Scientific Data

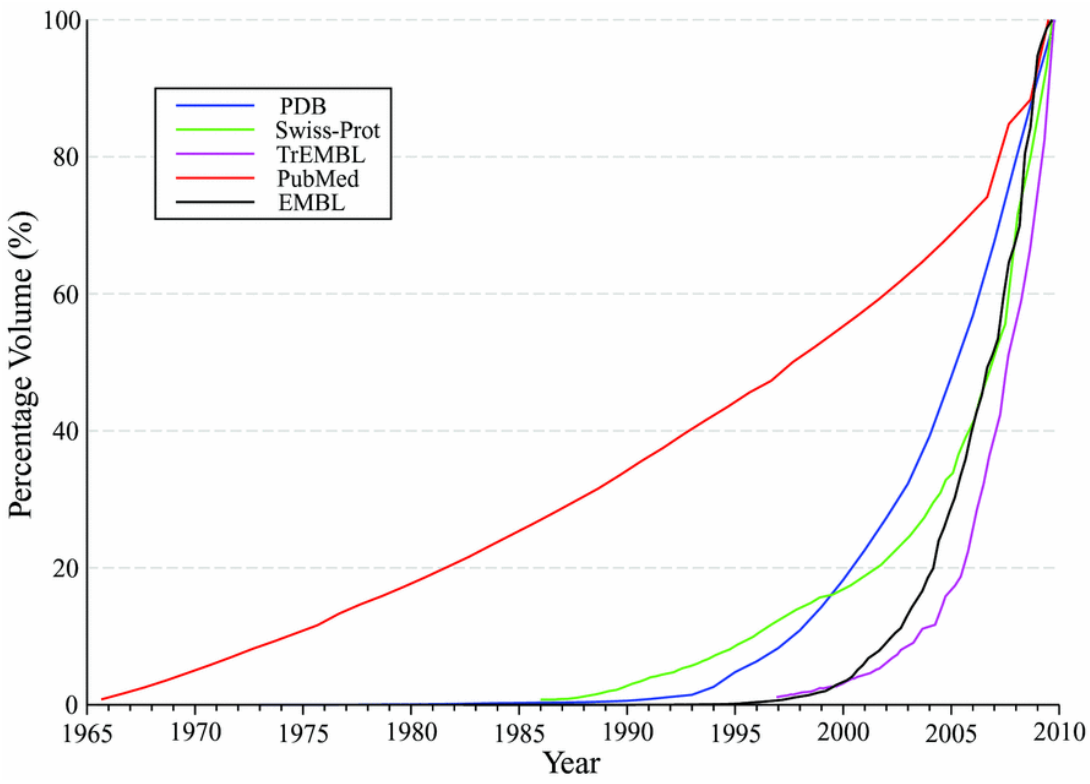
An open-access, peer-reviewed platform for data-focused publications

Andrew Hufton
Managing Editor, Data
Nature Publishing Group

nature publishing group

Data, data, data

Depositions of datasets in archives continue to grow, surpassing journal articles in biomedical research



Growth of biomedical research publications (red; current total >19 million), alongside the accumulation of research data, including nucleic acid sequences (black; current total ~163 million), computer-annotated protein sequences (magenta; current total 9 million), manually annotated protein sequences (green; current total 500,000) and protein structures (blue; current total 60,000)

Source: Biochemical Journal 2009 424, 317-333 - Teresa K. Attwood, Douglas B. Kell and others.

An illustrative example….

No. 4356 April 25, 1953 NATURE 737

equipment, and to Dr. G. E. R. Deacon and the captain and officers of H.R.S. *Discovery II* for their part in making the observations.

[1] Young, F. B., Gerard, H., and Jevons, W., *Phil. Mag.*, (1930).

[2] Longuet-Higgins, M. S., *Mon. Not. Roy. Astro. Soc., Geophys. Supp.*, (1949).

[3] Van Ark, W. S., Woods Hole Papers in Phys. Oceanog. Meteor., (1950).

[4] Ekman, V. W., *Arkiv. Mat. Astron. Fysik.* (Stockholm), (1904).

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

A Structure for Deoxyribose Nucleic Acid

WE wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A.). This structure has novel features which are of considerable biological interest.

