How to welcome the new era of public research data?

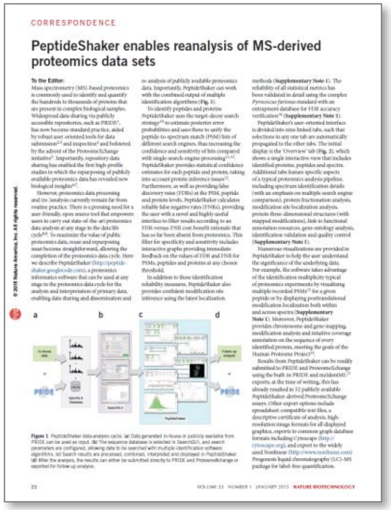
Harald Barsnes

Department of Biomedicine
& Department of Informatics

University of Bergen

PDA19
Gulbenkian - April 3rd 2019

Research is the sharing of knowledge, data and software!



CORRESPONDENCE

PeptideShaker enables reanalysis of MS-derived proteomics data sets

To the Editor:
Mass spectrometry (MS)-based proteomics is commonly used to identify and quantify the hundreds to thousands of proteins that are present in complex biological samples. Widespread data sharing via publicly accessible repositories, such as PRIDE[1], has now become standard practice, aided by robust user-oriented tools for data submission[2,3] and inspection[4] and bolstered by the advent of the ProteomeXchange initiative[5]. Importantly, repository data sharing has enabled the first high-profile studies in which the repurposing of publicly available proteomics data has revealed new biological insights[6,7].

However, proteomics data processing and (re-)analysis currently remain far from routine practice. There is a pressing need for a user-friendly, open source tool that empowers users to carry out state-of-the-art proteomics data analysis at any stage in the data life cycle[8,9]. To maximize the value of public proteomics data, reuse and repurposing must become straightforward, allowing the completion of the proteomics data cycle. Here we describe PeptideShaker (http://peptide-shaker.googlecode.com), a proteomics informatics software that can be used at any stage in the proteomics data cycle for the analysis and interpretation of primary data, enabling data sharing and dissemination and re-analysis of publicly available proteomics data. Importantly, PeptideShaker can work with the combined output of multiple identification algorithms (**Fig. 1**).

To identify peptides and proteins PeptideShaker uses the target-decoy search strategy[10] to estimate posterior error probabilities and uses these to unify the peptide-to-spectrum match (PSM) lists of different search engines, thus increasing the confidence and sensitivity of hits compared with single-search-engine processing[11,12]. PeptideShaker provides statistical confidence estimates for each peptide and protein, taking into account protein inference issues[13]. Furthermore, as well as providing false discovery rates (FDRs) at the PSM, peptide and protein levels, PeptideShaker calculates reliable false negative rates (FNRs), providing the user with a novel and highly useful interface to filter results according to an FDR-versus-FNR cost-benefit rationale that has so far been absent from proteomics. This filter for specificity and sensitivity includes interactive graphs providing immediate feedback on the values of FDR and FNR for PSMs, peptides and proteins at any chosen threshold.

In addition to these identification reliability measures, PeptideShaker also provides confident modification site inference using the latest localization methods (**Supplementary Note 1**). The reliability of all statistical metrics has been validated in detail using the complex *Pyrococcus furiosus* standard with an entrapment database for FDR accuracy verification[14] (**Supplementary Note 1**).

PeptideShaker's user-oriented interface is divided into nine linked tabs, such that selections in any one tab are automatically propagated to the other tabs. The initial display is the 'Overview' tab (**Fig. 2**), which shows a single interactive view that includes identified proteins, peptides and spectra. Additional tabs feature specific aspects of a typical proteomics analysis pipeline, including spectrum identification details (with an emphasis on multiple search engine comparison), protein fractionation analysis, modification site localization analysis, protein three-dimensional structures (with mapped modifications), link to functional annotation resources, gene ontology analysis, identification validation and quality control (**Supplementary Note 1**).

Numerous visualizations are provided in PeptideShaker to help the user understand the significance of the underlying data. For example, the software takes advantage of the identification multiplicity typical of proteomics experiments by visualizing multiple recorded PSMs[15] for a given peptide or by displaying posttranslational modification localization both within and across spectra (**Supplementary Note 1**). Moreover, PeptideShaker provides chromosome and gene mapping, modification analysis and intuitive coverage annotation on the sequence of every identified protein, meeting the goals of the Human Proteome Project[16].

Results from PeptideShaker can be readily submitted to PRIDE and ProteomeXchange using the built-in PRIDE and mzIdentML[17] exports; at the time of writing, this has already resulted in 52 publicly available PeptideShaker-derived ProteomeXchange assays. Other export options include spreadsheet-compatible text files, a descriptive certificate of analysis, high-resolution image formats for all displayed graphics, exports to common graph database formats including Cytoscape (http://cytoscape.org), and export to the widely used Nonlinear (http://www.nonlinear.com) Progenesis liquid chromatography (LC)-MS package for label-free quantification.

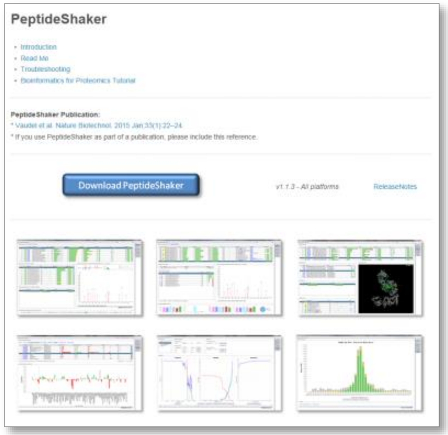
Figure 1 PeptideShaker data-analysis cycle. (a) Data generated in-house or publicly available from PRIDE can be used as input. (b) The sequence database is selected in SearchGUI, and search parameters are configured, allowing data to be searched with multiple identification software algorithms. (c) Search results are processed, combined, interpreted and displayed in PeptideShaker. (d) After the analysis, the results can either be submitted directly to PRIDE and ProteomeXchange or exported for follow-up analysis.

Scientific Paper

+

Underlying Data

+

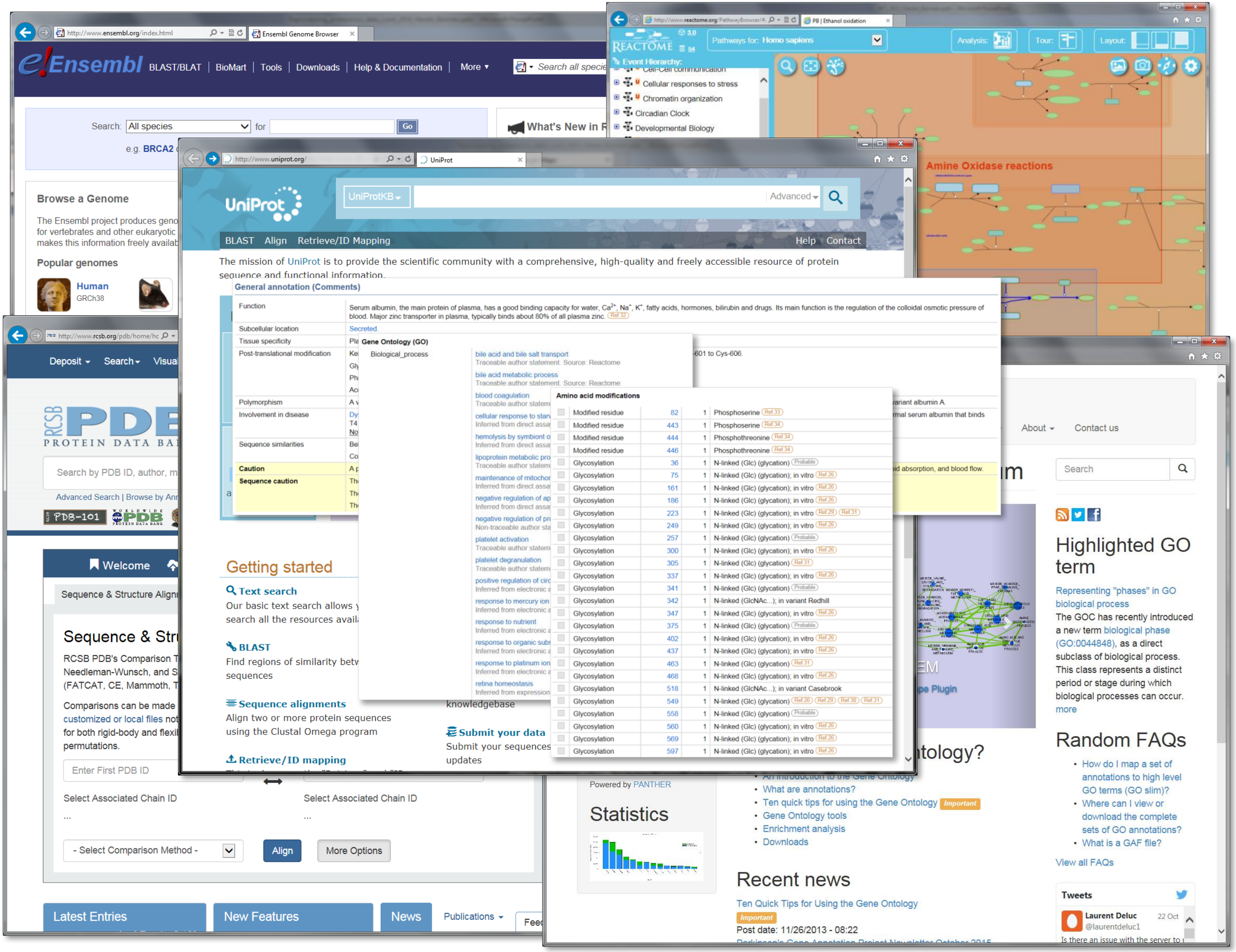


PeptideShaker

- Introduction
- Read Me
- Troubleshooting
- Bioinformatics for Proteomics Tutorial

PeptideShaker Publication:
* Vaudel et al. Nature Biotechnol. 2015 Jan;33(1):22–24.
* If you use PeptideShaker as part of a publication, please include this reference.

Software



http://www.ensembl.org/index.html — Ensembl Genome Browser

e!Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | More ▾ Search all specie

Search: All species for Go

e.g. BRCA2

What's New in R

Browse a Genome

The Ensembl project produces geno for vertebrates and other eukaryotic makes this information freely availab

Popular genomes

Human GRCh38

http://www.reactome.org/PathwayBrowser/# — PB | Ethanol oxidation

REACTOME v3.0 — Pathways for: Homo sapiens — Analysis: Tour: Layout:

Event Hierarchy:
- Cell-Cell communication
- Cellular responses to stress
- Chromatin organization
- Circadian Clock
- Developmental Biology

Amine Oxidase reactions

http://www.uniprot.org/ — UniProt

UniProt UniProtKB ▾ Advanced ▾

BLAST Align Retrieve/ID Mapping Help Contact

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

General annotation (Comments)

Function	Serum albumin, the main protein of plasma, has a good binding capacity for water, Ca^{2+}, Na^+, K^+, fatty acids, hormones, bilirubin and drugs. Its main function is the regulation of the colloidal osmotic pressure of blood. Major zinc transporter in plasma, typically binds about 80% of all plasma zinc. Ref.32
Subcellular location	Secreted.
Tissue specificity	Pla
Post-translational modification	Ke … -601 to Cys-606. Gly Ph Ac
Polymorphism	A v … variant albumin A.
Involvement in disease	Dy … rmal serum albumin that binds T4 No
Sequence similarities	Be Co
Caution	A p … id absorption, and blood flow.
Sequence caution	Th Th Th

Gene Ontology (GO)

Biological_process
- bile acid and bile salt transport — Traceable author statement. Source: Reactome
- bile acid metabolic process — Traceable author statement. Source: Reactome
- blood coagulation — Traceable author statem
- cellular response to starv — Inferred from direct assa
- hemolysis by symbiont o — Inferred from direct assa
- lipoprotein metabolic pro — Traceable author statem
- maintenance of mitochon — Inferred from direct assa
- negative regulation of ap — Inferred from direct assa
- negative regulation of pr — Non-traceable author sta
- platelet activation — Traceable author statem
- platelet degranulation — Traceable author statem
- positive regulation of circ — Inferred from electronic a
- response to mercury ion — Inferred from electronic a
- response to nutrient — Inferred from electronic a
- response to organic subs — Inferred from electronic a
- response to platinum ion — Inferred from electronic a
- retina homeostasis — Inferred from expression

Amino acid modifications

□	Modified residue	82	1	Phosphoserine Ref.33
□	Modified residue	443	1	Phosphoserine Ref.34
□	Modified residue	444	1	Phosphothreonine Ref.34
□	Modified residue	446	1	Phosphothreonine Ref.34
□	Glycosylation	36	1	N-linked (Glc) (glycation) Probable
□	Glycosylation	75	1	N-linked (Glc) (glycation); in vitro Ref.26
□	Glycosylation	161	1	N-linked (Glc) (glycation); in vitro Ref.26
□	Glycosylation	186	1	N-linked (Glc) (glycation); in vitro Ref.26
□	Glycosylation	223	1	N-linked (Glc) (glycation); in vitro Ref.29 Ref.31
□	Glycosylation	249	1	N-linked (Glc) (glycation); in vitro Ref.26
□	Glycosylation	257	1	N-linked (Glc) (glycation) Probable
□	Glycosylation	300	1	N-linked (Glc) (glycation); in vitro Ref.26
□	Glycosylation	305	1	N-linked (Glc) (glycation) Ref.31
□	Glycosylation	337	1	N-linked (Glc) (glycation); in vitro Ref.26
□	Glycosylation	341	1	N-linked (Glc) (glycation) Probable
□	Glycosylation	342	1	N-linked (GlcNAc...); in variant Redhill
□	Glycosylation	347	1	N-linked (Glc) (glycation); in vitro Ref.26
□	Glycosylation	375	1	N-linked (Glc) (glycation) Probable
□	Glycosylation	402	1	N-linked (Glc) (glycation); in vitro Ref.26
□	Glycosylation	437	1	N-linked (Glc) (glycation); in vitro Ref.26
□	Glycosylation	463	1	N-linked (Glc) (glycation) Ref.31
□	Glycosylation	468	1	N-linked (Glc) (glycation); in vitro Ref.26
□	Glycosylation	518	1	N-linked (GlcNAc...); in variant Casebrook
□	Glycosylation	549	1	N-linked (Glc) (glycation) Ref.26 Ref.29 Ref.30 Ref.31
□	Glycosylation	558	1	N-linked (Glc) (glycation) Probable
□	Glycosylation	560	1	N-linked (Glc) (glycation); in vitro Ref.26
□	Glycosylation	569	1	N-linked (Glc) (glycation); in vitro Ref.26
□	Glycosylation	597	1	N-linked (Glc) (glycation); in vitro Ref.26

Getting started

Text search — Our basic text search allows y search all the resources avail

BLAST — Find regions of similarity betw sequences

Sequence alignments — Align two or more protein sequences using the Clustal Omega program

Retrieve/ID mapping

knowledgebase

Submit your data — Submit your sequences updates

http://www.rcsb.org/pdb/home/hc

Deposit ▾ Search ▾ Visua

RCSB PDB PROTEIN DATA BA

Search by PDB ID, author, m

Advanced Search | Browse by Ann

PDB-101 wwPDB

Welcome

Sequence & Structure Align

Sequence & Str

RCSB PDB's Comparison T Needleman-Wunsch, and S (FATCAT, CE, Mammoth, T

Comparisons can be made customized or local files not for both rigid-body and flexi permutations.

Enter First PDB ID ⟷

Select Associated Chain ID … Select Associated Chain ID …

- Select Comparison Method - Align More Options

Latest Entries New Features News Publications ▾ Feed

About ▾ Contact us

Search

Highlighted GO term

Representing "phases" in GO biological process

The GOC has recently introduced a new term biological phase (GO:0044848), as a direct subclass of biological process. This class represents a distinct period or stage during which biological processes can occur. more

Random FAQs
- How do I map a set of annotations to high level GO terms (GO slim)?
- Where can I view or download the complete sets of GO annotations?
- What is a GAF file?

View all FAQs

Tweets — Laurent Deluc @laurentdeluc1 22 Oct — Is there an issue with the server to

ype Plugin

ntology?
- An introduction to the Gene Ontology
- What are annotations?
- Ten quick tips for using the Gene Ontology Important
- Gene Ontology tools
- Enrichment analysis
- Downloads

Powered by PANTHER

Statistics

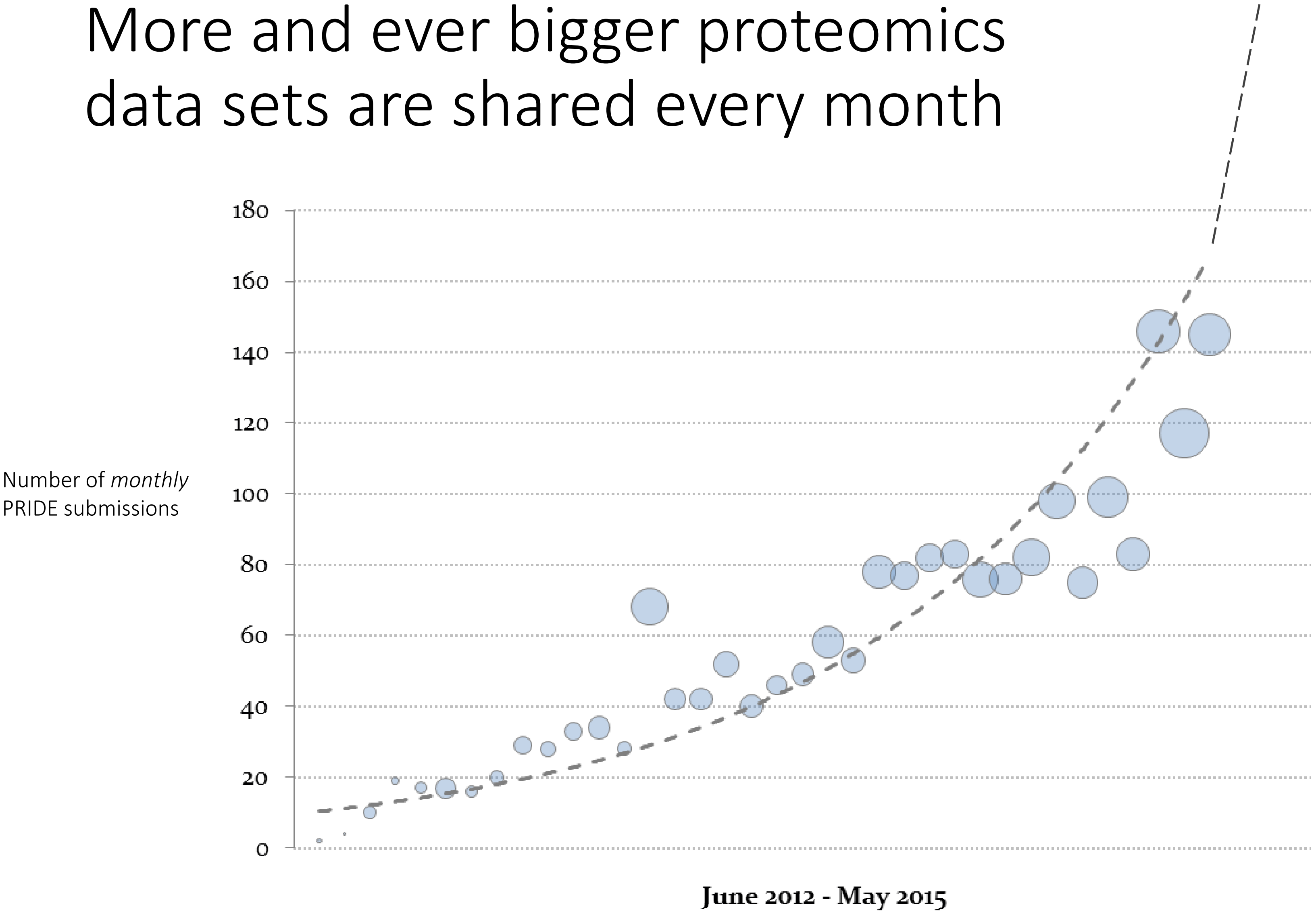
Recent news

Ten Quick Tips for Using the Gene Ontology Important

Post date: 11/26/2013 - 08:22

Parkinson's Gene Annotation Project Newsletter October 2015

More and ever bigger proteomics data sets are shared every month



2019: ~300 monthly

Vaudel, Proteomics, 2015

REVIEW

Exploring the potential of public proteomics data

Marc Vaudel[1], Kenneth Verheggen[2,3,4], Attila Csordas[5], Helge Ræder[6], Frode S. Berven[1,7], Lennart Martens[2,3,4], Juan A. Vizcaíno[5] and Harald Barsnes[1,6]*

[1] Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway
[2] Medical Biotechnology Center, VIB, Ghent, Belgium
[3] Department of Biochemistry, Ghent University, Ghent, Belgium
[4] Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium
[5] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[6] Department of Clinical Science, KG Jebsen Center for Diabetes Research, University of Bergen, Bergen, Norway
[7] Department of Clinical Medicine, KG Jebsen Centre for Multiple Sclerosis Research, University of Bergen, Bergen, Norway

In a global effort for scientific transparency, it has become feasible and good practice to share experimental data supporting novel findings. Consequently, the amount of publicly available MS-based proteomics data has grown substantially in recent years. With some notable exceptions, this extensive material has however largely been left untouched. The time has now come for the proteomics community to utilize this potential gold mine for new discoveries, and uncover its untapped potential. In this review, we provide a brief history of the sharing of proteomics data, showing ways in which publicly available proteomics data are already being (re-)used, and outline potential future opportunities based on four different usage types: use, reuse, reprocess, and repurpose. We thus aim to assist the proteomics community in stepping up to the challenge, and to make the most of the rapidly increasing amount of public proteomics data.

Received: July 13, 2015
Revised: August 25, 2015
Accepted: September 28, 2015

Keywords:
Bioinformatics / Computational proteomics / Data analysis / Databases / Data standards

1 Introduction

1.1 Background

Historically, a large proportion of the proteomics community was reticent to openly share the data they produced. However, the sharing of not only the knowledge obtained through proteomics experiments (through scientific publications), but also of the underlying data, has increasingly become standard practice, and is now even mandatory or strongly advised in many of the relevant scientific journals [1–3]. In addition, a number of funders (e.g. the Wellcome Trust and the NIH) are enforcing the public deposition of data from projects they fund as a way to maximize the value of the funds provided. As a result, the amount of publicly shared MS-based proteomics data has grown substantially, both in terms of number of submission and total data amount, as illustrated in Fig. 1.

Two key factors strongly contributed to this success: first, the sharing of the data has become much easier with the development of user-friendly tools and infrastructure; and second, the proteomics community, triggered by scientific journals and funders, has now agreed that it is good scientific practice to make the underlying data available when publishing novel findings.

There were several challenges to overcome in order to get to this point, see Fig. 2. The first of these challenges was the need for central and long-term public repositories to store the generated data. Several such generic repositories are now

Correspondence: Dr. Harald Barsnes, Proteomics Unit, Department of Biomedicine, University of Bergen, Jonas Liesvei 91, N-5009 Bergen, Norway
E-mail: harald.barsnes@uib.no
Fax: +47-55-58-63-60

Abbreviation: PSM, peptide to spectrum match

*Additional corresponding author: Dr. Juan A. Vizcaíno
E-mail: juan@ebi.ac.uk

Colour Online: See the article online to view Figs. 1–4 in colour.

How can we use the shared data?

1) Verify published findings

2) Reuse existing data or knowledge

3) Generate new knowledge

Dinosaur proteomics..?

Protein Sequences from Mastodon and *Tyrannosaurus Rex* Revealed by Mass Spectrometry

John M. Asara,[1,2]* Mary H. Schweitzer,[3] Lisa M. Freimark,[1] Matthew Phillips,[1] Lewis C. Cantley[1,4]

Interpreting Sequences from Mastodon and *T. rex*

J. ASARA *ET AL.* REPORTED THAT COLLAGEN proteins from well-preserved ancient fossil bones from a 160,000- to 600,000-year-old mastodon and a 68-million-year-old *T. rex* can be extracted and sequenced ("Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry," 13 April, p. 280). Tandem mass spectrometry (MS/MS) is an effective sequencing method for ancient fossils when DNA is not available. It has come to the original authors' attention that there are concerns regarding the reported sequences containing glycine (G) hydroxylation, as well as some positions of proline (P) hydroxylation. Although nonstandard postmortem

reported to be unstable (2, 3).

Ion trap mass spectrometers scan very fast and are highly sensitive but cannot resolve amino acids or combinations of modifications and amino acids that are near isobaric (same nominal mass), as stated in the original Report. It is sometimes difficult to determine the precise position of a modification from adjacent or nearby amino acid residues, since MS/MS spectra often lack sufficient site-specific fragment ions (*4*).

Hydroxylation of P to 4-hydroxyproline is a highly abundant modification that stabilizes the triple helical structure of collagen. Hydroxylation also occurs to a lesser extent on lysine (K) residues (*5*, *6*). In type I and type II collagens, these hydroxylation sites have been reported to exist nearly exclusively for P or K in the Y position of the collagen triplet repeat –GXY- (*7*, *8*). A singular exception, one P in human collagen I and II, is X position hydrox-

Asara *et al.* (2007) Science 316: 280-5.
Asara *et al.* (2007) Science 316: 1324-5.
Bern *et al.* (2009) JPR 9: 4328-32

research articles Journal of proteome research

Reanalysis of *Tyrannosaurus rex* Mass Spectra

Marshall Bern,*,† Brett S. Phinney,‡ and David Goldberg†

Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, California 94304, and Genome Center, University of California at Davis, Davis, California 95616

Received April 16, 2009

Asara et al. reported the detection of collagen peptides in a 68-million-year-old *Tyrannosaurus rex* bone by shotgun proteomics. This finding has been called into question as a possible statistical artifact. We reanalyze Asara et al.'s tandem mass spectra using a different search engine and different statistical tools. Our reanalysis shows a sample containing common laboratory contaminants, soil bacteria, and bird-like hemoglobin and collagen.

