Analyzing microbiome multi-omics data: Tools and challenges

Cecilia Noecker, PhD

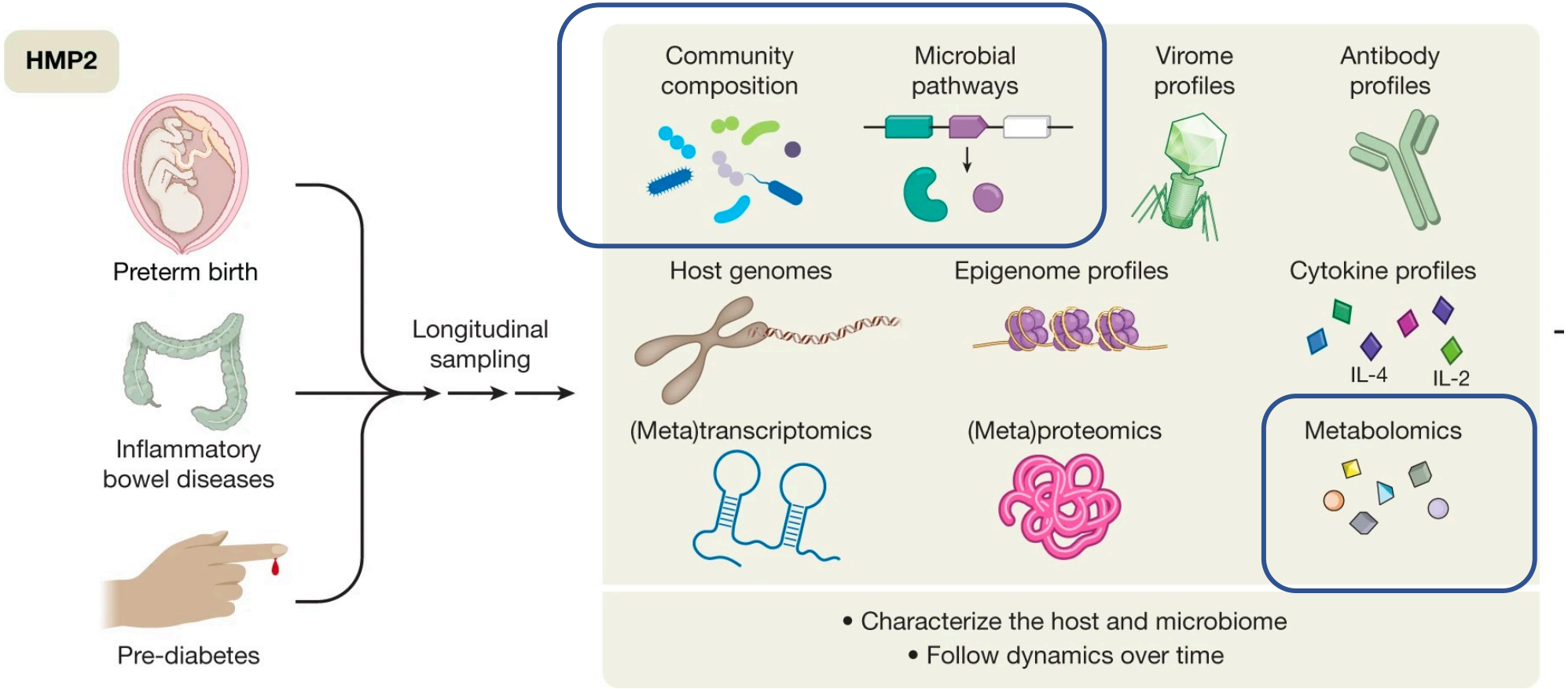
Turnbaugh Lab

May 25, 2021

Goals for today

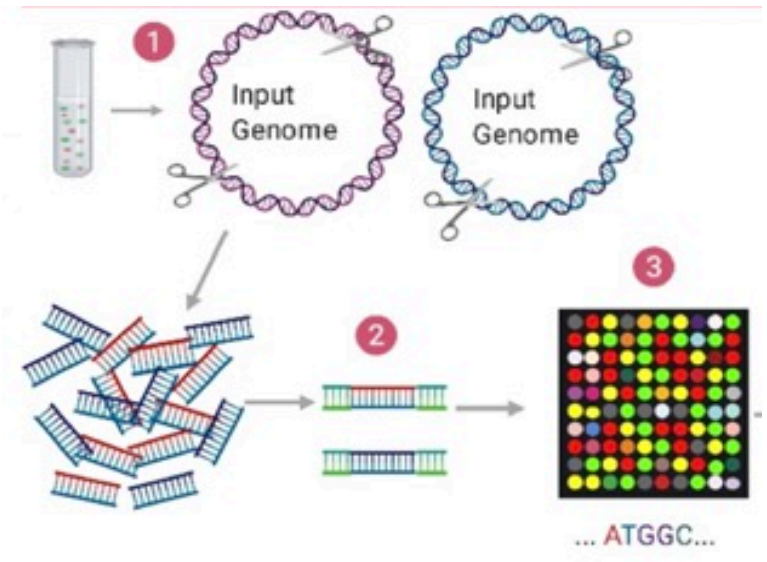
- Describe metagenomics and metabolomics technologies and data processing
