Multi-omics 2020-November-23 thomas.stoeger@northwestern.edu

Rationale: **Biology occurs at <u>different scales of complexity</u>**

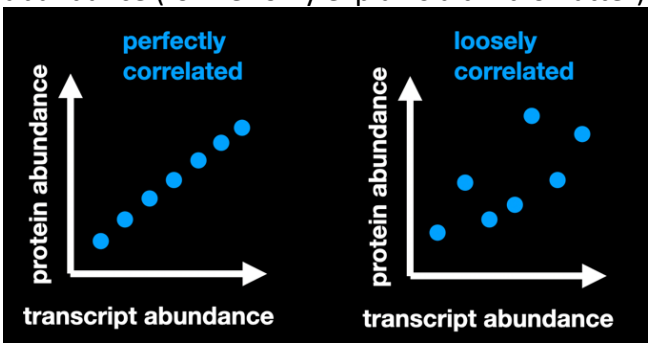
- Understand how they relate to each other
- Make educated guesses about one scale based from another scale
- Consistency across different scales can improve certain applications (e.g.: classifying patients according to disease severity, removing confounding effects)

<u>What is multi-omics?</u>

- Combination of different -omics approaches.
- Comparably small, but rapidly growing field.
    - Can be personally exciting.
    - Often absence of widely accepted tools.
    - Invites creativity: options (for combination) scale exponentially with datasets.

An exemplary tale: **<u>Do proteomes only mildly depend on transcriptomes?</u>**

- Take home: Be careful about **gene-specific scaling** and **measurement noise**.
- Claim: “Proteomes only mildly depend on transcriptomes.”
    - Based on observation that transcript abundance is only loosely correlated with protein abundance (former only explains a third of latter)



    - Widely believed until approx. 2010, and still sometimes mentioned.
    - (Retrospectively?) the claim appears quite unintuitive.
- Explanation:
    - Sequence properties of transcripts and proteins – mainly those that predict the stabilities of transcripts and proteins – uncouple transcript and protein abundance in a highly expected manner (this explains a further third of correlation between transcripts and proteins). Vogel et al., 2010 https://doi.org/10.1038/msb.2010.59 used transcriptomics + proteomics + sequence properties and predictive multivariate models
    - Mass spectrometry still has many **measurement inaccuracies** compared to transcriptomics. If accounting for these measurement inaccuracies, less than 10% of protein abundance remain unexplained by transcriptome and sequences properties relating to degradation of transcripts and proteins. Li et Biggin, 2015, https://doi.org/10.1126/science.aaa8332

Some unconventional, but ultimately useful, **<u>ideas for combining different aspects of biology</u>**:

- Take home: Take advantage of biology happening on different scales of complexity!
- Examples:
    - Use similarity in protein folds to predict interaction of proteins, and in extension virulence of different, new, strains of viruses. Lasso et al. 2019, https://doi.org/10.1016/j.cell.2019.08.005

- Use subcellular localization of transcripts to explain cell-to-cell variability in protein levels. Popovich et al. 2019, https://doi.org/10.1016/j.cels.2018.09.001
- Combine metabolomes, transcriptomes, proteomes, post-translational modifications to understand rapid vs. slow adaptation of bacteria to different nutrient sources. Buescher et al. 2012 https://doi.org/10.1126/science.1206871
- Predict the success of clinical trials by **disregarding published scientific literature** (which is biased) and instead re-evaluating support for involvement of genes in phenotypes of interest by only considering genome-wide assays on gene products and demonstrating effect of polymorphisms on gene products. Nelson et al. 2015, https://doi.org/10.1038/ng.3314
- Relationship between gene-specific properties, patents, and research grants. Oprea et al. 2018 https://doi.org/10.1038/nrd.2018.14
- Predicting reproducibility of individual claims in biomedical literature by tying claims to experimental approaches and social networks of individual authors. Danchev et al. 2019, https://doi.org/10.7554/eLife.43094
- Obtain rough overview on correlation between distinct entities that can be measured in a disease content, e.g.: microbiomes, transcriptomics, clinical parameters. Lloyd-Price et al. 2019 https://doi.org/10.1038/s41586-019-1237-9
- Combining distinct -omes in multivariate models can predict on influenza immunization better than metabolomes, proteomes, clinical labels, transcriptomes, microbiomes alone. Zhou et al. 2019, https://doi.org/10.1038/s41586-019-1236-x
- Identify **host factors that confound microbial studies** of human disease. Vujkovic-Cvijin et al. 2020, https://doi.org/10.1038/s41586-020-2881-9

Data organization – A practical challenge when working with (many) large datasets (including -omes).