RESEARCH PROFILE

Independent analysis of controversial *T. rex* data confirms findings

A recent *JPR* paper made John Asara's day. The researcher at Harvard Medical School and Beth Israel Deaconess Medical Center and his collaborator, Mary Schweitzer of North Carolina State University, have been embroiled in a controversy over a 68-million-year-old *Tyrannosaurus rex*. In 2007, they published data that indicated the dinosaur's bones contained collagen that closely matched that of birds (*Science* **2007**, DOI 10.1126/science.1137614). But their study was heavily criticized on several fronts, including the accusation that peptide matches to their MS data were statistically insignificant (see the *Analytical Chemistry* news story "Uproar over dinosaur data").

The *JPR* paper by Marshall Bern and colleagues at the Palo Alto Research Center, Inc., and the University of California Davis is the first independent

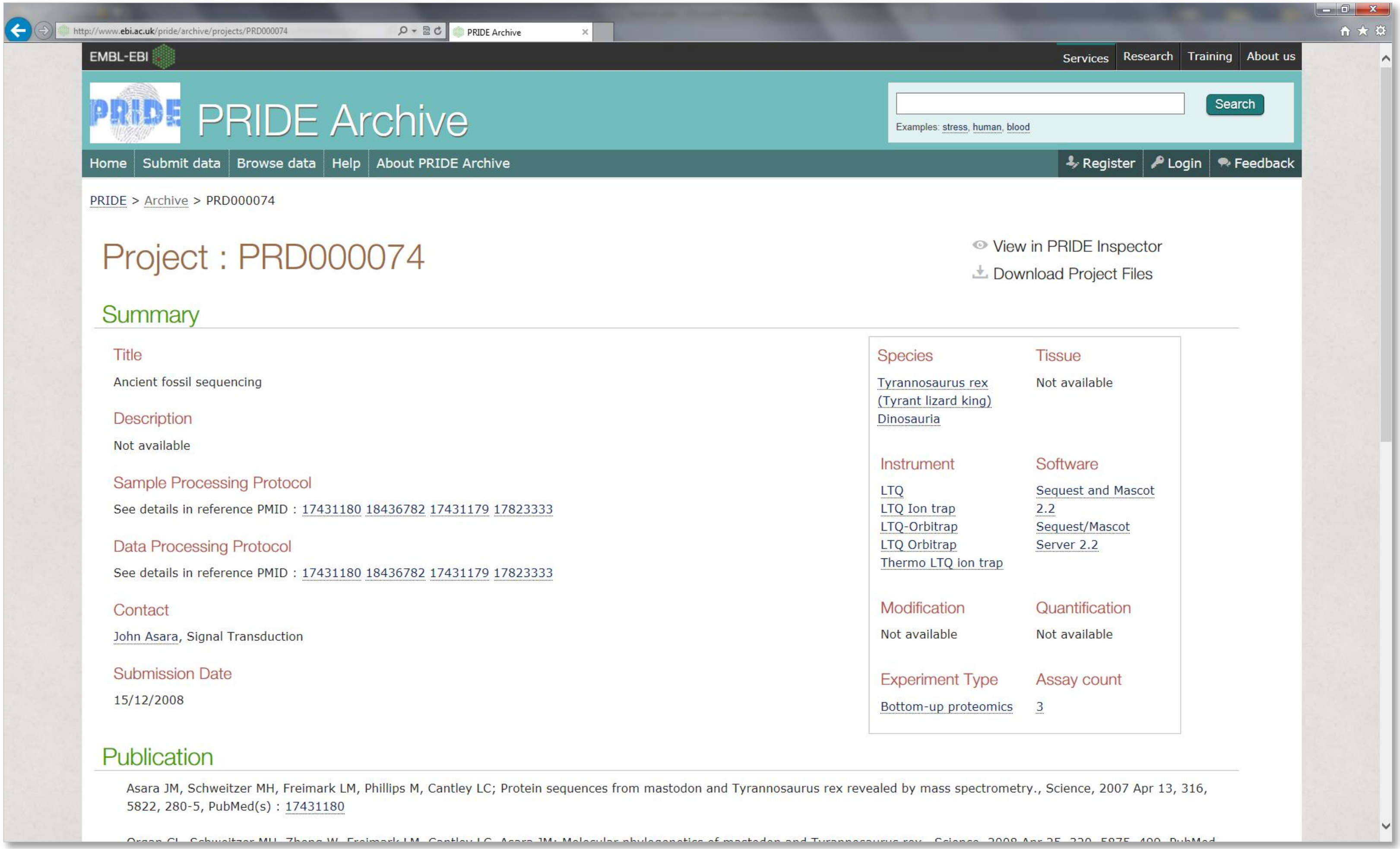
Presence of birdlike collagen and hemoglobin peptides have been confirmed by a second group of researchers in the controversial *T. rex* data set.

bolstering their analysis. In September 2008, Asara released only the *T. rex* spectra data set into the PRIDE database. He didn't release the control spectra from the soil sediment in the vicinity of the *T. rex* fossil (these events are chronicled in the *JPR* news story "A controversial data set stirs up even more controversy"). But Asara gave Bern and his team the

But 0.4
for Mas
Onic gi
matche
Of th
the Asa
GenBar
Consor
confirm
made t

PRIDE:
Project: PRD000074
Assay: accession 8633

Dinosaur proteomics..?   II



Project : PRD000074

Summary

Title

Ancient fossil sequencing

Description

Not available

Sample Processing Protocol

See details in reference PMID : 17431180 18436782 17431179 17823333

Data Processing Protocol

See details in reference PMID : 17431180 18436782 17431179 17823333

Contact

John Asara, Signal Transduction

Submission Date

15/12/2008

Species	Tissue
Tyrannosaurus rex (Tyrant lizard king) Dinosauria	Not available

Instrument	Software
LTQ LTQ Ion trap LTQ-Orbitrap LTQ Orbitrap Thermo LTQ ion trap	Sequest and Mascot 2.2 Sequest/Mascot Server 2.2

Modification	Quantification
Not available	Not available

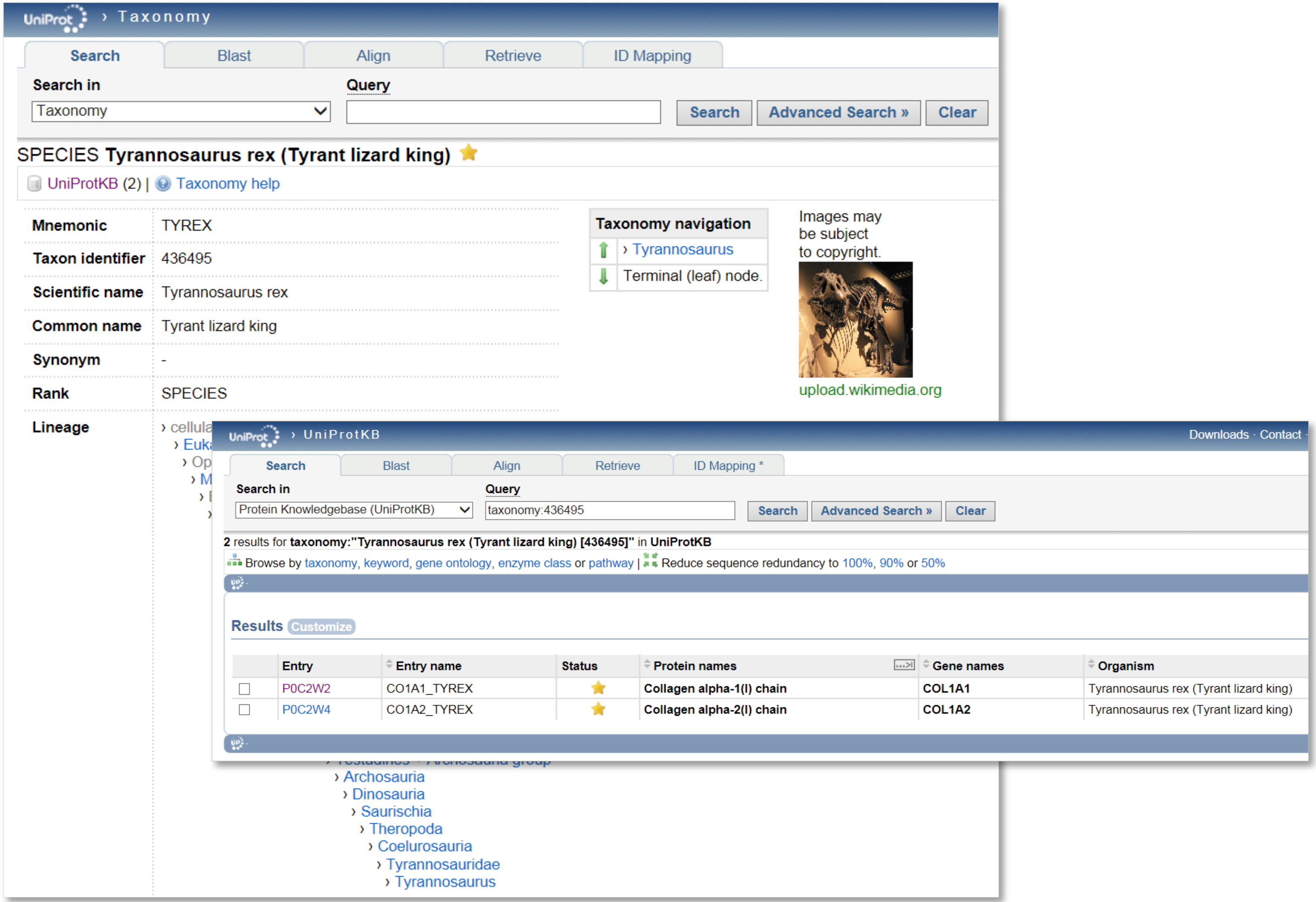
Experiment Type	Assay count
Bottom-up proteomics	3

Publication

Asara JM, Schweitzer MH, Freimark LM, Phillips M, Cantley LC; Protein sequences from mastodon and Tyrannosaurus rex revealed by mass spectrometry., Science, 2007 Apr 13, 316, 5822, 280-5, PubMed(s) : 17431180

Organ CL, Schweitzer MH, Zheng W, Freimark LM, Cantley LC, Asara JM; Molecular phylogenetics of mastodon and Tyrannosaurus rex., Science, 2008 Apr 25, 320, 5875, 499, PubMed

Dinosaur proteomics..?   III



UniProt › Taxonomy

Search | Blast | Align | Retrieve | ID Mapping

Search in: Taxonomy

Query:

Search | Advanced Search » | Clear

SPECIES **Tyrannosaurus rex (Tyrant lizard king)** ★

UniProtKB (2) | Taxonomy help

Mnemonic	TYREX
Taxon identifier	436495
Scientific name	Tyrannosaurus rex
Common name	Tyrant lizard king
Synonym	-
Rank	SPECIES
Lineage	› cellula › Euk › Op › M › I › › Testudines - Archosauria group › Archosauria › Dinosauria › Saurischia › Theropoda › Coelurosauria › Tyrannosauridae › Tyrannosaurus

Taxonomy navigation

⬆	› Tyrannosaurus
⬇	Terminal (leaf) node.

Images may be subject to copyright.

upload.wikimedia.org

UniProt › UniProtKB

Downloads · Contact ·

Search | Blast | Align | Retrieve | ID Mapping *

Search in: Protein Knowledgebase (UniProtKB)

Query: taxonomy:436495

Search | Advanced Search » | Clear

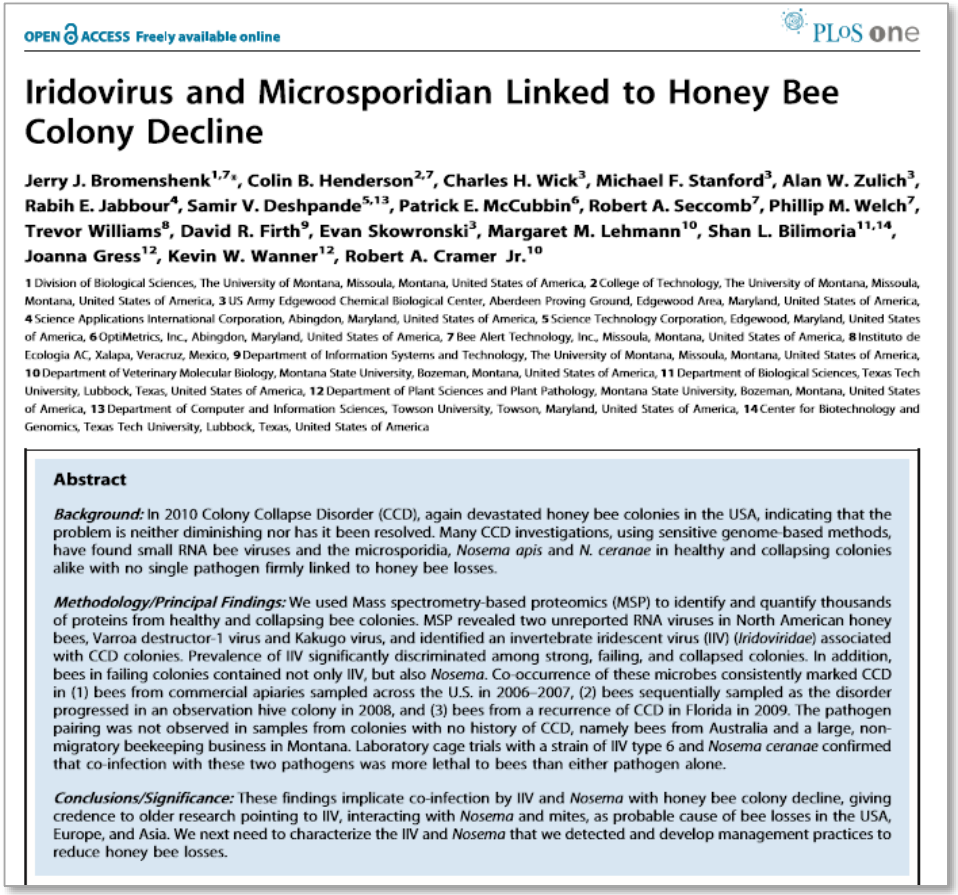
2 results for **taxonomy:"Tyrannosaurus rex (Tyrant lizard king) [436495]"** in **UniProtKB**

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50%

Results Customize

	Entry	Entry name	Status	Protein names	Gene names	Organism
□	P0C2W2	CO1A1_TYREX	★	**Collagen alpha-1(I) chain**	**COL1A1**	Tyrannosaurus rex (Tyrant lizard king)
□	P0C2W4	CO1A2_TYREX	★	**Collagen alpha-2(I) chain**	**COL1A2**	Tyrannosaurus rex (Tyrant lizard king)

Honey bee virus..?



OPEN ACCESS Freely available online PLoS one

Iridovirus and Microsporidian Linked to Honey Bee Colony Decline

Jerry J. Bromenshenk[1,7]*, Colin B. Henderson[2,7], Charles H. Wick[3], Michael F. Stanford[3], Alan W. Zulich[3], Rabih E. Jabbour[4], Samir V. Deshpande[5,13], Patrick E. McCubbin[6], Robert A. Seccomb[7], Phillip M. Welch[7], Trevor Williams[8], David R. Firth[9], Evan Skowronski[3], Margaret M. Lehmann[10], Shan L. Bilimoria[11,14], Joanna Gress[12], Kevin W. Wanner[12], Robert A. Cramer Jr.[10]

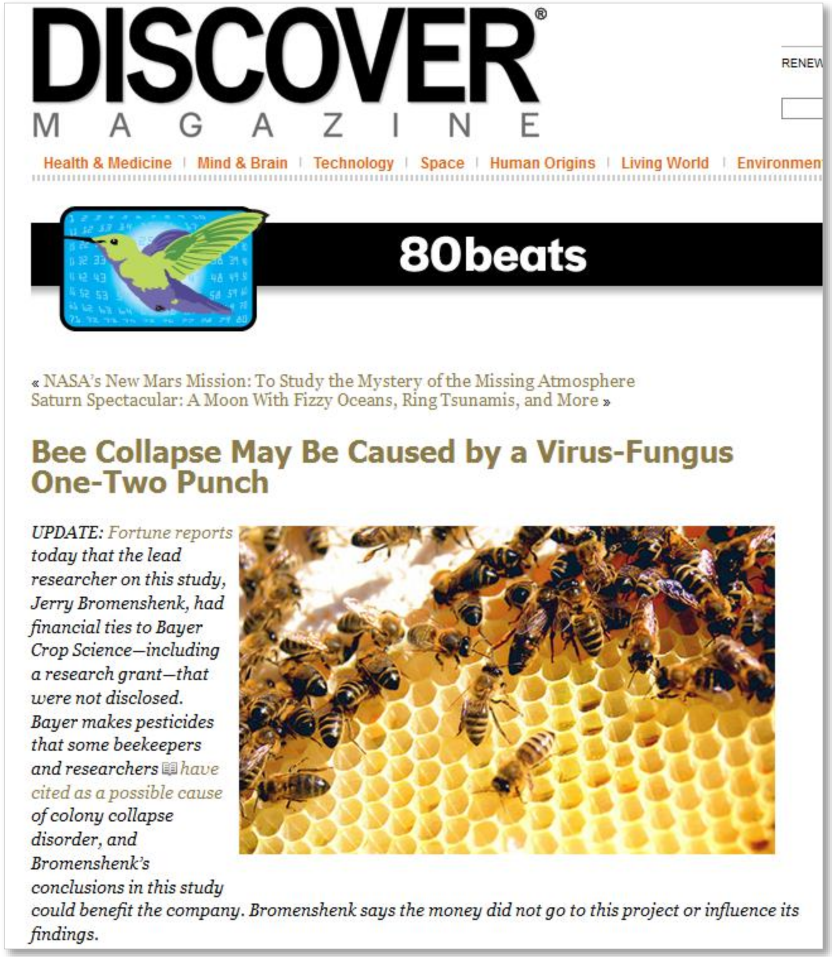
1 Division of Biological Sciences, The University of Montana, Missoula, Montana, United States of America, **2** College of Technology, The University of Montana, Missoula, Montana, United States of America, **3** US Army Edgewood Chemical Biological Center, Aberdeen Proving Ground, Edgewood Area, Maryland, United States of America, **4** Science Applications International Corporation, Abingdon, Maryland, United States of America, **5** Science Technology Corporation, Edgewood, Maryland, United States of America, **6** OptiMetrics, Inc., Abingdon, Maryland, United States of America, **7** Bee Alert Technology, Inc., Missoula, Montana, United States of America, **8** Instituto de Ecologia AC, Xalapa, Veracruz, Mexico, **9** Department of Information Systems and Technology, The University of Montana, Missoula, Montana, United States of America, **10** Department of Veterinary Molecular Biology, Montana State University, Bozeman, Montana, United States of America, **11** Department of Biological Sciences, Texas Tech University, Lubbock, Texas, United States of America, **12** Department of Plant Sciences and Plant Pathology, Montana State University, Bozeman, Montana, United States of America, **13** Department of Computer and Information Sciences, Towson University, Towson, Maryland, United States of America, **14** Center for Biotechnology and Genomics, Texas Tech University, Lubbock, Texas, United States of America

Abstract

Background: In 2010 Colony Collapse Disorder (CCD), again devastated honey bee colonies in the USA, indicating that the problem is neither diminishing nor has it been resolved. Many CCD investigations, using sensitive genome-based methods, have found small RNA bee viruses and the microsporidia, *Nosema apis* and *N. ceranae* in healthy and collapsing colonies alike with no single pathogen firmly linked to honey bee losses.

Methodology/Principal Findings: We used Mass spectrometry-based proteomics (MSP) to identify and quantify thousands of proteins from healthy and collapsing bee colonies. MSP revealed two unreported RNA viruses in North American honey bees, Varroa destructor-1 virus and Kakugo virus, and identified an invertebrate iridescent virus (IIV) (*Iridoviridae*) associated with CCD colonies. Prevalence of IIV significantly discriminated among strong, failing, and collapsed colonies. In addition, bees in failing colonies contained not only IIV, but also *Nosema*. Co-occurrence of these microbes consistently marked CCD in (1) bees from commercial apiaries sampled across the U.S. in 2006–2007, (2) bees sequentially sampled as the disorder progressed in an observation hive colony in 2008, and (3) bees from a recurrence of CCD in Florida in 2009. The pathogen pairing was not observed in samples from colonies with no history of CCD, namely bees from Australia and a large, non-migratory beekeeping business in Montana. Laboratory cage trials with a strain of IIV type 6 and *Nosema ceranae* confirmed that co-infection with these two pathogens was more lethal to bees than either pathogen alone.

Conclusions/Significance: These findings implicate co-infection by IIV and *Nosema* with honey bee colony decline, giving credence to older research pointing to IIV, interacting with *Nosema* and mites, as probable cause of bee losses in the USA, Europe, and Asia. We next need to characterize the IIV and *Nosema* that we detected and develop management practices to reduce honey bee losses.



DISCOVER® MAGAZINE

RENEW

Health & Medicine | Mind & Brain | Technology | Space | Human Origins | Living World | Environmen

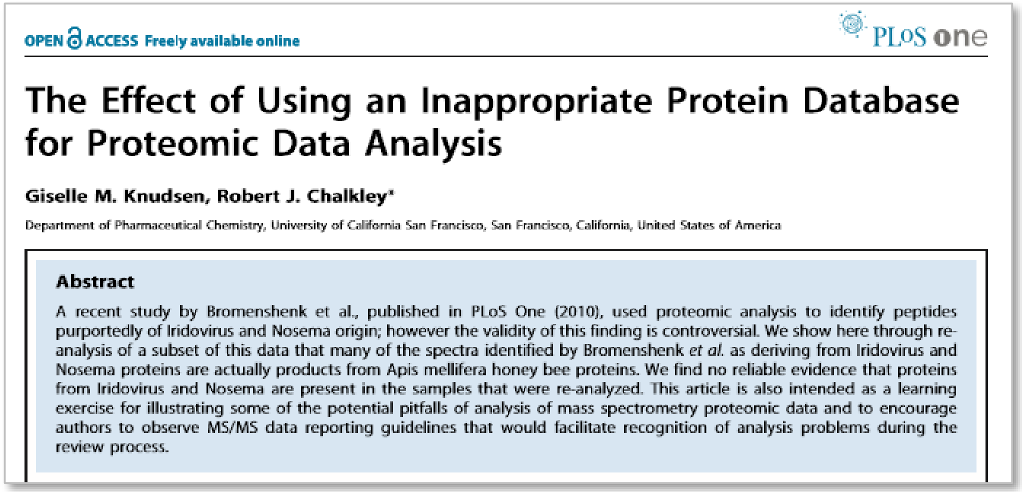
80beats

« NASA's New Mars Mission: To Study the Mystery of the Missing Atmosphere
Saturn Spectacular: A Moon With Fizzy Oceans, Ring Tsunamis, and More »

Bee Collapse May Be Caused by a Virus-Fungus One-Two Punch

UPDATE: Fortune reports today that the lead researcher on this study, Jerry Bromenshenk, had financial ties to Bayer Crop Science—including a research grant—that were not disclosed. Bayer makes pesticides that some beekeepers and researchers have cited as a possible cause of colony collapse disorder, and Bromenshenk's conclusions in this study could benefit the company. Bromenshenk says the money did not go to this project or influence its findings.

Honey bee virus..?   II



OPEN ACCESS Freely available online   PLoS one

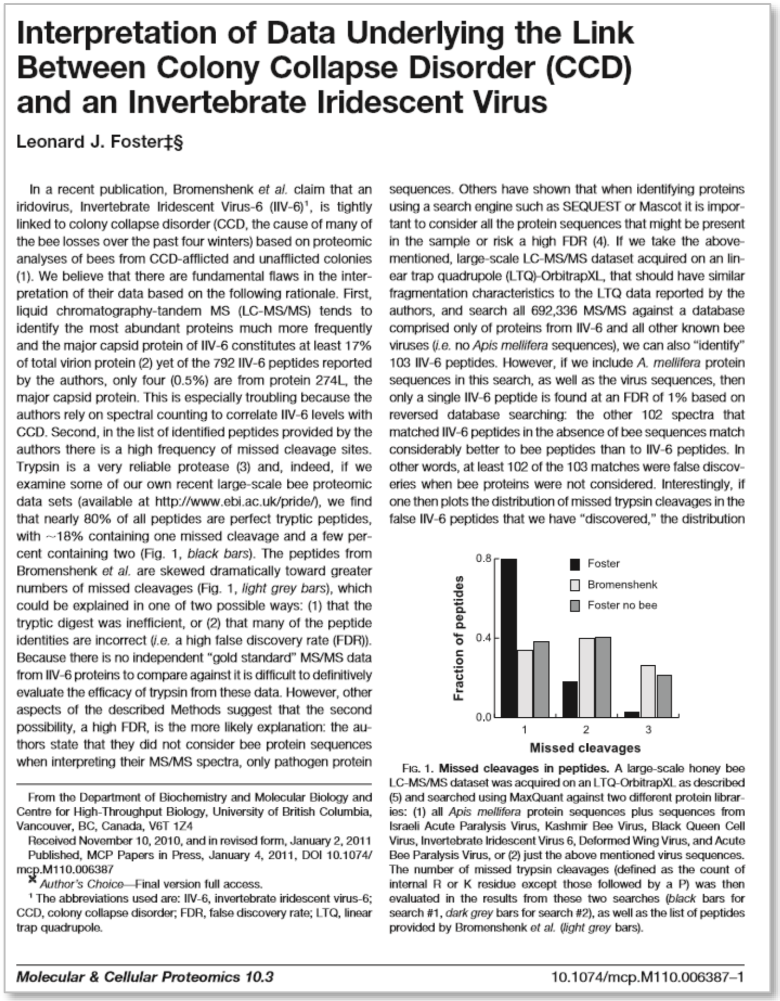
The Effect of Using an Inappropriate Protein Database for Proteomic Data Analysis

Giselle M. Knudsen, Robert J. Chalkley*

Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, California, United States of America

Abstract

A recent study by Bromenshenk et al., published in PLoS One (2010), used proteomic analysis to identify peptides purportedly of Iridovirus and Nosema origin; however the validity of this finding is controversial. We show here through re-analysis of a subset of this data that many of the spectra identified by Bromenshenk *et al.* as deriving from Iridovirus and Nosema proteins are actually products from Apis mellifera honey bee proteins. We find no reliable evidence that proteins from Iridovirus and Nosema are present in the samples that were re-analyzed. This article is also intended as a learning exercise for illustrating some of the potential pitfalls of analysis of mass spectrometry proteomic data and to encourage authors to observe MS/MS data reporting guidelines that would facilitate recognition of analysis problems during the review process.



