Developing Standards for Sequencing Methods Reporting

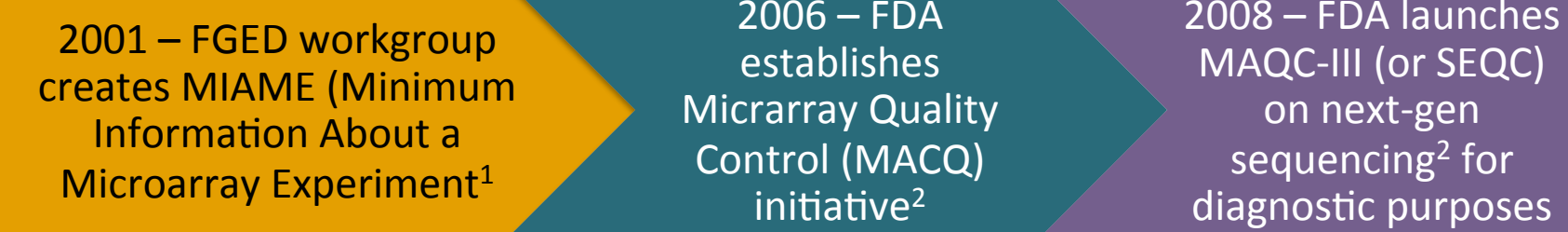Pamela L Shaw[1] & Simon M Lin[2,3]

Galter Health Sciences Library[1], Northwestern University Biomedical Informatics Center (NUBIC), Clinical & Translational Sciences Institute (NUCATS)[2],
Feinberg School of Medicine, Northwestern University, Chicago, IL 60611[1,2]
Director, Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI 54449[3]
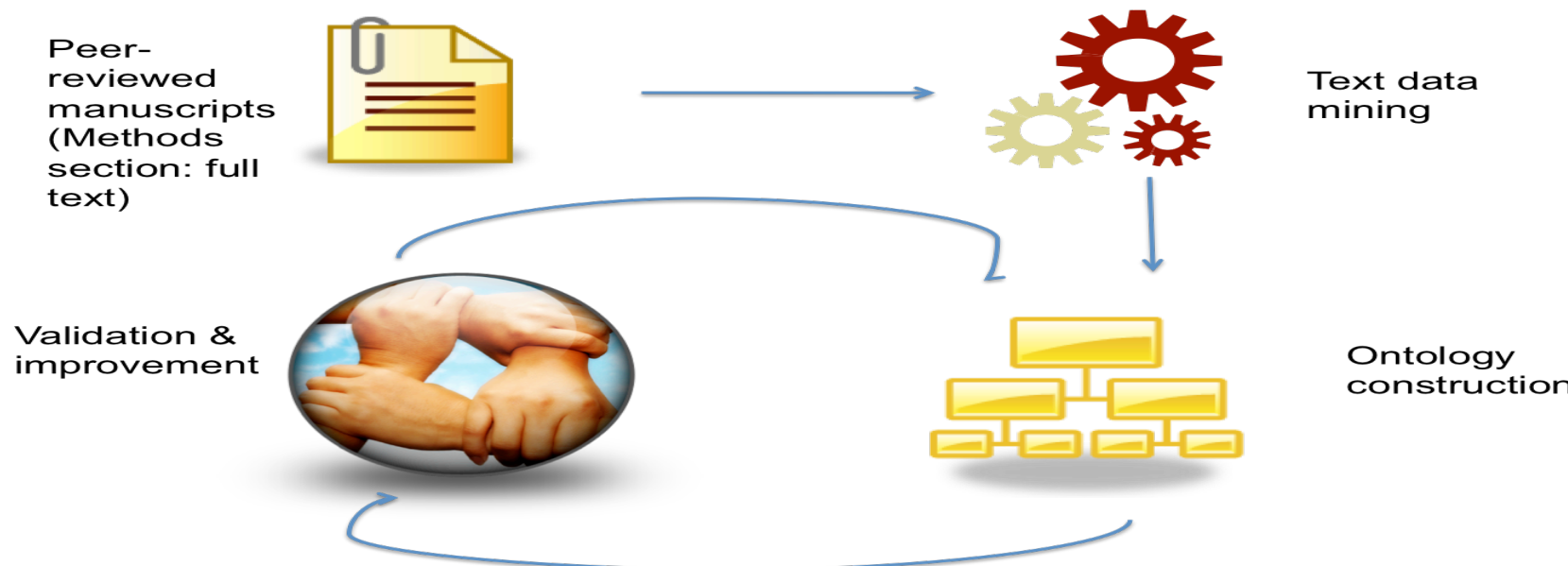
Objectives

