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?
- Setting up your laptops

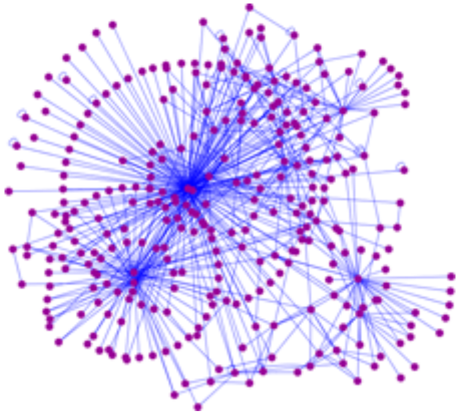
What about those of you that know Unix and Python very well?

- 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

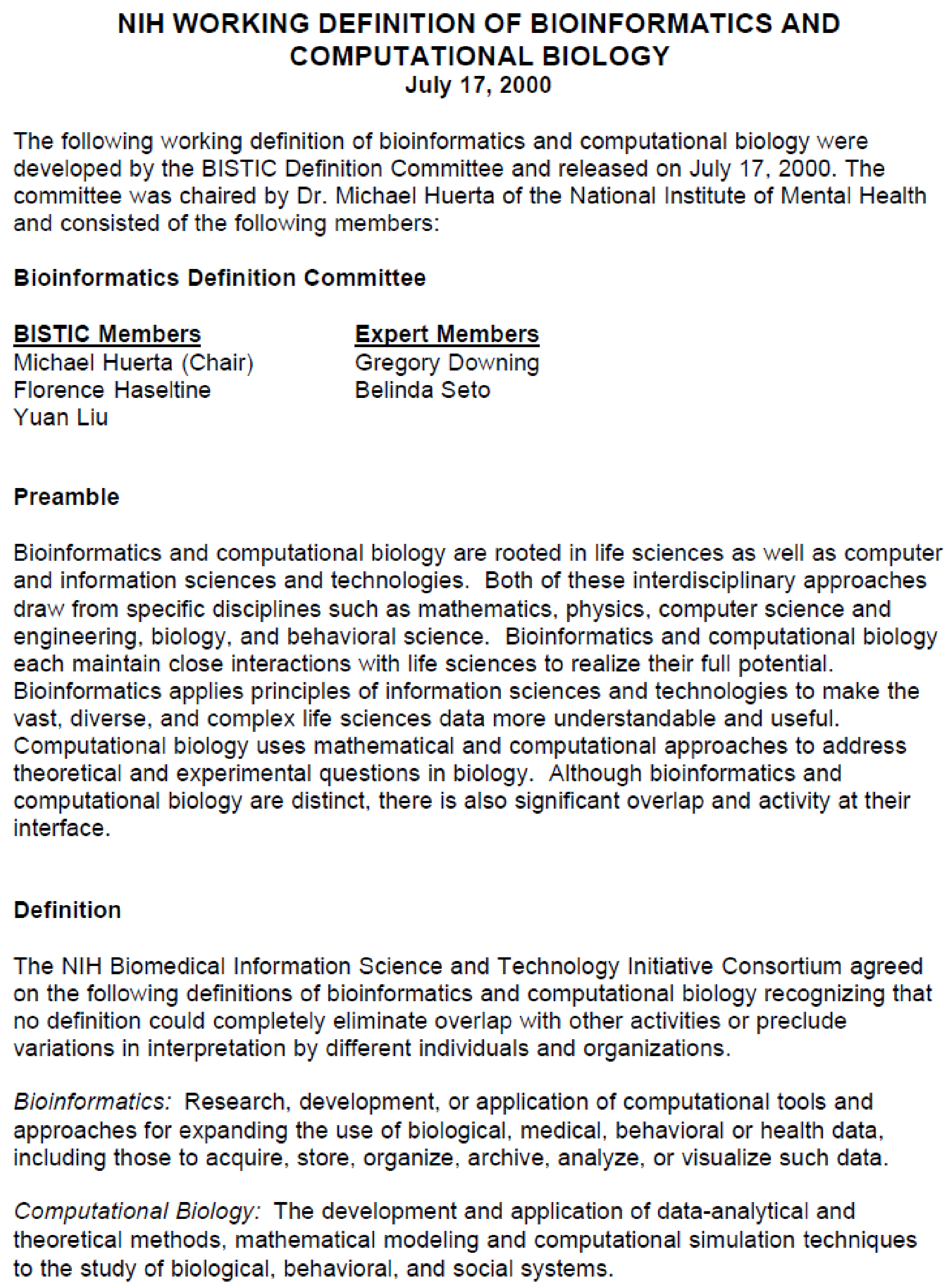
Biology in the 21st century is being transformed from a purely lab-based science to an information science as well.



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 2015 you need bioinformatics/CLS/informatics competence!!

Some examples

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

LETTERS

Genome-wide measurement of RNA secondary structure in yeast

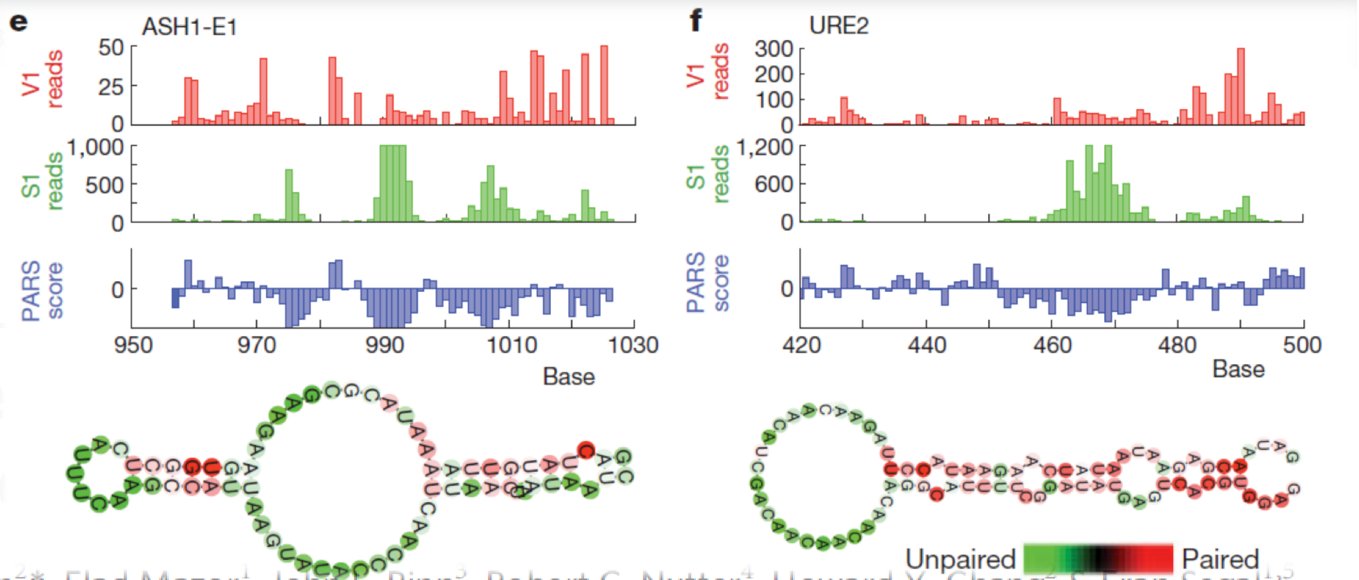
Michael Kertesz[1]*†, Yue Wan[2]*, Elad Mazor[1], John L. Rinn[3], Robert C. Nutter[4], Howard Y. Chang[2] & Eran Segal[1,5]

