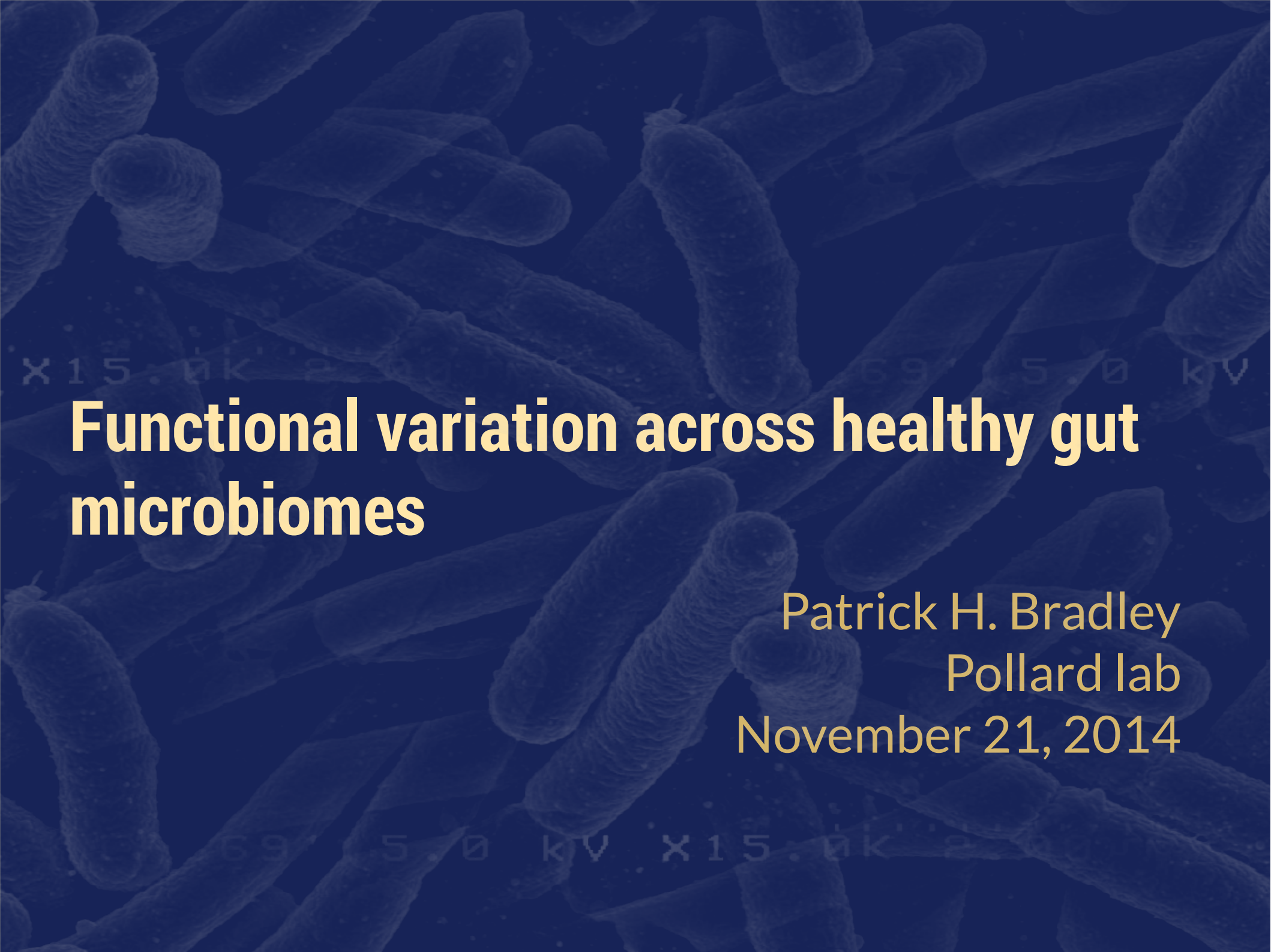
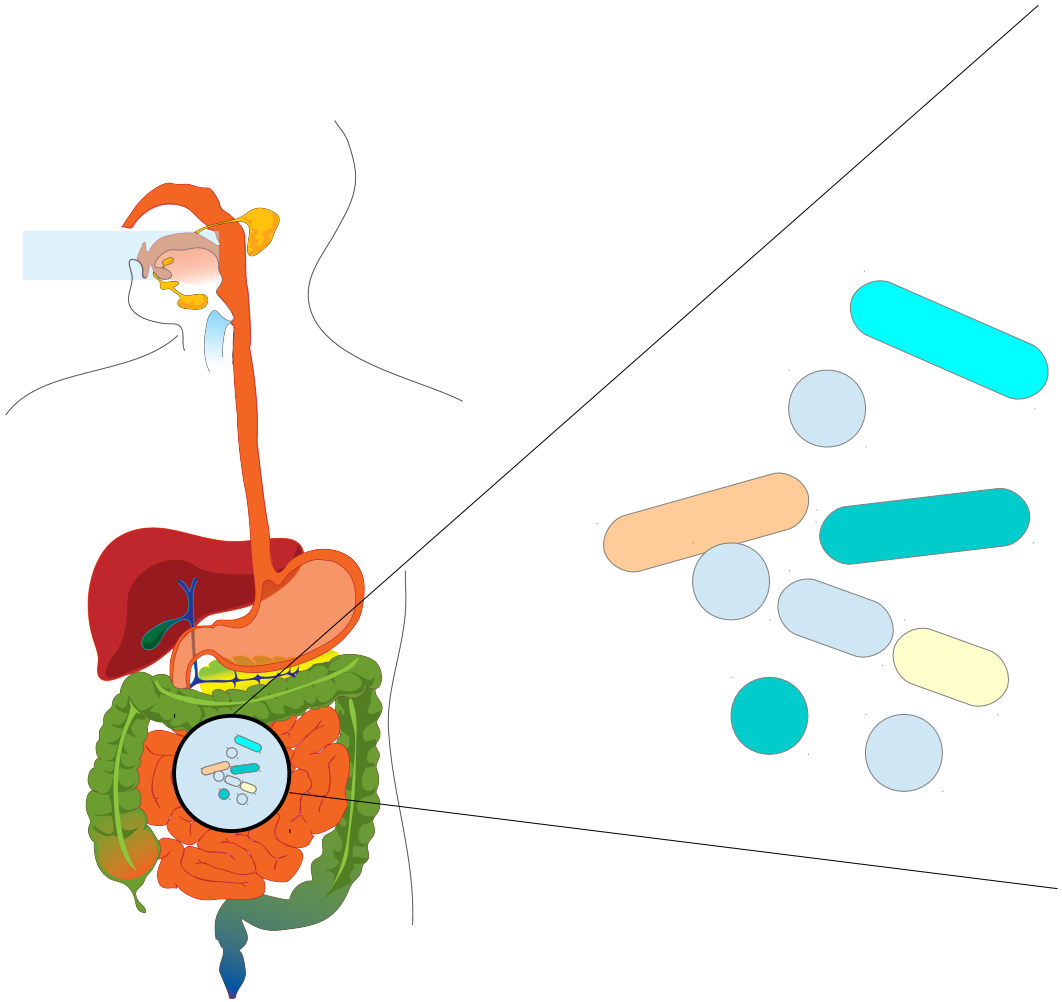


Functional variation across healthy gut microbiomes

Patrick H. Bradley
Pollard lab
November 21, 2014

How do bugs adapt to the gut?

