Oligotyping EF-1a amplicons:

A method for subspecies investigations of *Fusarium oxysporum*

9-8-17

Peter Henry

PhD Student

PIs: Tom Gordon and Johan Leveau

Dept. Plant Pathology at UC Davis

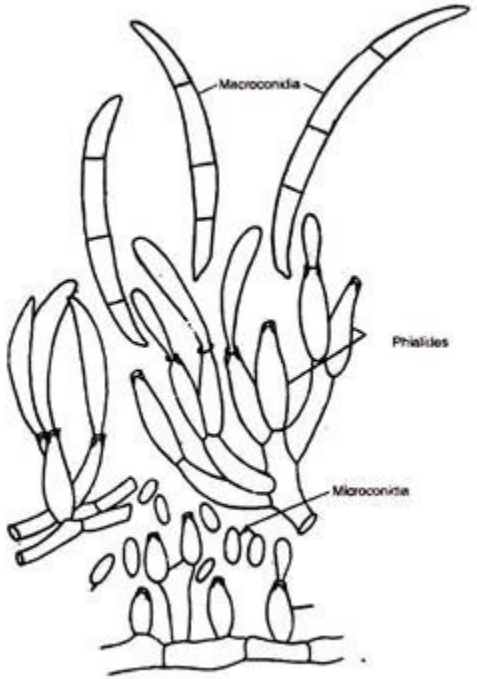
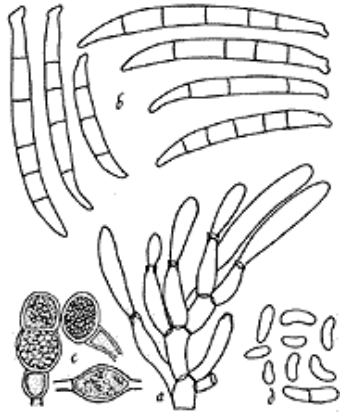


Fig. 3. Fusarium : Phialide, microconidia and macroconidia.



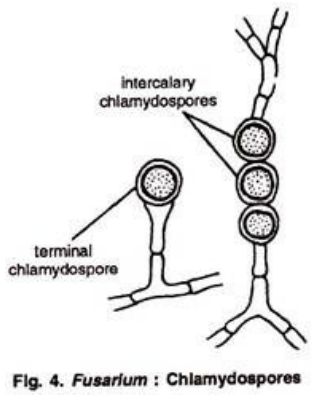
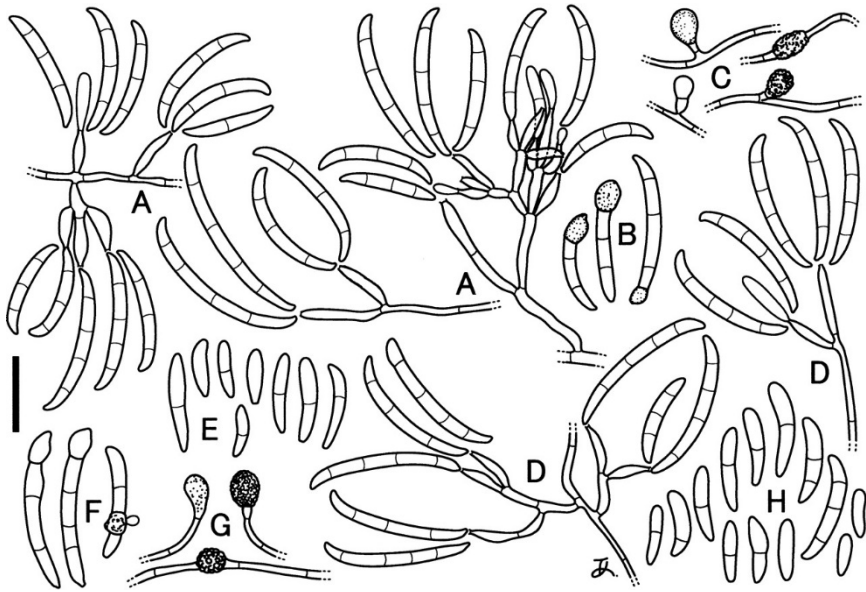
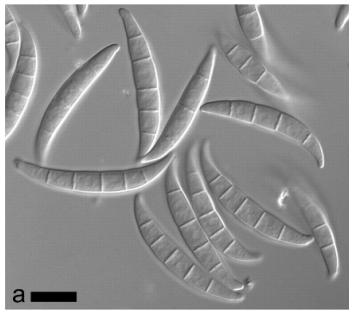
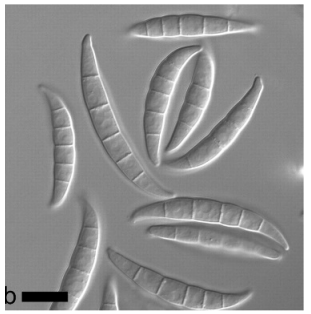


Fig. 4. *Fusarium* : Chlamydospores





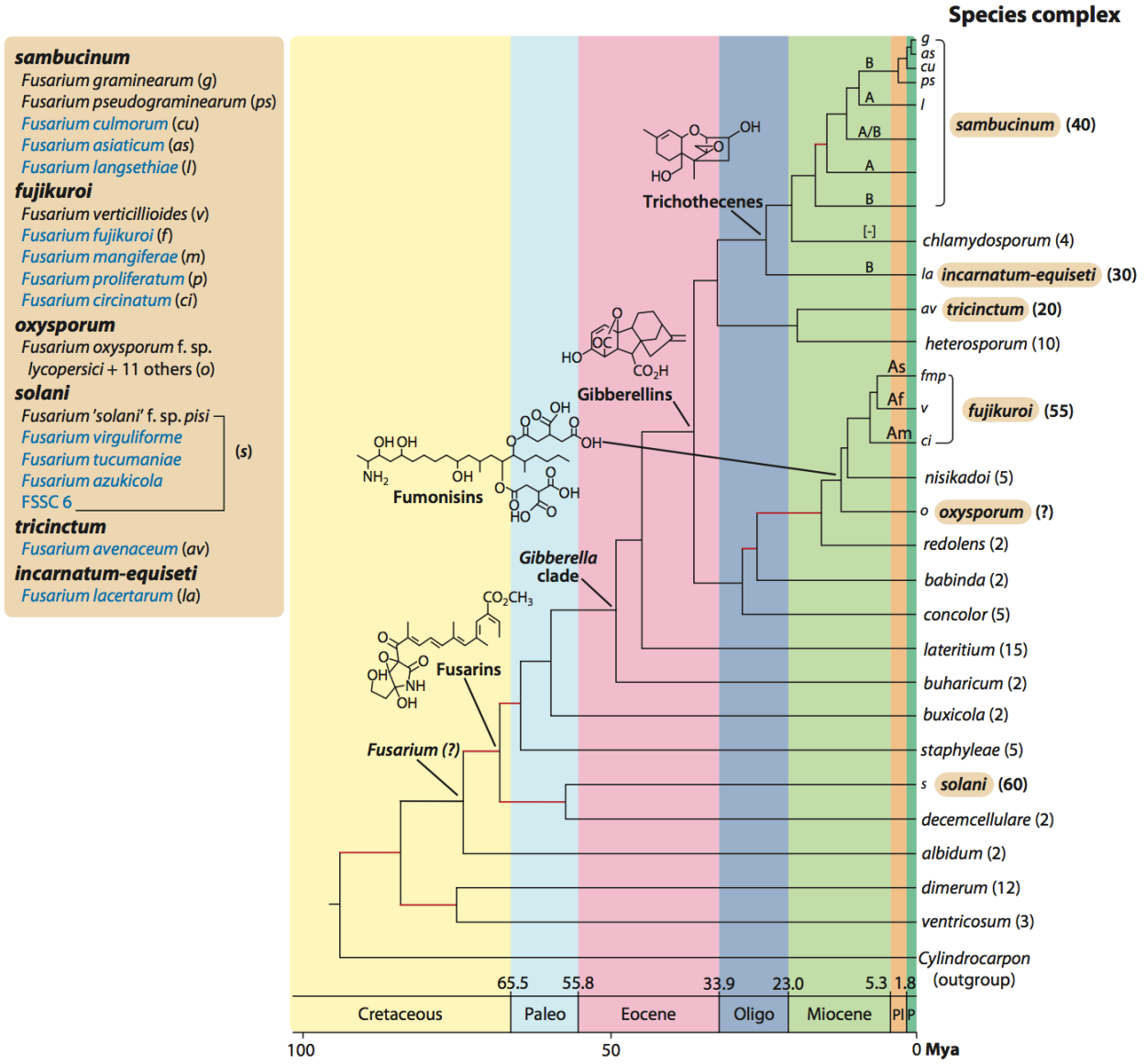
Overview



1. **Introduction to my study system**
    1. ***Fusarium***
    2. Background on the translation elongation factor 1-alpha locus
2. Explanation of oligotyping: getting more from your amplicons
    1. Shannon's Entropy
    2. Oligotyping
3. Pipeline & performance on mock communities
    1. vsearch – **inter**specific relative abundance
    2. Oligotyping – **intra**specific relative abundance
4. Where you can learn more

