# Prevalence and Patterns of Microarray Data Sharing

Heather A. Piwowar, Wendy W. Chapman

## Background

- Sharing research data is a cornerstone of science
- Prevalence of data sharing is not well understood
- The most comprehensive method for measuring occurrences of public data sharing is manual curation
- Our preliminary manual curation of 100 papers:
  - 30% of investigators publicly share their full microarray datasets:
    - 70% in NCBI's Gene Expression Omnibus (GEO)
    - 20% at EBI's ArrayExpress
    - 10% in smaller databases or lab or publisher websites.
- Manual curation is extremely time consuming
- Here we perform larger-scale automated estimates

## Limitations

- Prevalence depends on study selection criteria
- Some of the retrieved papers only reuse data, and thus are not relevant to the question of sharing. This may confound our results, but we estimate the effect is small.

## Conclusions

- Natural language processing techniques could be helpful for curation based on full-text data sharing statements
- Our automated approach yielded conservative estimates of 20% sharing within GEO and ArrayExpress
- Prevalence patterns suggest several promising hypotheses for understanding data sharing behavior
- We hope these preliminary results will inspire additional investigations into data sharing behavior, and in turn the development of effective policies and tools to facilitate this important aspect of scientific research.

## Acknowledgments

HAP funded by NLM Training Grant

*For more information: heather.piwowar@alumni.pitt.edu*
*Poster will be available at Nature Precedings.*

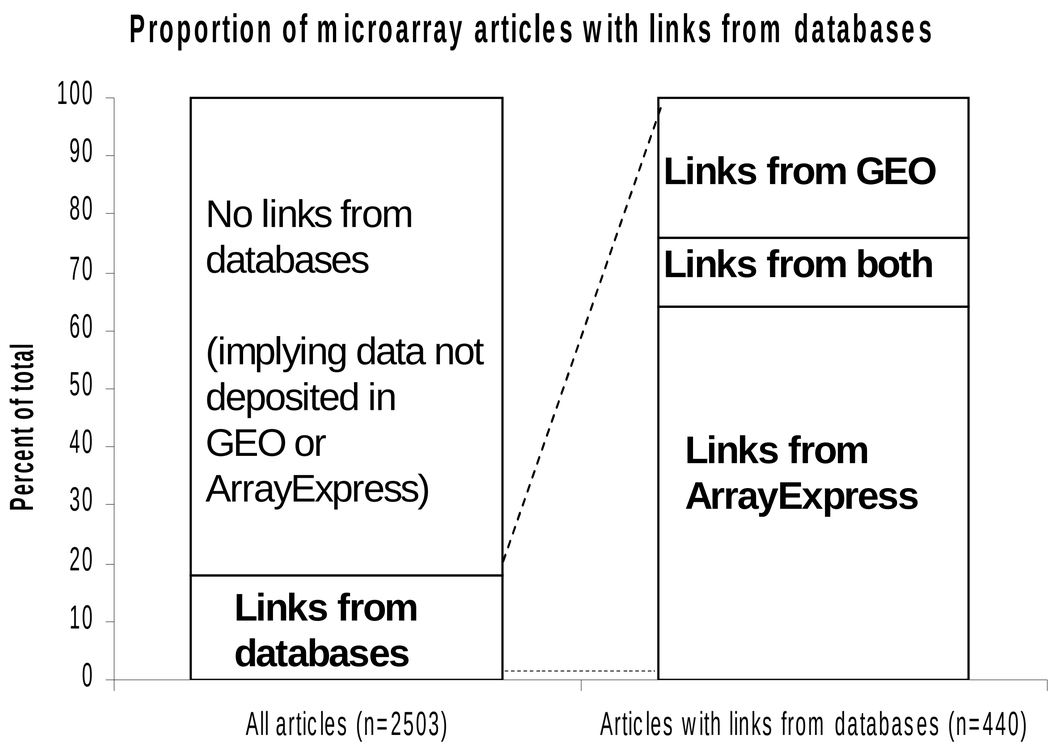
Nature Precedings : doi:10.1038/npre.2008.1701.1 : Posted 20 Mar 2008

## What proportion of microarray datasets are shared in public databases?

### About 20% of microarray studies had links from publicly available databases

- Extracted research articles with MeSH terms for both "Gene Expression Profiling" and "Oligonucleotide Array Sequence Analysis" published in 2006
- Searched GEO and ArrayExpress for links to these PubMed IDs to determine which articles had been credited as an originating data source

- 2503 articles
- 440 (18%) had links from either GEO or ArrayExpress:
  - 70% had links from GEO
  - 30% had links from ArrayExpress
  - Overlapping 12% from both




## Does sharing data correlate with other factors?

### Prevalence varies with time, publishing decisions, study topic and subjects, and funding source

- Prevalence of database links is 2.5x higher in 2006 than 2002.
- Studies with free full text at PubMed are 2.1x more likely to have links
- Studies published in a core clinical journal are 1.4x likely to have links as those published elsewhere
- Studies with human data were less likely to have a link when the trial was related to cancer (OR=0.8), but otherwise no different than studies on other species
- Studies funded by NIH were 2.0x more likely to have links than studies not funded by the NIH
- Increased number of funding sources correlates with a higher prevalence of sharing