LINCS: Example of Handling Multimodal data and Integration

Ajay Pillai

NHGRI

Breaking News

- New Funding Opportunity RM17-026: Establish an NIH Data Commons https://commonfund.nih.gov/sites/default/files/RM-17-026_CommonsPilotPhase.pdf
  - Cloud based
  - Define and Access FAIR* Biomedical Data
  - Start with TopMED, GTEx and Model Organism Databases

https://www.nhlbi.nih.gov/research/resources/nhlbi-precision-medicine-initiative/topmed

https://www.gtexportal.org/

https://www.genome.gov/10001837/model-organism-databases/

*Findable, Accessible, Interoperable and Reusable https://www.nature.com/articles/sdata201618

Example Use Case: Define peripheral blood-tissue molecular signatures of the 'Metabolically Healthy Obese' phenotype by linking clinical, genomic, and transcriptomic data in the TOPMed and GTEx datasets. Next query the Model Organism Databases (MODs) to ascertain whether 'loss-of-function' mutants or gene-knockout animals exhibit a similar 'metabolically healthy obese' phenotype as observed in humans.

Outline

- Introduction to LINCS data types
- Usable Data
- LINCS Data usable?
- Usable Software/Tools
- LINCS Software/Tools usable?
- One complete and important scientific story
- Conclusion

Two Ideas Today about the LINCS program

- LINCS has lots of data and I will try and show you what we are doing to make it useful to the world.

- Give you one example of very useful “theory” that came out of LINCS work with potentially important impact on Cancer drug discovery.
  - The right “theory” is extremely important in Big Data analytics.

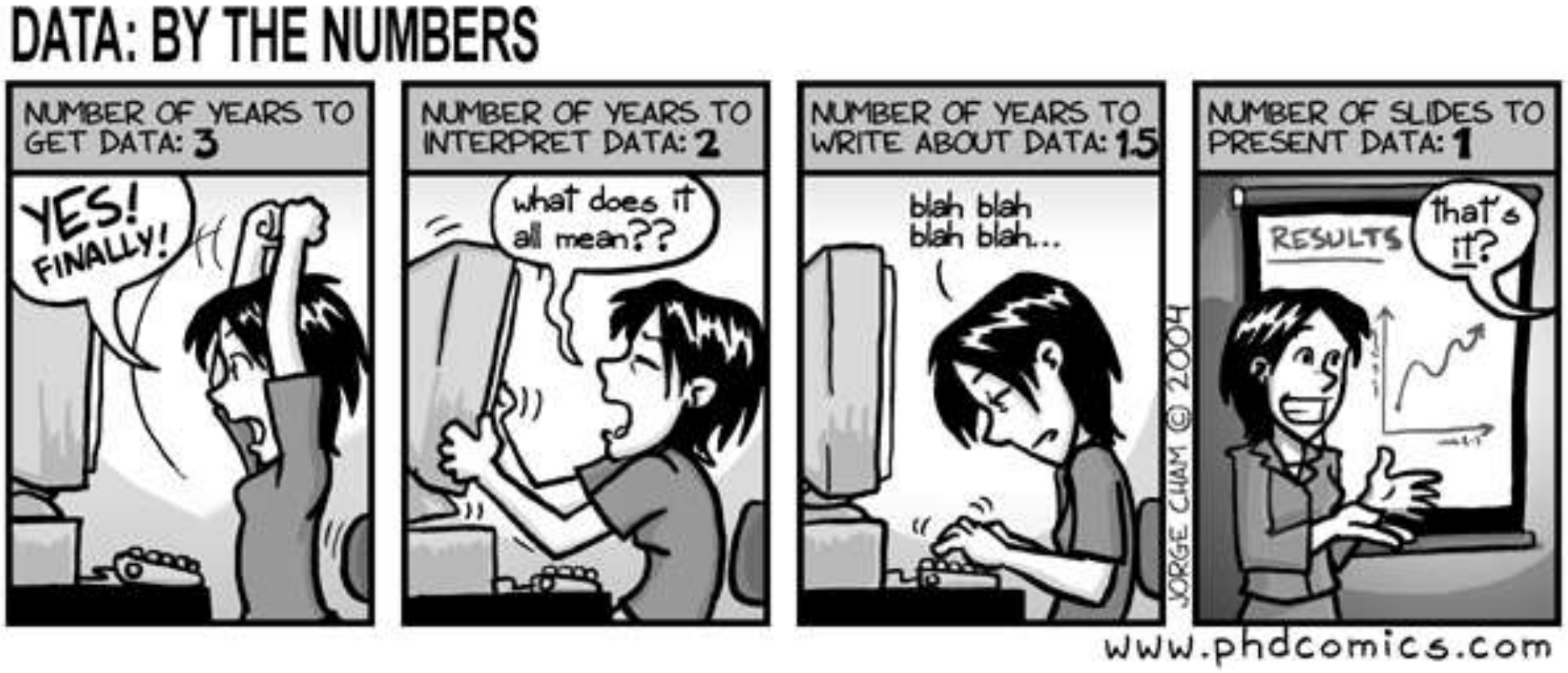
Introducing the breadth and scope of LINCS Data