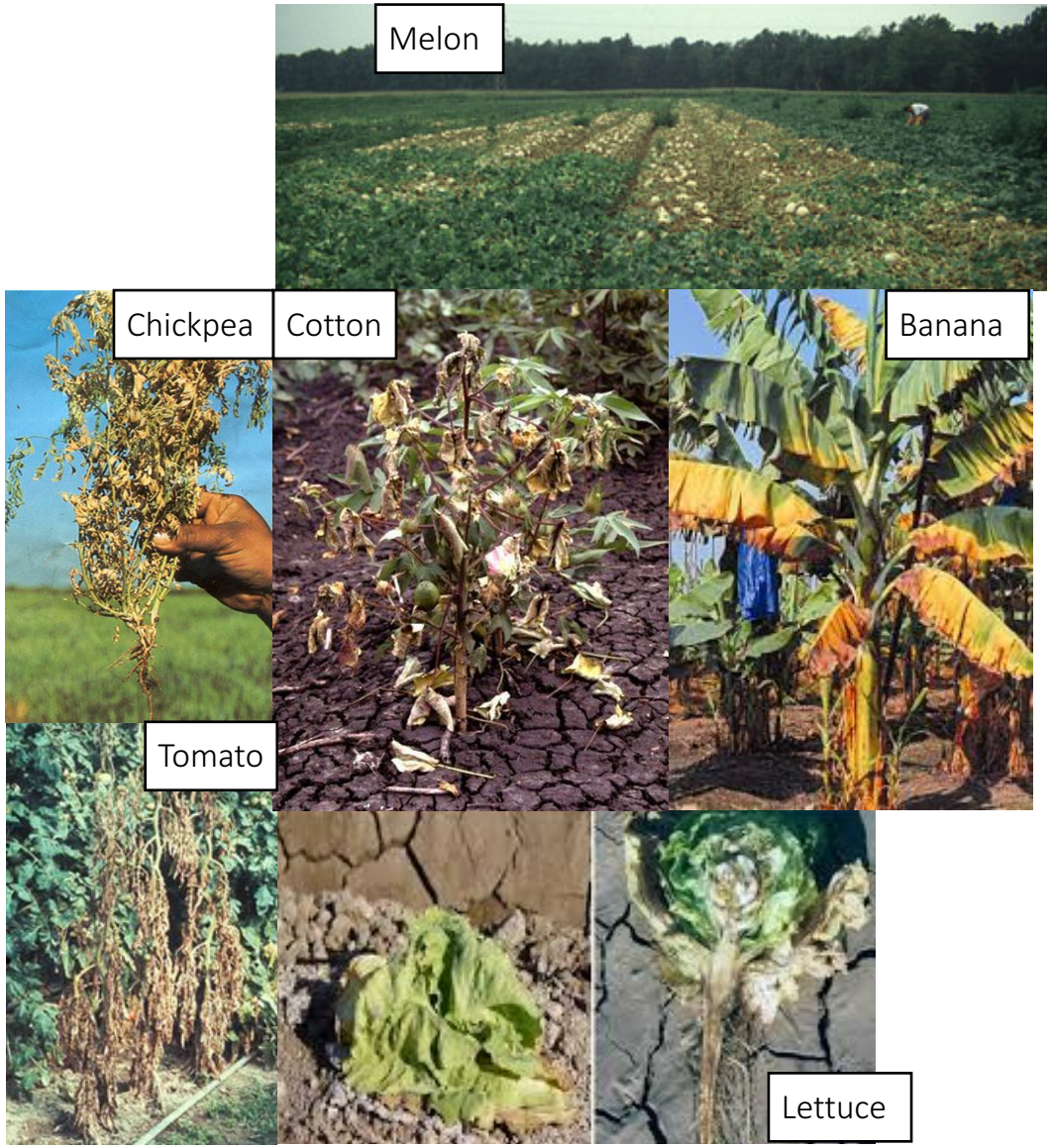
Fusarium

- A genus of Fungi containing:
  - Many economically important plant pathogens
  - Some human pathogens
  - Soil/plant-associated species



Fusarium oxysporum

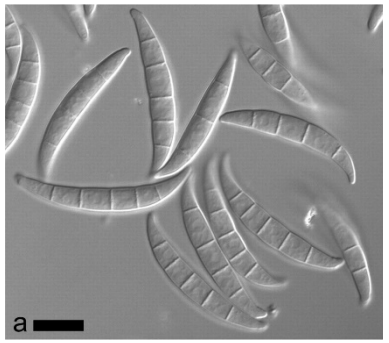
- Entire species = broad host range
  Over 70 reported hosts, primarily vegetable crops
- Individual strains = narrow host range
  Phenotypic characterization based on preferred host: “formae speciales”
- Characteristics:
  - Ubiquitous in soils across the globe
  - Highly diverse:
    - **Most strains are non-pathogenic**
    - Non-pathogenic strains can have disease suppressive effects
  - Main niche: root endophyte, primary consumer of dead plant tissue



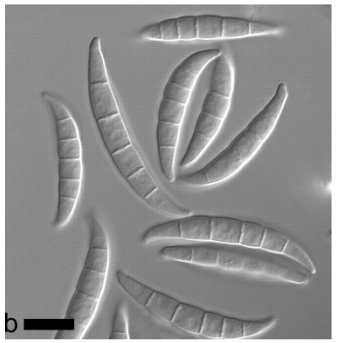
Fusarium oxysporum

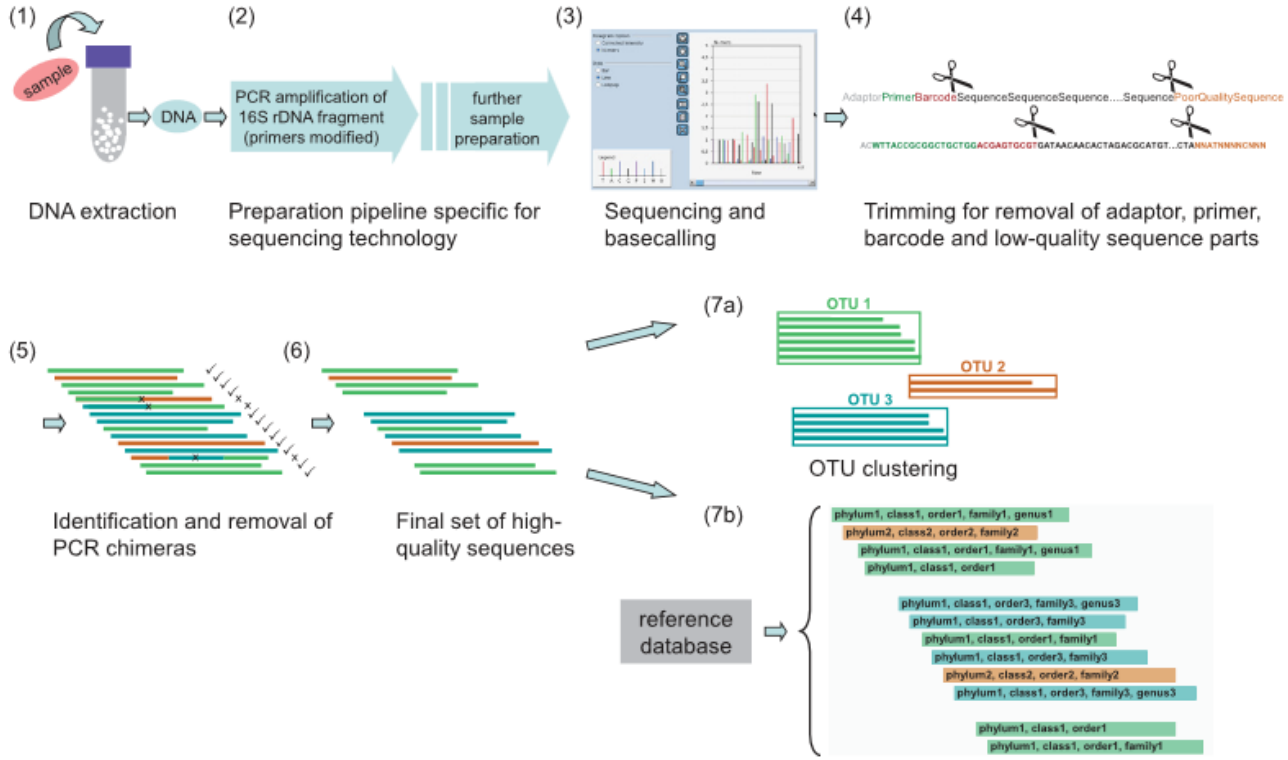
- What is the extent of diversity in soils?
- What environments are more diverse?
- What influences persistence in soil?
- How can we manage agricultural environments to decrease persistence of pathogenic strains?

