Analysis and assembly methods for microbiome sequencing data

Marcus Fedarko

UCSDCSE
Computer Science and Engineering

UC San Diego

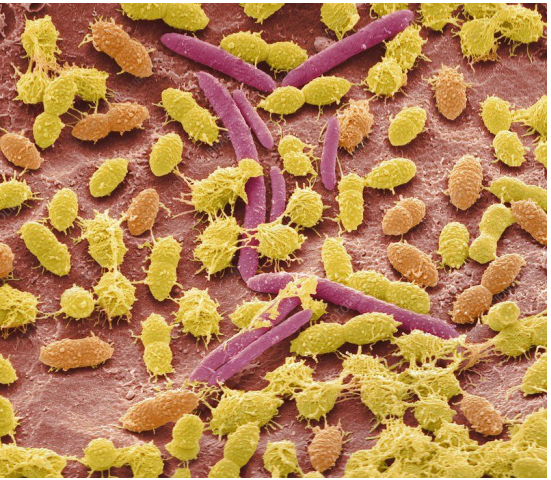
Analysis and assembly methods for **microbiome** sequencing data

Marcus Fedarko

UCSDCSE
Computer Science and Engineering

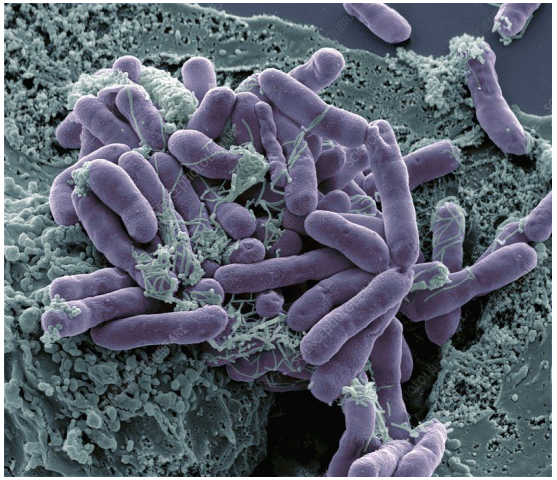
UC San Diego

Microbiomes



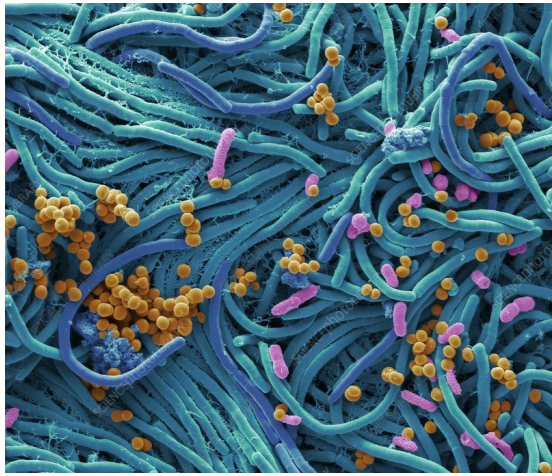
Steve Gschmeissner, Science Photo Library

“Scanning electron micrograph (SEM) of bacteria cultured from a sample of human faeces.”



Steve Gschmeissner, Science Photo Library

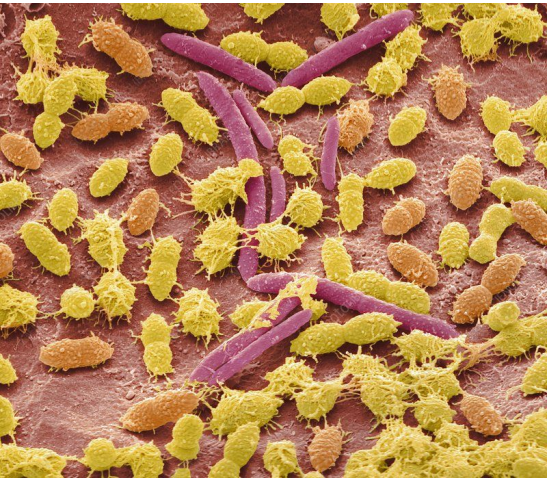
“Bacterial contamination, coloured scanning electron micrograph (SEM). *Escherichia coli* bacteria in a cell culture. This contamination has come from an unclean water source.”



Steve Gschmeissner, Science Photo Library

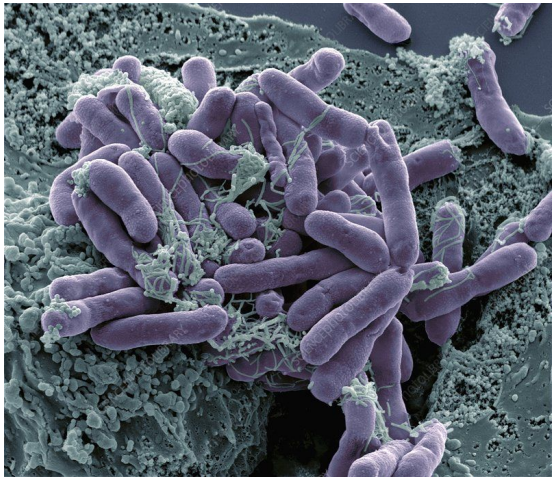
“Coloured scanning electron micrograph (SEM) of bacteria cultured from a mobile phone.”

