Timing, formatting and attribution for data access

Daniel Kohn kohnworkshop.com/

Myles Axton
Editor
Nature Genetics

LINCS data forum
Harvard Medical School
March 21th 2013

nature publishing group npg

Nature's mission statement written in 1869 still guides us today…

- First, to serve **scientists** through **prompt publication of significant advances** in any branch of science, and to provide a forum for the reporting and discussion of news and issues concerning science.

- Second, to ensure that the results of science are rapidly disseminated to the **public** throughout the world, in a fashion that conveys their **significance for knowledge, culture and daily life**.

Timing of access to human genome data

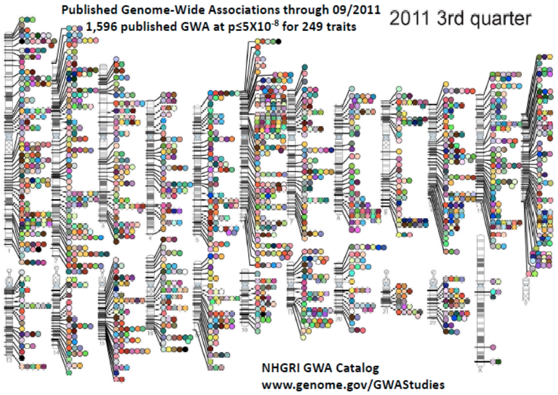
nature genetics

VOLUME 44 NUMBER 2 FEBRUARY 2012
www.nature.com/naturegenetics

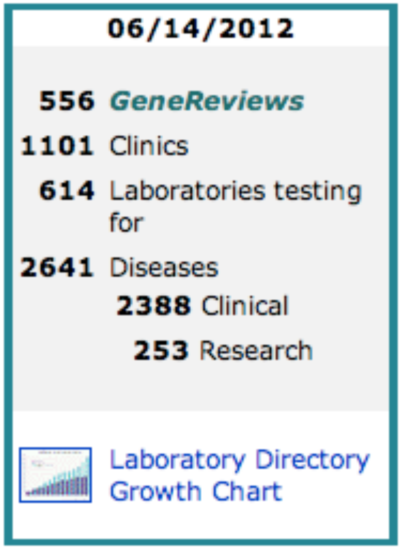
siRNA persistence in worms
Arabidopsis thaliana genome variation
Schistosoma haematobium genome

nature publishing group npg

Gene variants in human disease



GWAS
http://www.genome.gov/gwastudies
6/2/12
666 diseases and traits
1271 publications
313 in *Nat. Genet.*
1891/3869 $P<5x10^{-8}$ are in *Nat. Genet.*
6446 SNPs $P<10^{-5}$ total



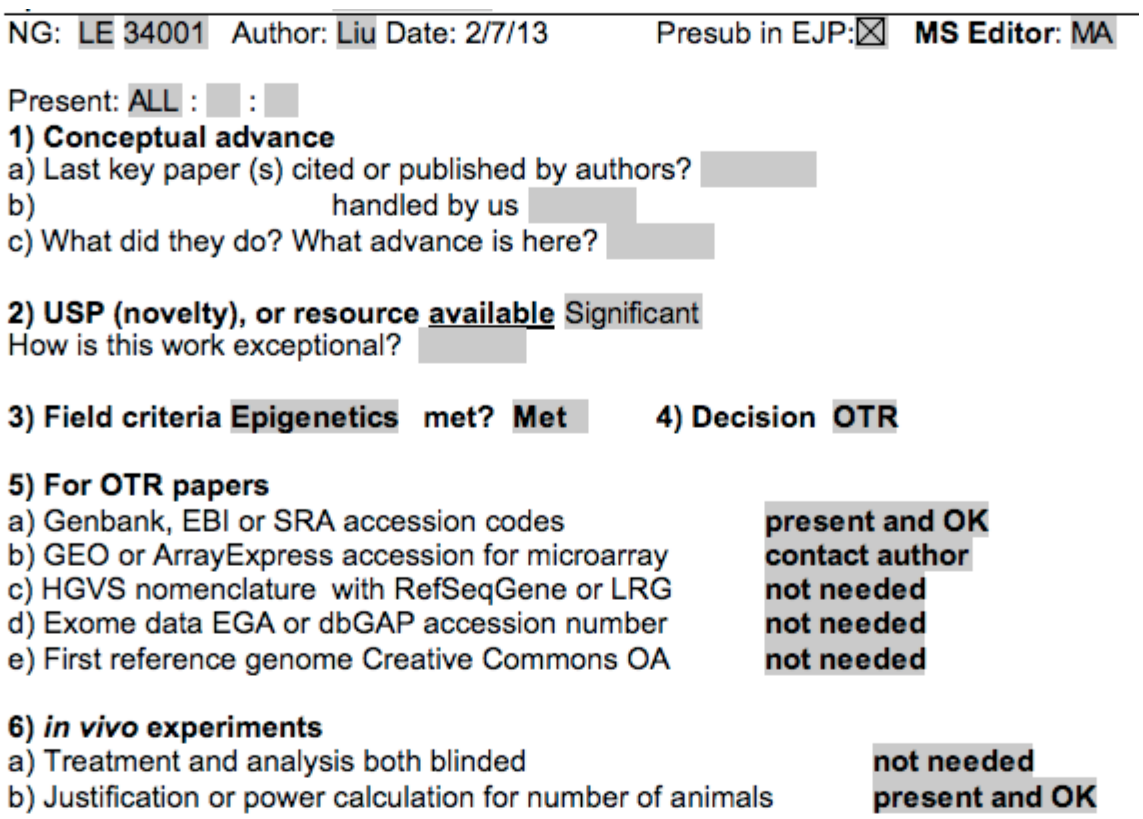
06/14/2012

556 ***GeneReviews***
1101 Clinics
614 Laboratories testing for
2641 Diseases
2388 Clinical
253 Research

Laboratory Directory Growth Chart

Mendelian disorders
www.genetests.org/

Initial manuscript assessment for peer review



NG: LE 34001 Author: Liu Date: 2/7/13 Presub in EJP:☒ **MS Editor**: MA

Present: ALL : :

1) Conceptual advance

a) Last key paper (s) cited or published by authors?

b) handled by us

c) What did they do? What advance is here?

2) USP (novelty), or resource <u>available</u> Significant

How is this work exceptional?