- Describe common challenges and solutions in multi-omics data analysis
- Formulate and test different types of hypotheses with omics data
  - Integrative pathway/reaction analysis
  - Predictive analysis with machine learning
- Practice wrangling and plotting data in R

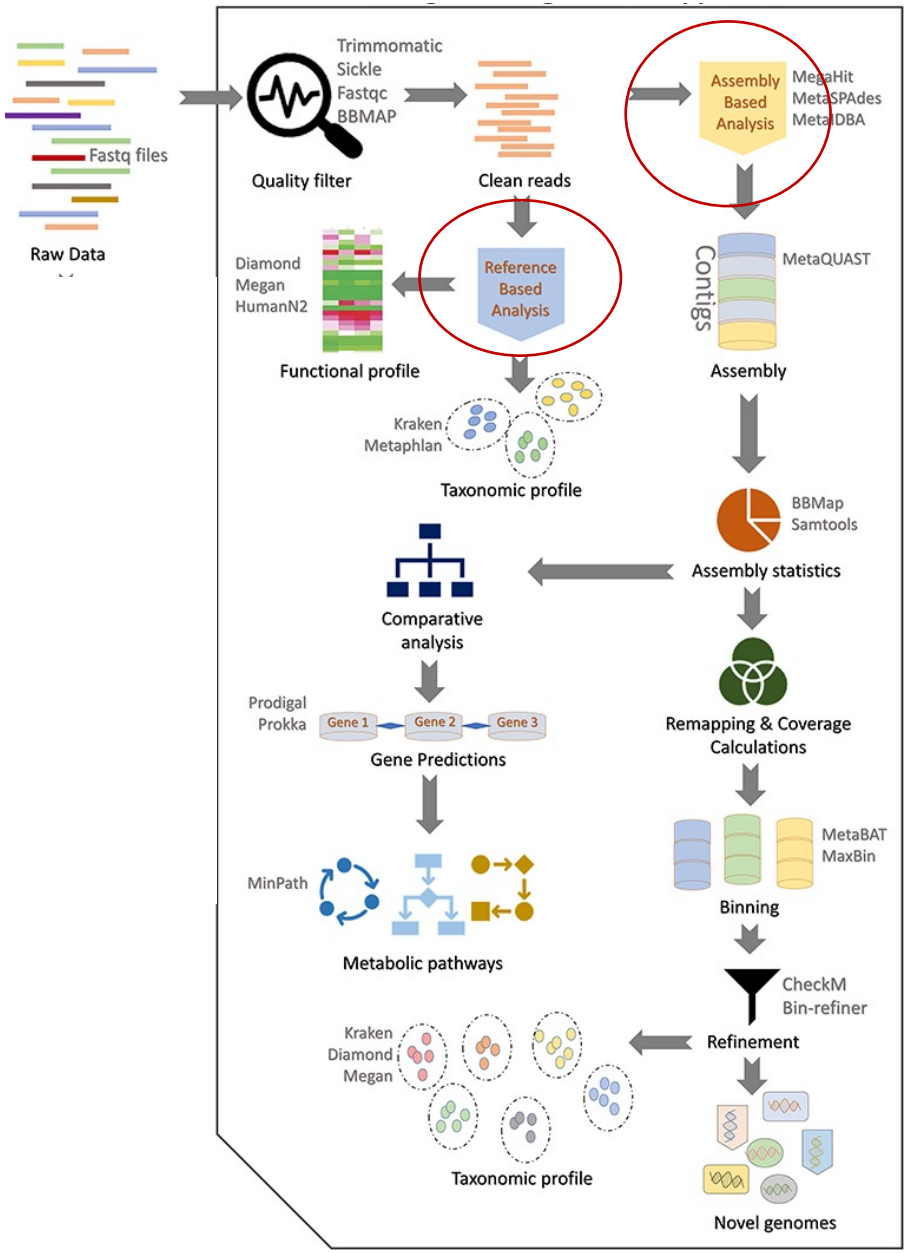
"Multi-omics"



