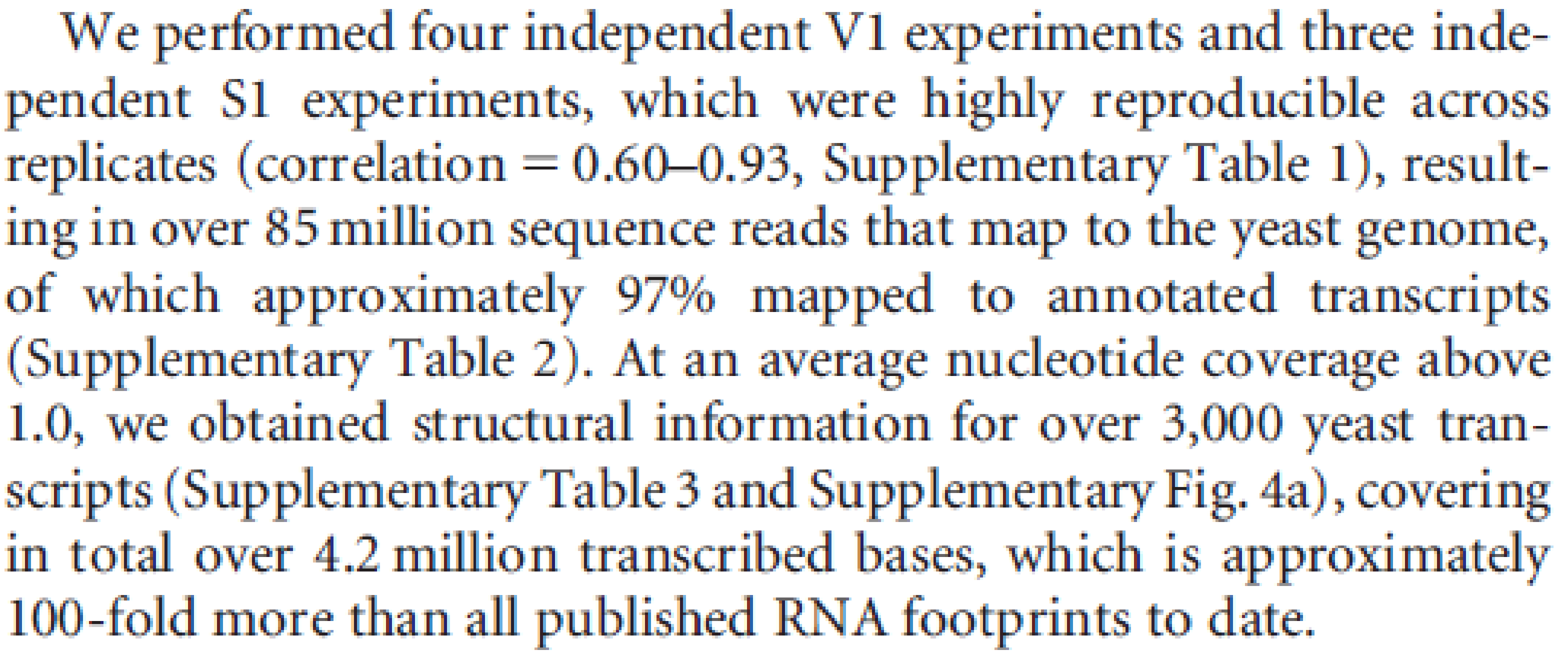
Genome-wide structure in y

Michael Kertesz[1]*†, Yue Wan[2]*, Elad Mazor[1], John L. Rinn[3], Robert C. Nutter[4], Howard Y. Chang[2] & Eran Segal[1,5]

The structures of RNA molecules are often important for their function and regulation[1–6], yet there are no experimental techniques for genome-scale measurement of RNA structure. Here we describe a novel strategy termed parallel analysis of RNA structure (PARS), which is based on deep sequencing fragments of RNAs that were treated with structure-specific enzymes, thus providing simultaneous *in vitro* profiling of the secondary structure of thousands of RNA species at single nucleotide resolution. We apply PARS to profile the secondary structure of the messenger RNAs (mRNAs) of the budding yeast *Saccharomyces cerevisiae* and obtain structural profiles for over 3,000 distinct transcripts. Analysis of these profiles reveals several RNA structural properties of yeast transcripts, including the existence of more secondary structure over coding regions compared with untranslated regions, a three-nucleotide periodicity of secondary structure across coding regions and an anti-correlation between the efficiency with which an mRNA is translated and the structure over its translation start site. PARS is readily applicable to other organisms and to profiling RNA structure in diverse conditions, thus enabling studies of the dynamics of secondary structure at a genomic scale.

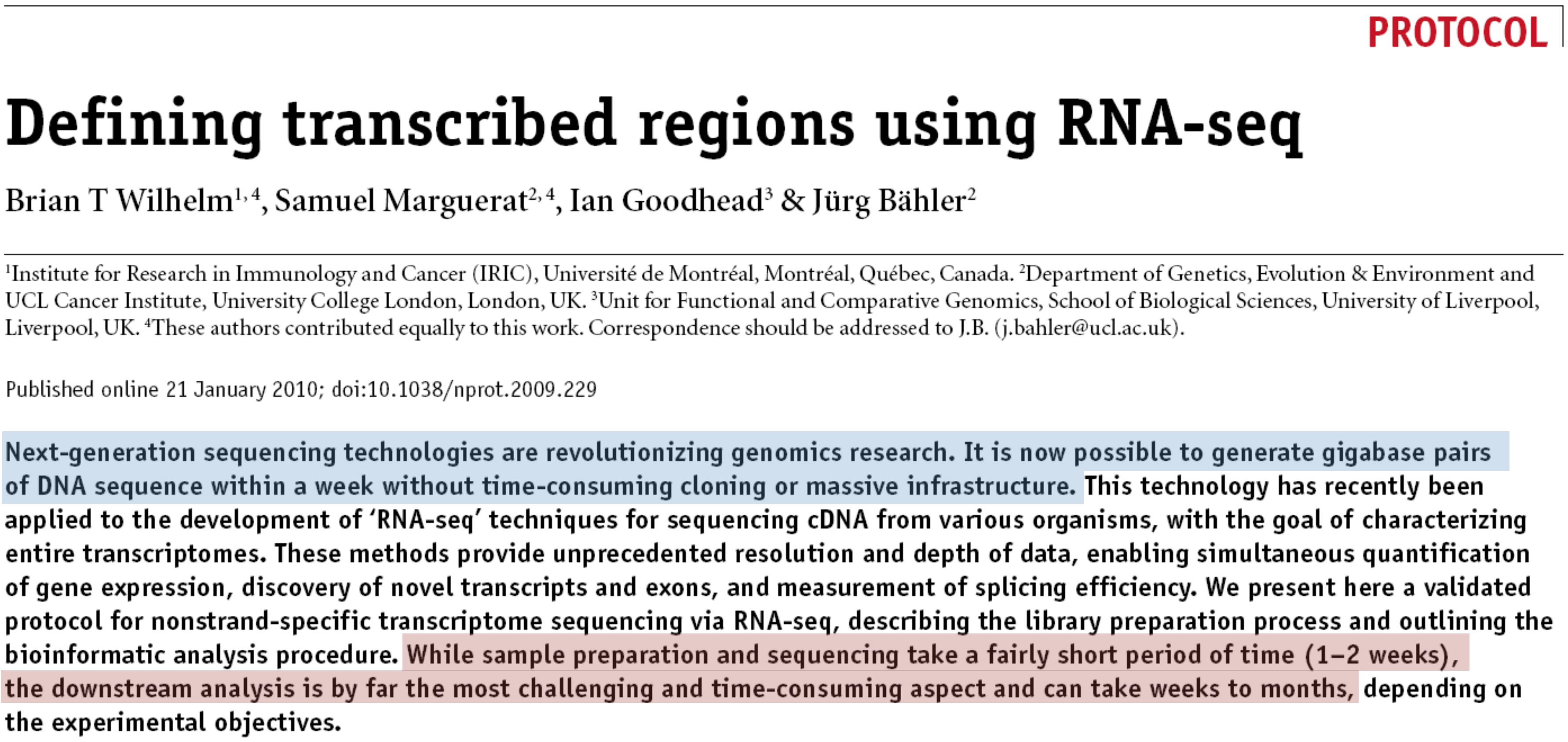
that typically have 5′ hydroxyl (Supplementary Fig. 3). Thus each observed cleavage site provides evidence that the cut nucleotide was in a double-stranded (for V1-treated samples) or single-stranded (for S1-treated samples) conformation. As a quantitative measure at nucleotide resolution representing the degree to which a nucleotide was in a double- or single-stranded conformation, we took the log ratio between the number of sequence reads obtained for each nucleotide in the V1 and S1 experiments. A higher (lower) log ratio, or PARS score, thus denotes a higher (lower) probability for a nucleotide to be in a double-stranded conformation.

We performed four independent V1 experiments and three independent S1 experiments, which were highly reproducible across replicates (correlation = 0.60–0.93, Supplementary Table 1), resulting in over 85 million sequence reads that map to the yeast genome, of which approximately 97% mapped to annotated transcripts (Supplementary Table 2). At an average nucleotide coverage above 1.0, we obtained structural information for over 3,000 yeast transcripts (Supplementary Table 3 and Supplementary Fig. 4a), covering in total over 4.2 million transcribed bases, which is approximately 100-fold more than all published RNA footprints to date.



We performed four independent V1 experiments and three independent S1 experiments, which were highly reproducible across replicates (correlation = 0.60–0.93, Supplementary Table 1), resulting in over 85 million sequence reads that map to the yeast genome, of which approximately 97% mapped to annotated transcripts (Supplementary Table 2). At an average nucleotide coverage above 1.0, we obtained structural information for over 3,000 yeast transcripts (Supplementary Table 3 and Supplementary Fig. 4a), covering in total over 4.2 million transcribed bases, which is approximately 100-fold more than all published RNA footprints to date.

Try to do this without (bio)informatics skills?



PROTOCOL

Defining transcribed regions using RNA-seq

Brian T Wilhelm[1,4], Samuel Marguerat[2,4], Ian Goodhead[3] & Jürg Bähler[2]

[1]Institute for Research in Immunology and Cancer (IRIC), Université de Montréal, Montréal, Québec, Canada. [2]Department of Genetics, Evolution & Environment and UCL Cancer Institute, University College London, London, UK. [3]Unit for Functional and Comparative Genomics, School of Biological Sciences, University of Liverpool, Liverpool, UK. [4]These authors contributed equally to this work. Correspondence should be addressed to J.B. (j.bahler@ucl.ac.uk).

Published online 21 January 2010; doi:10.1038/nprot.2009.229

Next-generation sequencing technologies are revolutionizing genomics research. It is now possible to generate gigabase pairs of DNA sequence within a week without time-consuming cloning or massive infrastructure. This technology has recently been applied to the development of 'RNA-seq' techniques for sequencing cDNA from various organisms, with the goal of characterizing entire transcriptomes. These methods provide unprecedented resolution and depth of data, enabling simultaneous quantification of gene expression, discovery of novel transcripts and exons, and measurement of splicing efficiency. We present here a validated protocol for nonstrand-specific transcriptome sequencing via RNA-seq, describing the library preparation process and outlining the bioinformatic analysis procedure. While sample preparation and sequencing take a fairly short period of time (1–2 weeks), the downstream analysis is by far the most challenging and time-consuming aspect and can take weeks to months, depending on the experimental objectives.

Lab: 1 week for one trained engineer?
Bioinformatics: Months of work!
This is the real research work?