Microbiomes



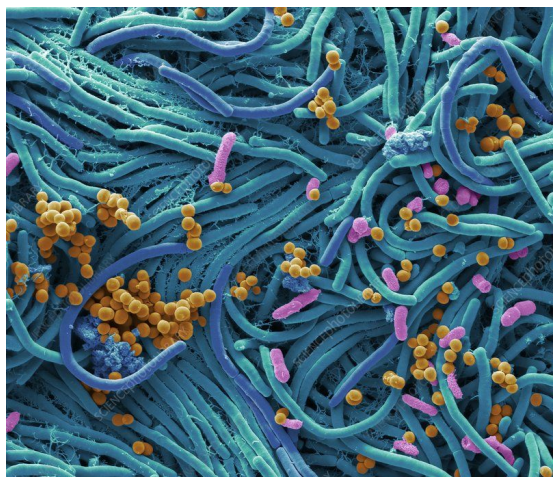
Steve Gschmeissner, Science Photo Library

“Scanning electron micrograph (SEM) of bacteria cultured from a sample of human faeces.”



Steve Gschmeissner, Science Photo Library

“Bacterial contamination, coloured scanning electron micrograph (SEM). *Escherichia coli* bacteria in a cell culture. This contamination has come from an unclean water source.”



Steve Gschmeissner, Science Photo Library

“Coloured scanning electron micrograph (SEM) of bacteria cultured from a mobile phone.”