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



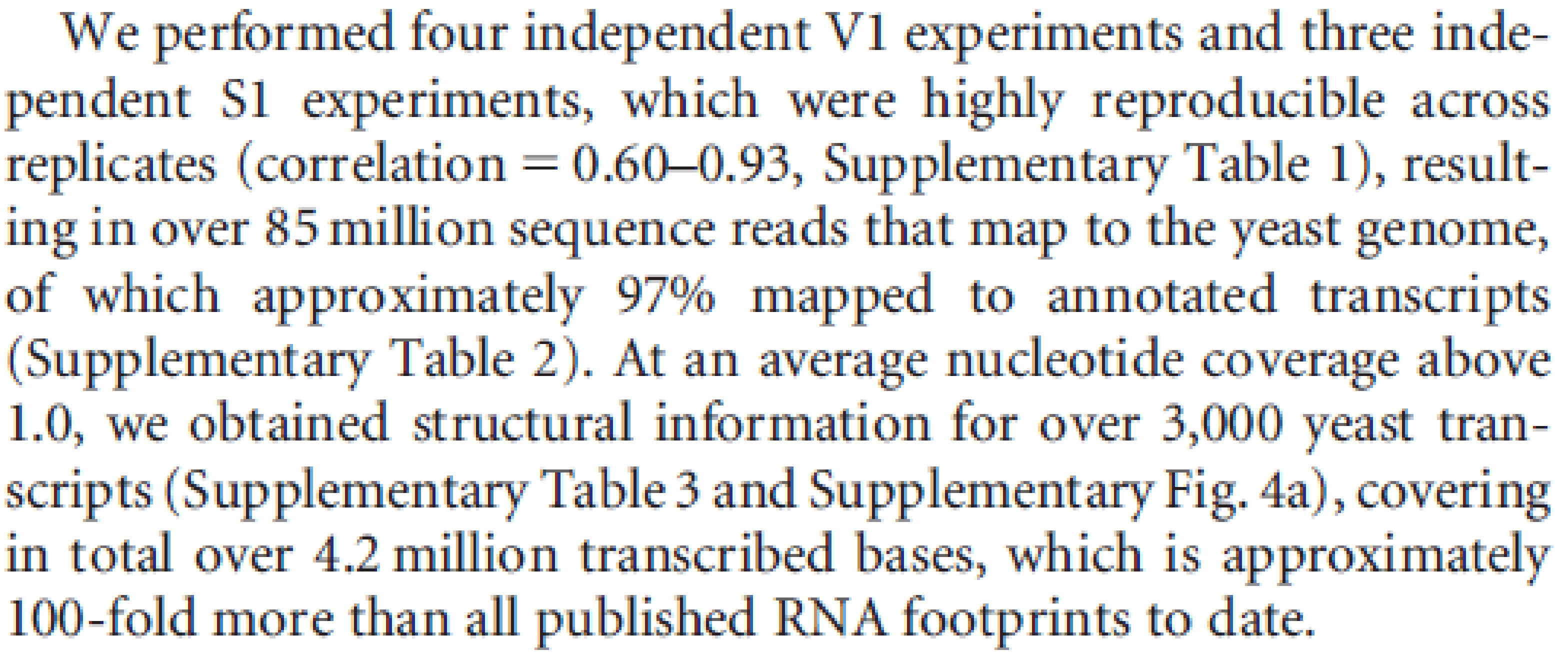
Genome-wide structure in y

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.

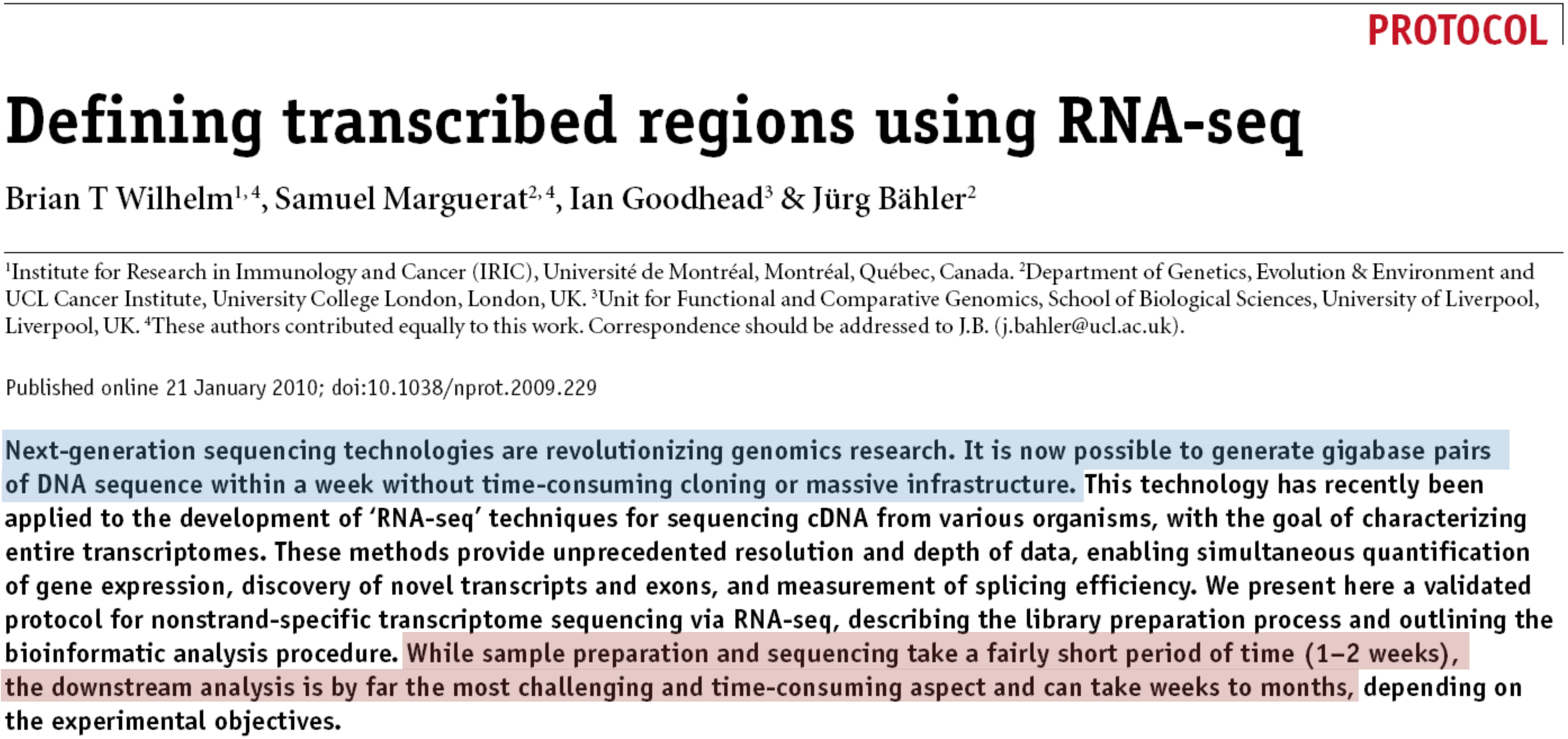
that typically have 5′ hydroxyl (Supplementary Fig. 3). Thus each observed cleavage site provides evidence that the cut nucleotide was in a double-stranded (for V1-treated samples) or single-stranded (for S1-treated samples) conformation. As a quantitative measure at nucleotide resolution representing 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 pa

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, Jianjun Cao, 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