Nat. Protoc. **5**, 256 (2010)

Vol 463 | 21 January 2010 | doi:10.1038/nature08696 nature

ARTICLES

The sequence and *de novo* assembly of the giant pan

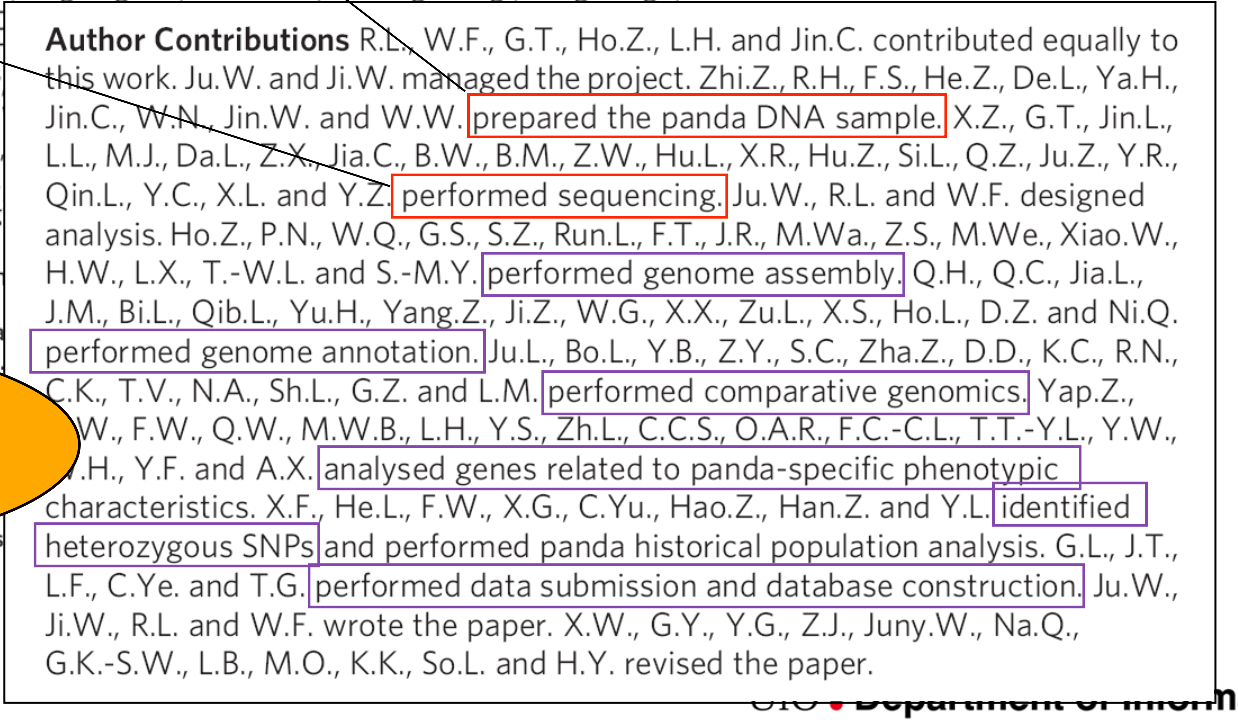
Ruiqiang Li[1,2]*, W Jing Cai[3,6]*, Quanfei Huang[1], Qingle Cai[1,7],
Bo Li[1], Yinqi Bai[1], Z Fuwen Wei[9], Heng Li[10], Min Jian[1], Jianwen Li[1],
Zhaolei Zhang[11], Rasm entao Yang[1], Zhaoling Xuan[1], Oliver A. Ryder[14],
Frederick Chi-Ching Leung[15], Yan Zhou , Xiao Sun[16], Yonggui Fu[17], Xiaodong Fang[1], Xiaosen Guo[1],
Bo Wang[1], Rong Hou[8], Fujun Shen[8], Bo Mu[1], Peixiang Ni[1], Runmao Lin[1], Wubin Qian[1], Guodong Wang[3,6], Chang Yu[1],
Wenhui Nie[6], Jinhuan Wang[6], Zhigang Wu[1], Huiqing Liang[1], Jiumeng Min[1,7], Qi Wu[9], Shifeng Cheng[1,7], Jue Ruan[1,3],
Mingw Wen[1], Binghang Liu[1], Xiaoli Ren[1], Huisong Zheng[1], Dong Dong[11],
Kathl g[1], C
Yingrui Li , Tommy T
Timing Gong[1], Hongde Liu[16], Dejin Zhang[16]
Yuanyuan Ren[1], Guojie Zhang[1,3,6], Michael
Yang Zheng[1,3], Yongyong Shi[5], Zhiqiang Li[5],
Feng Tian[1], Xiaoling Wang[1], Haiyin Wang[1],
Siu-Ming Yiu[22], Shiping Liu[23], Hemin Zhang
Junyi Wang[1], Nan Qin[1], Li Li[1], Jingxiang Li[1],
Maynard Olson[26], Xiuqing Zhang[1], Songgan

Using next-generation sequencing technology a giant pa (2 Ou using next-ge technologies genomes.

Travelled around in China and took blood samples from pandas

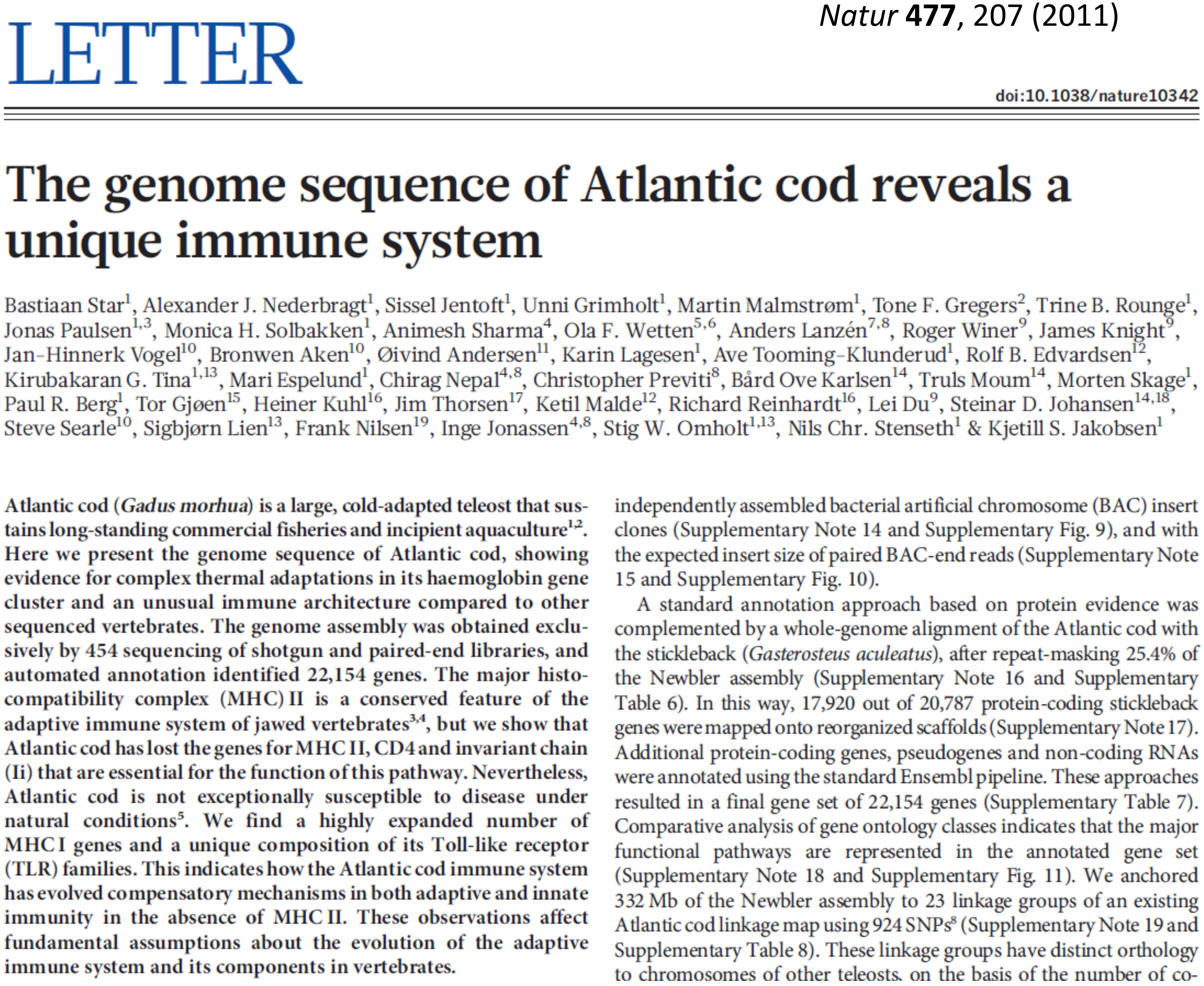
Wet lab?

Mostly bioinformatics, isn't it?



Author Contributions R.L., W.F., G.T., Ho.Z., L.H. and Jin.C. contributed equally to this work. Ju.W. and Ji.W. managed the project. Zhi.Z., R.H., F.S., He.Z., De.L., Ya.H., Jin.C., W.N., Jin.W. and W.W. prepared the panda DNA sample. X.Z., G.T., Jin.L., L.L., M.J., Da.L., Z.X., Jia.C., B.W., B.M., Z.W., Hu.L., X.R., Hu.Z., Si.L., Q.Z., Ju.Z., Y.R., Qin.L., Y.C., X.L. and Y.Z. performed sequencing. Ju.W., R.L. and W.F. designed analysis. Ho.Z., P.N., W.Q., G.S., S.Z., Run.L., F.T., J.R., M.Wa., Z.S., M.We., Xiao.W., H.W., L.X., T.-W.L. and S.-M.Y. performed genome assembly. Q.H., Q.C., Jia.L., J.M., Bi.L., Qib.L., Yu.H., Yang.Z., Ji.Z., W.G., X.X., Zu.L., X.S., Ho.L., D.Z. and Ni.Q. performed genome annotation. Ju.L., Bo.L., Y.B., Z.Y., S.C., Zha.Z., D.D., K.C., R.N., C.K., T.V., N.A., Sh.L., G.Z. and L.M. performed comparative genomics. Yap.Z., W., F.W., Q.W., M.W.B., L.H., Y.S., Zh.L., C.C.S., O.A.R., F.C.-C.L., T.T.-Y.L., Y.W., H., Y.F. and A.X. analysed genes related to panda-specific phenotypic characteristics. X.F., He.L., F.W., X.G., C.Yu., Hao.Z., Han.Z. and Y.L. identified heterozygous SNPs and performed panda historical population analysis. G.L., J.T., L.F., C.Ye. and T.G. performed data submission and database construction. Ju.W., Ji.W., R.L. and W.F. wrote the paper. X.W., G.Y., Y.G., Z.J., Juny.W., Na.Q., G.K.-S.W., L.B., M.O., K.K., So.L. and H.Y. revised the paper.

Natur **477**, 207 (2011)