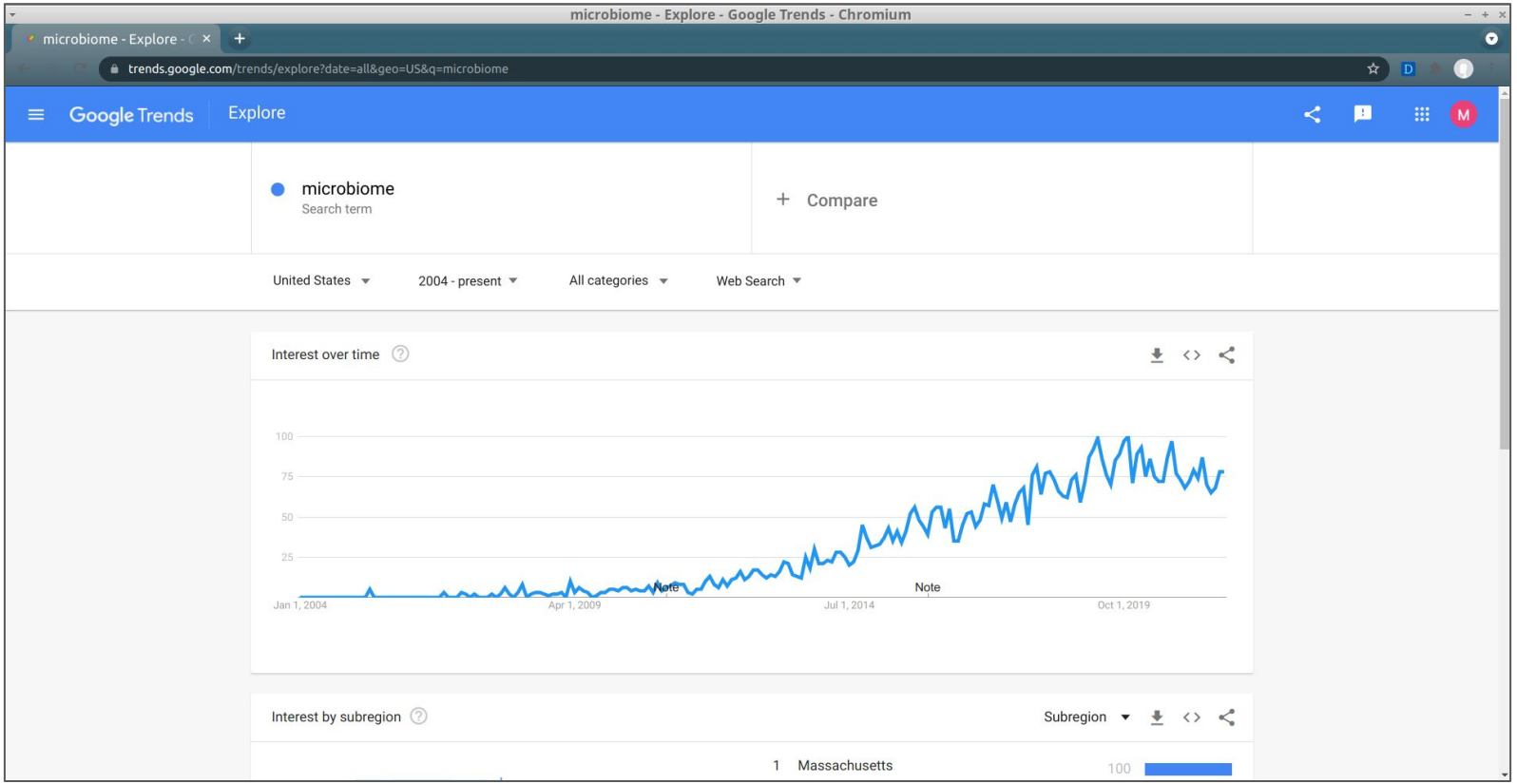
This talk

1. Introduction: Studying microbiomes
    a. Why bother?
    b. Why hasn't this research been more useful?
    c. Defining a goal for this talk
2. Culture-independent (a.k.a. sequencing-based) methods
    a. Marker gene sequencing
    b. Metagenome sequencing
3. Metagenome assembly
    a. Input (*reads*)
    b. Outputs (*contigs*, an *assembly graph*, …)
    c. Methods (*de novo* vs. reference-based, overlap graph vs. de Bruijn graph, …)
4. Future work: Solving the *strain separation problem*