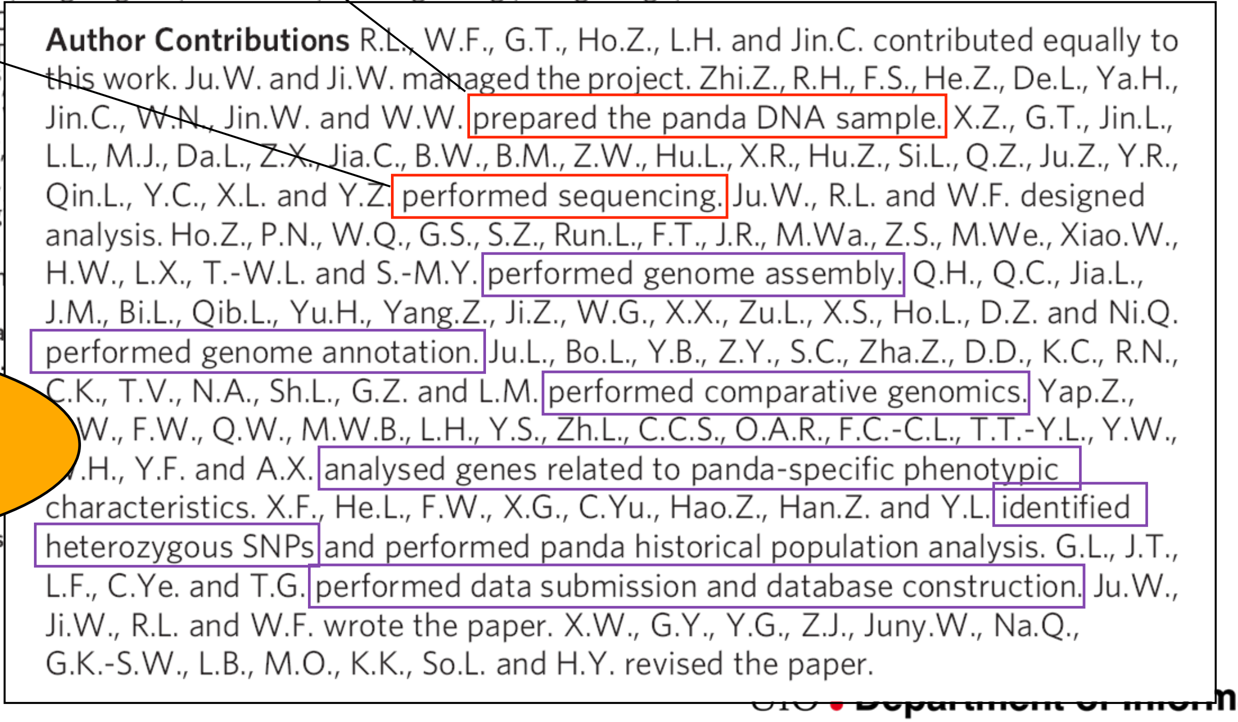
Using next-generation sequencing technology a giant pa (2.

Ou using next-ge technologies genomes.

Travelled around in China and took blood samples from pandas

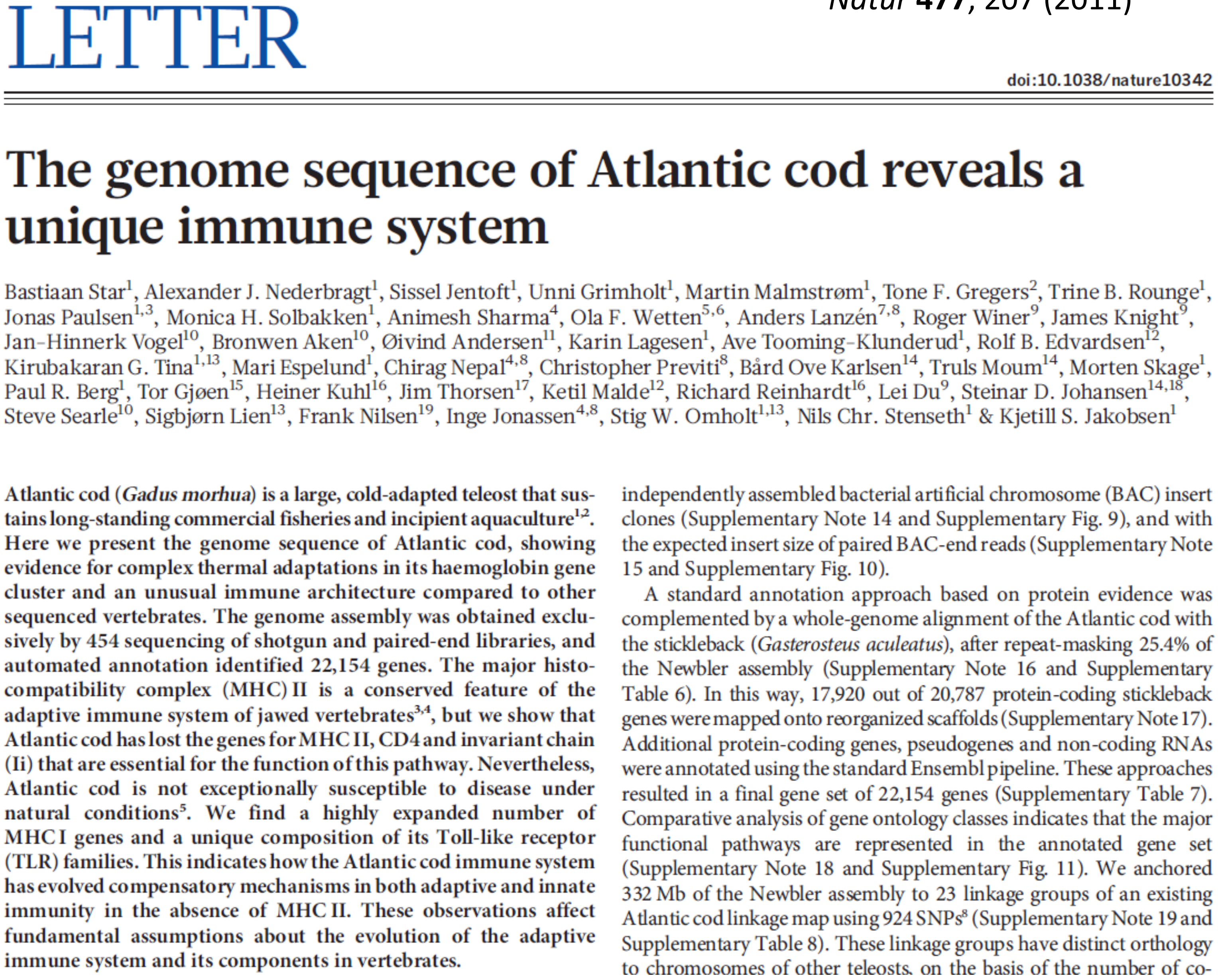
Wet lab?