3) Field criteria Epigenetics met? Met **4) Decision OTR**

5) For OTR papers

a) Genbank, EBI or SRA accession codes **present and OK**

b) GEO or ArrayExpress accession for microarray **contact author**

c) HGVS nomenclature with RefSeqGene or LRG **not needed**

d) Exome data EGA or dbGAP accession number **not needed**

e) First reference genome Creative Commons OA **not needed**

6) *in vivo* experiments

a) Treatment and analysis both blinded **not needed**

b) Justification or power calculation for number of animals **present and OK**

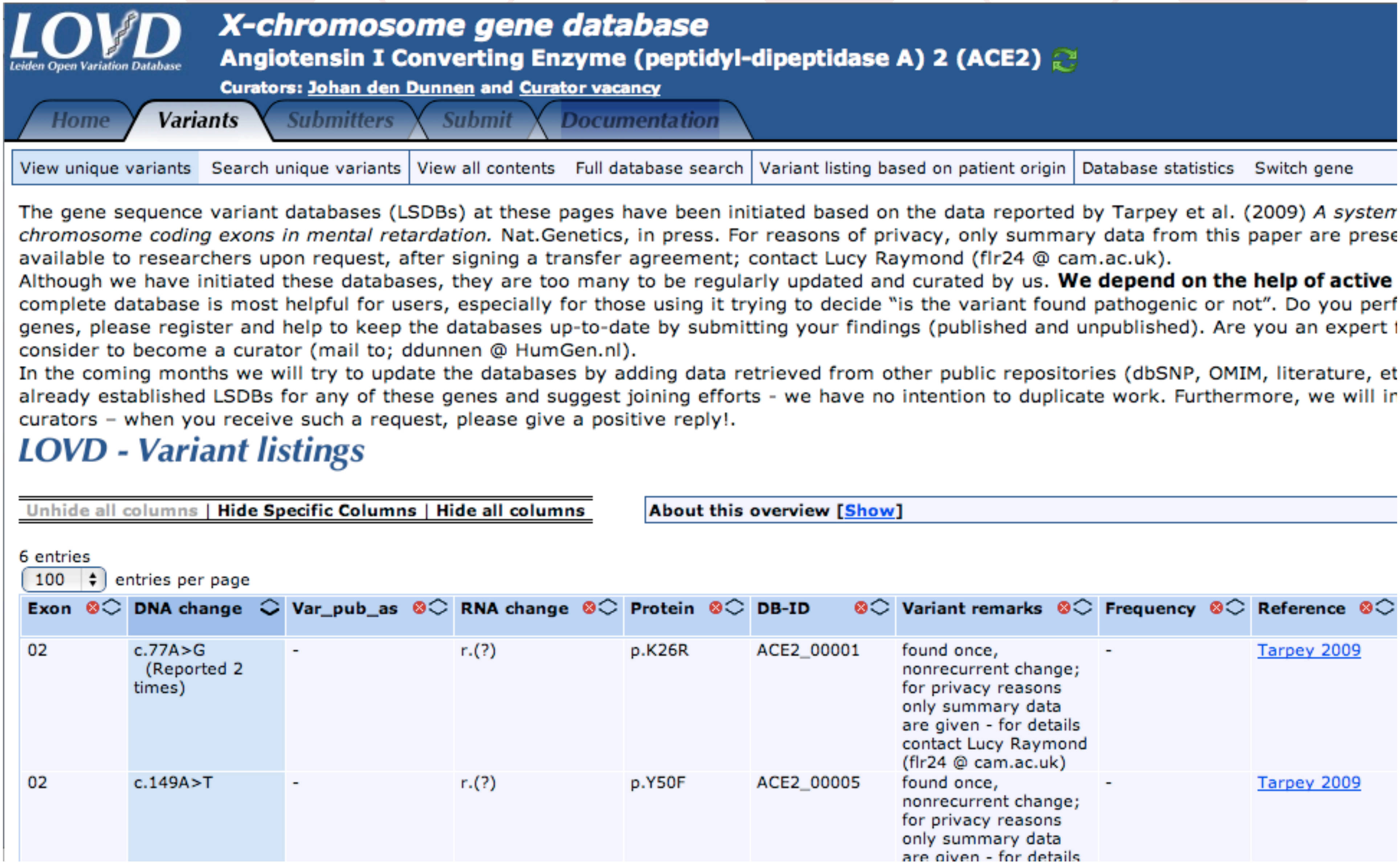
"Data available for referees => data available upon manuscript acceptance"

A quantum of attribution

nature genetics
VOLUME 36 NUMBER 7 JULY 2004
www.nature.com/naturegenetics
Knockout cattle
TFIIH in trichothiodystrophy
Kaposi virus reprograms cells

Alex Beard www.alexbeardstudio.com/

Coding variants deposited in gene database



X-chromosome gene database

Angiotensin I Converting Enzyme (peptidyl-dipeptidase A) 2 (ACE2)

Curators: Johan den Dunnen and Curator vacancy

View unique variants | Search unique variants | View all contents | Full database search | Variant listing based on patient origin | Database statistics | Switch gene

The gene sequence variant databases (LSDBs) at these pages have been initiated based on the data reported by Tarpey et al. (2009) *A systen*
chromosome coding exons in mental retardation. Nat.Genetics, in press. For reasons of privacy, only summary data from this paper are prese
available to researchers upon request, after signing a transfer agreement; contact Lucy Raymond (flr24 @ cam.ac.uk).
Although we have initiated these databases, they are too many to be regularly updated and curated by us. **We depend on the help of active**
complete database is most helpful for users, especially for those using it trying to decide "is the variant found pathogenic or not". Do you perf
genes, please register and help to keep the databases up-to-date by submitting your findings (published and unpublished). Are you an expert
consider to become a curator (mail to; ddunnen @ HumGen.nl).
In the coming months we will try to update the databases by adding data retrieved from other public repositories (dbSNP, OMIM, literature, et
already established LSDBs for any of these genes and suggest joining efforts - we have no intention to duplicate work. Furthermore, we will ir
curators – when you receive such a request, please give a positive reply!.