A structure for nucleic acid has already been proposed by Pauling and Corey[1]. They kindly made their manuscript available to us in advance of publication. Their model consists of three intertwined chains, with the phosphates near the fibre axis, and the bases on the outside. In our opinion, this structure is unsatisfactory for two reasons: (1) We believe that the material which gives the X-ray diagrams is the salt, not the free acid. Without the acidic hydrogen atoms it is not clear what forces would hold the structure together, especially as the negatively charged phosphates near the axis will repel each other. (2) Some of the van der Waals distances appear to be too small.

Another three-chain structure has also been suggested by Fraser (in the press). In his model the phosphates are on the outside and the bases on the inside, linked together by hydrogen bonds. This structure as described is rather ill-defined, and for this reason we shall not comment on it.

We wish to put forward a radically different structure for the salt of deoxyribose nucleic acid. This structure has two helical chains each coiled round the same axis (see diagram). We have made the usual chemical assumptions, namely, that each chain consists of phosphate diester groups joining β-D-deoxyribofuranose residues with 3′,5′ linkages. The two chains (but not their bases) are related by a dyad perpendicular to the fibre axis. Both chains follow right-handed helices, but owing to the dyad the sequences of the atoms in the two chains run in opposite directions. Each chain loosely resembles Furberg's[2] model No. 1; that is, the bases are on the inside of the helix and the phosphates on the outside. The configuration of the sugar and the atoms near it is close to Furberg's 'standard configuration', the sugar being roughly perpendicular to the attached base. There is a residue on each chain every 3·4 A. in the z-direction. We have assumed an angle of 36° between adjacent residues in the same chain, so that the structure repeats after 10 residues on each chain, that is, after 34 A. The distance of a phosphorus atom from the fibre axis is 10 A. As the phosphates are on the outside, cations have easy access to them.

The structure is an open one, and its water content is rather high. At lower water contents we would expect the bases to tilt so that the structure could become more compact.

The novel feature of the structure is the manner in which the two chains are held together by the purine and pyrimidine bases. The planes of the bases are perpendicular to the fibre axis. They are joined together in pairs, a single base from one chain being hydrogen-bonded to a single base from the other chain, so that the two lie side by side with identical z-co-ordinates. One of the pair must be a purine and the other a pyrimidine for bonding to occur. The hydrogen bonds are made as follows: purine position 1 to pyrimidine position 1; purine position 6 to pyrimidine position 6.

If it is assumed that the bases only occur in the structure in the most plausible tautomeric forms (that is, with the keto rather than the enol configurations) it is found that only specific pairs of bases can bond together. These pairs are: adenine (purine) with thymine (pyrimidine), and guanine (purine) with cytosine (pyrimidine).

In other words, if an adenine forms one member of a pair, on either chain, then on these assumptions the other member must be thymine; similarly for guanine and cytosine. The sequence of bases on a single chain does not appear to be restricted in any way. However, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined.

It has been found experimentally[4] that the ratio of the amounts of adenine to thymine, and the ratio of guanine to cytosine, are always very close to unity for deoxyribose nucleic acid.

It is probably impossible to build this structure with a ribose sugar in place of the deoxyribose, as the extra oxygen atom would make too close a van der Waals contact.

The previously published X-ray data[5,6] on deoxyribose nucleic acid are insufficient for a rigorous test of our structure. So far as we can tell, it is roughly compatible with the experimental data, but it must be regarded as unproved until it has been checked against more exact results. Some of these are given in the following communications. We were not aware of the details of the results presented there when we devised our structure, which rests mainly though not entirely on published experimental data and stereochemical arguments.

It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.

Full details of the structure, including the conditions assumed in building it, together with a set of co-ordinates for the atoms, will be published elsewhere.