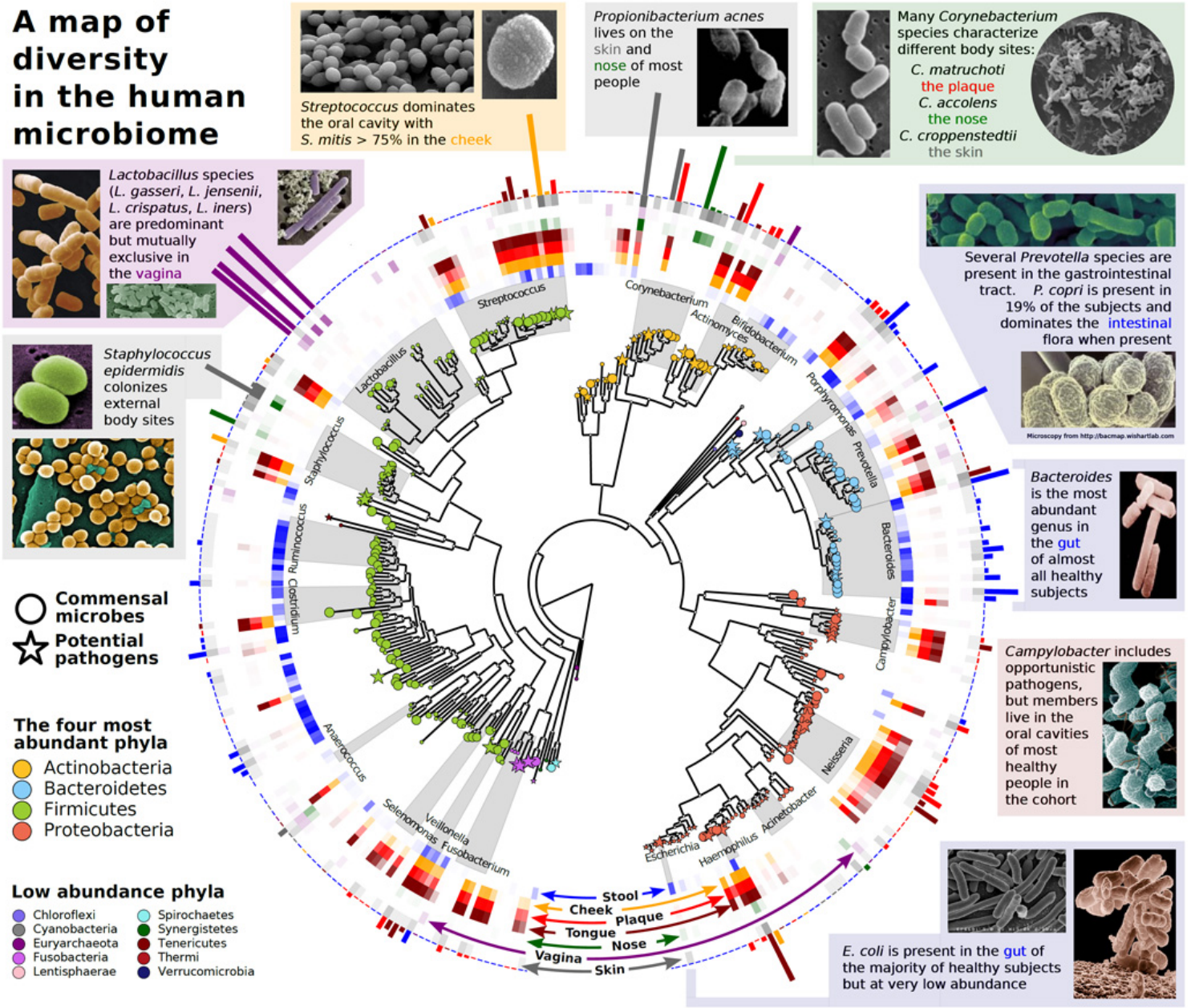


Are there some gene functions necessary for life in the human gut in general?

What about alternative strategies for colonizing the gut?

How do functions encoded by the microbiome vary across individuals?

Gut bacteria come from diverse taxa...

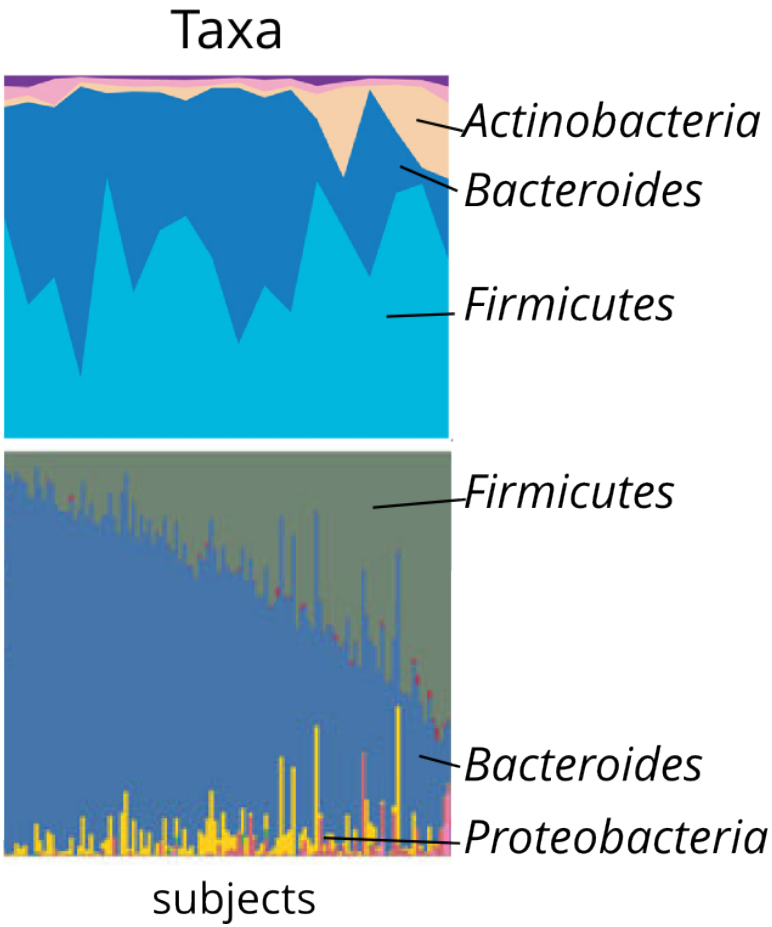


Morgan, Segata, Huttenhower, *Trends Genet* (2012)

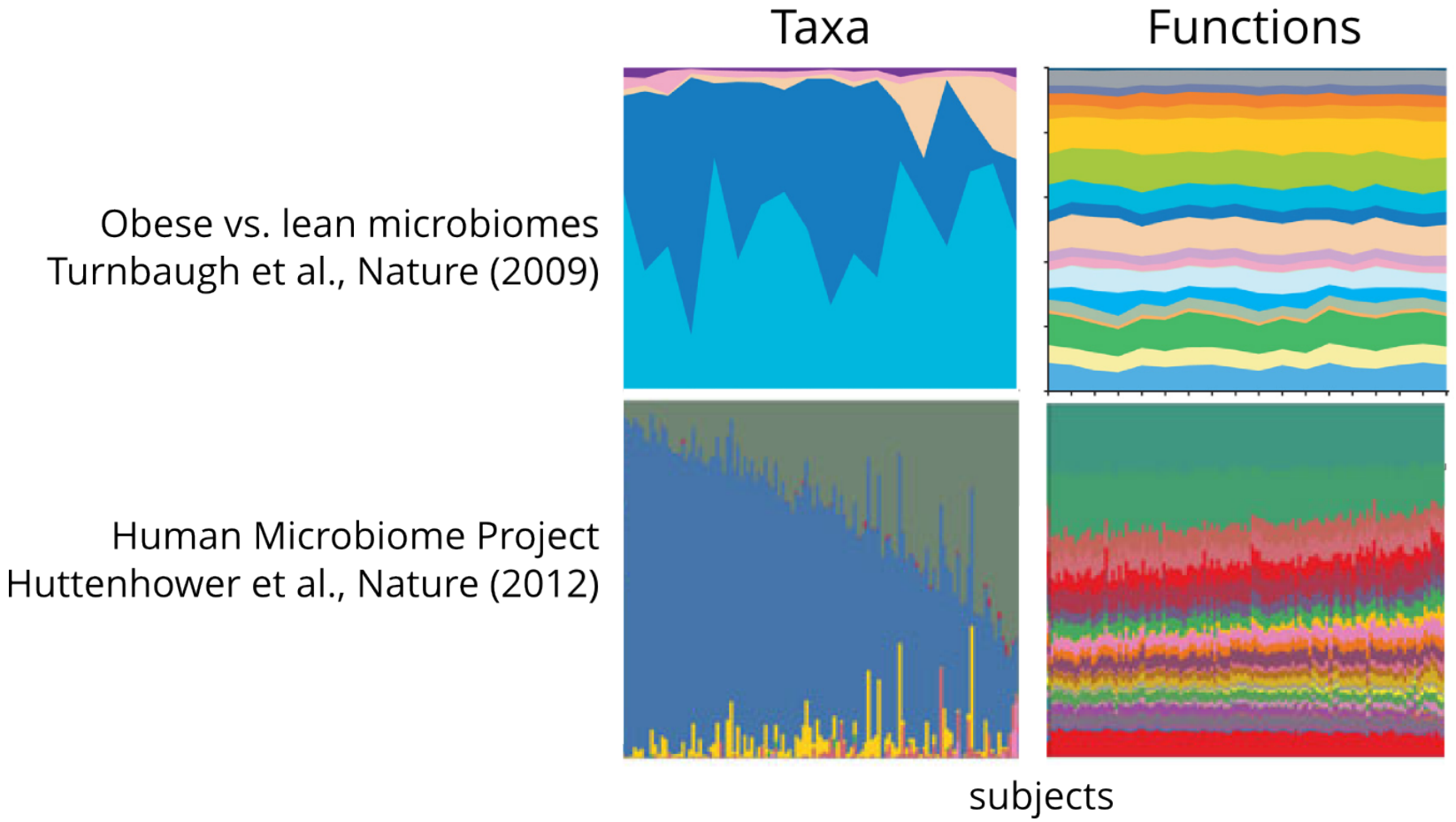
...and these taxa vary across individuals