LETTER

doi:10.1038/nature10342

The genome sequence of Atlantic cod reveals a unique immune system

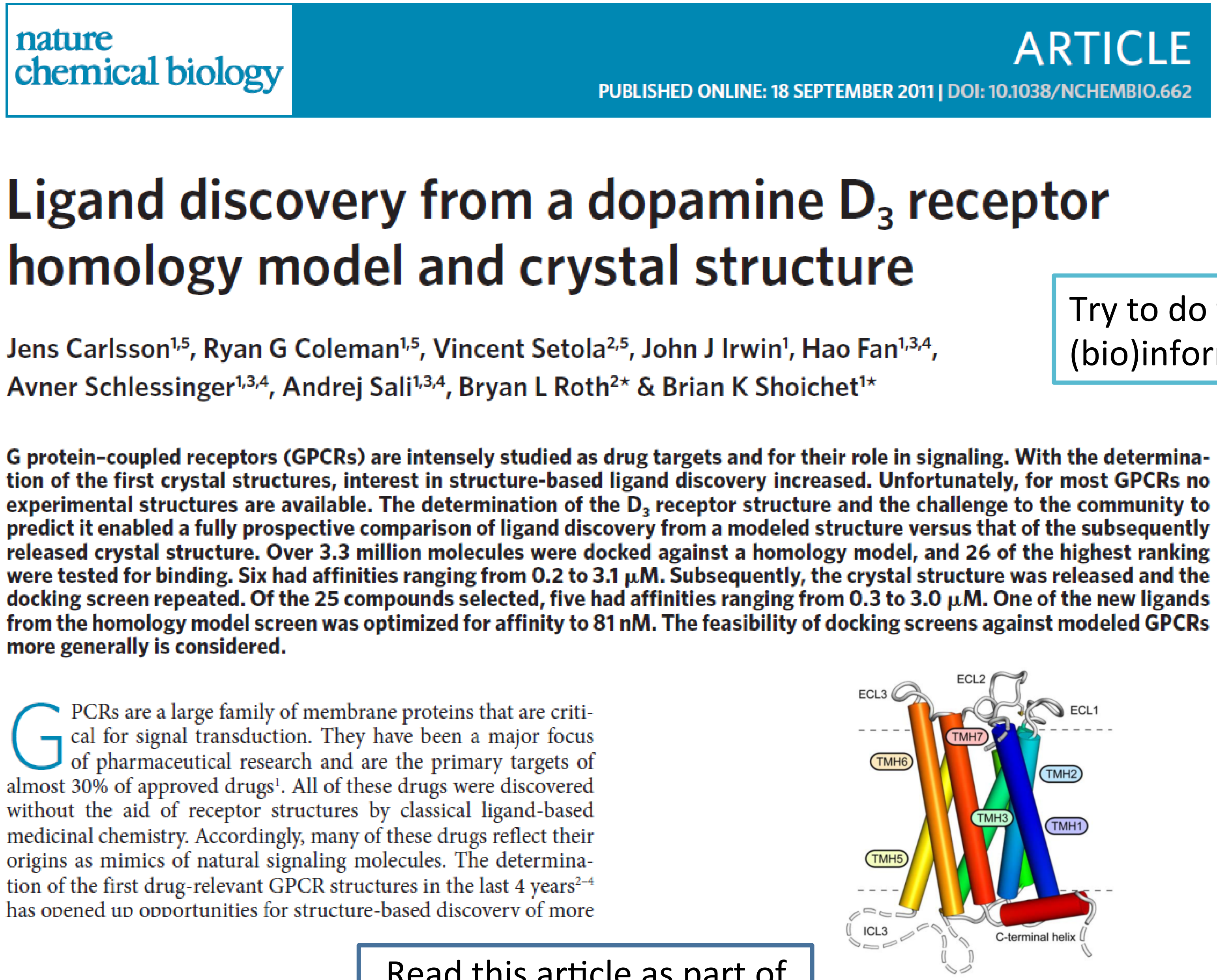
Bastiaan Star[1], Alexander J. Nederbragt[1], Sissel Jentoft[1], Unni Grimholt[1], Martin Malmstrøm[1], Tone F. Gregers[2], Trine B. Rounge[1], Jonas Paulsen[1,3], Monica H. Solbakken[1], Animesh Sharma[4], Ola F. Wetten[5,6], Anders Lanzén[7,8], Roger Winer[9], James Knight[9], Jan-Hinnerk Vogel[10], Bronwen Aken[10], Øivind Andersen[11], Karin Lagesen[1], Ave Tooming-Klunderud[1], Rolf B. Edvardsen[12], Kirubakaran G. Tina[1,13], Mari Espelund[1], Chirag Nepal[4,8], Christopher Previti[8], Bård Ove Karlsen[14], Truls Moum[14], Morten Skage[1], Paul R. Berg[1], Tor Gjøen[15], Heiner Kuhl[16], Jim Thorsen[17], Ketil Malde[12], Richard Reinhardt[16], Lei Du[9], Steinar D. Johansen[14,18], Steve Searle[10], Sigbjørn Lien[13], Frank Nilsen[19], Inge Jonassen[4,8], Stig W. Omholt[1,13], Nils Chr. Stenseth[1] & Kjetill S. Jakobsen[1]

Atlantic cod (*Gadus morhua*) is a large, cold-adapted teleost that sustains long-standing commercial fisheries and incipient aquaculture[1,2]. Here we present the genome sequence of Atlantic cod, showing evidence for complex thermal adaptations in its haemoglobin gene cluster and an unusual immune architecture compared to other sequenced vertebrates. The genome assembly was obtained exclusively by 454 sequencing of shotgun and paired-end libraries, and automated annotation identified 22,154 genes. The major histocompatibility complex (MHC) II is a conserved feature of the adaptive immune system of jawed vertebrates[3,4], but we show that Atlantic cod has lost the genes for MHC II, CD4 and invariant chain (Ii) that are essential for the function of this pathway. Nevertheless, Atlantic cod is not exceptionally susceptible to disease under natural conditions[5]. We find a highly expanded number of MHC I genes and a unique composition of its Toll-like receptor (TLR) families. This indicates how the Atlantic cod immune system has evolved compensatory mechanisms in both adaptive and innate immunity in the absence of MHC II. These observations affect fundamental assumptions about the evolution of the adaptive immune system and its components in vertebrates.

independently assembled bacterial artificial chromosome (BAC) insert clones (Supplementary Note 14 and Supplementary Fig. 9), and with the expected insert size of paired BAC-end reads (Supplementary Note 15 and Supplementary Fig. 10).

A standard annotation approach based on protein evidence was complemented by a whole-genome alignment of the Atlantic cod with the stickleback (*Gasterosteus aculeatus*), after repeat-masking 25.4% of the Newbler assembly (Supplementary Note 16 and Supplementary Table 6). In this way, 17,920 out of 20,787 protein-coding stickleback genes were mapped onto reorganized scaffolds (Supplementary Note 17). Additional protein-coding genes, pseudogenes and non-coding RNAs were annotated using the standard Ensembl pipeline. These approaches resulted in a final gene set of 22,154 genes (Supplementary Table 7). Comparative analysis of gene ontology classes indicates that the major functional pathways are represented in the annotated gene set (Supplementary Note 18 and Supplementary Fig. 11). We anchored 332 Mb of the Newbler assembly to 23 linkage groups of an existing Atlantic cod linkage map using 924 SNPs[8] (Supplementary Note 19 and Supplementary Table 8). These linkage groups have distinct orthology to chromosomes of other teleosts. on the basis of the number of co-

Jon K. Lærdahl,
Structural Bioinformatics



nature chemical biology

ARTICLE

PUBLISHED ONLINE: 18 SEPTEMBER 2011 | DOI: 10.1038/NCHEMBIO.662

Ligand discovery from a dopamine D_3 receptor homology model and crystal structure

Jens Carlsson[1,5], Ryan G Coleman[1,5], Vincent Setola[2,5], John J Irwin[1], Hao Fan[1,3,4], Avner Schlessinger[1,3,4], Andrej Sali[1,3,4], Bryan L Roth[2*] & Brian K Shoichet[1*]

G protein–coupled receptors (GPCRs) are intensely studied as drug targets and for their role in signaling. With the determination of the first crystal structures, interest in structure-based ligand discovery increased. Unfortunately, for most GPCRs no experimental structures are available. The determination of the D_3 receptor structure and the challenge to the community to predict it enabled a fully prospective comparison of ligand discovery from a modeled structure versus that of the subsequently released crystal structure. Over 3.3 million molecules were docked against a homology model, and 26 of the highest ranking were tested for binding. Six had affinities ranging from 0.2 to 3.1 μM. Subsequently, the crystal structure was released and the docking screen repeated. Of the 25 compounds selected, five had affinities ranging from 0.3 to 3.0 μM. One of the new ligands from the homology model screen was optimized for affinity to 81 nM. The feasibility of docking screens against modeled GPCRs more generally is considered.

GPCRs are a large family of membrane proteins that are critical for signal transduction. They have been a major focus of pharmaceutical research and are the primary targets of almost 30% of approved drugs[1]. All of these drugs were discovered without the aid of receptor structures by classical ligand-based medicinal chemistry. Accordingly, many of these drugs reflect their origins as mimics of natural signaling molecules. The determination of the first drug-relevant GPCR structures in the last 4 years[2–4] has opened up opportunities for structure-based discovery of more

Try to do this without (bio)informatics skills?

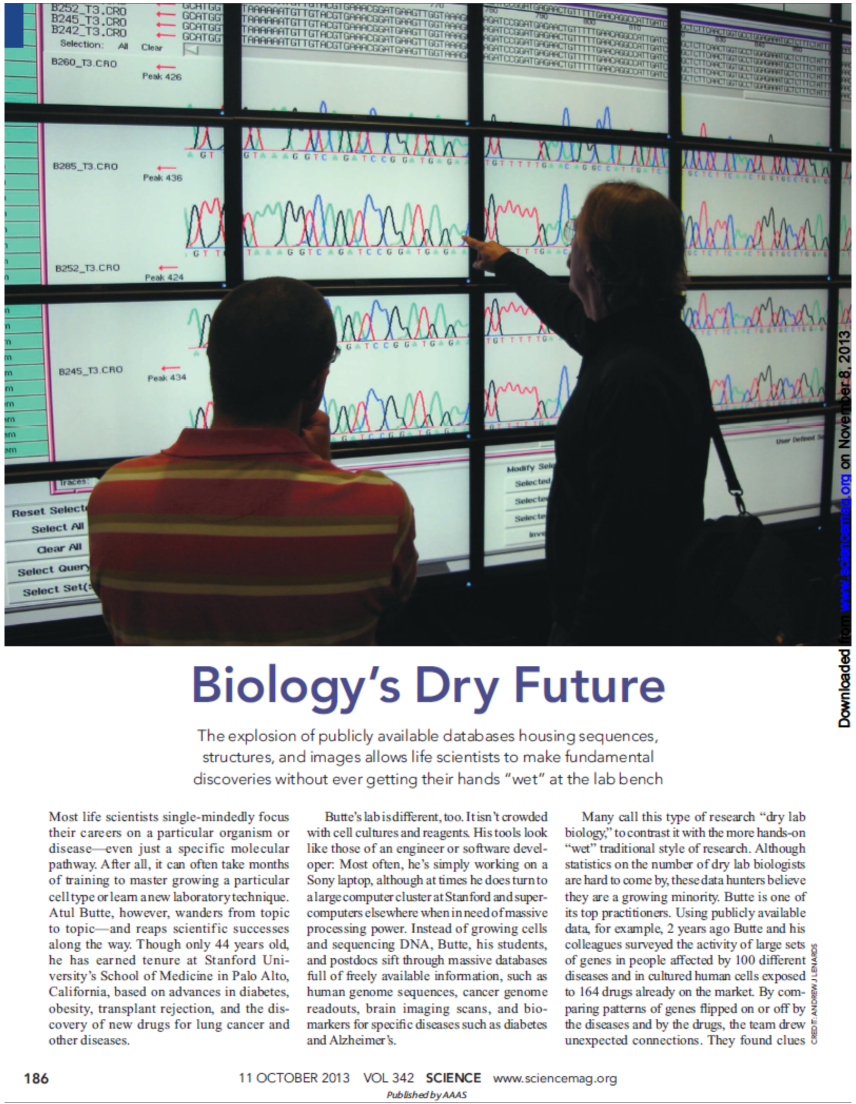
Read this article as part of the curriculum!

Oslo universitetssykehus

UiO : Department of Informatics
University of Oslo

Jon K. Lærdahl,
Structural Bioinformatics

No wet lab biology?


B252_T3.CRO
B245_T3.CRO
B242_T3.CRO
Selection: All Clear
B260_T3.CRO
Peak 426
B285_T3.CRO
Peak 436
B252_T3.CRO
Peak 424
B245_T3.CRO
Peak 434
Reset Select
Select All
Clear All
Select Query
Select Set
Modify Sel
Selected
User Defined

Biology's Dry Future

The explosion of publicly available databases housing sequences, structures, and images allows life scientists to make fundamental discoveries without ever getting their hands "wet" at the lab bench

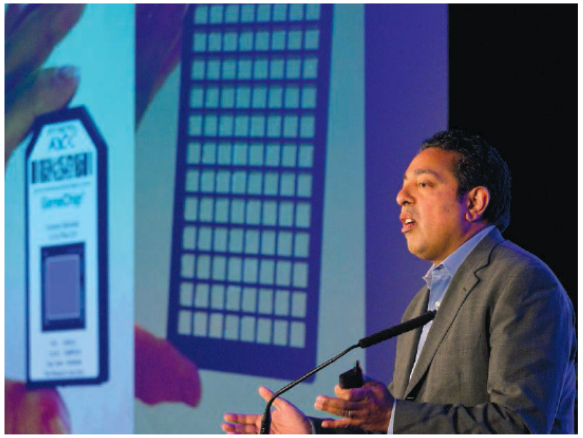
Most life scientists single-mindedly focus their careers on a particular organism or disease—even just a specific molecular pathway. After all, it can often take months of training to master growing a particular cell type or learn a new laboratory technique. Atul Butte, however, wanders from topic to topic—and reaps scientific successes along the way. Though only 44 years old, he has earned tenure at Stanford University's School of Medicine in Palo Alto, California, based on advances in diabetes, obesity, transplant rejection, and the discovery of new drugs for lung cancer and other diseases.

Butte's lab is different, too. It isn't crowded with cell cultures and reagents. His tools look like those of an engineer or software developer: Most often, he's simply working on a Sony laptop, although at times he does turn to a large computer cluster at Stanford and supercomputers elsewhere when in need of massive processing power. Instead of growing cells and sequencing DNA, Butte, his students, and postdocs sift through massive databases full of freely available information, such as human genome sequences, cancer genome readouts, brain imaging scans, and biomarkers for specific diseases such as diabetes and Alzheimer's.