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

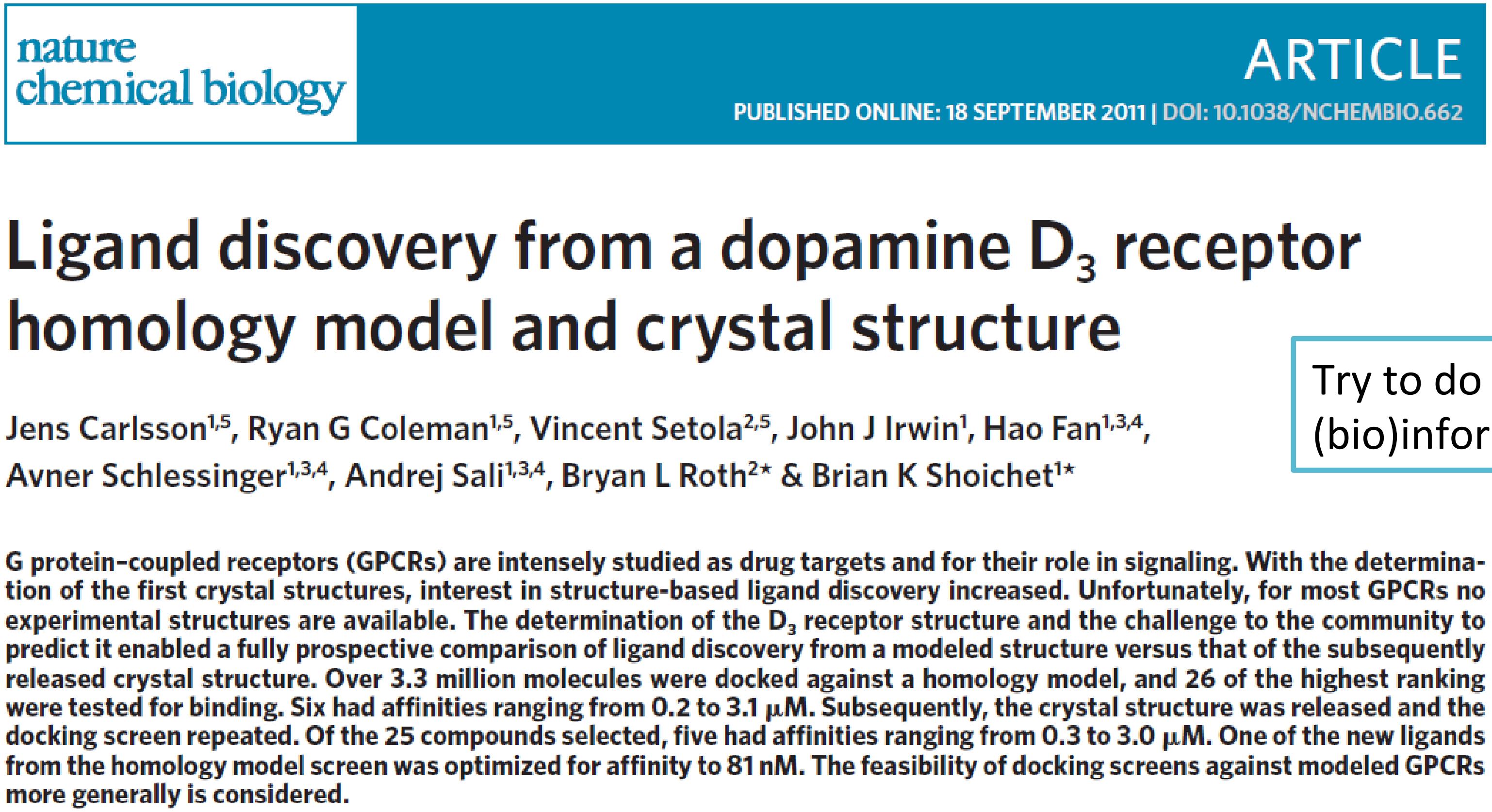
The genome sequence of Atlantic cod reveals a unique immune system

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-



nature chemical biology

ARTICLE

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

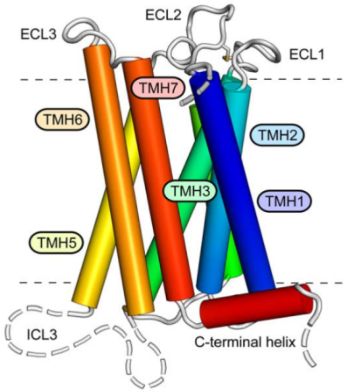
Ligand discovery from a dopamine D_3 receptor homology model and crystal structure

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.

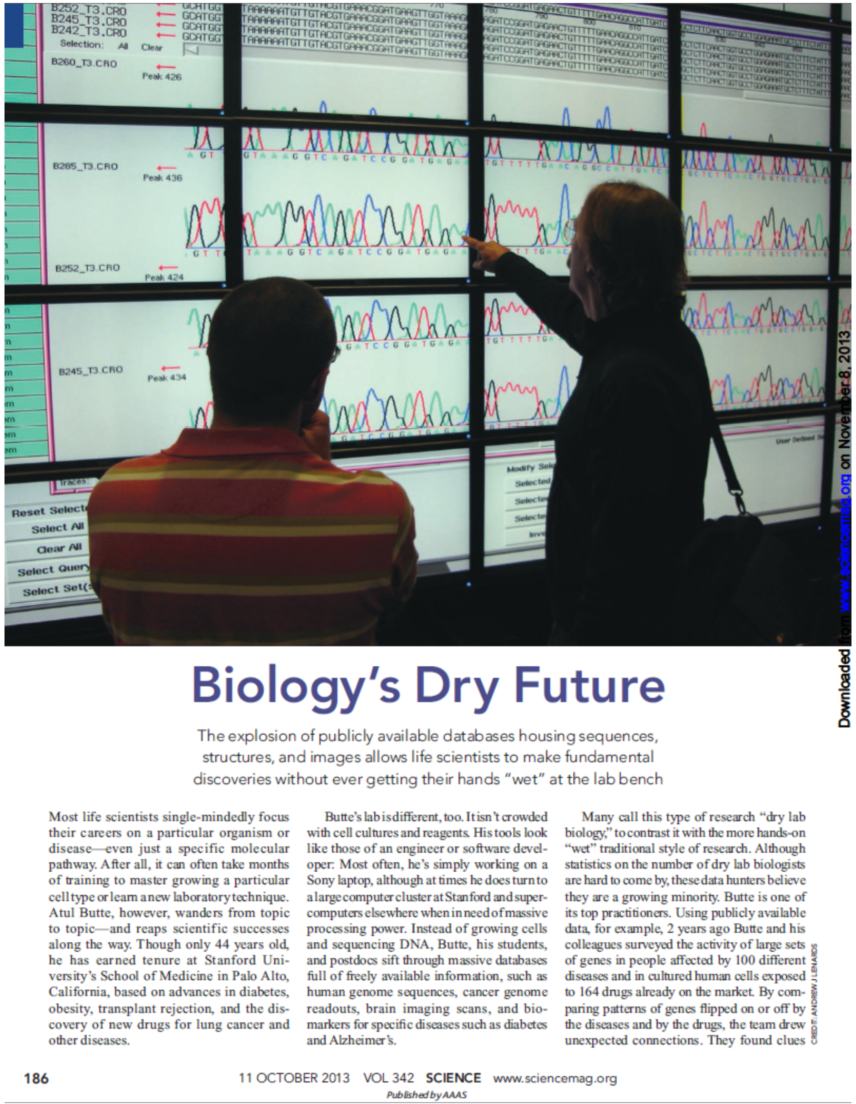
Try to do this without (bio)informatics skills?

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



Read this article as part of the curriculum!

No wet lab biology?



