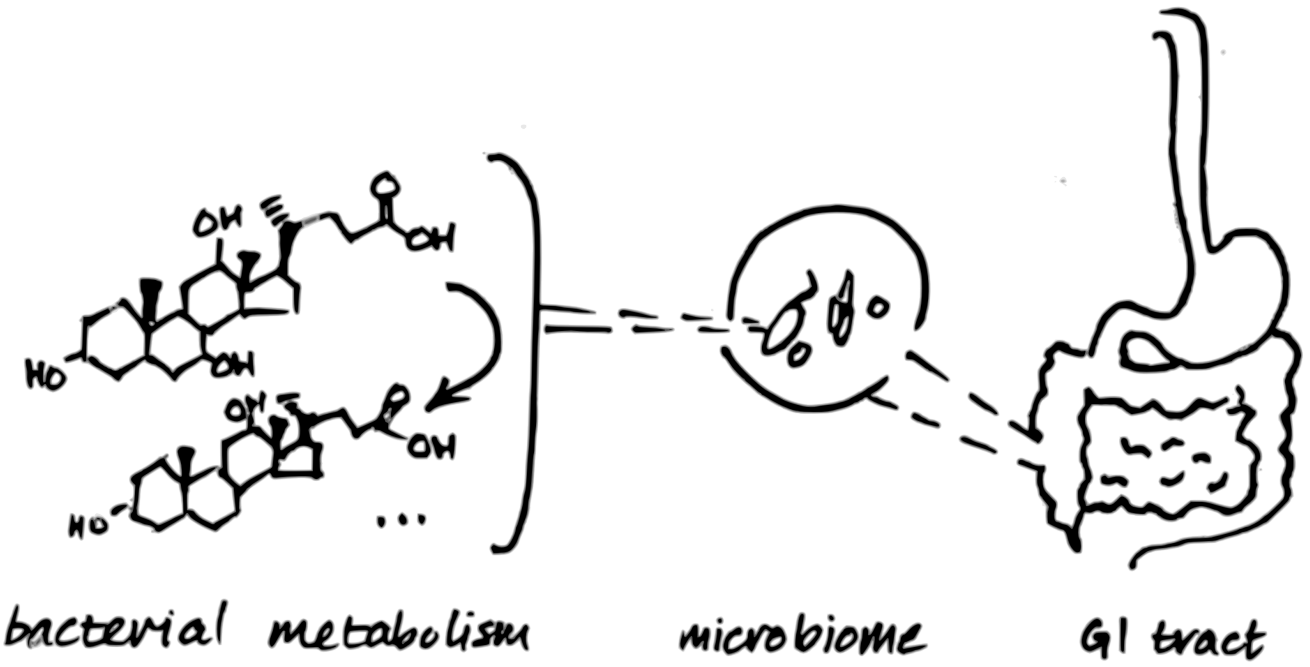
Assessing functional variability across metagenomes

Patrick H. Bradley
Pollard Lab
2015 Mar 30

The human gut microbiome encodes a wealth of gene families

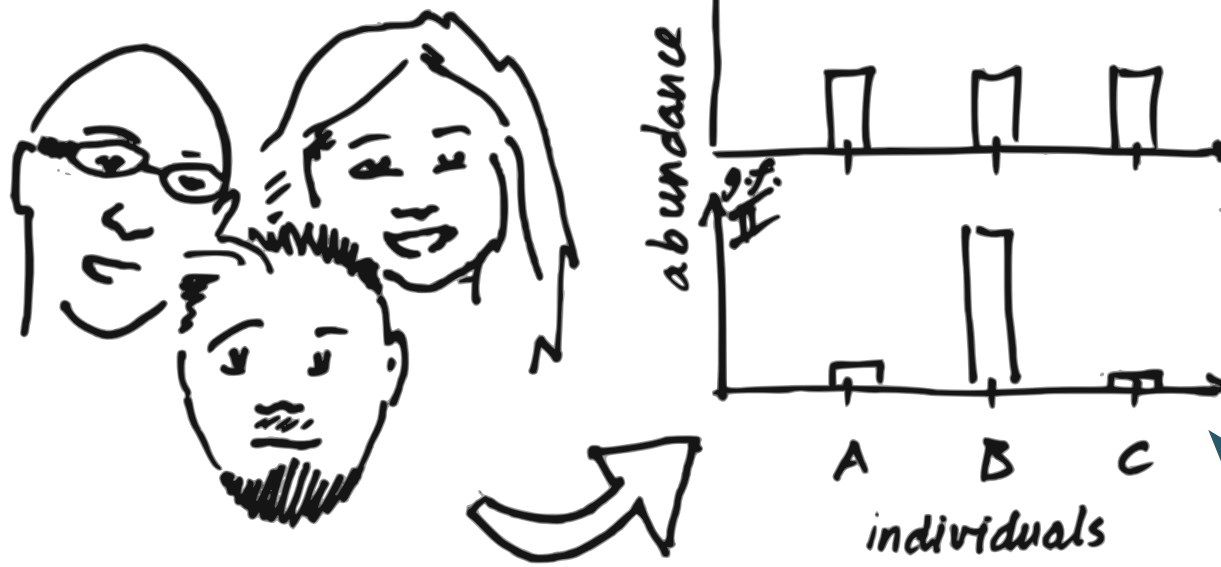


Which of these gene families allow particular bacteria to thrive in the gut, either

- **in general**, or
- in **different niches**?

The variance of gene families could help us understand selection in the gut

Looking across healthy individuals...