LOVD - Variant listings

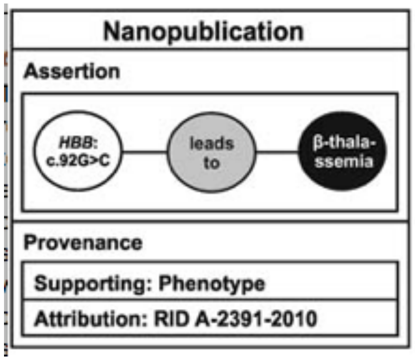
Unhide all columns | **Hide Specific Columns** | **Hide all columns**

About this overview [Show]

6 entries
100 entries per page

Exon	DNA change	Var_pub_as	RNA change	Protein	DB-ID	Variant remarks	Frequency	Reference
02	c.77A>G (Reported 2 times)	-	r.(?)	p.K26R	ACE2_00001	found once, nonrecurrent change; for privacy reasons only summary data are given - for details contact Lucy Raymond (flr24 @ cam.ac.uk)	-	Tarpey 2009
02	c.149A>T	-	r.(?)	p.Y50F	ACE2_00005	found once, nonrecurrent change; for privacy reasons only summary data are given - for details	-	Tarpey 2009

Microattribution and nanopublication as means to incentivize the placement of human genome variation data into the public domain



Nanopublication

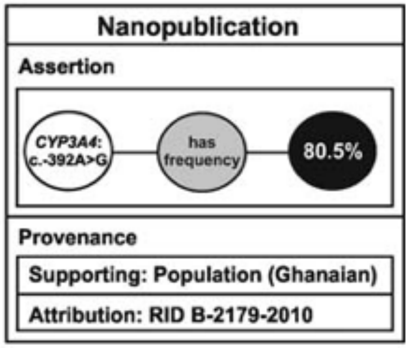
Assertion

HBB: c.92G>C — leads to — β-thalassemia

Provenance

Supporting: Phenotype

Attribution: RID A-2391-2010



Nanopublication

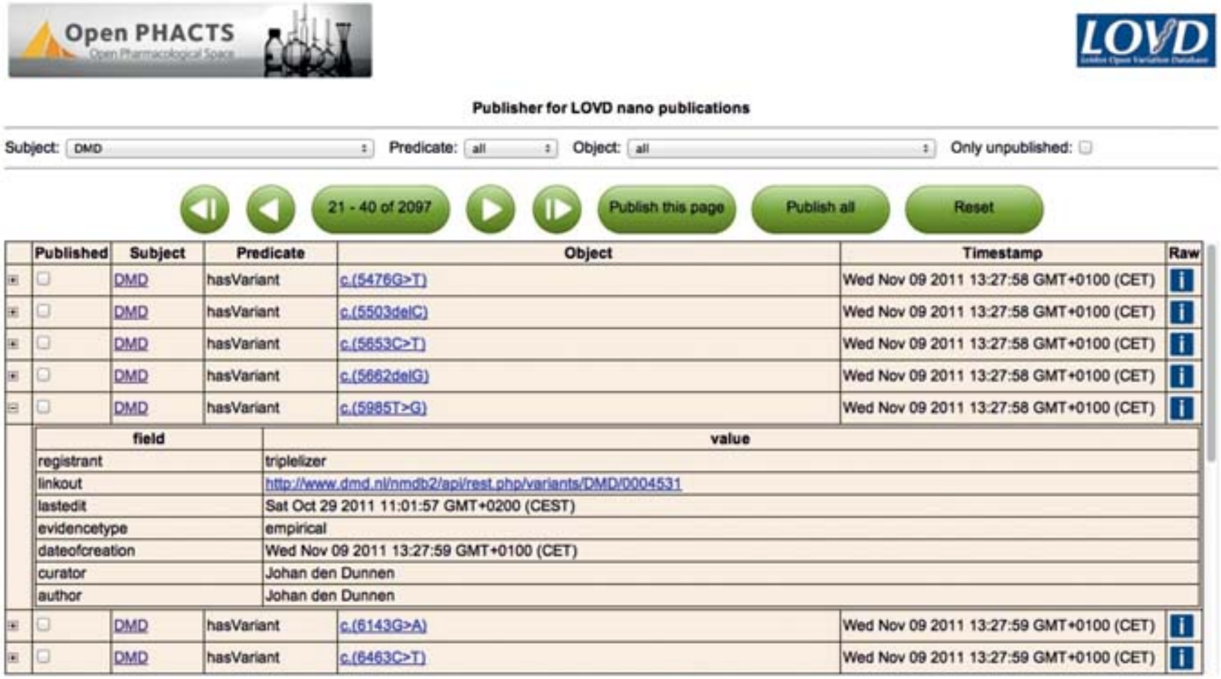
Assertion

CYP3A4: c.-392A>G — has frequency — 80.5%

Provenance

Supporting: Population (Ghanaian)

Attribution: RID B-2179-2010



Publisher for LOVD nano publications

Subject: DMD Predicate: all Object: all Only unpublished: □

	Published	Subject	Predicate	Object	Timestamp	Raw
⊞	□	DMD	hasVariant	c.(5476G>T)	Wed Nov 09 2011 13:27:58 GMT+0100 (CET)	i
⊞	□	DMD	hasVariant	c.(5503delC)	Wed Nov 09 2011 13:27:58 GMT+0100 (CET)	i
⊞	□	DMD	hasVariant	c.(5653C>T)	Wed Nov 09 2011 13:27:58 GMT+0100 (CET)	i
⊞	□	DMD	hasVariant	c.(5662delG)	Wed Nov 09 2011 13:27:58 GMT+0100 (CET)	i
⊟	□	DMD	hasVariant	c.(5985T>G)	Wed Nov 09 2011 13:27:58 GMT+0100 (CET)	i