We are much indebted to Dr. Jerry Donohue for constant advice and criticism, especially on interatomic distances. We have also been stimulated by a knowledge of the general nature of the unpublished experimental results and ideas of Dr. M. H. F. Wilkins, Dr. R. E. Franklin and their co-workers at

This figure is purely diagrammatic. The two ribbons symbolize the sugar-phosphate chains, and the horizontal rods the pairs of bases holding the chains together. The vertical line marks the fibre axis

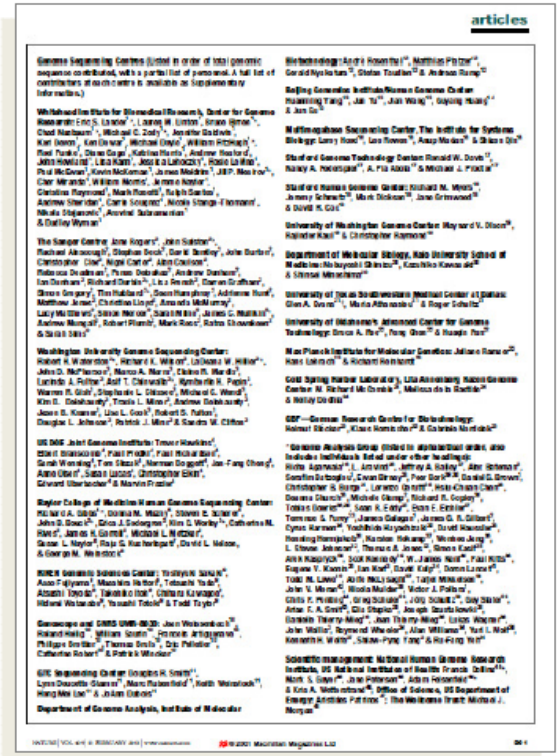
DNA Structure 1953

1 Page
2 Authors
1 Figure
no data

An illustrative example….

Human Genome 2001

62 Pages, 150 Authors,
49 Figure, 27 tables


articles

Genome Sequencing Centres (Listed in order of total genomic sequence contributed, with a partial list of personnel. A full list of contributors at each centre is available as Supplementary Information.)

Whitehead Institute for Biomedical Research, Center for Genome Research: Eric S. Lander*, Lauren M. Linton, Bruce Birren*, Chad Nusbaum*, Michael C. Zody*, Jennifer Baldwin, Keri Devon*, Ken Dewar*, Michael Doyle*, William FitzHugh*, Rod Funke*, Diana Gage*, Katrina Harris*, Andrew Heaford*, John Howland*, L. Ioan Kann*, Jessica Lehoczky*, Rosie LeVine*, Paul McEwan*, Kevin McKernan*, James Meldrim*, Jill P. Mesirov*, Cher Miranda*, William Morris*, Jerome Naylor*, Christina Raymond*, Mark Rosetti*, Ralph Santos*, Andrew Sheridan*, Carrie Sougnez*, Nicole Stange-Thomann*, Nikola Stojanovic*, Aravind Subramanian*, Dudley Wyman*

The Sanger Centre: Jane Rogers*, Jan Sulston*, Rachael Ainscough*, Stephen Beck*, David Bentley*, John Burton*, Christopher Clee*, Nigel Carter*, Alan Coulson*, Rebecca Deadman*, Piroo Deloukas*, Andrew Dunham*, Ian Dunham*, Richard Durbin*, Lisa French*, Darren Grafham*, Simon Gregory*, Tim Hubbard*, Sean Humphray*, Adrienne Hunt*, Matthew Jones*, Christine Lloyd*, Amanda McMurray*, Lucy Matthews*, Simon Mercer*, Sarah Milne*, James C. Mullikin*, Andrew Mungall*, Robert Plumb*, Mark Ross*, Ratna Shownkeen* & Sarah Sims*

Washington University Genome Sequencing Center: Robert H. Waterston*, Richard K. Wilson*, LaDeana W. Hillier*, John D. McPherson*, Marco A. Marra*, Elaine R. Mardis*, Lucinda A. Fulton*, Asif T. Chinwalla*, Kymberlie H. Pepin*, Warren R. Gish*, Stephanie L. Chissoe*, Michael C. Wendl*, Kim D. Delehaunty*, Tracie L. Miner*, Andrew Delehaunty*, Jason B. Kramer*, Lisa L. Cook*, Robert S. Fulton*, Douglas L. Johnson*, Patrick J. Minx* & Sandra W. Clifton*

US DOE Joint Genome Institute: Trevor Hawkins*, Elbert Branscomb*, Paul Predki*, Paul Richardson*, Sarah Wenning*, Tom Slezak*, Norman Doggett*, Jan-Fang Cheng*, Anna Olsen*, Susan Lucas*, Christopher Elkin*, Edward Uberbacher* & Marvin Frazier*