Invariable gene families could point to functions necessary for life in the gut in general

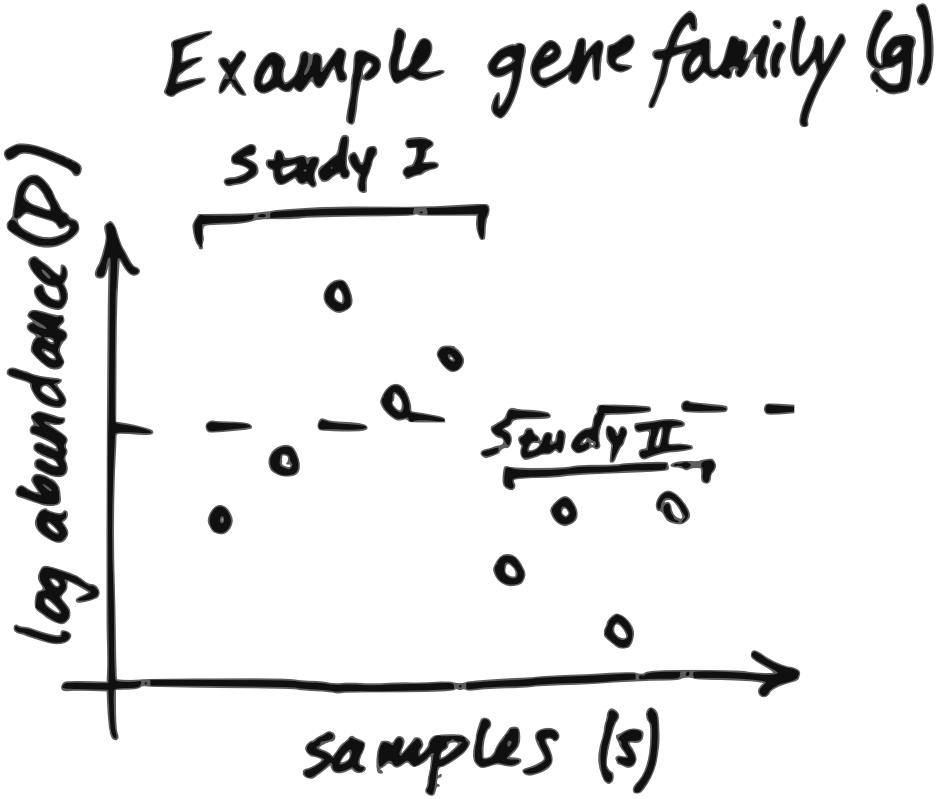
Variable gene families might be important in specific niches

Study design

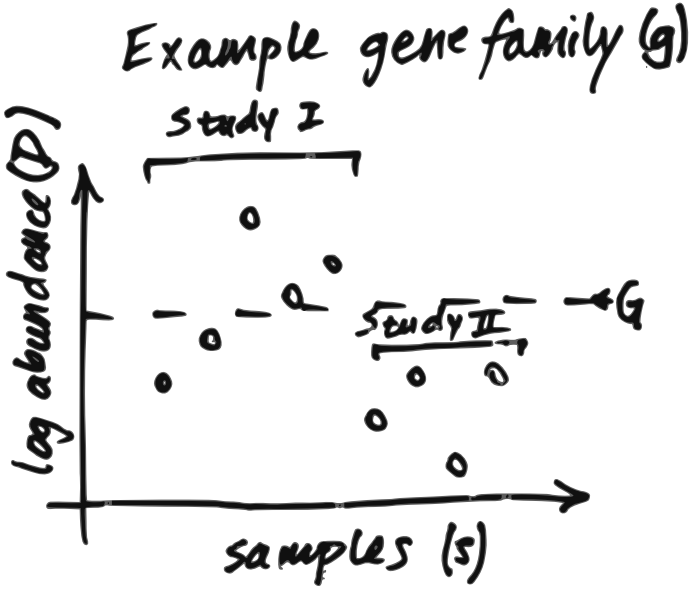
globe adapted from jasondavies.com

- We gather **shotgun sequencing** samples from either explicitly healthy people (HMP) or controls from case-control studies (MetaHIT, GC, T2D)
- Total **n = 53**
- Studies rarefied to 20M reads, then mapped to **~6K KEGG Orthology families** using Shotmap (Nayfach et al, submitted)

We fit a model to all of these data as follows



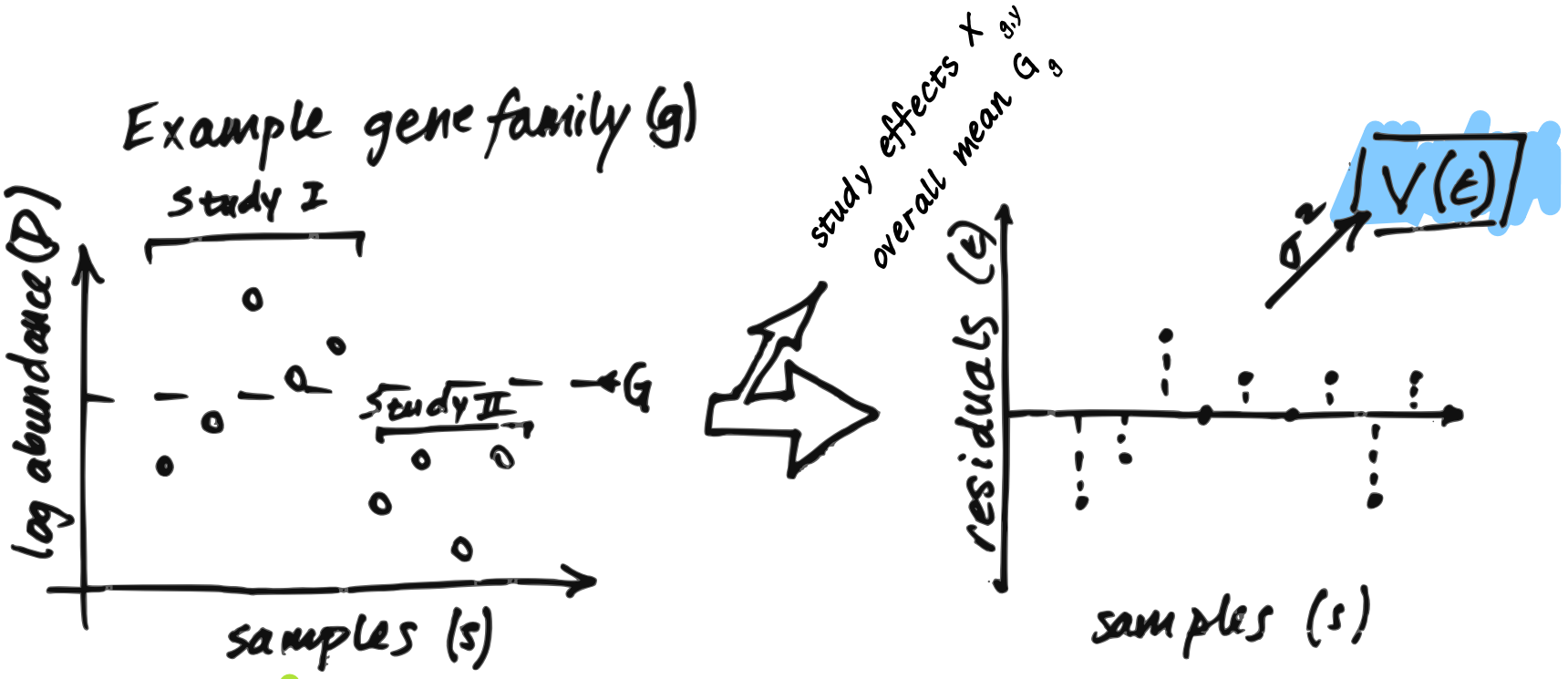
We fit a model to all of these data as follows