Many call this type of research "dry lab biology," to contrast it with the more hands-on "wet" traditional style of research. Although statistics on the number of dry lab biologists are hard to come by, these data hunters believe they are a growing minority. Butte is one of its top practitioners. Using publicly available data, for example, 2 years ago Butte and his colleagues surveyed the activity of large sets of genes in people affected by 100 different diseases and in cultured human cells exposed to 164 drugs already on the market. By comparing patterns of genes flipped on or off by the diseases and by the drugs, the team drew unexpected connections. They found clues

CREDIT: ANDREW J LENARDS

Published by AAAS



"I'm like a **kid in a candy store.** There is so much we can do."

—Atul Butte, Stanford University School of Medicine

Science, **342**, 186 (2013)

Exome sequencing detects disease-causing SNVs and CNVs in Primary Immunodeficiencies

Hanne Sørmo Sorte, PhD student

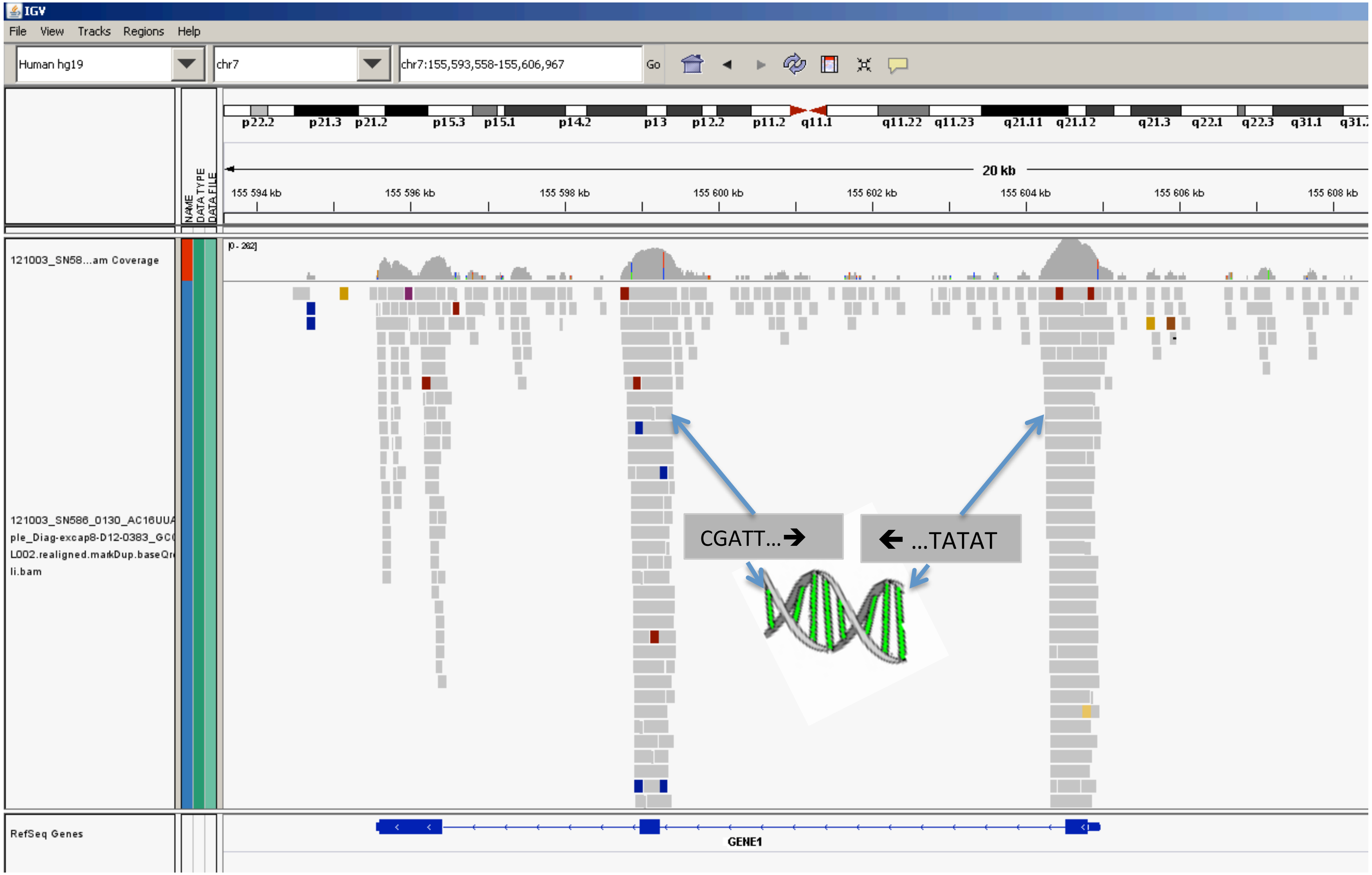
Department of Medical Genetics

Oslo University Hospital and University of Oslo

Oslo, Norway

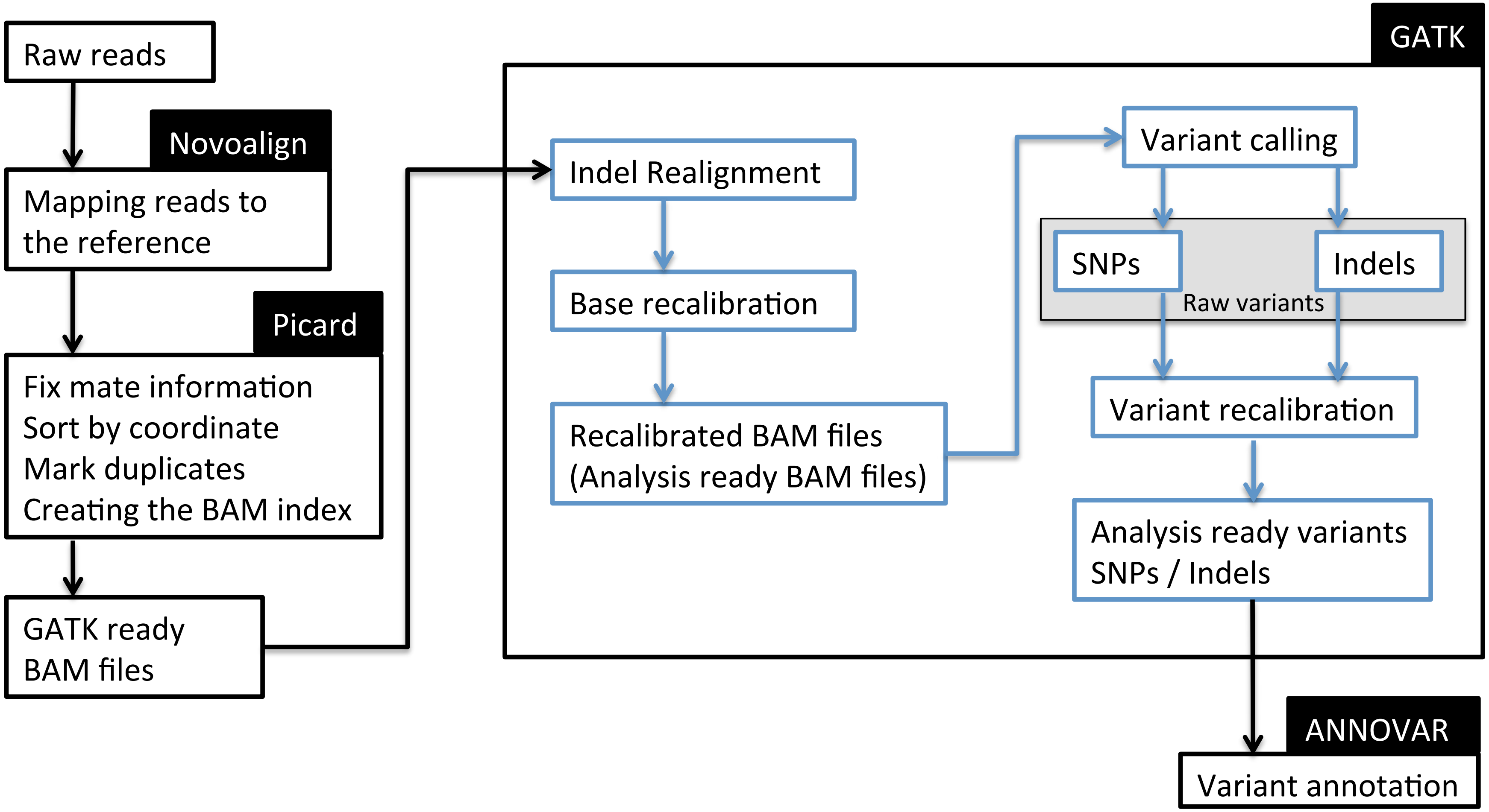
Oslo University Hospital

UiO : University of Oslo


IGV
File View Tracks Regions Help
Human hg19
chr7
chr7:155,593,558-155,606,967
Go
p22.2 p21.3 p21.2 p15.3 p15.1 p14.2 p13 p12.2 p11.2 q11.1 q11.22 q11.23 q21.11 q21.12 q21.3 q22.1 q22.3 q31.1 q31.
20 kb
155 594 kb
155 596 kb
155 598 kb
155 600 kb
155 602 kb
155 604 kb
155 606 kb
155 608 kb
NAME
DATA TYPE
DATA FILE
121003_SN58...am Coverage
[0 - 282]
121003_SN586_0130_AC16UUA
ple_Diag-excap8-D12-0383_GC
L002.realigned.markDup.baseQ
li.bam
CGATT... →
← ...TATAT
RefSeq Genes
GENE1

- Mapping to the reference genome
- IGV: Visualization tool – chromosome w/ tracks ex RefSeq genes
- Exonic/intronic/intergenic regions

Bioinformatics



Raw reads

Novoalign

Mapping reads to the reference

Picard

Fix mate information
Sort by coordinate
Mark duplicates
Creating the BAM index

GATK ready BAM files

GATK

Indel Realignment

Base recalibration

Recalibrated BAM files (Analysis ready BAM files)

Variant calling

SNPs

Indels

Raw variants

Variant recalibration

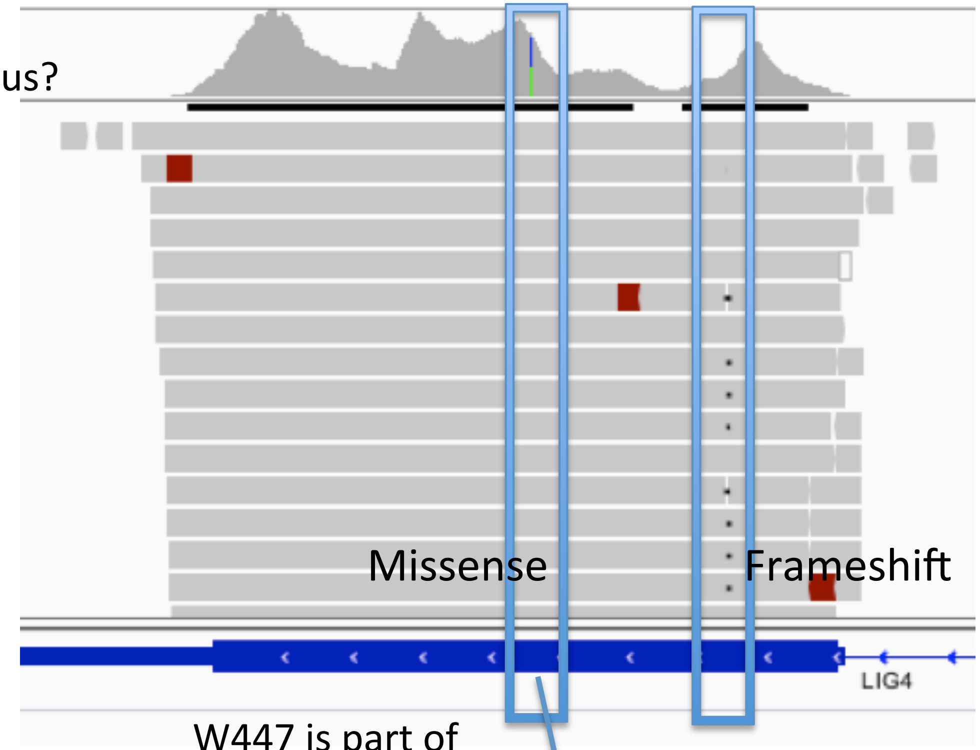
Analysis ready variants SNPs / Indels

ANNOVAR

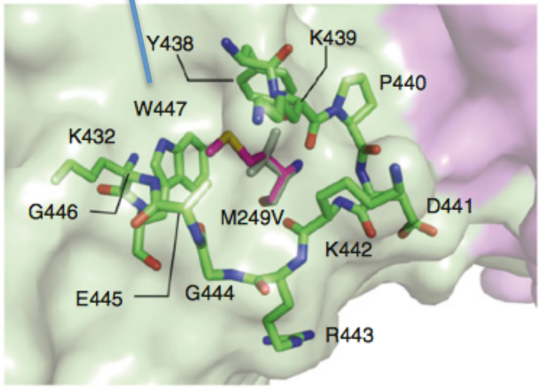
Variant annotation

Solved case - example of clinical utility

- 4 year old boy, healthy until 2 ½
- Feb -13: Anemia and thrombocytopenia – virus? normal lymphocytes/IgG
- Apr-13:
  - respiratory distress ->
  - OUS: pneumocystis pneumonia
  - $T^{low}B^{-}NK^{+}$ + low IgG – treated
  - Chronic Rota virus + parvoB19
    – not able to cure
- Fall -13: falling T-cell - develop full but untypical immunodeficiency ->
  - Specific genetic tests negative
  - HSCT transplantation?
  - Pretreatment conditioning
- Exome sequencing: LIG4 (DNA ligase IV)
  - impaired DNA ds break rejoining
  - Few reported; different presentation



W447 is part of the catalytic pocket



Confirmed by radiosensitivity assay -> HSCT w/correct preconditioning -> Now completely healthy!!