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

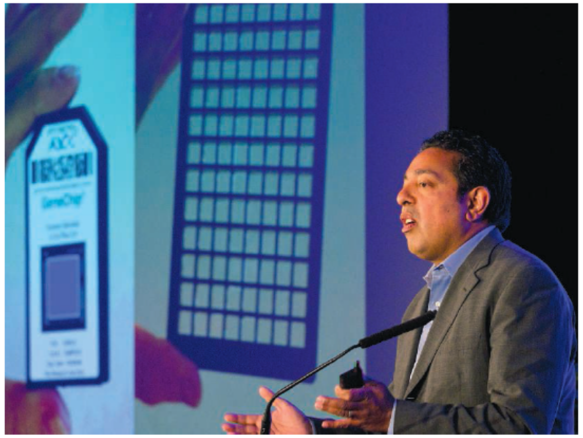
Most life scientists single-mindedly focus their careers on a particular organism or disease—even just a specific molecular pathway. After all, it can often take months of training to master growing a particular cell type or learn a new laboratory technique. Atul Butte, however, wanders from topic to 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

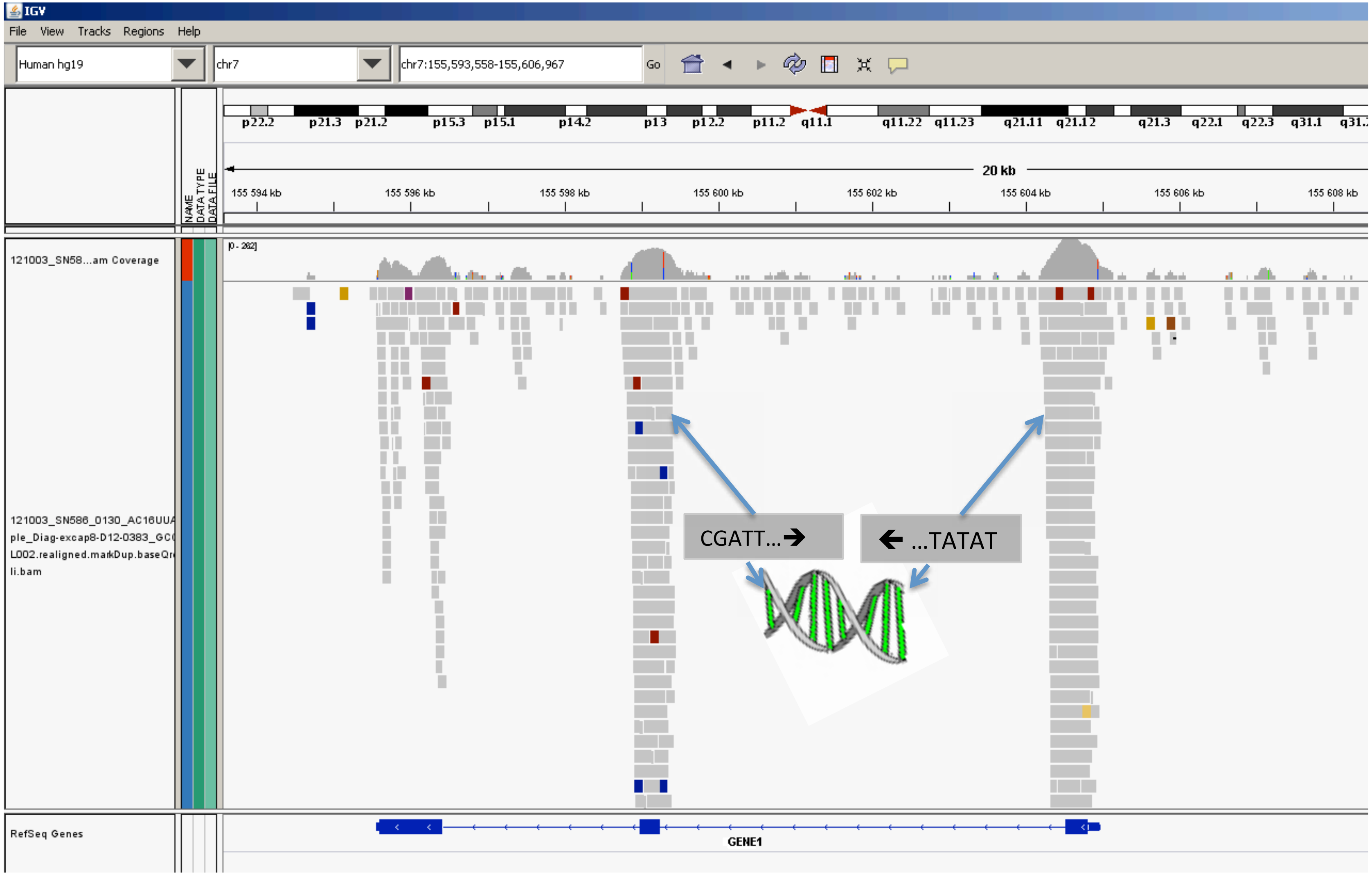
Department of Medical Genetics

Oslo University Hospital and University of Oslo

Oslo, Norway

Oslo University Hospital

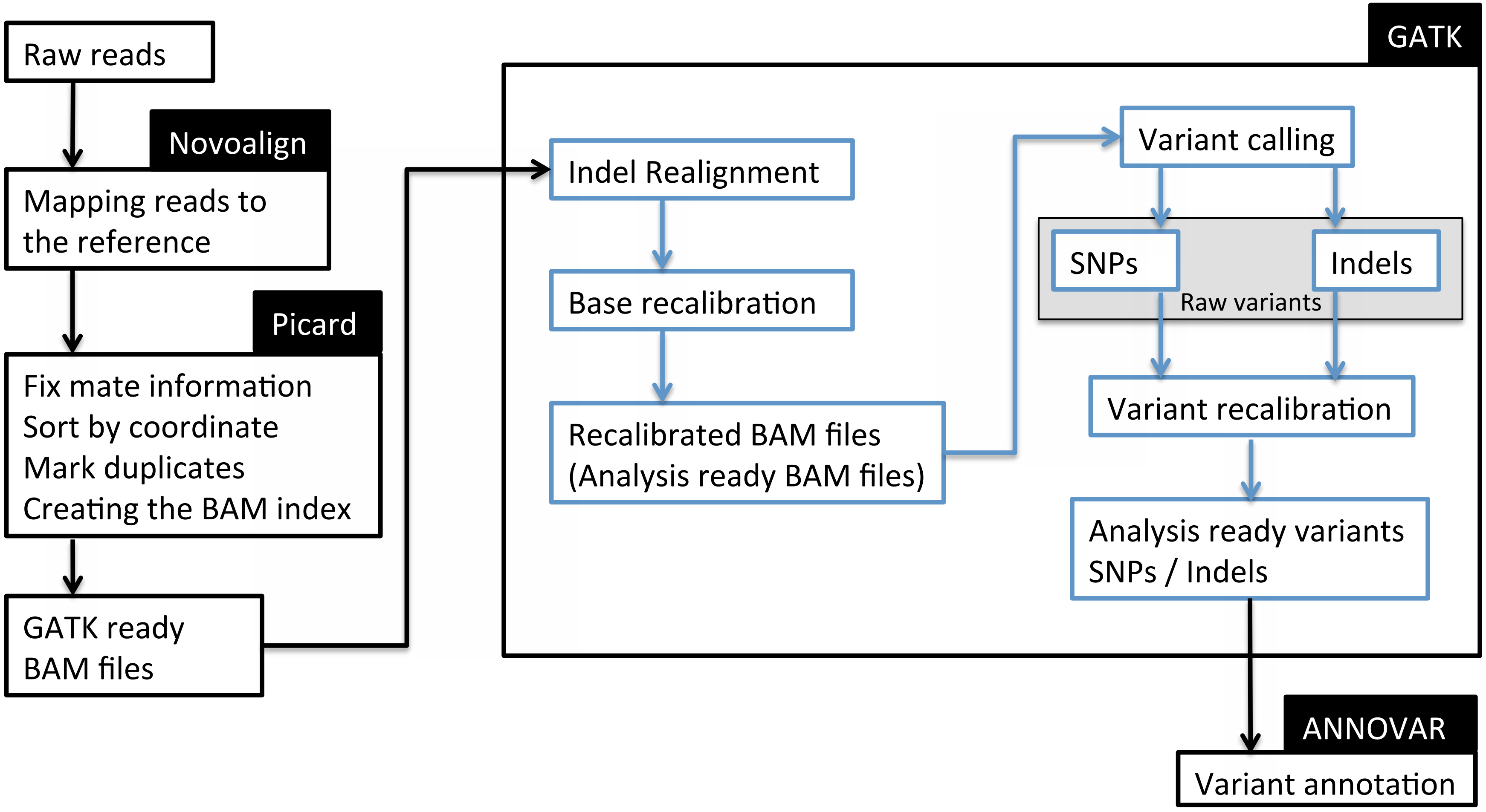
UiO : University of Oslo



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
[0 - 282]
121003_SN58...am Coverage
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