Model: $$D_{g,s} = G_g + \sum_{y \in Y} I_{y,s} X_{g,y} + \epsilon_{g,s}$$

- $D_{g,s}$: log abundance (gene g in sample s)
- G_g: overall mean of gene g
- $I_{y,s}$: indicator variable
- $X_{g,y}$: study effect for gene g in study y
- $\epsilon_{g,s}$: residuals

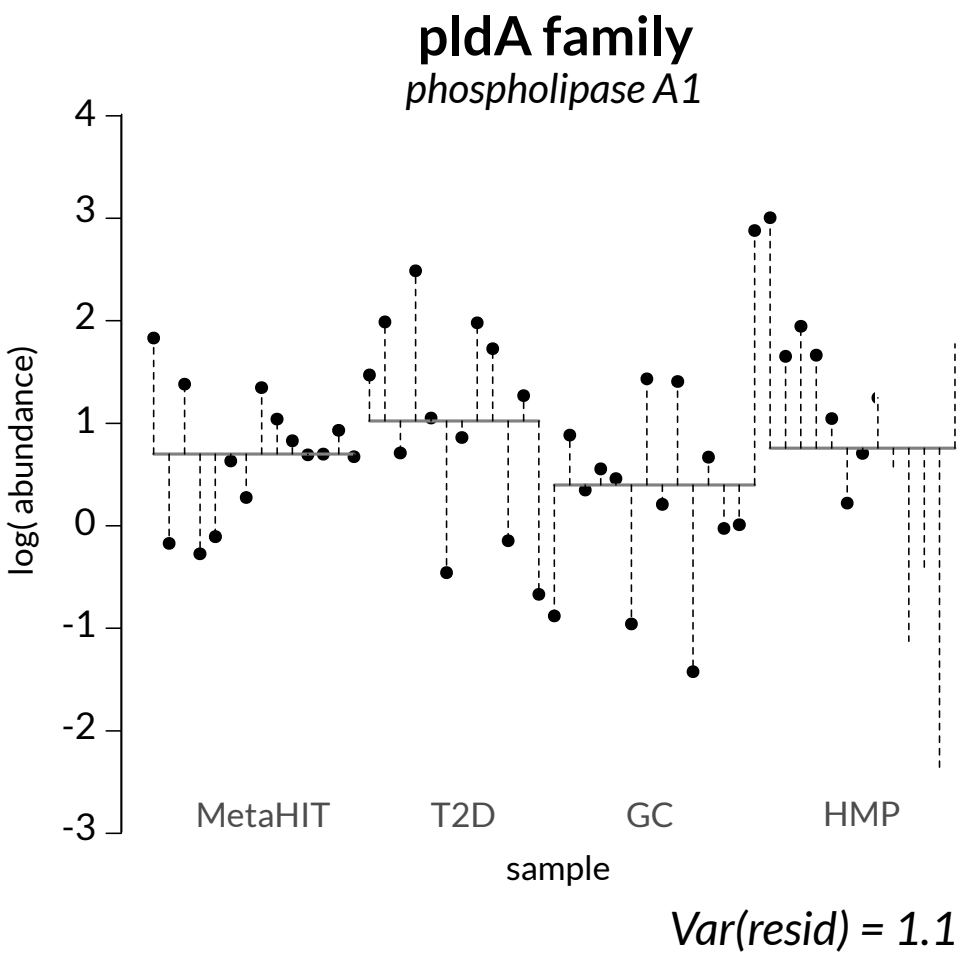
We fit a model to all of these data as follows