- A cube with a lot of randomly distributed holes.
  - Sub-cubes that are filled-out.
- On the cell type axis we have (all human):
  - Cell lines, iPS, iPS-derived and differentiated, primary cells
- On the perturbation axis we have:
  - Drugs/small molecules, genetic, disease (ALS, SMA)
- On the assay axis we have:
  - Transcriptomics (L1000, RNASeq), Proteomics (MS-based, Phospho, RPPA, DIA/SWATH), Epigenetic (ATACSeq, Chromatin Profiling), Imaging (Microenvironment Microarray, Robotic Microscopy, Live Cell Imaging)
- Not being a complete matrix of cells makes the project challenging for data integration, but also opens it up for a more realistic data integration problem.

Breadth and Scope of LINCS Data

- We have the largest collection of systematically generated data, using standard protocols on a wide range of cell types:
  - Transcriptomics data (>1M profiles of reduced representation)
  - Phospho-proteomics data (this is also undertaken in DIA/SWATH mode)
  - Microenvironment Perturbation data

LINCS Data: plagued by standard challenges



NIH LINCS PROGRAM

Outline

- Introduction to LINCS data types
- Usable Data
- LINCS Data usable?
- Usable Software/Tools
- LINCS Software/Tools usable?
- One complete and important story
- Conclusion

What is usable data?

- Helping prepare for the ‘fourth paradigm’ in biology.
- Demand rises if data is discoverable, accessible, usable.
- What do we know about usable data?
  - Posting data or links is not usable.
  - Mix of data types and formats WITHOUT associated publications & context is not usable.
  - No usability without sustainability.

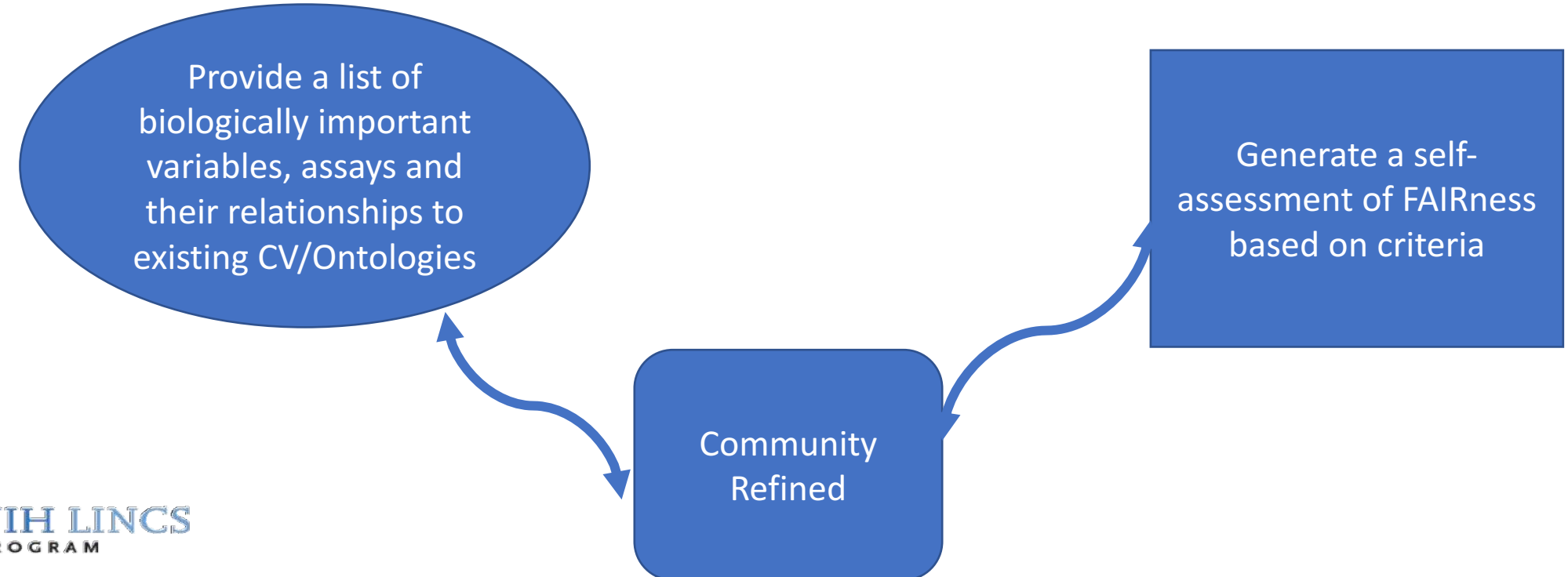
- The FAIR metric assesses some of these challenges:
  - Findable, Accessible, Interpretable, Reusable https://www.force11.org/group/fairgroup/fairprinciples

