Beginning to Analyzing Function

Gail Rosen

Functional Annotation

- **dictionary definition of "to annotate":**
  - "to make or furnish critical or explanatory notes or comment"
- **some of what this includes for genomics**
  - gene product names
  - functional characteristics of gene products
  - physical characteristics of gene/protein/genome
  - overall metabolic profile of the organism
- **elements of the annotation process**
  - gene finding
  - homology searches
  - functional assignment
  - ORF management
  - data availability

ORF

- Open Reading Frame

1. ATG CAA TGG GGA AAT GTT ACC AGG TCC GAA CTT ATT GAG GTA AGA CAG ATT TAA
2. A TGC AAT GGG GAA ATG TTA CCA GGT CCG AAC TTA TTG AGG TAA GAC AGA TTT AA
3. AT GCA ATG GGG AAA TGT TAC CAG GTC CGA ACT TAT TGA GGT AAG ACA GAT TTA A

Sample sequence showing three different reading frames. Start codons are highlighted in purple, and stop codons are highlighted in red.

Annotation Pipeline

Generation of Open Reading Frames

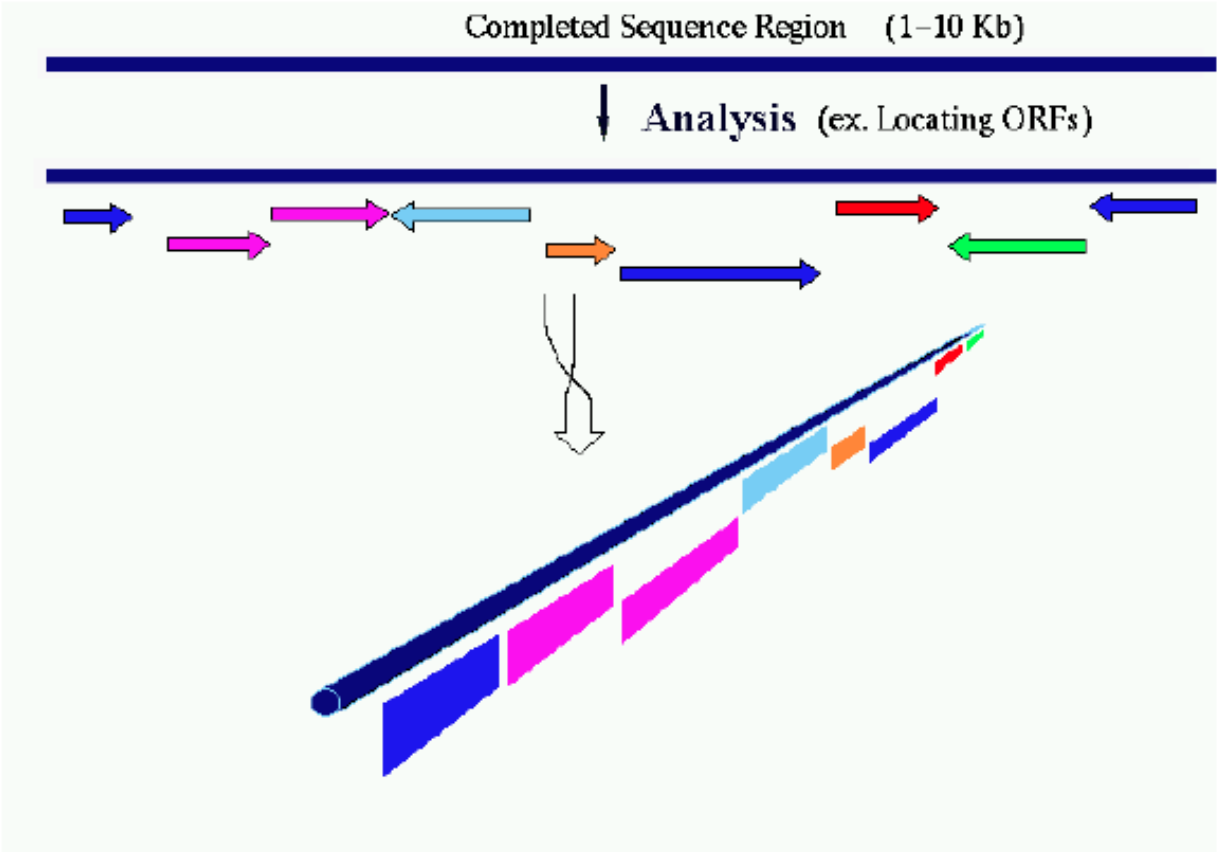
Homology Searches

Putative ID
Frameshift Detection
Ambiguity Report

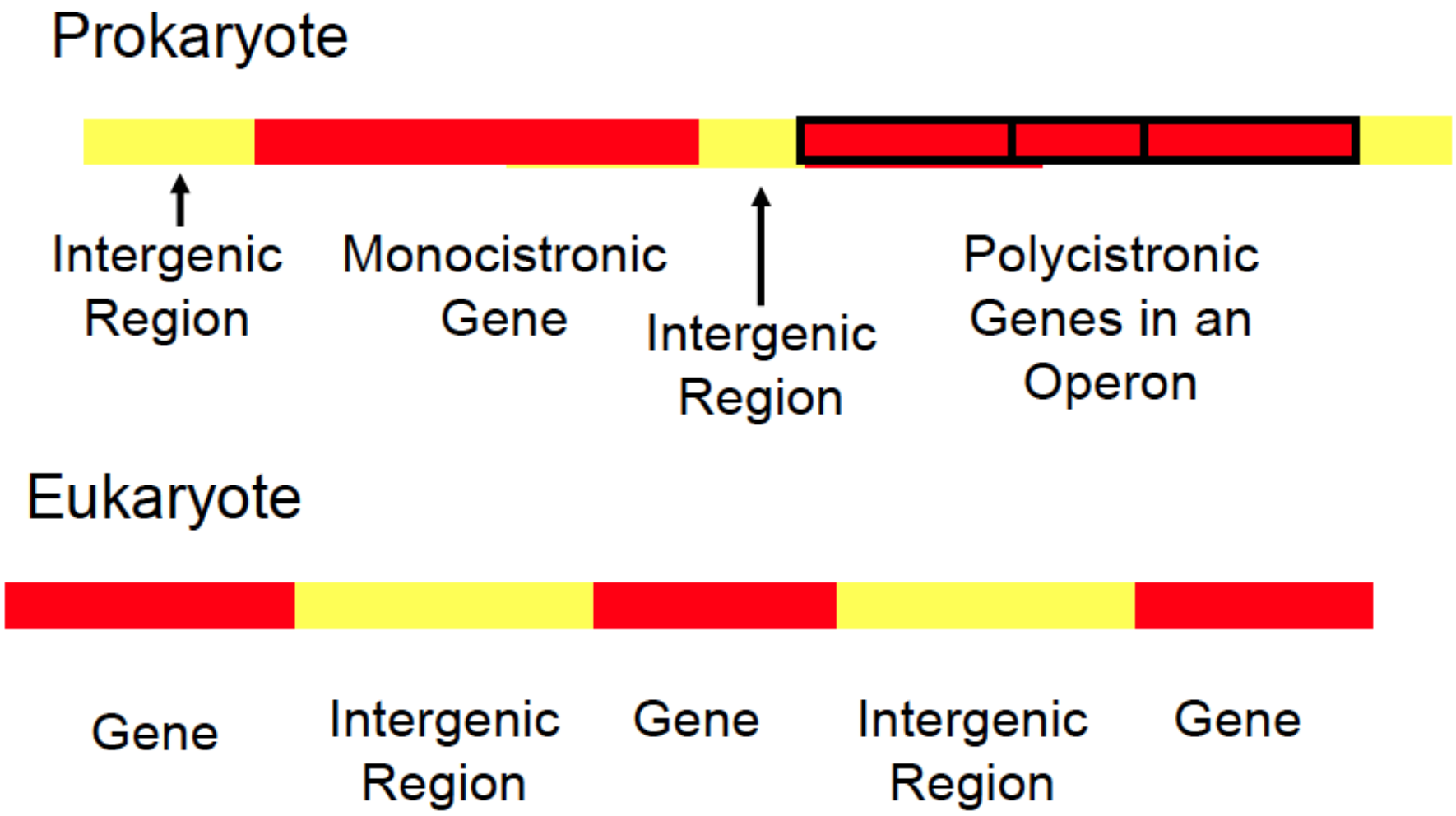
Role Assignment
Metabolic Pathways
Gene Families