We need a high-throughput tool to assess changes in relative abundance of *Fusarium oxysporum* strains…

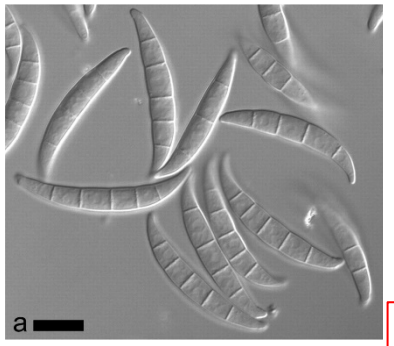


Introduction

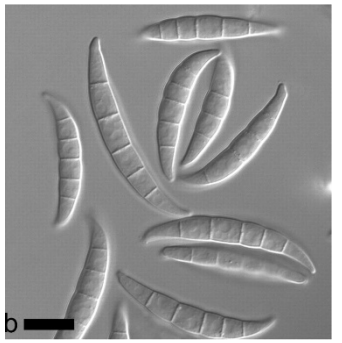




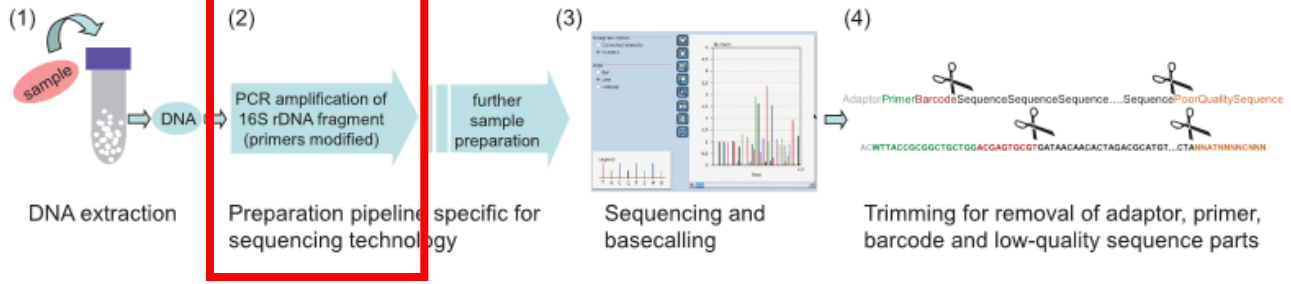
- I want to monitor changes in relative abundance of *Fusarium oxysporum* sequence types
- Typical microbiome analyses of fungi cannot evaluate sub-species diversity

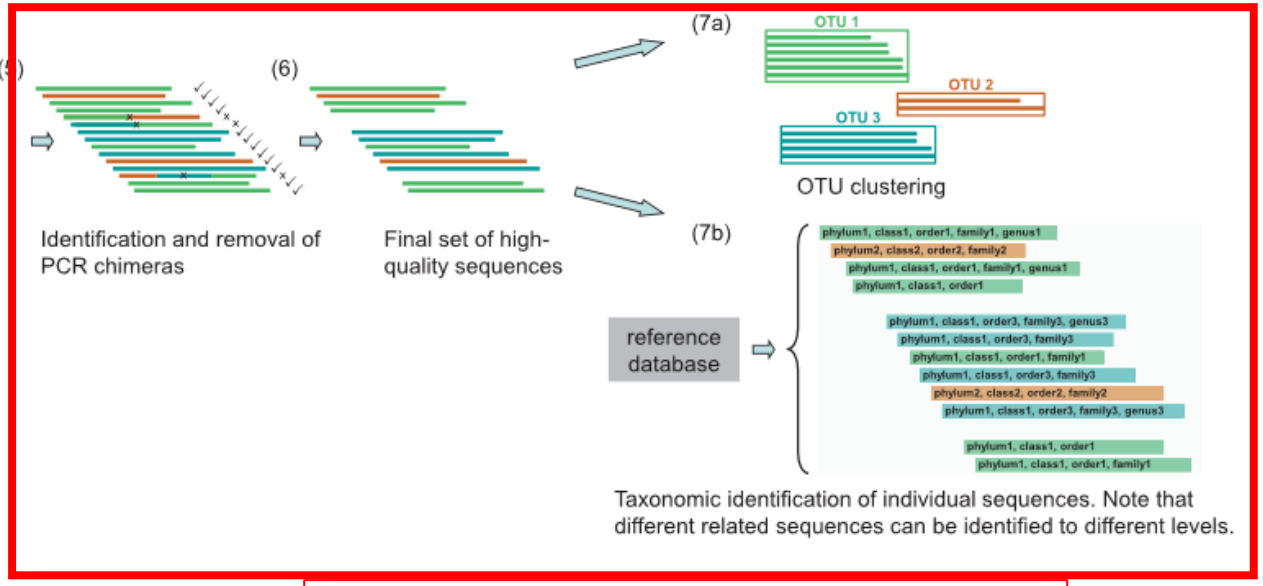


Introduction



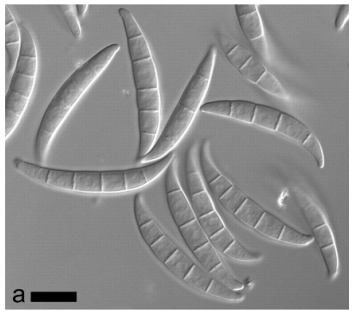
Target EF-1a with PCR



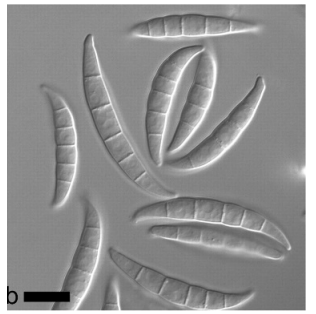


Customize data analysis for this locus

- I want to monitor changes in relative abundance of *Fusarium oxysporum* sequence types
- Typical microbiome analyses of fungi cannot evaluate sub-species diversity