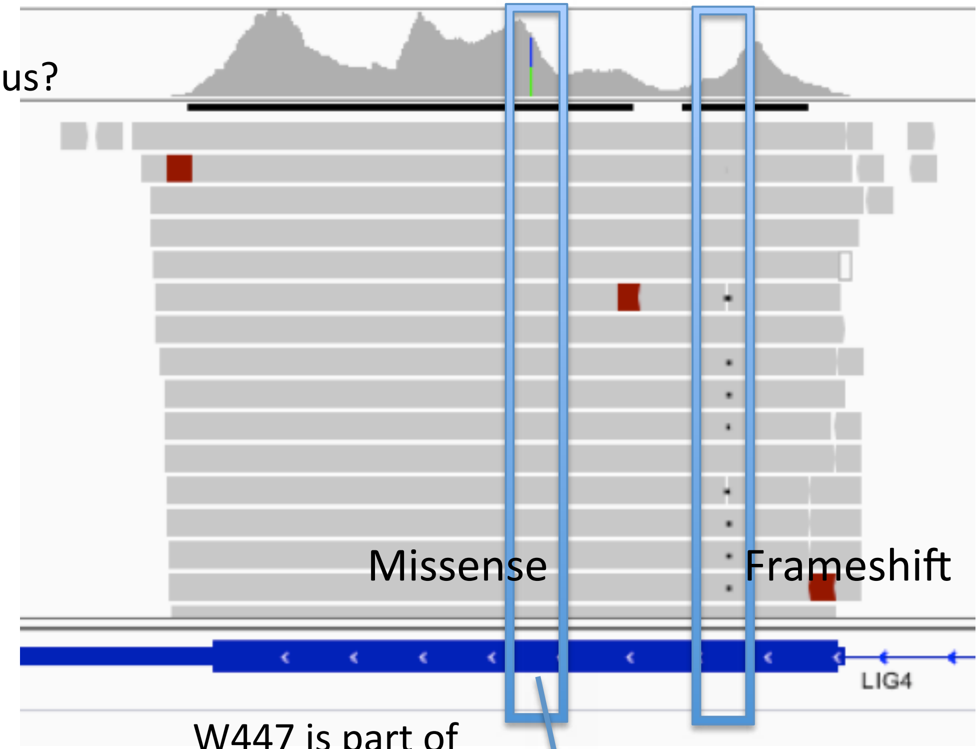


Solved case - example of clinical utility

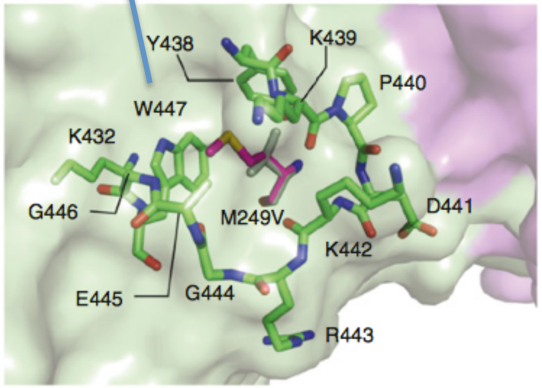
- 4 year old boy, healthy until 2 ½
- Feb -13: Anemia and thrombocytopenia – virus?

  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!!

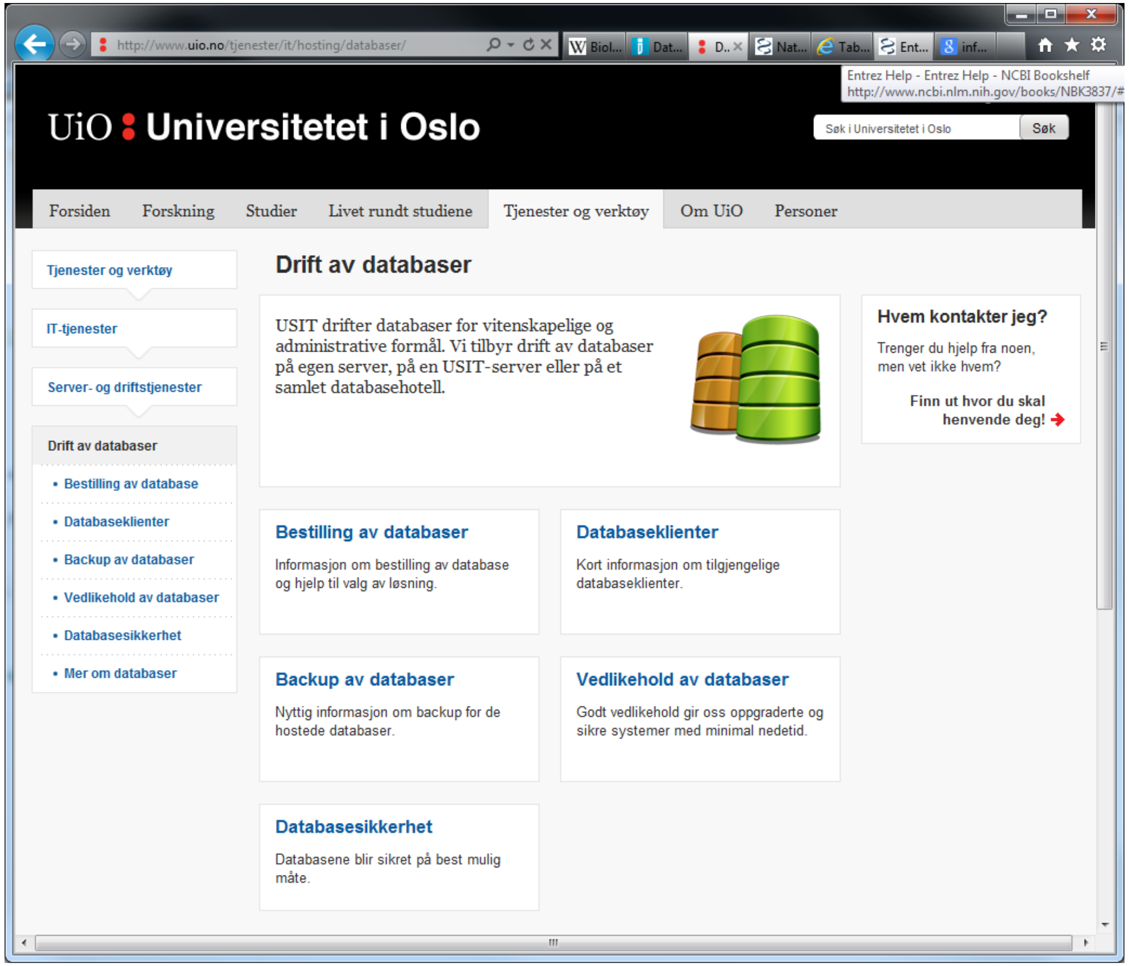
Database

- 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)


http://www.uio.no/tjenester/it/hosting/databaser/
Biol... Dat... D... Nat... Tab... Ent... inf...
Entrez Help - Entrez Help - NCBI Bookshelf
http://www.ncbi.nlm.nih.gov/books/NBK3837/#
UiO : Universitetet i Oslo
Søk i Universitetet i Oslo
Søk
Forsiden
Forskning
Studier
Livet rundt studiene
Tjenester og verktøy
Om UiO
Personer
Tjenester og verktøy
IT-tjenester
Server- og driftstjenester
Drift av databaser
• Bestilling av database
• Databaseklienter
• Backup av databaser
• Vedlikehold av databaser
• Databasesikkerhet
• Mer om databaser

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.