Procter et al *Nature* 2019

Generating and processing metagenomic data





Bharti & Grimm
Briefings in Bioinformatics 2019

Generating and processing metabolomics data

Most common: LC-MS/MS

Samples

Prep/extract metabolites

Separate compounds

Mass spectrometer

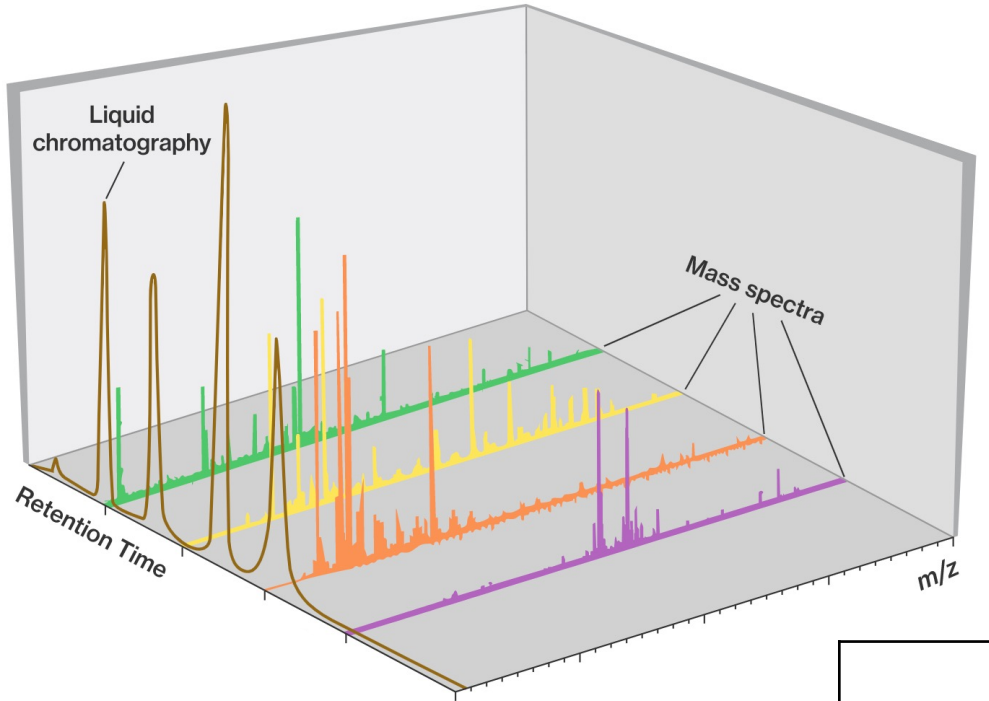
Ionize

Separate by m/z

Fragment

Separate by m/z

Detect masses



Group similar features across samples

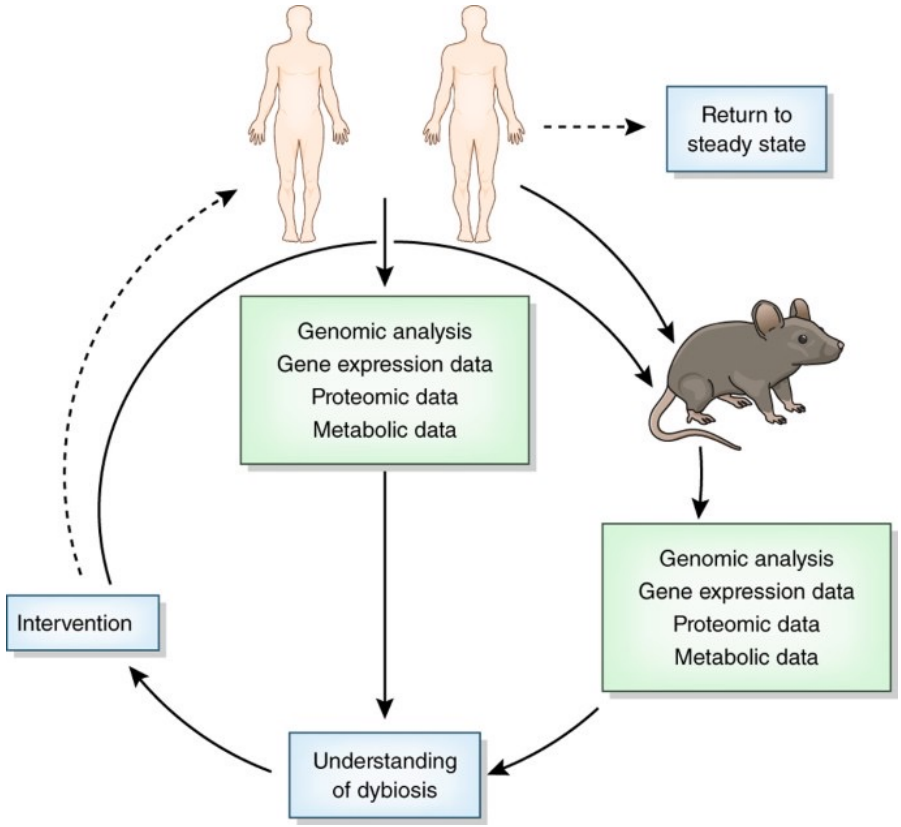
Search against a database of known metabolites

	sample1	sample2	sample3
proline			
lactate			
unknown1			
unknown2			
…			

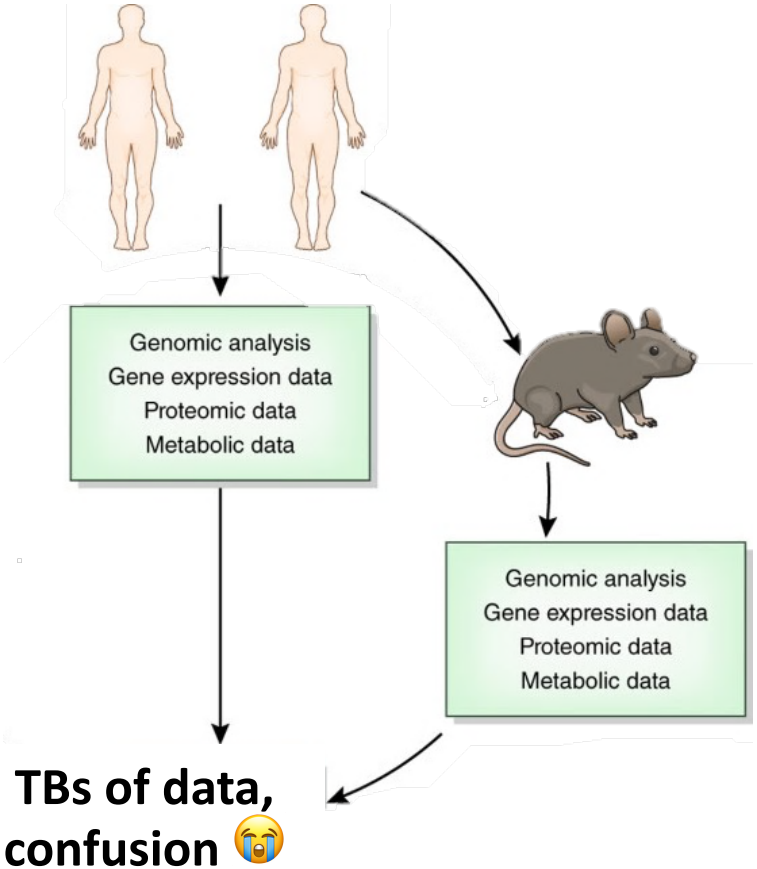
Image: CourseHero

Multi-omic analysis is not a solved problem

Expectation



Reality



TBs of data, confusion 😭

Gilbert et al *Nature* 2018

Challenge #1 in multi-omics analysis: Define your questions

- What general patterns can be observed?
  - Ordination plots, association analyses
- What features change as a result of an intervention?
  - Regression, time series/change point analysis
- Can we identify biomarkers and/or make predictions about clinical features?
  - Predictive models/machine learning
- Can subjects be grouped into categories based on their molecular profiles?
  - Clustering
- Is there evidence of particular molecular mechanisms? (e.g. are certain microbes producing or consuming a metabolite?)
  - Mechanistic models

Also important for study design!!