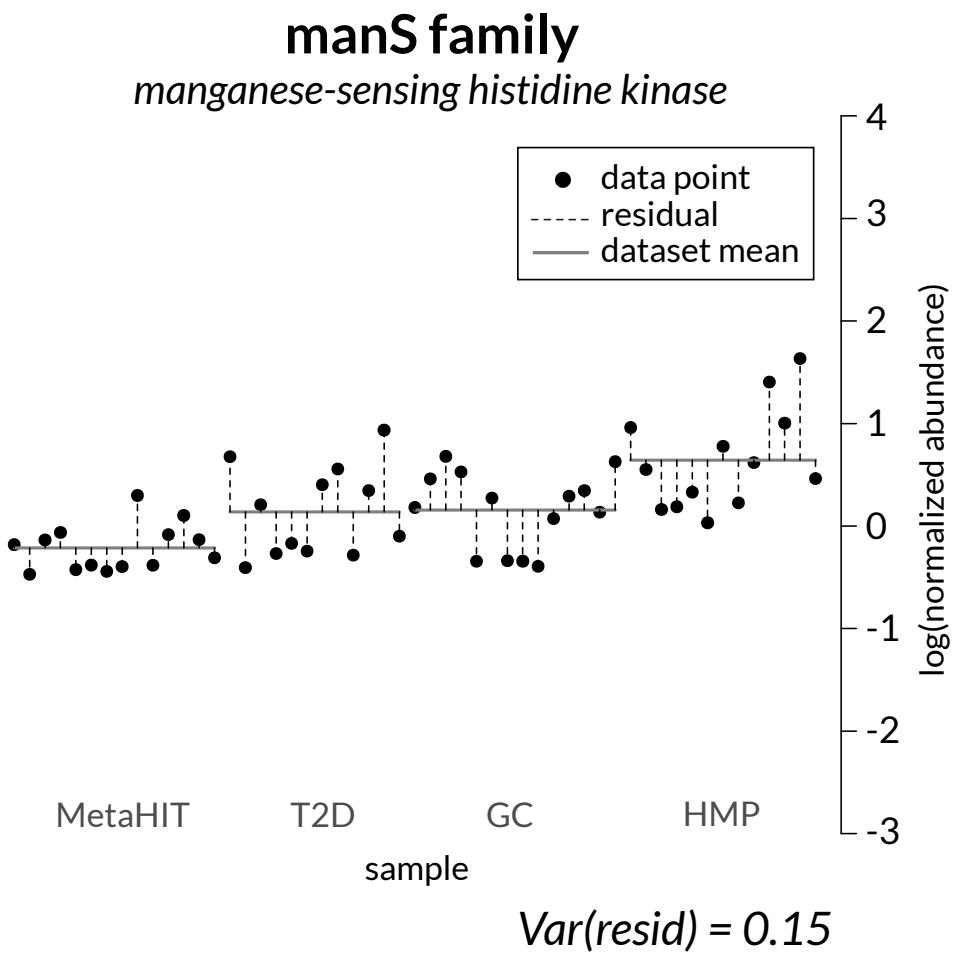


Model: $D_{g,s} = G_g + \sum_{y \in Y} I_{y,s} X_{g,y} + \epsilon_{g,s}$

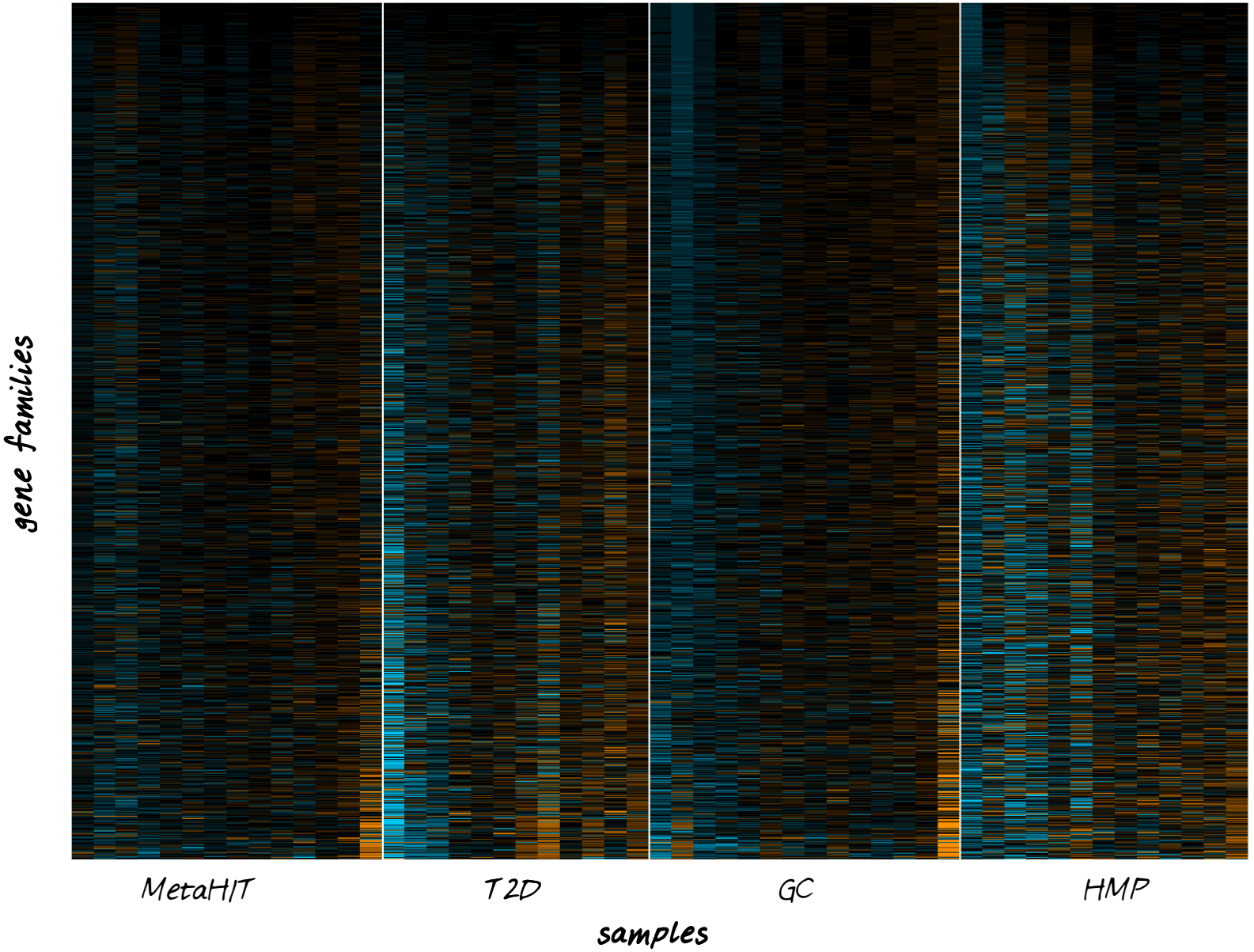
- $D_{g,s}$: log abundance (gene g in sample s)
- G_g: overall mean of gene g
- $I_{y,s}$: Indicator variable
- $X_{g,y}$: Study effect for gene g in study y
- $\epsilon_{g,s}$: residuals

Residual variances of real gene families





Overview of residual variances



How do we tell if these residual variances are higher/lower than expected?