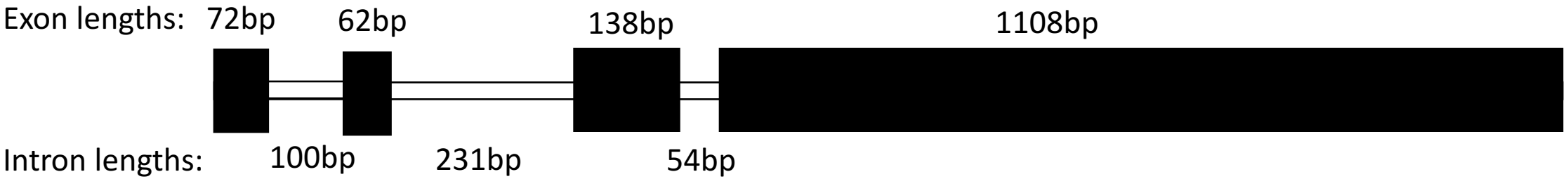


Overview



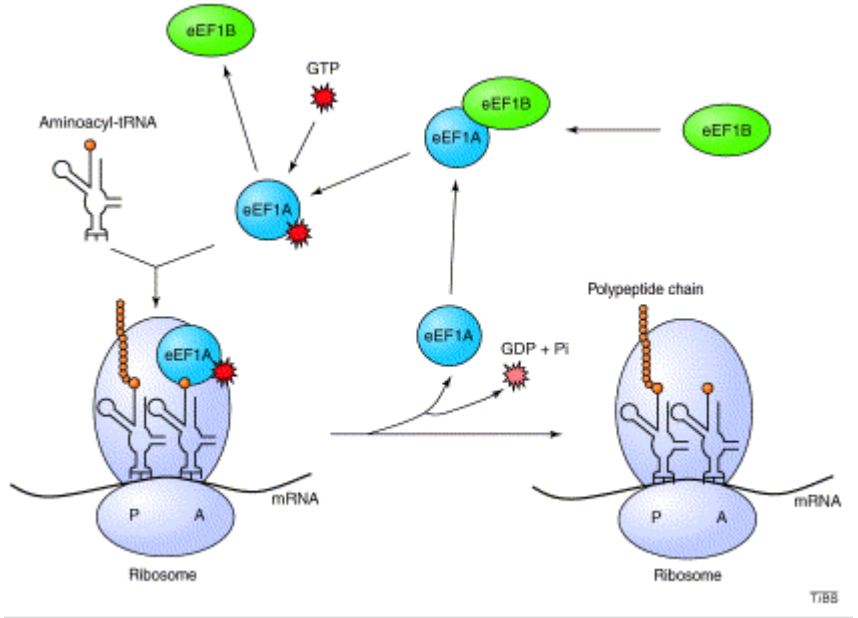
1. **Introduction to my study system**
    1. *Fusarium*
    2. **Background on the translation elongation factor 1-alpha locus**
2. Explanation of oligotyping: getting more from your amplicons
    1. Shannon's Entropy
    2. Oligotyping
3. Pipeline & performance on mock communities
    1. vsearch – **inter**specific relative abundance
    2. Oligotyping – **intra**specific relative abundance
4. Where you can learn more

Translation Elongation factor 1-alpha



- **Conserved, single copy gene in eukaryotes**

- Eukaryotes have two elongation factor genes.
- Elongation factor 1 has two subunits: alpha and beta-gamma
  - Elongation factor 1 alpha regulates entry of tRNAs into an available site on a ribosome
  - Elongation factor beta-gamma facilitates release of GDP from the alpha subunit



DOI: http://dx.doi.org/10.1016/j.tibs.2003.11.006

History of EF-1a use in phylogenetics

Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 7749–7754, July 1996
Evolution

The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny

SANDRA L. BALDAUF*†, JEFFREY D. PALMER‡, AND W. FORD DOOLITTLE*

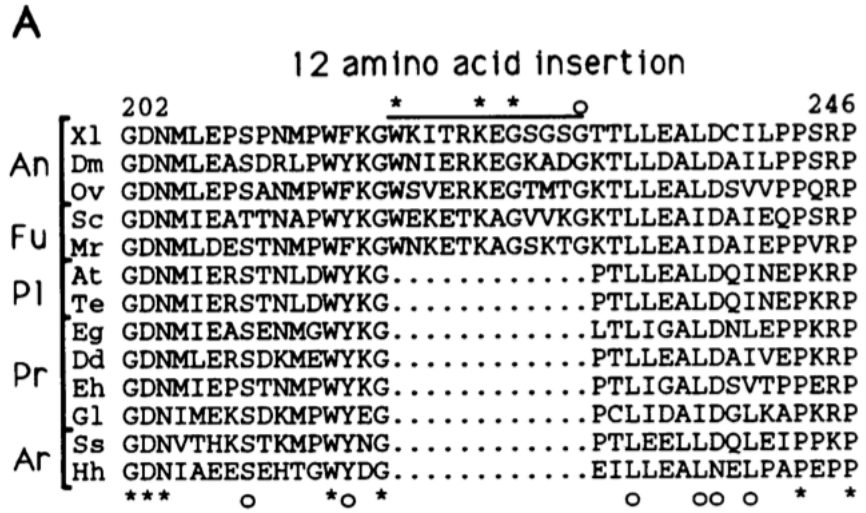
*Canadian Institute for Advanced Research and Department of Biochemistry, Dalhousie University, Halifax, NS B3H 4H7, Canada; and ‡Department of Biology, Indiana University, Bloomington, IN 47405

Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 11558–11562, December 1993
Evolution

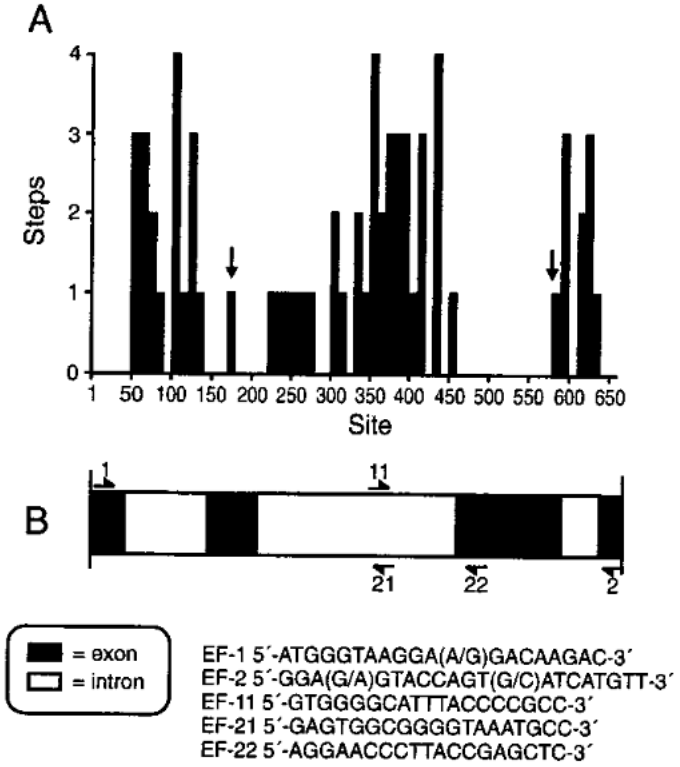
Animals and fungi are each other's closest relatives: Congruent evidence from multiple proteins

SANDRA L. BALDAUF* AND JEFFREY D. PALMER†

*Institute for Marine Biosciences, National Research Council of Canada, Halifax, NS, Canada, B3H 3Z1; and †Department of Biology, Indiana University, Bloomington, IN 47405



History of EF-1a use in phylogenetics



Proc. Natl. Acad. Sci. USA
Vol. 95, pp. 2044–2049, March 1998
Applied Biological Sciences

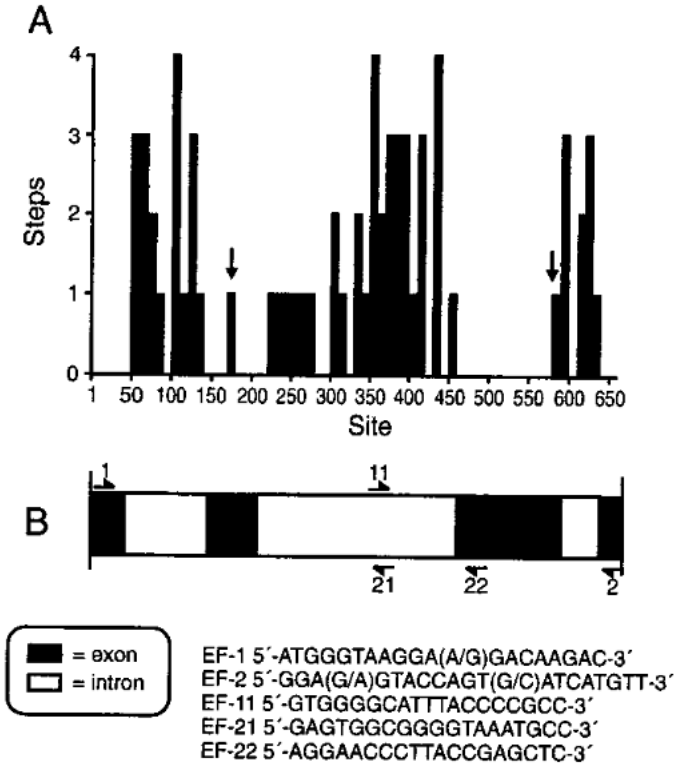
Multiple evolutionary origins of the fungus causing Panama disease of banana: Concordant evidence from nuclear and mitochondrial gene genealogies

KERRY O'DONNELL*†‡, H. CORBY KISTLER†§, ELIZABETH CIGELNIK*, AND RANDY C. PLOETZ¶

*National Center for Agricultural Utilization Research, U.S. Department of Agriculture–Agricultural Research Service, 1815 North University Street, Peoria, IL 61604; §Department of Plant Pathology, University of Florida, Gainesville, FL 32611; and ¶University of Florida, Tropical Research and Education Center, 18905 SW 280th Street, Homestead, FL 33031-3314.

