# Data Sharing: Data Explorations Workshop

Feb 2008

Katrina L. Kelner
Deputy Editor, Life Sciences
Science Magazine

Science

AAAS

# *Science*’s Data Access Policy (Short Version)

All data necessary for a reader to understand and assess the conclusions of the manuscript must be available to any reader

# Where the Data Reside

- Included in print version of the paper
- Included in Supplementary Online Material (pdf) (freely available)
- If too large, deposit in public database (standard formatting)
    - GenBank (genome sequences)
    - Protein DataBase (protein structures)
    - GEO (gene expression studies)
    - Climate data

# What if the data set is large but there is no database?

- Not desirable
- The author may host the data on his or her web site if they certify that it will remain there, unchanged for 5 years.
- In addition, a copy of the data must be sent to Science on a CD for archiving.

# Science's Data Access Policy: (Details)

After publication, all data necessary to understand, assess, and extend the conclusions of the manuscript must be available to any reader of *Science*.

# Science's Data Access Policy: (Details)

We recognize that discipline-specific conventions or special circumstances may occasionally apply, and we will consider these in negotiating compliance with requests.

# Science's Data Access Policy: (Details)

For further information about accessibility of data and materials, see the following resources:

Cech, T. R. (2003), *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*.

American Psychological Association, *Responsible Conduct of Research: Data Sharing and Data Archiving*.

National Science Foundation Policy on Data Sharing.

# Ambiguity #1: What are “data”?

--Raw data: X-ray film, traces, or counts from machine

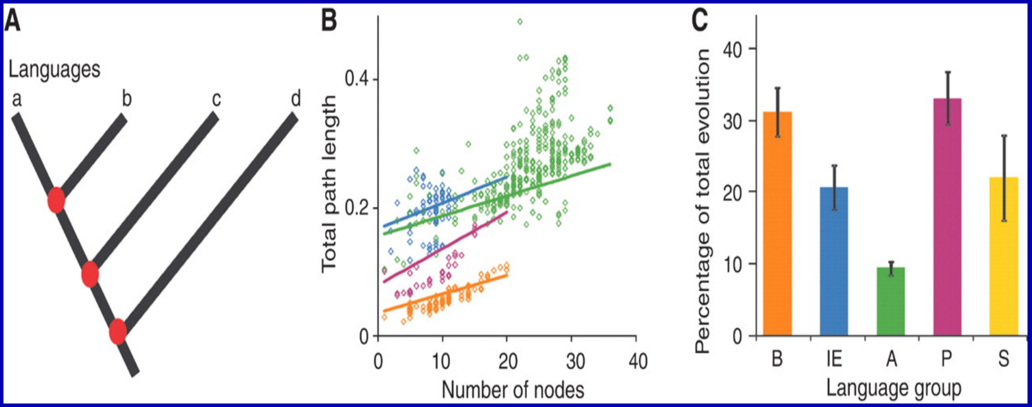
--Processed data, sequence data

--Summaries of all data values

--Tables and figures prepared for publication

# Example of Ambiguity #1: What are data?

- Atkinson et al., Languages Evolve in Punctuational Bursts
- A study of Bantu, Indo-European, Austronesian, and Polynesian languages shows that up to one-third of their words arose in rapid evolutionary bursts from the predecessor tongue