Obese vs. lean microbiomes
Turnbaugh et al., Nature (2009)

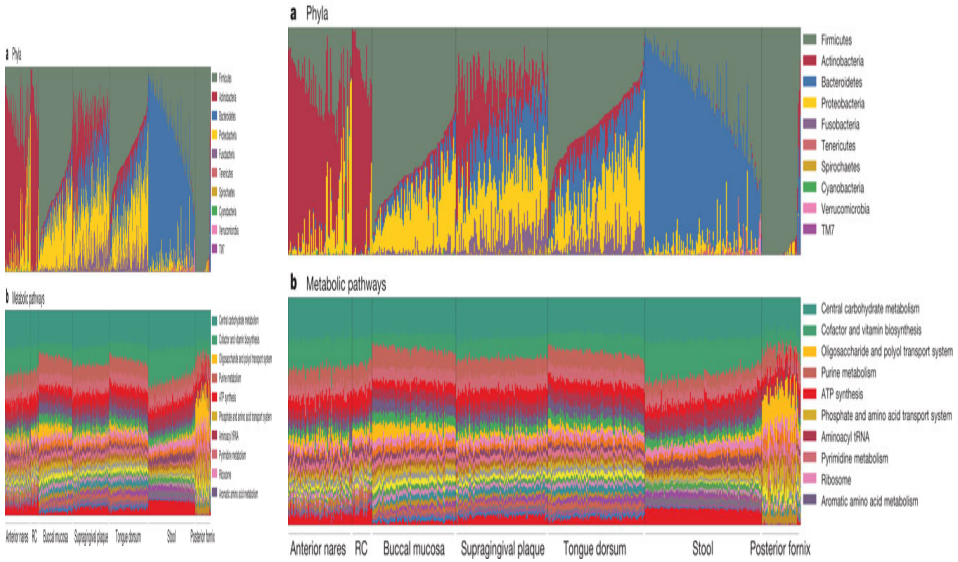
Human Microbiome Project
Huttenhower et al., Nature (2012)



Previous studies have found less variability at the level of gene function



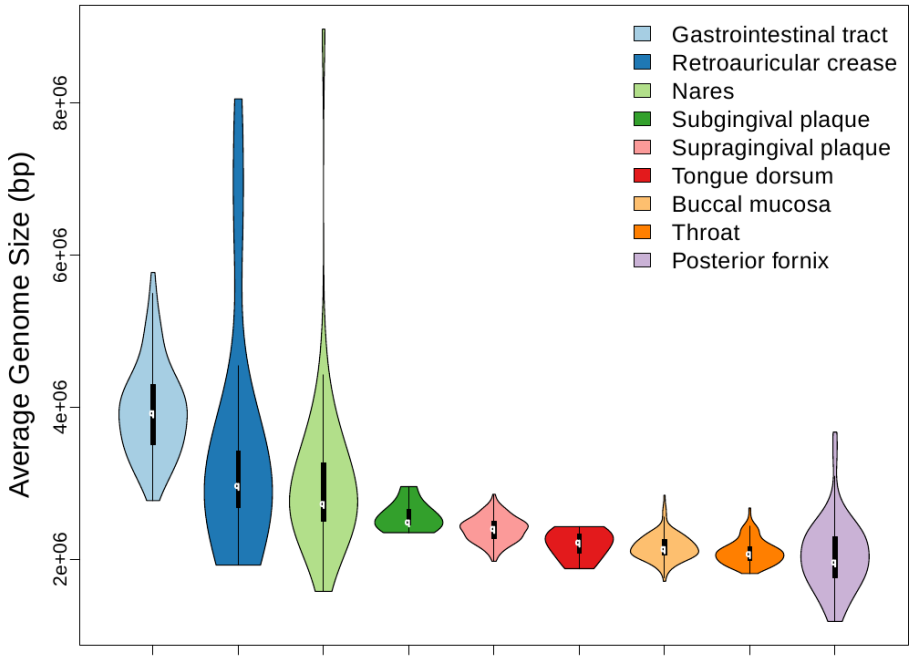
There are some limitations to these claims

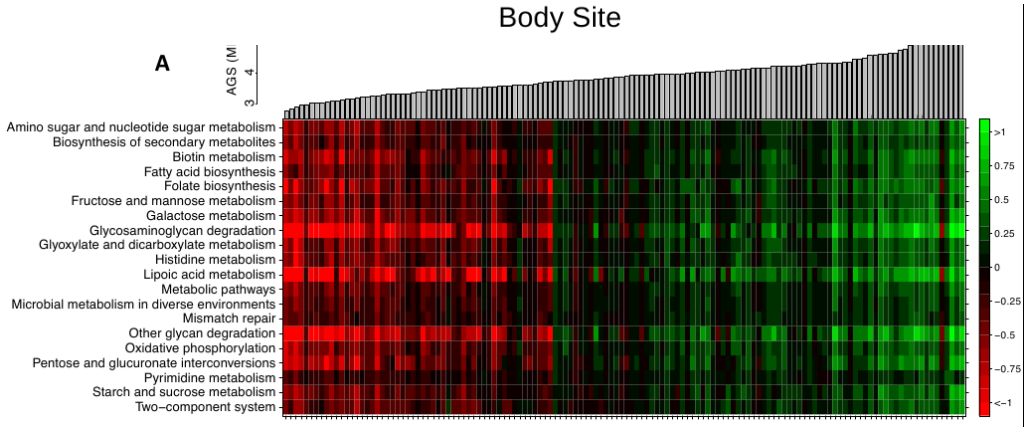


HMP Consortium, *Nature* (2012)

- Lots of "housekeeping," highly conserved genes and pathways
- Not corrected for average genome size
- Mean and variance correlated
- No statistical testing – is this what we expect?