Baylor College of Medicine Human Genome Sequencing Center: Richard A. Gibbs*, Donna M. Muzny*, Steven E. Scherer*, John B. Bouck*, Erica J. Sodergren*, Kim C. Worley*, Catherine M. Rives*, James H. Gorrell*, Michael L. Metzker*, Susan L. Naylor*, Raju S. Kucherlapati*, David L. Nelson*, & George M. Weinstock*

RIKEN Genomic Sciences Center: Yoshiyuki Sakaki*, Asao Fujiyama*, Masahira Hattori*, Tetsushi Yada*, Atsushi Toyoda*, Takehiko Itoh*, Chiharu Kawagoe*, Hidemi Watanabe*, Yasushi Totoki* & Todd Taylor*

Genoscope and CNRS UMR-8030: Jean Weissenbach*, Roland Heilig*, William Saurin*, Francis Artiguenave*, Philippe Brottier*, Thomas Bruls*, Eric Pelletier*, Catherine Robert*, & Patrick Wincker*

GTC Sequencing Center: Douglas R. Smith*, Lynn Doucette-Stamm*, Mario Rubenfield*, Keith Weinstock*, Hong Mei Lee* & JoAnn Dubois*

Department of Genome Analysis, Institute of Molecular Biotechnology: Andre Rosenthal*, Matthias Platzer*, Gerald Nyakatura*, Stefan Taudien* & Andreas Rump*

Beijing Genomics Institute/Human Genome Center: Huanming Yang*, Jian Yu*, Jun Wang*, Guyang Huang* & Jun Gu*

McDonnell Genome Sequencing Center, The Institute for Systems Biology: Leroy Hood*, Lee Rowen*, Anup Madan* & Shizen Qin*

Stanford Genome Technology Center: Ronald W. Davis*, Nancy A. Hillier*, A. Fritz Abola*, & Michael J. Proctor*

Stanford Human Genome Center: Richard M. Myers*, Jeremy Schmutz*, Mark Dickson*, Jane Grimwood* & David R. Cox*

University of Washington Genome Center: Maynard V. Olson*, Kajuna Kaul* & Christopher Raymond*

Department of Molecular Biology, Keio University School of Medicine: Nobuyoshi Shimizu*, Kazuhiko Kawasaki* & Shinsei Minoshima*

University of Texas Southwestern Medical Center at Dallas: Glen A. Evans*, Maria Athanasiou* & Roger Schultz*

University of Wisconsin Affymetrix Center for Genome Technology: Simon A. Fry*, Peng Gao* & Xuanjie Fan*

Max Planck Institute for Molecular Genetics: Juliane Rosner*, Hans Lehrach* & Richard Reinhardt*

Cold Spring Harbor Laboratory, Lita Annenberg Hazen Genome Center: W. Richard McCombie*, Melissa de la Bastide* & Nancy Dredge*