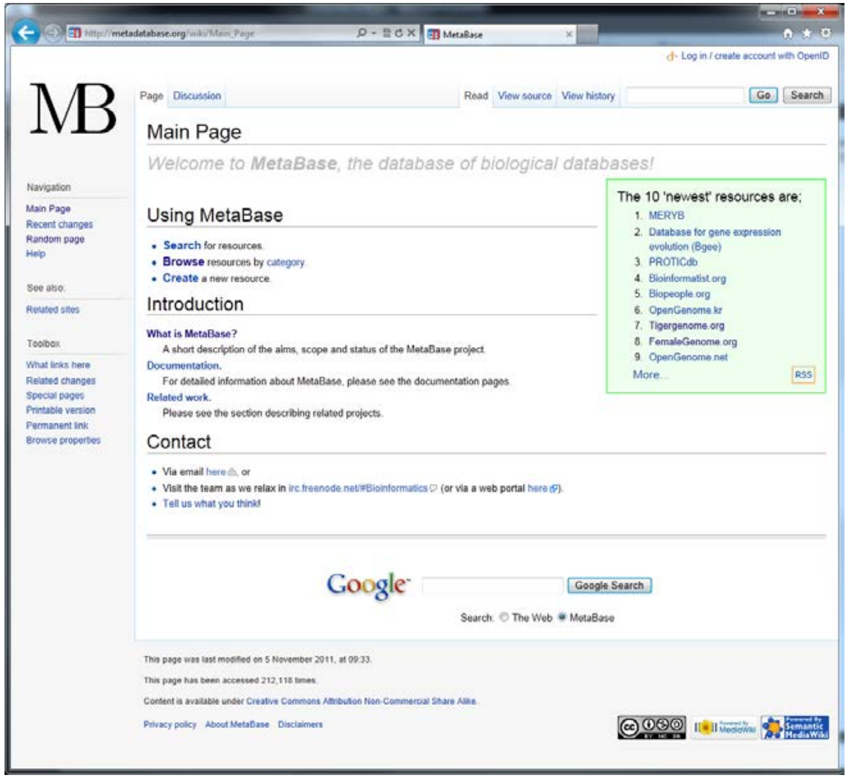
Vedlikehold av databaser

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...



Main Page

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

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.

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!

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...

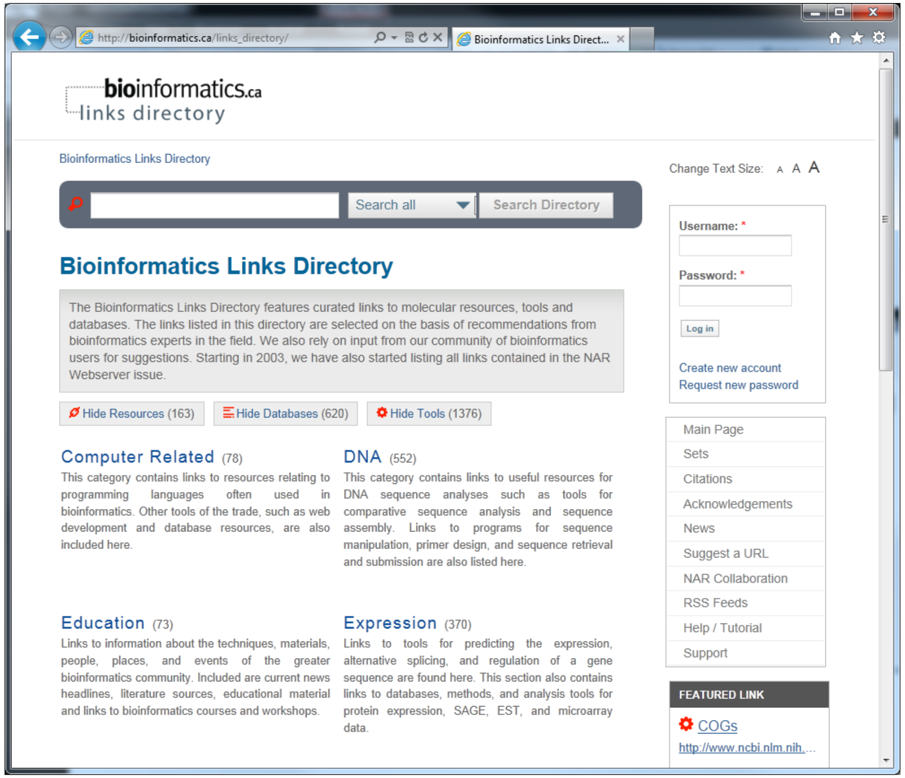
This page was last modified on 5 November 2011, at 09:33.

This page has been accessed 212,118 times.

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

MetaBase, the database of biological databases (>1800 entries)

- http://metadatabase.org



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 (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.

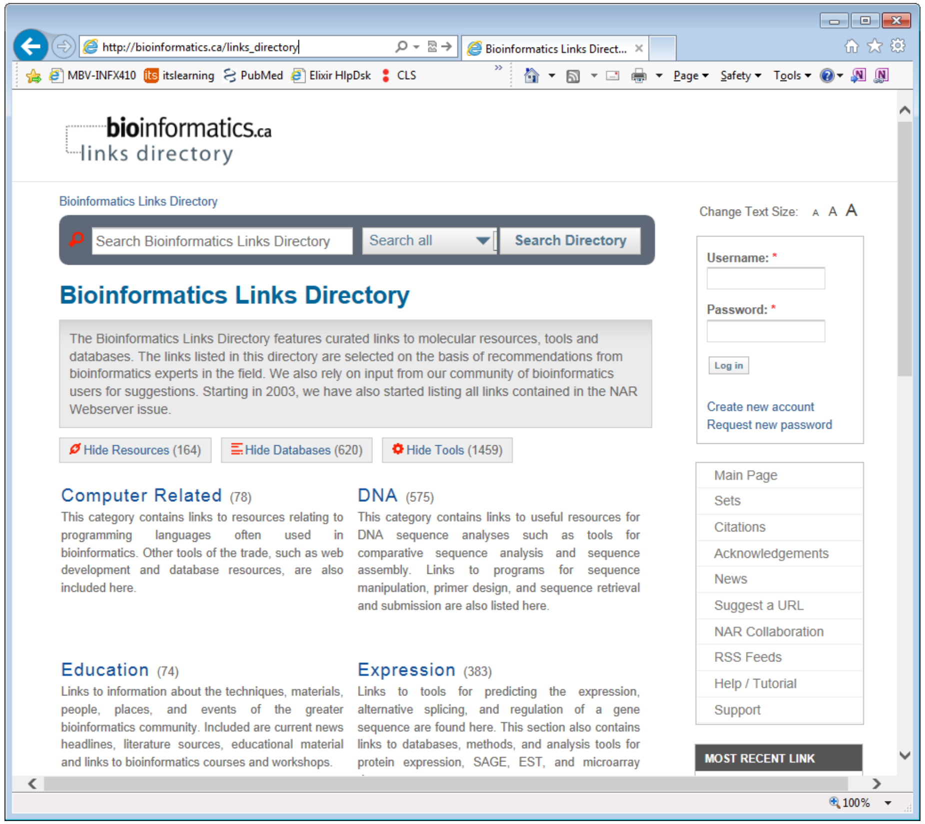
Expression (370)