There are some limitations to these claims

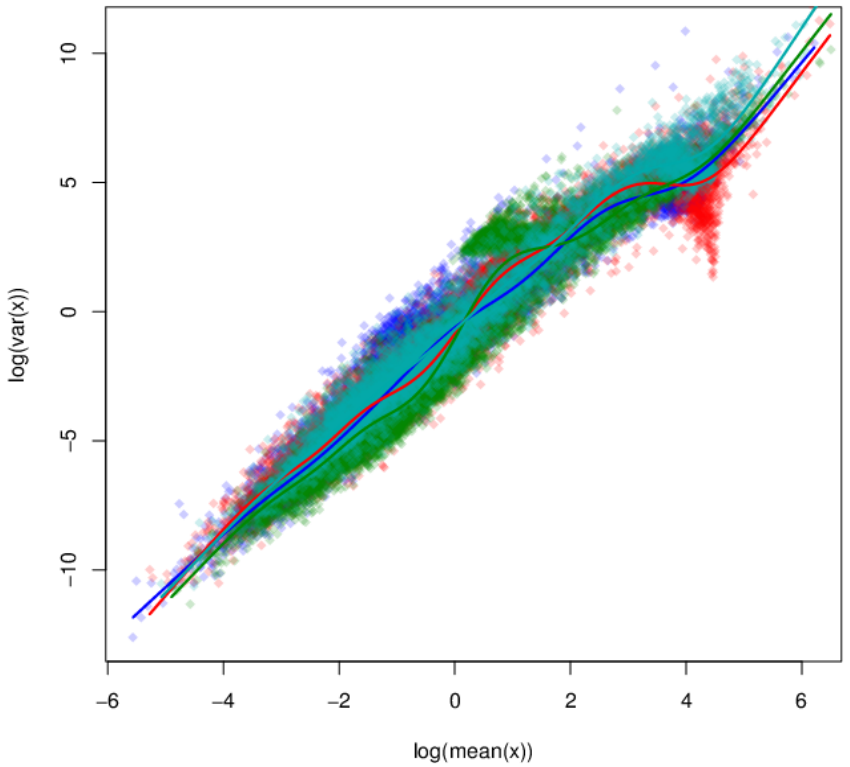




- Lots of “housekeeping,” highly conserved genes and pathways
- Not corrected for average genome size
- Mean and variance correlated
- No statistical testing – is this what we expect?

Nayfach & Pollard, in review

There are some limitations to these claims



- Lots of “housekeeping,” highly conserved genes and pathways
- Not corrected for average genome size
- Mean and variance correlated
- No statistical testing – is this what we expect?

How can we formalize this idea?

- Typically, one would use the **mean** as a test statistic (when, e.g., comparing two groups)
- Here, we use the **variance** as our test statistic

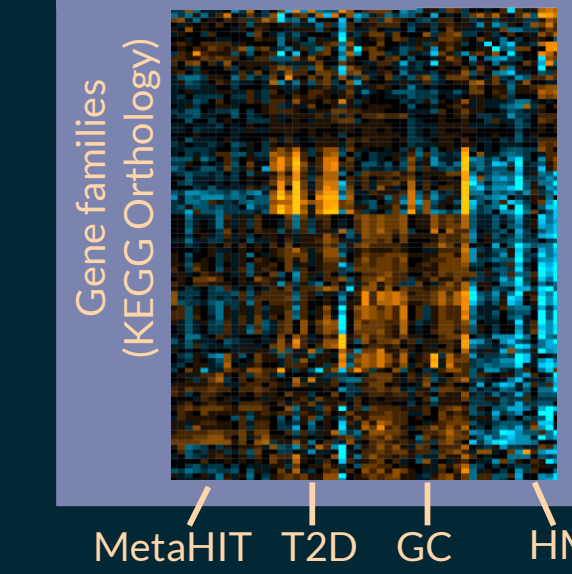
What are we using for data?

Healthy individuals and controls from case-control studies

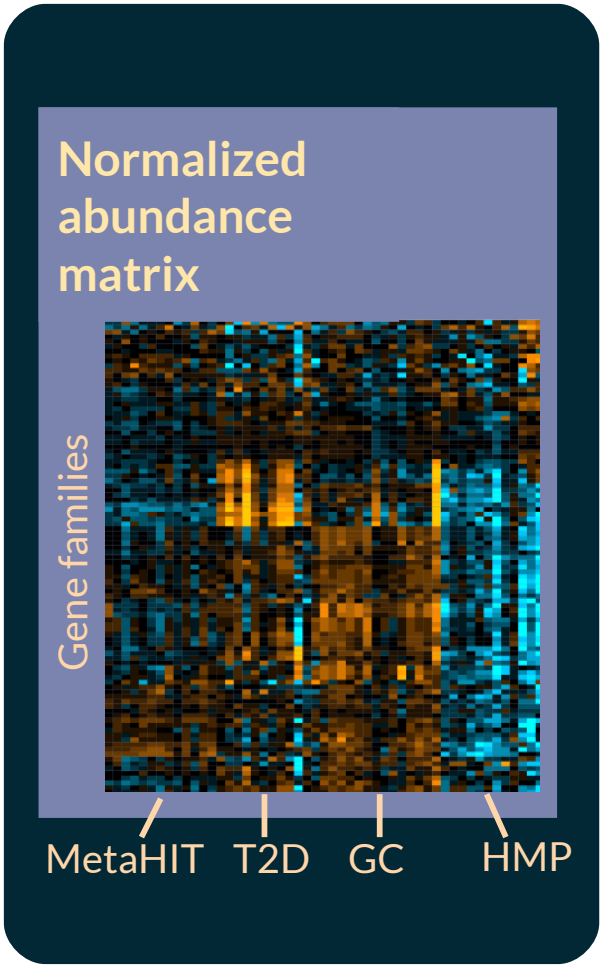
- **MetaHIT** (Spanish cohort, non-IBD, *n* = 14)
- **HMP** (American cohort, all healthy, *n* = 13)
- **Glucose Control** (**GC**: European, controls, *n* = 14)
- **Type II Diabetes** (**T2D**: Chinese, normal glucose tolerance, *n* = 12)

Process with Shotmap, MicrobeCensus

Normalized abundance matrix



How do we model this data?

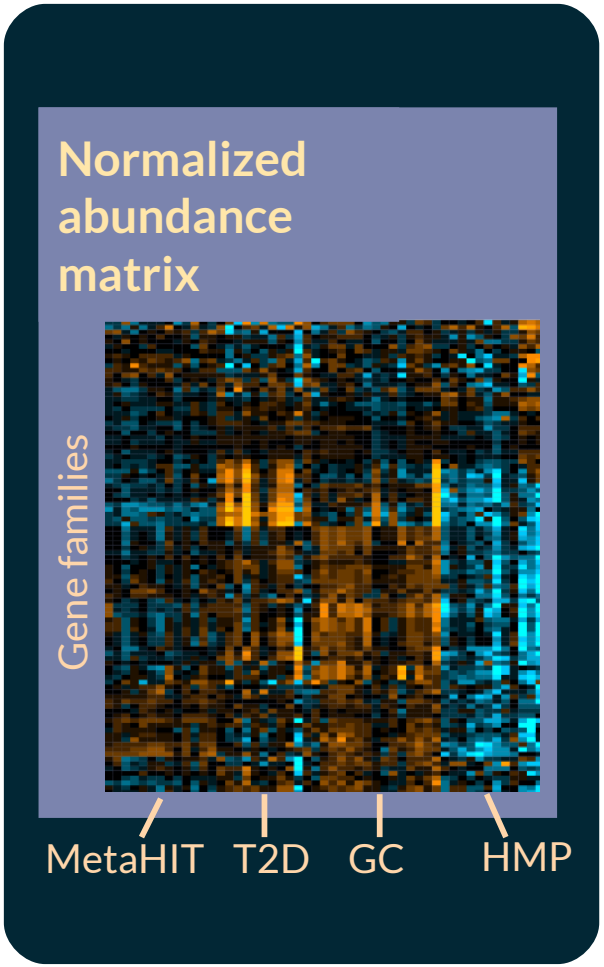


Each gene family is modeled as a function of the dataset:

$$\log(\mathrm{Y}_{g,d}) = \beta_{g,0} + \sum_{d \in D} I(d)\beta_{g,d} + \epsilon$$

- $\log(\mathrm{Y}_{g,d})$: Normalized abundance of gene *g* in dataset *d*
- $\beta_{g,0}$: Overall "mean effect" for gene *g*
- $I(d)$: Indicator for dataset *d*
- $\beta_{g,d}$: Mean "offset" for gene *g* in dataset *d*
- ϵ: Residual variation

How do we model this data?