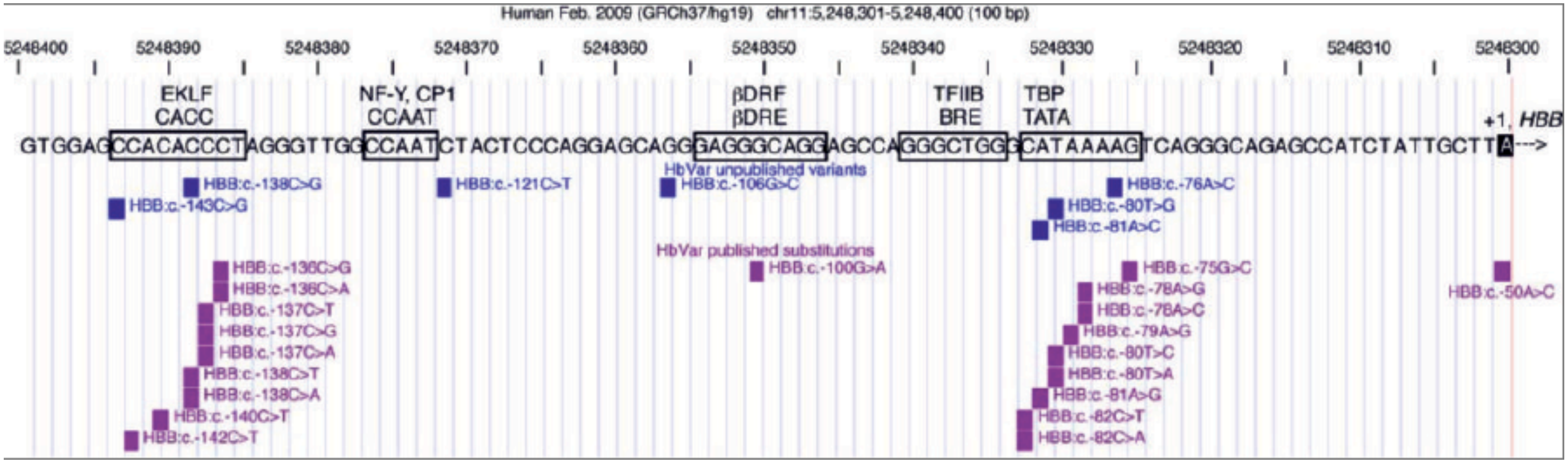
field	value
registrant	triplelizer
linkout	http://www.dmd.nl/nmdb2/api/rest.php/variants/DMD/0004531
lastedit	Sat Oct 29 2011 11:01:57 GMT+0200 (CEST)
evidencetype	empirical
dateofcreation	Wed Nov 09 2011 13:27:59 GMT+0100 (CET)
curator	Johan den Dunnen
author	Johan den Dunnen

	Published	Subject	Predicate	Object	Timestamp	Raw
⊞	□	DMD	hasVariant	c.(6143G>A)	Wed Nov 09 2011 13:27:59 GMT+0100 (CET)	i
⊞	□	DMD	hasVariant	c.(6463C>T)	Wed Nov 09 2011 13:27:59 GMT+0100 (CET)	i

Human Mutation
Volume 33, Issue 11, pages 1503-1512, 23 JUL 2012 DOI: 10.1002/humu.22144
http://onlinelibrary.wiley.com/doi/10.1002/humu.22144/full#fig3

Human Variome Microattribution Review (Hemoglobin)

http://www.bx.psu.edu/~giardine/

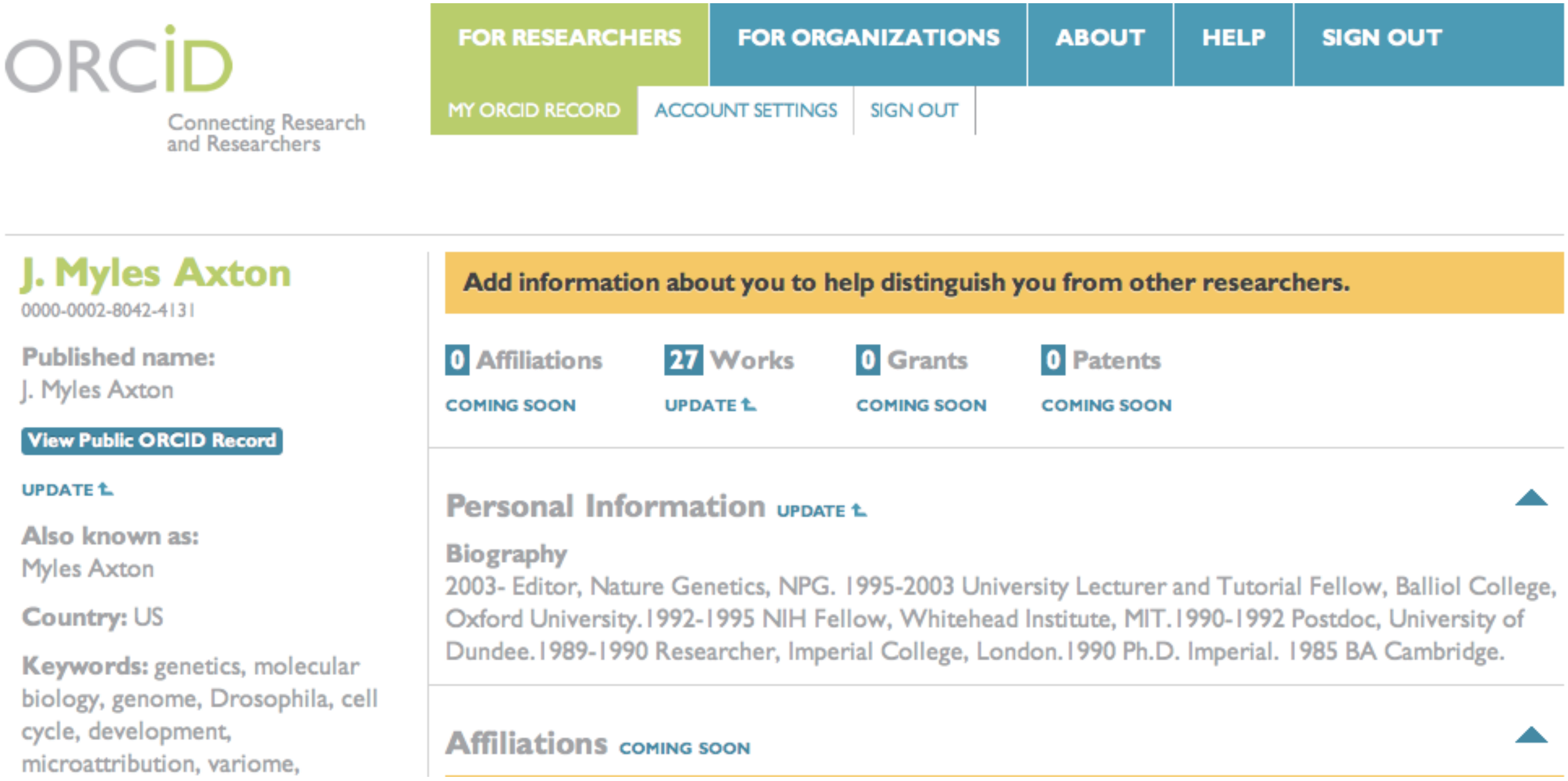


Microattribution : "giving database accessions the same citation conventions and indices that journal articles currently enjoy"
http://en.wikipedia.org/wiki/Microattribution

Giardine, B. *et al. Nat. Genet.* ***43,*** 295–301 (2011) doi:10.1038/**ng.785**

nature publishing group npg

A unique identifier allows you to control your reputation



J. Myles Axton

0000-0002-8042-4131

Published name:
J. Myles Axton

View Public ORCID Record

UPDATE

Also known as:
Myles Axton

Country: US

Keywords: genetics, molecular biology, genome, Drosophila, cell cycle, development, microattribution, variome,

Add information about you to help distinguish you from other researchers.