Interpretation of Data Underlying the Link Between Colony Collapse Disorder (CCD) and an Invertebrate Iridescent Virus

Leonard J. Foster‡§

In a recent publication, Bromenshenk *et al.* claim that an iridovirus, Invertebrate Iridescent Virus-6 (IIV-6)[1], is tightly linked to colony collapse disorder (CCD, the cause of many of the bee losses over the past four winters) based on proteomic analyses of bees from CCD-afflicted and unafflicted colonies (1). We believe that there are fundamental flaws in the interpretation of their data based on the following rationale. First, liquid chromatography-tandem MS (LC-MS/MS) tends to identify the most abundant proteins much more frequently and the major capsid protein of IIV-6 constitutes at least 17% of total virion protein (2) yet of the 792 IIV-6 peptides reported by the authors, only four (0.5%) are from protein 274L, the major capsid protein. This is especially troubling because the authors rely on spectral counting to correlate IIV-6 levels with CCD. Second, in the list of identified peptides provided by the authors there is a high frequency of missed cleavage sites. Trypsin is a very reliable protease (3) and, indeed, if we examine some of our own recent large-scale bee proteomic data sets (available at http://www.ebi.ac.uk/pride/), we find that nearly 80% of all peptides are perfect tryptic peptides, with ~18% containing one missed cleavage and a few percent containing two (Fig. 1, *black bars*). The peptides from Bromenshenk *et al.* are skewed dramatically toward greater numbers of missed cleavages (Fig. 1, *light grey bars*), which could be explained in one of two possible ways: (1) that the tryptic digest was inefficient, or (2) that many of the peptide identities are incorrect (*i.e.* a high false discovery rate (FDR)). Because there is no independent "gold standard" MS/MS data from IIV-6 proteins to compare against it is difficult to definitively evaluate the efficacy of trypsin from these data. However, other aspects of the described Methods suggest that the second possibility, a high FDR, is the more likely explanation: the authors state that they did not consider bee protein sequences when interpreting their MS/MS spectra, only pathogen protein sequences. Others have shown that when identifying proteins using a search engine such as SEQUEST or Mascot it is important to consider all the protein sequences that might be present in the sample or risk a high FDR (4). If we take the above-mentioned, large-scale LC-MS/MS dataset acquired on an linear trap quadrupole (LTQ)-OrbitrapXL, that should have similar fragmentation characteristics to the LTQ data reported by the authors, and search all 692,336 MS/MS against a database comprised only of proteins from IIV-6 and all other known bee viruses (*i.e.* no *Apis mellifera* sequences), we can also "identify" 103 IIV-6 peptides. However, if we include *A. mellifera* protein sequences in this search, as well as the virus sequences, then only a single IIV-6 peptide is found at an FDR of 1% based on reversed database searching: the other 102 spectra that matched IIV-6 peptides in the absence of bee sequences match considerably better to bee peptides than to IIV-6 peptides. In other words, at least 102 of the 103 matches were false discoveries when bee proteins were not considered. Interestingly, if one then plots the distribution of missed trypsin cleavages in the false IIV-6 peptides that we have "discovered," the distribution

FIG. 1. **Missed cleavages in peptides.** A large-scale honey bee LC-MS/MS dataset was acquired on an LTQ-OrbitrapXL as described (5) and searched using MaxQuant against two different protein libraries: (1) all *Apis mellifera* protein sequences plus sequences from Israeli Acute Paralysis Virus, Kashmir Bee Virus, Black Queen Cell Virus, Invertebrate Iridescent Virus 6, Deformed Wing Virus, and Acute Bee Paralysis Virus, or (2) just the above mentioned virus sequences. The number of missed trypsin cleavages (defined as the count of internal R or K residue except those followed by a P) was then evaluated in the results from these two searches (*black* bars for search #1, *dark grey* bars for search #2), as well as the list of peptides provided by Bromenshenk *et al.* (*light grey* bars).

From the Department of Biochemistry and Molecular Biology and Centre for High-Throughput Biology, University of British Columbia, Vancouver, BC, Canada, V6T 1Z4

Received November 10, 2010, and in revised form, January 2, 2011

Published, MCP Papers in Press, January 4, 2011, DOI 10.1074/mcp.M110.006387

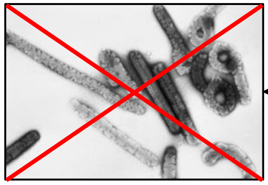
✖ *Author's Choice*—Final version full access.

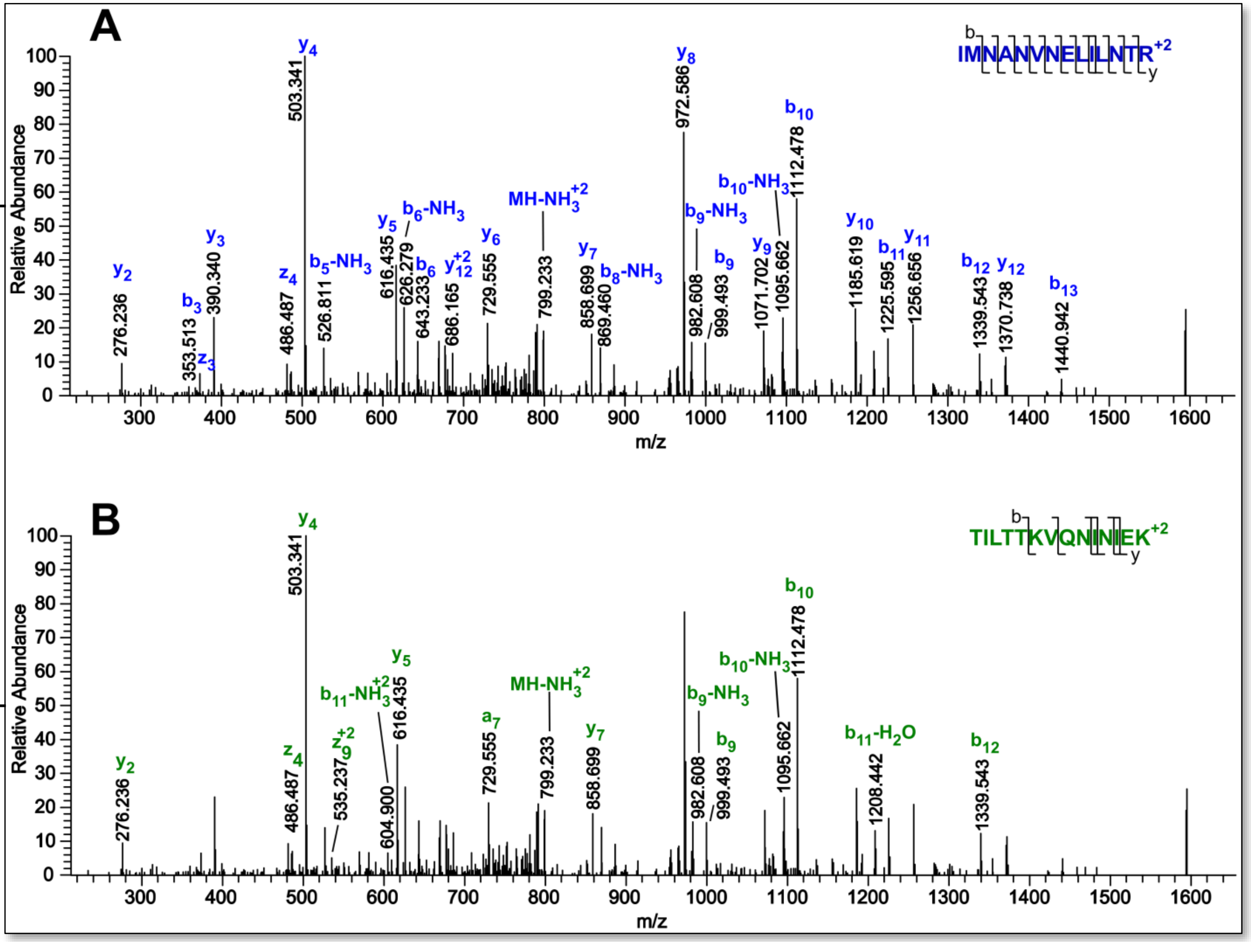
[1] The abbreviations used are: IIV-6, invertebrate iridescent virus-6; CCD, colony collapse disorder; FDR, false discovery rate; LTQ, linear trap quadrupole.

Molecular & Cellular Proteomics 10.3   10.1074/mcp.M110.006387–1

- **Big pitfall**: Search database composed of <u>only</u> virus proteins, i.e. <u>No honey bee proteins at all!</u>

Honey bee virus..?   III





Knudsen and Chalkley, PLoS One, 2011

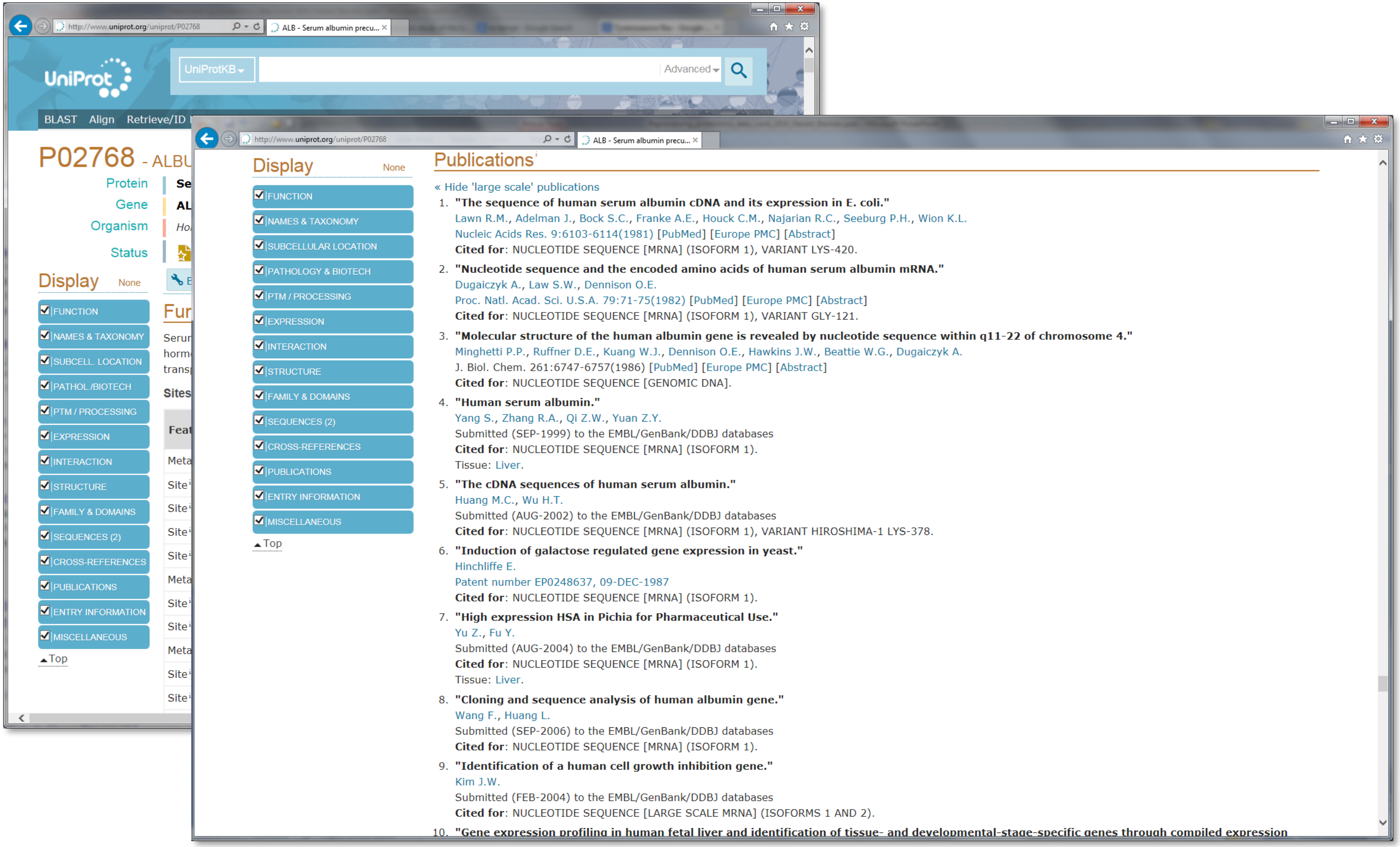
How can we use the shared data?

1) Verify published findings

2) **Reuse existing data or knowledge**

3) Generate new knowledge

Proteomics/protein databases

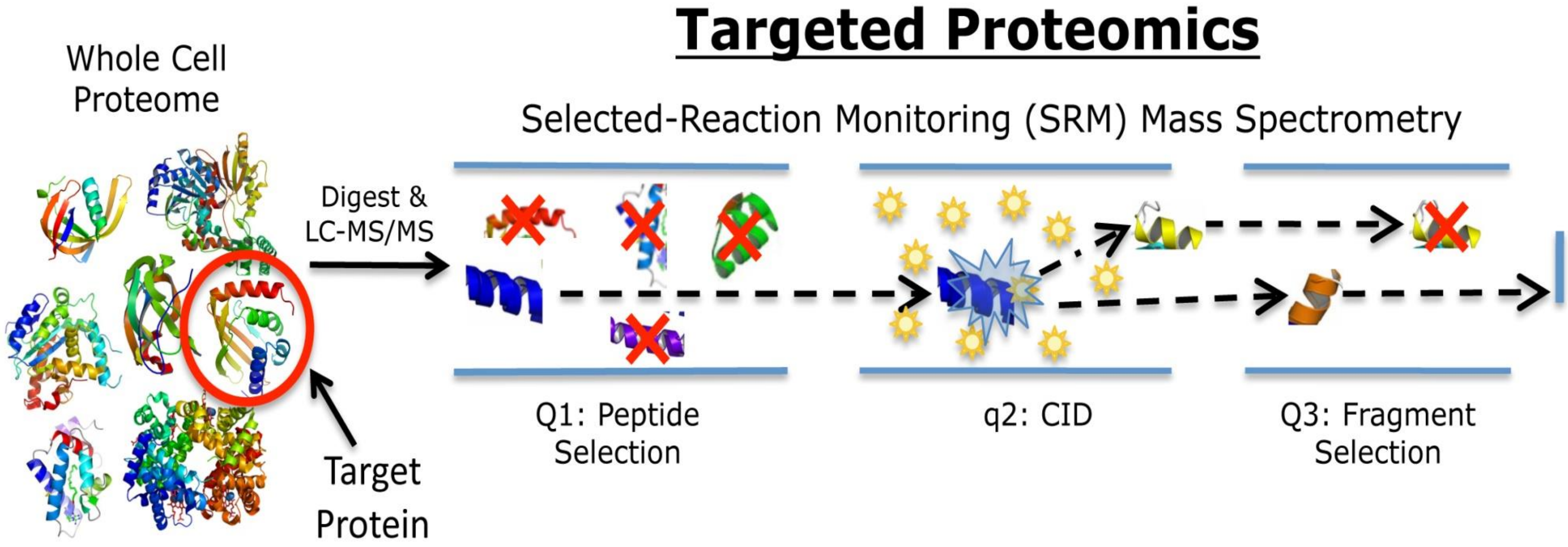


Publications

« Hide 'large scale' publications

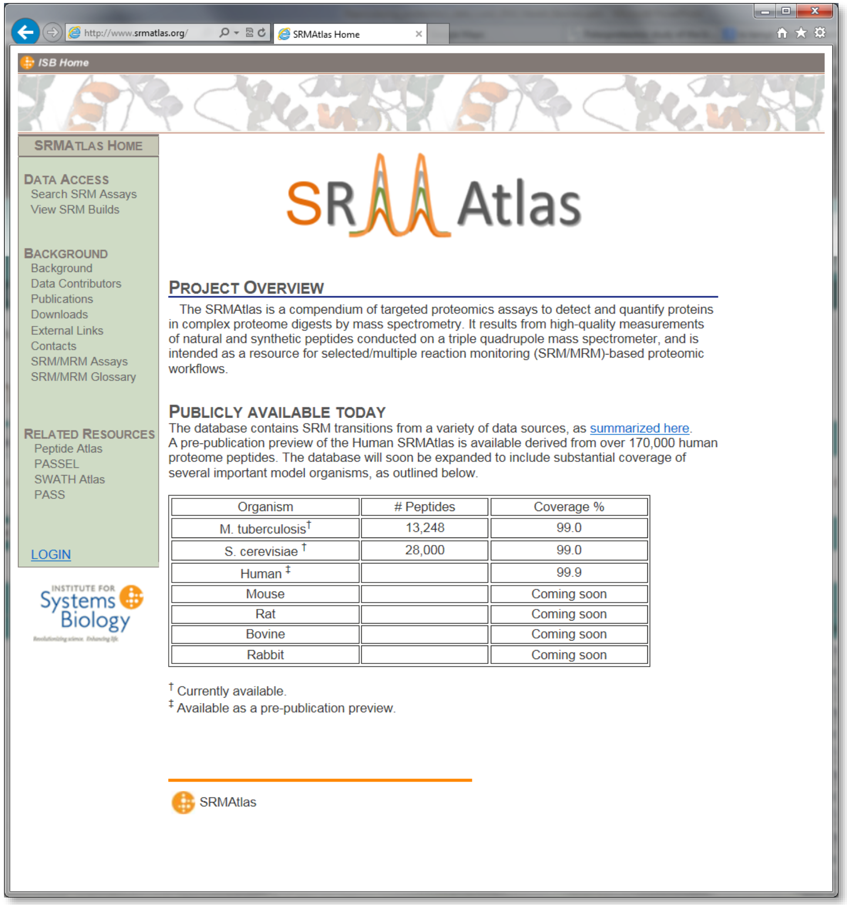
1. **"The sequence of human serum albumin cDNA and its expression in E. coli."**
   Lawn R.M., Adelman J., Bock S.C., Franke A.E., Houck C.M., Najarian R.C., Seeburg P.H., Wion K.L.
   Nucleic Acids Res. 9:6103-6114(1981) [PubMed] [Europe PMC] [Abstract]
   Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1), VARIANT LYS-420.
2. **"Nucleotide sequence and the encoded amino acids of human serum albumin mRNA."**
   Dugaiczyk A., Law S.W., Dennison O.E.
   Proc. Natl. Acad. Sci. U.S.A. 79:71-75(1982) [PubMed] [Europe PMC] [Abstract]
   Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1), VARIANT GLY-121.
3. **"Molecular structure of the human albumin gene is revealed by nucleotide sequence within q11-22 of chromosome 4."**
   Minghetti P.P., Ruffner D.E., Kuang W.J., Dennison O.E., Hawkins J.W., Beattie W.G., Dugaiczyk A.
   J. Biol. Chem. 261:6747-6757(1986) [PubMed] [Europe PMC] [Abstract]
   Cited for: NUCLEOTIDE SEQUENCE [GENOMIC DNA].
4. **"Human serum albumin."**
   Yang S., Zhang R.A., Qi Z.W., Yuan Z.Y.
   Submitted (SEP-1999) to the EMBL/GenBank/DDBJ databases
   Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
   Tissue: Liver.
5. **"The cDNA sequences of human serum albumin."**
   Huang M.C., Wu H.T.
   Submitted (AUG-2002) to the EMBL/GenBank/DDBJ databases
   Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1), VARIANT HIROSHIMA-1 LYS-378.
6. **"Induction of galactose regulated gene expression in yeast."**
   Hinchliffe E.
   Patent number EP0248637, 09-DEC-1987
   Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
7. **"High expression HSA in Pichia for Pharmaceutical Use."**
   Yu Z., Fu Y.
   Submitted (AUG-2004) to the EMBL/GenBank/DDBJ databases
   Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
   Tissue: Liver.
8. **"Cloning and sequence analysis of human albumin gene."**
   Wang F., Huang L.
   Submitted (SEP-2006) to the EMBL/GenBank/DDBJ databases
   Cited for: NUCLEOTIDE SEQUENCE [MRNA] (ISOFORM 1).
9. **"Identification of a human cell growth inhibition gene."**
   Kim J.W.
   Submitted (FEB-2004) to the EMBL/GenBank/DDBJ databases
   Cited for: NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA] (ISOFORMS 1 AND 2).
10. **"Gene expression profiling in human fetal liver and identification of tissue- and developmental-stage-specific genes through compiled expression**

Reusing targeted proteomics data



http://newscenter.lbl.gov/wp-content/uploads/Petzold-Targeted-Proteomics.jpg

Reusing targeted proteomics data II



SRMATLAS HOME

DATA ACCESS
Search SRM Assays
View SRM Builds

BACKGROUND
Background
Data Contributors
Publications
Downloads
External Links
Contacts
SRM/MRM Assays
SRM/MRM Glossary

RELATED RESOURCES
Peptide Atlas
PASSEL
SWATH Atlas
PASS

LOGIN

Institute for Systems Biology

SRMAtlas

PROJECT OVERVIEW

The SRMAtlas is a compendium of targeted proteomics assays to detect and quantify proteins in complex proteome digests by mass spectrometry. It results from high-quality measurements of natural and synthetic peptides conducted on a triple quadrupole mass spectrometer, and is intended as a resource for selected/multiple reaction monitoring (SRM/MRM)-based proteomic workflows.

PUBLICLY AVAILABLE TODAY

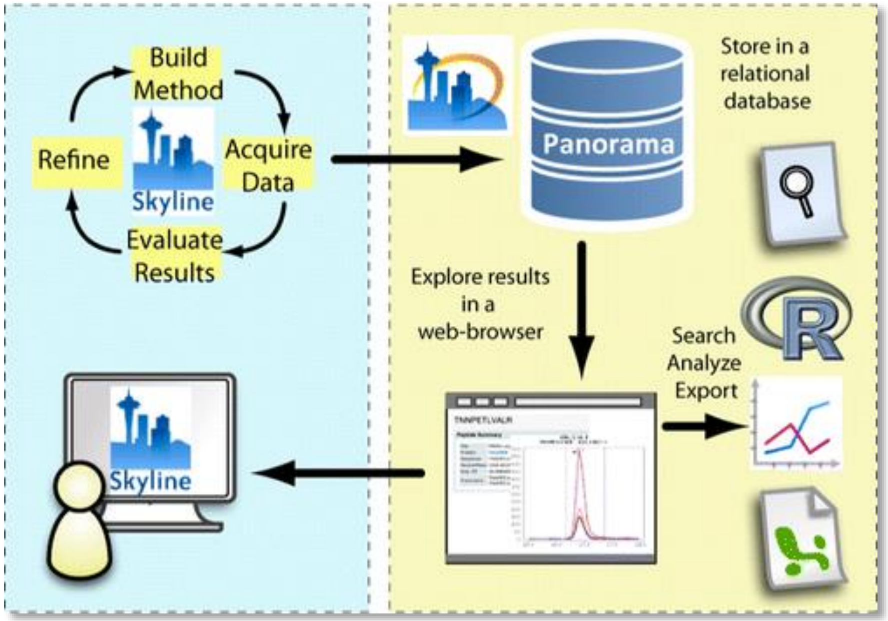
The database contains SRM transitions from a variety of data sources, as summarized here.
A pre-publication preview of the Human SRMAtlas is available derived from over 170,000 human proteome peptides. The database will soon be expanded to include substantial coverage of several important model organisms, as outlined below.

Organism	# Peptides	Coverage %
M. tuberculosis†	13,248	99.0
S. cerevisiae †	28,000	99.0
Human ‡		99.9
Mouse		Coming soon
Rat		Coming soon
Bovine		Coming soon
Rabbit		Coming soon

† Currently available.
‡ Available as a pre-publication preview.

SRMAtlas

http://www.srmatlas.org



Vagisha *et al.*: J. Proteome Res., 2014, 13 (9), pp 4205–4210

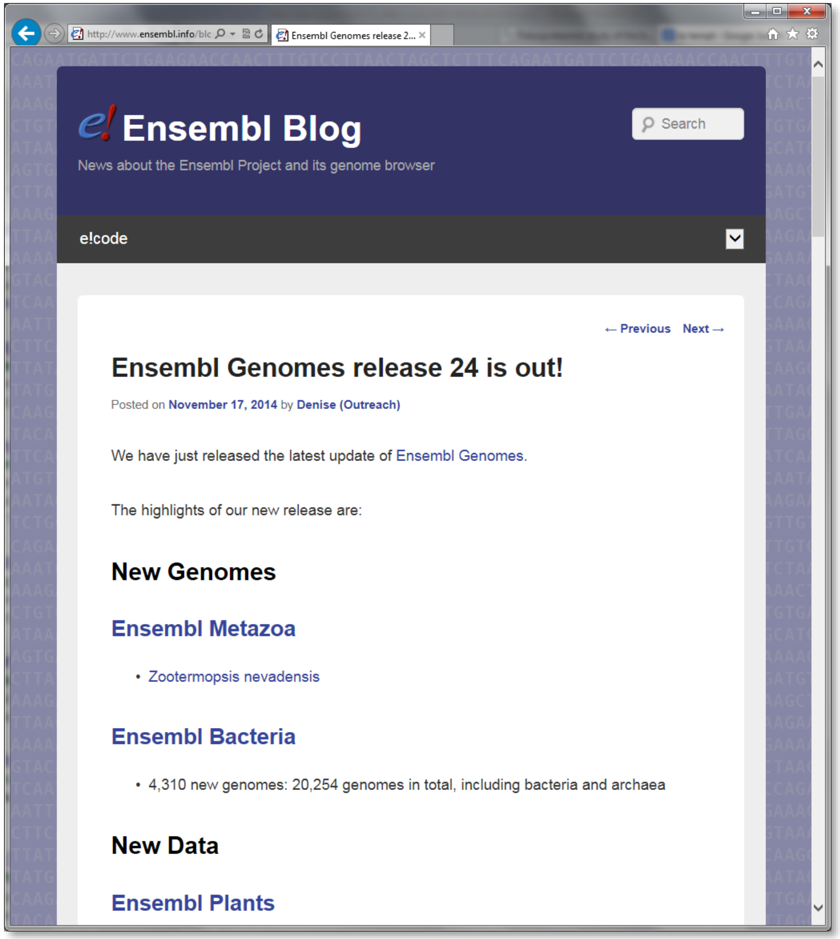
How can we use the shared data?