The Fourth Paradigm: Data-Intensive Scientific Discovery
Hey, Tansley, Tolle

Is LINCS data usable?

Findable, Accessible, Interoperable, and Reusable (FAIR)

- Lots of metadata available.
- LINCS has begun a self-assessment of FAIRness
  - Challenge: How do you assess if a clear explanation of the variables has been provided?



NIH LINCS PROGRAM

FAIR Principle	Description
F	Standardized IDs are used to identify dataset
F	Using intuitive search terms, the dataset appears in the first page of search engine results
FR	(meta)data are assigned a globally unique and eternally persistent identifier ~ Research Resource Identifiers (RRIDs) are used in associated publications
A	The dataset's metadata use vocabularies that are compliant with the FAIR principles
A	The dataset is retrievable by a standardized protocol
A	The dataset is available in a human-readable format
A	The dataset is available in a standard machine-accessible format
AI	The dataset is available in standard and useful format(s) that is/are interoperable with popular analysis tools
AR	User-selected subsets of the full dataset can be downloaded
I	Metadata are linked to other datasets, vocabularies and ontologies
R	The meta(data) are sufficiently complete to permit effective reuse ~ for LINCS: all Level 1 metadata are present
R	A tutorial page is provided for the dataset to describe the format of the dataset ~ A standard structured metadata description (e.g. HCLS dataset description) is associated with the resource
R	Information is provided describing how to cite the dataset
R	A description of the methods used to acquire the data is provided
R	Licensing & Version information is provided on the dataset's landing page
R	Tools that can be used to analyze the dataset are listed on the dataset's landing page
R	Contact information is provided for the originator(s) of the dataset

FAIR Principle	Description
F	Standardized IDs are used to identify the tool
F	Using intuitive search terms, the tool appears in the first page of search engine results
FA	The tool is hosted in one or more well-used repositories
A	An explicit commitment has been made to maintain all versions of the tool to be available indefinitely ~ Funding is available to fulfill this commitment
A	Source code is shared on a public repository
A	Code is written in an open-source, free programming language
AIR	The tool inputs standard data format(s) consistent with community practice
AIR	All previous versions of the tool are made available
AR	Web-based version is available (in addition to desktop version)
R	Source code is documented
R	Code libraries are up to date
R	Pipelines that use the tool have been standardized and provide detailed usage guidelines
R	A tutorial, case study, example datasets is provided for the tool
R	Licensing, citation, version information is provided on the tool's landing page
R	A paper about the tool has been published
R	Contact information is provided for the originator(s) of the tool
R	A mechanism for reporting problems with the tool is available

Is LINCS data usable ctd: can we go beyond?

- What does a scientist need:
  - First and foremost needs help in answering the question: “is this data relevant to me?”
  - We started down this path: we have a UI that supports workflows based on scientific questions stated in english: http://lincsproject.org/LINCS/tools
  - We have tools, data, and exploration based on publications: http://lincs.hms.harvard.edu/fallahi-sichani-molsystbiol-2017/

Outline

- Introduction to LINCS data types
- Usable Data
- LINCS Data usable?
- Usable Software/Tools
- LINCS Software/Tools usable?
- One complete and important story
- Conclusion

What are usable software/tools?

- It should,
  - Accomplish goals, responsive, consistent behavior, efficient, engaging.
- Scientific Software:
  - Is it implementing what is missing from existing software?
  - Do you collect user feedback?
  - Use standard data formats
  - Are you ready for data growth?
  - Expose parameters (not all)
  - Provide logs (configuration & technical)
  - Avoid complex setup
  - Tutorial
  - Long-term availability

Are LINCS software/tools usable?

