Creative Commons CC0 for data and (some of) BioMed Central's other initiatives in open data

July 2012

Iain Hrynaszkiewicz
Publisher (Open Science), BioMed Central
iain.hrynaszkiewicz@biomedcentral.com

BioMed Central
The Open Access Publisher

BioMed Central open data initiatives

1. Data journals and article types
2. Open Data Award
3. Data deposition, citation, and linking
4. Data workflow integration (LabArchives partnership)
5. Data licensing
6. Human subjects – confidentiality and consent
7. Guidance and best practice
8. Data formats and standards

BioMed Central open data initiatives

1. **Data journals and article types**
2. Open Data Award
3. **Data deposition, citation, and linking**
4. Data workflow integration (LabArchives partnership)
5. **Data licensing**
6. Human subjects – confidentiality and consent
7. Guidance and best practice
8. Data formats and standards

Data notes (papers)

BMC Research Notes

(GIGA)n SCIENCE

Open Network Biology

TRIALS

Data notes: "[B]riefly describe a biomedical data set or database, with the data being readily accessible and attributed to a source" http://bit.ly/y3Jb3b

Data notes: "[E]xceptional datasets deposited in our *GigaScience* repository that have been selected for further peer-review" http://bit.ly/yPBsAA

Models: "[N]etwork-based models of living systems linked to the corresponding coherent datasets upon which the models are based." http://bit.ly/otktKZ

Research: E.g. The International Stroke Trial database http://www.trialsjournal.com/content/12/1/101

BioMed Central
The Open Access Publisher

Data journal with integrated repository – launching July 2012

http://www.gigasciencejournal.com

华大基因
BGI

DataCite

"GigaScience aims to revolutionize data dissemination, organization, understanding, and use. An online open-access open-data journal, we publish 'big-data' studies from the entire spectrum of life and biomedical sciences. To achieve our goals, the journal has a novel publication format: one that links standard manuscript publication with an extensive database that hosts all associated data and provides data analysis tools and cloud-computing resources."

BioMed Central
The Open Access Publisher

http://gigadb.org/

Mouse methylomes

Here we present 18 genome-wide DNA methylation profiles of wild type and Thymine DNA glycosylase (*Tdg*) knockout cells, which serve as an excellent murine methylome resource. The 18 samples represent 6 biological cohorts: 6 samples were derived from mouse embryonic stem cells (3 *Tdg+/-*, 3 *Tdg-/-*), 6 samples were from mouse neural precursor cells (3 *Tdg+/-*, 3 *Tdg-/-*) and 6 samples were obtained from mouse embryonic fibroblasts (3 *Tdg+/+*, 3 *Tdg-/-*).

Next generation sequencing was performed on the libraries using an Illumina GAIIx for each sample. Paired end alignment against the mouse genome (Build NCBIM37) was performed using BWA (v0.5.8), and filtering to remove those reads failing to map was performed using SAMtools (v0.1.9) and a custom perl script. The Bioconductor (v2.7) package MeDIPs (v1.0.0) was used to normalize for size of the sequence library, done by calculating reads per million in tiled windows across the genome. Fragment length normalization was performed using a custom perl script. Wig tracks representing library size normalized alignment were generated using a combination of MEDIPS and custom R scripts. In addition to the total alignment wig track, strand specific wig tracks were also generated, enabling the user to infer whether the MeDIP signal is derived by methylation on the forward and/or reverse strand.

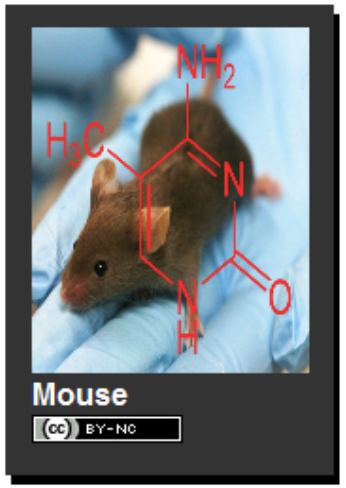
The MeDIP-seq data were processed using the analysis pipeline MeDUSA (Methylated DNA Utility for Sequence Analysis). MeDUSA brings together numerous software packages to perform a full analysis of MeDIP-seq data, including sequence alignment, quality control, and determination and annotation of differentially methylated regions.