1) Verify published findings

2) Reuse existing data or knowledge

3) **Generate new knowledge**

Databases improve



Ensembl Blog

News about the Ensembl Project and its genome browser

Ensembl Genomes release 24 is out!

Posted on November 17, 2014 by Denise (Outreach)

We have just released the latest update of Ensembl Genomes.

The highlights of our new release are:

New Genomes

Ensembl Metazoa

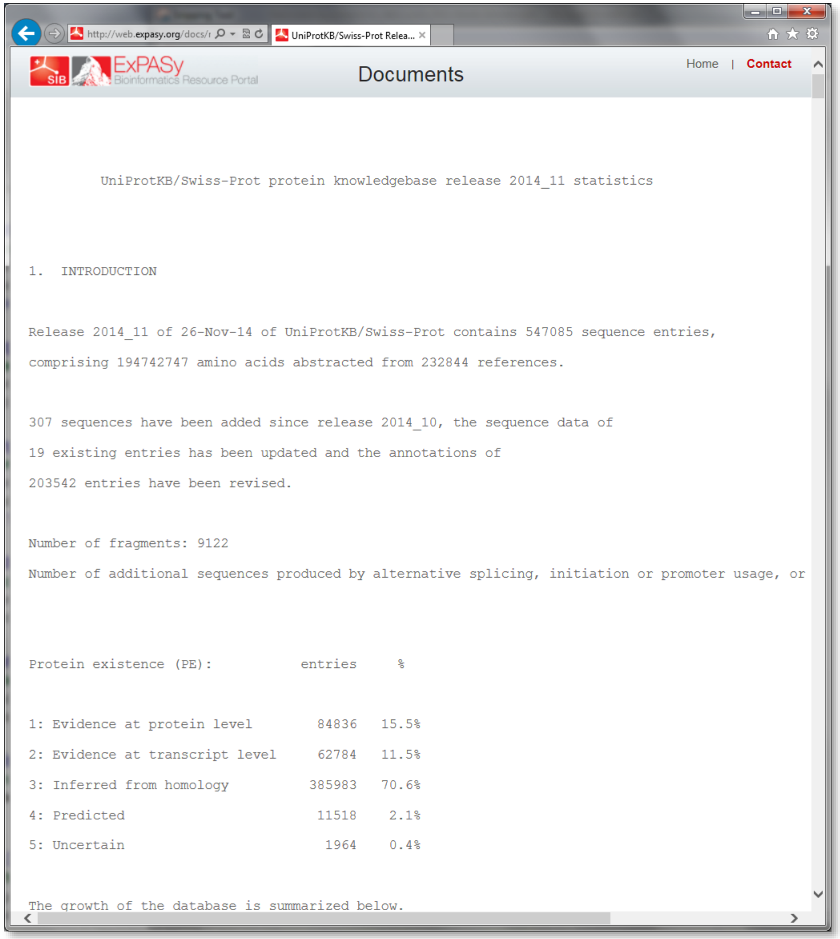
- Zootermopsis nevadensis

Ensembl Bacteria

- 4,310 new genomes: 20,254 genomes in total, including bacteria and archaea

New Data

Ensembl Plants

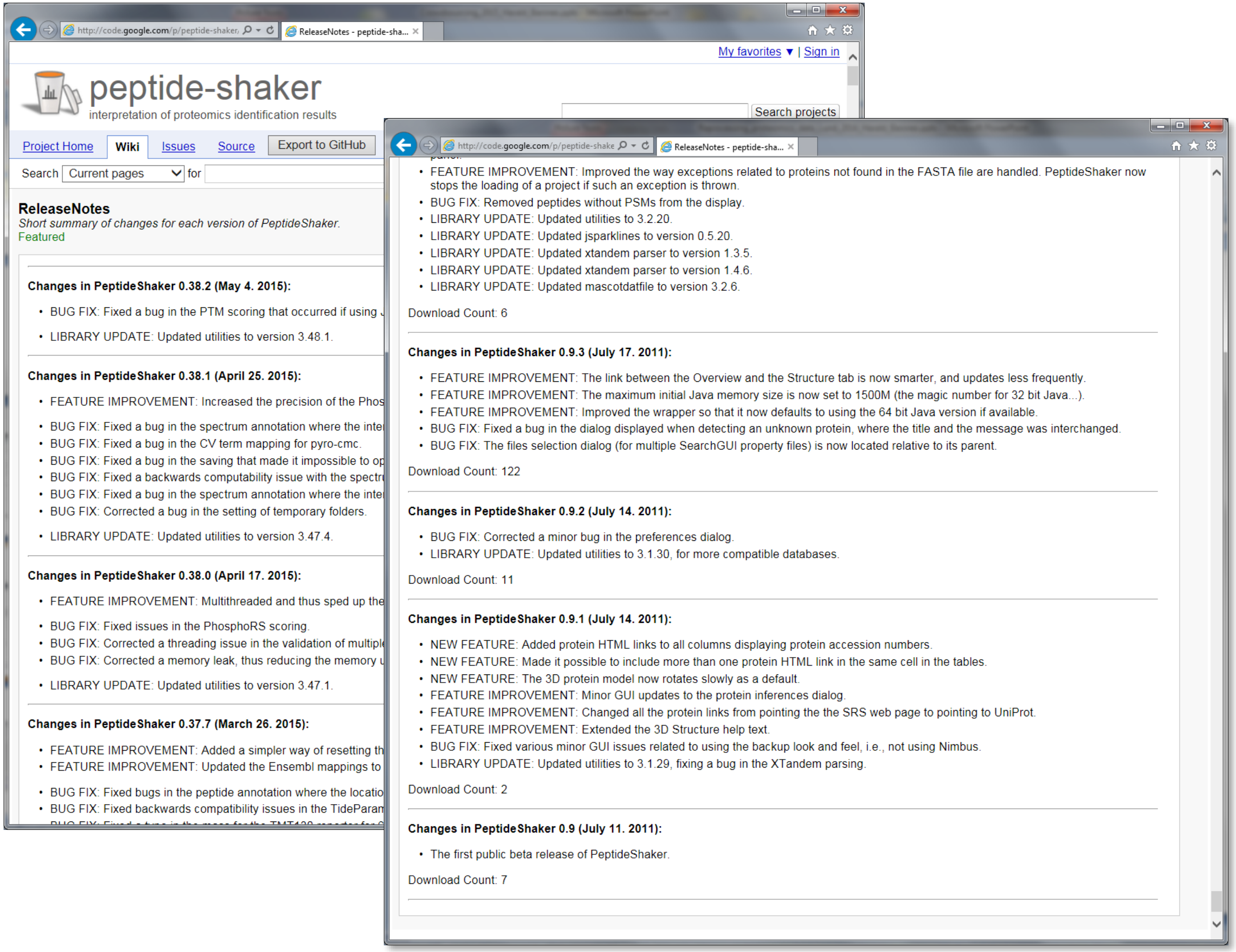


http://web.expasy.org/docs/r
UniProtKB/Swiss-Prot Relea...
ExPASy
Bioinformatics Resource Portal
Home | Contact

Documents

```
        UniProtKB/Swiss-Prot protein knowledgebase release 2014_11 statistics


1.  INTRODUCTION

Release 2014_11 of 26-Nov-14 of UniProtKB/Swiss-Prot contains 547085 sequence entries,
comprising 194742747 amino acids abstracted from 232844 references.

307 sequences have been added since release 2014_10, the sequence data of
19 existing entries has been updated and the annotations of
203542 entries have been revised.

Number of fragments: 9122
Number of additional sequences produced by alternative splicing, initiation or promoter usage, or


Protein existence (PE):           entries     %

1: Evidence at protein level        84836   15.5%
2: Evidence at transcript level     62784   11.5%
3: Inferred from homology          385983   70.6%
4: Predicted                        11518    2.1%
5: Uncertain                         1964    0.4%

The growth of the database is summarized below.
```

Software improves



peptide-shaker

interpretation of proteomics identification results

ReleaseNotes

Short summary of changes for each version of PeptideShaker.

Featured

Changes in PeptideShaker 0.38.2 (May 4. 2015):

- BUG FIX: Fixed a bug in the PTM scoring that occurred if using J
- LIBRARY UPDATE: Updated utilities to version 3.48.1.

Changes in PeptideShaker 0.38.1 (April 25. 2015):

- FEATURE IMPROVEMENT: Increased the precision of the Phos
- BUG FIX: Fixed a bug in the spectrum annotation where the inte
- BUG FIX: Fixed a bug in the CV term mapping for pyro-cmc.
- BUG FIX: Fixed a bug in the saving that made it impossible to op
- BUG FIX: Fixed a backwards computability issue with the spectr
- BUG FIX: Fixed a bug in the spectrum annotation where the inte
- BUG FIX: Corrected a bug in the setting of temporary folders.
- LIBRARY UPDATE: Updated utilities to version 3.47.4.

Changes in PeptideShaker 0.38.0 (April 17. 2015):

- FEATURE IMPROVEMENT: Multithreaded and thus sped up the
- BUG FIX: Fixed issues in the PhosphoRS scoring.
- BUG FIX: Corrected a threading issue in the validation of multipl
- BUG FIX: Corrected a memory leak, thus reducing the memory u
- LIBRARY UPDATE: Updated utilities to version 3.47.1.

Changes in PeptideShaker 0.37.7 (March 26. 2015):

- FEATURE IMPROVEMENT: Added a simpler way of resetting th
- FEATURE IMPROVEMENT: Updated the Ensembl mappings to
- BUG FIX: Fixed bugs in the peptide annotation where the locatio
- BUG FIX: Fixed backwards compatibility issues in the TideParam

http://code.google.com/p/peptide-shake
ReleaseNotes - peptide-sha...

- FEATURE IMPROVEMENT: Improved the way exceptions related to proteins not found in the FASTA file are handled. PeptideShaker now stops the loading of a project if such an exception is thrown.
- BUG FIX: Removed peptides without PSMs from the display.
- LIBRARY UPDATE: Updated utilities to 3.2.20.
- LIBRARY UPDATE: Updated jsparklines to version 0.5.20.
- LIBRARY UPDATE: Updated xtandem parser to version 1.3.5.
- LIBRARY UPDATE: Updated xtandem parser to version 1.4.6.
- LIBRARY UPDATE: Updated mascotdatfile to version 3.2.6.

Download Count: 6

Changes in PeptideShaker 0.9.3 (July 17. 2011):

- FEATURE IMPROVEMENT: The link between the Overview and the Structure tab is now smarter, and updates less frequently.
- FEATURE IMPROVEMENT: The maximum initial Java memory size is now set to 1500M (the magic number for 32 bit Java...).
- FEATURE IMPROVEMENT: Improved the wrapper so that it now defaults to using the 64 bit Java version if available.
- BUG FIX: Fixed a bug in the dialog displayed when detecting an unknown protein, where the title and the message was interchanged.
- BUG FIX: The files selection dialog (for multiple SearchGUI property files) is now located relative to its parent.

Download Count: 122

Changes in PeptideShaker 0.9.2 (July 14. 2011):

- BUG FIX: Corrected a minor bug in the preferences dialog.
- LIBRARY UPDATE: Updated utilities to 3.1.30, for more compatible databases.

Download Count: 11

Changes in PeptideShaker 0.9.1 (July 14. 2011):

- NEW FEATURE: Added protein HTML links to all columns displaying protein accession numbers.
- NEW FEATURE: Made it possible to include more than one protein HTML link in the same cell in the tables.
- NEW FEATURE: The 3D protein model now rotates slowly as a default.
- FEATURE IMPROVEMENT: Minor GUI updates to the protein inferences dialog.
- FEATURE IMPROVEMENT: Changed all the protein links from pointing the the SRS web page to pointing to UniProt.
- FEATURE IMPROVEMENT: Extended the 3D Structure help text.
- BUG FIX: Fixed various minor GUI issues related to using the backup look and feel, i.e., not using Nimbus.
- LIBRARY UPDATE: Updated utilities to 3.1.29, fixing a bug in the XTandem parsing.

Download Count: 2

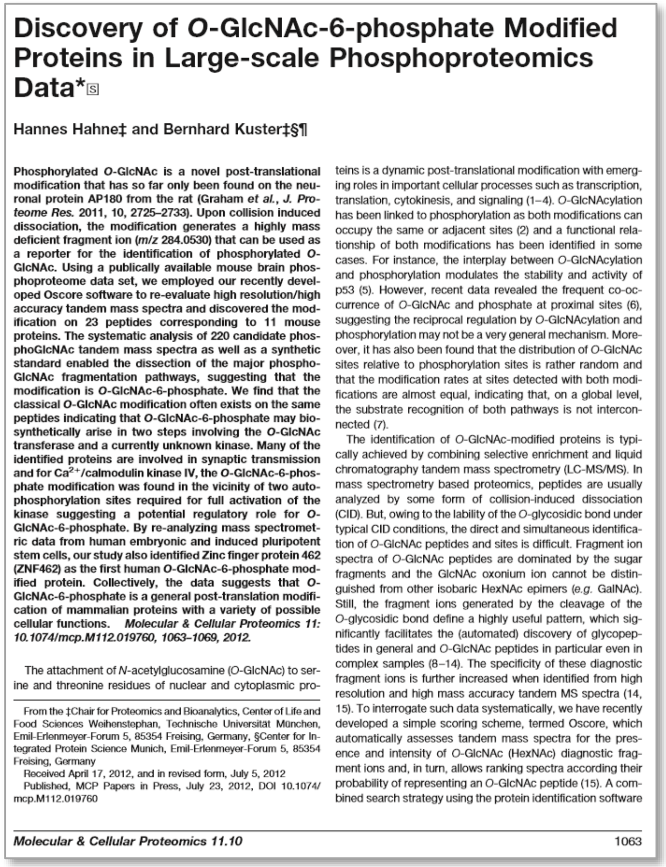
Changes in PeptideShaker 0.9 (July 11. 2011):

- The first public beta release of PeptideShaker.

Download Count: 7

Reprocess to find new post-translational modifications

- Reprocess raw data with new hypotheses in mind (not taken into account by the original authors)



Discovery of *O*-GlcNAc-6-phosphate Modified Proteins in Large-scale Phosphoproteomics Data*

Hannes Hahne‡ and Bernhard Kuster‡§¶

Phosphorylated *O*-GlcNAc is a novel post-translational modification that has so far only been found on the neuronal protein AP180 from the rat (Graham *et al.*, *J. Proteome Res.* 2011, 10, 2725–2733). Upon collision induced dissociation, the modification generates a highly mass deficient fragment ion (*m/z* 284.0530) that can be used as a reporter for the identification of phosphorylated *O*-GlcNAc. Using a publically available mouse brain phosphoproteome data set, we employed our recently developed Oscore software to re-evaluate high resolution/high accuracy tandem mass spectra and discovered the modification on 23 peptides corresponding to 11 mouse proteins. The systematic analysis of 220 candidate phosphoGlcNAc tandem mass spectra as well as a synthetic standard enabled the dissection of the major phospho-GlcNAc fragmentation pathways, suggesting that the modification is *O*-GlcNAc-6-phosphate. We find that the classical *O*-GlcNAc modification often exists on the same peptides indicating that *O*-GlcNAc-6-phosphate may biosynthetically arise in two steps involving the *O*-GlcNAc transferase and a currently unknown kinase. Many of the identified proteins are involved in synaptic transmission and for Ca^{2+}/calmodulin kinase IV, the *O*-GlcNAc-6-phosphate modification was found in the vicinity of two autophosphorylation sites required for full activation of the kinase suggesting a potential regulatory role for *O*-GlcNAc-6-phosphate. By re-analyzing mass spectrometric data from human embryonic and induced pluripotent stem cells, our study also identified Zinc finger protein 462 (ZNF462) as the first human *O*-GlcNAc-6-phosphate modified protein. Collectively, the data suggests that *O*-GlcNAc-6-phosphate is a general post-translation modification of mammalian proteins with a variety of possible cellular functions. *Molecular & Cellular Proteomics 11: 10.1074/mcp.M112.019760, 1063–1069, 2012.*

The attachment of *N*-acetylglucosamine (*O*-GlcNAc) to serine and threonine residues of nuclear and cytoplasmic proteins is a dynamic post-translational modification with emerging roles in important cellular processes such as transcription, translation, cytokinesis, and signaling (1–4). *O*-GlcNAcylation has been linked to phosphorylation as both modifications can occupy the same or adjacent sites (2) and a functional relationship of both modifications has been identified in some cases. For instance, the interplay between *O*-GlcNAcylation and phosphorylation modulates the stability and activity of p53 (5). However, recent data revealed the frequent co-occurrence of *O*-GlcNAc and phosphate at proximal sites (6), suggesting the reciprocal regulation by *O*-GlcNAcylation and phosphorylation may not be a very general mechanism. Moreover, it has also been found that the distribution of *O*-GlcNAc sites relative to phosphorylation sites is rather random and that the modification rates at sites detected with both modifications are almost equal, indicating that, on a global level, the substrate recognition of both pathways is not interconnected (7).

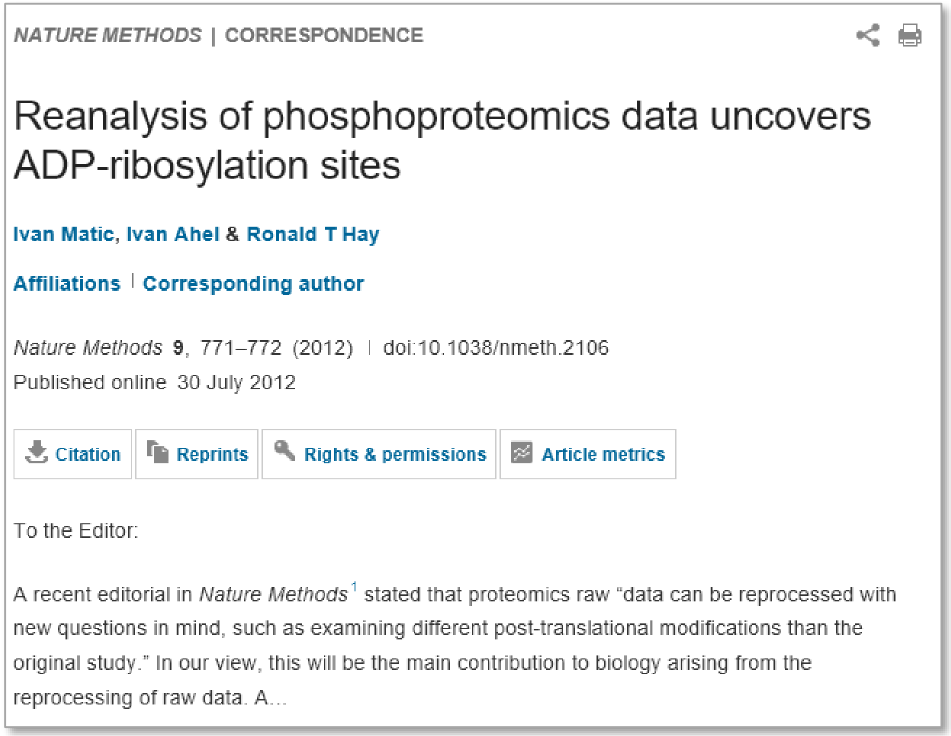
The identification of *O*-GlcNAc-modified proteins is typically achieved by combining selective enrichment and liquid chromatography tandem mass spectrometry (LC-MS/MS). In mass spectrometry based proteomics, peptides are usually analyzed by some form of collision-induced dissociation (CID). But, owing to the lability of the *O*-glycosidic bond under typical CID conditions, the direct and simultaneous identification of *O*-GlcNAc peptides and sites is difficult. Fragment ion spectra of *O*-GlcNAc peptides are dominated by the sugar fragments and the GlcNAc oxonium ion cannot be distinguished from other isobaric HexNAc epimers (*e.g.* GalNAc). Still, the fragment ions generated by the cleavage of the *O*-glycosidic bond define a highly useful pattern, which significantly facilitates the (automated) discovery of glycopeptides in general and *O*-GlcNAc peptides in particular even in complex samples (8–14). The specificity of these diagnostic fragment ions is further increased when identified from high resolution and high mass accuracy tandem MS spectra (14, 15). To interrogate such data systematically, we have recently developed a simple scoring scheme, termed Oscore, which automatically assesses tandem mass spectra for the presence and intensity of *O*-GlcNAc (HexNAc) diagnostic fragment ions and, in turn, allows ranking spectra according their probability of representing an *O*-GlcNAc peptide (15). A combined search strategy using the protein identification software

From the ‡Chair for Proteomics and Bioanalytics, Center of Life and Food Sciences Weihenstephan, Technische Universität München, Emil-Erlenmeyer-Forum 5, 85354 Freising, Germany, §Center for Integrated Protein Science Munich, Emil-Erlenmeyer-Forum 5, 85354 Freising, Germany

Received April 17, 2012, and in revised form, July 5, 2012

Published, MCP Papers in Press, July 23, 2012, DOI 10.1074/mcp.M112.019760

Molecular & Cellular Proteomics 11.10 1063



NATURE METHODS | CORRESPONDENCE

Reanalysis of phosphoproteomics data uncovers ADP-ribosylation sites

Ivan Matic, Ivan Ahel & Ronald T Hay

Affiliations | Corresponding author

Nature Methods **9**, 771–772 (2012) | doi:10.1038/nmeth.2106

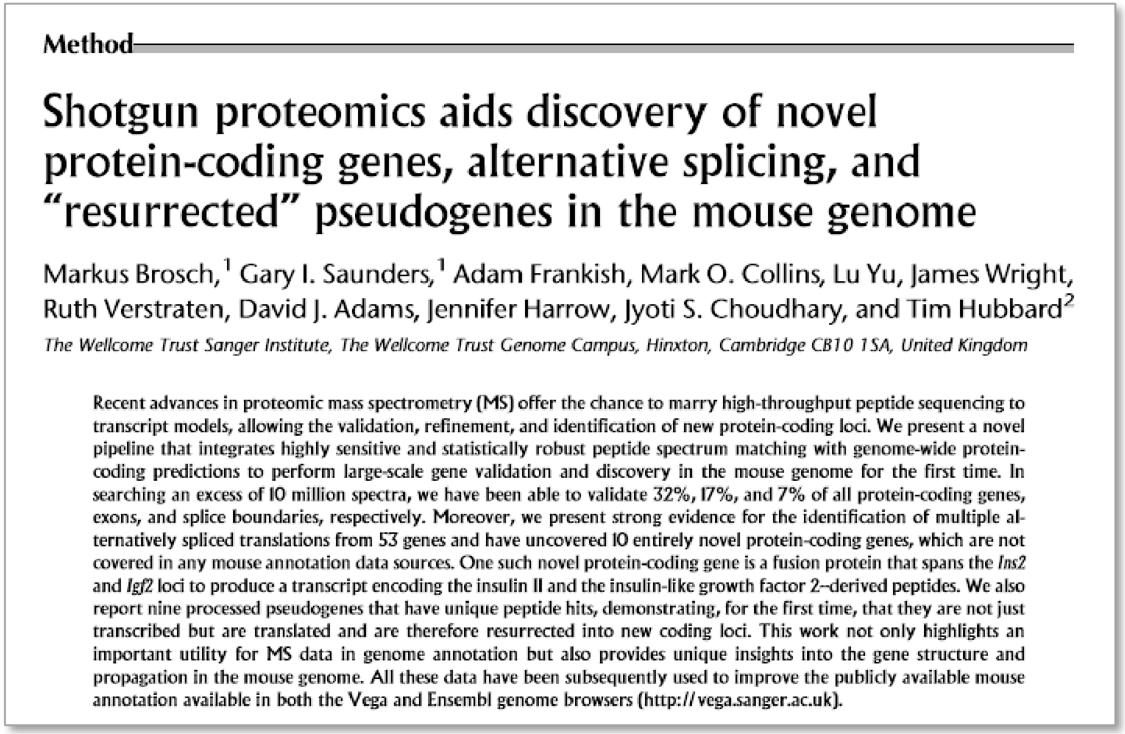
Published online 30 July 2012

To the Editor:

A recent editorial in *Nature Methods*[1] stated that proteomics raw "data can be reprocessed with new questions in mind, such as examining different post-translational modifications than the original study." In our view, this will be the main contribution to biology arising from the reprocessing of raw data. A...

Reprocess to improve genome annotations

- Reprocessing raw mass spectrometry data
  - Validate existing genes
  - Find new splice isoforms, pseudogenes, etc.