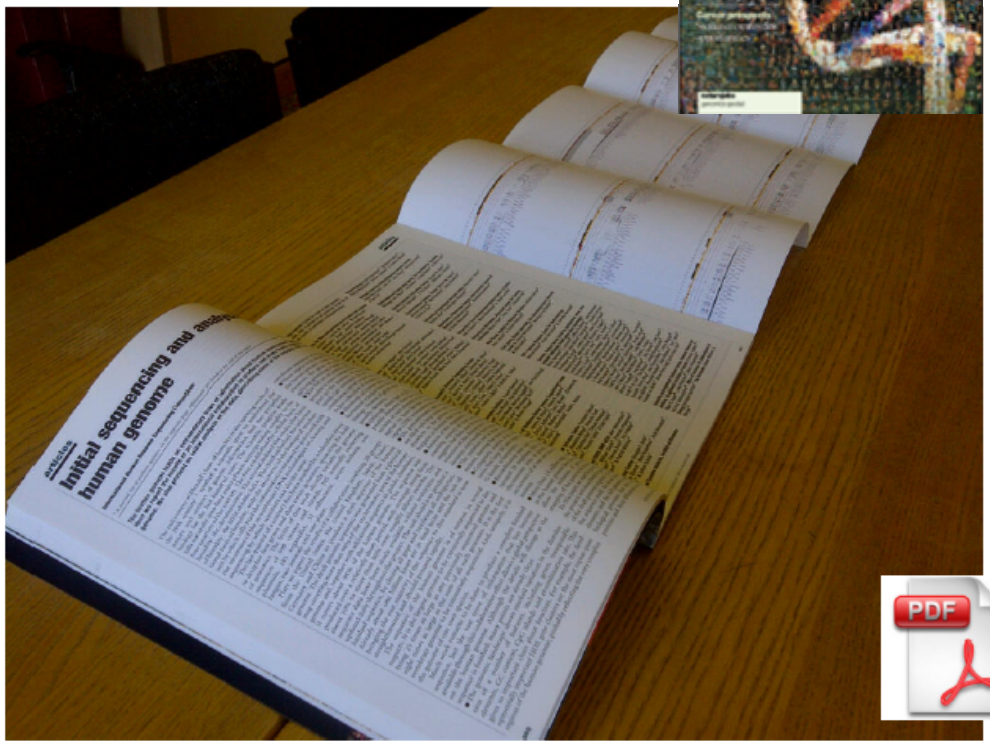
DKFZ—German Research Centre for Biotechnology: Helmut Blöcker*, Klaus Homuchka* & Gabriele Monfeldt*

Genome Analysis Group (listed in alphabetical order, also includes individuals listed under other headings): Richa Agarwala†, L. Aravind*, Jeffrey A. Bailey*§, Alex Bateman*, Serafim Batzoglou*, Ewan Birney*, Peer Bork*¶, Daniel G. Brown*, Christopher B. Burge*, Lorenzo Cerutti*, Hsia-Chun Chen*, Deanna Church*#, Michele Clamp*, Richard R. Copley*, Tobias Doerks*‖, Scott K. Eddy*, Eran E. Eichler*, Terrence S. Furey*‖, James Galagan*, James G. R. Gilbert*, Cyrus Harmon*, Yoshihide Hayashizaki*¶, David Haussler*, Henning Hermjakob*, Karsten Hokamp*, Wonhee Jang*, L. Steven Johnson*, Theresa A. Jones*, Simon Kasif*, Arik Kaspryzk*, Scott Kennedy*, W. James Kent*, Paul Kitts*, Eugene V. Koonin*, Jan Korf*, David Kulp*, Doron Lancet*, Todd M. Lowe*, Ariel McLysaght*, Tarjei Mikkelsen*, John V. Moran*, Nicola Mulder*, Victor J. Pollara*, Chris P. Ponting*, Greg Schuler*, Jörg Schultz*, Guy Slater*, Arian F. A. Smit*, Elia Stupka*, Joseph Szustakowski*, Danielle Thierry-Mieg*, Jean Thierry-Mieg*, Lukas Wagner*, John Wallis*, Raymond Wheeler*, Alan Williams*, Yuri I. Wolf*, Kenneth H. Wolfe*, Shiaw-Pyng Yang*** & Ru-Fang Yeh*

Scientific management: National Human Genome Research Institute, US National Institutes of Health: Francis Collins*‖‖, Mark S. Guyer*‖, Jane Peterson*‖, Adam Felsenfeld*‖; & Kris A. Wetterstrand*‖; Office of Science, US Department of Energy: Anthony Palmisano*‖; The Wellcome Trust: Michael J. Morgan*

articles

initial sequencing and analysis of the human genome



Existing challenges

- Data producers do not necessarily get appropriate credit for their work
- The peer review process at research journals is not ideal for ensuring data release and data standards
- Data and info about datasets often ends in supp. material
- Valuable datasets are not released