- LINCS data pipelines:
  - Validated by independent group (Data Coordination Center)
  - Documentation of the pipeline and standardization (as much as possible)
  - Available as open-source in public locations like github

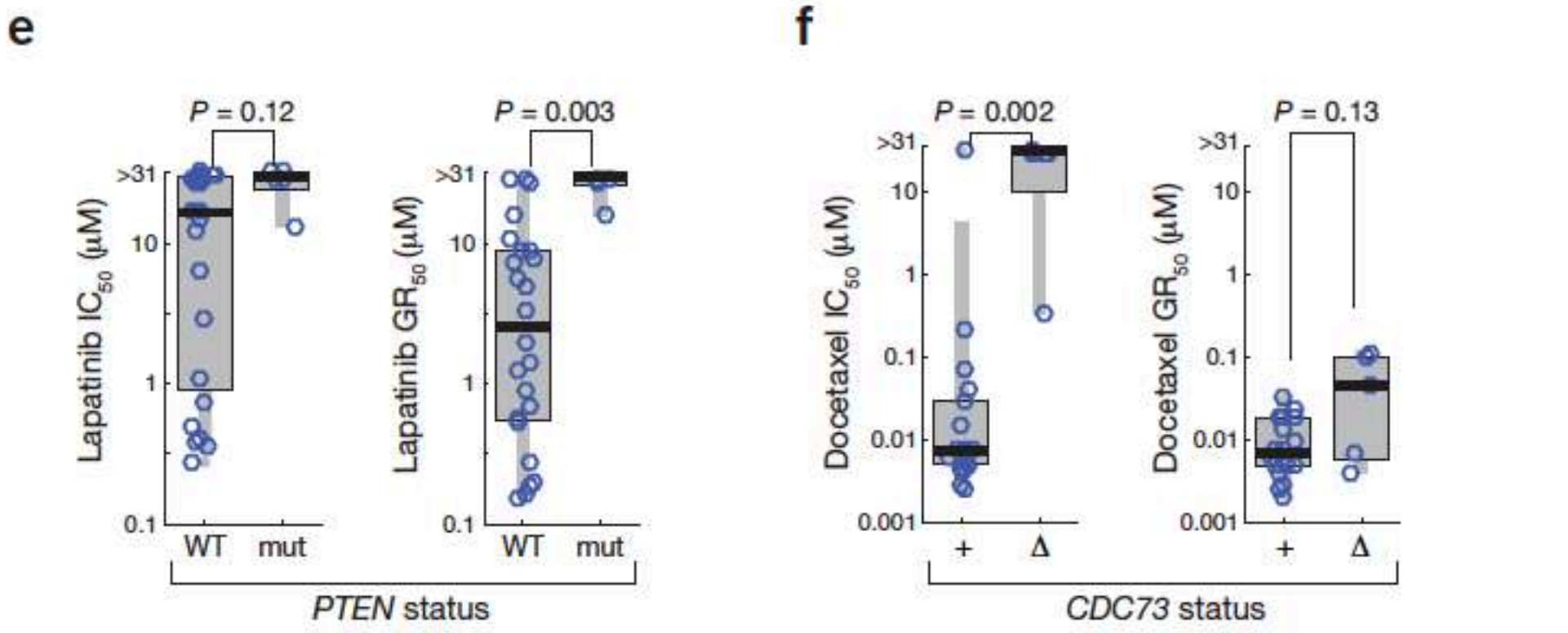
- LINCS analytical tools:
  - Connections to link together LINCS and non-LINCS data:
    - Tools like ENRICHR and OMICSintegrator
  - Shiny Apps
  - Jupyter Notebooks

Outline

- Introduction to LINCS data types
- Usable Data
- LINCS Data usable?
- Usable Software/Tools
- LINCS Software/Tools usable?
- One complete and important story
- Conclusion

GR (Growth Rate) Metrics: In Cancer Chemotherapy

Hafner etal, Nat Biotech, 35(6), June 2017



False Negative association by looking at IC50

False Positive association by looking at IC50

GR Metrics: What they achieve

- GR metrics have the potential to improve pre-clinical pharmacology
  - IC50 leads to multiple FP & FN
    - cell division rates vary in a systematic manner with tumor type and genotype.
  - GR metrics correct not only for this but also for arbitrary differences in protocol that affect cell-division number
  - GR metrics reduce the impact of confounders, cross-study reproducibility is improved.
  - Efficacy as measured by GR_{max} and potency as measured by GR50 differ at a biological level and are associated with largely *non-overlapping* genetic alterations