0 Affiliations	27 Works	0 Grants	0 Patents
COMING SOON	UPDATE	COMING SOON	COMING SOON

Personal Information UPDATE

Biography
2003- Editor, Nature Genetics, NPG. 1995-2003 University Lecturer and Tutorial Fellow, Balliol College, Oxford University.1992-1995 NIH Fellow, Whitehead Institute, MIT.1990-1992 Postdoc, University of Dundee.1989-1990 Researcher, Imperial College, London.1990 Ph.D. Imperial. 1985 BA Cambridge.

Affiliations COMING SOON

Nature is a founding partner of ORCID

nature publishing group npg

Access to data

nature
genetics
VOLUME 37 NUMBER 5 MAY 2005
www.nature.com/naturegenetics
Epistasis by microarray
Imprinting in preeclampsia
Combinatorial miRNA

Eve Stockton www.evestockton.com/

nature publishing group npg

It's not data, it's *my* data!

"Scientists are trained to recognize that correlation is not causation, that no conclusions should be drawn simply on the basis of correlation between X and Y (it could just be a coincidence). Instead, you must understand the underlying mechanisms that connect the two. Once you have a model, you can connect the data sets with confidence. Data without a model is just noise.

The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. **Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all."**

Chris Anderson - *Wired* 06.23.08, The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

http://www.wired.com/science/discoveries/magazine/16-07/pb_theory

"Chris Anderson, the editor of *Wired* magazine wrote in 2008 that the sheer volume of data would obviate the need for theory , and even the scientific method.

…But …..these views are badly mistaken. **The numbers have not way of speaking for themselves. We speak for them. We imbue them with meaning.**

Big data will produce progress – eventually. How quickly it does and whether we regress in the meantime, will depend on us.

Our biological instincts are not always very well adapted to the information-rich modern world. Unless we work *actively* to become aware of the biases we introduce, the returns to additional information may be minimal – or diminishing.

Meanwhile, if the quantity of information is increasing by 2.5 quintillion bytes per day, the amount of *useful* information almost certainly isn't. **Most of it is just noise, and the noise is increasing faster than the signal."**

Nate Silver – *The Signal and the Noise*. Penguin Press NY, 2012

EDITORIAL

nature genetics

It's not about the data

Researchers, funders and journals are in broad agreement that data must be accessible to support the conclusions of scientific publications and for the research to have impact. What is lacking is agreement on timing, formatting and attribution.

nature genetics **| volume 44 | number 2 | February 2012 | 111**

While keeping up pressure for access to data resources (“No second thoughts about data access”; *Nat. Genet. 43, 389, 2011)* we have been advocating the use of **citable data management plans** in line with the proposals of major funding agencies.

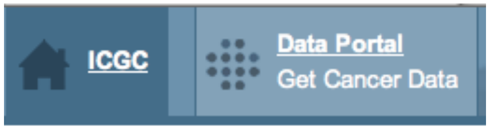
The plan finally matures into a ‘**data descriptor**’, which we define as a user guide to the resources, accession codes and use conditions accompanying a completed project or publication.

Most repositories are designed for specific assay types, necessitating the fragmentation of complex datasets.
Metadata formatting will be needed to ensure that biomedical research datasets become interoperable.
This solution is the overarching **ISA** framework, where the acronym stands for ‘**Investigation**’ (the project context), ‘**Study**’ (a unit of research) and ‘**Assay**’ (analytical measurement).

nature publishing group npg

Evolution of data management plans for the 39 International Cancer Genome Consortium projects

STAGE 0
34 studies with data portal, *Nature* ‘marker paper’ explaining the data release strategy.
http://dcc.icgc.org/
Nature 15 April 2010 doi:10.1038/**nature08987**
International network of cancer genome projects



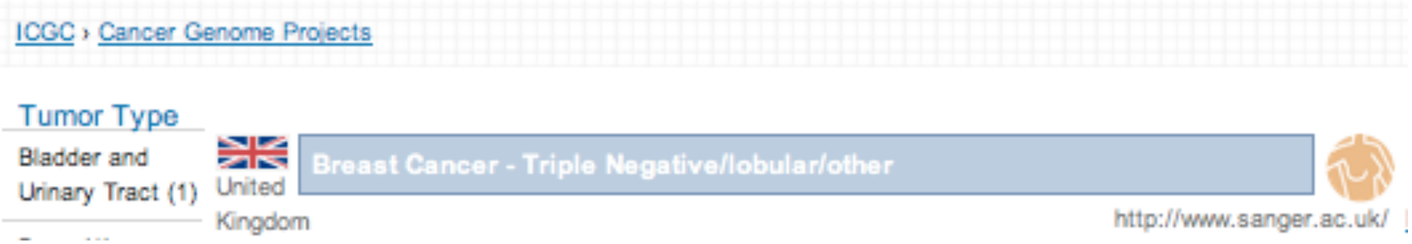
International
Cancer Genome
Consortium

STAGE 1
2 studies explain the project
Gastric cancer, intestinal and diffuse type – China
Oral cancer- gingivobuccal – India

STAGE 2
2 Studies have a detailed data management plan
Breast_Carcinoma-WTSI-UK-1
Pancreatic_Cancer-OICR-CA-1

STAGE 3
1 Study has a data descriptor in dbGAP, database with accession code **phs000370.v1.p1**
an associated *Science* publication and 883 sequence data depositions in SRA database
July 28 2011 DOI: 10.1126/**science.1208130**
The Mutational Landscape of Head and Neck Squamous Cell Carcinoma
Nicolas Stransky *et al.*

STAGE 2 – Data management plan



ICGC › Cancer Genome Projects

Tumor Type

Bladder and Urinary Tract (1)