A Role for Data Publications

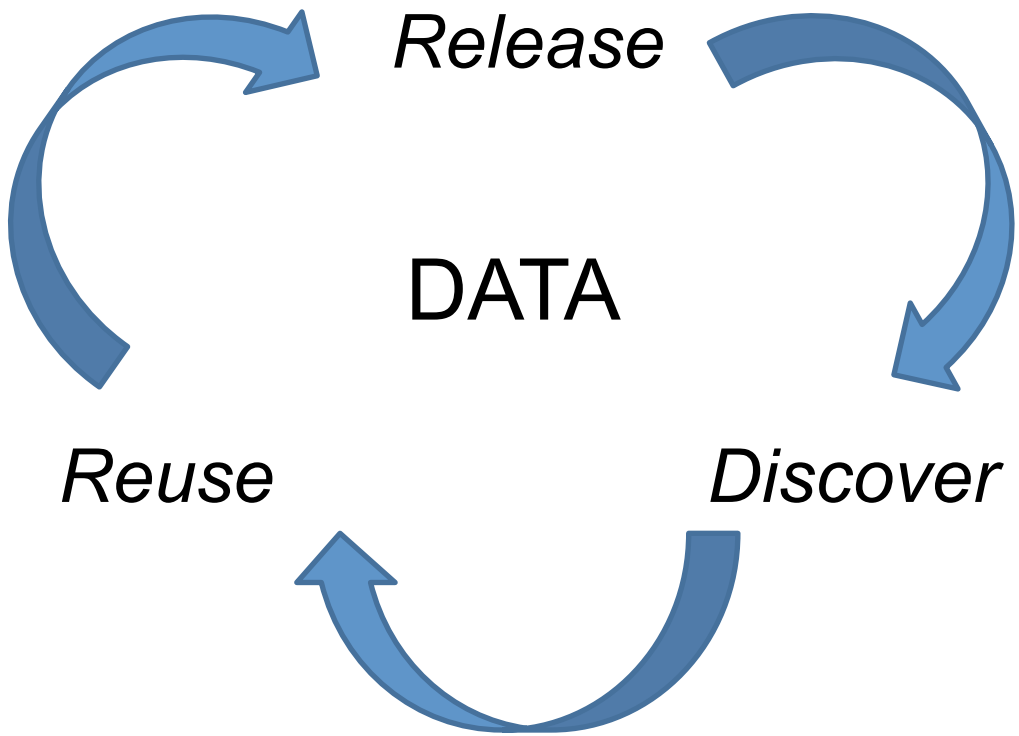
GigaScience
F1000R
ZooKeys
Ecological Archives

and soon *Scientific Data…*

Why more data?

- To ensure that the scientific process is transparent and reproducible

- To accelerate new discoveries, and reduce unnecessarily repeated work


Release
DATA
Reuse
Discover

If we want scientists to release their data, we need to provide a credit mechanism

If we want scientists to release their data, we need to provide a credit mechanism

PUBLICATION

If want released data to be reusable, we need critical evaluation to verify experimental rigor and the completeness of the description

If want released data to be reusable, we need critical evaluation to verify experimental rigor and the completeness of the description