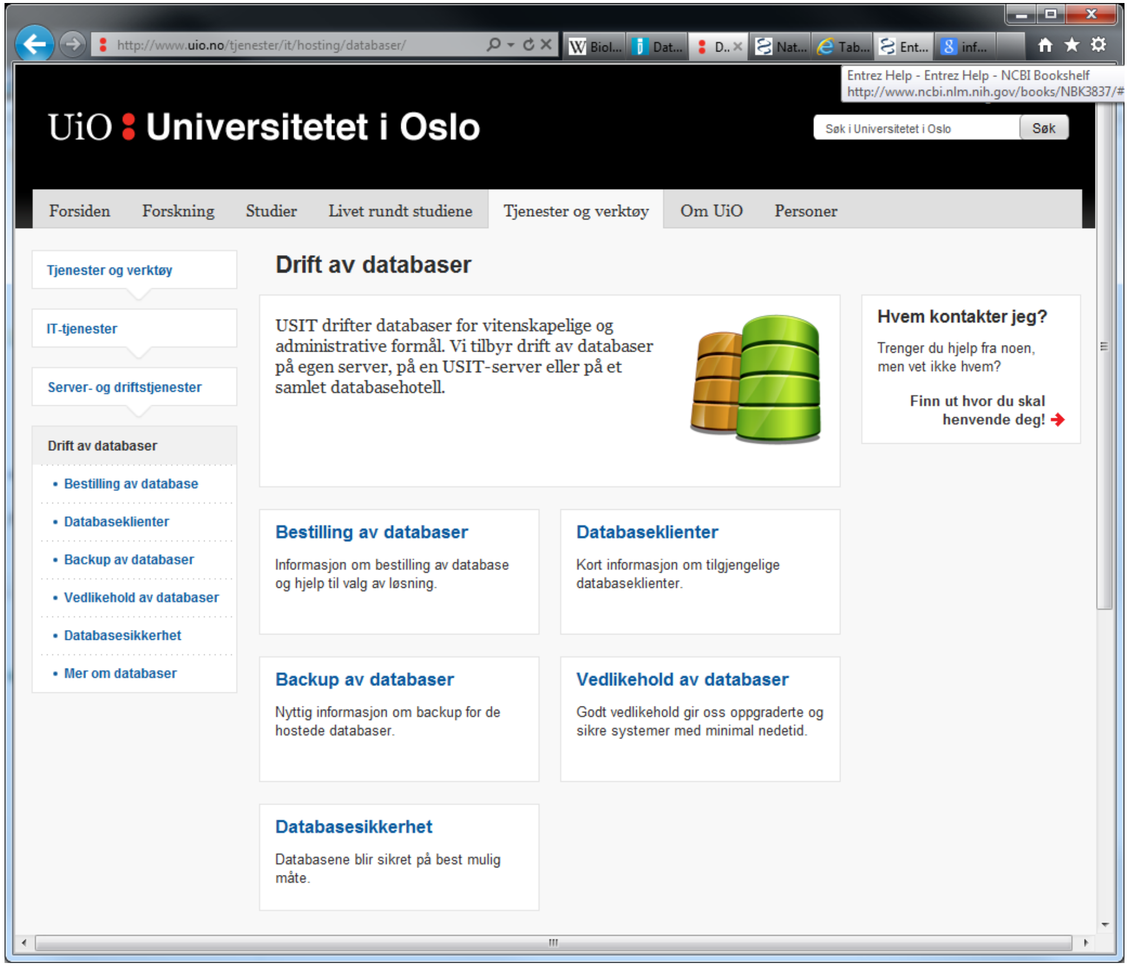
Database

- Organized collection of data/information, in computer-readable form
- Defining characteristics
  - the contents
  - the ontology (list of valid terms and their definition, vocabulary)
  - logical structure (interrelationship among the data)
  - data format
  - routes for data retrieval, data presentation or analysis
  - links to other databases, references to original publication data etc.

A.M. Lesk, *Introduction to Bioinformatics*

Making your own database?

Talk to an informatician! (or USIT at UiO)



Drift av databaser

USIT drifter databaser for vitenskapelige og administrative formål. Vi tilbyr drift av databaser på egen server, på en USIT-server eller på et samlet databasehotell.

Hvem kontakter jeg?

Trenger du hjelp fra noen, men vet ikke hvem?

Finn ut hvor du skal henvende deg! →

Bestilling av databaser

Informasjon om bestilling av database og hjelp til valg av løsning.

Databaseklienter

Kort informasjon om tilgjengelige databaseklienter.

Backup av databaser

Nyttig informasjon om backup for de hostede databaser.

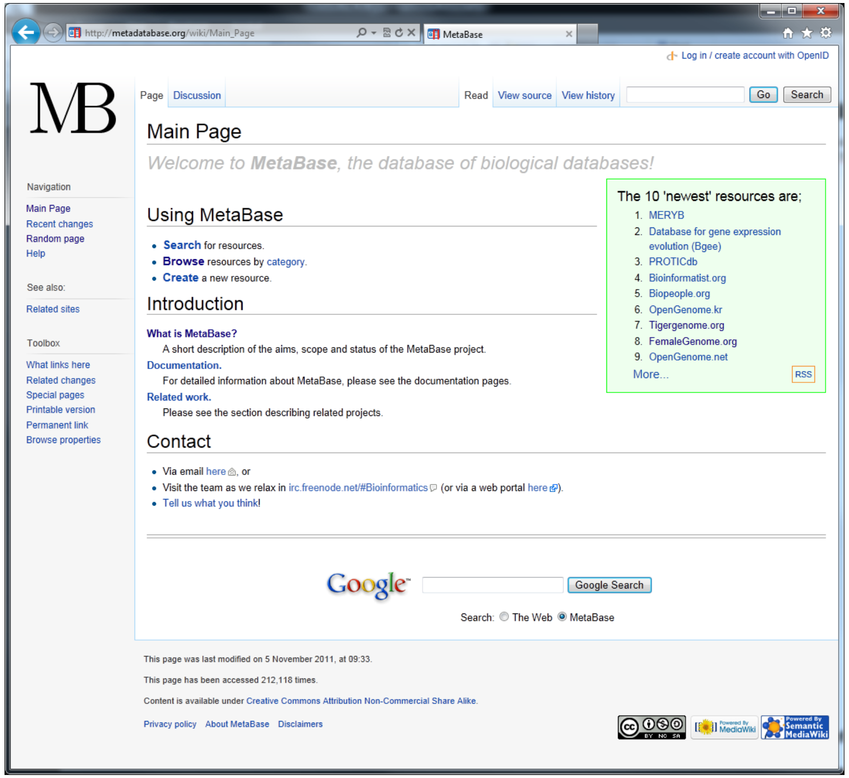
Vedlikehold av databaser

Godt vedlikehold gir oss oppgraderte og sikre systemer med minimal nedetid.

Databasesikkerhet

Databasene blir sikret på best mulig måte.

A lot of biological databases already available...



http://metadatabase.org/wiki/Main_Page
MetaBase
Log in / create account with OpenID
MB
Page
Discussion
Read
View source
View history
Go
Search

Main Page

*Welcome to **MetaBase**, the database of biological databases!*

Navigation

- Main Page
- Recent changes
- Random page
- Help

See also:

- Related sites

Toolbox

- What links here
- Related changes
- Special pages
- Printable version
- Permanent link
- Browse properties

Using MetaBase

- **Search** for resources.
- **Browse** resources by category.
- **Create** a new resource.

Introduction

What is MetaBase?
A short description of the aims, scope and status of the MetaBase project.

Documentation.
For detailed information about MetaBase, please see the documentation pages.

Related work.
Please see the section describing related projects.

The 10 'newest' resources are;

1. MERYB
2. Database for gene expression evolution (Bgee)
3. PROTICdb
4. Bioinformatist.org
5. Biopeople.org
6. OpenGenome.kr
7. Tigergenome.org
8. FemaleGenome.org
9. OpenGenome.net

More...

RSS

Contact

- Via email here, or
- Visit the team as we relax in irc.freenode.net/#Bioinformatics (or via a web portal here).
- Tell us what you think!

Google [ ] Google Search

Search: ○ The Web ● MetaBase

This page was last modified on 5 November 2011, at 09:33.

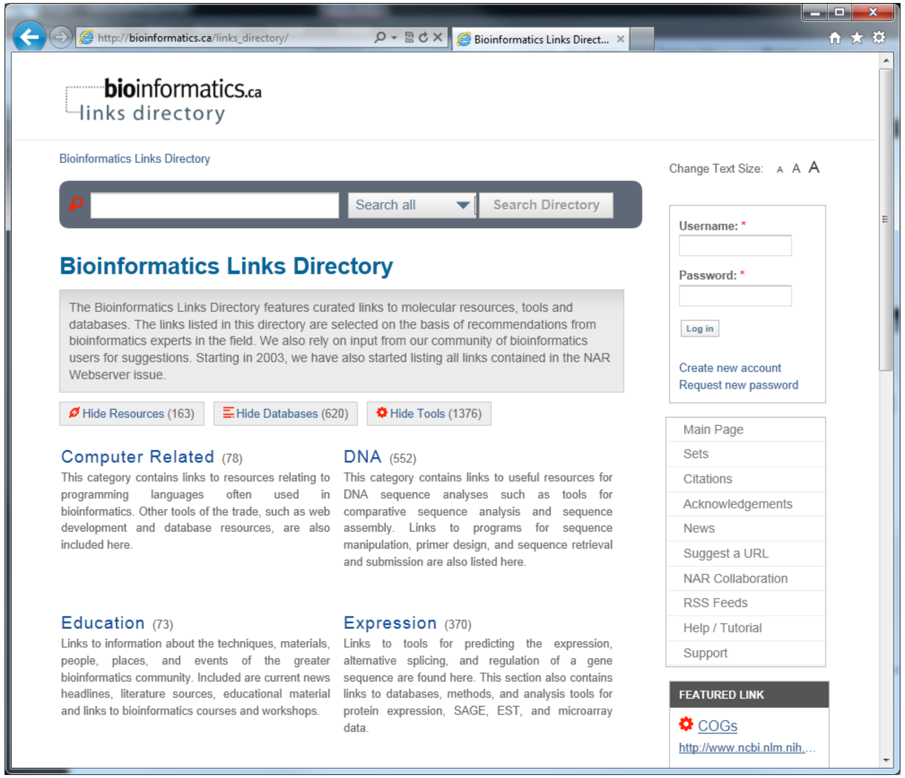
This page has been accessed 212,118 times.

Content is available under Creative Commons Attribution Non-Commercial Share Alike.

Privacy policy About MetaBase Disclaimers

MetaBase, the database of biological databases (1800 entries)

- http://metadatabase.org



http://bioinformatics.ca/links_directory/
Bioinformatics Links Direct...
bioinformatics.ca
links directory

Bioinformatics Links Directory

Search all ▼ Search Directory

Change Text Size: A A A

Bioinformatics Links Directory

The Bioinformatics Links Directory features curated links to molecular resources, tools and databases. The links listed in this directory are selected on the basis of recommendations from bioinformatics experts in the field. We also rely on input from our community of bioinformatics users for suggestions. Starting in 2003, we have also started listing all links contained in the NAR Webserver issue.