United Kingdom

Breast Cancer - Triple Negative/lobular/other

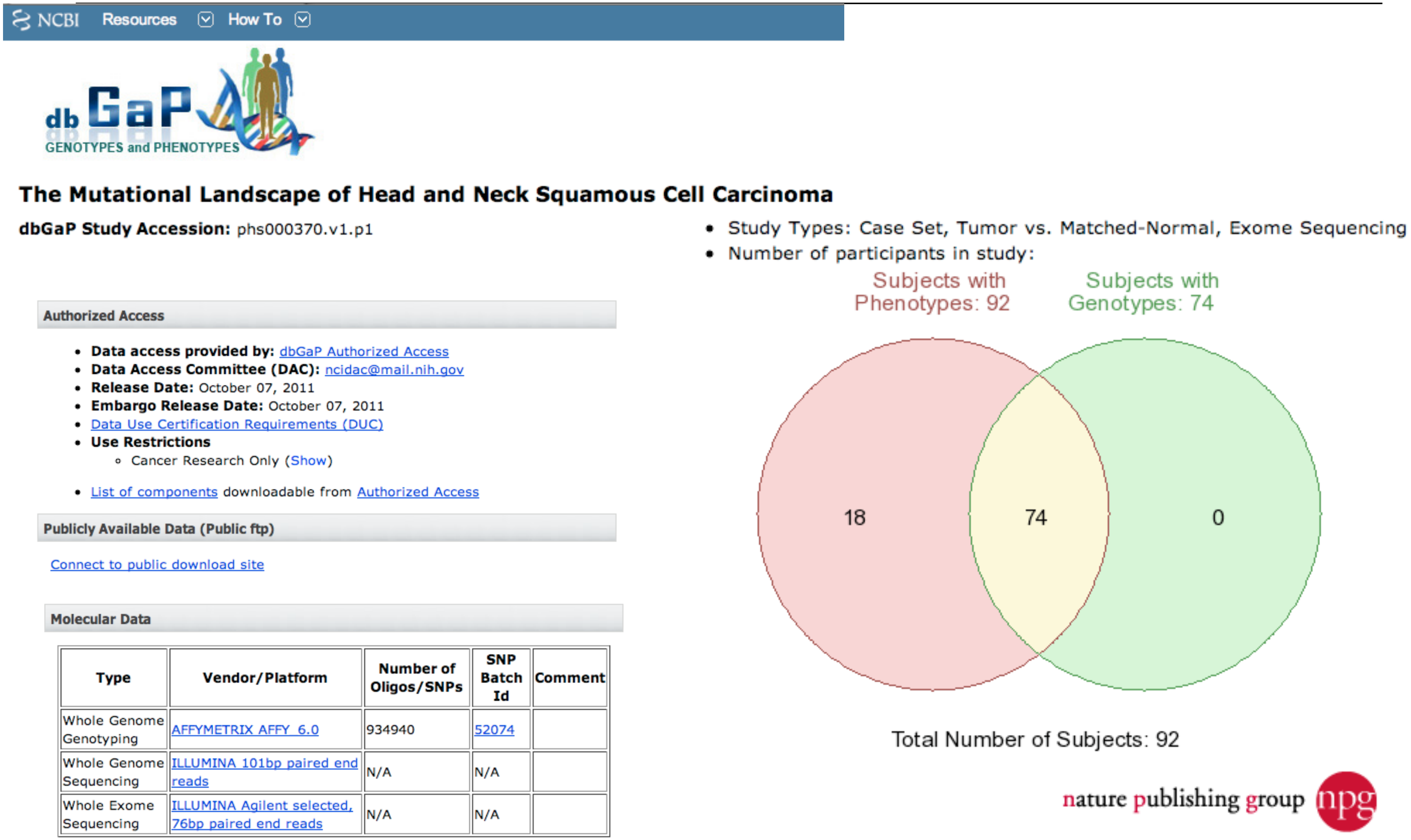
http://www.sanger.ac.uk/

Whole genome sequencing of 100 tumor/normal pairs.
Time limits for publication moratoriums: All data shall become free of a publication moratorium when either the data is published by the ICGC member project or one year after the specified quantity of data
(e.g. genome dataset from 100 tumours per project) has been released via the ICGC database or other public databases.
In all cases data shall be free of a publication moratorium two years after its initial release.

Project summary: http://www.icgc.org/

Files in directory sv_sangerBreast.txt:
matched_sample_id --- Unique identifier for the control matched to the tumour sample.
tumour_sample_id --- Unique identifier for the tumour sample donated by the donor.
variant_type --- Type of mutation/variation.
assembly_version --- Version of reference genome assembly.
chr_from --- Name of the donor chromosome containing the mutation/variation.
chr_from_bkpt --- Breakpoint position of the mutation/variation on the donor chromosome.
chr_from_strand --- Donor chromome strand.
chr_to --- Name of the acceptor chromosome containing the mutation/variation.
chr_to_bkpt --- Breakpoint position of the mutation/variation on the acceptor chromosome.
chr_to_strand --- Acceptor chromosome strand.

STAGE 3 – Data descriptor



The Mutational Landscape of Head and Neck Squamous Cell Carcinoma

dbGaP Study Accession: phs000370.v1.p1

- Study Types: Case Set, Tumor vs. Matched-Normal, Exome Sequencing
- Number of participants in study:

Subjects with Phenotypes: 92

Subjects with Genotypes: 74

Total Number of Subjects: 92

Authorized Access

- **Data access provided by:** dbGaP Authorized Access
- **Data Access Committee (DAC):** ncidac@mail.nih.gov
- **Release Date:** October 07, 2011
- **Embargo Release Date:** October 07, 2011
- Data Use Certification Requirements (DUC)
- **Use Restrictions**
  - Cancer Research Only (Show)
- List of components downloadable from Authorized Access

Publicly Available Data (Public ftp)

Connect to public download site

Molecular Data

Type	Vendor/Platform	Number of Oligos/SNPs	SNP Batch Id	Comment
Whole Genome Genotyping	AFFYMETRIX AFFY 6.0	934940	52074	
Whole Genome Sequencing	ILLUMINA 101bp paired end reads	N/A	N/A	
Whole Exome Sequencing	ILLUMINA Agilent selected, 76bp paired end reads	N/A	N/A	

Figure 1: The ISA framework in action in the stem cell–based system of the Harvard Stem Cell Institute (HSCI).