PEER REVIEW

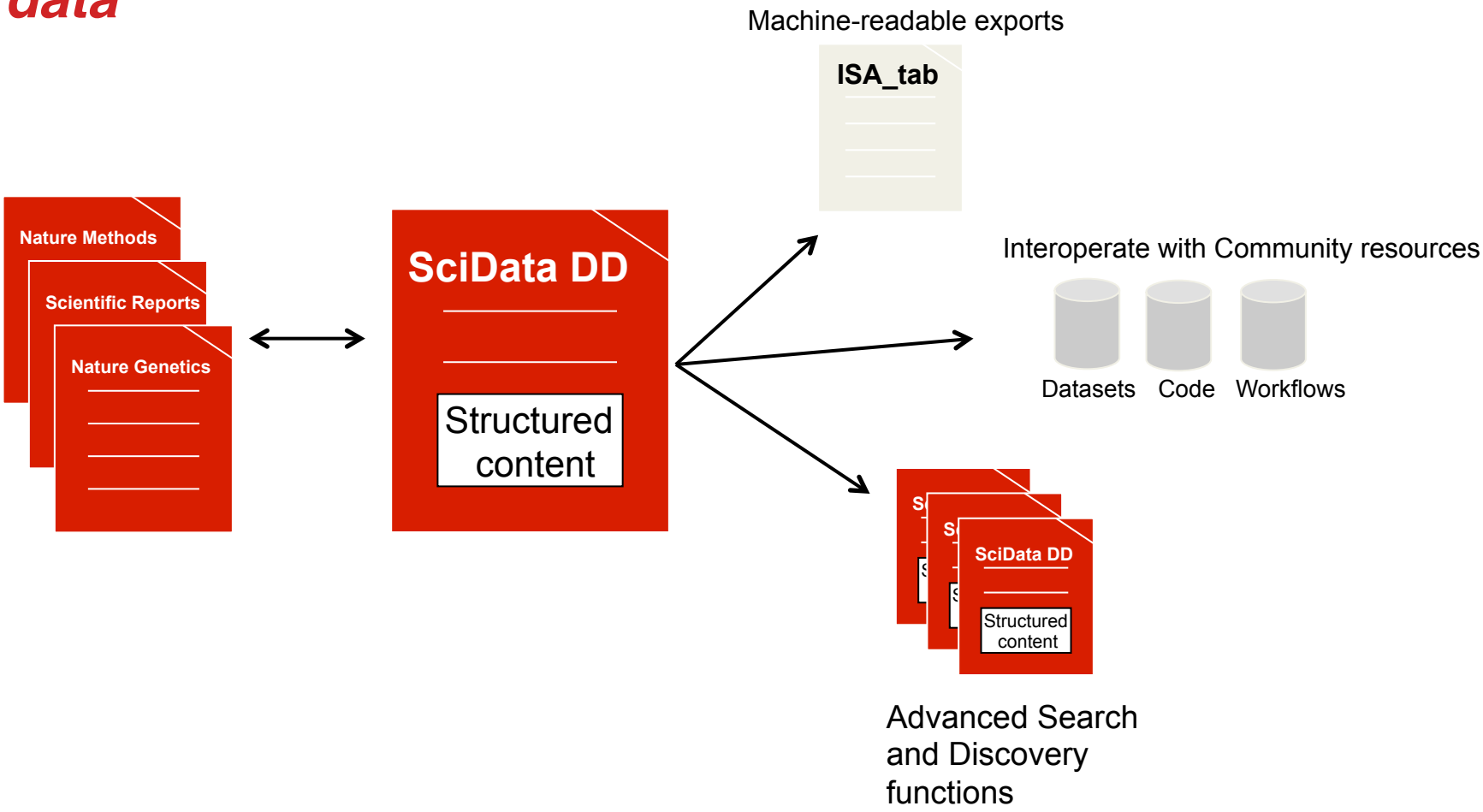
If we want scientists to be able to find datasets that will accelerate their research, then datasets need to be searchable & discoverable

If we want scientists to be able to find datasets that will accelerate their research, then datasets need to be searchable & discoverable