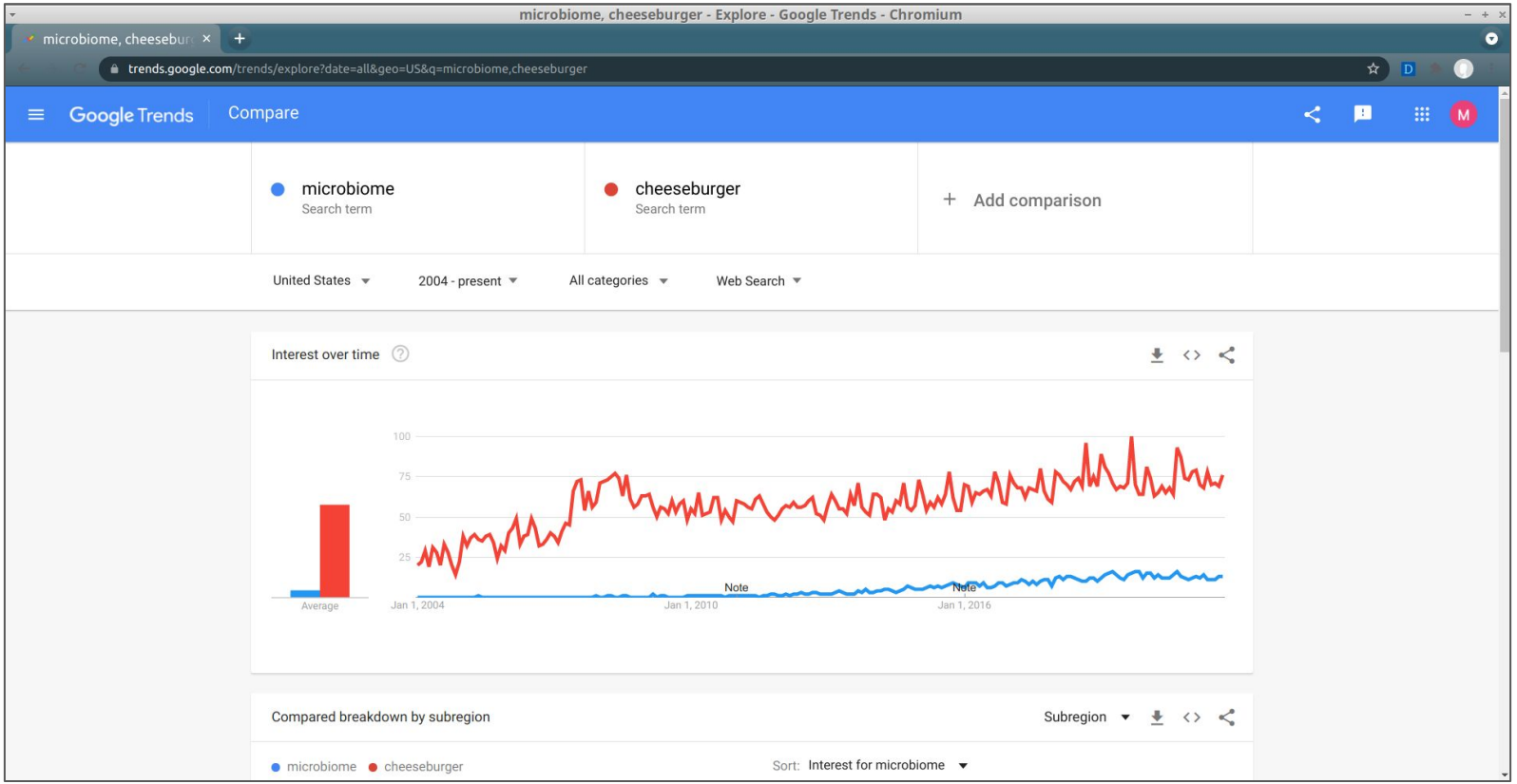
This talk

1. **Introduction: Studying microbiomes**
   a. Why bother?
   b. Why hasn't this research been more useful?
   c. Defining a goal for this talk
2. Culture-independent (a.k.a. sequencing-based) methods
   a. Marker gene sequencing
   b. Metagenome sequencing
3. Metagenome assembly
   a. Input (*reads*)
   b. Outputs (*contigs*, an *assembly graph*, …)
   c. Methods (*de novo* vs. reference-based, overlap graph vs. de Bruijn graph, …)
4. Future work: Solving the *strain separation problem*

Introduction: Why bother studying microbiomes?



Introduction: Why bother studying microbiomes?



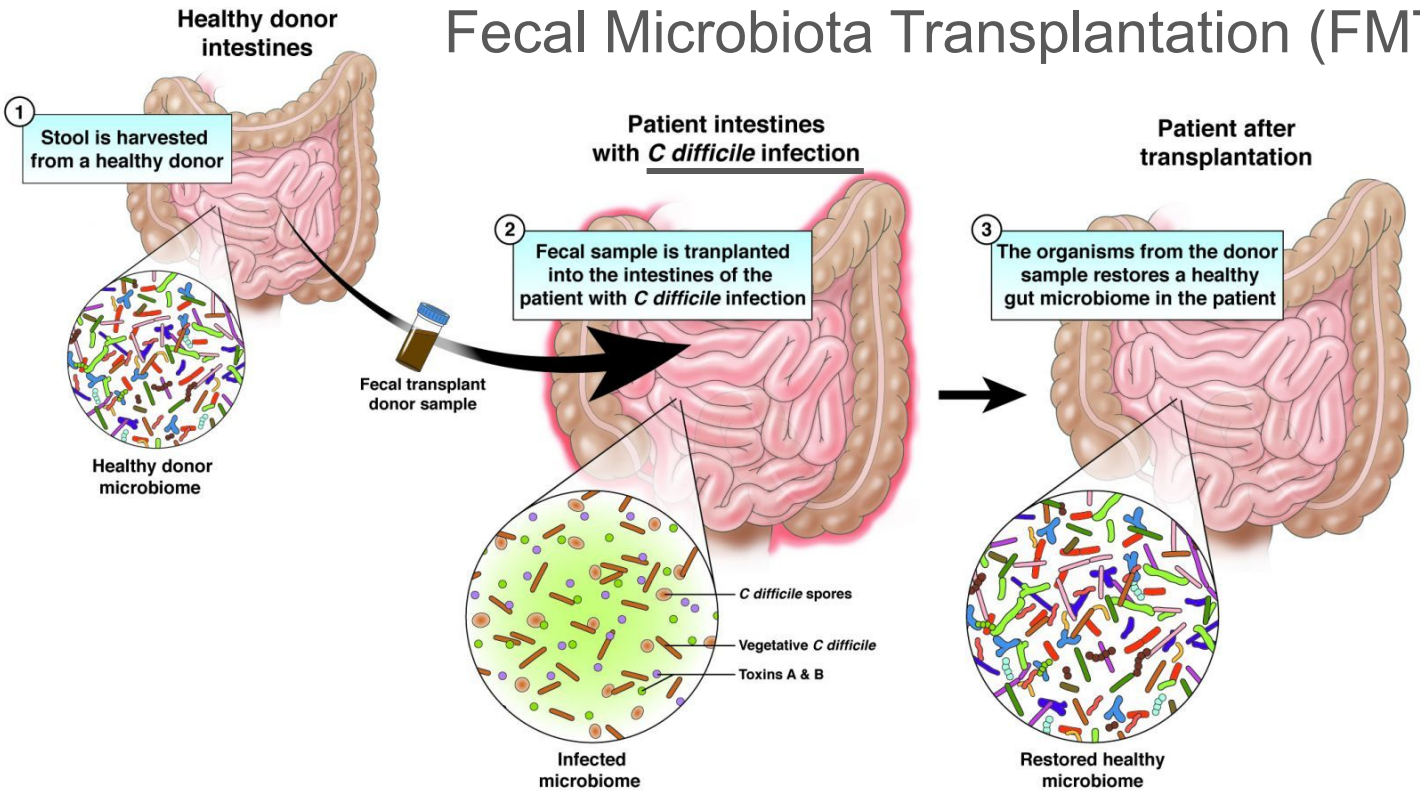
Introduction: Why bother studying microbiomes?