Two possible options for the null

1) Generate data from a null distribution, then calculate test statistics
2) Generate a bootstrap distribution of (real) test statistics, then center and scale them using the null distribution

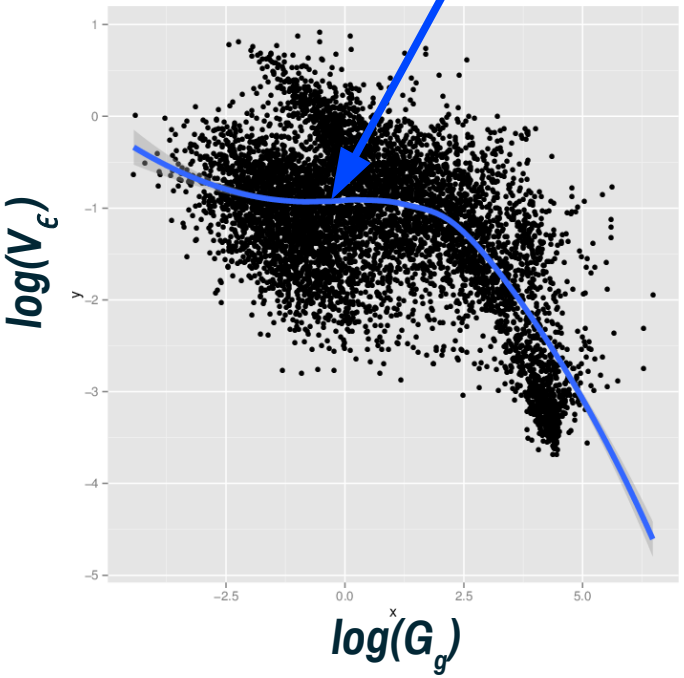
First option: generate data from the null

Model: $D_{g,s} = G_g + \sum_{y \in Y} I_{y,s} X_{g,y} + \epsilon_{g,s}$

$\epsilon_{g,s}$: generated from $N(0, \sqrt{(E(V(\epsilon) \mid G_g))})$

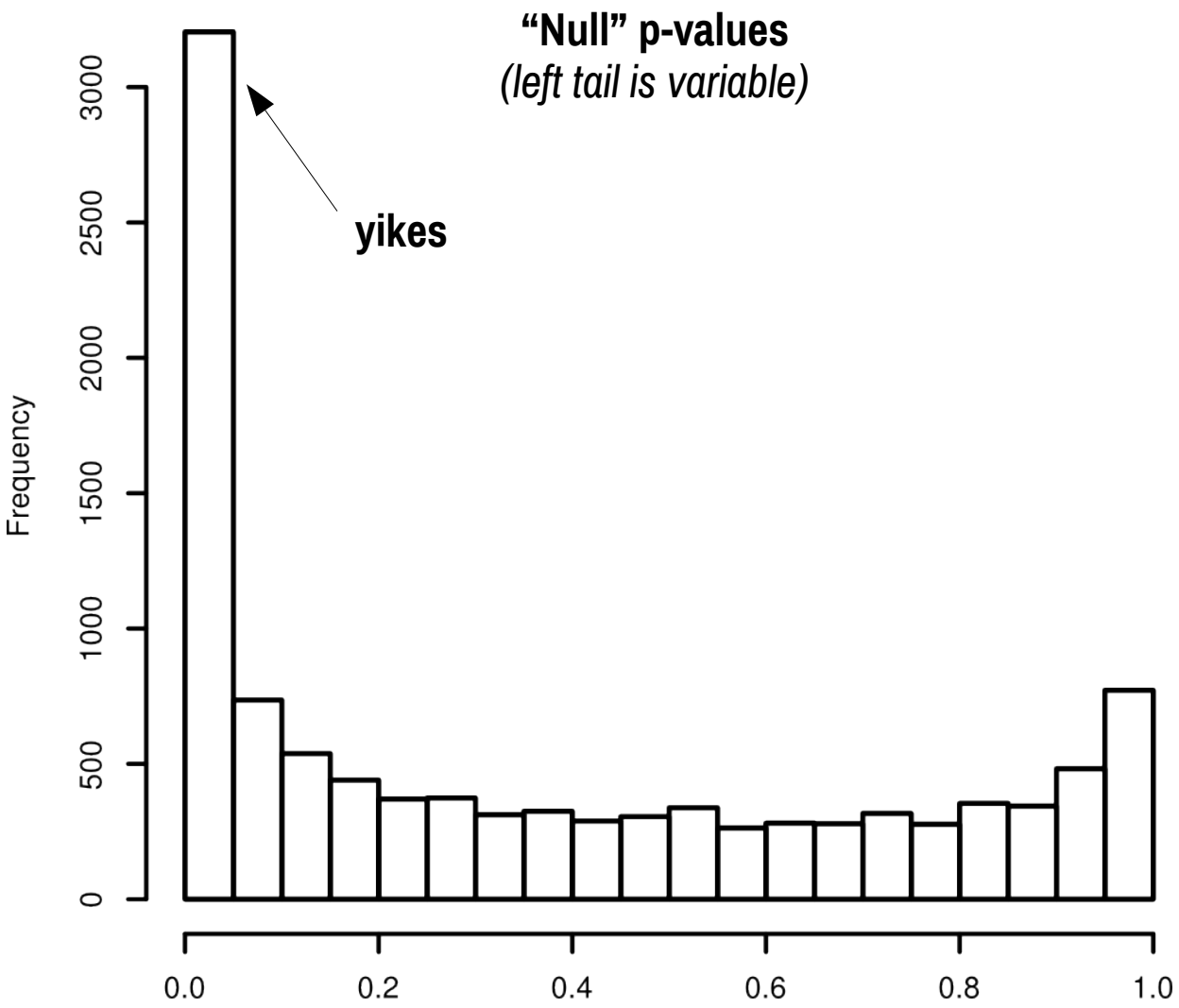
from what was fit originally

Assume that under the null, each gene's residual variance is what you would expect based on its mean:



Problem

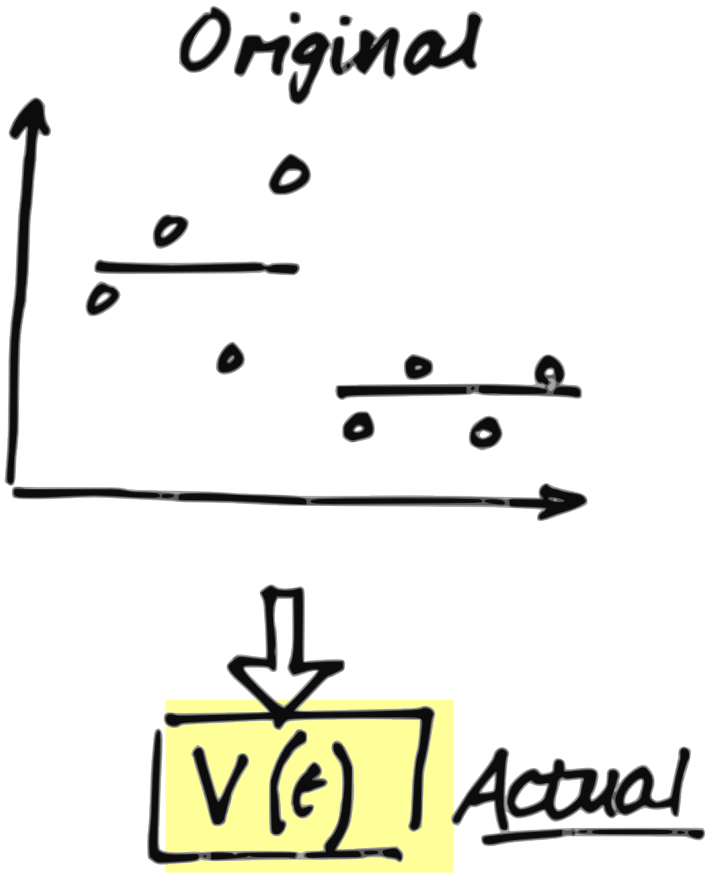
- We evaluated this on a dataset that we *know* to be null
  - Generate fake dataset, purely based on sampling with replacement from a real sample
- Result: **anti-conservative bias**
- Reasons?
  - Breaks dependence between gene families
  - Underestimates variability around null (related)