# Data included in Suppl. Online Material

Indo-European

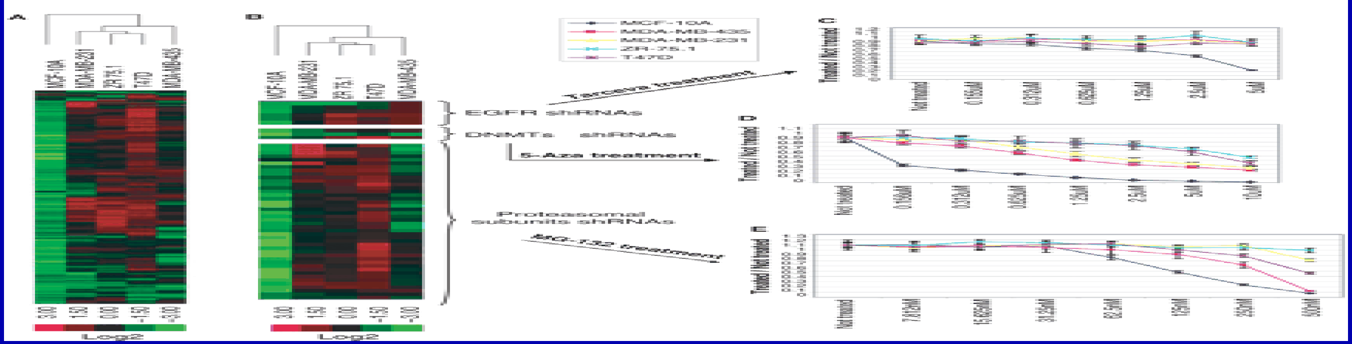
Afghan, Afrikaans, Albanian_Top, Armenian_Mod, Baluchi, Bengali, Brazilian, Breton_List, Bulgarian, Byelorussian, Catalan, Czech_E, Danish, Dutch_List, English_ST, Faroese, Flemish, French, French_Creole_C, Frisian, German_ST, Greek_Mod, Gujarati, Gypsy_Gk, Hindi, Hittite, Icelandic_ST, Irish_A, Italian, Kashmiri, Lusatian_U, Macedonian, Marathi, Nepali_List, Ossetic, Panjabi_ST, Penn_Dutch, Persian_List, Polish, Portuguese_ST, Provencal, Riksmal, Rumanian_List, Russian, Sardinian_L, Serbocroatian, Singhalese, Slovak, Slovenian, Spanish, Swedish_Up, Tadzik, Takitaki, Tocharian_B, Ukrainian, Vlach, Wakhi, Walloon

all, and, animal, ashes, at, back, bad, bark (of a tree), because, belly, big, bird, to bite, black, blood, to blow (wind), bone, tobreathe, to burn (intransitive), child (young), cloud, cold (weather), to come, to count, to cut, day (not night), to die, to dig, dirty, dog, to drink, dry (substance), dull (knife), dust, ear, earth (soil), to eat, egg, eye, to fall (drop), far, fat (substance), father, to fear, feather (large), few, to fight, fire, fish, five, to float, to flow, flower, to fly, fog, foot, four, to freeze, fruit, to give, good, grass, green, guts, hair, hand, he, head, to hear, heart, heavy, here, to hit, hold (in hand), how, to hunt game), husband, i, ice, if, in, to kill, know (facts), lake, to laugh, leaf, left (hand), leg, to lie (on side), to live, liver, long, louse, man(male), many, meat (flesh), moon, ………..

# **Ambiguity #2:** How to present huge datasets, in Suppl. Online Material?

Silva et al. **Profiling Essential Genes in Human Mammary Cells by Multiplex RNAi Screening**

Systematic inhibition of gene expression with RNA interference screening reveals genes essential for growth and survival of tumor cells, potentially leading to new cancer drugs.



# Data in Suppl. Online Material

- **Supporting Tables S3 and S5 to S8.** Additional data tables, packaged as worksheets in an Excel workbook. File is packaged as a compressed archive, in *.zip format
- Five 9 x 817 tables, listing genes and results for each gene
- Public database would be better

## Ambiguity #3: Sometimes sharing data violates the law

- Frayling et al. A Common Variant in the *FTO* Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity
- Accession numbers for deposited sequence variants from dbSNP are Exon3_A 69374768, 3_UTR_A 69374769, 3_UTR_B 69374770, and 3_UTR_G 69374771.
- The Wellcome Trust case-control Consortium Data Access Committee will consider applications for access to genotype data and samples. Access to data will be granted to qualified investigator for appropriate use. The Committee is concerned only with access to the core, anonymised, genotype data and samples generated by this study, and the only phenotypic information held by the Consortium is that which is implied by membership of a particular case or control group.

# Obstacles to seamless data sharing

- Non-uniformity of format
- Lack of public databases
- Author’s desire to capture first rewards from data
- Technical barriers to sharing
- Privacy laws
- MTAs
- Field-specific conventions (computer code, ecological studies)

# Proposition

The traditional print-based scientific paper is no longer the optimal format for presenting peer-reviewed scientific results.