"During the Eastern Jin dynasty (AD 300–400 years), '*Zhou Hou Bei Ji Fang*', a well-known monograph of traditional Chinese medicine (TCM) written by Hong Ge, recorded a case of **treating patients with food poisoning or severe diarrhea by ingesting human fecal suspension** (known as yellow soup or Huang-Long decoction)."

H. Du, T.-t. Kuang, S. Qiu, T. Xu, C.-L. G. Huan, G. Fan, and Y. Zhang. Fecal medicines used in traditional medical system of China: a systematic review of their names, original species, traditional uses, and modern investigations. *Chinese medicine, 14*(1):1–16, 2019.

American Gastroenterological Association. Fecal Microbiota Transplanation (FMT): Overview. https://gastro.org/practice-guidance/gi-patient-center/topic/fecal-microbiota-transplantation-fmt/.

Introduction: Why bother studying microbiomes?



Introduction: Why bother studying microbiomes?

aga

Clinical Gastroenterology and Hepatology

Volume 9, Issue 12, December 2011, Pages 1044-1049

Perspective

Treating *Clostridium difficile* Infection With Fecal Microbiota Transplantation

Johan S. Bakken *, Thomas Borody ‡, Lawrence J. Brandt §, Joel V. Brill ||, Daniel C. Demarco ¶, Marc Alaric Franzos #, Colleen Kelly **, Alexander Khoruts ‡‡, Thomas Louie §§, Lawrence P. Martinelli |||, Thomas A. Moore ¶¶ ✉, George Russell ###, Christina Surawicz ***, Fecal Microbiota Transplantation Workgroup

ALIMENTARY TRACT: ORIGINAL ARTICLES

Fecal Microbiota Transplantation for the Treatment of *Clostridium difficile* Infection

A Systematic Review

Cammarota, Giovanni MD; Ianiro, Gianluca MD; Gasbarrini, Antonio MD

Fecal Microbiota Transplant for Treatment of *Clostridium difficile* Infection in Immunocompromised Patients

Colleen R. Kelly, MD, FACG,[1] Chioma Ihunnah, MD, MPH,[1] Monika Fischer, MD, MSCR,[2] Alexander Khoruts, MD,[3] Christina Surawicz, MD, MACG,[4] Anita Afzali, MD, MPH,[4] Olga Aroniadis, MD,[5] Amy Barto, MD,[6] Thomas Borody, MD, PhD, FACG,[7] Andrea Giovanelli, BS,[8] Shelley Gordon, MD, PhD,[9] Michael Gluck, MD,[10] Elizabeth L. Hohmann, MD,[11] Dina Kao, MD,[12] John Y. Kao, MD,[13] Daniel P. McQuillen, MD,[6] Mark Mellow, MD, FACG,[14] Kevin M. Rank, MD,[3] Krishna Rao, MD,[13] Arnab Ray, MD,[15] Margot A. Schwartz, MD, MPH,[10] Namita Singh, MD,[16] Neil Stollman, MD, FACG,[8] David L. Suskind, MD,[16] Stephen M. Vindigni, MD, MPH,[4] Ilan Youngster, MD,[11] and Lawrence Brandt, MD, MACG[5]

J Med Life. 2016 Apr-Jun; 9(2): 160–162.

PMCID: PMC4863507

PMID: 27453747

Fecal transplantation – the new, inexpensive, safe, and rapidly effective approach in the treatment of gastrointestinal tract diseases

R Oprita,* M Bratu,* B Oprita,* and B Diaconescu*

Preliminary Communication

Oral, Capsulized, Frozen Fecal Microbiota Transplantation for Relapsing *Clostridium difficile* Infection

Ilan Youngster, MD, MMSc; George H. Russell, MD, MSc; Christina Pindar, BA; Tomer Ziv-Baran, PhD; Jenny Sauk, MD; Elizabeth L. Hohmann, MD

Introduction: Why bother studying microbiomes?

ORIGINAL ARTICLE

Helicobacter pylori Infection and the Risk of Gastric Carcinoma

Julie Parsonnet, M.D., Gary D. Friedman, M.D., M.S., Daniel P. Vandersteen, Yuan Chang, M.D., Joseph H. Vogelman, D.E.E., Norman Orentreich, M.D., and Richard K. Sibley, M.D.