For more information see:
The Ensembl HEROIC portal, for wig tracks displaying normalized read depth:
http://projects.ensembl.org/heroic/ or http://www2.cancer.ucl.ac.uk/medicalgenomics/tdg_web/trackList.php

Analysis tools also available at http://www2.cancer.ucl.ac.uk/medicalgenomics/medusa/

download

readme
readme.txt



Citation

In accordance with our terms of use, please cite this dataset as:
Wilson, G; Dharmi, P; Saito, Y; Cortázar, D; Kunz, C; Schär, P; Beck, S (2012): Resources for the MeDUSA (Methylated DNA Utility for Sequence Analysis) MeDIP-seq computational analysis pipeline for the identification of differentially methylated regions, and associated methylome data from 18 wild-type and mutant mouse ES, NP and MEF cells. GigaScience. http://dx.doi.org/10.5524/100035

Accession codes associated with this data:
NCBI BioProject PRJNA138437
NCBI Study SRP005934
NCBI GEO GSE27468

Data **deposition**, citation, and linking

DataCite

Helping you to find,
access, and reuse data

http://datacite.org/repolist

Repositories

This list is a working document, initated via a collaboration between the British Library, BioMed Central and the Digital Curation Centre, that aims to capture the growing number of repositories for research data. It is provided for information purposes only: DataCite provides no endorsements as to the quality or suitability of the repositories listed. We encourage community participation in developing this resource. Please contact us to suggest changes or additions.

Repository	Website	Subject area(s)	Funding model	Deposit restrictions	Access restrictions	Li... ag...
Domain-specific and general data repositories (multiple formats accepted)						
		Technical sciences, climate,			Data is only displayed if permission has been granted by the owner who has	N

http://www.biomedcentral.com/about/supportingdata

But where can I put my data if not in additional files?
Additional files (supplementary materials) are a viable option for certain types of small-scale data files and BioMed Central authors are encouraged to include data as additional files where possible. But for large datasets and to meet data availability policies of some communities and institutions, data need to be hosted in a repository. We are therefore keen to provide our authors with as much information as possible on where they can deposit their data, so it can potentially be linked to their publication(s). Well established and widely supported databases exist for certain types of data such as nucleic acid sequences, protein sequences, and atomic coordinates, and these are already included in relevant journals' instructions for authors. But there are many other repositories, which could potentially link data to our publications. Therefore, to complement the development of this new article section, we have been collaborating with DataCite, the British Library and the Digital Curation Centre to develop and maintain a list of domain and institution-specific repositories – repositories which accept a variety of data file types and assign a variety of unique, permanent identifiers for deposited data. This list is available on the DataCite website and is linked from the instructions for authors of the relevant journals. Community participation in developing this resource is strongly encouraged. Please contact DataCite to suggest changes and additions to the repository list.

BioMed Central
The Open Access Publisher

Data deposition, **citation**, and linking

References

All references, including URLs, must be numbered consecutively, in square brackets, in the order in which they are cited in the text, followed by any in tables or legends. Each reference must have an individual reference number. Please avoid excessive referencing. If automatic numbering systems are used, the reference numbers must be finalized and the bibliography must be fully formatted before submission.

Only articles, datasets and abstracts that have been published or are in press, or are available through public e-print/preprint servers, may be cited; unpublished abstracts, unpublished data and personal communications should not be included in the reference list, but may be included in the text and referred to as "unpublished observations" or "personal communications" giving the names of the involved researchers. Obtaining permission to quote personal communications and unpublished data from the cited colleagues is the responsibility of the author. Footnotes are not allowed, but endnotes are permitted. Journal abbreviations follow Index Medicus/MEDLINE. Citations in the reference list should include all named authors, up to the first 30 before adding '*et al.*'.

Examples of the *BMC Research Notes* reference style

Dataset with persistent identifier

Zheng, L-Y; Guo, X-S; He, B; Sun, L-J; Peng, Y; Dong, S-S; Liu, T-F; Jiang, S; Ramachandran, S; Liu, C-M; Jing, H-C (2011): Genome data from sweet and grain sorghum (Sorghum bicolor). *GigaScience*. http://dx.doi.org/10.5524/100012.

http://www.biomedcentral.com/bmcresnotes/authors/instructions/researcharticle#formatting-references

BioMed Central
The Open Access Publisher

Data deposition, citation, and **linking**

- 'Availability of supporting data' section
- Data must be permanently available – DOI, handle or URL*
- Optional at BMC – journals can require, encourage or omit the section
- Journals include *GigaScience*, *BMC Research Notes*, *Retrovirology*

http://www.biomedcentral.com/about/supportingdata

The following format for the 'Availability of supporting data' section is required when data are available in an open access repository:

"The data set(s) supporting the results of this article is(are) available in the [repository name] repository, [unique persistent identifier/link for dataset(s)]."