DNA Motifs
Regulatory Elements
Repetitive Sequences

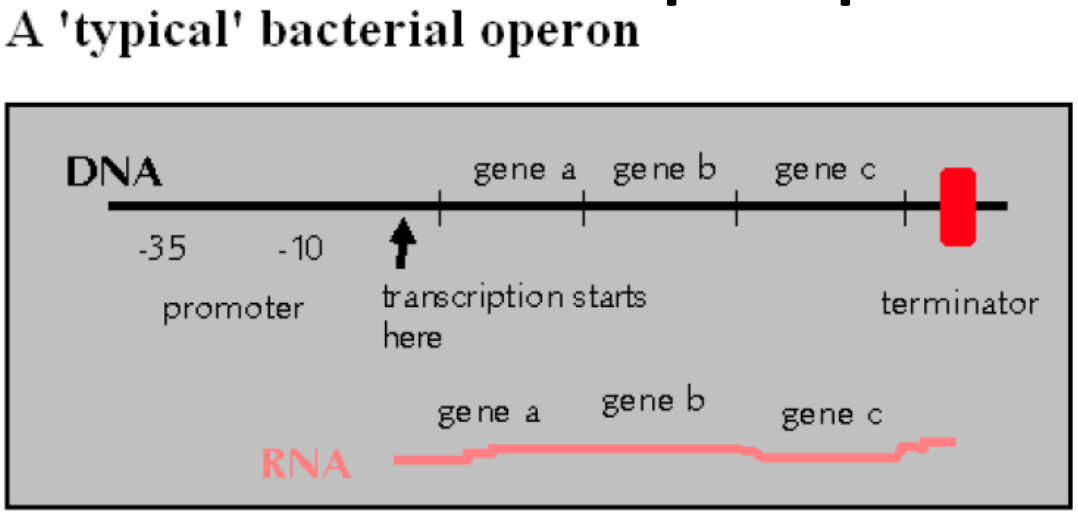
Comparative Genomics

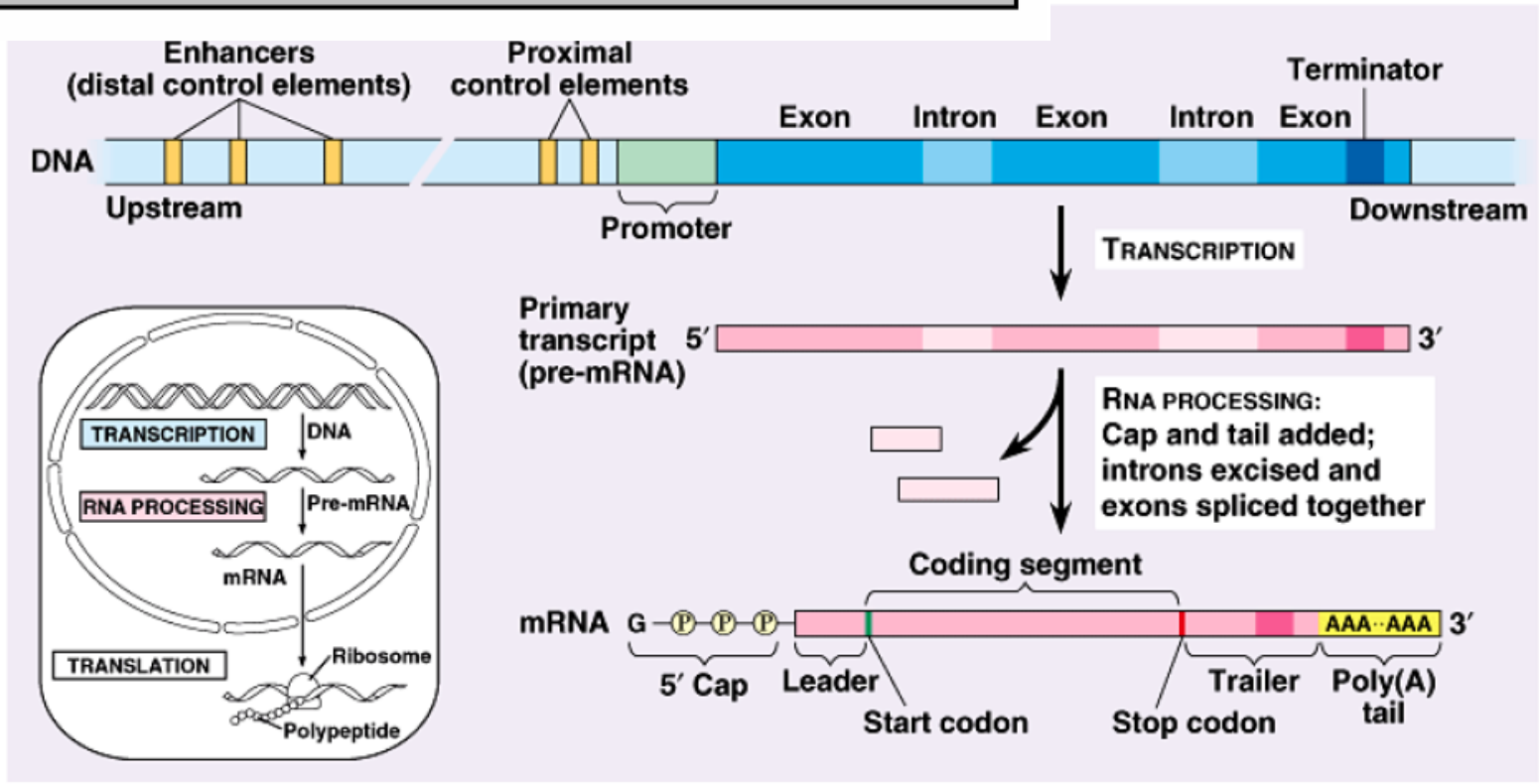


Genome Structure



Transcript “processing”





Annotating

Two main types of data used in defining gene structure:

Prediction based: algorithms designed to find genes/gene structures based on nucleotide sequence and composition