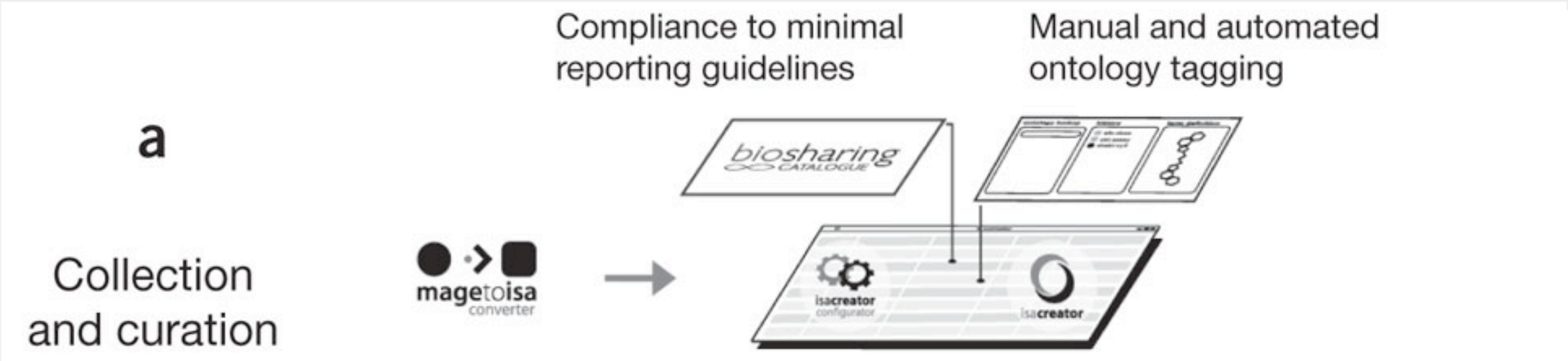
From

Toward interoperable bioscience data

Susanna-Assunta Sansone, **Philippe Rocca-Serra**, **Dawn Field**, **Eamonn Maguire**, **Chris Taylor**, **Oliver Hofmann**, **Hong Fang**, **Steffen Neumann**, **Weida Tong**, **Linda Amaral-Zettler**, **Kimberly Begley**, **Tim Booth**, **Lydie Bougueleret**, **Gully Burns**, **Brad Chapman**, **Tim Clark**, **Lee-Ann Coleman**, **Jay Copeland**, **Sudeshna Das**, **Antoine de Daruvar**, **Paula de Matos**, **Ian Dix**, **Scott Edmunds**, **Chris T Evelo**, **Mark J Forster** + ***et al.***

Nature Genetics **44**, 121–126 (2012) | doi:10.1038/ng.1054

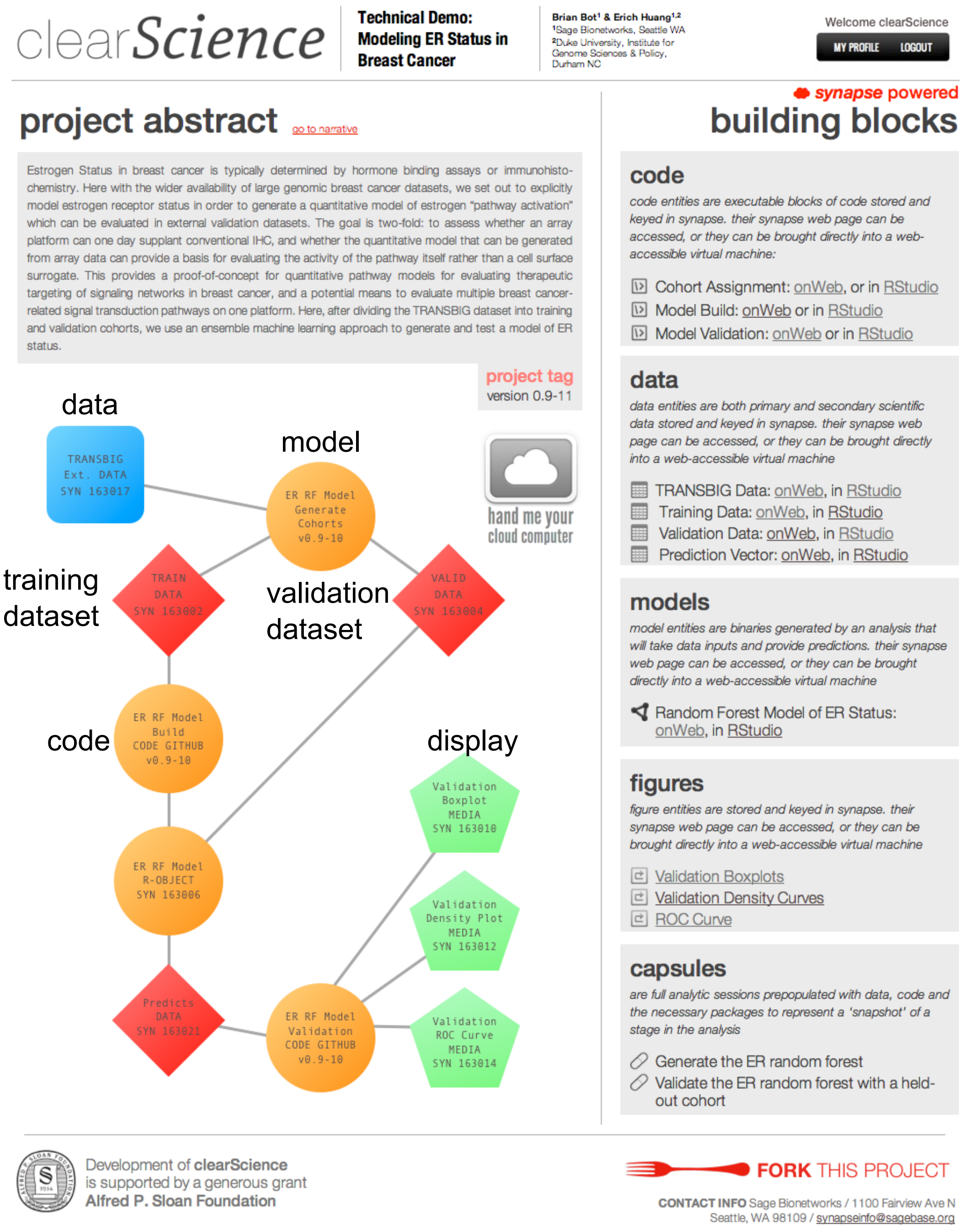
Figure 1: The ISA framework in action in the stem cell–based system of the Harvard Stem Cell Institute (HSCI).