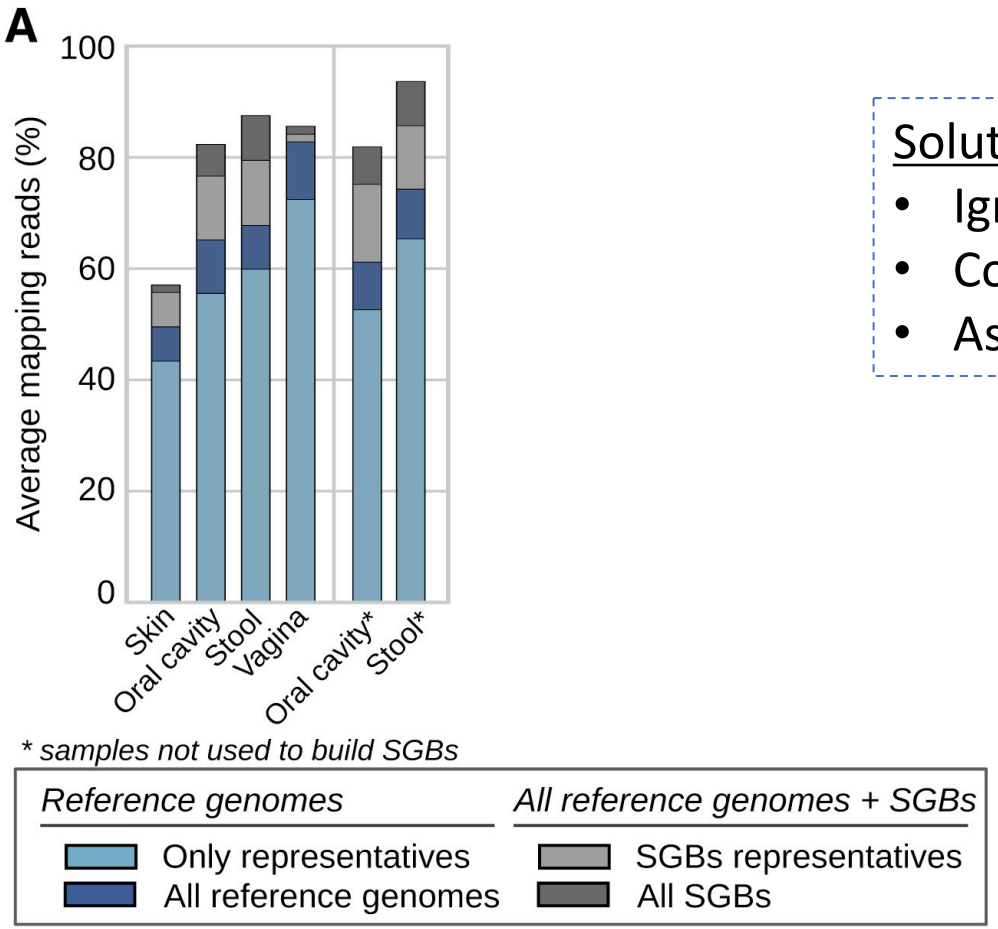
Some other computational challenges…

- Reference database completeness
- Quantification and compositionality
- Dimensionality ($n << p$)

Reference database incompleteness

Metagenomics

Alignment to known genomes:



Solutions:

- Ignore unknowns
- Continually expanding databases
- Assemble genomes, analyze those

