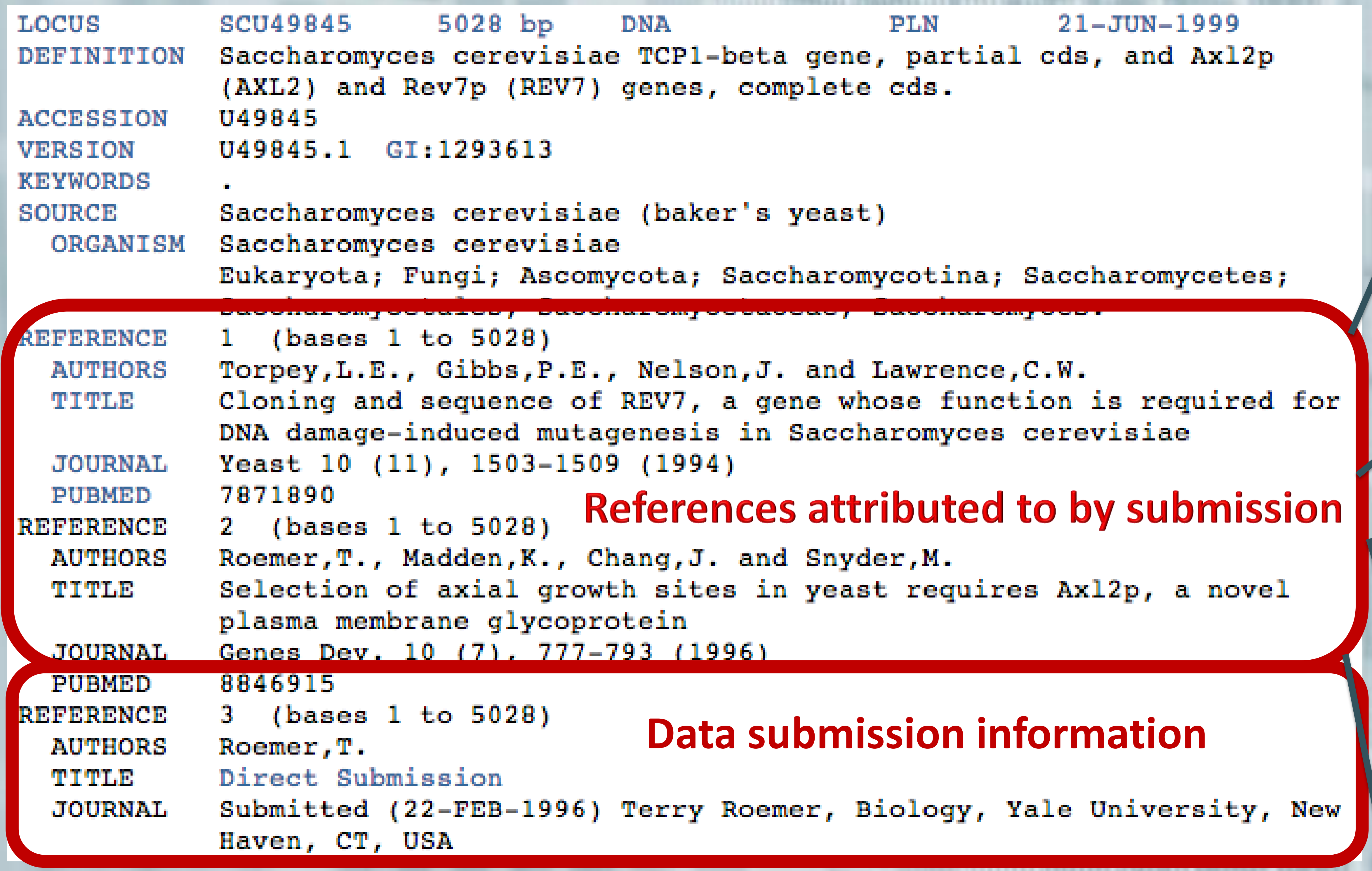
Attributions from Data Authors to Publications: Implications for Data Curation

Jian Qin, Mark Costa, Jun Wang
School of Information Studies
Syracuse University

GenBank Metadata

A typical GenBank record includes a DNA/RNA sequence and the metadata describing the sequence data. Each GenBank record contains one data submission and a metadata section describes the date, type, source, definition, authors of the sequence data as well as the metadata for publication(s) or patent detailing the discovery and provenance of the sequence.

Metadata section of GenBank record



```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
```

References attributed to by submission

Data submission information

Attribution patterns

One-to-one: the submission referenced one publication and the submitter is one of the authors of the publication referenced.

One-to-many: this pattern appears to have two variations:
a) Many submissions (records) from the same group or author point to the same publication and the data submitter is also one of the authors for the reference cited;
b) One data submission references more than one publication in the same record.

Non-overlapping: there is no overlap between the authors for the publication(s) referenced and for the sequence submitted. The data submitter appears to have used biomaterial from others to generate the DNA sequences.

Resubmission: a record has two or more submissions in which the old submission is replaced by the newer one with attribution made to the older reference(s) and submission(s).

Challenges and Implications

- Loose metadata standards, including entity resolution, make analysis difficult.
- Analyzing attribution patterns in the data is not only useful for retrieval, but also for assessment.
- Quantitative, algorithm-based analysis is improved by quality data curation and hindered by poor data quality.
- The costs of assessments using data repositories decreases with higher quality data curation practices and increases with lower quality curation practices.

National Science Foundation
WHERE DISCOVERIES BEGIN