31). Results from the present study demonstrate that the EF-1α gene, with 95% of the signal derived from intron sequences, possesses 50% more phylogenetic information than the mtSSU rDNA. In sharp contrast to the exons, which lack indels and possess only two phylogenetically informative sites within the 36 taxon matrix, EF-1α introns appear to be under relaxed evolutionary constraints as inferred from the substitutional pattern (Fig. 1*A*) that includes indels. However, the low level of homoplasy observed, coupled with the concordance of EF-1α and mtSSU rDNA gene trees, indicate that informative sites within the EF-1α gene are not saturated.

History of EF-1a use in phylogenetics

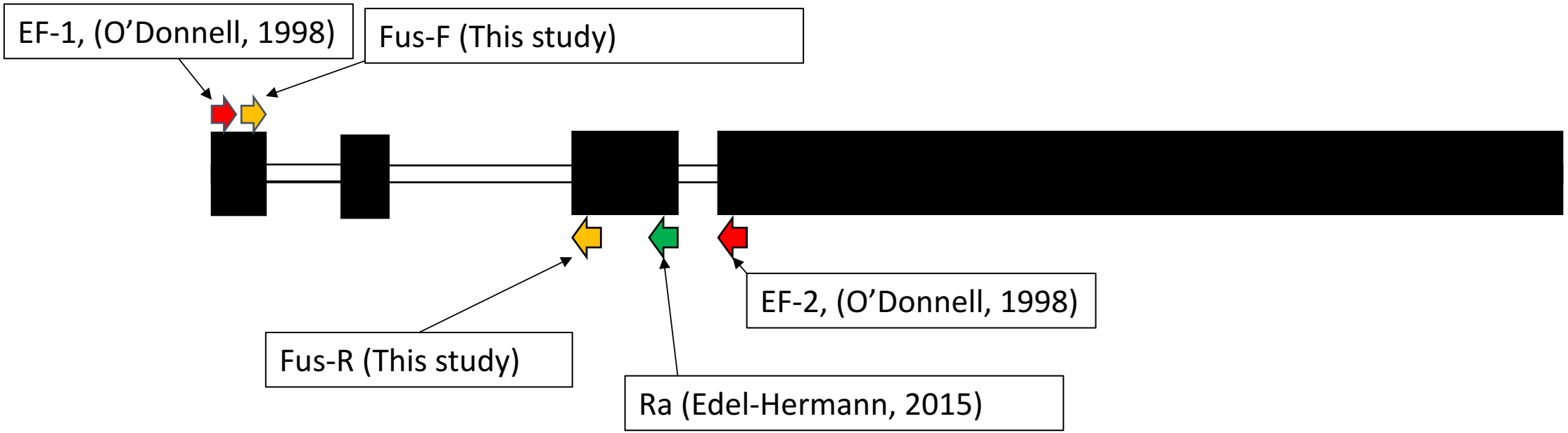


- Subsequent to O'Donnell et al. (1998), EF-1a became widely adopted for phylogenetic comparisons in *Fusarium*

- 2,725 sequences in my custom database:
  - 1,651 from Fusarium-ID database
  - 449 from NCBI Genbank
  - 625 from 14 recent publications

- **Final database = 718 unique EF-1a sequences**

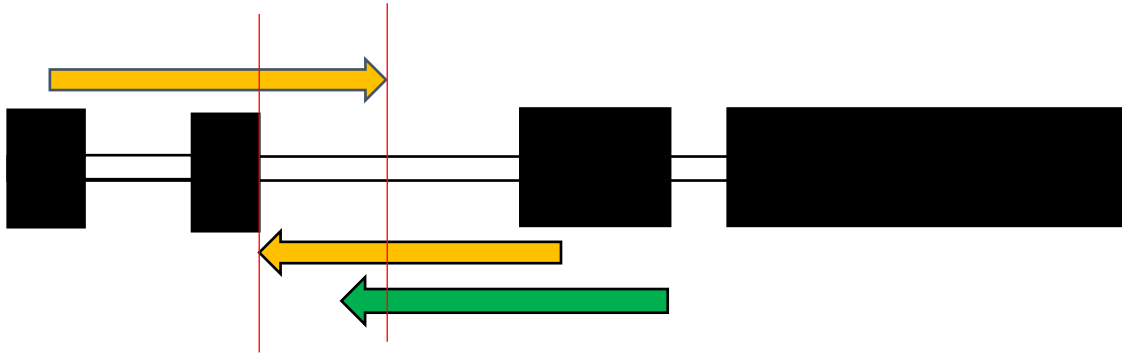
Primers

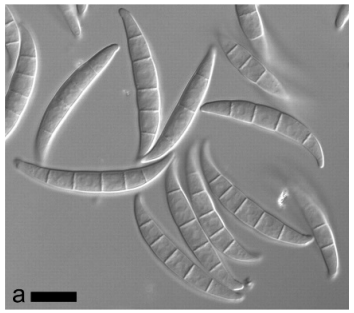


- Fa/Ra amplicon sizes (including primers):
  - *Fusarium oxysporum*: ~572bp
  - *Fusarium solani*: ~612 bp

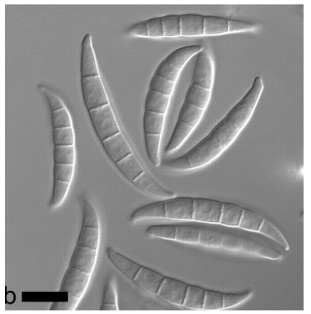
- Fus-F/Fus-R amplicon sizes (including primers):
  - *Fusarium oxysporum*: ~450bp
  - *Fusarium solani*: ~495bp

Illumina MiSeq 300bp paired end sequencing:





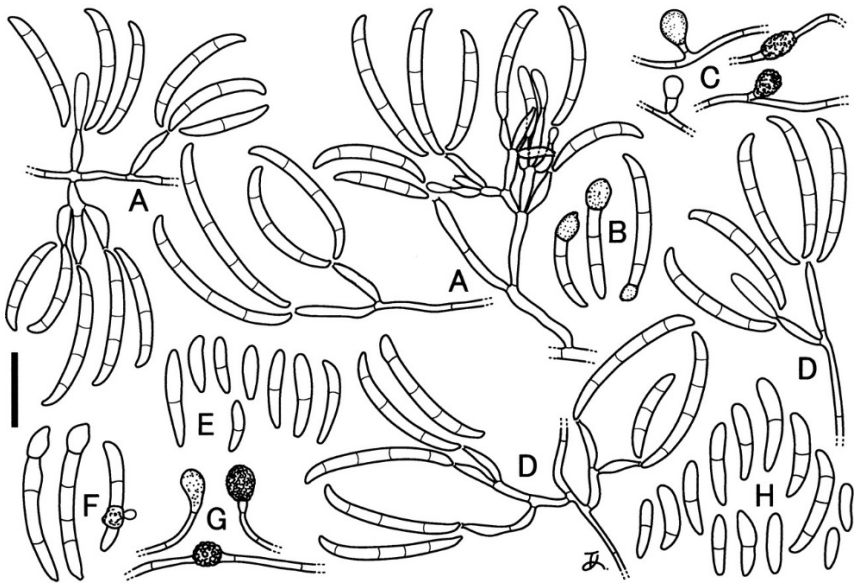
Overview



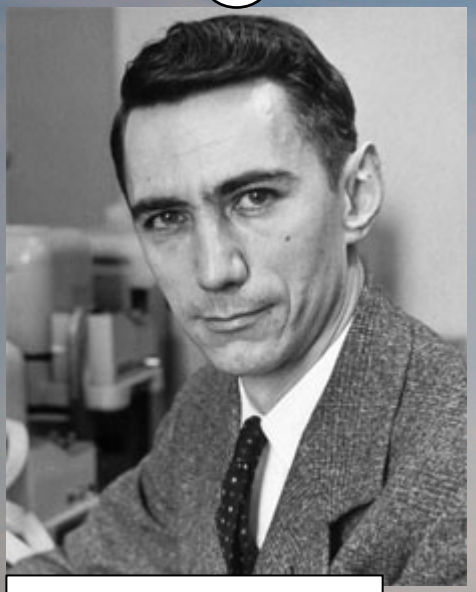
1. Introduction to my study system
    1. *Fusarium*
    2. Background on the translation elongation factor 1-alpha locus
2. **Explanation of oligotyping: getting more from your amplicons**
    1. **Shannon's Entropy**
    2. Oligotyping
3. Pipeline & performance on mock communities
    1. vsearch – **inter**specific relative abundance
    2. Oligotyping – **intra**specific relative abundance
4. Where you can learn more

Takeaway ideas

- **Oligotyping can extract additional information from your amplicons**
  - Analyzes positional entropy in aligned reads
- It is first necessary to classify reads
  - dbcAmplicons
  - Qiime
  - Mothur
- Reads corresponding to a single species or genus can be aligned and oligotyped



What is the relationship between predictability and information?

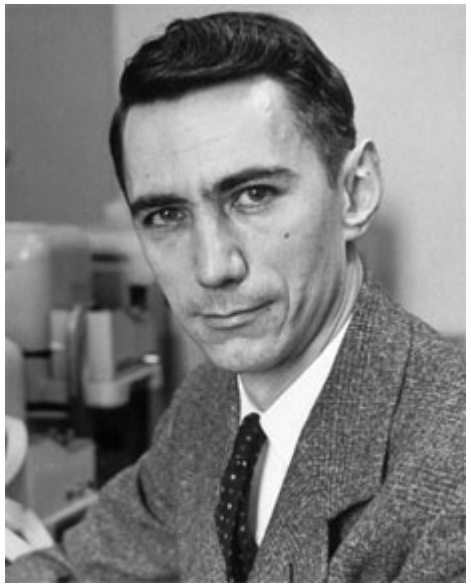


Claude Shannon

Shannon's entropy: *Uncertainty has greater information value*

$$H(x) = \sum_{i=1}^{n} -P(x_i) * \log_2(P(x_i))$$

- H(x) = total entropy
- P(x) – probability of "x"
- Example: A fair coin (50% heads, 50% tails)
  - f(x) for heads (or tails), = 0.5
    - Heads: $-0.5*\log_2(0.5) = -0.5*-1 = 0.5$
    - Tails: $-0.5*\log_2(0.5) = -0.5*-1 = 0.5$
  - H(x) = 0.5 + 0.5 = 1



Claude Shannon

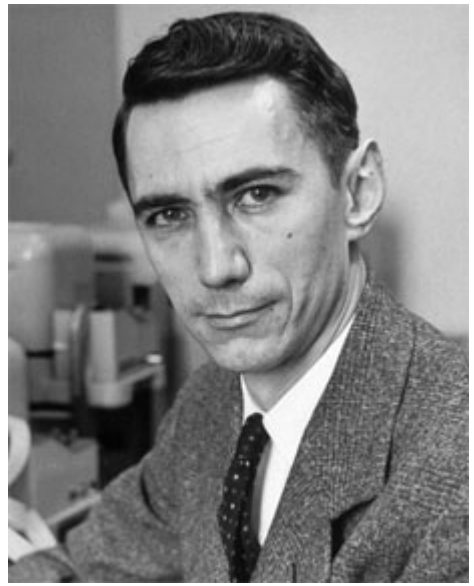
- **Equal probabilities for all possibilities have the highest Shannon entropy**

Shannon's entropy:
Uncertainty has greater information value

$$H(x) = \sum_{i=1}^{n} -P(x_i) * \log_2(P(x_i))$$

- H(x) = total entropy
- P(x) – probability of "x"
- Example: An unfair coin (20% heads, 80% tails)
  - Heads: $-0.2*\log_2(0.2) = 0.46$
  - Tails: $-0.8*\log_2(0.8) = 0.25$
- H(x) = 0.46 + 0.25 = 0.71

- **Unequal probabilities have lower Shannon entropy**

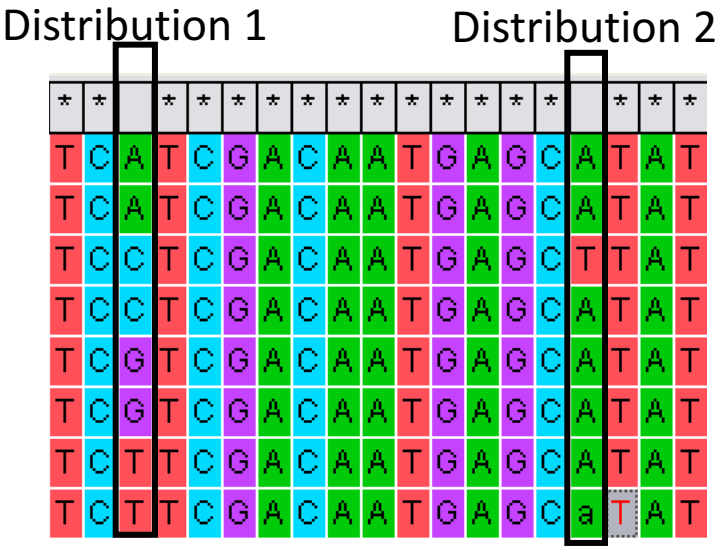


Claude Shannon

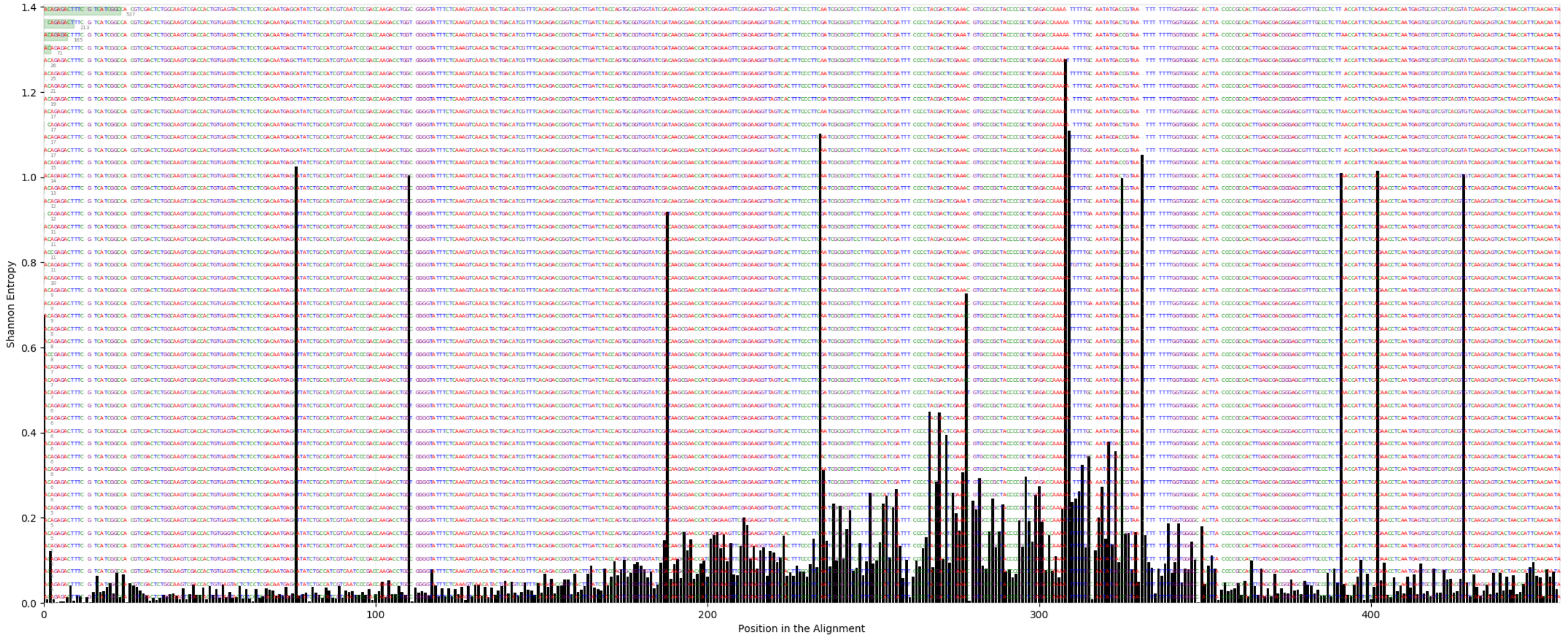
Shannon's entropy:
Uncertainty has greater information value