Method

Shotgun proteomics aids discovery of novel protein-coding genes, alternative splicing, and "resurrected" pseudogenes in the mouse genome

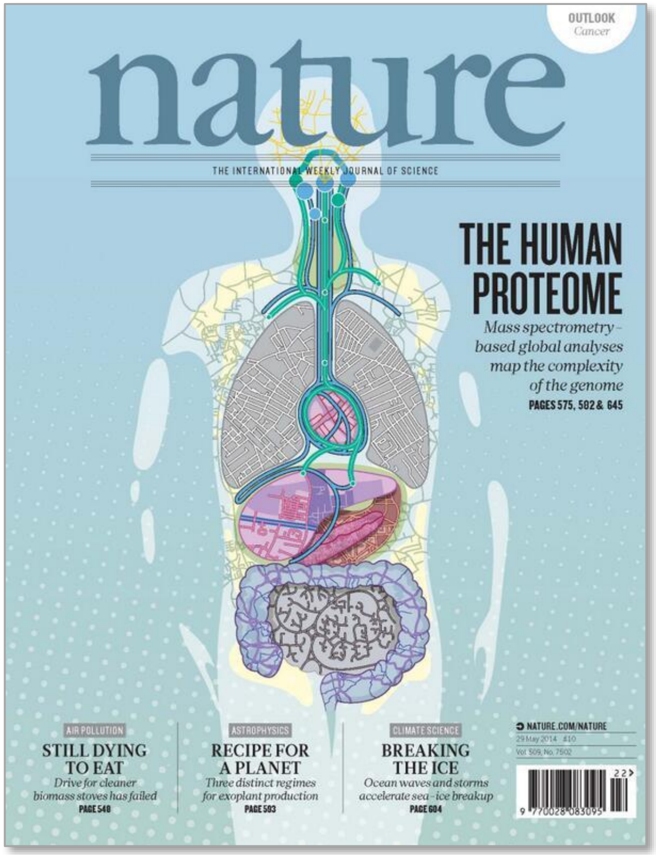
Markus Brosch,[1] Gary I. Saunders,[1] Adam Frankish, Mark O. Collins, Lu Yu, James Wright, Ruth Verstraten, David J. Adams, Jennifer Harrow, Jyoti S. Choudhary, and Tim Hubbard[2]

The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom

Recent advances in proteomic mass spectrometry (MS) offer the chance to marry high-throughput peptide sequencing to transcript models, allowing the validation, refinement, and identification of new protein-coding loci. We present a novel pipeline that integrates highly sensitive and statistically robust peptide spectrum matching with genome-wide protein-coding predictions to perform large-scale gene validation and discovery in the mouse genome for the first time. In searching an excess of 10 million spectra, we have been able to validate 32%, 17%, and 7% of all protein-coding genes, exons, and splice boundaries, respectively. Moreover, we present strong evidence for the identification of multiple alternatively spliced translations from 53 genes and have uncovered 10 entirely novel protein-coding genes, which are not covered in any mouse annotation data sources. One such novel protein-coding gene is a fusion protein that spans the *Ins2* and *Igf2* loci to produce a transcript encoding the insulin II and the insulin-like growth factor 2–derived peptides. We also report nine processed pseudogenes that have unique peptide hits, demonstrating, for the first time, that they are not just transcribed but are translated and are therefore resurrected into new coding loci. This work not only highlights an important utility for MS data in genome annotation but also provides unique insights into the gene structure and propagation in the mouse genome. All these data have been subsequently used to improve the publicly available mouse annotation available in both the Vega and Ensembl genome browsers (http://vega.sanger.ac.uk).

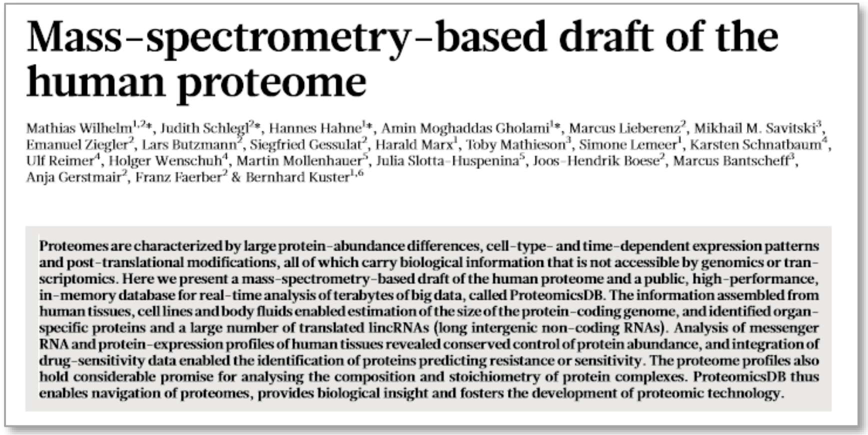
- 53 genes alternatively transcribed
- 10 new protein coding genes

Drafts of the human proteome



OUTLOOK
Cancer
nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
THE HUMAN PROTEOME
Mass spectrometry-based global analyses map the complexity of the genome
PAGES 575, 582 & 645
AIR POLLUTION
STILL DYING TO EAT
Drive for cleaner biomass stoves has failed
PAGE 540
ASTROPHYSICS
RECIPE FOR A PLANET
Three distinct regimes for exoplant production
PAGE 503
CLIMATE SCIENCE
BREAKING THE ICE
Ocean waves and storms accelerate sea-ice breakup
PAGE 604
NATURE.COM/NATURE
29 May 2014 £10
Vol 509, No. 7502

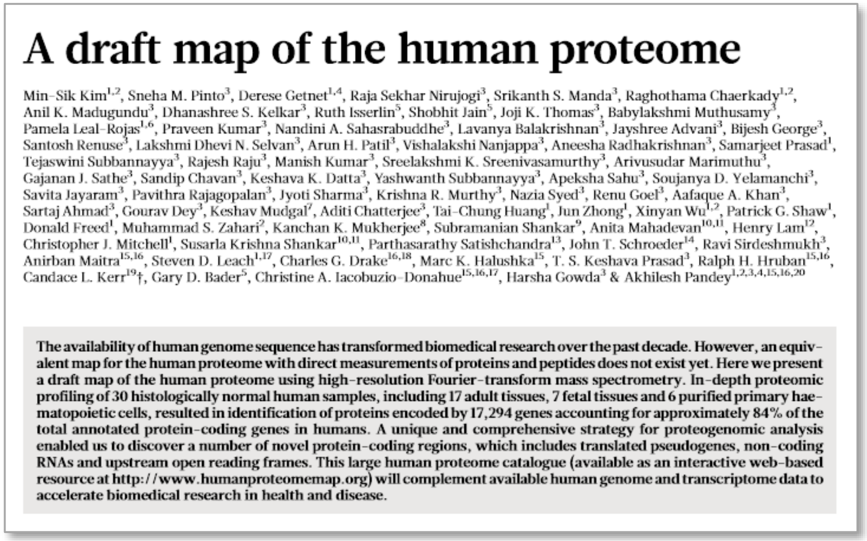
Nature cover May 2014



Mass-spectrometry-based draft of the human proteome

Mathias Wilhelm[1,2]*, Judith Schlegl[2]*, Hannes Hahne[1]*, Amin Moghaddas Gholami[1]*, Marcus Lieberenz[2], Mikhail M. Savitski[3], Emanuel Ziegler[2], Lars Butzmann[2], Siegfried Gessulat[2], Harald Marx[1], Toby Mathieson[3], Simone Lemeer[1], Karsten Schnatbaum[4], Ulf Reimer[4], Holger Wenschuh[4], Martin Mollenhauer[5], Julia Slotta-Huspenina[5], Joos-Hendrik Boese[2], Marcus Bantscheff[3], Anja Gerstmair[2], Franz Faerber[2] & Bernhard Kuster[1,6]

Proteomes are characterized by large protein-abundance differences, cell-type- and time-dependent expression patterns and post-translational modifications, all of which carry biological information that is not accessible by genomics or transcriptomics. Here we present a mass-spectrometry-based draft of the human proteome and a public, high-performance, in-memory database for real-time analysis of terabytes of big data, called ProteomicsDB. The information assembled from human tissues, cell lines and body fluids enabled estimation of the size of the protein-coding genome, and identified organ-specific proteins and a large number of translated lincRNAs (long intergenic non-coding RNAs). Analysis of messenger RNA and protein-expression profiles of human tissues revealed conserved control of protein abundance, and integration of drug-sensitivity data enabled the identification of proteins predicting resistance or sensitivity. The proteome profiles also hold considerable promise for analysing the composition and stoichiometry of protein complexes. ProteomicsDB thus enables navigation of proteomes, provides biological insight and fosters the development of proteomic technology.

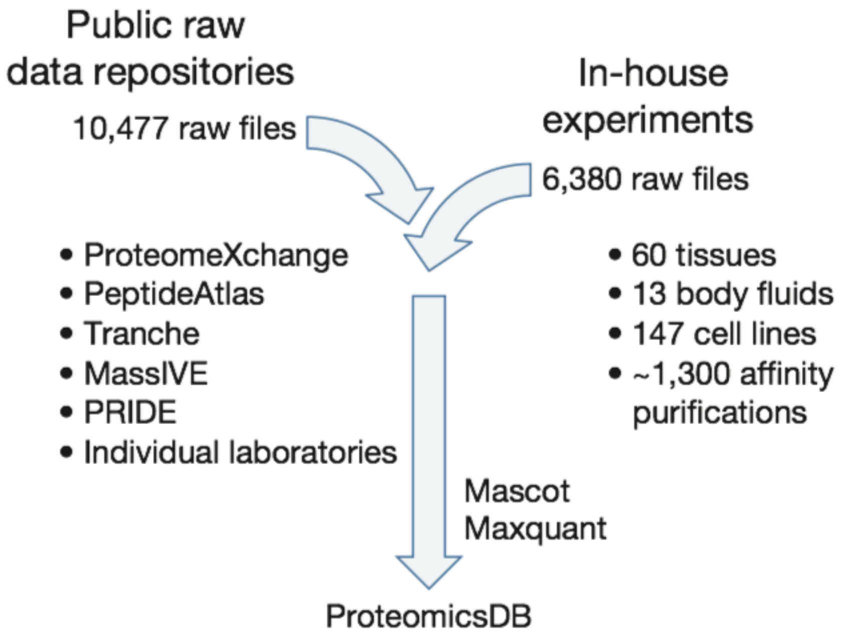


A draft map of the human proteome

Min-Sik Kim[1,2], Sneha M. Pinto[3], Derese Getnet[1,4], Raja Sekhar Nirujogi[3], Srikanth S. Manda[3], Raghothama Chaerkady[1,2], Anil K. Madugundu[3], Dhanashree S. Kelkar[3], Ruth Isserlin[5], Shobhit Jain[5], Joji K. Thomas[3], Babylakshmi Muthusamy[3], Pamela Leal-Rojas[1,6], Praveen Kumar[3], Nandini A. Sahasrabuddhe[3], Lavanya Balakrishnan[3], Jayshree Advani[3], Bijesh George[3], Santosh Renuse[3], Lakshmi Dhevi N. Selvan[3], Arun H. Patil[3], Vishalakshi Nanjappa[3], Aneesha Radhakrishnan[3], Samarjeet Prasad[1], Tejaswini Subbannayya[3], Rajesh Raju[3], Manish Kumar[3], Sreelakshmi K. Sreenivasamurthy[3], Arivusudar Marimuthu[3], Gajanan J. Sathe[3], Sandip Chavan[3], Keshava K. Datta[3], Yashwanth Subbannayya[3], Apeksha Sahu[3], Soujanya D. Yelamanchi[3], Savita Jayaram[3], Pavithra Rajagopalan[3], Jyoti Sharma[3], Krishna R. Murthy[3], Nazia Syed[3], Renu Goel[3], Aafaque A. Khan[3], Sartaj Ahmad[3], Gourav Dey[3], Keshav Mudgal[7], Aditi Chatterjee[3], Tai-Chung Huang[1], Jun Zhong[1], Xinyan Wu[1,2], Patrick G. Shaw[1], Donald Freed[1], Muhammad S. Zahari[2], Kanchan K. Mukherjee[8], Subramanian Shankar[9], Anita Mahadevan[10,11], Henry Lam[12], Christopher J. Mitchell[1], Susarla Krishna Shankar[10,11], Parthasarathy Satishchandra[13], John T. Schroeder[14], Ravi Sirdeshmukh[3], Anirban Maitra[15,16], Steven D. Leach[1,17], Charles G. Drake[16,18], Marc K. Halushka[15], T. S. Keshava Prasad[3], Ralph H. Hruban[15,16], Candace L. Kerr[19]†, Gary D. Bader[5], Christine A. Iacobuzio-Donahue[15,16,17], Harsha Gowda[3] & Akhilesh Pandey[1,2,3,4,15,16,20]

The availability of human genome sequence has transformed biomedical research over the past decade. However, an equivalent map for the human proteome with direct measurements of proteins and peptides does not exist yet. Here we present a draft map of the human proteome using high-resolution Fourier-transform mass spectrometry. In-depth proteomic profiling of 30 histologically normal human samples, including 17 adult tissues, 7 fetal tissues and 6 purified primary haematopoietic cells, resulted in identification of proteins encoded by 17,294 genes accounting for approximately 84% of the total annotated protein-coding genes in humans. A unique and comprehensive strategy for proteogenomic analysis enabled us to discover a number of novel protein-coding regions, which includes translated pseudogenes, non-coding RNAs and upstream open reading frames. This large human proteome catalogue (available as an interactive web-based resource at http://www.humanproteomemap.org) will complement available human genome and transcriptome data to accelerate biomedical research in health and disease.

Draft of the human proteome



Public raw data repositories

10,477 raw files

- ProteomeXchange
- PeptideAtlas
- Tranche
- MassIVE
- PRIDE
- Individual laboratories

In-house experiments

6,380 raw files

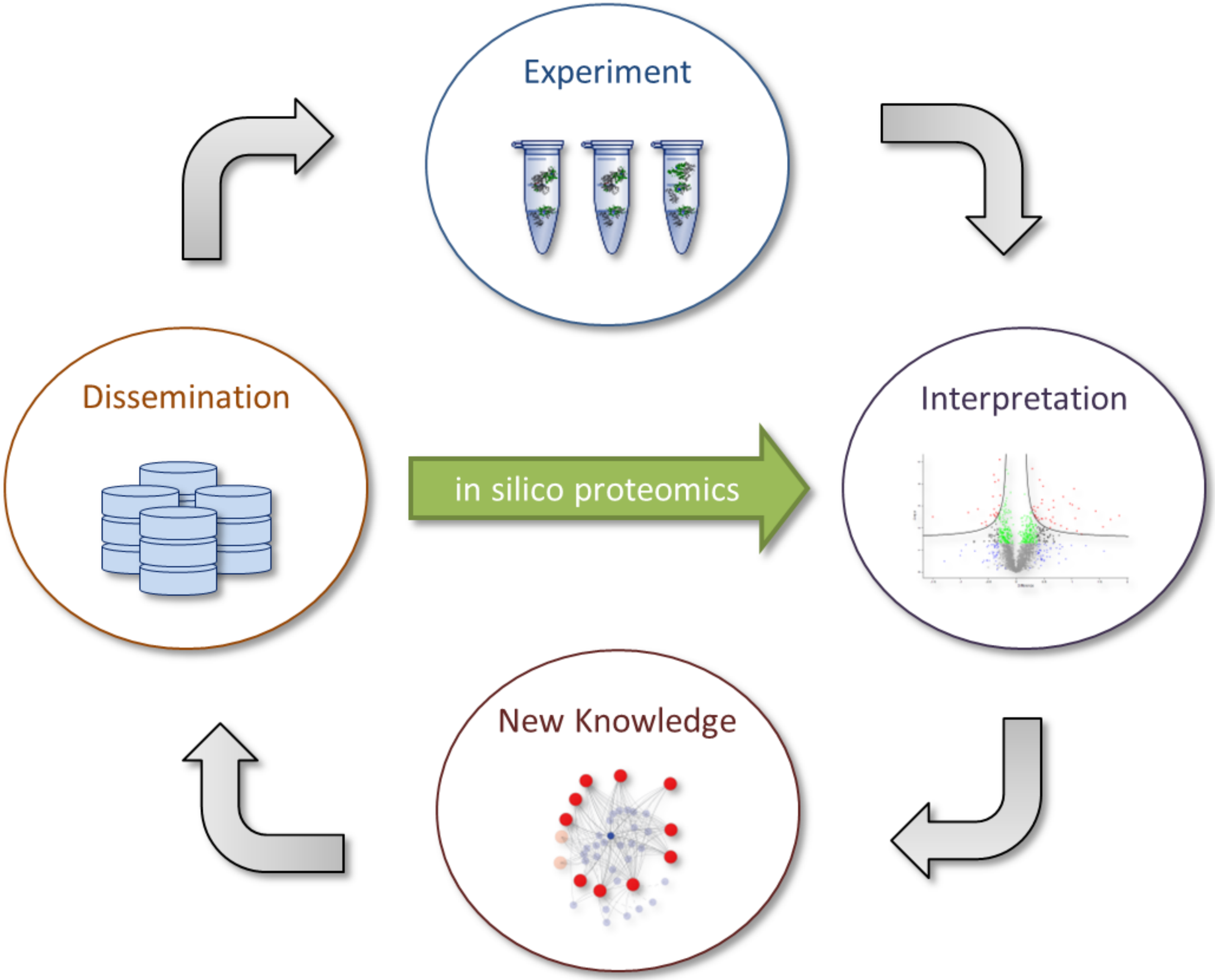
- 60 tissues
- 13 body fluids
- 147 cell lines
- ~1,300 affinity purifications

Mascot
Maxquant

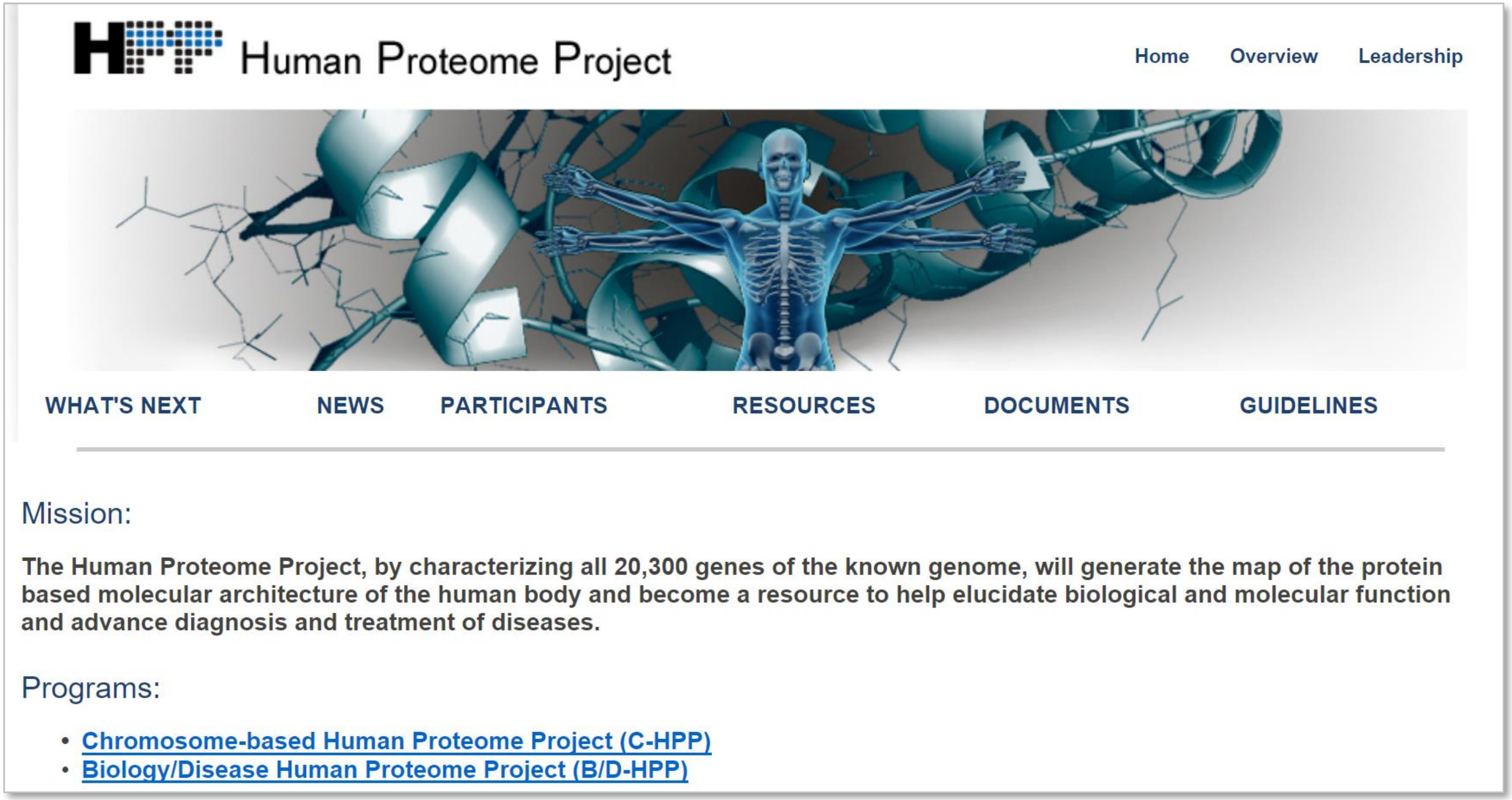
ProteomicsDB

Wilhelm *et al.*, *Nature*, 2014

Public data makes *in silico* proteomics possible!



Vaudel, Proteomics, 2015

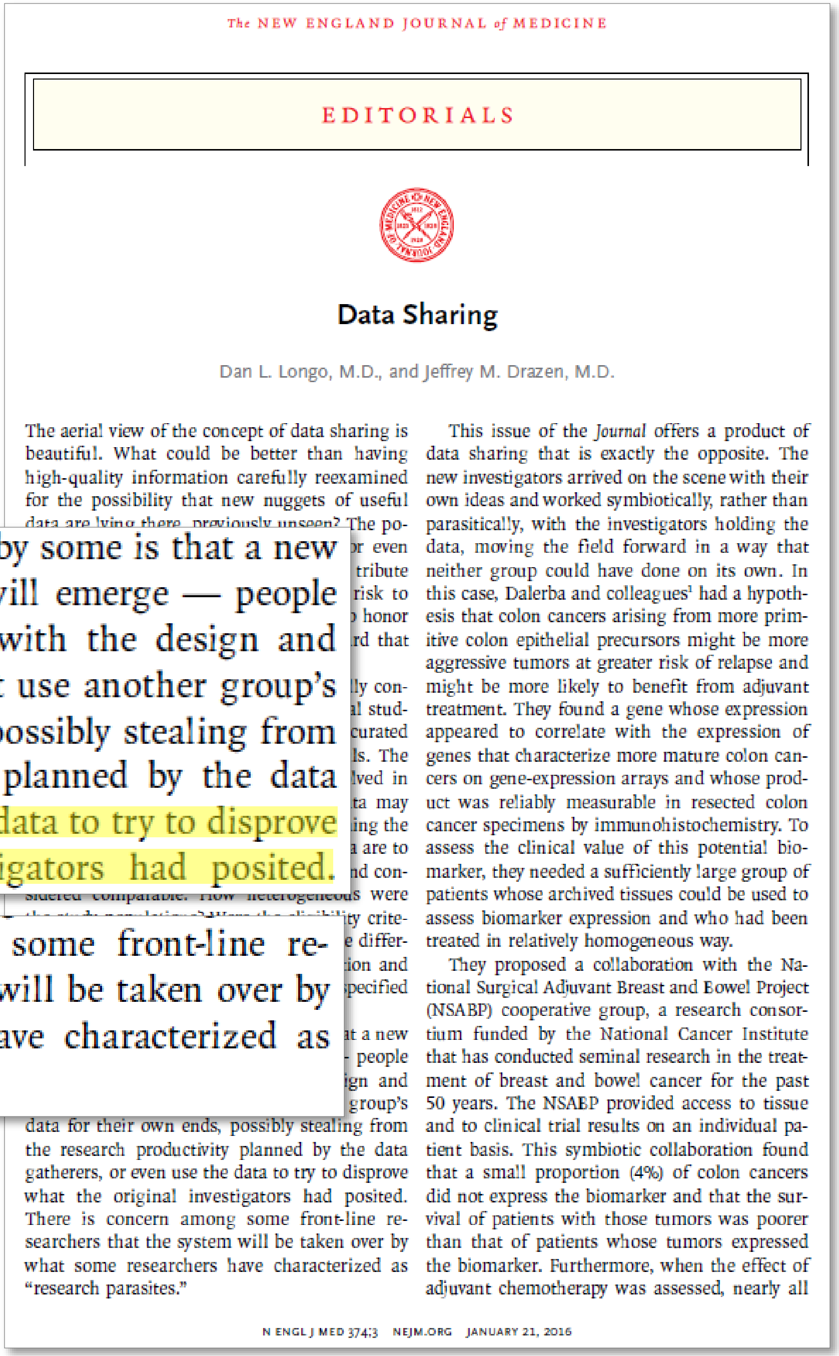


Mission:

The Human Proteome Project, by characterizing all 20,300 genes of the known genome, will generate the map of the protein based molecular architecture of the human body and become a resource to help elucidate biological and molecular function and advance diagnosis and treatment of diseases.

Programs:

- Chromosome-based Human Proteome Project (C-HPP)
- Biology/Disease Human Proteome Project (B/D-HPP)



EDITORIALS

Data Sharing

Dan L. Longo, M.D., and Jeffrey M. Drazen, M.D.