- Take home: Know some principles early enough to **save you – or your computational collaborator – multiple months of avoidable work** later on.
- Know the basics:
  - How to organize data in tables, so that they are computationally well traceable? Broman et Woo 2017 https://doi.org/10.1080/00031305.2017.1375989: Be consistent, write dates like YYYY-MM-DD, do not leave any cells empty, **put just one thing in a cell**, organize data as single rectangle (with subjects as rows and variables as columns, and with a single header row), **create a data dictionary**, do not include calculations in the raw data files, do not use font color for highlighting data, choose good names for things, make backups, use data validation to avoid data entry errors, save the data as plain text files
  - Eleven tips for working with large data. Nowogrodzki 2020 https://www.nature.com/articles/d41586-020-00062-z: Cherish your data, visualize the information, **show your workflow**, **use version control**, record metadata, automate-automate-automate, make computing time count, capture your environment, don't download the data, start early with data organization, get help
- If working with data personally:
  - Consider reading a textbook on data integration – which is a discipline that existed independently of biology, and has many problems already solved; recommend: Principles of Data Integration by Doan, Halevy, Ives
  - Consider including/adding field-specific conventions, of which you might have heard, e.g.: tidy data (https://en.wikipedia.org/wiki/Tidy_data)
  - **Wrappers**: Do not access data directly from your analytical code, and instead access it through a function that sits between analytical code and data. This way you can focus on analytics, while allowing data to grow (or be moved to other storage location or storage type when project matures). Also, this strategy will allow you to optimize speed of data access only when needed.

Analytical approaches to combine multiple datasets

- Take home: Be cautious and **know how you know how well approaches work**.
- Some standardized tools – particularly around common -omes, such as those relating to gene expression, e.g.: github.com/mikelove/awesome-multi-omics, but less for more creative multi-omes
- Some different general strategies; Rappoport et Shamir, 2018, https://doi.org/10.1093/nar/gky889 and Stuart et Satija, 2019, https://doi.org/10.1038/s41576-019-0093-7

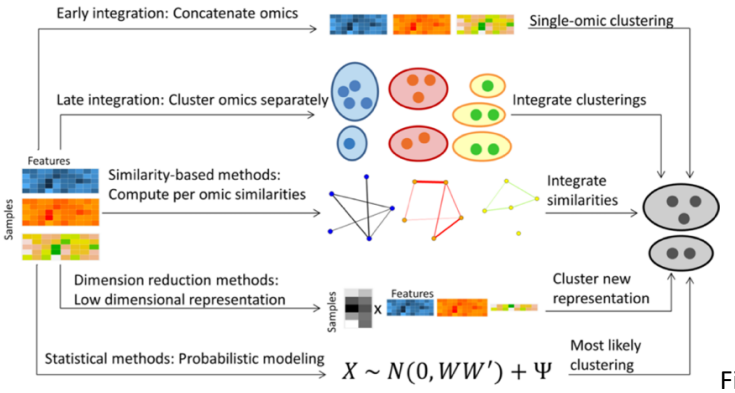


Figure from Rappoport et Shamir

- Performance can ultimately depend on how good underlying assumptions match biology (e.g.: if "true" observations are conserved across distinct modalities of the data – or if they are not); see above review by Rappoport et Shamir for example. This also means that there might not be something like a "best" approach.
- Addressing how well an approach works:
    - **Performance of a modle might not depend much on computational choice**, and instead might reflect absence of having observed the right data modality; E.g.: From field of sociology (regretfully such studies are rare/absent from most domains of biology), Salganik et al., 2020, https://doi.org/10.1073/pnas.1915006117 had 160 scientific teams competing on constructing predictive models from the same massive multimodal data set through many different analytical approaches, and realized that the approaches not only had similar performances, but that their mistakes appeared for the same families (suggesting that lack of observing/measuring right modality, rather than analytical details, were problem)
    - Do **cross-validation**, a data-scientific approach used independent of -omics: Leave out some samples, develop model, and see how well it explains observations among the samples that have not been used to develop the model. E.g.: McCabe et al, 2019, https://doi.org/10.1093/bib/bbz070

How similar are genes to each other when assessed through the lenses of distinct -omes?

- It is possible to assess similarity between two genes from an individual -ome, e.g.: genes with a similar differential expression across perturbations can be said to be more similar to each other; likewise, it is possible to define similarities by known biology (e.g.: if genes are involved in same pathways)
- In next 45 minutes: Ask whether distinct lenses (e.g.: gene expression, pathways, diseases) identify the same pairs of genes to be similar and if/when they provided additional information.
- You can download preprocessed similarities from https://maayanlab.cloud/Harmonizome or (better as faster, but limited to some -omes) from https://northwestern.box.com/s/dvrxld7ioe6jgm2srrs7rj8mzkqpkhsa . For example code see https://github.com/tstoeger/course_multi_omics/blob/main/how_well_do_different_views_on_biology_correlate.ipynb
- **Be creative!**
- Afterwards, interested people can optionally share their results/conclusions.