$$H(x) = \sum_{i=1}^{n} -P(x_i) * \log_2(P(x_i))$$

- **What about DNA?**
- *Distribution 1:*
  - -0.25*$\log_2$(0.25) = 0.5
  - -0.25*$\log_2$(0.25) = 0.5
  - -0.25*$\log_2$(0.25) = 0.5
  - -0.25*$\log_2$(0.25) = 0.5
  - H(x) = 2
- *Distribution 2:*
  - -0.875*$\log_2$(0.875) = 0.17
  - -0.125*$\log_2$(0.125) = 0.38
  - H(x) = 0.55
- *Distribution 3 (not shown):*
  - -0.998*$\log_2$(0.998) = 0.003
  - -0.002*$\log_2$(0.002) = 0.017
  - H(x) = 0.02



Shannon’s entropy for every position in an alignment

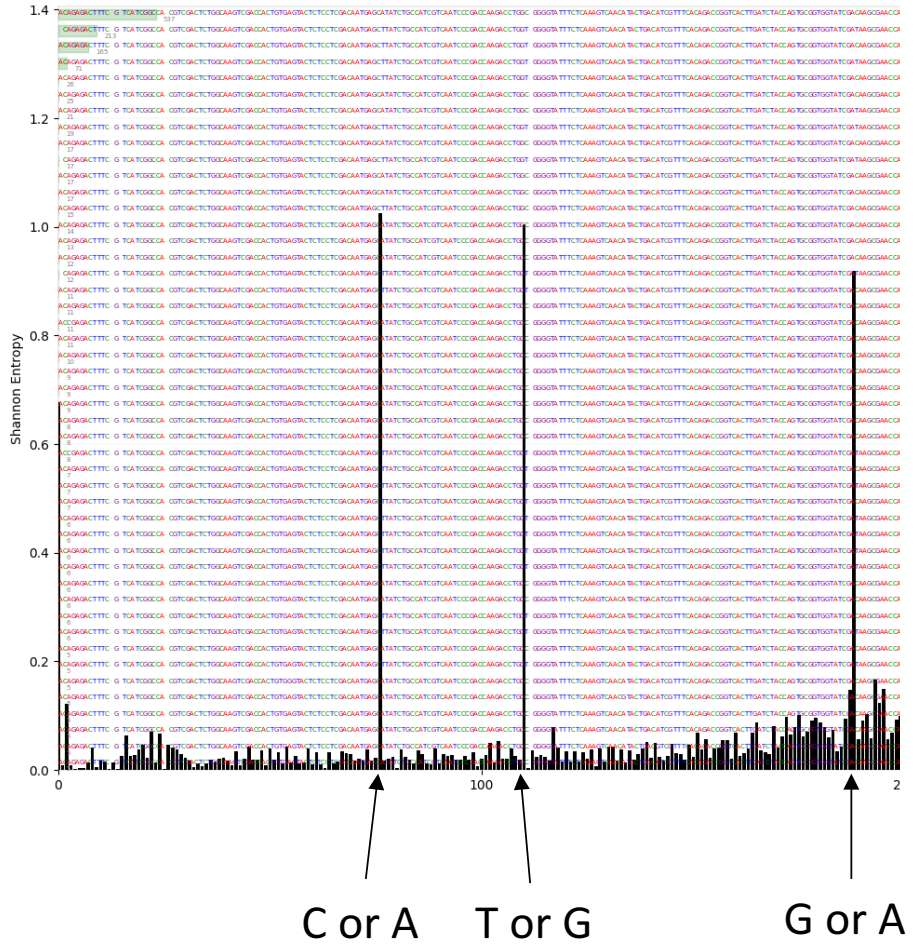


The Entropy to Oligotype continuum

1. Determine entropy for every position
2. Compare nucleotides at entropic positions for every read
    1. CTA = 66%
    2. CGG = 20%
    3. AGA = 13%

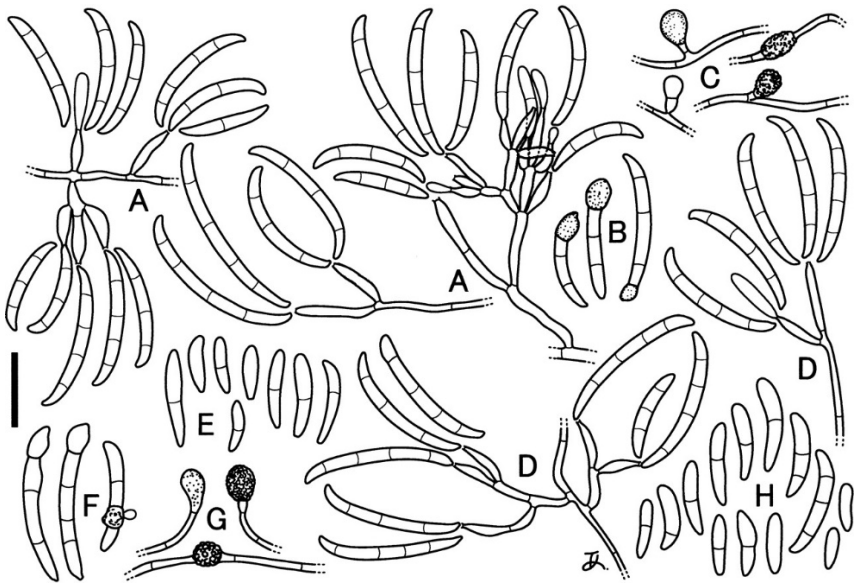
    Min. abundance threshold

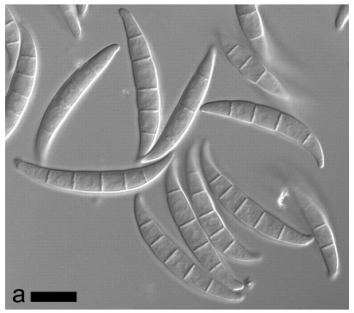
    4. AGG = 0.5%
    5. CGA = 0.5%
3. Set minimum abundance threshold
4. Estimate relative abundances of oligotypes
5. Reconstruct sequences for each oligotype



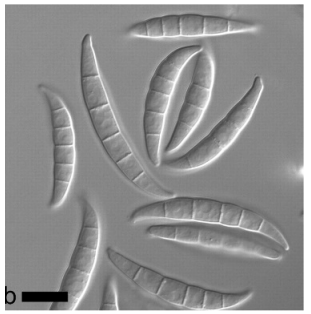
Takeaway ideas

- **Oligotyping can extract additional information from your amplicons**
  - Analyzes positional entropy in aligned reads
- It is first necessary to classify reads
  - dbcAmplicons
  - Qiime
  - Mothur
- Reads corresponding to a single species or genus can be aligned and oligotyped