The aerial view of the concept of data sharing is beautiful. What could be better than having high-quality information carefully reexamined for the possibility that new nuggets of useful data are lying there, previously unseen? The po-

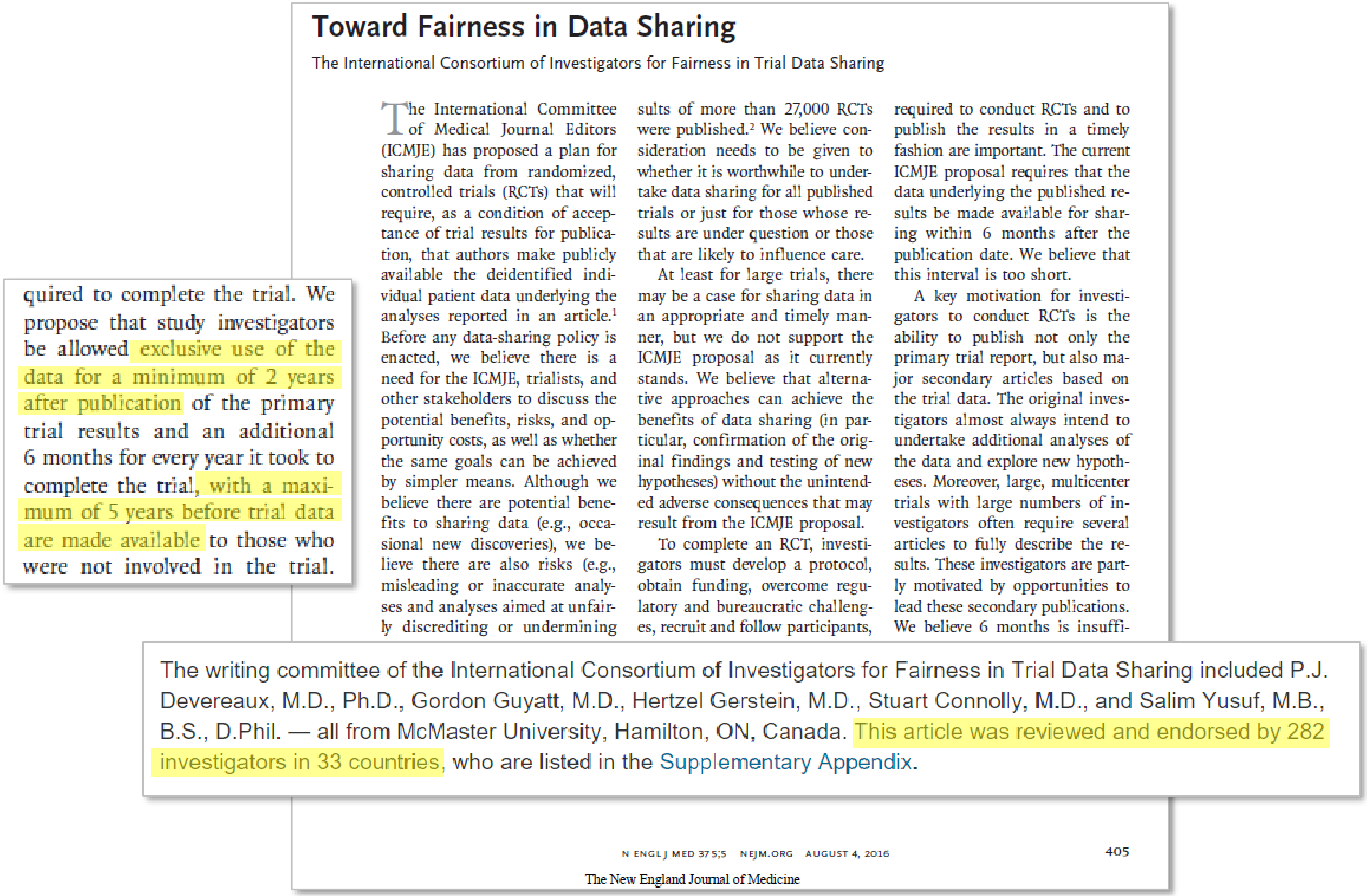
data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited. There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as "research parasites."

This issue of the *Journal* offers a product of data sharing that is exactly the opposite. The new investigators arrived on the scene with their own ideas and worked symbiotically, rather than parasitically, with the investigators holding the data, moving the field forward in a way that neither group could have done on its own. In this case, Dalerba and colleagues[1] had a hypothesis that colon cancers arising from more primitive colon epithelial precursors might be more aggressive tumors at greater risk of relapse and might be more likely to benefit from adjuvant treatment. They found a gene whose expression appeared to correlate with the expression of genes that characterize more mature colon cancers on gene-expression arrays and whose product was reliably measurable in resected colon cancer specimens by immunohistochemistry. To assess the clinical value of this potential biomarker, they needed a sufficiently large group of patients whose archived tissues could be used to assess biomarker expression and who had been treated in relatively homogeneous way.

They proposed a collaboration with the National Surgical Adjuvant Breast and Bowel Project (NSABP) cooperative group, a research consortium funded by the National Cancer Institute that has conducted seminal research in the treatment of breast and bowel cancer for the past 50 years. The NSABP provided access to tissue and to clinical trial results on an individual patient basis. This symbiotic collaboration found that a small proportion (4%) of colon cancers did not express the biomarker and that the survival of patients with those tumors was poorer than that of patients whose tumors expressed the biomarker. Furthermore, when the effect of adjuvant chemotherapy was assessed, nearly all

A second concern held by some is that a new class of research person will emerge — people who had nothing to do with the design and execution of the study but use another group's data for their own ends, possibly stealing from the research productivity planned by the data gatherers, or even use the data to try to disprove what the original investigators had posited.

There is concern among some front-line researchers that the system will be taken over by what some researchers have characterized as "research parasites."



Toward Fairness in Data Sharing

The International Consortium of Investigators for Fairness in Trial Data Sharing

The International Committee of Medical Journal Editors (ICMJE) has proposed a plan for sharing data from randomized, controlled trials (RCTs) that will require, as a condition of acceptance of trial results for publication, that authors make publicly available the deidentified individual patient data underlying the analyses reported in an article.[1] Before any data-sharing policy is enacted, we believe there is a need for the ICMJE, trialists, and other stakeholders to discuss the potential benefits, risks, and opportunity costs, as well as whether the same goals can be achieved by simpler means. Although we believe there are potential benefits to sharing data (e.g., occasional new discoveries), we believe there are also risks (e.g., misleading or inaccurate analyses and analyses aimed at unfairly discrediting or undermining

sults of more than 27,000 RCTs were published.[2] We believe consideration needs to be given to whether it is worthwhile to undertake data sharing for all published trials or just for those whose results are under question or those that are likely to influence care.

At least for large trials, there may be a case for sharing data in an appropriate and timely manner, but we do not support the ICMJE proposal as it currently stands. We believe that alternative approaches can achieve the benefits of data sharing (in particular, confirmation of the original findings and testing of new hypotheses) without the unintended adverse consequences that may result from the ICMJE proposal.

To complete an RCT, investigators must develop a protocol, obtain funding, overcome regulatory and bureaucratic challenges, recruit and follow participants, required to conduct RCTs and to publish the results in a timely fashion are important. The current ICMJE proposal requires that the data underlying the published results be made available for sharing within 6 months after the publication date. We believe that this interval is too short.

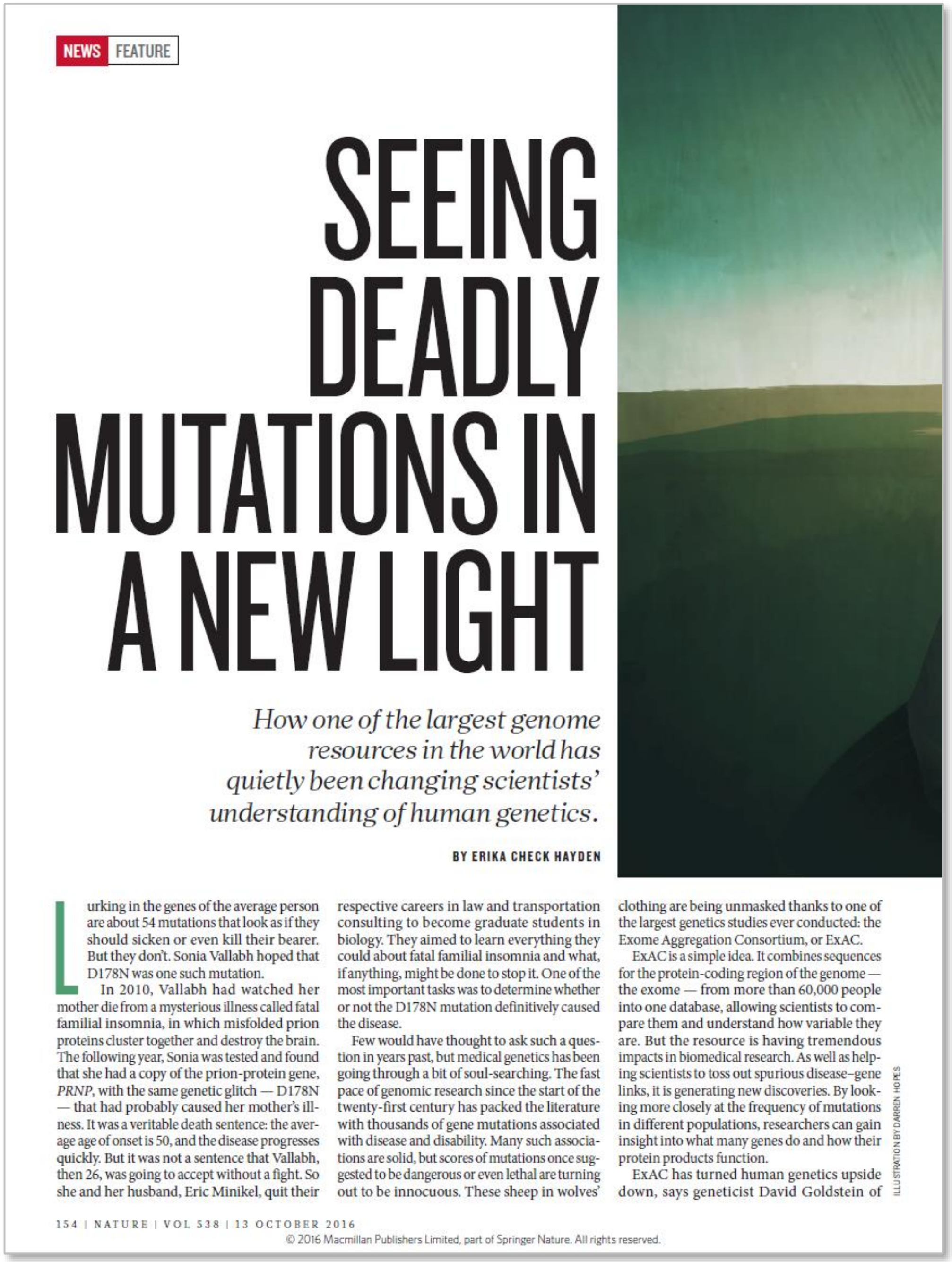
A key motivation for investigators to conduct RCTs is the ability to publish not only the primary trial report, but also major secondary articles based on the trial data. The original investigators almost always intend to undertake additional analyses of the data and explore new hypotheses. Moreover, large, multicenter trials with large numbers of investigators often require several articles to fully describe the results. These investigators are partly motivated by opportunities to lead these secondary publications. We believe 6 months is insuffi-

quired to complete the trial. We propose that study investigators be allowed exclusive use of the data for a minimum of 2 years after publication of the primary trial results and an additional 6 months for every year it took to complete the trial, with a maximum of 5 years before trial data are made available to those who were not involved in the trial.

The writing committee of the International Consortium of Investigators for Fairness in Trial Data Sharing included P.J. Devereaux, M.D., Ph.D., Gordon Guyatt, M.D., Hertzel Gerstein, M.D., Stuart Connolly, M.D., and Salim Yusuf, M.B., B.S., D.Phil. — all from McMaster University, Hamilton, ON, Canada. This article was reviewed and endorsed by 282 investigators in 33 countries, who are listed in the Supplementary Appendix.

N ENGL J MED 375;5 NEJM.ORG AUGUST 4, 2016 405

The New England Journal of Medicine



NEWS FEATURE

SEEING DEADLY MUTATIONS IN A NEW LIGHT

How one of the largest genome resources in the world has quietly been changing scientists' understanding of human genetics.

BY ERIKA CHECK HAYDEN

Lurking in the genes of the average person are about 54 mutations that look as if they should sicken or even kill their bearer. But they don't. Sonia Vallabh hoped that D178N was one such mutation.

In 2010, Vallabh had watched her mother die from a mysterious illness called fatal familial insomnia, in which misfolded prion proteins cluster together and destroy the brain. The following year, Sonia was tested and found that she had a copy of the prion-protein gene, *PRNP*, with the same genetic glitch — D178N — that had probably caused her mother's illness. It was a veritable death sentence: the average age of onset is 50, and the disease progresses quickly. But it was not a sentence that Vallabh, then 26, was going to accept without a fight. So she and her husband, Eric Minikel, quit their respective careers in law and transportation consulting to become graduate students in biology. They aimed to learn everything they could about fatal familial insomnia and what, if anything, might be done to stop it. One of the most important tasks was to determine whether or not the D178N mutation definitively caused the disease.

Few would have thought to ask such a question in years past, but medical genetics has been going through a bit of soul-searching. The fast pace of genomic research since the start of the twenty-first century has packed the literature with thousands of gene mutations associated with disease and disability. Many such associations are solid, but scores of mutations once suggested to be dangerous or even lethal are turning out to be innocuous. These sheep in wolves' clothing are being unmasked thanks to one of the largest genetics studies ever conducted: the Exome Aggregation Consortium, or ExAC.

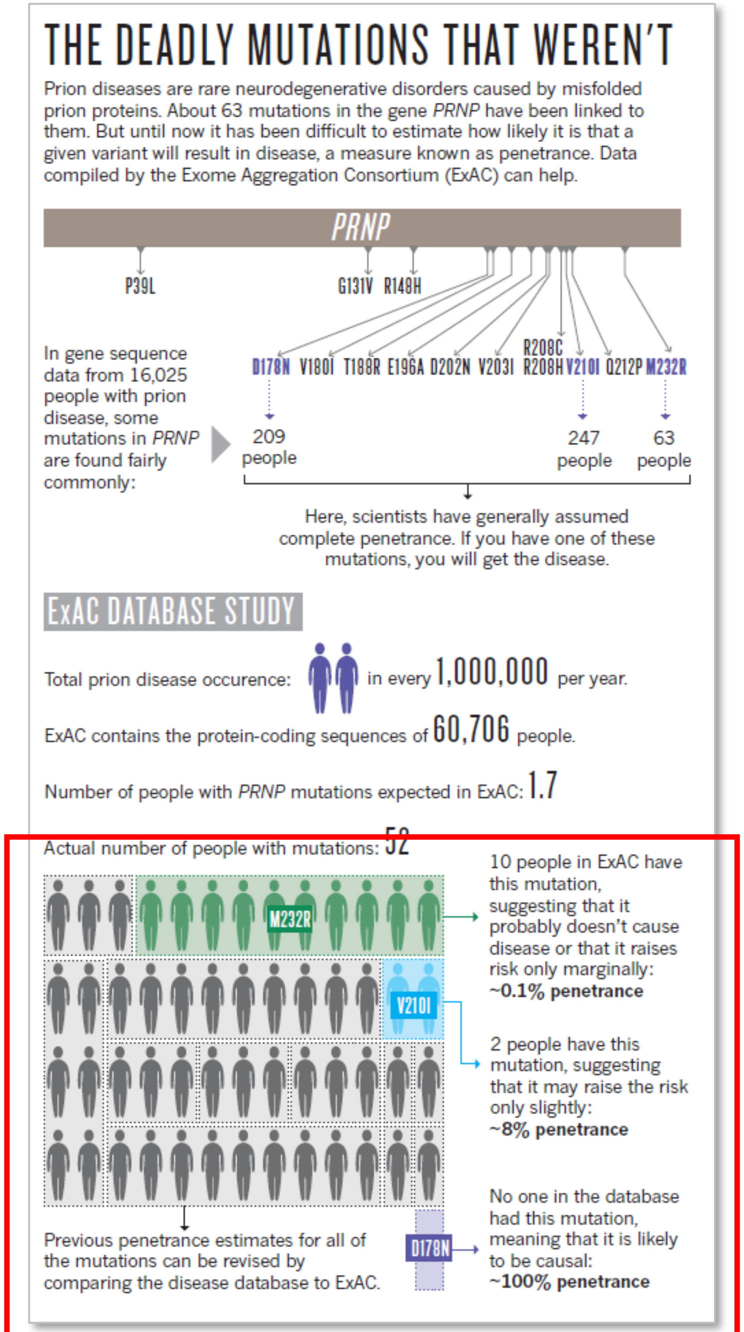
ExAC is a simple idea. It combines sequences for the protein-coding region of the genome — the exome — from more than 60,000 people into one database, allowing scientists to compare them and understand how variable they are. But the resource is having tremendous impacts in biomedical research. As well as helping scientists to toss out spurious disease–gene links, it is generating new discoveries. By looking more closely at the frequency of mutations in different populations, researchers can gain insight into what many genes do and how their protein products function.

ExAC has turned human genetics upside down, says geneticist David Goldstein of

ILLUSTRATION BY DARREN HOPES

THE DEADLY MUTATIONS THAT WEREN'T

Prion diseases are rare neurodegenerative disorders caused by misfolded prion proteins. About 63 mutations in the gene *PRNP* have been linked to them. But until now it has been difficult to estimate how likely it is that a given variant will result in disease, a measure known as penetrance. Data compiled by the Exome Aggregation Consortium (ExAC) can help.



In gene sequence data from 16,025 people with prion disease, some mutations in *PRNP* are found fairly commonly:

Here, scientists have generally assumed complete penetrance. If you have one of these mutations, you will get the disease.

ExAC DATABASE STUDY

Total prion disease occurence: 2 in every 1,000,000 per year.

ExAC contains the protein-coding sequences of 60,706 people.

Number of people with *PRNP* mutations expected in ExAC: 1.7

Actual number of people with mutations: 52

10 people in ExAC have this mutation, suggesting that it probably doesn't cause disease or that it raises risk only marginally: **~0.1% penetrance**

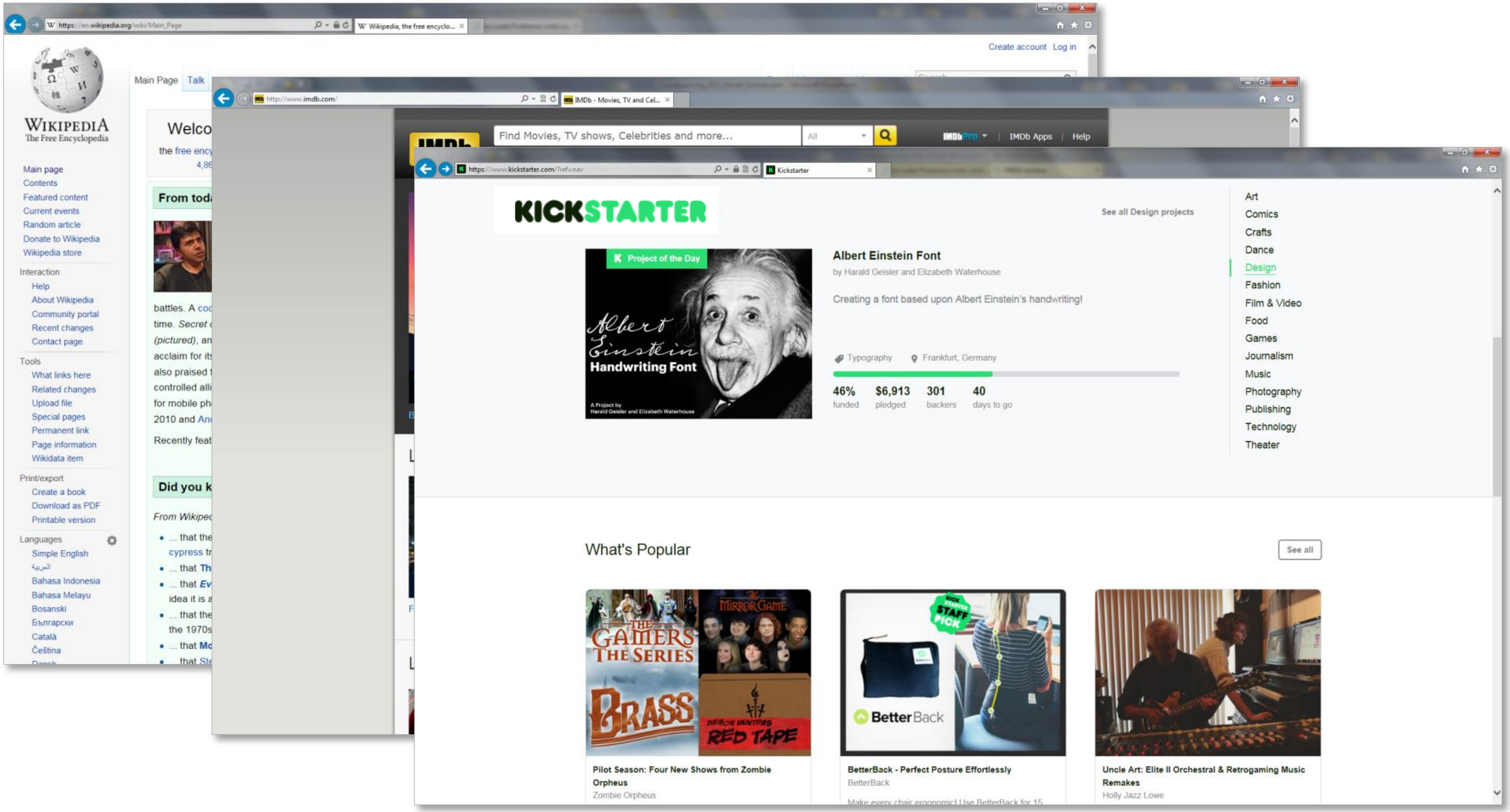
2 people have this mutation, suggesting that it may raise the risk only slightly: **~8% penetrance**

No one in the database had this mutation, meaning that it is likely to be causal: **~100% penetrance**

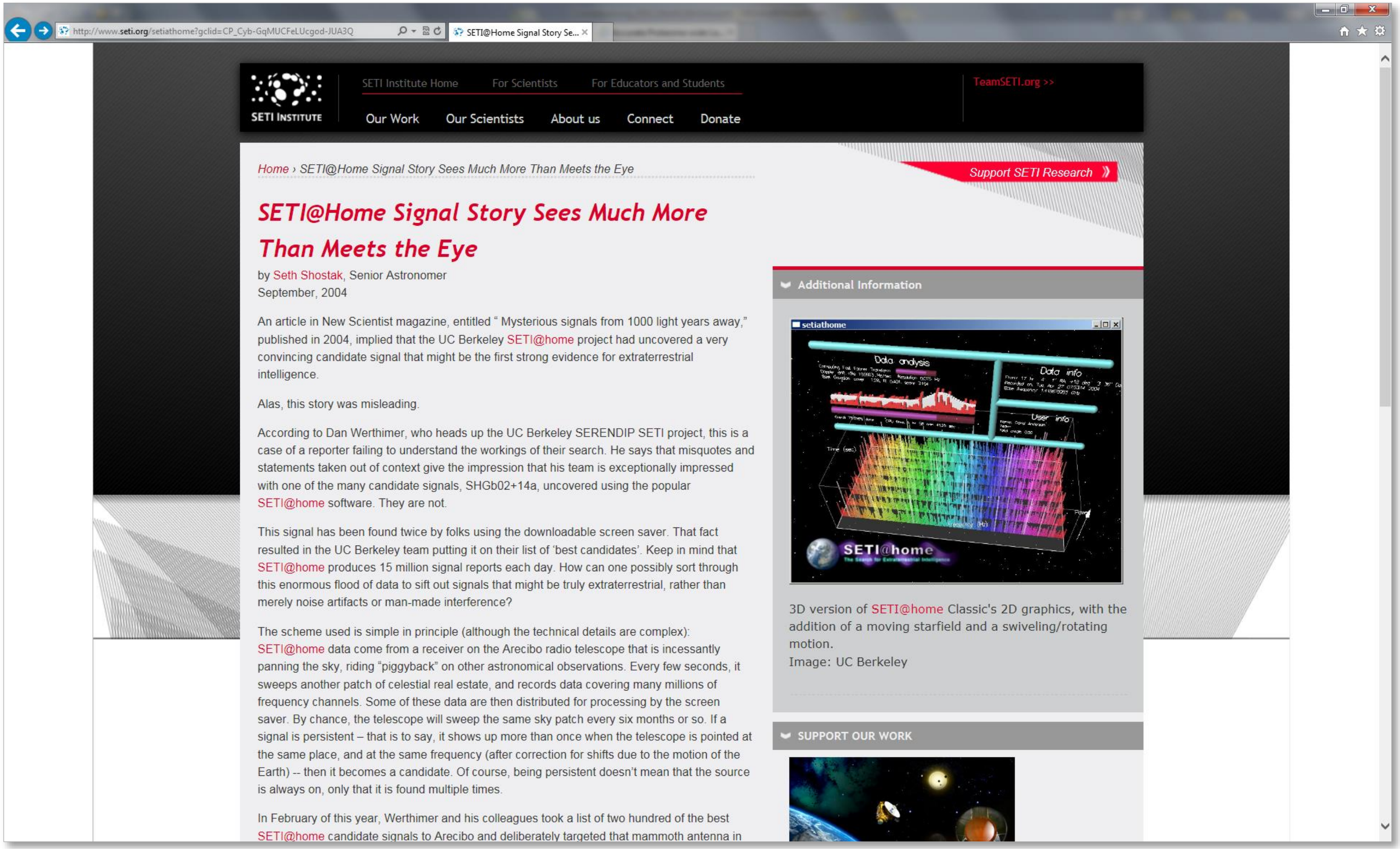
Previous penetrance estimates for all of the mutations can be revised by comparing the disease database to ExAC.

Crowdsourcing is the process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community, rather than from traditional employees or suppliers.

Crowdsourcing - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Crowdsourcing**



SETI@HOME



Home › SETI@Home Signal Story Sees Much More Than Meets the Eye

SETI@Home Signal Story Sees Much More Than Meets the Eye

by Seth Shostak, Senior Astronomer
September, 2004

An article in New Scientist magazine, entitled " Mysterious signals from 1000 light years away," published in 2004, implied that the UC Berkeley SETI@home project had uncovered a very convincing candidate signal that might be the first strong evidence for extraterrestrial intelligence.

Alas, this story was misleading.

According to Dan Werthimer, who heads up the UC Berkeley SERENDIP SETI project, this is a case of a reporter failing to understand the workings of their search. He says that misquotes and statements taken out of context give the impression that his team is exceptionally impressed with one of the many candidate signals, SHGb02+14a, uncovered using the popular SETI@home software. They are not.