The following format is required when data are included as additional files:

"The data set(s) supporting the results of this article is(are) included within the article (and its additional file(s))"

We also recommend that the data set(s) be cited, where appropriate in the manuscript, and included in the reference list.

Journals requiring or encouraging the inclusion of the 'Availability of supporting data' section

A list of BioMed Central journals that encourage or require authors to include this section can be found in the table below.

Journals that include 'Availability of supporting data' section in their research articles

Journal	*Required or encouraged?*	*Specific repository required?*	*Applies to articles submitted from*
Annals of Clinical Microbiology and Antimicrobials	Encouraged	n/a	November 2011
BMC Research Notes	Encouraged	n/a	August 2011
Cell & Bioscience	Encouraged	n/a	December 2011
Clinical Epigenetics	Encouraged	n/a	December 2011
Flavour	Encouraged	n/a	January 2012
Frontiers in Zoology	Encouraged	n/a	December 2011
GigaScience	Required	*GigaScience* database (contact the editors)	July 2011
Gut Pathogens	Encouraged	n/a	November 2011
Open Network Biology	Required	*Open Network Biology* repository (contact the editors)	July 2011
Retrovirology	Required	n/a	November 2011
Silence	Required	n/a	December 2011

BioMed Central
The Open Access Publisher

Creative Commons and journal data

- BMC open data statement August 2010 expressed support for public domain dedication of data in journals
- Most content is CC-BY – requires a change to standard license agreement
- Not compliant with the "Panton Principles" OKF definitions

http://blogs.openaccesscentral.com/blogs/bmcblog/resource/opendatastatementdraft.pdf

BioMed Central
The Open Access Publisher

Copyright and data

If data = numerical representation of facts then they are *generally* not copyrightable, but...

- Many levels of data/derived digital data
- Jurisdictional differences (e.g. US vs. Australian law; EU database rights)

= ambiguity about legal status of content

Licenses and waivers for data

- Licenses are for asserting rights; waivers are for giving them up
- Restrictions on data transfer, integration, reuse slow down research
- Attribution stacking problematic for large/combined datasets

Ball A: **How to License Research Data** 2011
http://www.dcc.ac.uk/resources/how-guides/license-research-data

Why Creative Commons CC0?

- **interoperability:** CC0 is human and machine-readable
- **universality:** CC0 is global and universal and widely recognized
- **simplicity:** no need for humans to make, and respond to, individual data requests

Schaeffer P: **Why does Dryad use CC0?** http://blog.datadryad.org/2011/10/05/why-does-dryad-use-cc0/

BioMed Central
The Open Access Publisher

BioMed Central Blog

http://bit.ly/h6PVoO

MONDAY MAR 21, 2011

Dear Scientist, Help us to put open principles into practice

31 tweets retweet

On 2nd September 2010 BioMed Central issued a draft position statement in support of open data, which provides a number of recommendations – and aspirations – for the role of publishers in promoting reproducible research by increasing scientific data sharing, data reuse and open data.

In the months since the BioMed Central statement, a number of high-level policy statements on data sharing have emerged, including the Oxford Data Sharing Statement, a report to the European Commission, and a joint pledge from a consortium of 17 major public health research funding agencies.

These are positive developments but there are many details and practicalities still be to worked out. So rather than 'why share data?', the question now is 'how?'. A number of projects, initiatives and working groups have formed around issues of data sharing and reproducible research, but a need has been identified for shared understanding on three key issues affecting authors, editors, publishers and funders of life science research.

Licenses and waivers
BioMed Central believes that open data should mean that data are freely available on the public internet permitting any user to download, copy, analyse, re-process, pass them to software or use them for any other purpose without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Around the same time BioMed Central's open data statement was issued Heather Piwowar and Peter Murray-Rust of the Open Knowledge Foundation sent enquiries to a number publishers – Nature, PLoS, BioMed Central included – about the openness of the data published in their journals. The enquiry applied to raw data and meta data published as supplementary material (additional files), and to data extractable from article full text, tables and figures. They established that much data are freely available for harvesting, under Creative Commons attribution licenses, but are not yet fully openly-available according to the Panton Principles for Open Data in Science. Many members of the scientific community have signed in support of these principles; it is time to put them into practice (**see proposed goal #1**).