Bioinformatics methods are becoming common in biomedical publication. No guidelines exist for reporting of bioinformatics and data analysis methods to guide authors who publish research in genome sequencing-specifically ChIP-Seq and RNA-Seq experiments-so duplication of data analysis methods by other labs is difficult, if not impossible. This project mined sequencing methods text in order to recommend best-practices data analysis reporting.

Background - Timeline

- 2001 – FGED workgroup creates MIAME (Minimum Information About a Microarray Experiment[1]
- 2006 – FDA establishes Micrarray Quality Control (MACQ) initiative[2]
- 2008 – FDA launches MAQC-III (or SEQC) on next-gen sequencing[2] for diagnostic purposes

Methods - Overview

Ontology Construction Process

Peer-reviewed manuscripts (Methods section: full text)

Validation & improvement
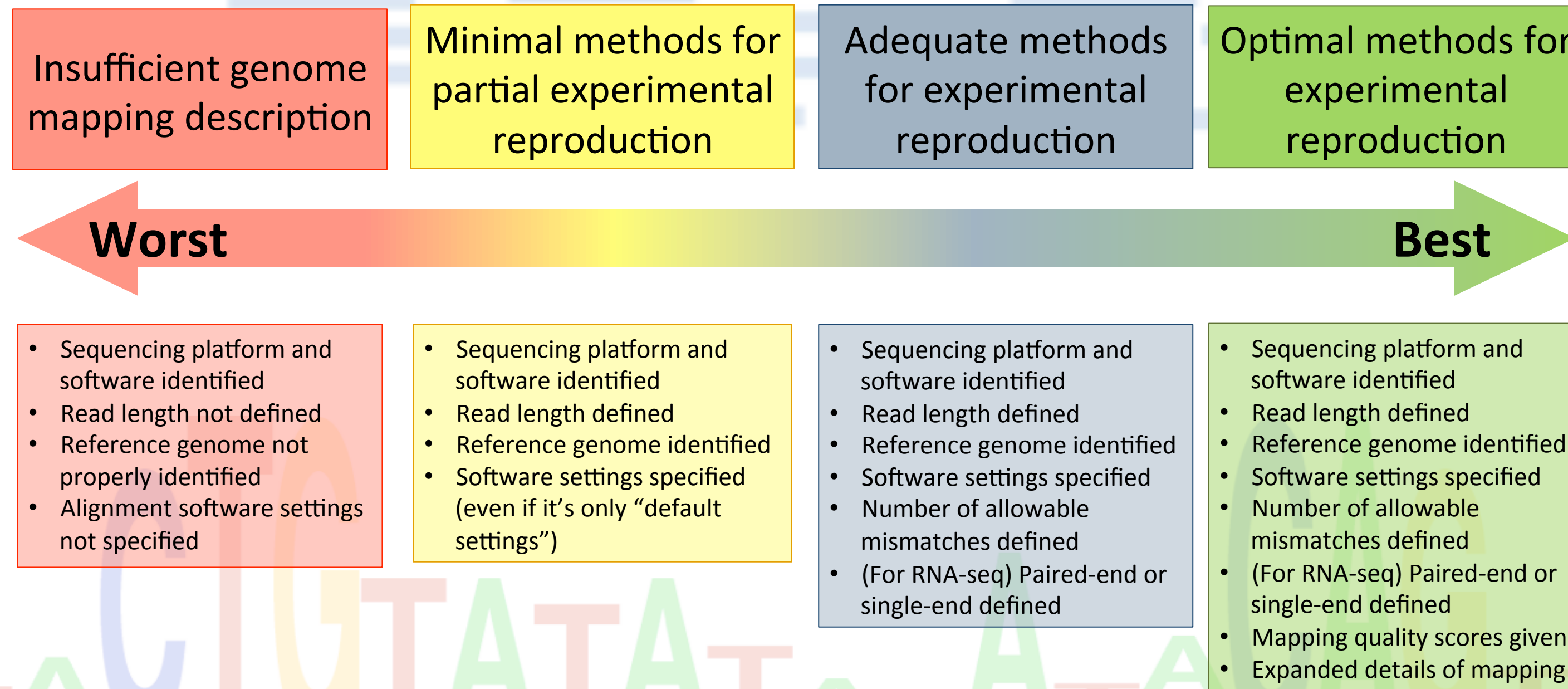
Text data mining