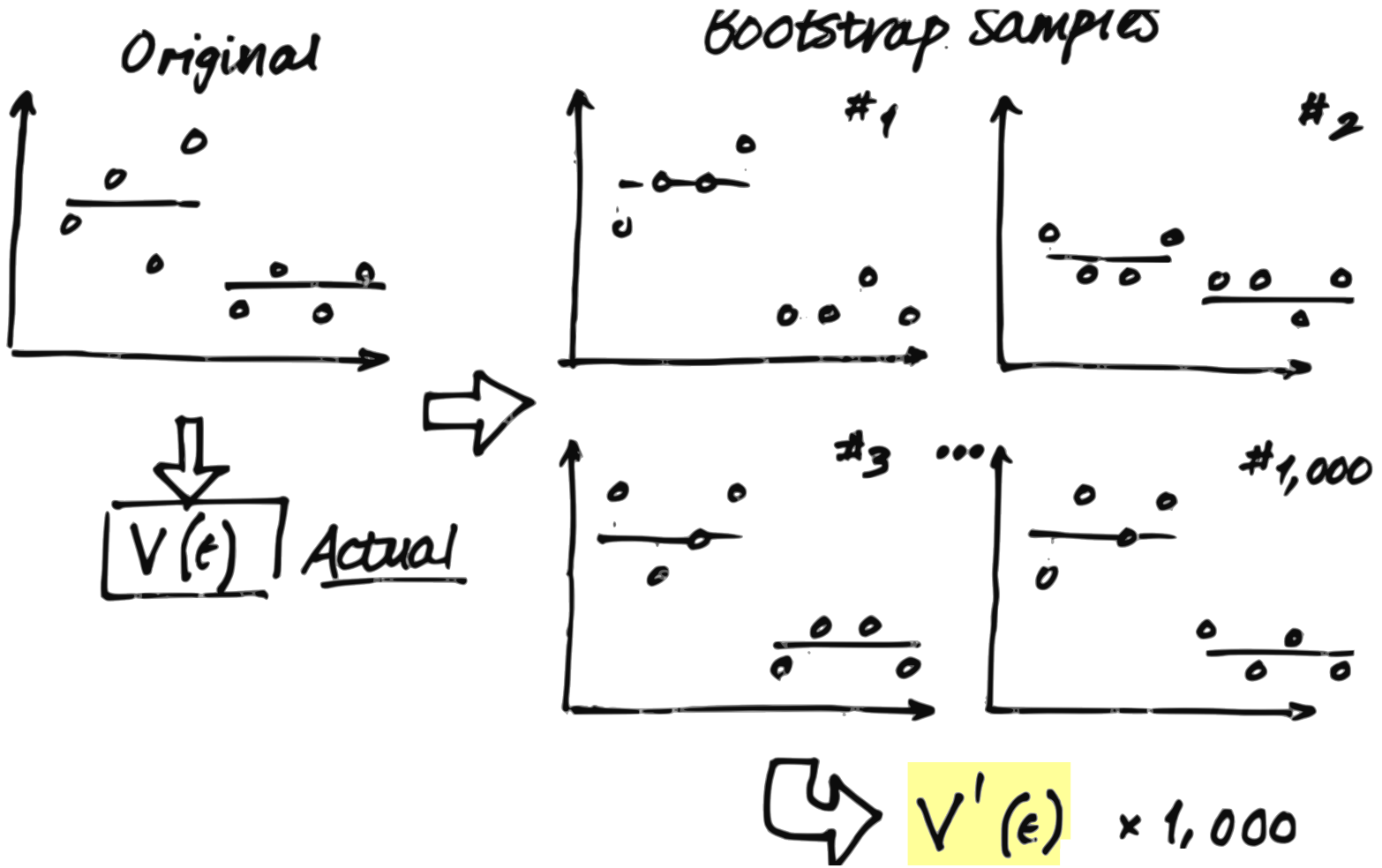
New desiderata for null distribution

- Under the null, assume reads just being drawn from the exact same population with replacement => **Poisson**
- But **also**, gene families aren't independent from one another. Our null should take that into account as well.