Sequence similarity (DNA and protein): alignment to mRNA sequences (ESTs) and proteins from the same species or related species; identification of domains and motifs

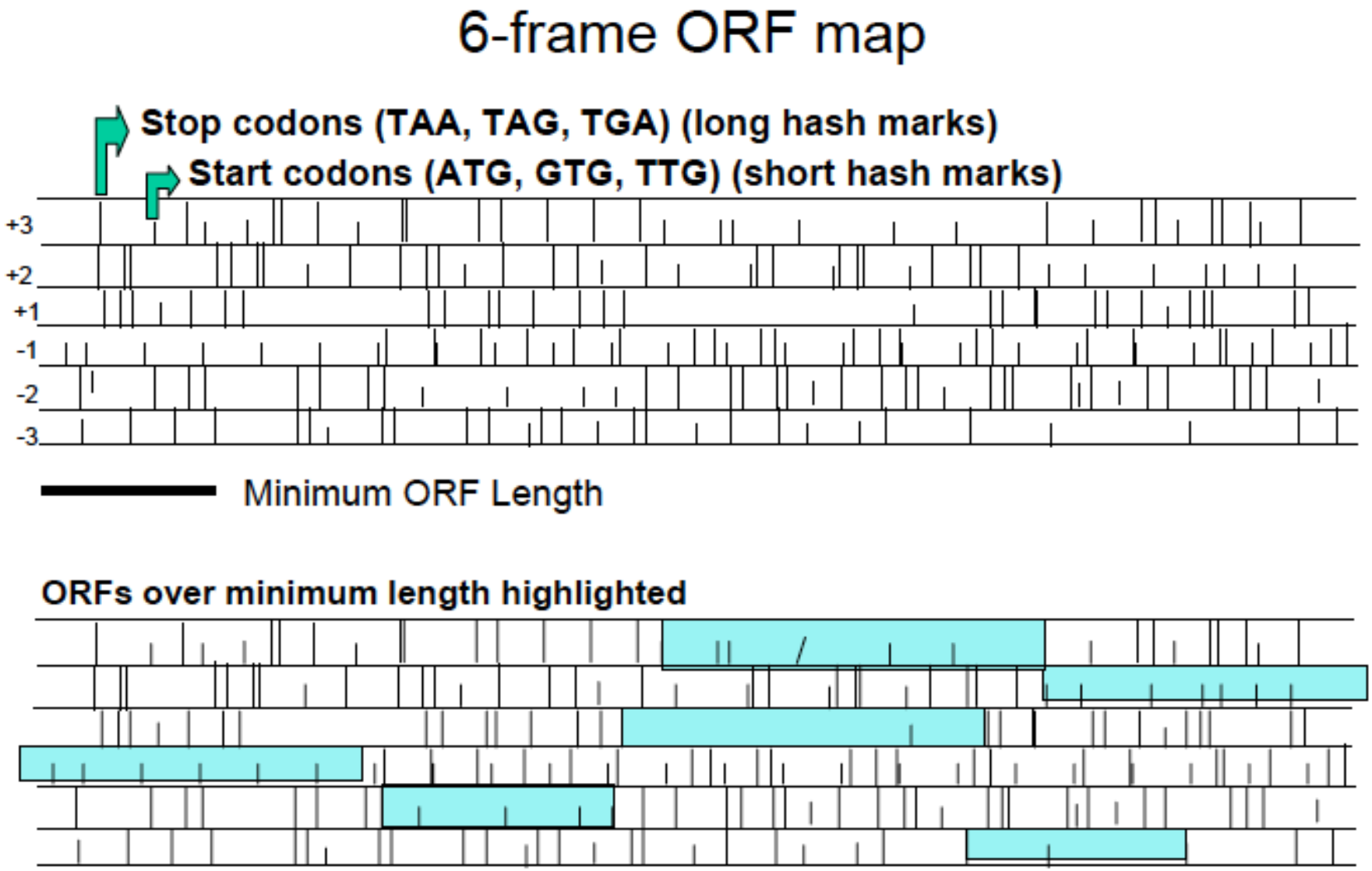
First Step: ORF Finding (traditionally whole organisms)

Running a Gene-finder is a two-part process

1) Train Gene finder for the organism you have sequenced.