Supplementary (additional) files
Editors and publishers are acutely aware of the limited pool of peer reviewers who are increasingly called upon to help try and ensure the integrity of the published record. The online availability of research data as a supplementary (additional) files has prompted debate about the role of peer review in this non-written material, and indeed the role of journals in publishing this material. (**see proposed goal #2**).

BioMed Central
The Open Access Publisher

BioMed Central Blog

http://bit.ly/n4U348

MONDAY AUG 08, 2011

Report from the Publishing Open Data Working Group meeting, 17th June 2011

21 tweets retweet

On 17th June BioMed Central held a Publishing Open Data Working Group meeting, proposed in the spring, in London, UK. This post is a summary report from the meeting, including the next steps for the stakeholders involved. The meeting has also been reported by Alex Ball on the Digital Curation Centre blog. Many thanks to all the attendees acknowledged below for their contributions. While an important reason for convening the meeting was to stimulate debate amongst authors, editors, publishers, funders and librarians, it's excellent to report that there are a number of mutually agreeable ways forward on all three of the meeting's proposed goals. Please note that the actions and views stated do not necessarily represent the views of all attendees.

Goal 1: Establish a process and policy for implementing a variable publishers'/authors' license agreement, allowing public domain dedication of data and data elements of scientific articles

A common misconception about implementing Creative Commons CC0 for data published within or alongside scientific articles seems to have been that it applies to all scientists' data, not just those submitted to a journal. This goal pertains only to content which researchers already publish. By implementing a variable license agreement (with CC0 for data and a Creative Commons Attribution license for creative and written works), we would be asking authors to only apply different terms of use to some parts of what they already publish. Journal and publisher policies for the availability of all (that is, including unpublished) underlying data are important, but are a distinct issue, discussed as part of Goal 3.

The International Stroke Trial database, published by Sandercock *et al*. in *Trials* in April 2011, for example, includes a brief article describing a large clinical dataset, and the dataset is an accompanying CSV file. With a variable license agreement, the data (the CSV file) would be available for reuse without a legal requirement for attribution. Scientific norms of citation would still apply and, for any future aggregated use in, for example, a systematic review and meta-analysis, it would undoubtedly be scientifically (culturally) essential for the source data to be cited – even if not legally required – to ensure credibility.

The consensus of the group was that a feasible approach to the implementation of a variable license agreement would be to specify that, from a specific date, any author submitting to a journal/publisher agrees to dedicate the data elements of their article and supplementary material to the public domain. This was a key proposal in the BioMed Central Open Data statement, but the group felt that much more detail on the process, policy and implications of variable licensing for published articles and

Implementation in journal publishing

- Removes ambiguity about legal status of content published under CC-BY
- New license agreement would specify date from which it would apply
- Some relatively minor technical, legal, and operational implications

Hrynaszkiewicz I, Cockerill MJ: **Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals.** *BMC Research Notes* 2012 (in press)

BioMed Central
The Open Access Publisher

Proposed new license statement

"© 2012 <Author> et al.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data included in this article, its reference list(s) and its additional files, are distributed under the terms of the Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/; http://www.biomedcentral.com/about/access)."

But what do we mean by data?

- Definitions vary quite widely
- For implementation, general guidelines with some specific examples needed
- Examples in journal articles/additional files include tabular data, XML, graphical data points, bibliographic data (including reference lists), RDF

CC0 uses cases in journal publishing

- Text mining e.g. Testing of analysis tools against data harvested from journals
- Open bibliography – diversification and democratisation of impact measures
- Faster progress in areas where lack of public (combinable) datasets are hampering research e.g. EvoMRI

Open by default – opt out

- Public domain not always possible
- Non-standard licenses needed, as already happens for e.g. US government employees
- Very few changes to standard procedures and author behaviour needed for implementation

FAQs?

- Will I risk loss of credit (citations)?
- Will I put competitors at an advantage?
- Will plagiarism be more likely?
- Will I lose any right to express wishes about future uses of my data?
- ....?