Second option: bootstrap test statistics

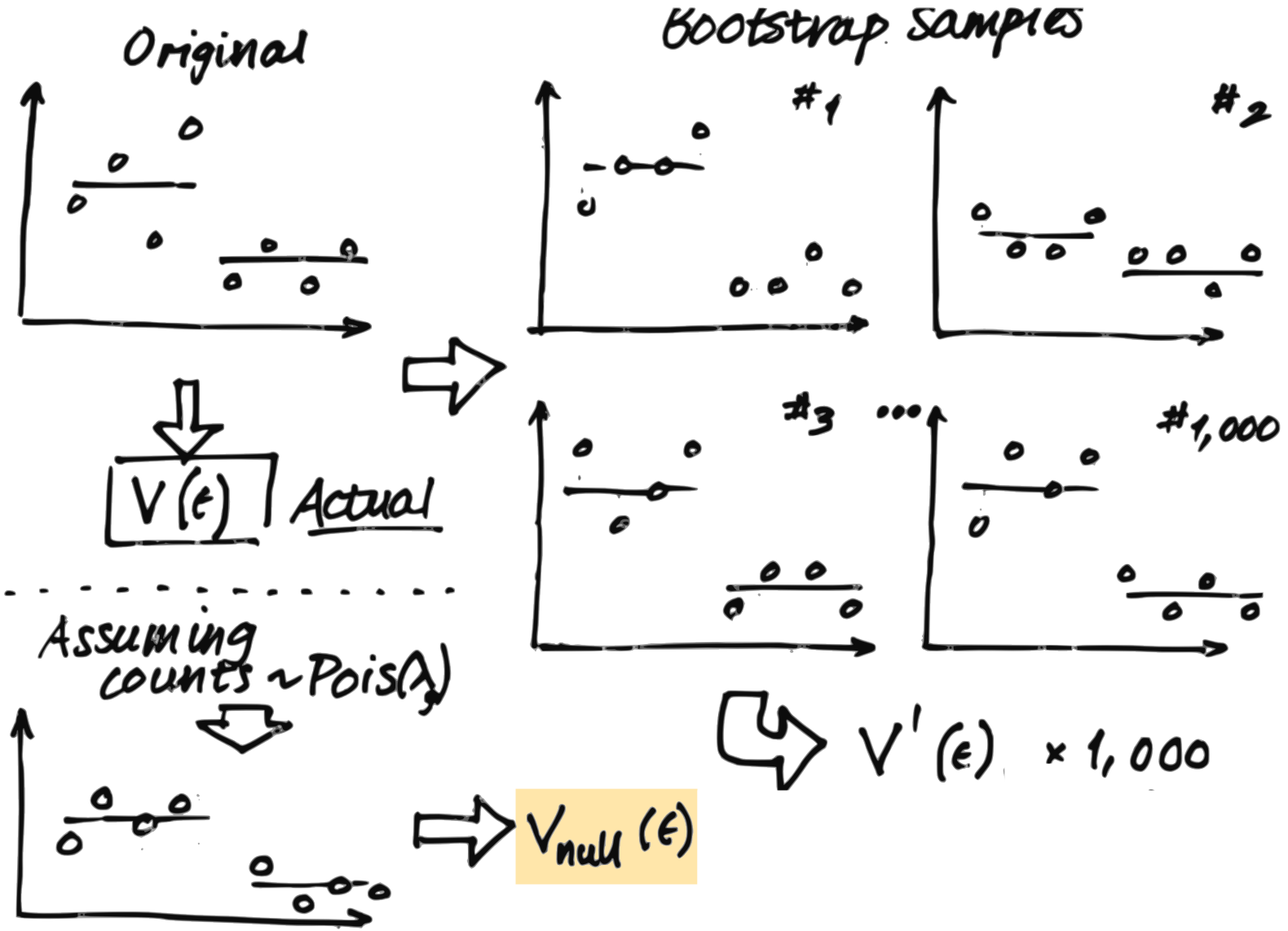


Second option: bootstrap test statistics

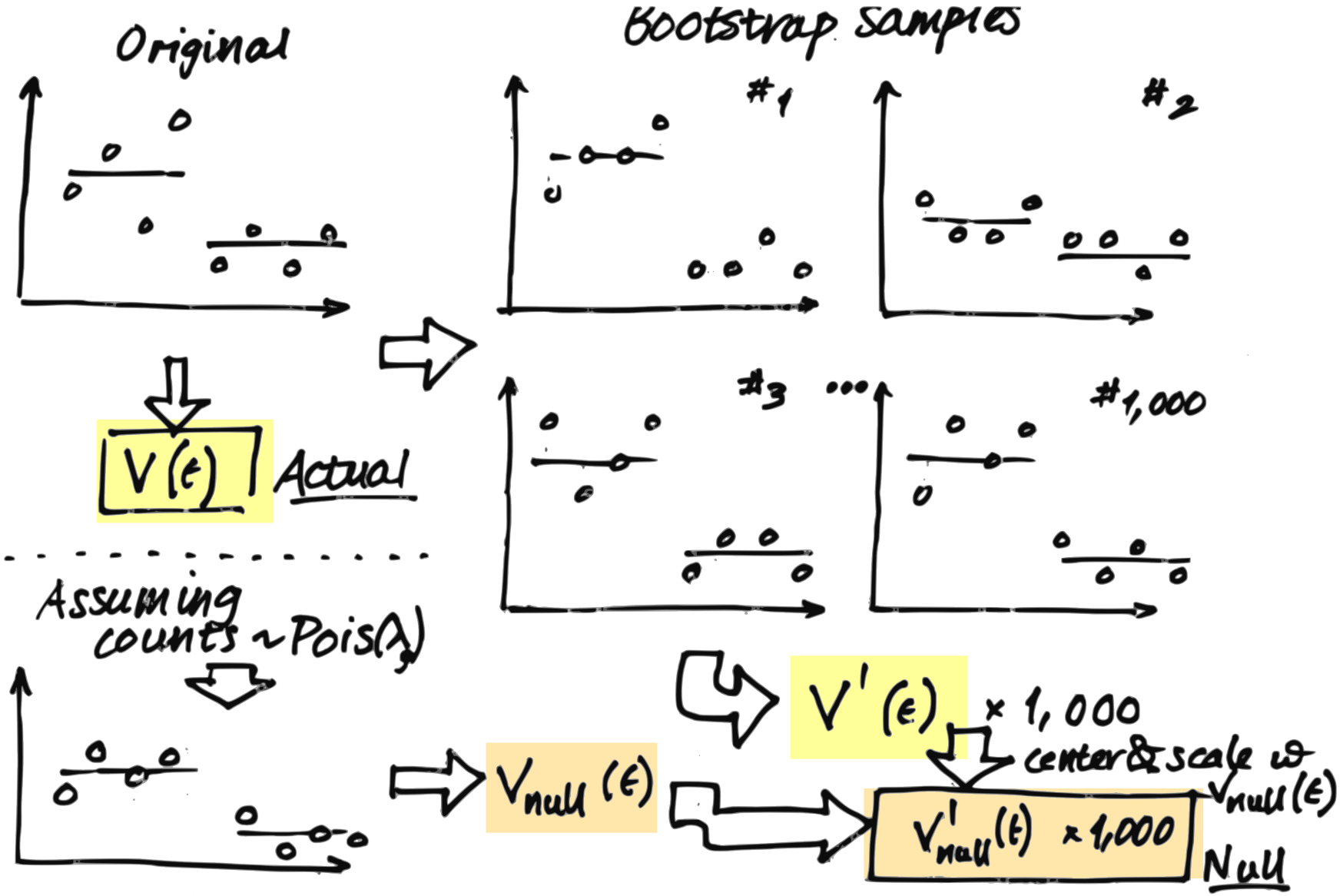


N.B. Here, I'm showing just one gene family, but the key is that they're all bootstrapped together (i.e. we're drawing **entire samples** with replacement). This is how we are trying to preserve gene-to-gene correlations.