Each gene family is modeled as a function of the dataset:

$$\log(\mathrm{Y}_{g,d}) = \beta_{g,0} + \sum_{d \in D} I(d)\beta_{g,d} + \epsilon$$

- $\log(\mathrm{Y}_{g,d})$: Normalized abundance of gene *g* in dataset *d*
- $\beta_{g,0}$: Overall "mean effect" for gene *g*
- $I(d)$: Indicator for dataset *d*
- $\beta_{g,d}$: Mean "offset" for gene *g* in dataset *d*
- ϵ: Residual variation

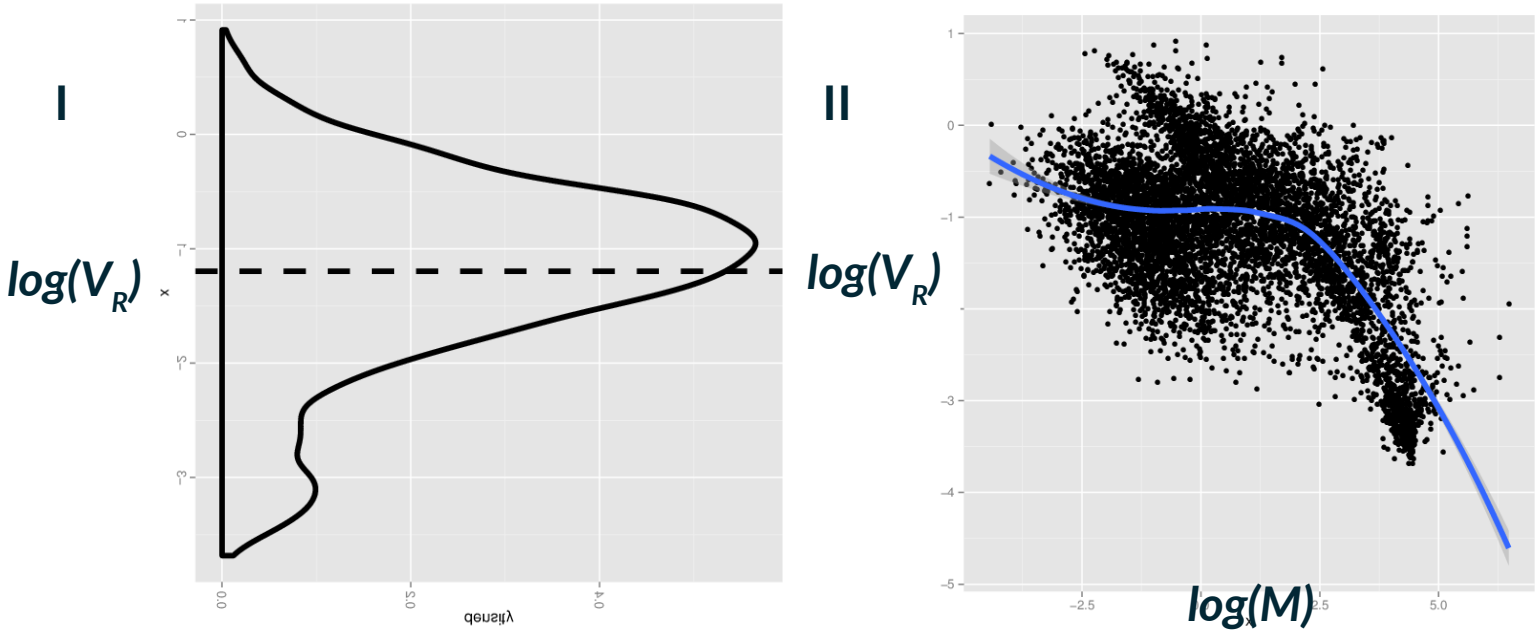
Test statistic $V_R = \dfrac{\sum \epsilon^2}{(n-1)}$

What is our null hypothesis?

Two potential options:

I. The residual variance V_R is equal to its expected value:
$V_R = E(V_R)$

II. The residual variance V_R is equal to its expected value *given the mean abundance of that gene*: $V_R = E(V_R \mid M)$



We can get a background distribution by simulating data from the model we fit earlier

$$\log(\mathrm{Y}_{g,d}) = \underbrace{\beta_{g,0} + \sum_{d \in D} I(d)\beta_{g,d}} + \epsilon$$

Fit from original data

Drawn from a normal distribution N(0, σ)

We can get a background distribution by simulating data from the model we fit earlier

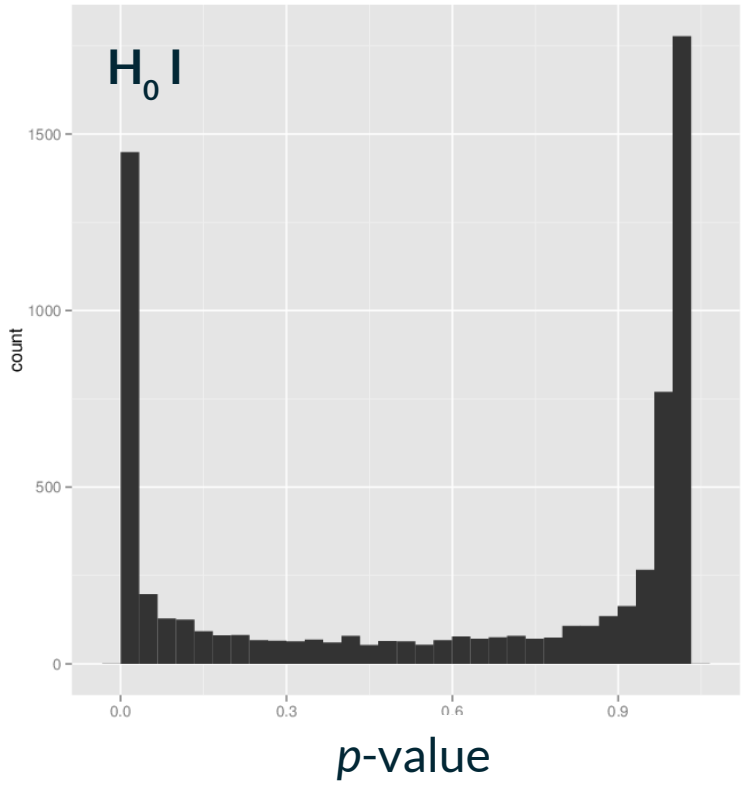
$$\log(\mathrm{Y}_{g,d}) = \underbrace{\beta_{g,0} + \sum_{d \in D} I(d)\beta_{g,d}}_{\text{Fit from original data}} + \underbrace{\epsilon}_{\text{Drawn from a normal distribution } N(0, \sigma)}$$

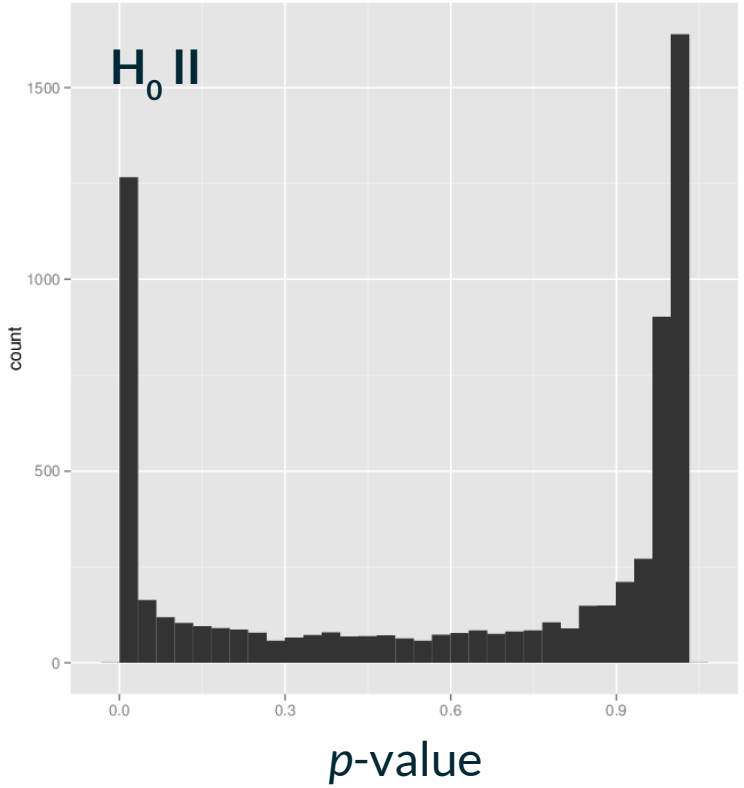
How we choose the standard deviation σ depends on the null hypothesis:

H_0 I: $\sigma \sim N(\sqrt{E(V_R)}, \hat{se})$

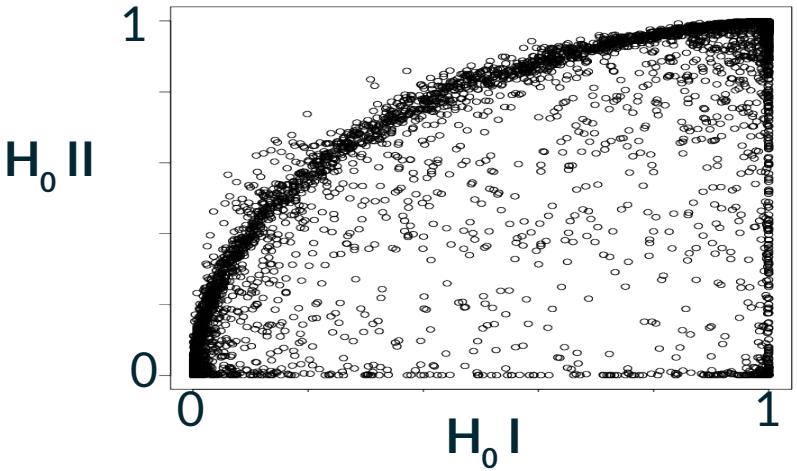
H_0 II: $\sigma \sim N(\sqrt{E(V_R|M)}, \hat{se})$

Both null hypotheses give related results



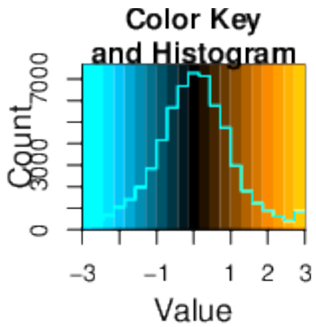


Agreement between models:



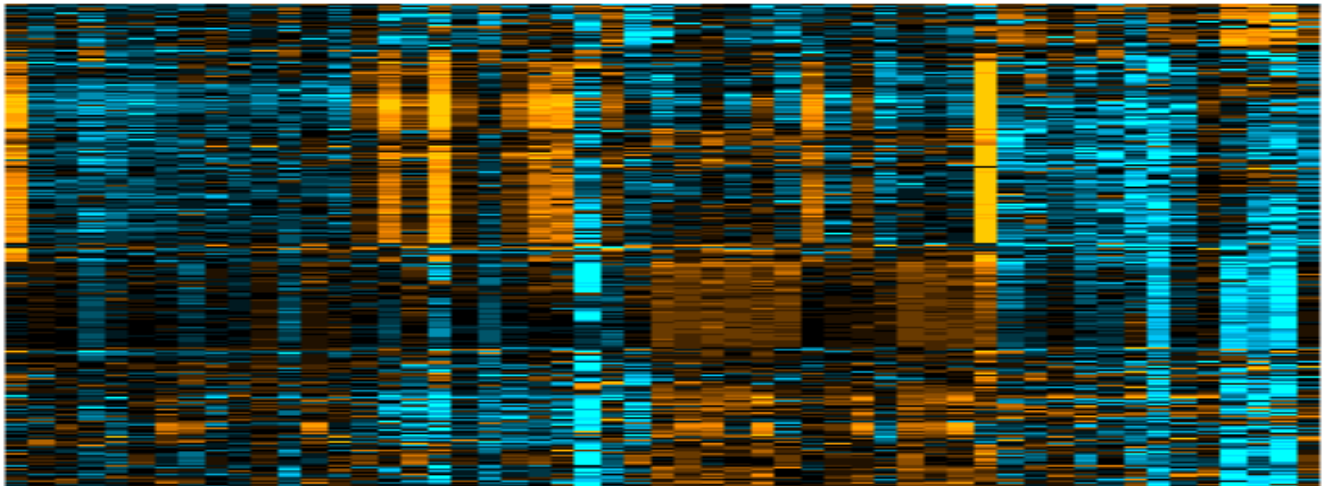
10,000 ft. view of significant results

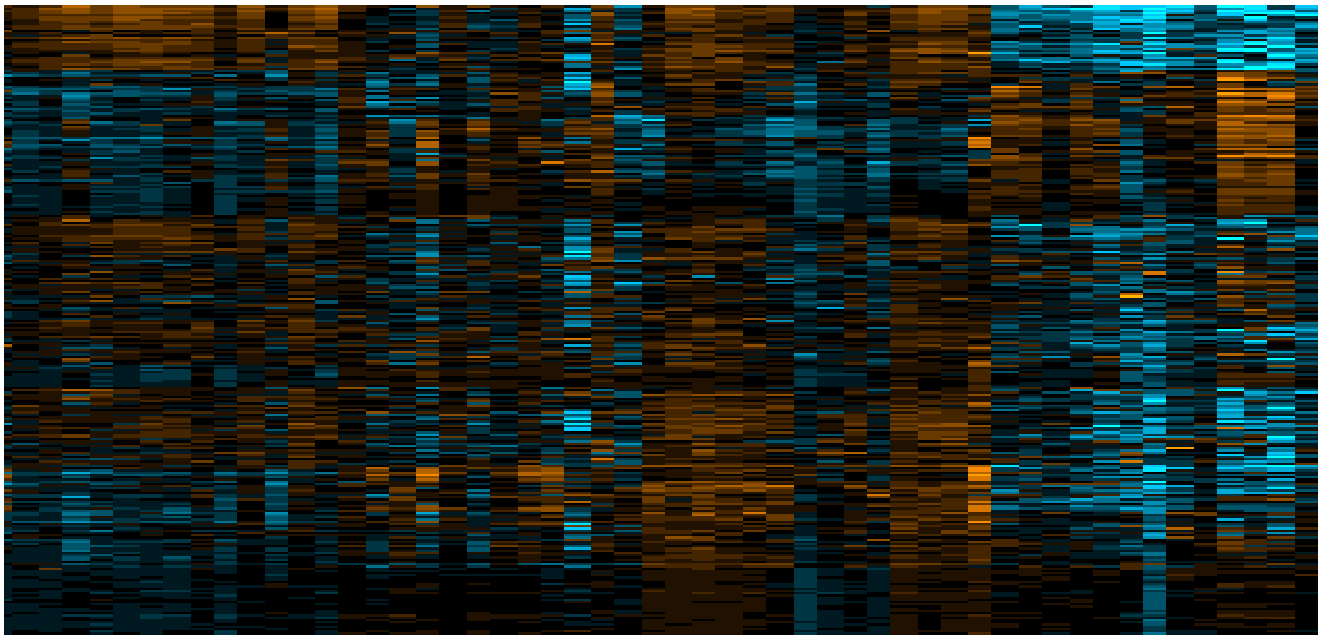
H_0 II: Significantly variable



H_0 II: Significantly invariable

Note: dataset effects and overall mean subtracted out





H_0 I: significantly variable enrichments

- **Transport**
  - Glu/Asp transport system
  - PTS system, N-Ac-galactosamine-specific II component
- **Nitrate metabolism**
  - Denitrification, nitrate => nitrogen
  - Dissimilatory nitrate reduction, nitrate => ammonia