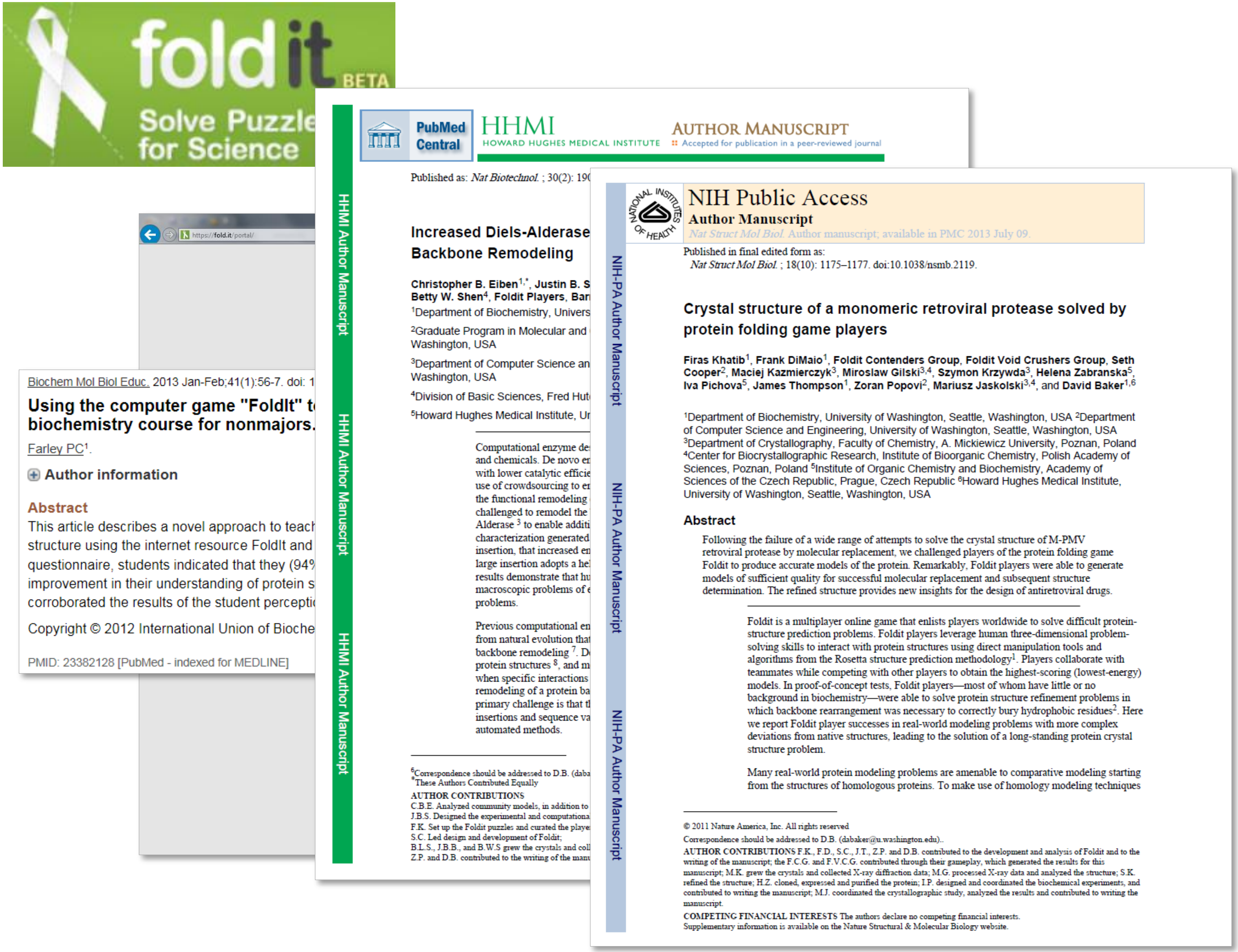
This signal has been found twice by folks using the downloadable screen saver. That fact resulted in the UC Berkeley team putting it on their list of 'best candidates'. Keep in mind that SETI@home produces 15 million signal reports each day. How can one possibly sort through this enormous flood of data to sift out signals that might be truly extraterrestrial, rather than merely noise artifacts or man-made interference?

The scheme used is simple in principle (although the technical details are complex): SETI@home data come from a receiver on the Arecibo radio telescope that is incessantly panning the sky, riding "piggyback" on other astronomical observations. Every few seconds, it sweeps another patch of celestial real estate, and records data covering many millions of frequency channels. Some of these data are then distributed for processing by the screen saver. By chance, the telescope will sweep the same sky patch every six months or so. If a signal is persistent – that is to say, it shows up more than once when the telescope is pointed at the same place, and at the same frequency (after correction for shifts due to the motion of the Earth) -- then it becomes a candidate. Of course, being persistent doesn't mean that the source is always on, only that it is found multiple times.

In February of this year, Werthimer and his colleagues took a list of two hundred of the best SETI@home candidate signals to Arecibo and deliberately targeted that mammoth antenna in

Additional Information

3D version of SETI@home Classic's 2D graphics, with the addition of a moving starfield and a swiveling/rotating motion.
Image: UC Berkeley



Biochem Mol Biol Educ. 2013 Jan-Feb;41(1):56-7. doi: 1

Using the computer game "FoldIt" t
biochemistry course for nonmajors.

Farley PC[1].

⊕ Author information

Abstract

This article describes a novel approach to teach
structure using the internet resource FoldIt and
questionnaire, students indicated that they (94%
improvement in their understanding of protein s
corroborated the results of the student percepti

Copyright © 2012 International Union of Bioche

PMID: 23382128 [PubMed - indexed for MEDLINE]

Published as: *Nat Biotechnol.* ; 30(2): 190

Increased Diels-Alderase
Backbone Remodeling

Christopher B. Eiben[1,*], Justin B. S
Betty W. Shen[4], Foldit Players, Bar

[1]Department of Biochemistry, Univers

[2]Graduate Program in Molecular and
Washington, USA

[3]Department of Computer Science an
Washington, USA

[4]Division of Basic Sciences, Fred Hut

[5]Howard Hughes Medical Institute, Ur

Computational enzyme de
and chemicals. De novo en
with lower catalytic efficie
use of crowdsourcing to en
the functional remodeling
challenged to remodel the
Alderase [3] to enable additi
characterization generated
insertion, that increased en
large insertion adopts a hel
results demonstrate that hu
macroscopic problems of e
problems.

Previous computational en
from natural evolution that
backbone remodeling [7]. D
protein structures [8], and m
when specific interactions
remodeling of a protein ba
primary challenge is that t
insertions and sequence va
automated methods.

[6]Correspondence should be addressed to D.B. (daba
[*]These Authors Contributed Equally

AUTHOR CONTRIBUTIONS

C.B.E. Analyzed community models, in addition to
J.B.S. Designed the experimental and computationa
F.K. Set up the Foldit puzzles and curated the playe
S.C. Led design and development of Foldit;
B.L.S., J.B.B., and B.W.S grew the crystals and coll
Z.P. and D.B. contributed to the writing of the manu

NIH Public Access
Author Manuscript
Nat Struct Mol Biol. Author manuscript; available in PMC 2013 July 09.

Published in final edited form as:
Nat Struct Mol Biol. ; 18(10): 1175–1177. doi:10.1038/nsmb.2119.

Crystal structure of a monomeric retroviral protease solved by protein folding game players

Firas Khatib[1], Frank DiMaio[1], Foldit Contenders Group, Foldit Void Crushers Group, Seth Cooper[2], Maciej Kazmierczyk[3], Miroslaw Gilski[3,4], Szymon Krzywda[3], Helena Zabranska[5], Iva Pichova[5], James Thompson[1], Zoran Popovi[2], Mariusz Jaskolski[3,4], and David Baker[1,6]

[1]Department of Biochemistry, University of Washington, Seattle, Washington, USA [2]Department of Computer Science and Engineering, University of Washington, Seattle, Washington, USA [3]Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University, Poznan, Poland [4]Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland [5]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Prague, Czech Republic [6]Howard Hughes Medical Institute, University of Washington, Seattle, Washington, USA

Abstract

Following the failure of a wide range of attempts to solve the crystal structure of M-PMV retroviral protease by molecular replacement, we challenged players of the protein folding game Foldit to produce accurate models of the protein. Remarkably, Foldit players were able to generate models of sufficient quality for successful molecular replacement and subsequent structure determination. The refined structure provides new insights for the design of antiretroviral drugs.

Foldit is a multiplayer online game that enlists players worldwide to solve difficult protein-structure prediction problems. Foldit players leverage human three-dimensional problem-solving skills to interact with protein structures using direct manipulation tools and algorithms from the Rosetta structure prediction methodology[1]. Players collaborate with teammates while competing with other players to obtain the highest-scoring (lowest-energy) models. In proof-of-concept tests, Foldit players—most of whom have little or no background in biochemistry—were able to solve protein structure refinement problems in which backbone rearrangement was necessary to correctly bury hydrophobic residues[2]. Here we report Foldit player successes in real-world modeling problems with more complex deviations from native structures, leading to the solution of a long-standing protein crystal structure problem.

Many real-world protein modeling problems are amenable to comparative modeling starting from the structures of homologous proteins. To make use of homology modeling techniques

© 2011 Nature America, Inc. All rights reserved

Correspondence should be addressed to D.B. (dabaker@u.washington.edu)..

AUTHOR CONTRIBUTIONS F.K., F.D., S.C., J.T., Z.P. and D.B. contributed to the development and analysis of Foldit and to the writing of the manuscript; the F.C.G. and F.V.C.G. contributed through their gameplay, which generated the results for this manuscript; M.K. grew the crystals and collected X-ray diffraction data; M.G. processed X-ray data and analyzed the structure; S.K. refined the structure; H.Z. cloned, expressed and purified the protein; I.P. designed and coordinated the biochemical experiments, and contributed to writing the manuscript; M.J. coordinated the crystallographic study, analyzed the results and contributed to writing the manuscript.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.
Supplementary information is available on the Nature Structural & Molecular Biology website.

nature COMMUNICATIONS

ARTICLE

Received 18 Apr 2016 | Accepted 12 Jul 2016 | Published 16 Sep 2016

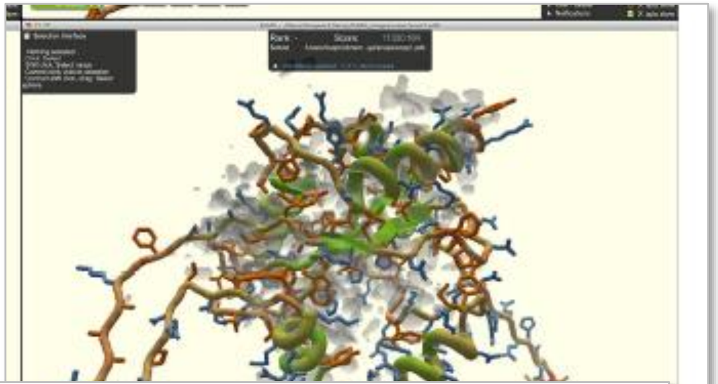
DOI: 10.1038/ncomms12549

OPEN

Determining crystal structures through crowdsourcing and coursework

Scot
Seth
Phili
Finn
Davi

We
Intro
spac
we h
unde
remo
Anal
histic
study
densi



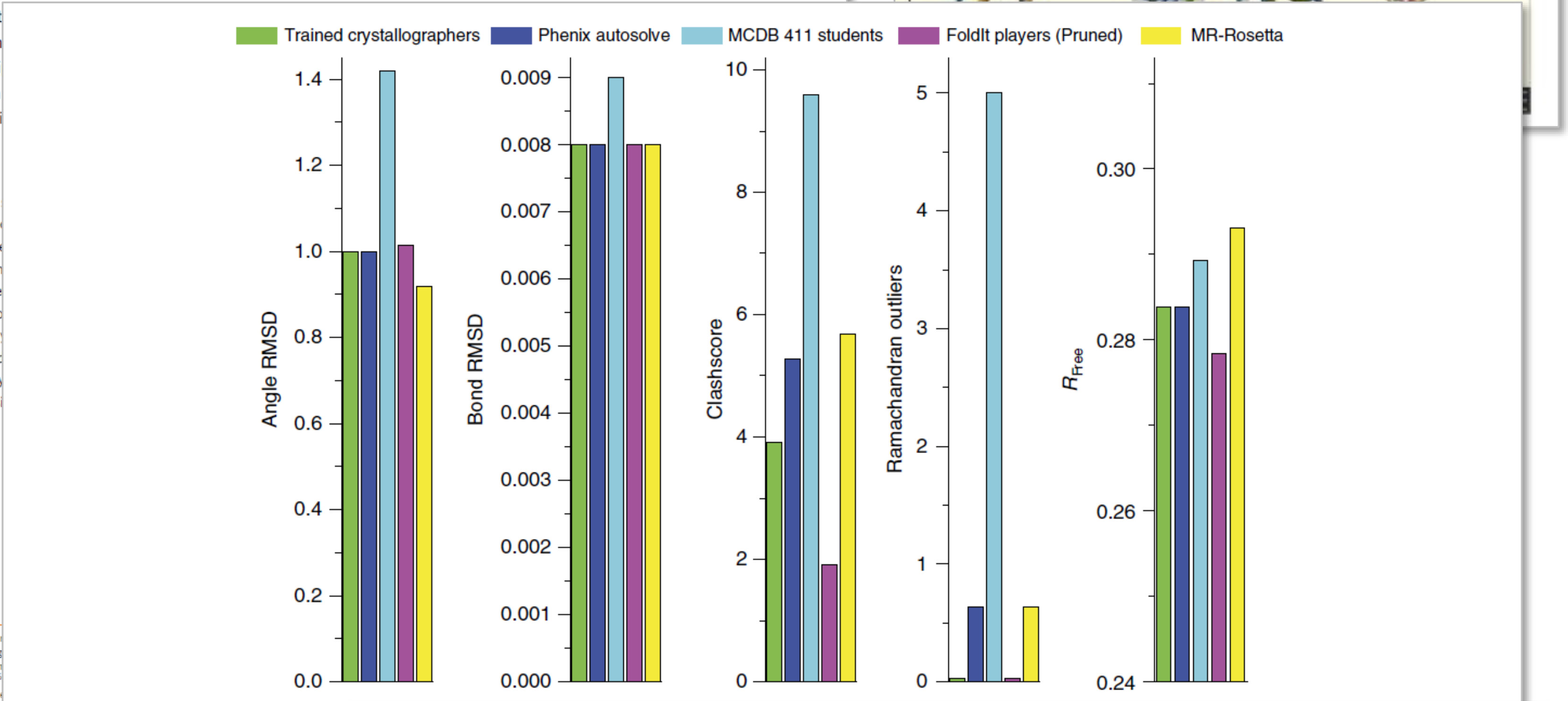


Figure 2 | Model-building comp
Foldit structures. In all cases, low

Here we show that Foldit players can build structural models at least as effectively as trained crystallographers and state-of-the-art automated methods, enabling a novel crowd-powered strategy for solving high-accuracy crystal structures. Combined with the

[1]Depa
Michig
Califor
USA. [6]
Scienc
USA. [9]
Michig
Arbor,
(ICS-4
Institut
Massa
(email:

†A list of consortium members appears at the end of the paper.

eterna

Solve Puzzles. Invent Medicine.

Empowering citizen scientists to invent medicine

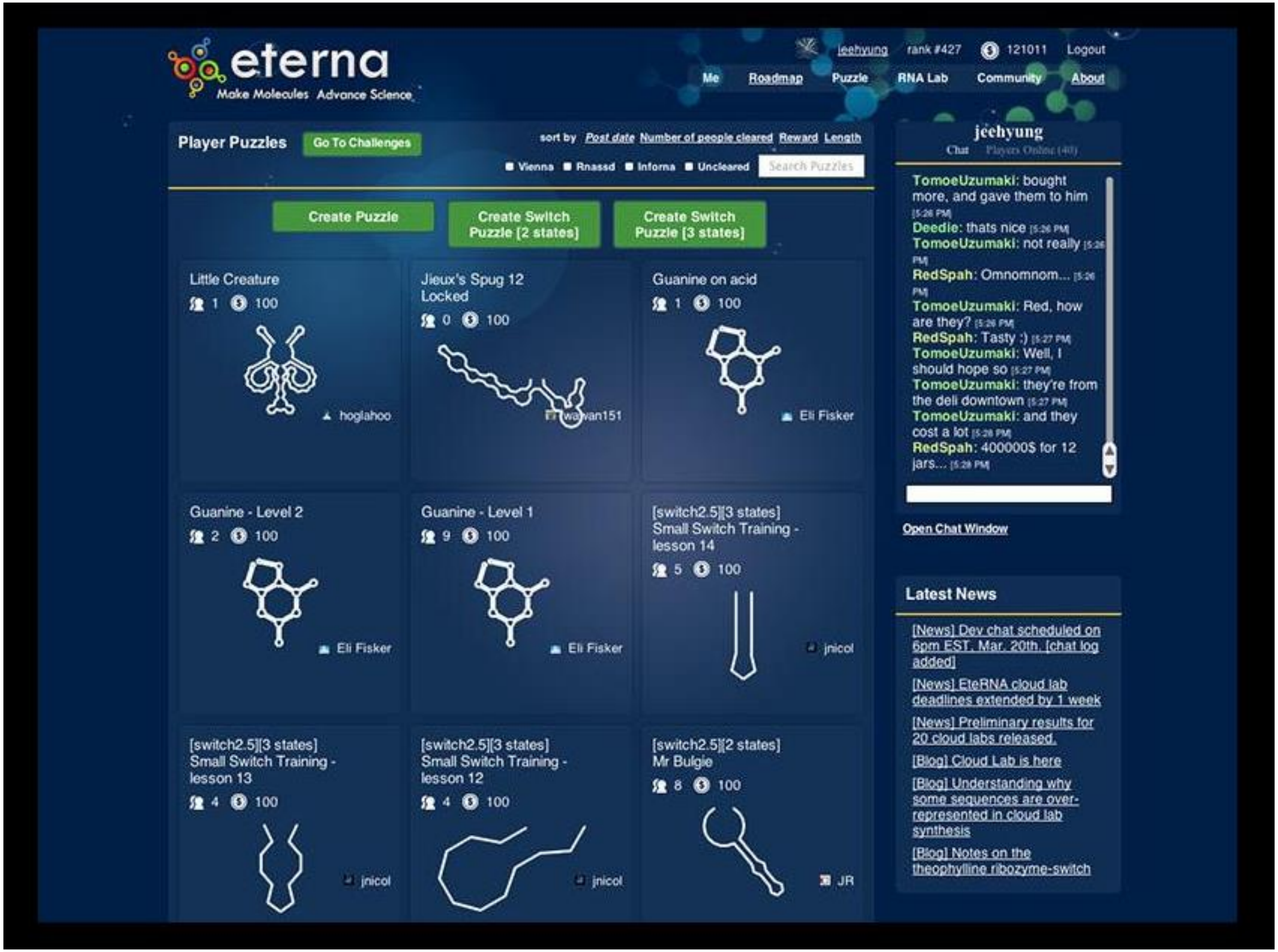
Solve puzzles to design molecular medicines.

Get feedback from real experiments at Stanford's School of Medicine.

Work together to **write papers** for scientific peer review.

Propose your own puzzles to advance research and **invent medicine**.

www.eternagame.org



eterna
Make Molecules Advance Science
jeehyung rank #427 121011 Logout
Me Roadmap Puzzle RNA Lab Community About

Player Puzzles

Go To Challenges
sort by Post date Number of people cleared Reward Length
■ Vienna ■ Rnassd ■ Inforna ■ Uncleared
Search Puzzles
Create Puzzle
Create Switch Puzzle [2 states]
Create Switch Puzzle [3 states]

Little Creature
1 100
hoglahoo

Jieux's Spug 12 Locked
0 100
wawan151

Guanine on acid
1 100
Eli Fisker

Guanine - Level 2
2 100
Eli Fisker

Guanine - Level 1
9 100
Eli Fisker

[switch2.5][3 states] Small Switch Training - lesson 14
5 100
jnicol

[switch2.5][3 states] Small Switch Training - lesson 13
4 100
jnicol

[switch2.5][3 states] Small Switch Training - lesson 12
4 100
jnicol

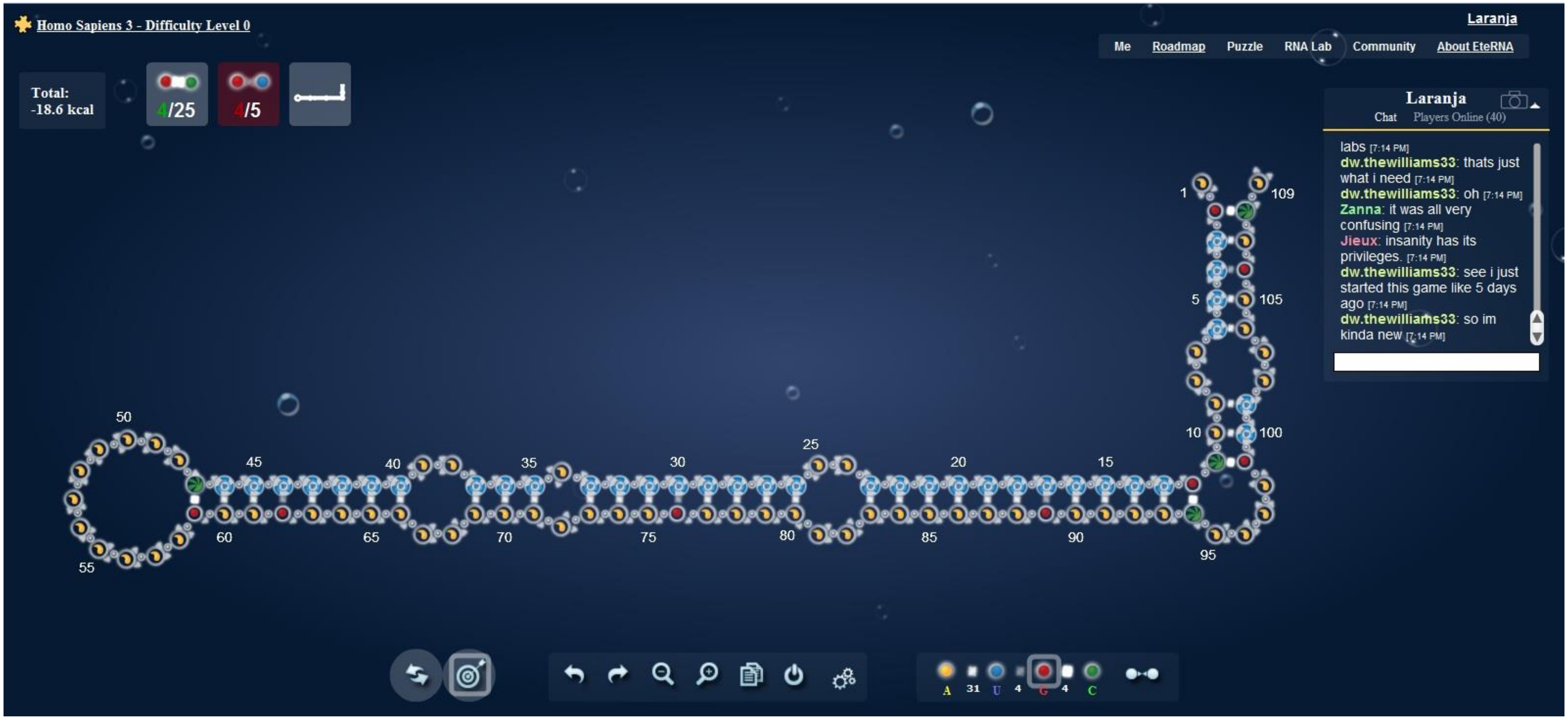
[switch2.5][2 states] Mr Bulgie
8 100
JR

jeehyung

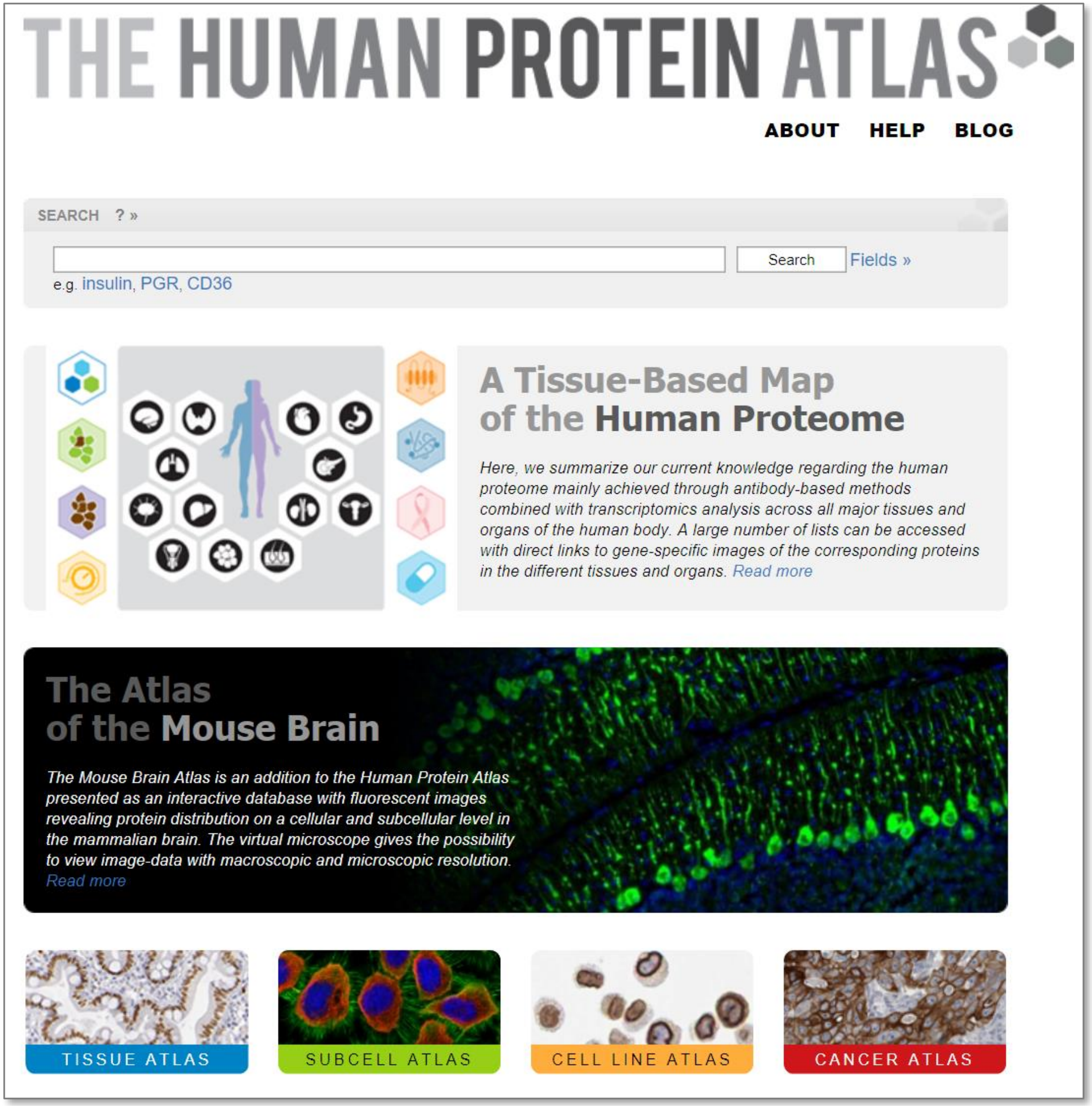
TomoeUzumaki: bought more, and gave them to him [5:26 PM]
Deedie: thats nice [5:26 PM]
TomoeUzumaki: not really [5:26 PM]
RedSpah: Omnomnom... [5:26 PM]
TomoeUzumaki: Red, how are they? [5:26 PM]
RedSpah: Tasty :) [5:27 PM]
TomoeUzumaki: Well, I should hope so [5:27 PM]
TomoeUzumaki: they're from the deli downtown [5:27 PM]
TomoeUzumaki: and they cost a lot [5:28 PM]
RedSpah: 400000$ for 12 jars... [5:28 PM]