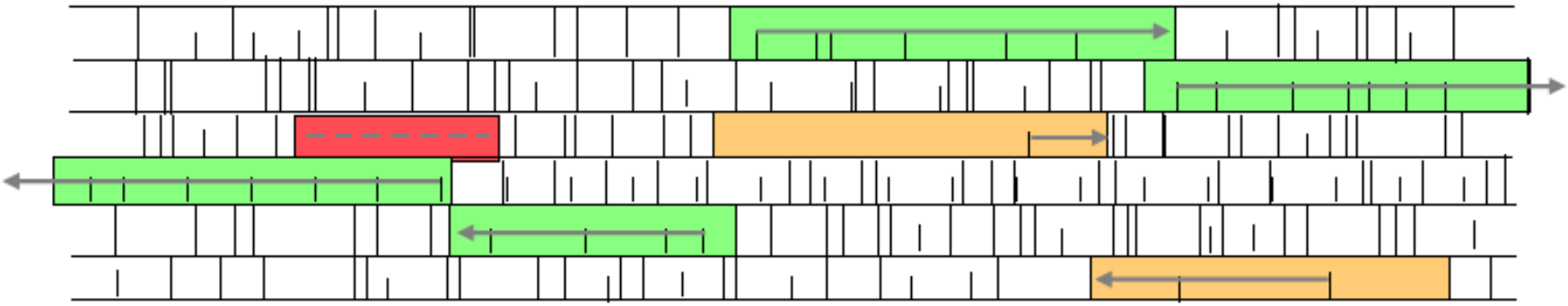
2) Run the trained Gene finder on the completed sequence.

Candidate Genes

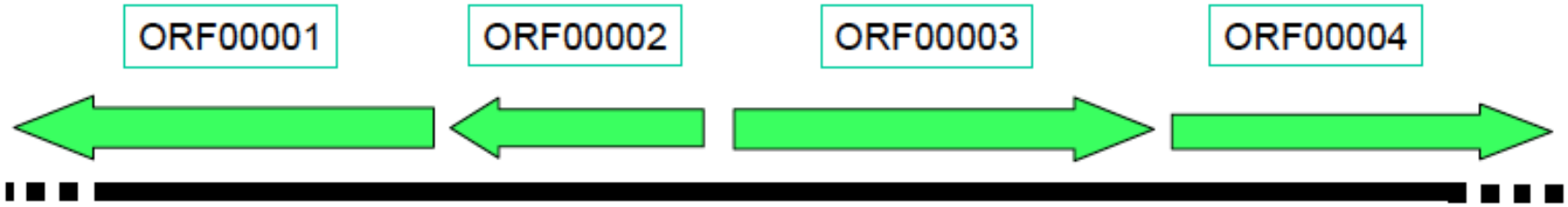


Annotate ORFs

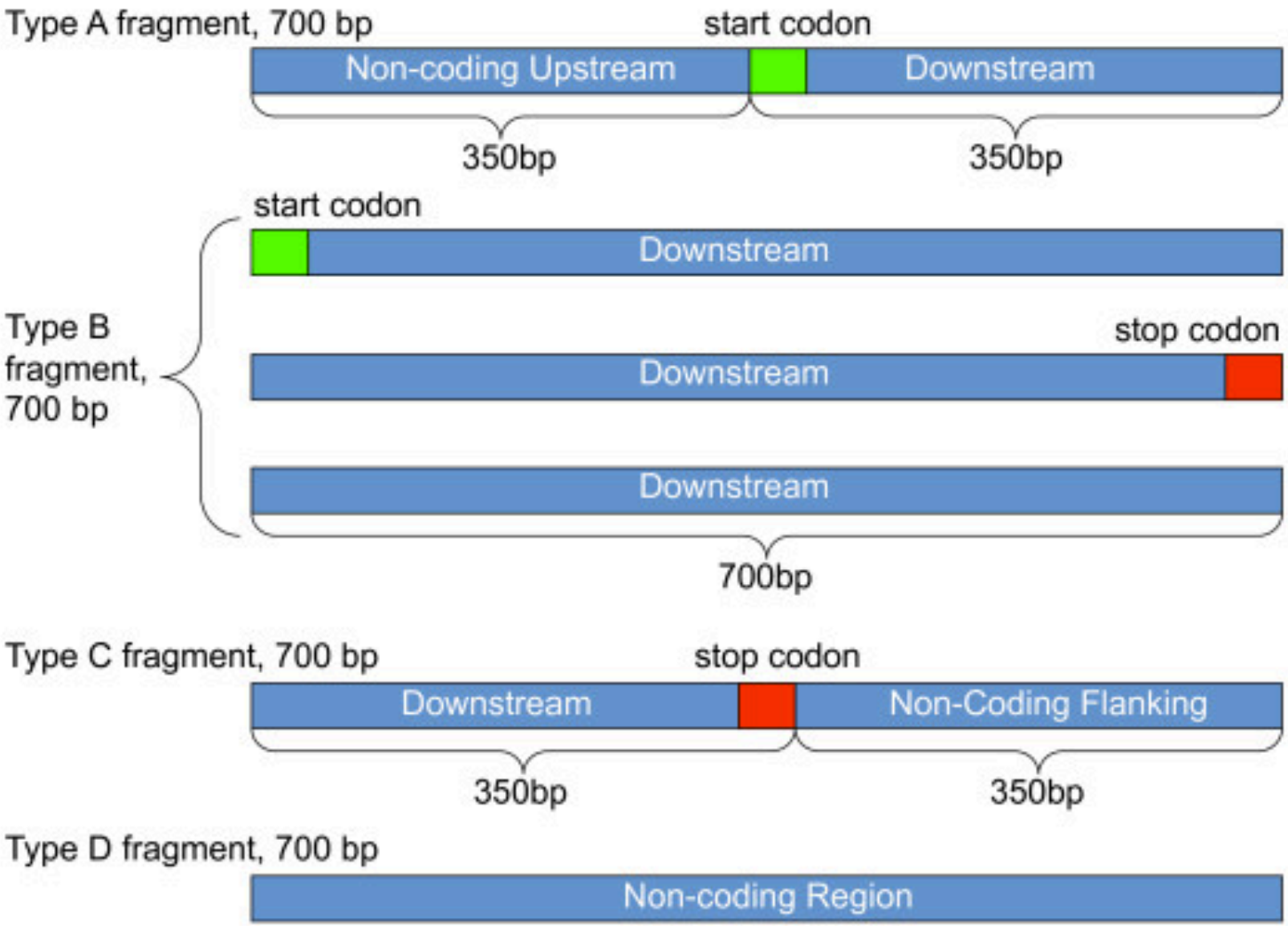
Possible translations represented by arrows, moving from start to stop, the dotted line represents an ORF with no start site.



Glimmer chooses the set of likely genes.