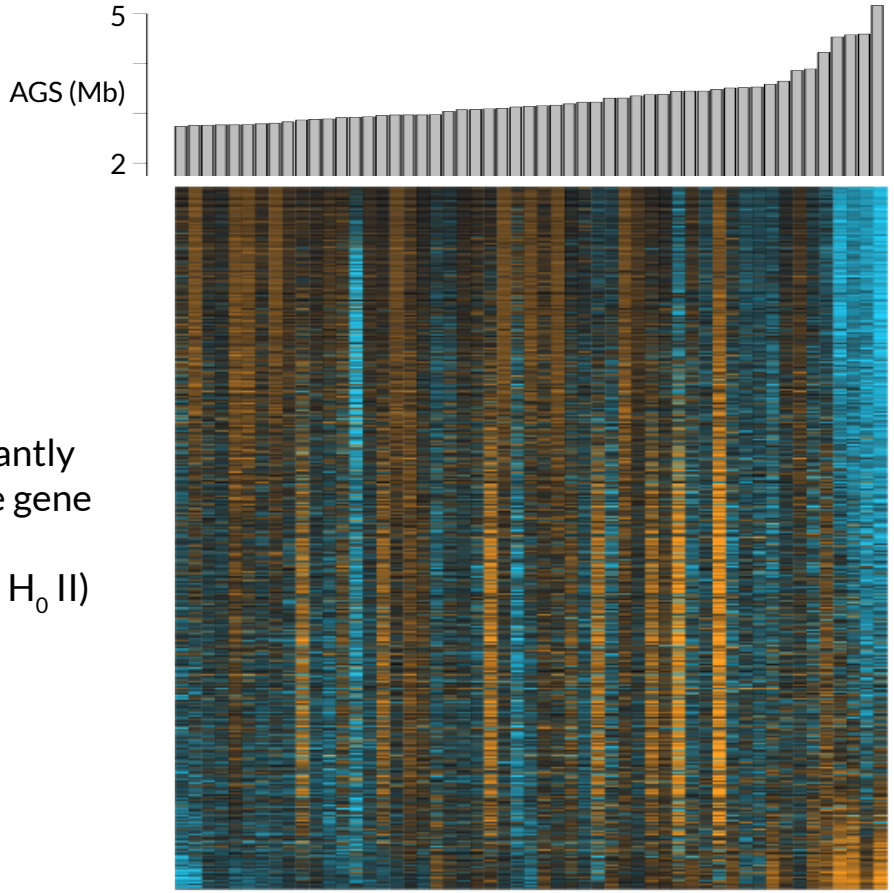
$q < .25$

H_0 II: significantly variable enrichments

- **Sugar transport**
  - ABC transport systems for:
    - Maltose/maltodextrin
    - Raffinose/stachyose/melibiose
    - L-arabinose/lactose
    - N-acetylglucosamine
    - Methyl-galactoside
    - Arabinogalactan oligomer/maltooligosaccharides
  - PTS systems: arbutin-like, mannose-specific, NAGal-specific
- **Oligopeptide transport system**
- **Nitrate metabolism**
  - Denitrification, dissimilatory nitrate reduction

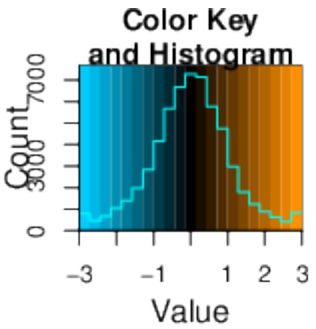
$q < .25$

Most variable genes correlate with AGS

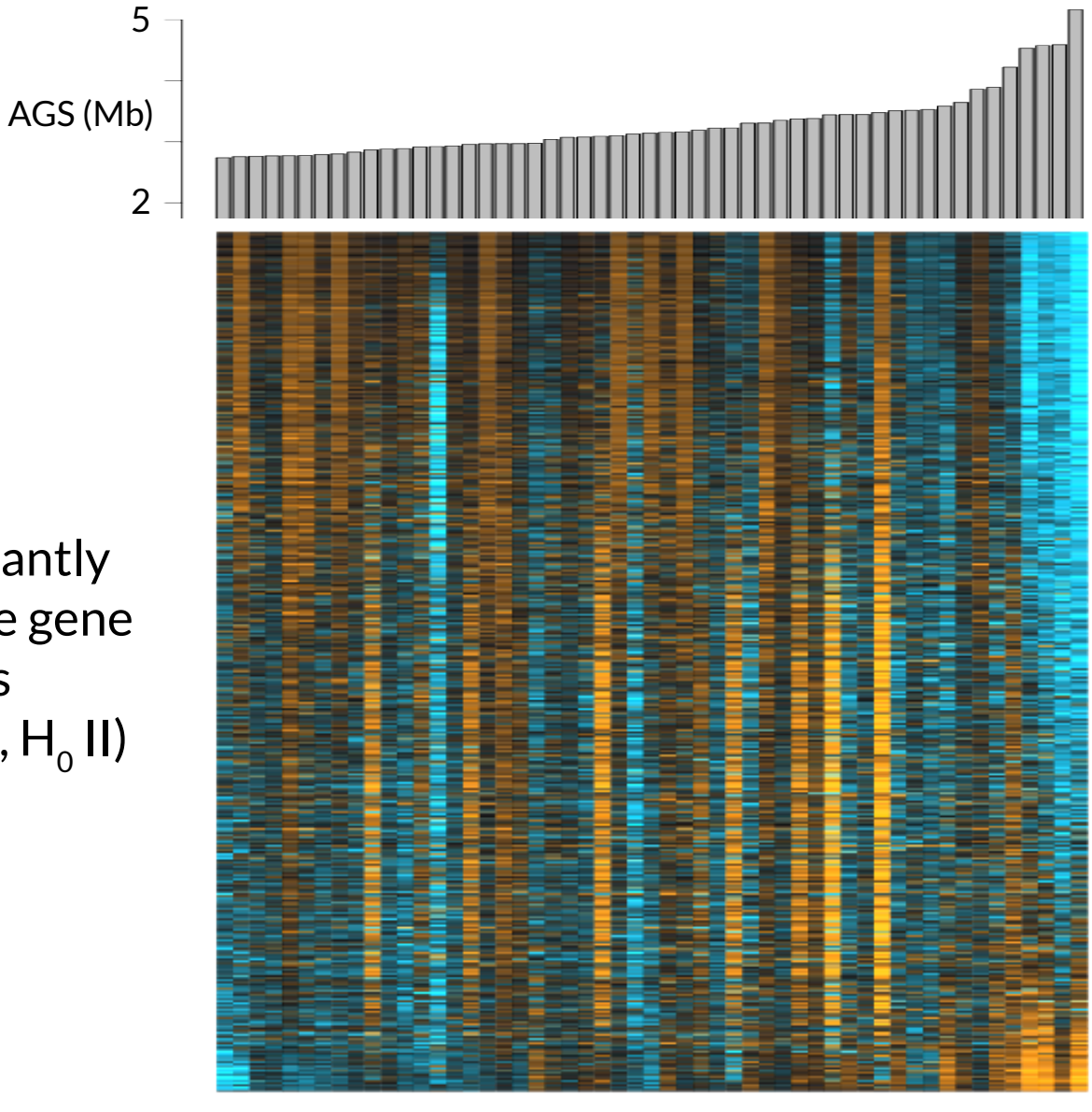


Significantly variable gene families ($q \leq 0.1$, H_0 II)

Metagenomes

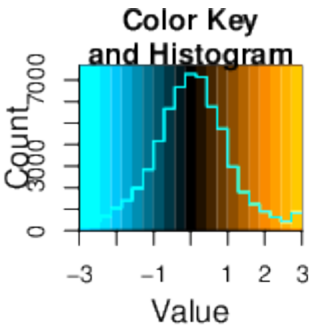


Most variable genes correlate with AGS



Significantly variable gene families ($q \leq 0.1$, H_0 II)