Infection with *Helicobacter pylori* Strains Possessing *cagA* Is Associated with an Increased Risk of Developing Adenocarcinoma of the Stomach[1]

Martin J. Blaser,[2] Guillermo I. Perez-Perez, Harry Kleanthous, Timothy L. Cover, Richard M. Peek, P. H. Chyou, Grant N. Stemmermann, and Abraham Nomura

Helicobacter pylori Persistence: an Overview of Interactions between *H. pylori* and Host Immune Defenses

Holly M. Scott Algood[1] and Timothy L. Cover[1,2,3,*]

The gastric microbiome, its interaction with *Helicobacter pylori*, and its potential role in the progression to stomach cancer

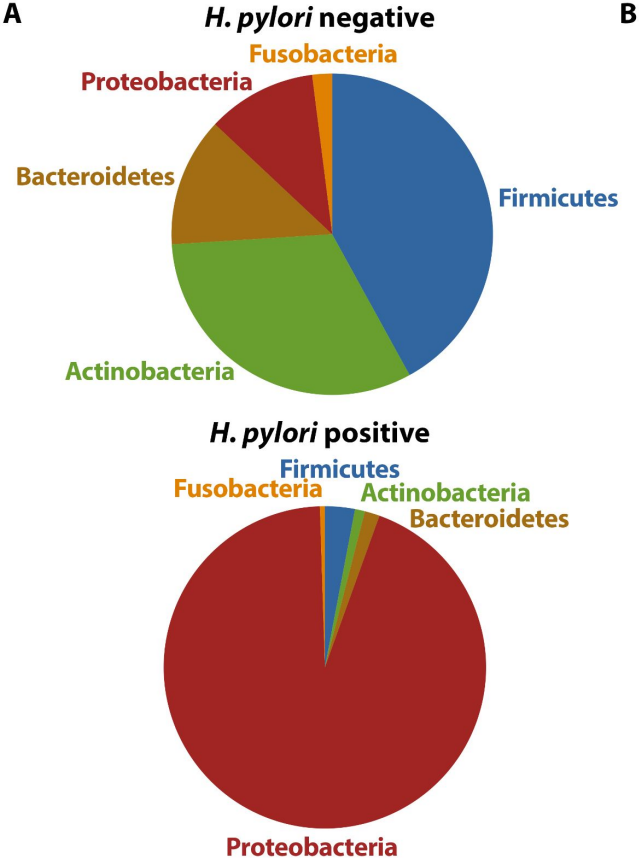
Jennifer M. Noto[1]*, Richard M. Peek, Jr.[1,2,3]

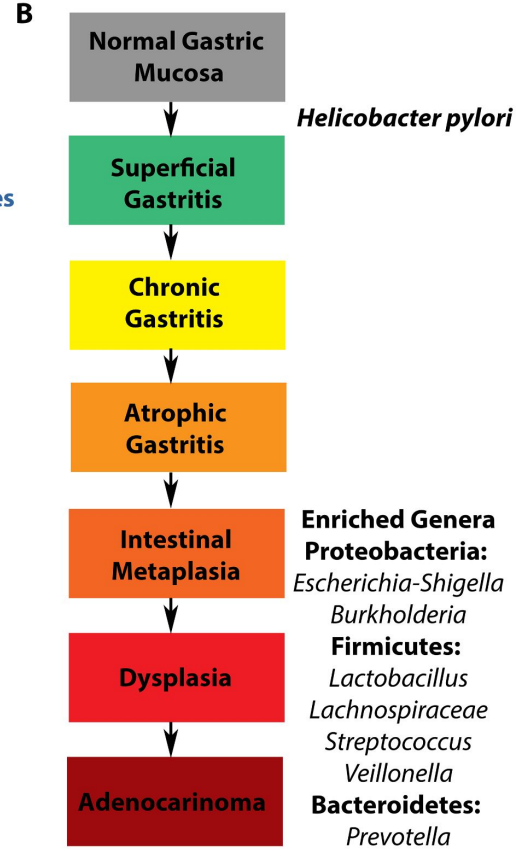
Introduction: Why bother studying microbiomes?

The gastric microbiome, its interaction with *Helicobacter pylori*, and its potential role in the progression to stomach cancer

Jennifer M. Noto[1]*, Richard M. Peek, Jr.[1,2,3]

Introduction: Why bother studying microbiomes?





> "*H. pylori*-negative individuals harbor a microbiota that is more complex and highly diverse compared to *H. pylori*-positive individuals. [...] Following infection with *H. pylori*, Proteobacteria and specifically *H. pylori* dominate the gastric microbiota. This leads to the development of chronic gastritis."

The gastric microbiome, its interaction with *Helicobacter pylori*, and its potential role in the progression to stomach cancer

Jennifer M. Noto[1]*, Richard M. Peek, Jr.[1,2,3]

Introduction: Why is it always *C. diff* and *H. pylori*?

"There are two well-documented diseases in the microbiome field that link a microbial biomarker with causation in disease: ***Helicobacter pylori*-associated peptic ulceration and gastric cancer** (Parsonnet et al., 1991) and ***Clostridium* (or *Clostridioides*) *difficile* infection-associated diarrhea** (Gupta et al., 2016)."

"However, causal inferences between complex microbiomes and other inflammatory, metabolic, neoplastic, and neuro-behavioral disorders have been **neither compelling nor conclusive** [...]"