More Complicated for Metagenomics



Functional Assignments

Name

Descriptive common name for the protein, with as much specificity as the evidence supports; gene symbol.

Role

Describe what the protein is doing in the cell and why.

Associated information:

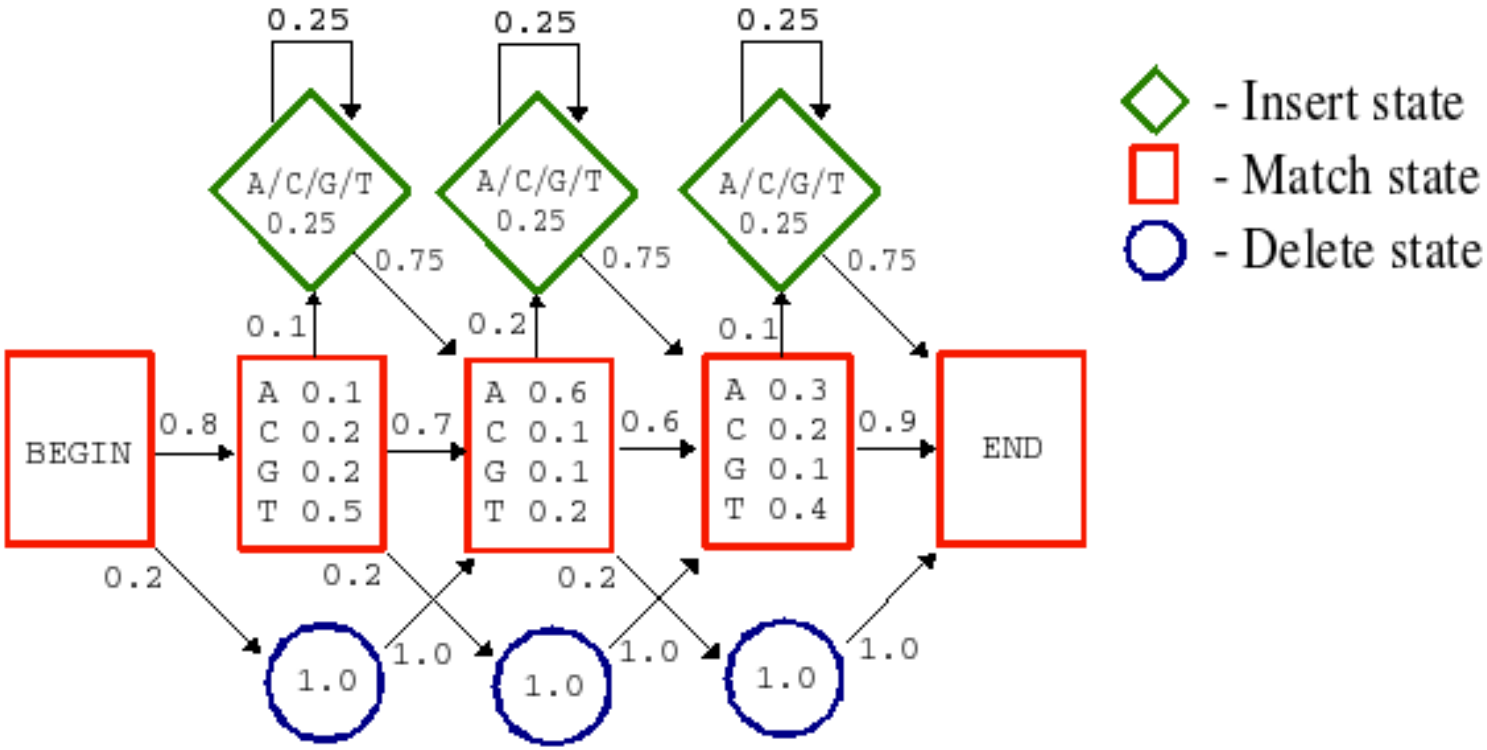
Supporting evidence:Domain and motifs
EC number if protein is an enzyme.
Paralogous family membership.

Evidence for Gene Function

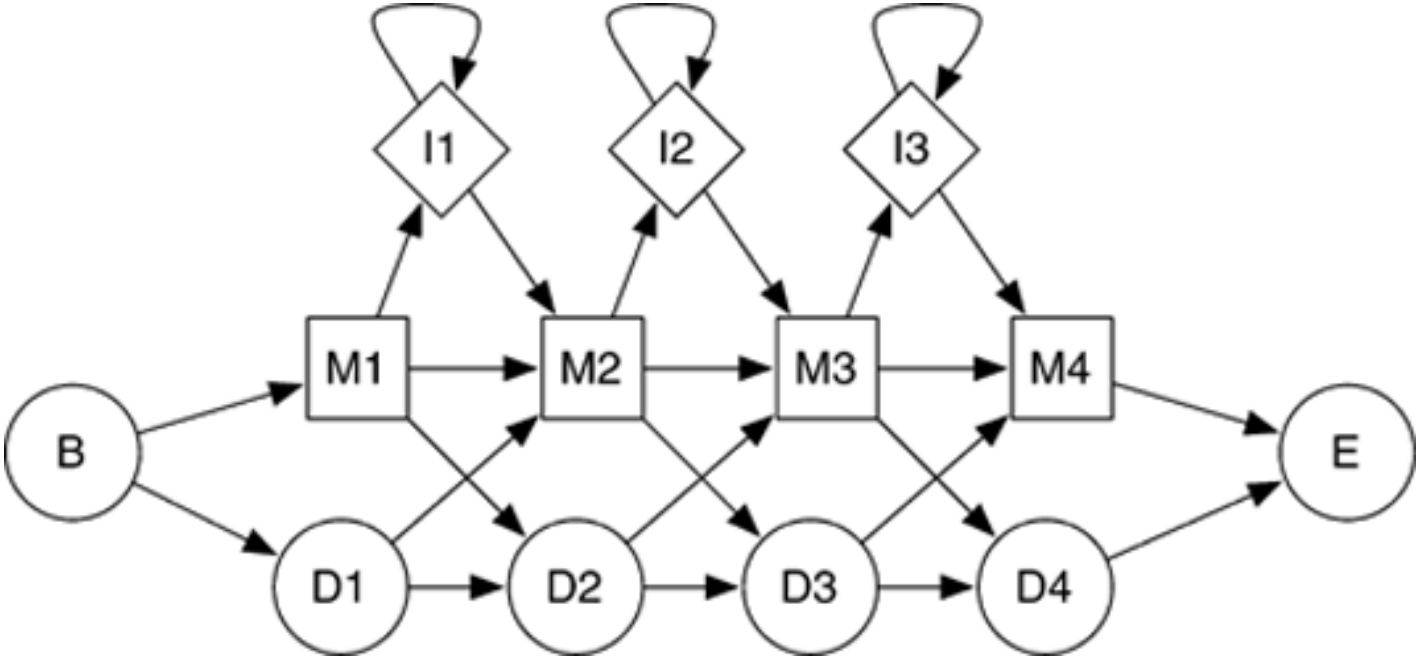
- PROSITE Motifs
  - collection of protein motifs associated with active sites, binding sites, etc.
  - help in classifying genes into functional families when HMMs for that family have not been built