Hide Resources (163) Hide Databases (620) Hide Tools (1376)

Computer Related (78)
This category contains links to resources relating to programming languages often used in bioinformatics. Other tools of the trade, such as web development and database resources, are also included here.

DNA (552)
This category contains links to useful resources for DNA sequence analyses such as tools for comparative sequence analysis and sequence assembly. Links to programs for sequence manipulation, primer design, and sequence retrieval and submission are also listed here.

Education (73)
Links to information about the techniques, materials, people, places, and events of the greater bioinformatics community. Included are current news headlines, literature sources, educational material and links to bioinformatics courses and workshops.

Expression (370)
Links to tools for predicting the expression, alternative splicing, and regulation of a gene sequence are found here. This section also contains links to databases, methods, and analysis tools for protein expression, SAGE, EST, and microarray data.

Username: *

Password: *

Log in

Create new account
Request new password

- Main Page
- Sets
- Citations
- Acknowledgements
- News
- Suggest a URL
- NAR Collaboration
- RSS Feeds
- Help / Tutorial
- Support

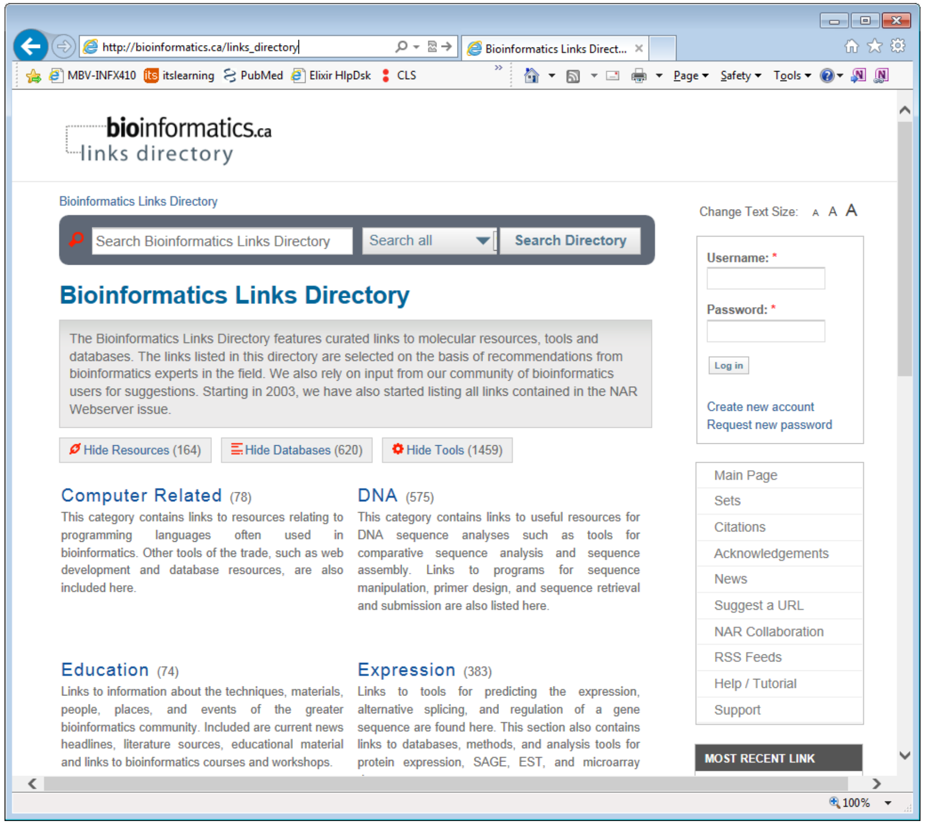
FEATURED LINK

COGs
http://www.ncbi.nlm.nih....

bioinformatics.ca – links directory (623 databases)

- http://bioinformatics.ca/links_directory

btw, the **bio**informatics.ca links directory is an excellent resource



Bioinformatics Links Directory

The Bioinformatics Links Directory features curated links to molecular resources, tools and databases. The links listed in this directory are selected on the basis of recommendations from bioinformatics experts in the field. We also rely on input from our community of bioinformatics users for suggestions. Starting in 2003, we have also started listing all links contained in the NAR Webserver issue.

Computer Related (78)

This category contains links to resources relating to programming languages often used in bioinformatics. Other tools of the trade, such as web development and database resources, are also included here.

DNA (575)

This category contains links to useful resources for DNA sequence analyses such as tools for comparative sequence analysis and sequence assembly. Links to programs for sequence manipulation, primer design, and sequence retrieval and submission are also listed here.

Education (74)

Links to information about the techniques, materials, people, places, and events of the greater bioinformatics community. Included are current news headlines, literature sources, educational material and links to bioinformatics courses and workshops.

Expression (383)

Links to tools for predicting the expression, alternative splicing, and regulation of a gene sequence are found here. This section also contains links to databases, methods, and analysis tools for protein expression, SAGE, EST, and microarray

bioinformatics.ca – links directory

- http://bioinformatics.ca/links_directory
- Currently
  - 1549 tools
  - 623 databases
  - 174 “resources”
- The problem is not to find a tool or database, but to know what is “gold” and what is “junk”

Some important centres for bioinformatics

- National Center for Biotechnology Information (NCBI)
  - part of the US National Library of Medicine (NLM), a branch of the National Institutes of Health
  - located in Bethesda, Maryland
- European Bioinformatics Institute (EMBL-EBI)
  - part of part of European Molecular Biology Laboratory (EMBL)
  - located in Hinxton, Cambridgeshire, UK

NCBI databases

- Provided the GenBank DNA sequence database since 1992
- Online Mendelian Inheritance in Man (OMIM) - known diseases with a genetic component and links to genes
  - started early 1960s as a book
  - online version, OMIM, since 1987
  - on the WWW by NCBI in 1995
  - currently >22,000 entries (14,400 genes)
- EST - nucleotide database subset that contains only Expressed Sequence Tag records
- Gene - genes and associated information for a number of organisms in addition to and including human
- Protein sequence database - collection of protein sequence entries compiled from a variety of sources including Swiss-Prot, PIR, PRF, PDB, and translations from annotated coding regions in GenBank and RefSeq
- PubMed - access to over 15 million citations from MEDLINE and additional life sciences journals
- SNP - repository for both single nucleotide substitutions and short deletion and insertion polymorphisms

All data is publicly available

NCBI databases

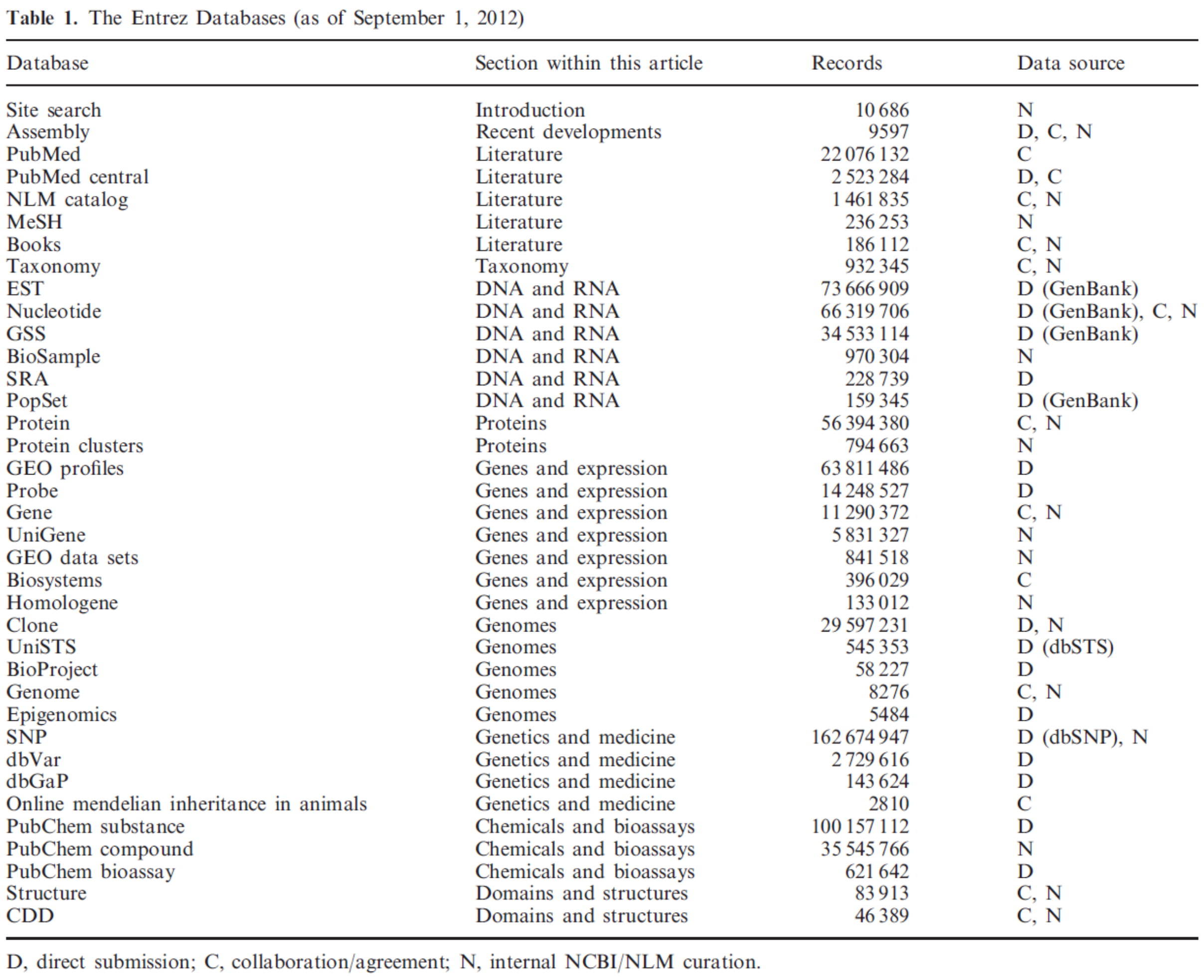


Table 1. The Entrez Databases (as of September 1, 2012)

Database	Section within this article	Records	Data source
Site search	Introduction	10 686	N
Assembly	Recent developments	9597	D, C, N
PubMed	Literature	22 076 132	C
PubMed central	Literature	2 523 284	D, C
NLM catalog	Literature	1 461 835	C, N
MeSH	Literature	236 253	N
Books	Literature	186 112	C, N
Taxonomy	Taxonomy	932 345	C, N
EST	DNA and RNA	73 666 909	D (GenBank)
Nucleotide	DNA and RNA	66 319 706	D (GenBank), C, N
GSS	DNA and RNA	34 533 114	D (GenBank)
BioSample	DNA and RNA	970 304	N
SRA	DNA and RNA	228 739	D
PopSet	DNA and RNA	159 345	D (GenBank)
Protein	Proteins	56 394 380	C, N
Protein clusters	Proteins	794 663	N
GEO profiles	Genes and expression	63 811 486	D
Probe	Genes and expression	14 248 527	D
Gene	Genes and expression	11 290 372	C, N
UniGene	Genes and expression	5 831 327	N
GEO data sets	Genes and expression	841 518	N
Biosystems	Genes and expression	396 029	C
Homologene	Genes and expression	133 012	N
Clone	Genomes	29 597 231	D, N
UniSTS	Genomes	545 353	D (dbSTS)
BioProject	Genomes	58 227	D
Genome	Genomes	8276	C, N
Epigenomics	Genomes	5484	D
SNP	Genetics and medicine	162 674 947	D (dbSNP), N
dbVar	Genetics and medicine	2 729 616	D
dbGaP	Genetics and medicine	143 624	D
Online mendelian inheritance in animals	Genetics and medicine	2810	C
PubChem substance	Chemicals and bioassays	100 157 112	D
PubChem compound	Chemicals and bioassays	35 545 766	N
PubChem bioassay	Chemicals and bioassays	621 642	D
Structure	Domains and structures	83 913	C, N
CDD	Domains and structures	46 389	C, N

D, direct submission; C, collaboration/agreement; N, internal NCBI/NLM curation.