J. Walter, A. M. Armet, B. B. Finlay, and F. Shanahan. Establishing or Exaggerating Causality for the Gut Microbiome: Lessons from Human Microbiota-Associated Rodents. *Cell, 180*(2):221–232, 2020.

Introduction: Obesity and the gut microbiome

An obesity-associated gut microbiome with increased capacity for energy harvest

Peter J. Turnbaugh[1], Ruth E. Ley[1], Michael A. Mahowald[1], Vincent Magrini[2], Elaine R. Mardis[1,2] & Jeffrey I. Gordon[1]

Introduction: Obesity and the gut microbiome

An obesity-associated gut microbiome with increased capacity for energy harvest

Peter J. Turnbaugh[1], Ruth E. Ley[1], Michael A. Mahowald[1], Vincent Magrini[2], Elaine R. Mardis[1,2] & Jeffrey I. Gordon[1]

"Adult germ-free [wild-type] C57BL/ 6J mice were colonized (by gavage) with a microbiota harvested from the caecum of **obese (ob/ob)** or **lean (+/+)** donors (1 donor and 4–5 germ-free recipients per treatment group per experiment; two independent experiments). […]"

(**Context: ob/ob mice** have a specific mutation "[…] that produces a stereotyped, fully penetrant obesity phenotype"; **+/+ mice** lack this mutation.)

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D., & Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proceedings of the national academy of sciences*, *102*(31), 11070-11075.

Introduction: Obesity and the gut microbiome

An obesity-associated gut microbiome with increased capacity for energy harvest

Peter J. Turnbaugh[1], Ruth E. Ley[1], Michael A. Mahowald[1], Vincent Magrini[2], Elaine R. Mardis[1,2] & Jeffrey I. Gordon[1]

"Adult germ-free [wild-type] C57BL/ 6J mice were colonized (by gavage) with a microbiota harvested from the caecum of **obese (ob/ob)** or **lean (+/+)** donors (1 donor and 4–5 germ-free recipients per treatment group per experiment; two independent experiments). […]"

Results: "Strikingly, mice colonized with an **ob/ob microbiota** exhibited a **significantly greater percentage increase in body fat over two weeks** than mice colonized with a **+/+ microbiota** […]"

Introduction: Obesity and the gut microbiome

An obesity-associated gut microbiome with increased capacity for energy harvest

Peter J. Turnbaugh[1], Ruth E. Ley[1], Michael A. Mahowald[1], Vincent Magrini[2], Elaine R. Mardis[1,2] & Jeffrey I. Gordon[1]

Gut microbiome, obesity, and metabolic dysfunction

Herbert Tilg[1] and Arthur Kaser[2]

Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome

Marc A. Sze, Patrick D. Schloss

The Human Microbiome and Obesity: Moving beyond Associations

Padma Maruvada [1], Vanessa Leone [2], Lee M. Kaplan [3], Eugene B. Chang [2]

GASTROINTESTINAL INFECTIONS: EDITED BY MITCHELL COHEN

Obesity and the human microbiome

Ley, Ruth E

A Taxonomic Signature of Obesity in the Microbiome? Getting to the Guts of the Matter

Mariel M. Finucane, Thomas J. Sharpton, Timothy J. Laurent, Katherine S. Pollard

Introduction: obiome

An obesity-a robiome
with increas ergy harvest

Peter J. Turnbaugh[1], Ruth E. Ley R. Mardis[1,2] & Jeffrey I. Gordon[1]