- InterPro
  - Brings together HMMs (both TIGR and Pfam) Prosite motifs and other forms of motif/domain clustering
  - Results in motif “signatures” for families or functions
  - GO terms have been assigned to many of these

Markov Chains



Profile HMMs



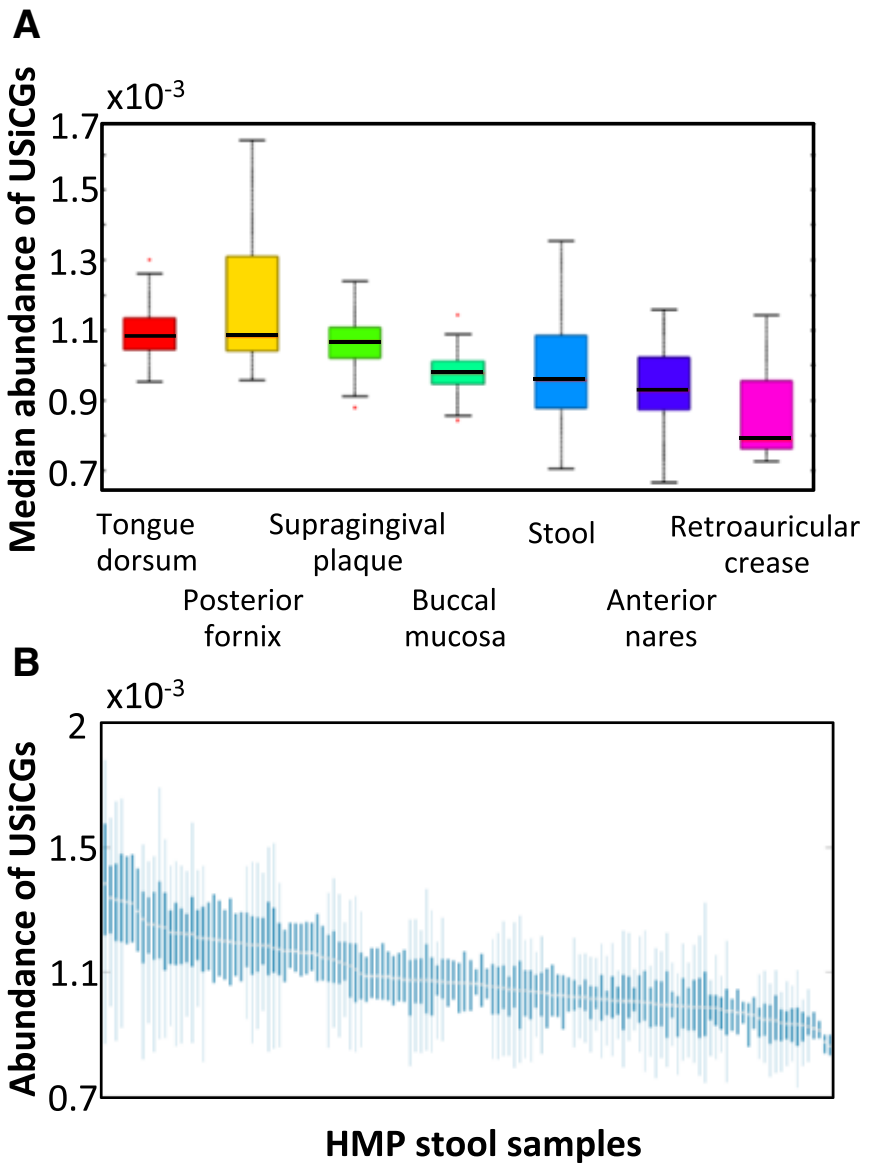
* Viterbi: Find labels given model and sequence
* Forward/Backward Algorithm: Find Probability of label at a certain position given model and sequence

Next

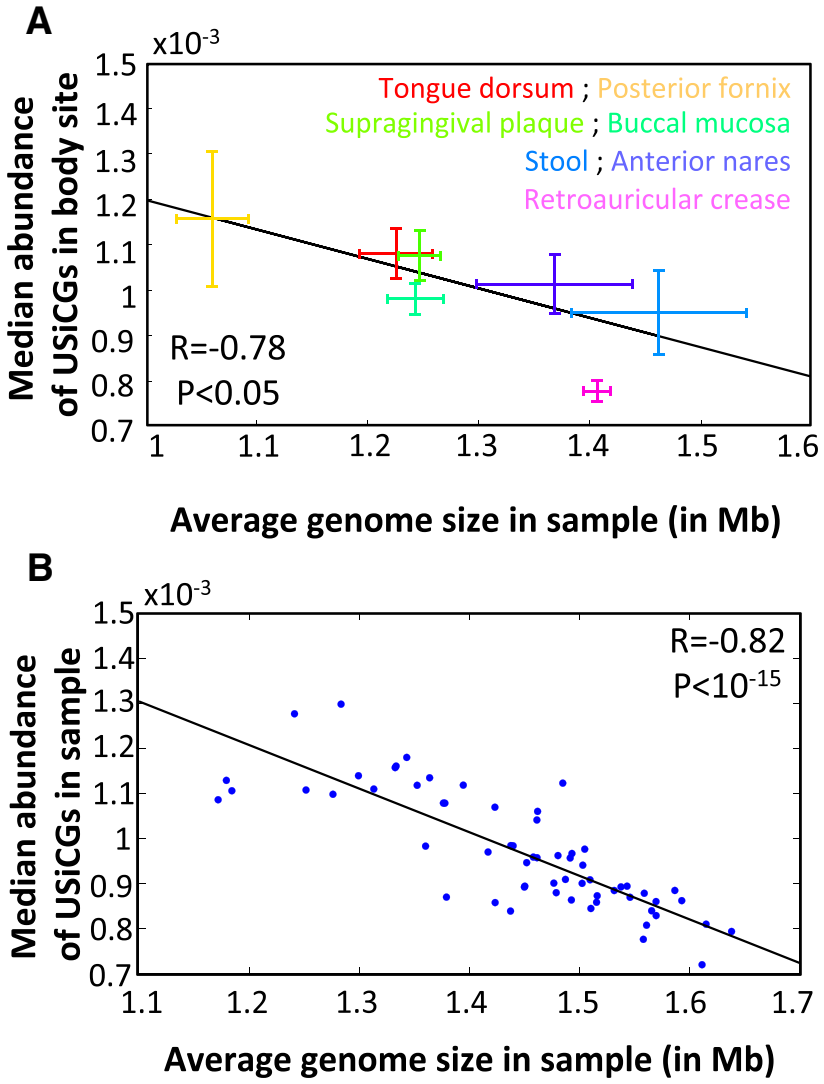
- Have genes annotated
- How to see their “differential abundance” between samples