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.

bioinformatics.ca – links directory (623 databases)

- http://bioinformatics.ca/links_directory

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


http://bioinformatics.ca/links_directory
Bioinformatics Links Direct...
MBV-INFX410
itslearning
PubMed
Elixir HlpDsk
CLS
Page
Safety
Tools
100%

bioinformatics.ca
links directory

Bioinformatics Links Directory

Search 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 (164) | Hide Databases (620) | Hide Tools (1459)

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

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

MOST RECENT LINK

bioinformatics.ca – links directory

- http://bioinformatics.ca/links_directory
- Currently
    - 1548 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 >23,000 entries (15,000 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

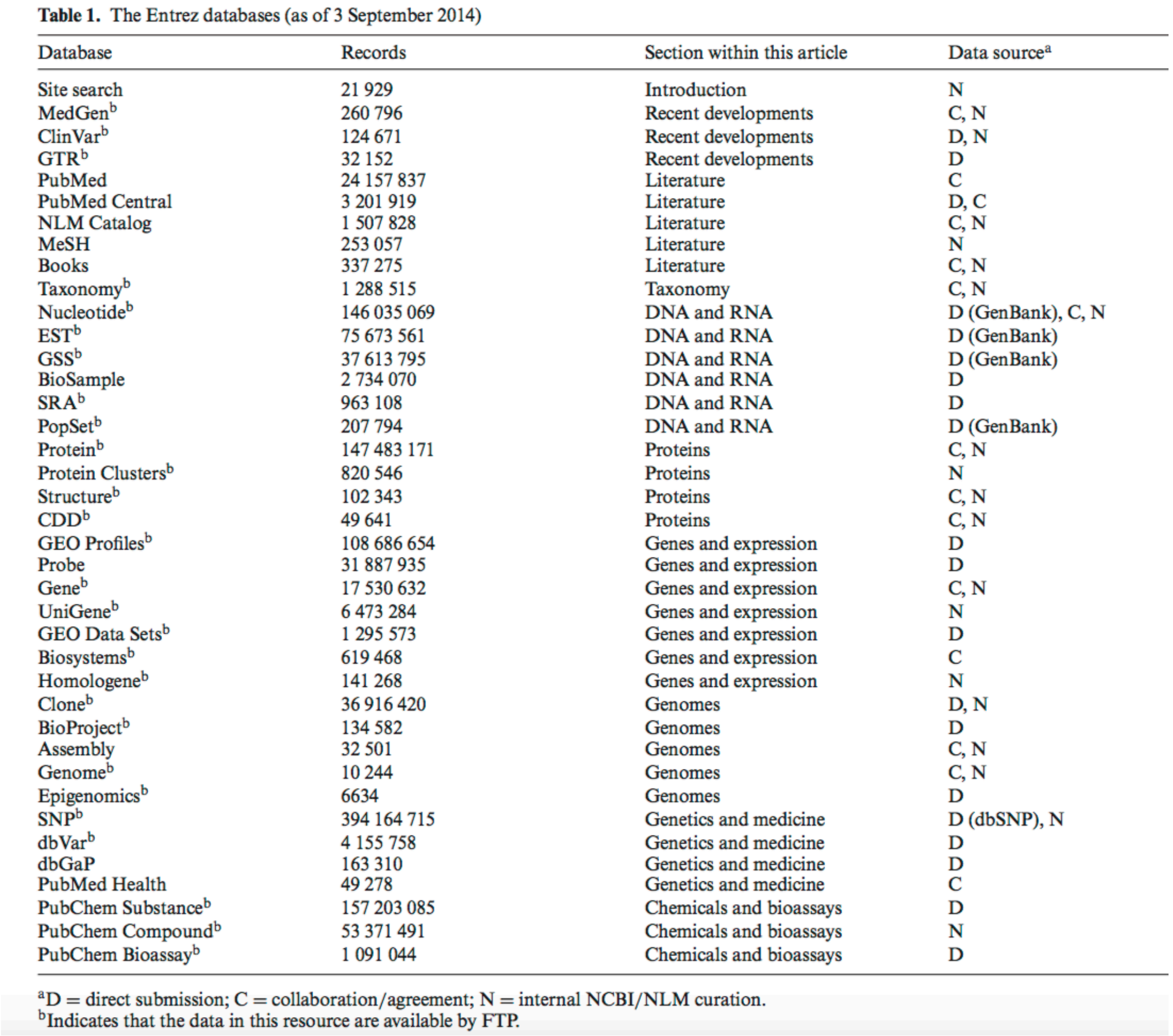


Table 1. The Entrez databases (as of 3 September 2014)

Database	Records	Section within this article	Data source[a]
Site search	21 929	Introduction	N
MedGen[b]	260 796	Recent developments	C, N
ClinVar[b]	124 671	Recent developments	D, N
GTR[b]	32 152	Recent developments	D
PubMed	24 157 837	Literature	C
PubMed Central	3 201 919	Literature	D, C
NLM Catalog	1 507 828	Literature	C, N
MeSH	253 057	Literature	N
Books	337 275	Literature	C, N
Taxonomy[b]	1 288 515	Taxonomy	C, N
Nucleotide[b]	146 035 069	DNA and RNA	D (GenBank), C, N
EST[b]	75 673 561	DNA and RNA	D (GenBank)
GSS[b]	37 613 795	DNA and RNA	D (GenBank)
BioSample	2 734 070	DNA and RNA	D
SRA[b]	963 108	DNA and RNA	D
PopSet[b]	207 794	DNA and RNA	D (GenBank)
Protein[b]	147 483 171	Proteins	C, N
Protein Clusters[b]	820 546	Proteins	N
Structure[b]	102 343	Proteins	C, N
CDD[b]	49 641	Proteins	C, N
GEO Profiles[b]	108 686 654	Genes and expression	D
Probe	31 887 935	Genes and expression	D
Gene[b]	17 530 632	Genes and expression	C, N
UniGene[b]	6 473 284	Genes and expression	N
GEO Data Sets[b]	1 295 573	Genes and expression	D
Biosystems[b]	619 468	Genes and expression	C
Homologene[b]	141 268	Genes and expression	N
Clone[b]	36 916 420	Genomes	D, N
BioProject[b]	134 582	Genomes	D
Assembly	32 501	Genomes	C, N
Genome[b]	10 244	Genomes	C, N
Epigenomics[b]	6634	Genomes	D
SNP[b]	394 164 715	Genetics and medicine	D (dbSNP), N
dbVar[b]	4 155 758	Genetics and medicine	D
dbGaP	163 310	Genetics and medicine	D
PubMed Health	49 278	Genetics and medicine	C
PubChem Substance[b]	157 203 085	Chemicals and bioassays	D
PubChem Compound[b]	53 371 491	Chemicals and bioassays	N
PubChem Bioassay[b]	1 091 044	Chemicals and bioassays	D

[a] D = direct submission; C = collaboration/agreement; N = internal NCBI/NLM curation.
[b] Indicates that the data in this resource are available by FTP.

40 databases that together contains 1.3 billion records

Nucleic Acids Res. **43**, D6 (2015)