Latest News

- [News] Dev chat scheduled on 6pm EST, Mar. 20th. [chat log added]
- [News] EteRNA cloud lab deadlines extended by 1 week
- [News] Preliminary results for 20 cloud labs released.
- [Blog] Cloud Lab is here
- [Blog] Understanding why some sequences are over-represented in cloud lab synthesis
- [Blog] Notes on the theophylline ribozyme-switch



labs [7:14 PM]
dw.thewilliams33: thats just what i need [7:14 PM]
dw.thewilliams33: oh [7:14 PM]
Zanna: it was all very confusing [7:14 PM]
Jieux: insanity has its privileges. [7:14 PM]
dw.thewilliams33: see i just started this game like 5 days ago [7:14 PM]
dw.thewilliams33: so im kinda new [7:14 PM]

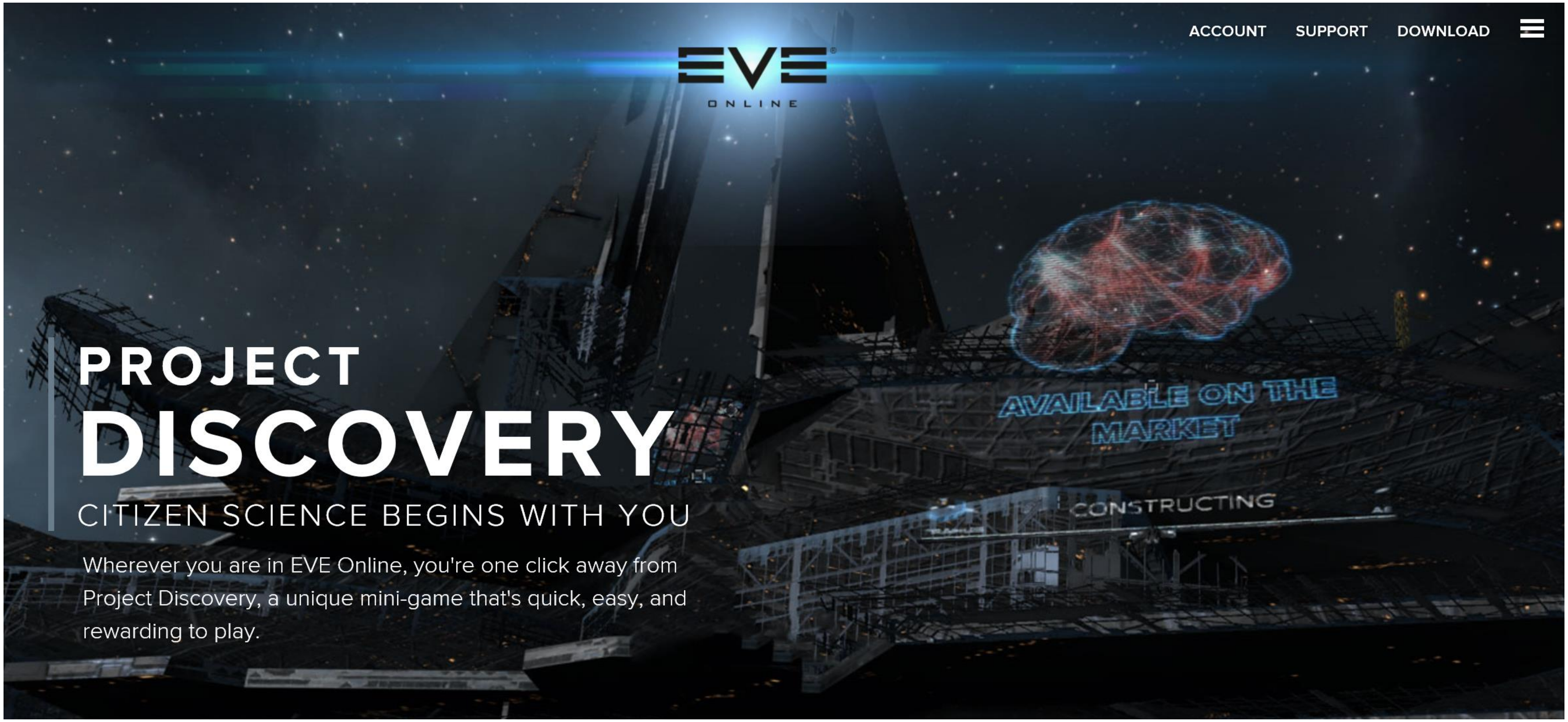


A Tissue-Based Map of the Human Proteome

Here, we summarize our current knowledge regarding the human proteome mainly achieved through antibody-based methods combined with transcriptomics analysis across all major tissues and organs of the human body. A large number of lists can be accessed with direct links to gene-specific images of the corresponding proteins in the different tissues and organs. Read more

The Atlas of the Mouse Brain

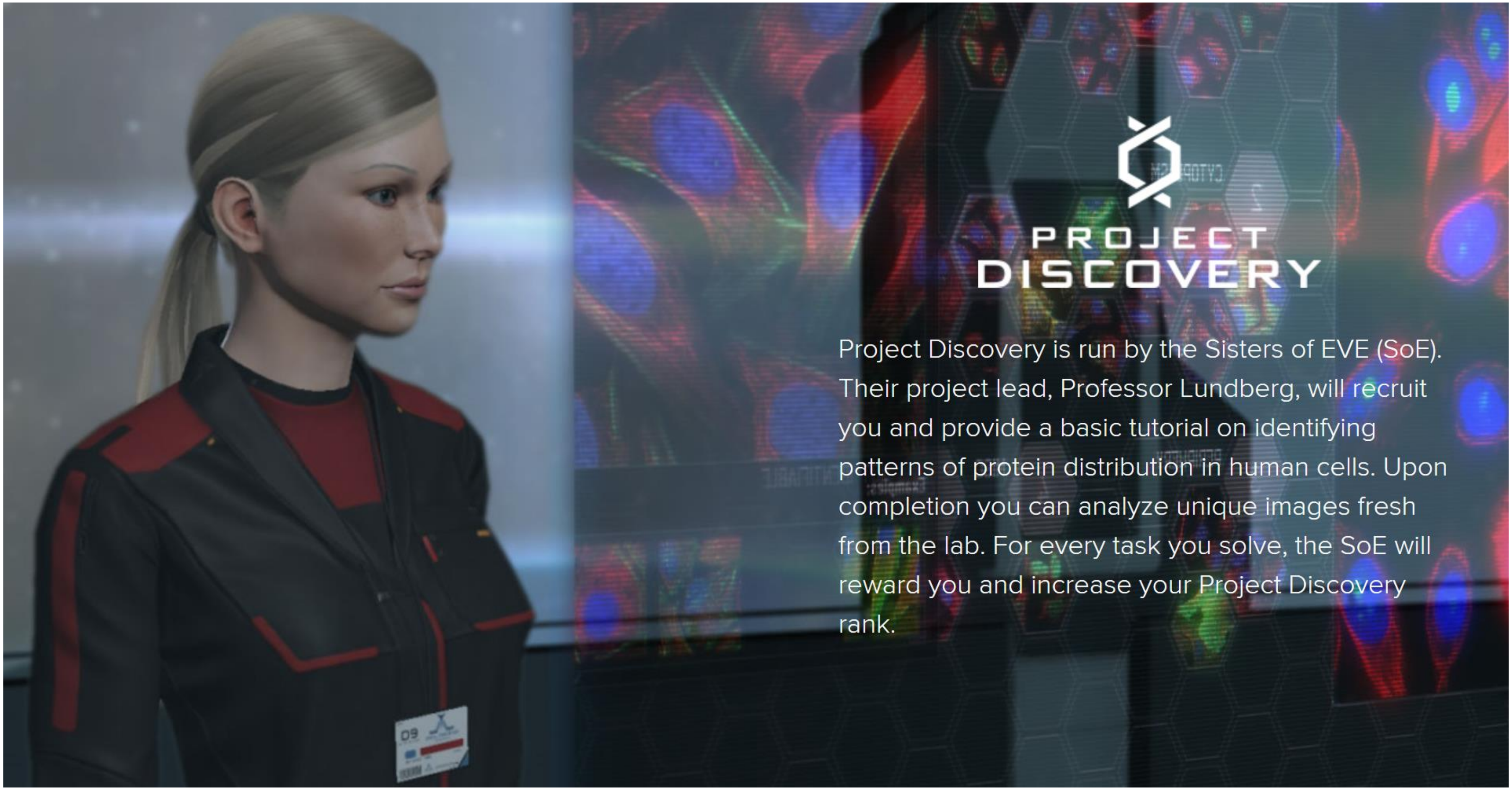
The Mouse Brain Atlas is an addition to the Human Protein Atlas presented as an interactive database with fluorescent images revealing protein distribution on a cellular and subcellular level in the mammalian brain. The virtual microscope gives the possibility to view image-data with macroscopic and microscopic resolution. Read more



PROJECT DISCOVERY

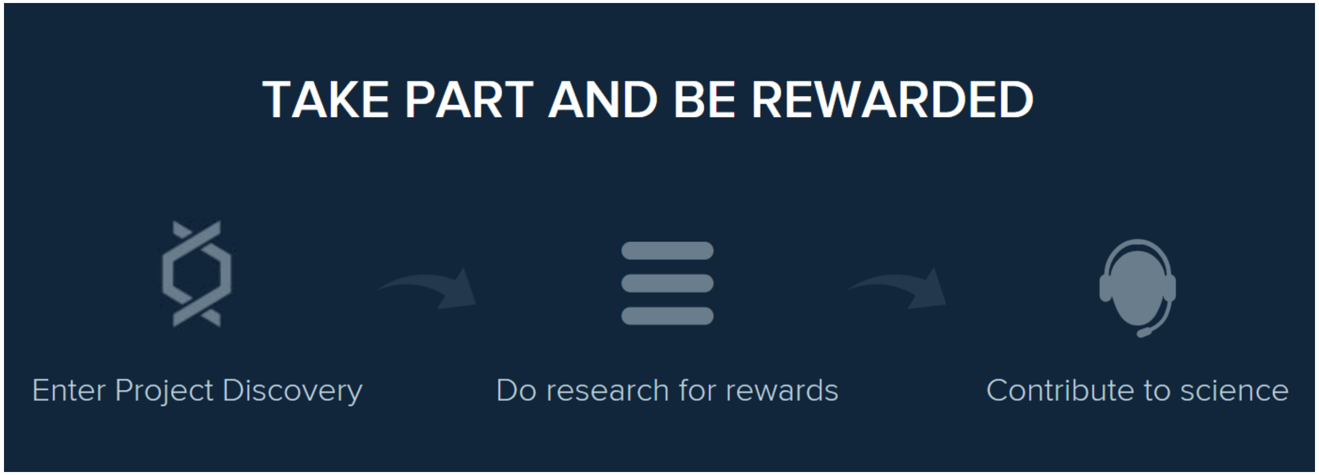
CITIZEN SCIENCE BEGINS WITH YOU

Wherever you are in EVE Online, you're one click away from Project Discovery, a unique mini-game that's quick, easy, and rewarding to play.



PROJECT DISCOVERY

Project Discovery is run by the Sisters of EVE (SoE). Their project lead, Professor Lundberg, will recruit you and provide a basic tutorial on identifying patterns of protein distribution in human cells. Upon completion you can analyze unique images fresh from the lab. For every task you solve, the SoE will reward you and increase your Project Discovery rank.

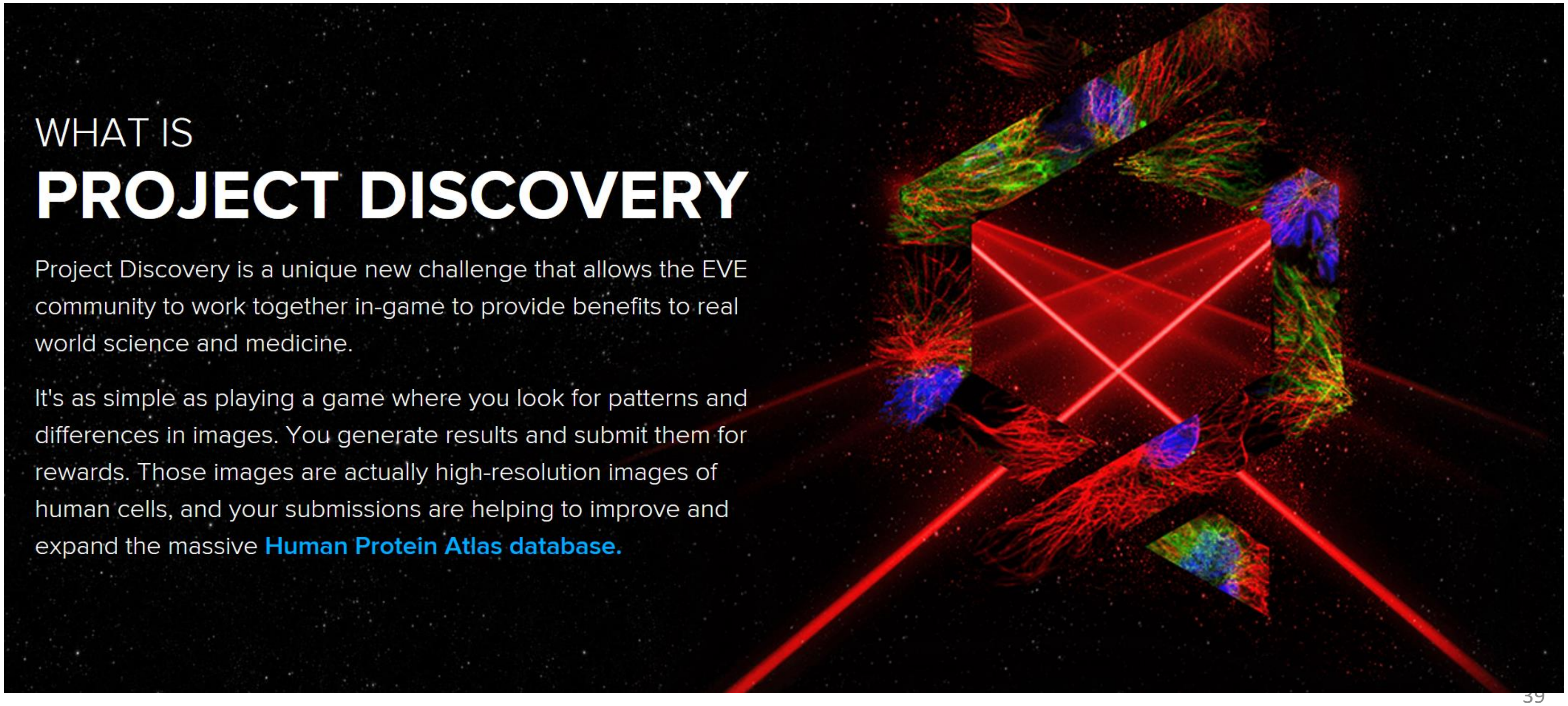


TAKE PART AND BE REWARDED

Enter Project Discovery

Do research for rewards

Contribute to science

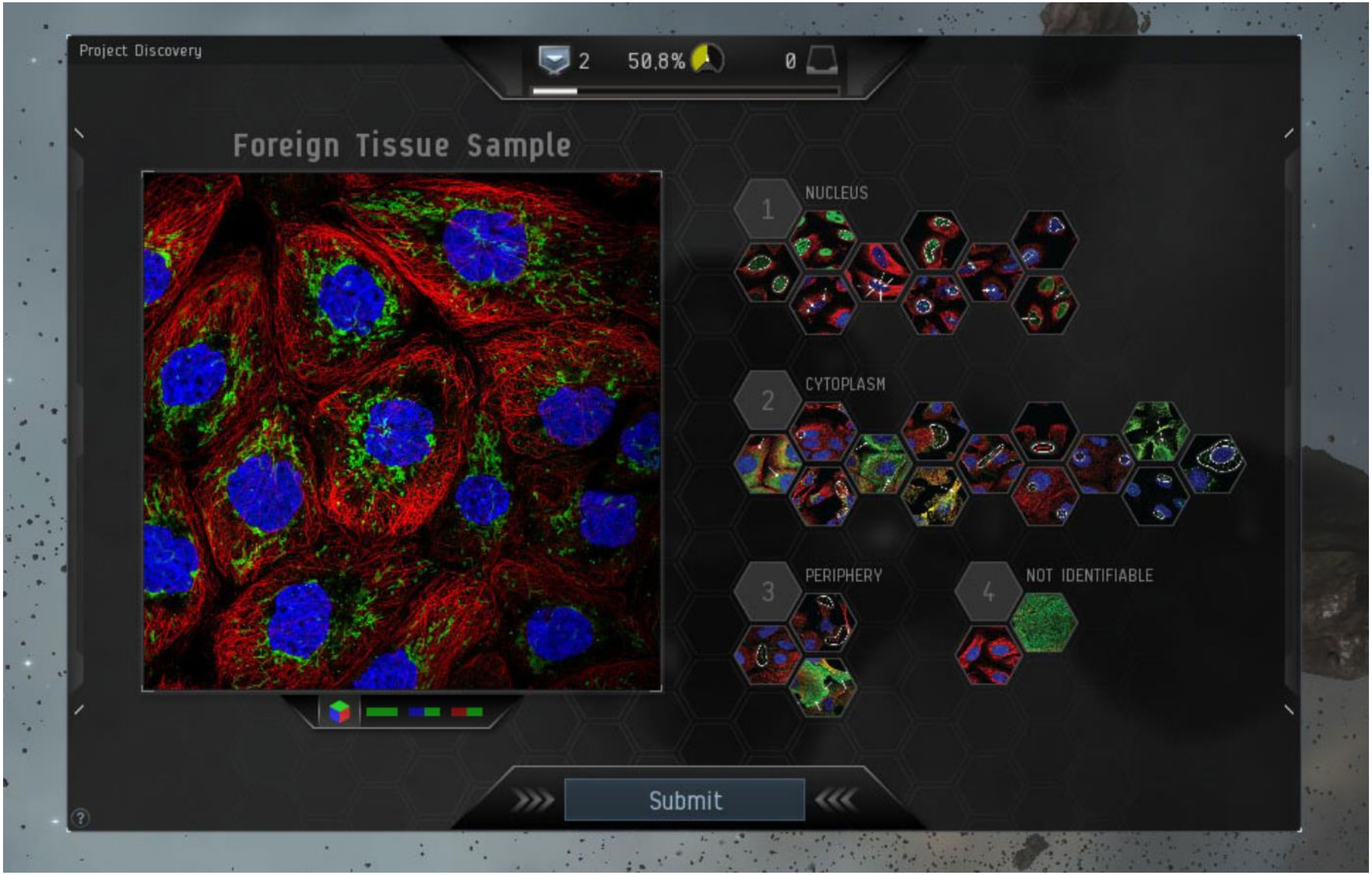


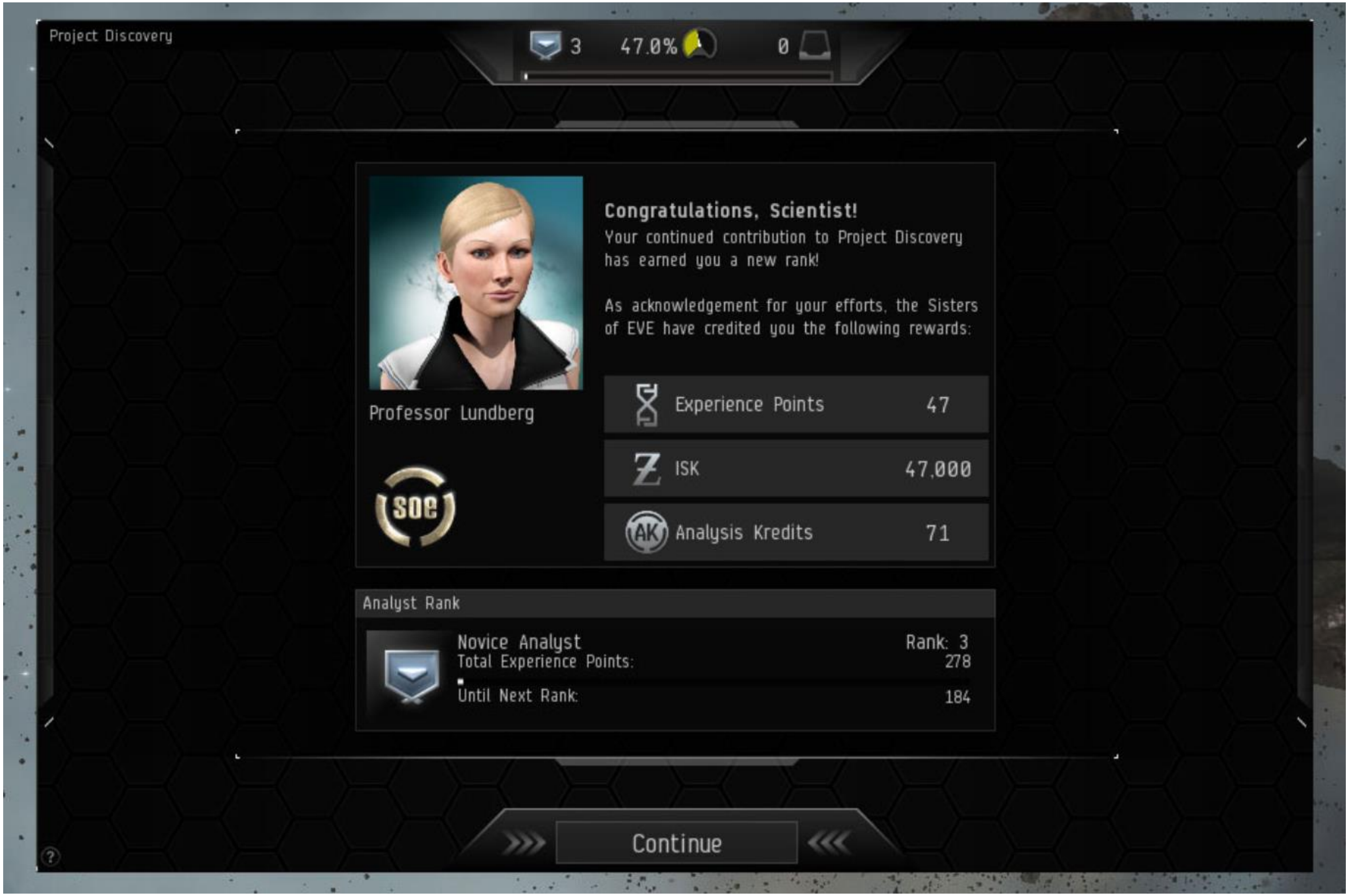
WHAT IS

PROJECT DISCOVERY

Project Discovery is a unique new challenge that allows the EVE community to work together in-game to provide benefits to real world science and medicine.

It's as simple as playing a game where you look for patterns and differences in images. You generate results and submit them for rewards. Those images are actually high-resolution images of human cells, and your submissions are helping to improve and expand the massive **Human Protein Atlas database.**


Project Discovery
2
50,8%
0
Foreign Tissue Sample
1
NUCLEUS
2
CYTOPLASM
3
PERIPHERY
4
NOT IDENTIFIABLE
Submit
?



Congratulations, Scientist!
Your continued contribution to Project Discovery has earned you a new rank!

As acknowledgement for your efforts, the Sisters of EVE have credited you the following rewards:

Experience Points	47
ISK	47,000
Analysis Kredits	71

Professor Lundberg

Analyst Rank

Novice Analyst — Rank: 3
Total Experience Points: 278
Until Next Rank: 184

J.R.R. Tolkien, A Conversation with Smaug

https://www.flickr.com/photos/fantasy-art-and-portraits/2884954207

Here is treasure of unlimited size, with all dragons chased away – now what will you do?

HOW
WHY
HOW
WHEN
WHERE
WHO
WHERE
WHERE
WHY
WHO
WHO
WHEN
WHAT
WHAT
WHY
WHAT
WHAT
WHEN
HOW
WHEN
WHO
WHERE
WHY
HOW
?