37 databases that together contains over 690 million records

Nucleic Acids Res. **41**, D8 (2013)

EMBL-EBI databases

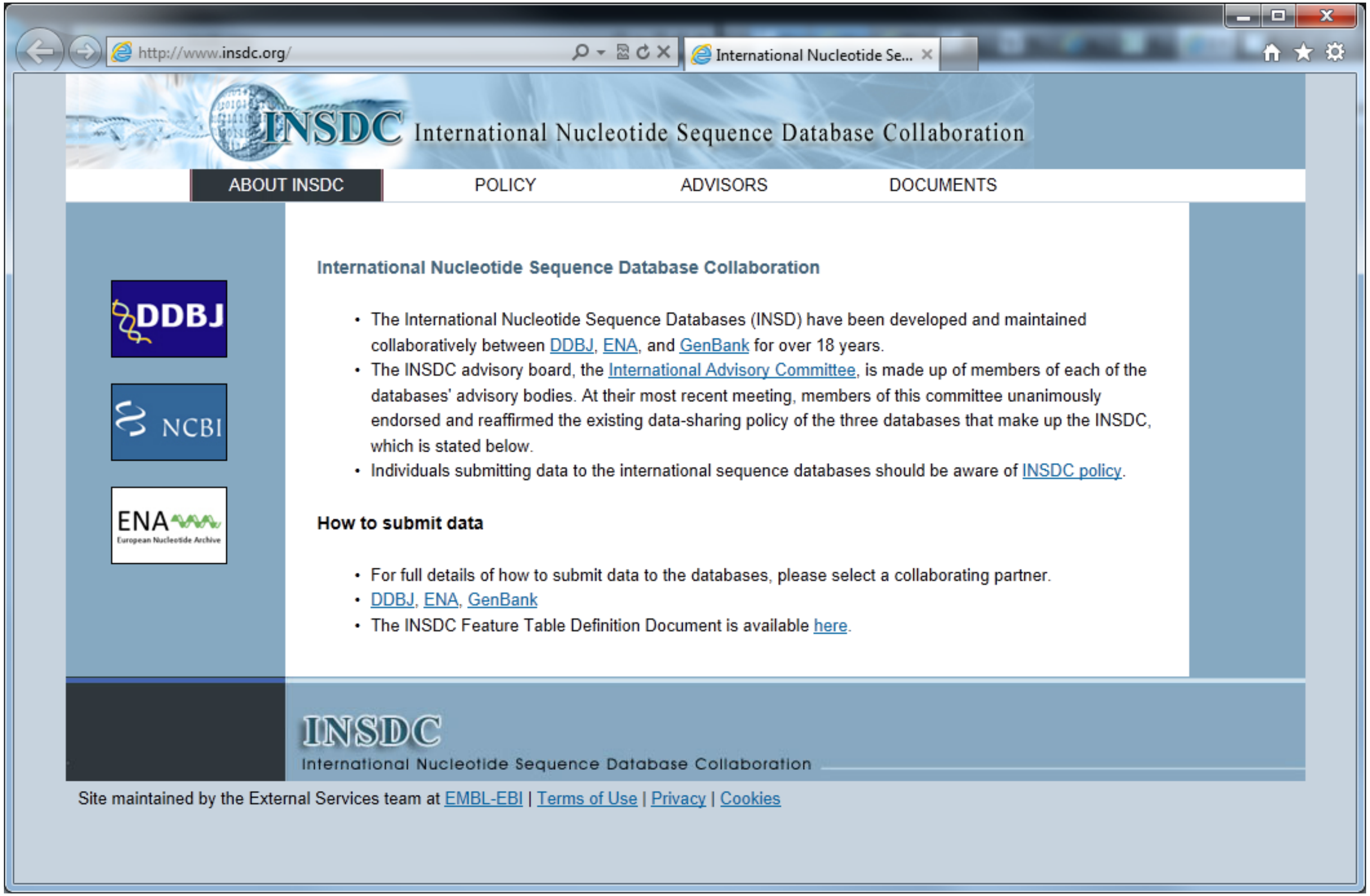
- European Nucleotide Archive (ENA) nucleotide sequence database
- Ensembl - automatic and manually curated annotation on selected eukaryotic (vertebrate) genomes
- Ensembl Genomes – Ensembl for "all other organisms"
- UniProt – protein sequence and functional information
- ChEMBL – database of bioactive compounds
- IntAct - repository of molecular interactions, including protein-protein, protein-small molecule and protein-nucleic acid interactions
- CiteXplore – 25 million literature abstracts including PubMed, Agricola & patents
- Gene Ontology (GO) - controlled vocabulary to describe gene and gene product attributes in any organism
- Gene Ontology Annotation (GOA) – GO annotations for proteins in UniProt

All data is publicly available

GenBank

- a comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation
- *all* publicly available DNA sequences
- submissions from authors
  - web-based BankIt
  - standalone program Sequin
- submissions from EST and other high-throughput sequencing projects
- daily exchange of data with ENA and DNA Data Bank of Japan (DDBJ)
  - all sequences submitted to DDBJ, ENA, or GenBank will end up in all 3 databases within few days

INSDC


http://www.insdc.org/
International Nucleotide Se...
INSDC International Nucleotide Sequence Database Collaboration
ABOUT INSDC
POLICY
ADVISORS
DOCUMENTS
DDBJ
NCBI
ENA European Nucleotide Archive

International Nucleotide Sequence Database Collaboration

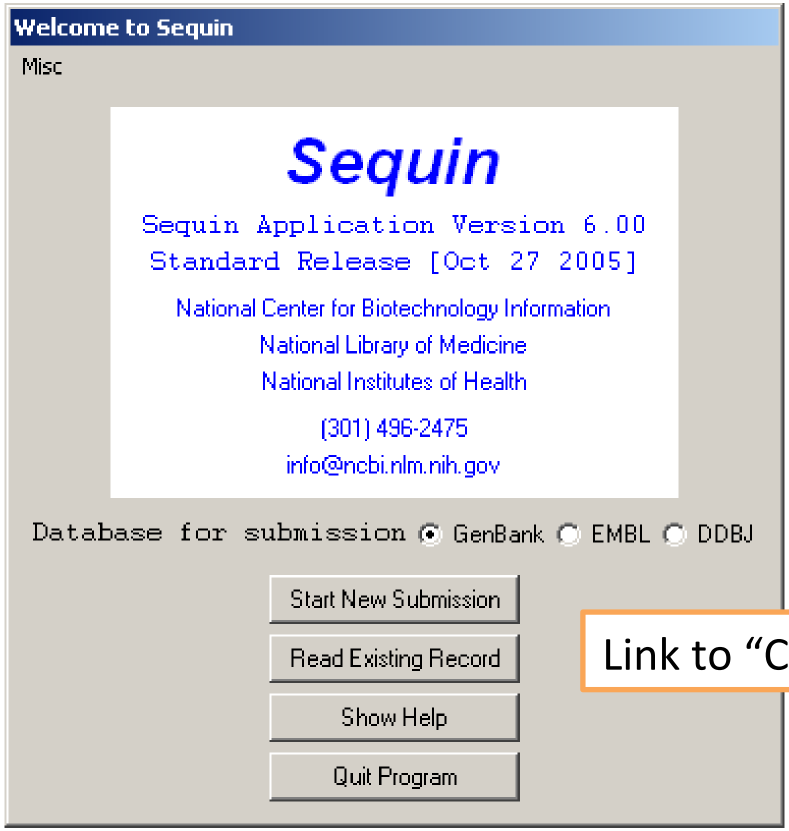
- The International Nucleotide Sequence Databases (INSD) have been developed and maintained collaboratively between DDBJ, ENA, and GenBank for over 18 years.
- The INSDC advisory board, the International Advisory Committee, is made up of members of each of the databases' advisory bodies. At their most recent meeting, members of this committee unanimously endorsed and reaffirmed the existing data-sharing policy of the three databases that make up the INSDC, which is stated below.
- Individuals submitting data to the international sequence databases should be aware of INSDC policy.

How to submit data

- For full details of how to submit data to the databases, please select a collaborating partner.
- DDBJ, ENA, GenBank
- The INSDC Feature Table Definition Document is available here.

Site maintained by the External Services team at EMBL-EBI | Terms of Use | Privacy | Cookies

Sequin – for submitting to GenBank



Welcome to Sequin

Misc

Sequin

Sequin Application Version 6.00
Standard Release [Oct 27 2005]

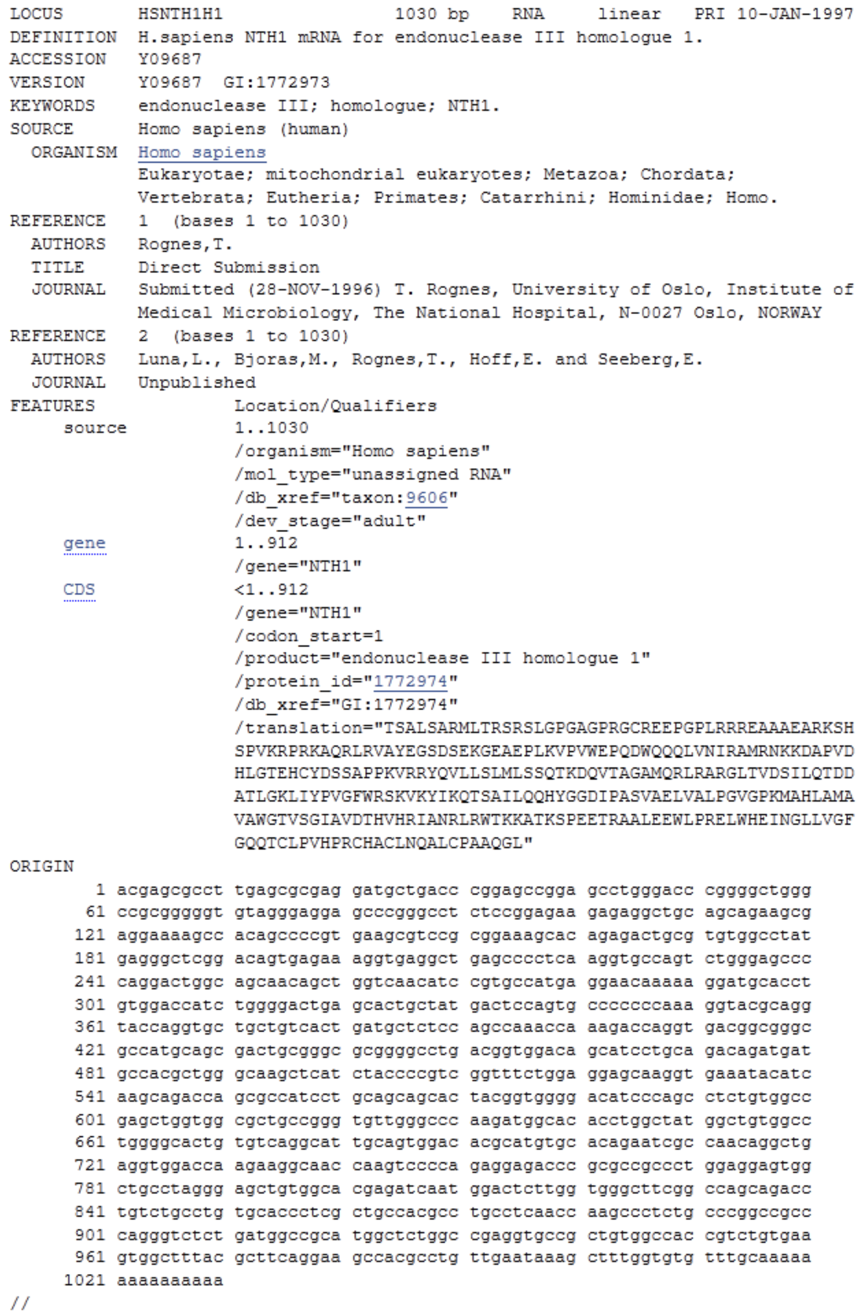
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health

(301) 496-2475
info@ncbi.nlm.nih.gov

Database for submission ● GenBank ○ EMBL ○ DDBJ

Link to "Create a submission"