Ontology construction

Methods - Searches

- PubMed searched in a 2 year window for ChIP-seq or RNA-seq
- Full text Methods (or Supplemental text, if relevant) of resultant articles searched for methods describing bioinformatics or data analysis—specifically: mapping reads to reference genome
  - Exclusion criteria:
    - No mention of genome mapping step
    - Manuscript employed previously-described methods
   - Manuscript introduced a new mapping protocol

Methods – Coding and Sorting

- Methods text along with citation information was entered into Excel then read into a MySQL database for text mining
- Methods were mined, identified, and sorted into categories:

Insufficient genome mapping description	Minimal methods for partial experimental reproduction	Adequate methods for experimental reproduction	Optimal methods for experimental reproduction

Worst ←——————————————→ **Best**

• Sequencing platform and software identified • Read length not defined • Reference genome not properly identified • Alignment software settings not specified	• Sequencing platform and software identified • Read length defined • Reference genome identified • Software settings specified (even if it's only "default settings")	• Sequencing platform and software identified • Read length defined • Reference genome identified • Software settings specified • Number of allowable mismatches defined • (For RNA-seq) Paired-end or single-end defined	• Sequencing platform and software identified • Read length defined • Reference genome identified • Software settings specified • Number of allowable mismatches defined • (For RNA-seq) Paired-end or single-end defined • Mapping quality scores given • Expanded details of mapping

Results (thus far)

DNA crosslinked with protein for ChIP-Seq

- Papers meeting criteria:
  - 28 for ChIP-Seq
  - 23 for RNA-Seq
- Only ~15% of ChIP-Seq papers provided optimal methods detail; nearly 50% of RNA-Seq manuscripts provided optimal detail
- Initial summaries presented to FDA for ChIP-Seq in December 2010 and for RNA-Seq in December 2011
- Working groups to be formed to assess and recommend protocol ontologies
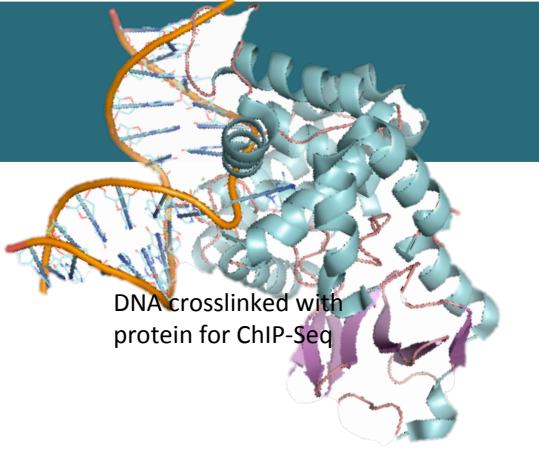
Conclusions Reached (thus far)

- GOAL: "It is about consistency and reproducibility by a single method, not about the comparison of methods." - Simon Lin
- Reproducible detail must be improved
- This is a work-in-progress that requires voluntary participation from sequencing & bioinformatics professionals from all institutions

Sources Cited

1. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet. 2001;29(4):365-71. PubMed PMID: 11726920.
2. U.S. Food and Drug Administration. Microarray Quality Control (MAQC) [updated 09/29/2011]. Available from: http://www.fda.gov/ScienceResearch/BioinformaticsTools/MicroarrayQualityControlProject/.