Overview

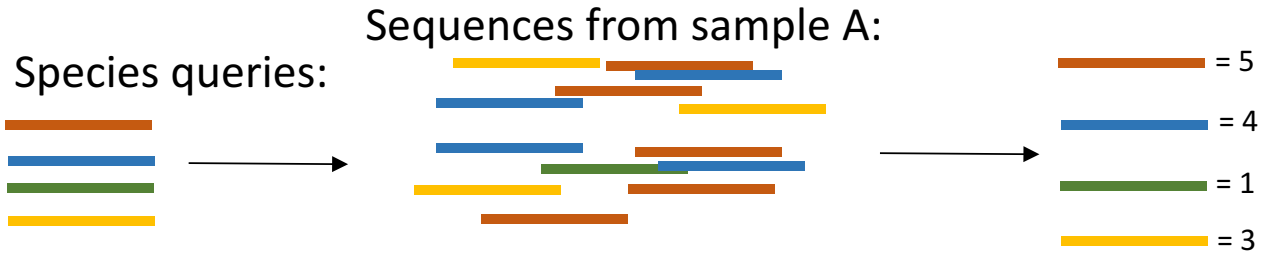


1. Introduction to my study system
    1. *Fusarium*
    2. Background on the translation elongation factor 1-alpha locus
2. Explanation of oligotyping: getting more from your amplicons
    1. Shannon's Entropy
    2. Oligotyping
3. **Pipeline & performance on mock communities**
    1. vsearch – **inter**specific relative abundance
    2. Oligotyping – **intra**specific relative abundance
4. Where you can learn more

Pipeline overview

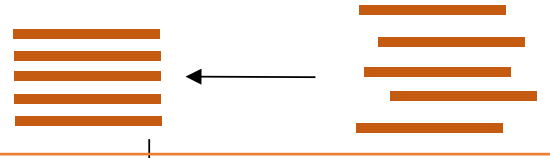
- Bin sequences by barcode:
  - Join paired-end reads (pear)
  - Split reads by barcode (fastx_toolkit)
  - Convert fastq to fasta format (fastx_toolkit)
- Count reads for each species
  - Query samples with species-specific vsearch parameters
- Gather subspecies diversity
  - Extract sequences from a given species (e.g. *F. oxysporum*)
  - Align extracted sequences (muscle)
  - Trim primer/barcode regions manually (MEGA7)
  - Conduct entropy analysis (oligotype)
  - Generate oligotypes (oligotype)

Count reads for each *Fusarium* species:

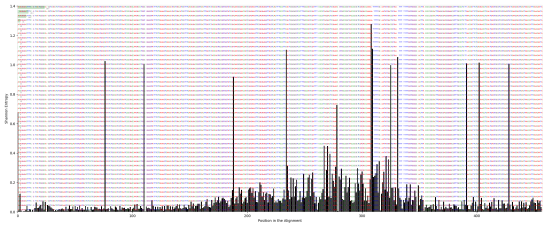


vsearch: obtains read counts per species by % identity

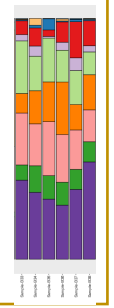
Align/analyze entropy for all reads of a species:



muscle:



entropy-analysis:



oligotyping: sequence type relative abundance

- I didn't write any programs used in this pipeline.

PeerJ

VSEARCH: a versatile open source tool for metagenomics

Torbjørn Rognes[1,2], Tomáš Flouri[3,4], Ben Nichols[5], Christopher Quince[5,6] and Frédéric Mahé[7,8]

1792–1797 *Nucleic Acids Research, 2004, Vol. 32, No. 5*
DOI: 10.1093/nar/gkh340

MUSCLE: multiple sequence alignment with high accuracy and high throughput

Robert C. Edgar*

Methods in Ecology and Evolution

British Ecological Society

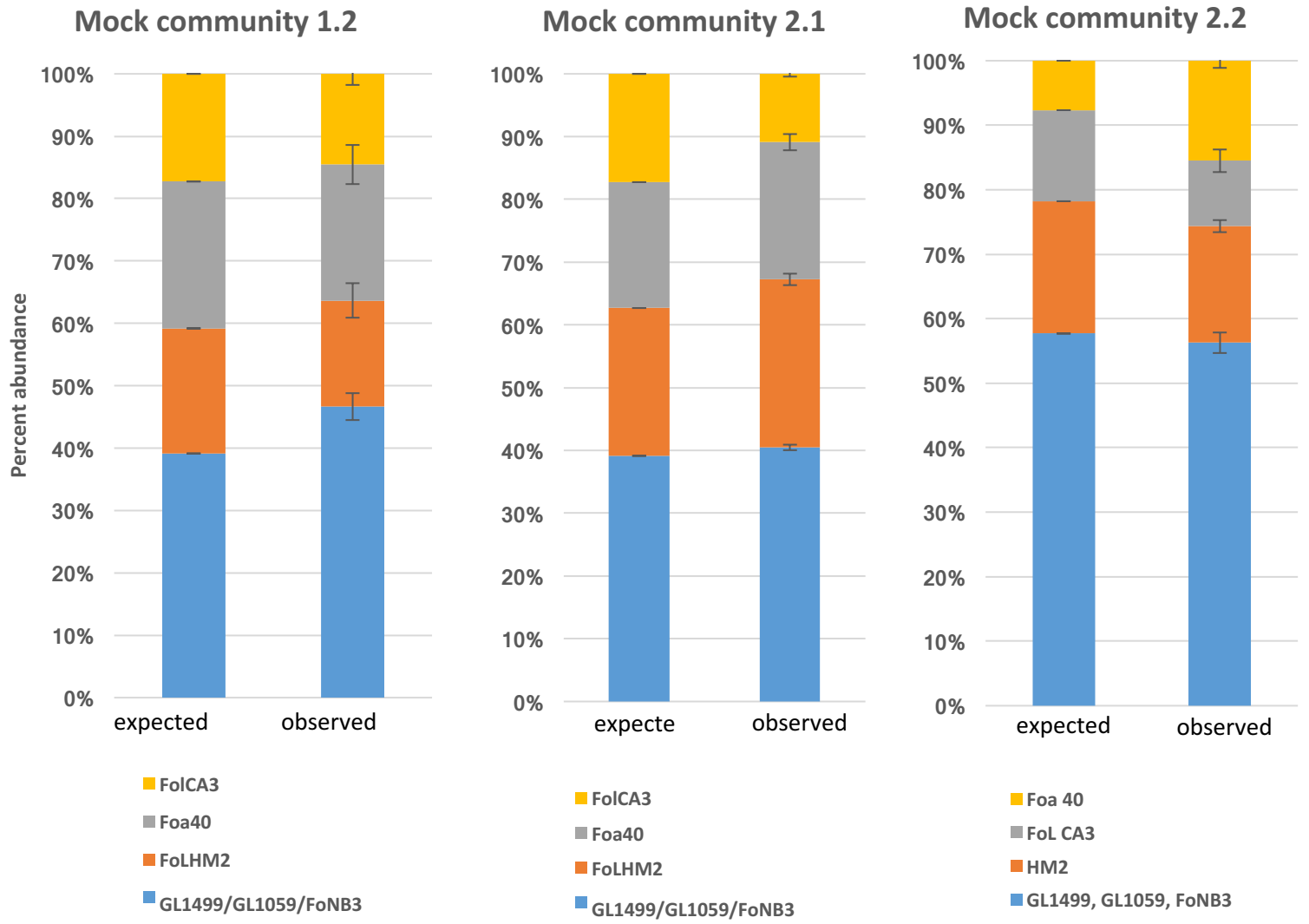
Methods in Ecology and Evolution 2013, **4**, 1111–1119 doi: 10.1111/2041-210X.12114

Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data

A. Murat Eren*, Loïs Maignien, Woo Jun Sul, Leslie G. Murphy, Sharon L. Grim, Hilary G. Morrison and Mitchell L. Sogin

Estimating relative abundance: Oligotyping

- Mock communities:
  - 4 EF-1a sequence types
  - *Fusarium oxysporum* only
- Expected abundance calculated based on DNA concentration
- Averages taken from two technical reps
- **Sequences are identical to the input**



Where you can learn more

- **Eren lab website:**
  - http://merenlab.org/software/oligotyping/
- **Frontiers in Microbiology Research topic**: New insights into microbial ecology through subtle nucleotide variation
  - http://journal.frontiersin.org/researchtopic/2427/new-insights-into-microbial-ecology-through-subtle-nucleotide-variation