Normalizing Gene Abundances

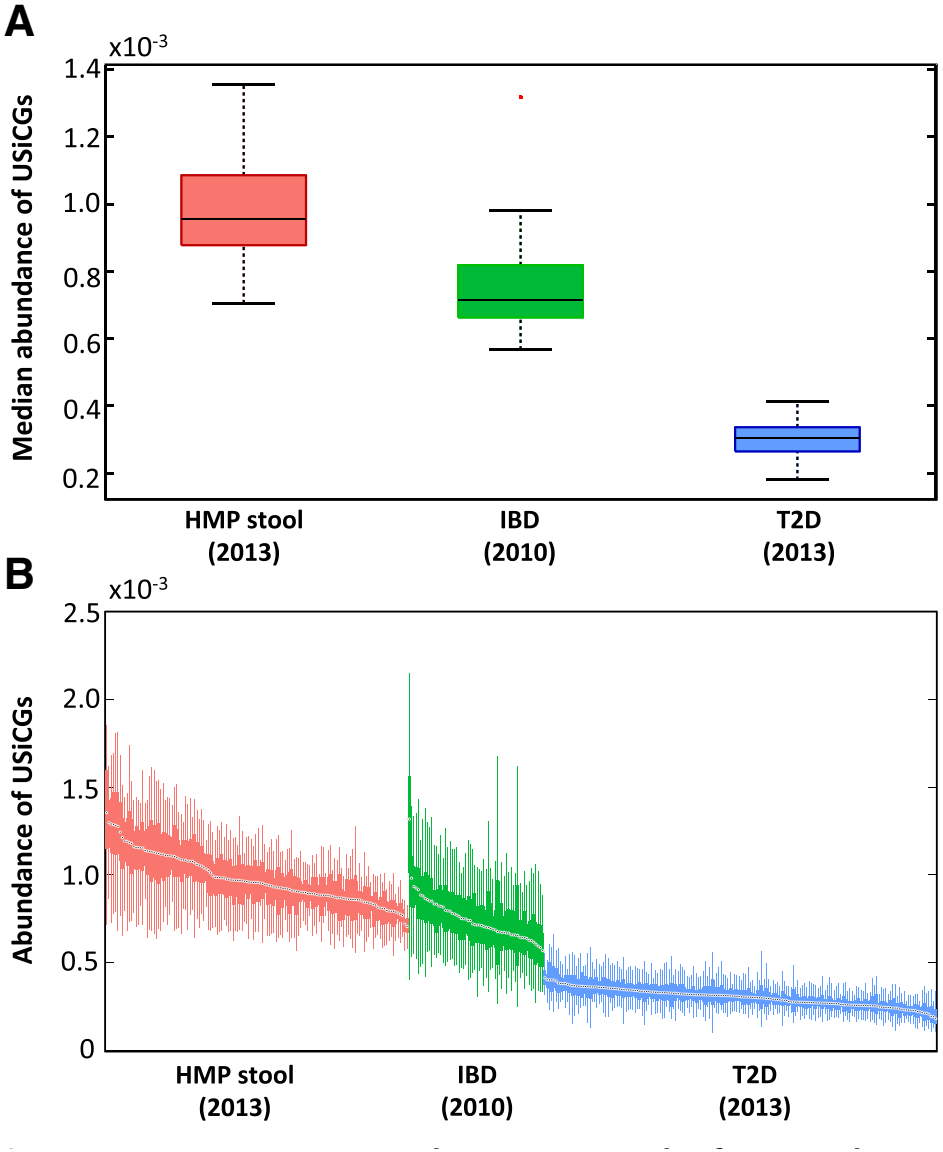
- Multiple copies of genes?
- Single-copy marker genes



RA lower in samples with larger genomes



Inter- and intra- studies



MUSiCC

- Describe the abundance of each gene in the microbiome, not as its relative abundance in the sample, but rather as the average copy number of this gene across all microbial cells in the sampled community.
- Description of the gene content of a ‘typical microbe’ from that sample. This profile therefore has a clear biological interpret-ation, and can be reliably and meaningfully compared across samples.