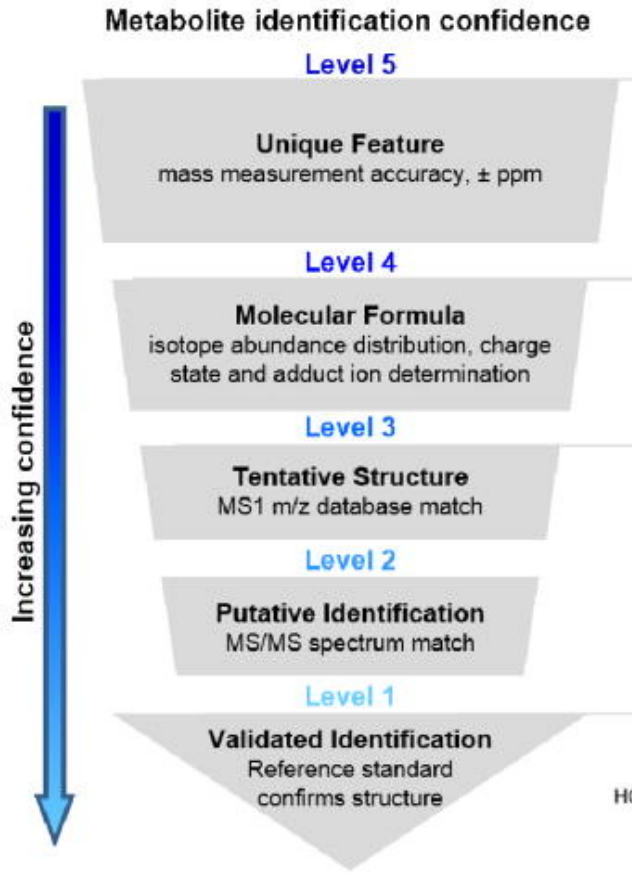
Pasolli 2019

Reference database incompleteness

Metabolomics

Database identification of features:

0.5-5% of LC-MS features typically match a known database standard at Levels 1-3



Solutions:

- Ignore unknowns
- Continually expanding databases
- Cheminformatics methods to group related features and learn/predict identities of unknowns (Molecular networking, CSI-FingerID, Mummichog)

Schrimpe-Rutledge *J ASMS* 2016

Quantification: What do our data values mean?

Metagenomics

Read mapping counts are influenced by:
1) True abundance in sample
2) Library size (sequencing depth)
3) Gene length/genome size
4) Mappability
5) Compositionality

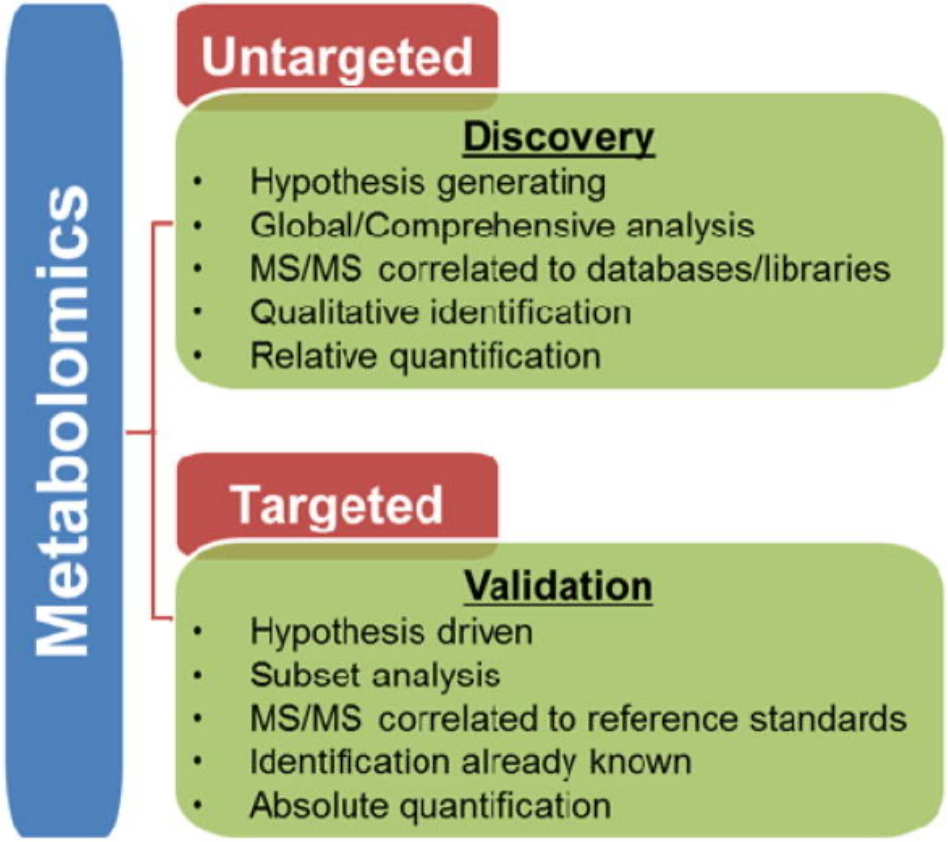
→Normalization is important!