Metagenomes



H_0 I: significantly invariable enrichments

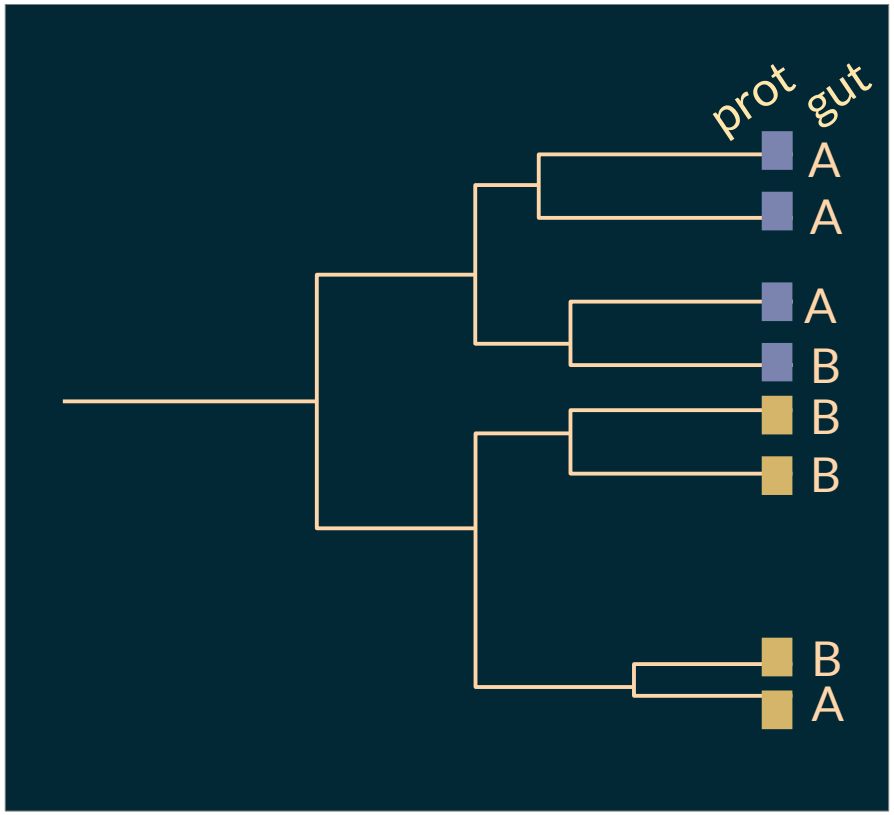
- Glycolysis, pentose-phosphate pathway
- Amino acid biosynthesis
  - Lys, Val, Ile biosynthesis
  - Ile biosynthesis from Thr
- Nucleotide sugar biosynthesis
  - IMP => ADP, GDP
- **C5 isoprenoid biosynthesis, non-mevalonate**
- **Vitamin biosynthesis**
  - B12 biosynthesis
  - Thiamine biosynthesis
  - Ascorbate biosynthesis from glucose-1-P
- F-type ATPase
- Ribosome (bacteria, archaea)
- **Transport**
  - Branched-chain amino acid transport
  - Peptide/nickel transport
- **Sec (secretion) system**
- Aminoacyl-tRNA biosynthesis (euks and proks)
- **PleC-PleD (cell fate control) 2-component regulatory system**
- **Trehalose biosynthesis, D-glucose 1P => trehalose**
- Exon junction complex (EJC) *[eukaryotic, low abundance]*

$q < .25$

H_0 II: significantly invariable enrichments

- Ribosome (bacterial and archaeal)
- Aminoacyl-tRNA biosynthesis (euks and proks)
- **Vitamin B12 biosynthesis**
- **C5 isoprenoid biosynthesis, non-mevalonate**
- F-type ATPase
- Isoleucine biosynthesis from threonine
- **Sec (secretion) system**
- A few *(low-abundance)* eukaryotic-specific modules:
  - BRCA1-associated genome surveillance complex (BASC)
  - Exon junction complex (EJC)
  - Nuclear pore complex

Phylogenetic logistic regression



- Method from Ives & Garland, 2010
- Are two variables significantly associated, *taking into account tree structure?*

Procedure for PLR

- Build 16S tree
  - Used Clustal Omega to build initial alignments
  - Cleaned alignment (25%+ non-gaps, 25%+ ID letters)
  - Distances calculated and then averaged to yield distance matrix of organisms
  - Distance matrix assembled using NJ
- Used KEGG annotations for presence/absence of a gene family in a given organism
- Used Lozupone et al. annotations of gut vs. non-gut adapted organisms
- Tested for phylogenetically-surprising (PLR) and non-surprising (Wilcox test) associations
  - For PLR, can look at:
    - Presence/absence
    - Number of representatives per genome
- Set FDR at 10%, then look for significantly high overlap with (in)variable genes

Significant overlaps found between gut-associated and (in)variable genes

Gut-associated and/or gut-depleted

Variable/invariable	Presence/absence	Abundance	Wilcox
Variable (H_0 I: mean)	1	1	1
Variable (H_0 II: loess)	1	1	1
Invariable (H_0 I: mean)	**1.6E-9**	0.64	**0.0011**
Invariable (H_0 II: loess)	**2.2E-8**	0.83	0.60

NOTE: Different from previous findings
Here, restricting to gene families that were present in at least 3 organisms tested

$q < .1$

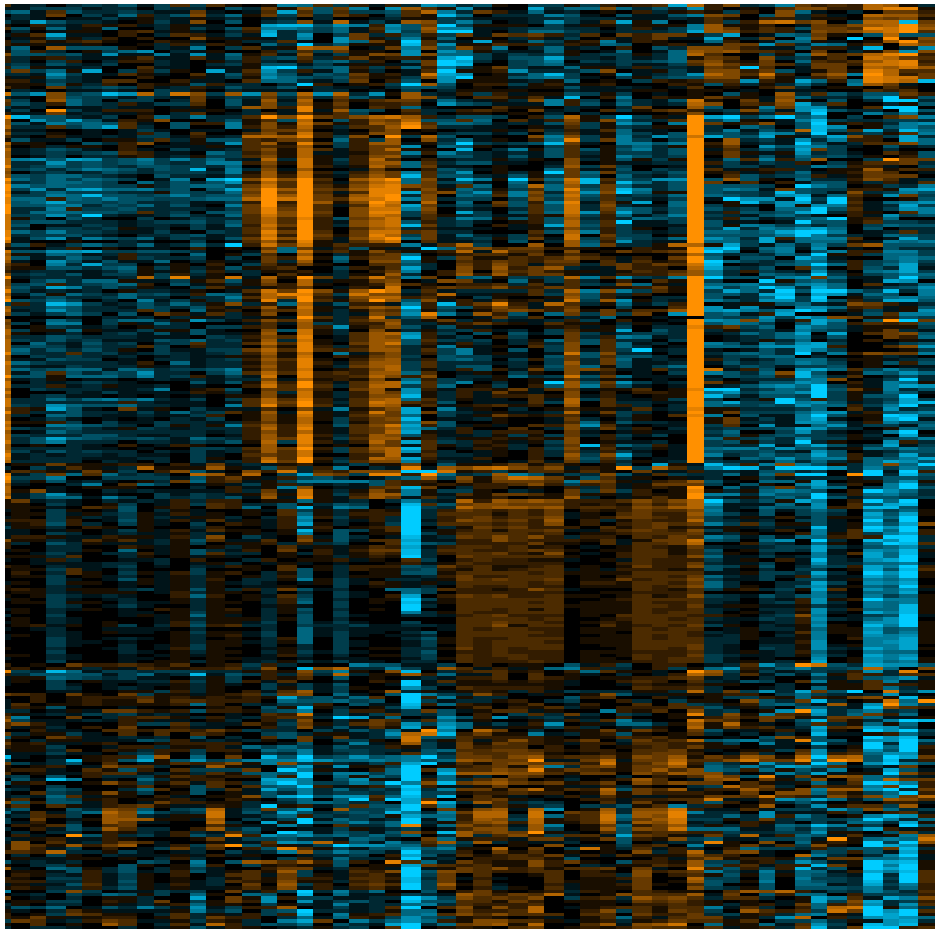
Issues with expanded 16S tree

Conclusions

- We have a method for identifying significantly variable and invariable gene families (and pathways)
- In a set of healthy human gut microbiota, variable genes/pathways tended to correlate with average genome size
- Invariable genes/pathways included not only housekeeping genes, but also vitamin biosynthesis, secretion, and isoprenoid biosynthesis
- Variable and invariable gene families may be linked to inheritance

Thanks!

- Katie Pollard
- Pollard lab
  - Stephen Nayfach
  - Stacia Wyman
  - Svetlana Lyalina
  - Josh Ladau
- Thomas Sharpton
- Wall and Hernandez labs
- Pathway meeting attendees



Lys/Arg/Orn transport
Glu/Asp transport
PTS system: NAGal, glucitol/sorbitol, galactosamine
Type 2 secretion system
Type 3 secretion system, EHEC/EPEC pathogenicity signature
Nitrate metabolism

Raffinose/stachyose/melibiose transport system
L-Arabinose/lactose transport system
N-Acetylglucosamine transport system
Methyl-galactoside transport system
Oligopeptide transport system
PTS system: Mannose
Bacterial proteasome
Arabinogalactan oligomer/maltooligosaccharide transport system
Methanogenesis
Multidrug/hemolysin transport system