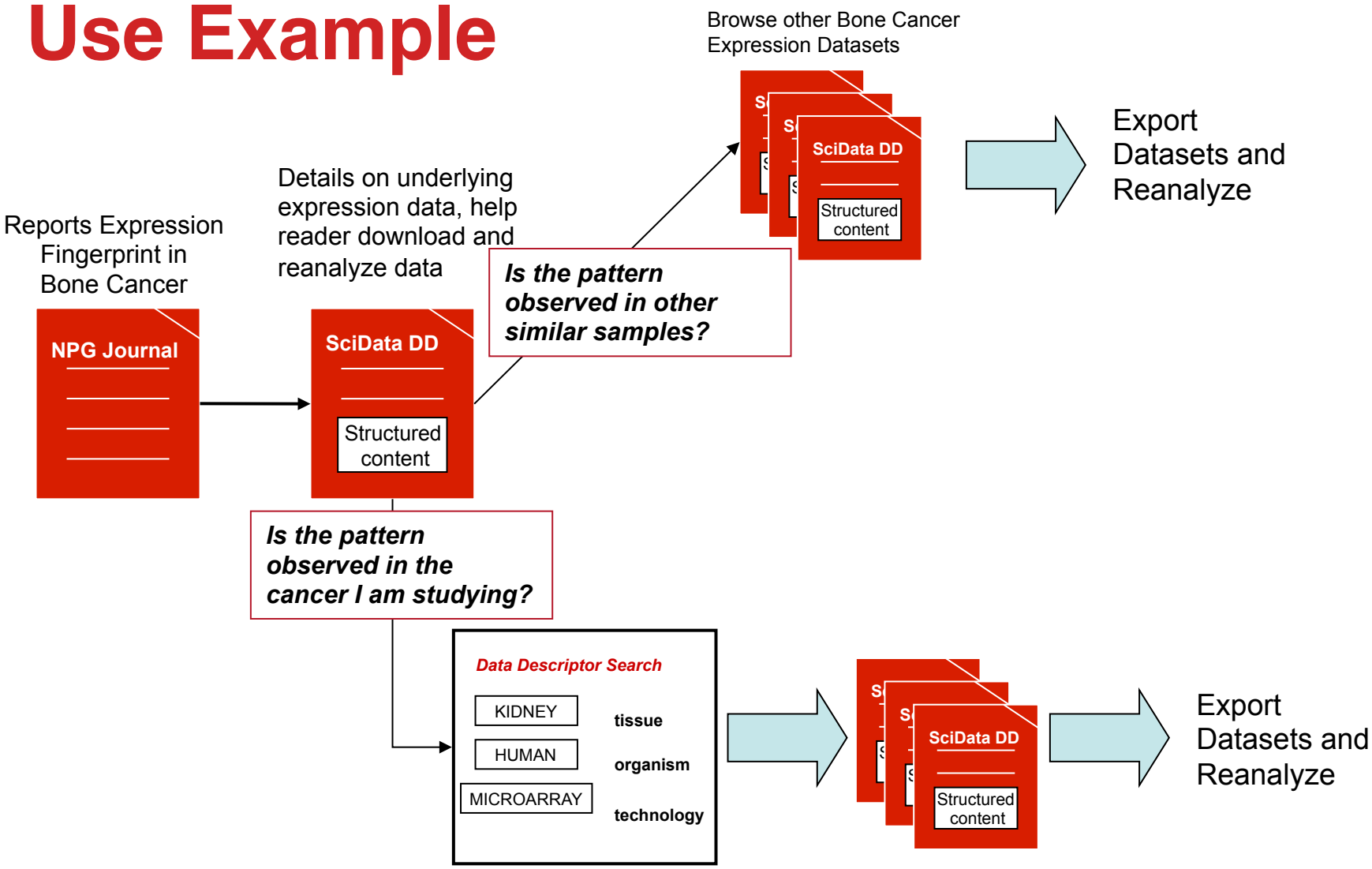
STANDARDIZED,
CURATED DESCRIPTIONS

The Data Publication

seamless integration between research findings and data



Use Example


Browse other Bone Cancer Expression Datasets
SciData DD
Structured content
Export Datasets and Reanalyze
Details on underlying expression data, help reader download and reanalyze data
Reports Expression Fingerprint in Bone Cancer
NPG Journal
SciData DD
Structured content
Is the pattern observed in other similar samples?
Is the pattern observed in the cancer I am studying?
Data Descriptor Search
KIDNEY
tissue
HUMAN
organism
MICROARRAY
technology
SciData DD
Structured content
Export Datasets and Reanalyze

Scientific Data Concept Overview

Credit: citeable, peer-reviewed mechanism for recording credit for dataset creation.

Rapid peer-review and publication evaluate the technical quality of the procedures used to generate the data, the reuse value of the resulting datasets, the completeness of the data description, and alignment of with existing community standards.

Reuse: complete and standardized descriptions help others discover and reuse your data, and discover your related research articles.

Diverse data-types: *Scientific Data* will initially focus on experimental and observational datasets from the life, biomedical and environmental science communities, but will be open to content from a wide range of experimental disciplines.

Cooperation with a broad set of public, community-recognized repositories will promote existing standards, and **integration** with generalist scientific repositories like Dryad and FigShare will make it easy to deposit diverse data-types.

Open access: Data Descriptors will be released freely to the public under an open access license.

Complementary: Data Descriptors can be used to describe both datasets that are associated with a traditional research article, as well as standalone datasets.

Focused Scope

complements both journal articles and repository records

Includes

- Highly detailed, reproducible methods descriptions
- Quality control & technical validation experiments
- Searchable, machine-readable meta-data

Does Not Include

- In depth analysis or tests of hypotheses
- New scientific conclusions
- Exploratory analysis (e.g. clustering)

Thanks!

Andrew Hufton andrew.hufton@nature.com
Managing Editor, Data