Normalization by estimated genome equivalents:

0.8 copies/genome

- MicrobeCensus
- MUSiCC

Quantification: What do our data values mean?



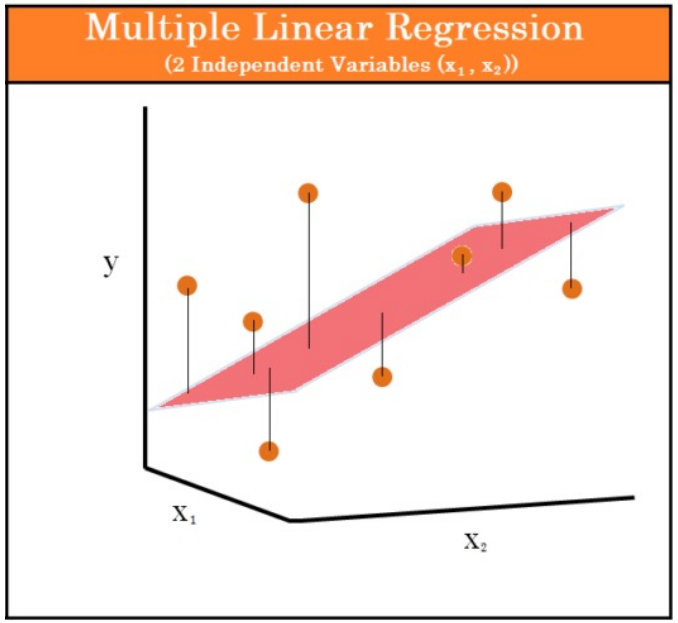
Schrimpe-Rutledge *J ASMS* 2016

Dimensionality (the *n* << *p* problem)

n samples

	met1	met2	met3	taxon1
sample 1				
sample 2				
sample 3				

p features



Standard statistical analysis: Model the outcome as a function of our features/predictors

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i.$$

Disease severity = baseline + metabolite 1 effect + metabolite2 effect + taxon1 effect + … + residual error

Too many parameters → many possible models!

Solutions:

- Methods that define other constraints to choose the "best" model (regularization, Lasso regression)
- Ask different questions (aware of multiple hypothesis issues) (differential abundance, clustering)
- Reduce *p* with dimensional reduction tools (PCA)
- Reduce *p* by making use of relevant biological prior knowledge (pathway analysis, mechanistic modeling)

Today's dataset

	Non-IBD control	CD	UC
Discovery cohort (PRISM)	34	68	53
Validation cohort (LifeLines DEEP and NLIBD)	22	20	23

Multi-omic screening of stool samples

Metagenomic shotgun sequencing: Microbial taxa, genes and pathways

+

Four LC–MS metabolomics methods: Lipids, polar metabolites, free fatty acids and bile acids

Data subset:

- Cincinnati study site: 31 subjects, 160 samples
- MGH study site: 28 subjects, 143 samples

2 vignettes:

- Can IBD-associated metabolite shifts be explained by relevant microbial taxa and genes?
- How well do the microbiome & metabolome predict IBD status?

Franzosa *Nature Microbiology* 2019

Some additional resources

- Microbiome omics analysis:
  - https://www.sciencedirect.com/science/article/pii/S193152441630127X
- biobakery omics tools:
  - https://elifesciences.org/articles/65088
  - https://github.com/biobakery/biobakery/wiki
- Metabolomics workflows:
  - https://www.sciencedirect.com/science/article/pii/S0003267018306354
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5110944/
- tidymodels machine learning (many tutorials and examples):
  - https://www.tidymodels.org
- Using mechanistic models to make sense of microbiome data:
  - https://www.nature.com/articles/s41564-019-0491-9