BankIt is web-based alternative



```
LOCUS       HSNTH1H1                1030 bp    RNA     linear   PRI 10-JAN-1997
DEFINITION  H.sapiens NTH1 mRNA for endonuclease III homologue 1.
ACCESSION   Y09687
VERSION     Y09687  GI:1772973
KEYWORDS    endonuclease III; homologue; NTH1.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
            Vertebrata; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1030)
  AUTHORS   Rognes,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (28-NOV-1996) T. Rognes, University of Oslo, Institute of
            Medical Microbiology, The National Hospital, N-0027 Oslo, NORWAY
REFERENCE   2  (bases 1 to 1030)
  AUTHORS   Luna,L., Bjoras,M., Rognes,T., Hoff,E. and Seeberg,E.
  JOURNAL   Unpublished
FEATURES             Location/Qualifiers
     source          1..1030
                     /organism="Homo sapiens"
                     /mol_type="unassigned RNA"
                     /db_xref="taxon:9606"
                     /dev_stage="adult"
     gene            1..912
                     /gene="NTH1"
     CDS             <1..912
                     /gene="NTH1"
                     /codon_start=1
                     /product="endonuclease III homologue 1"
                     /protein_id="1772974"
                     /db_xref="GI:1772974"
                     /translation="TSALSARMLTRSRSLGPGAGPRGCREEPGPLRRREAAAEARKSH
                     SPVKRPRKAQRLRVAYEGSDSEKGEAEPLKVPVWEPQDWQQQLVNIRAMRNKKDAPVD
                     HLGTEHCYDSSAPPKVRRYQVLLSLMLSSQTKDQVTAGAMQRLRARGLTVDSILQTDD
                     ATLGKLIYPVGFWRSKVKYIKQTSAILQQHYGGDIPASVAELVALPGVGPKMAHLAMA
                     VAWGTVSGIAVDTHVHRIANRLRWTKKATKSPEETRAALEEWLPRELWHEINGLLVGF
                     GQQTCLPVHPRCHACLNQALCPAAQGL"
ORIGIN
        1 acgagcgcct tgagcgcgag gatgctgacc cggagccgga gcctgggacc cggggctggg
       61 ccgcgggggt gtagggagga gcccgggcct ctccggagaa gagaggctgc agcagaagcg
      121 aggaaaagcc acagccccgt gaagcgtccg cggaaagcac agagactgcg tgtggcctat
      181 gagggctcgg acagtgagaa aggtgaggct gagcccctca aggtgccagt ctgggagccc
      241 caggactggc agcaacagct ggtcaacatc cgtgccatga ggaacaaaaa ggatgcacct
      301 gtggaccatc tggggactga gcactgctat gactccagtg ccccccccaaa ggtacgcagg
      361 taccaggtgc tgctgtcact gatgctctcc agccaaacca aagaccaggt gacggcgggc
      421 gccatgcagc gactgcgggc gcgggggcctg acggtggaca gcatcctgca gacagatgat
      481 gccacgctgg gcaagctcat ctaccccgtc ggtttctgga ggagcaaggt gaaatacatc
      541 aagcagacca gcgccatcct gcagcagcac tacggtgggg acatcccagc ctctgtggcc
      601 gagctggtgg cgctgccggg tgttgggccc aagatggcac acctggctat ggctgtggcc
      661 tggggcactg tgtcaggcat tgcagtggac acgcatgtgc acagaatcgc caacaggctg
      721 aggtggacca agaaggcaac caagtcccca gaggagaccc gcgccgccct ggaggagtgg
      781 ctgcctaggg agctgtggca cgagatcaat ggactcttgg tgggcttcgg ccagcagacc
      841 tgtctgcctg tgcaccctcg ctgccacgcc tgcctcaacc aagccctctg cccggccgcc
      901 cagggtctct gatggccgca tggctctggc cgaggtgccg ctgtggccac cgtctgtgaa
      961 gtggctttac gcttcaggaa gccacgcctg ttgaataaag ctttggtgtg tttgcaaaaa
     1021 aaaaaaaaaa
//
```

Entry in GenBank format

April 2011:

- 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions
- 191,401,393,188 bases in 62,715,288 sequence records in the WGS

Growth of GenBank

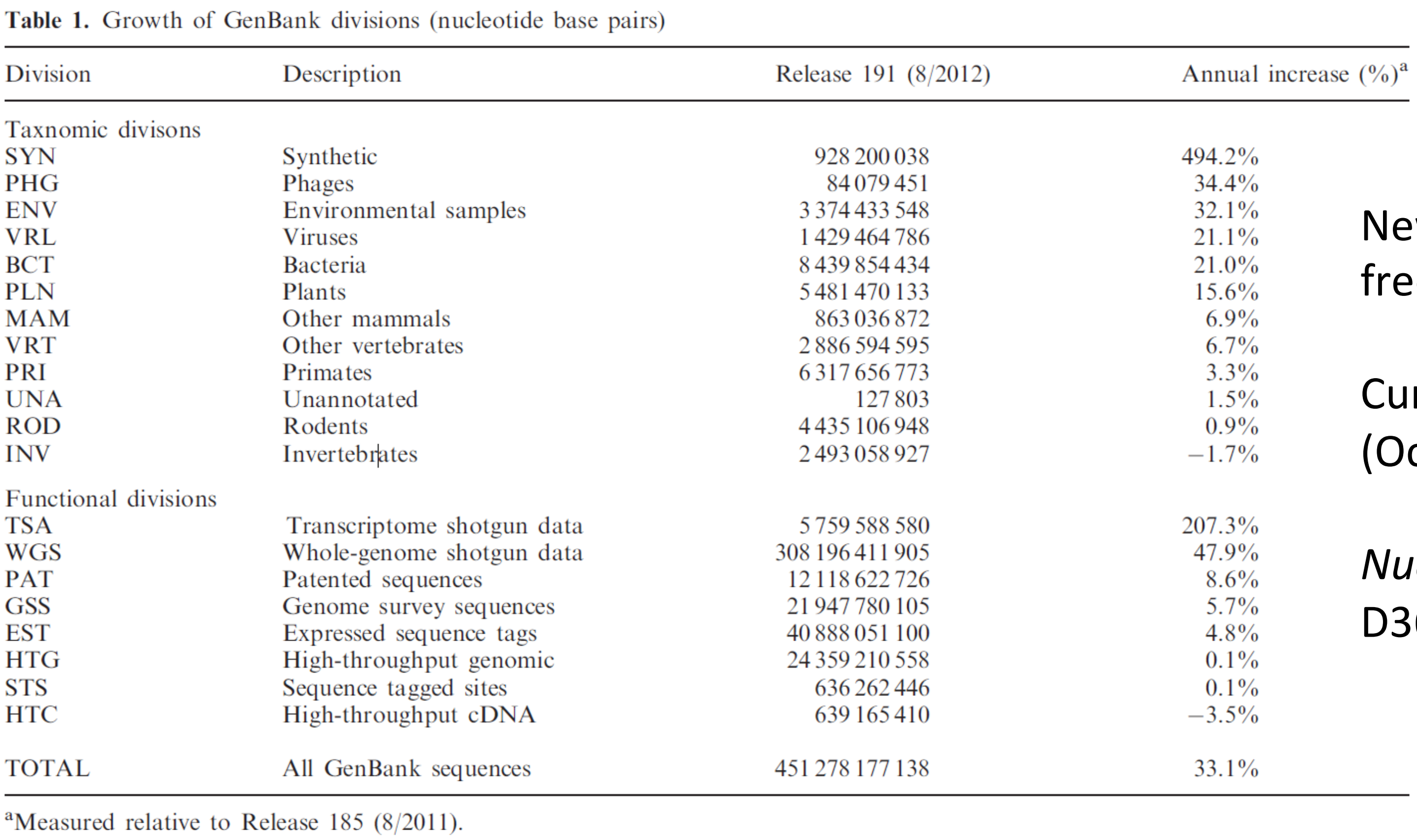


Table 1. Growth of GenBank divisions (nucleotide base pairs)

Division	Description	Release 191 (8/2012)	Annual increase (%)[a]
Taxnomic divisons			
SYN	Synthetic	928 200 038	494.2%
PHG	Phages	84 079 451	34.4%
ENV	Environmental samples	3 374 433 548	32.1%
VRL	Viruses	1 429 464 786	21.1%
BCT	Bacteria	8 439 854 434	21.0%
PLN	Plants	5 481 470 133	15.6%
MAM	Other mammals	863 036 872	6.9%
VRT	Other vertebrates	2 886 594 595	6.7%
PRI	Primates	6 317 656 773	3.3%
UNA	Unannotated	127 803	1.5%
ROD	Rodents	4 435 106 948	0.9%
INV	Invertebrates	2 493 058 927	−1.7%
Functional divisions			
TSA	Transcriptome shotgun data	5 759 588 580	207.3%
WGS	Whole-genome shotgun data	308 196 411 905	47.9%
PAT	Patented sequences	12 118 622 726	8.6%
GSS	Genome survey sequences	21 947 780 105	5.7%
EST	Expressed sequence tags	40 888 051 100	4.8%
HTG	High-throughput genomic	24 359 210 558	0.1%
STS	Sequence tagged sites	636 262 446	0.1%
HTC	High-throughput cDNA	639 165 410	−3.5%
TOTAL	All GenBank sequences	451 278 177 138	33.1%

[a] Measured relative to Release 185 (8/2011).

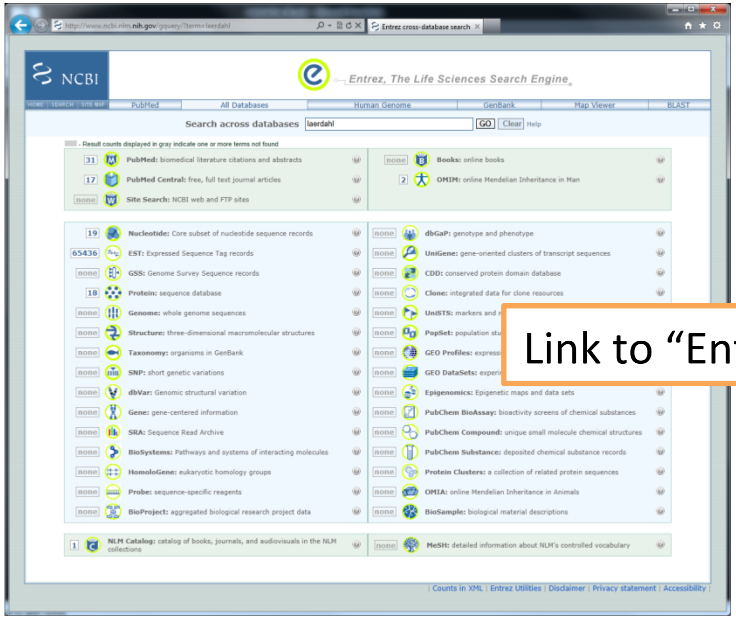
New release frequency: 2 months

Current release is 204 (Oct, 2014)

Nucleic Acids Res. **41**, D36 (2013)

NCBI Entrez retrieval system

- Entrez is the most widely used interface for information retrieval from the NCBI databases
  - search engine
  - web portal
  - global query of all (35?) NCBI databases



http://www.ncbi.nlm.nih.gov/gquery/?term=laerdahl

Entrez cross-database search

Search across databases: laerdahl

- Result counts displayed in gray indicate one or more terms not found

31	**PubMed:** biomedical literature citations and abstracts	none	**Books:** online books
17	**PubMed Central:** free, full text journal articles	2	**OMIM:** online Mendelian Inheritance in Man
none	**Site Search:** NCBI web and FTP sites		

19	**Nucleotide:** Core subset of nucleotide sequence records	none	**dbGaP:** genotype and phenotype
65436	**EST:** Expressed Sequence Tag records	none	**UniGene:** gene-oriented clusters of transcript sequences
none	**GSS:** Genome Survey Sequence records	none	**CDD:** conserved protein domain database
18	**Protein:** sequence database	none	**Clone:** integrated data for clone resources
none	**Genome:** whole genome sequences	none	**UniSTS:** markers and r
none	**Structure:** three-dimensional macromolecular structures	none	**PopSet:** population stu
none	**Taxonomy:** organisms in GenBank	none	**GEO Profiles:** expressi
none	**SNP:** short genetic variations	none	**GEO DataSets:** experi
none	**dbVar:** Genomic structural variation	none	**Epigenomics:** Epigenetic maps and data sets
none	**Gene:** gene-centered information	none	**PubChem BioAssay:** bioactivity screens of chemical substances
none	**SRA:** Sequence Read Archive	none	**PubChem Compound:** unique small molecule chemical structures
none	**BioSystems:** Pathways and systems of interacting molecules	none	**PubChem Substance:** deposited chemical substance records
none	**HomoloGene:** eukaryotic homology groups	none	**Protein Clusters:** a collection of related protein sequences
none	**Probe:** sequence-specific reagents	none	**OMIA:** online Mendelian Inheritance in Animals
none	**BioProject:** aggregated biological research project data	none	**BioSample:** biological material descriptions

1	**NLM Catalog:** catalog of books, journals, and audiovisuals in the NLM collections	none	**MeSH:** detailed information about NLM's controlled vocabulary

| Counts in XML | Entrez Utilities | Disclaimer | Privacy statement | Accessibility |

Link to "Entrez" and search for "Laerdahl"