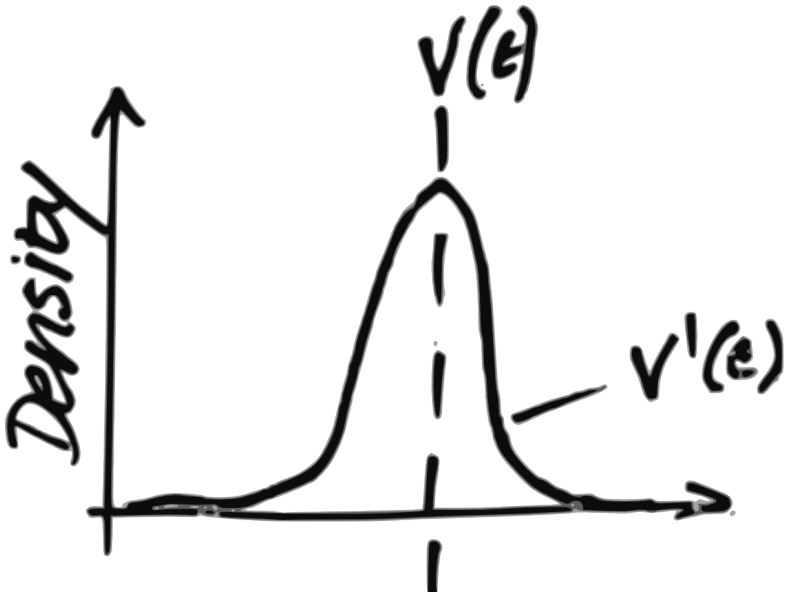
Second option: center and scale



Second option: center and scale



More detail on centering and scaling



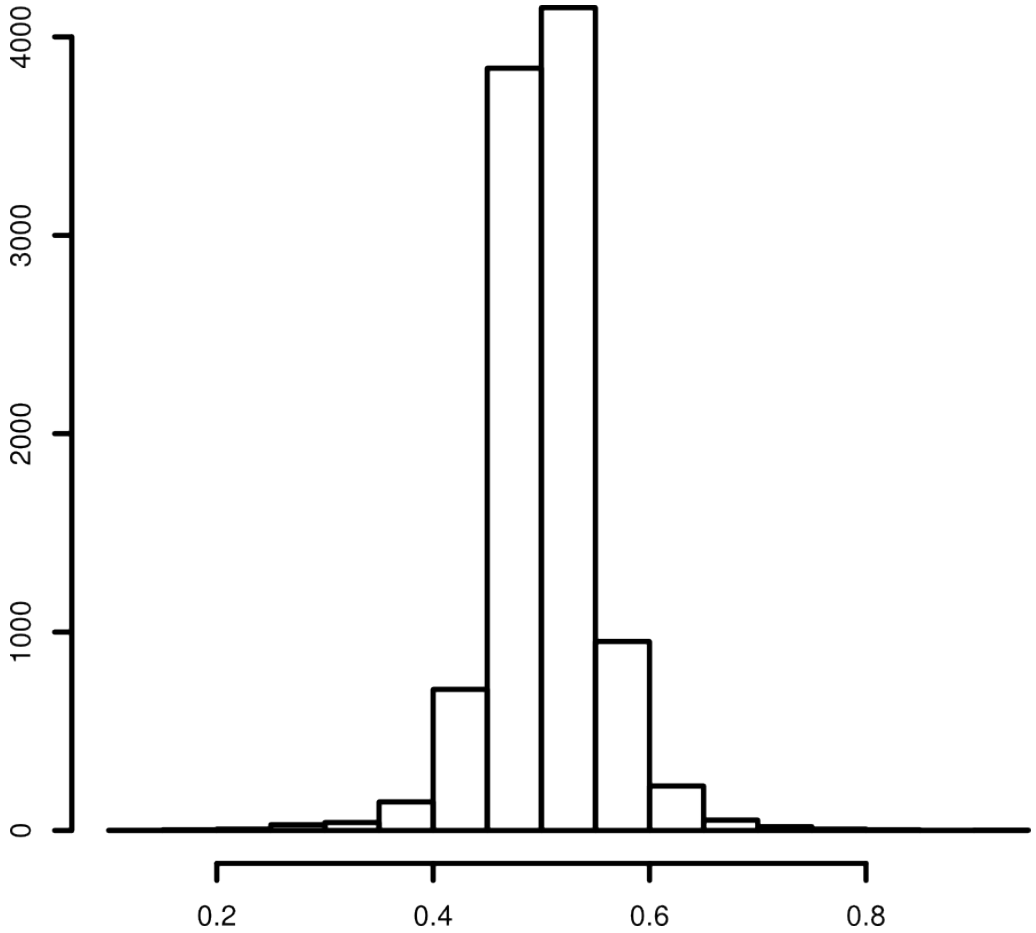
Center & scale $V'(\varepsilon)$ to have

$$\mu(V'(\varepsilon)) = V_{null}(\varepsilon)$$

$$\sigma(V'(\varepsilon)) = \sqrt{counts}$$

An alternative technique would be to try to generate V'_{null} by generating null data, but then you would break the gene-to-gene correlations unless you had some kind of explicit model for them.

Appears to work properly



- Results on same “null” dataset to left
- p-values centered around 0.5
- (note, still not a flat distribution; thus the test may actually be slightly conservative?)

Next steps

- Make sure I'm still able to make discoveries on real data
- Right now still working with counts; need to add in normalizations for:
  - average family length
  - average genome size
- Rewrite and submit

Thanks

- Stephen Nayfach
- Katie Pollard
- iSEEM folks
- funding sources