Important pharmacogenomic associations are not those associated with low IC50 values, but rather those that result in the *most negative GR value at clinically relevant drug concentrations*

Growth Response (GR) Metrics: LINCS Manuscript Collection

- Fallahi-Sichani M, Honarnejad S, Heiser LM, Gray JW, Sorger PK. Metrics other than potency reveal systematic variation in responses to cancer drugs. Nature chemical biology. 2013;9(11):708-14. Epub 2013/09/10. doi: 10.1038/nchembio.1337 nchembio.1337 [pii]. PMID: 24013279; PMCID: PMC3947796.
- Hafner M, Niepel M, Chung M, Sorger PK. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. Nat Methods. 2016;13(6):521-7. Epub 2016/05/03. doi: 10.1038/nmeth.3853. PMID: 27135972; PMCID: PMC4887336.
- Marc Hafner, Mario Niepel, Kartik Subramanian, and Peter K. Sorger "*Designing drug response experiments and quantifying their results*" (2017) Current Protocols in Chemical Biology, in press
- Mario Niepel, Marc Hafner, Mirra Chung, and Peter K. Sorger "*Measuring cancer drug sensitivity and resistance in cultured cells"*. (2017) Current Protocols in Chemical Biology, in press.
- Mario Niepel, Marc Hafner, Qiaonan Duan, Zichen Wang, Evan O. Paull, Mirra Chung, Xiaodong Lu, Joshua M. Stuart, Todd R. Golub, Aravind Subramanian, Avi Ma'ayan, and Peter K. Sorger "*Common and cell-type specific responses to anti-cancer drugs revealed by high throughput transcript profiling* in review. Nature Communications, in revision.
- Marc Hafner, Laura M. Heiser, Elizabeth H. Williams, Mario Niepel, Nicholas J. Wang, James E. Korkola, Joe W. Gray and Peter K. Sorger "*Quantification of sensitivity and resistance of breast cancer cell lines to anti-cancer drugs using GR metrics"* Scientific Data, in press.
- Marc Hafner, Mario Niepel and Peter K. Sorger "*Improving preclinical cancer pharmacogenomics by using alternative drug sensitivity metrics*" in review (2017) Nature Biotechnology, in press.
- Nicholas A Clark, Marc Hafner, Michal Kouril, Elizabeth H Williams, Jeremy L Muhlich, Marcin Pilarczyk, Mario Niepel, Peter K Sorger, Mario Medvedovic "*GRcalculator: an online tool for calculating and mining dose-response data"* BMC Bioinformatics, in review.
- Mario Niepel, Marc Hafner, Mirra Chung and Peter K. Sorger "*DyeDrop, a rapid and minimally disruptive staining method for phenotyping cells in drug response assays"* submitted.

Outline

- Introduction to LINCS data types
- Usable Data
- LINCS Data usable?
- Usable Software/Tools
- LINCS Software/Tools usable?
- One complete and important story
- Conclusion

Conclusions

- LINCS has already established standards and approaches that other programs and projects can adopt.
  - A number of new programs are working with us to adopt these.

- We as a community, including LINCS, have a long way to go to enable the fourth-paradigm in biology.
  - A number of key elements, however, are lining up.

Future Directions & Unmet Challenges

- Formatting and making it useful:
  - Things yet to be done: put our data through “tidy-ing”, e.g. OpenRefine http://openrefine.org and tidyr https://blog.rstudio.org/2014/07/22/introducing-tidyr/
- Investigating logins and creating a user community.
- More work on “what does a scientist need?”

Future Directions & Unmet Challenges 2

- Started down the path of Docker Containers:
  - This looks to be essential
  - Could be challenging for Imaging and perhaps for SWATH proteomics

- Connections to other people's tools:
  - A pipeline that can pass data for further analysis to non-LINCS tools.
  - This requires 'standardizing' the output formats for ALL LINCS tools.

- Continue to re-visit the question: "what does a scientist need?"

Future Directions & Unmet Challenges 3

- Availability on the Cloud

- Lookout for a new BD2K FOA on community-developed models for cloud deployment:
  - Hope to create efficiencies
  - Hope to standardize principles
  - Develop solutions that are usable for ALL NIH communities (aka ICs)