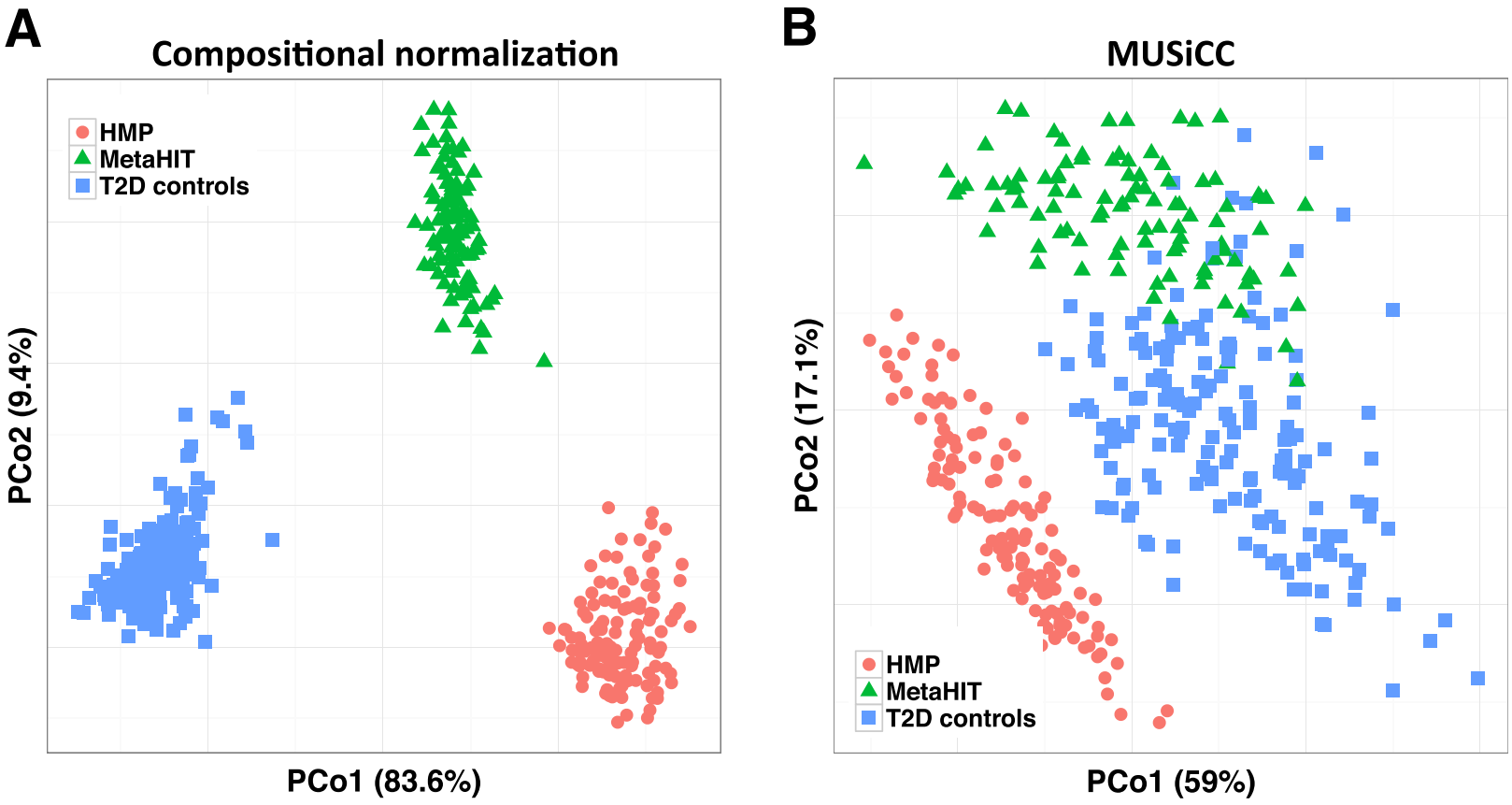
Goal

- The measured abundance of each gene in a sample (after correcting for gene length) is dividedby the median abundance of the USiCGs in the sample (NOT by the sum of all abundances)

Algorithm

- Learn Fold-change between the abundance of each USiCG and the mean abundance of all USiCGs in the sample
- five-fold cross-validation (CV) scheme to learn an Elastic-Net regularized linear model that predicts the fold-change of each USiCG in the sample.

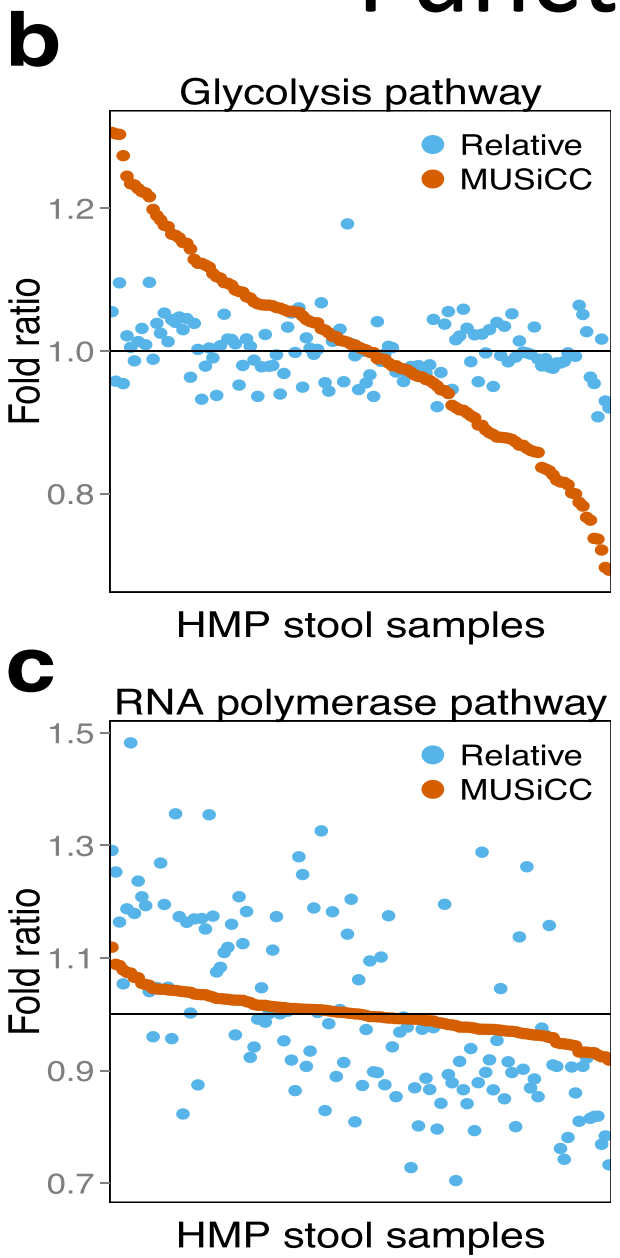
Post-MUSiCC


A
Compositional normalization
HMP
MetaHIT
T2D controls
PCo2 (9.4%)
PCo1 (83.6%)
B
MUSiCC
PCo2 (17.1%)
HMP
MetaHIT
T2D controls
PCo1 (59%)

Applying to Functional Pathways

- For each KEGG pathway, the variation in the profile was calculated as the coefficient of variation (CoV), defined as the standard variation in the abundance across samples divided by the mean abundance.

Functional


b
Glycolysis pathway
Relative
MUSiCC
1.2
1.0
0.8
Fold ratio
HMP stool samples
c
RNA polymerase pathway
Relative
MUSiCC
1.5
1.3
1.1
0.9
0.7
Fold ratio
HMP stool samples