Configurable spreadsheet-like editing environment

nature publishing group npg

Journal and database working together on peer review of transparent data



clearScience

Technical Demo: Modeling ER Status in Breast Cancer

Brian Bot[1] & Erich Huang[1,2]
[1]Sage Bionetworks, Seattle WA
[2]Duke University, Institute for Genome Sciences & Policy, Durham NC

project abstract

go to narrative

Estrogen Status in breast cancer is typically determined by hormone binding assays or immunohistochemistry. Here with the wider availability of large genomic breast cancer datasets, we set out to explicitly model estrogen receptor status in order to generate a quantitative model of estrogen "pathway activation" which can be evaluated in external validation datasets. The goal is two-fold: to assess whether an array platform can one day supplant conventional IHC, and whether the quantitative model that can be generated from array data can provide a basis for evaluating the activity of the pathway itself rather than a cell surface surrogate. This provides a proof-of-concept for quantitative pathway models for evaluating therapeutic targeting of signaling networks in breast cancer, and a potential means to evaluate multiple breast cancer-related signal transduction pathways on one platform. Here, after dividing the TRANSBIG dataset into training and validation cohorts, we use an ensemble machine learning approach to generate and test a model of ER status.

project tag
version 0.9-11

data
TRANSBIG Ext. DATA SYN 163017
model
ER RF Model Generate Cohorts v0.9-10
training dataset
TRAIN DATA SYN 163002
validation dataset
VALID DATA SYN 163004
code
ER RF Model Build CODE GITHUB v0.9-10
ER RF Model R-OBJECT SYN 163006
Predicts DATA SYN 163021
ER RF Model Validation CODE GITHUB v0.9-10
display
Validation Boxplot MEDIA SYN 163010
Validation Density Plot MEDIA SYN 163012
Validation ROC Curve MEDIA SYN 163014
hand me your cloud computer

synapse powered

building blocks

code

code entities are executable blocks of code stored and keyed in synapse. their synapse web page can be accessed, or they can be brought directly into a web-accessible virtual machine:

- Cohort Assignment: onWeb, or in RStudio
- Model Build: onWeb or in RStudio
- Model Validation: onWeb or in RStudio

data

data entities are both primary and secondary scientific data stored and keyed in synapse. their synapse web page can be accessed, or they can be brought directly into a web-accessible virtual machine

- TRANSBIG Data: onWeb, in RStudio
- Training Data: onWeb, in RStudio
- Validation Data: onWeb, in RStudio
- Prediction Vector: onWeb, in RStudio

models

model entities are binaries generated by an analysis that will take data inputs and provide predictions. their synapse web page can be accessed, or they can be brought directly into a web-accessible virtual machine

- Random Forest Model of ER Status: onWeb, in RStudio

figures

figure entities are stored and keyed in synapse. their synapse web page can be accessed, or they can be brought directly into a web-accessible virtual machine

- Validation Boxplots
- Validation Density Curves
- ROC Curve

capsules

are full analytic sessions prepopulated with data, code and the necessary packages to represent a 'snapshot' of a stage in the analysis

- Generate the ER random forest
- Validate the ER random forest with a held-out cohort

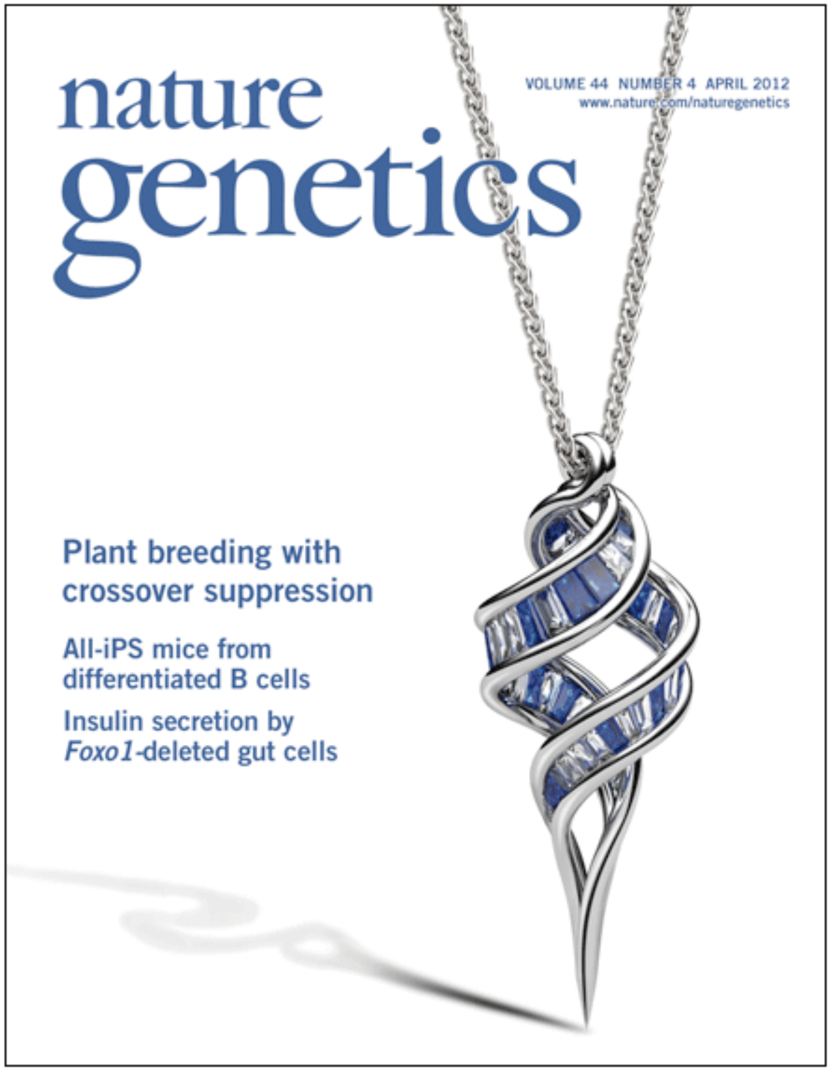
Development of **clearScience** is supported by a generous grant **Alfred P. Sloan Foundation**

FORK THIS PROJECT

CONTACT INFO Sage Bionetworks / 1100 Fairview Ave N
Seattle, WA 98109 / synapseinfo@sagebase.org

Data at GEO
Curated data in cloud
Programs at R
Project pipeline
Referee comments at journal
Open comments on programs, datasets and analysis
Link to DOI of refereed and published paper

Thank you!



nature genetics

VOLUME 44 NUMBER 4 APRIL 2012
www.nature.com/naturegenetics

Plant breeding with crossover suppression

All-iPS mice from differentiated B cells

Insulin secretion by *Foxo1*-deleted gut cells

Alexander Davis http://www.rowandavis.com/

nature publishing group npg