Gut microbiom lic dysfunction

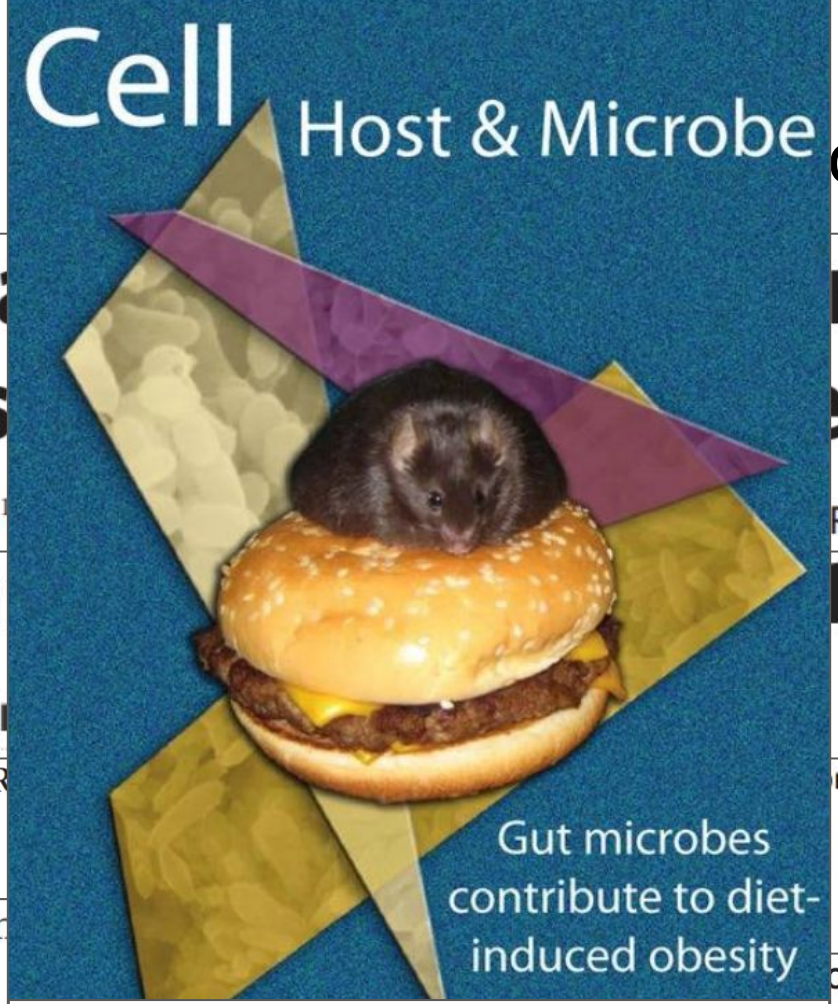
Herbert Tilg[1] and Arthu

Looking for a Signal in the Noise: R ITED BY MITCHELL COHEN
Microbiome human microbiome

Marc A. Sze, Patrick D. Schloss

The Human Microbiome ar of Obesity in the Microbiome?
beyond Associations e Matter

Padma Maruvada [1], Vanessa Leone [2], Lee M. Kaplan [3], Eugene B . Laurent, Katherine S. Pollard



Turnbaugh, P. J. (2017). Microbes and diet-induced obesity: fast, cheap, and out of control. *Cell Host & Microbe*, *21*(3), 278-281.

Introduction: Obesity and the gut microbiome?

Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome

Marc A. Sze, Patrick D. Schloss

Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA

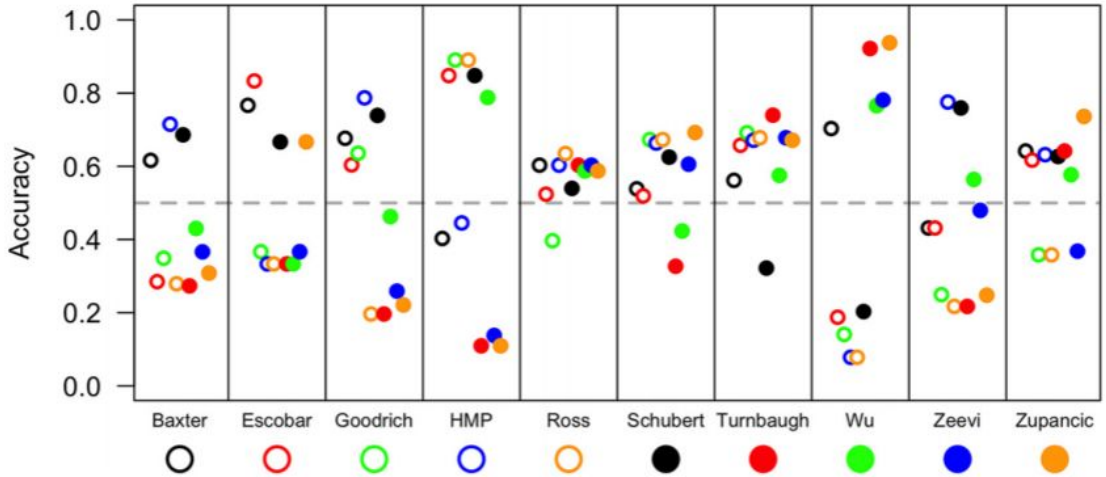


FIG 5 Overall accuracy with which each study predicted nonobese and obese individuals on the basis of that study's random forest machine learning model applied to each of the other studies. HMP, Human Microbiome Project.

Introduction: Why is it always *C. diff* and *H. pylori*?

Microbiome science needs a healthy dose of scepticism

The human microbiome in health and disease: hype or hope

Gwen Falony[a,b], Doris Vandeputte[a,b], Clara Caenepeel[c], Sara Vieira-Silva[a,b], Tanine Daryoush[a,b], Séverine Vermeire[c] and Jeroen Raes[a,b]

Separating the microbiome from the hyperbolome

Fergus Shanahan

Establishing or Exaggerating Causality for the Gut Microbiome: Lessons from Human Microbiota-Associated Rodents

Jens Walter,[1,2,3,4,8,*] Anissa M. Armet,[1,8] B. Brett Finlay,[5,6,7] and Fergus Shanahan[3]

Introduction: Why is this so difficult?

- Challenges common to many areas of science
  - Limited sample sizes
  - Institutional and personal biases against the publication of null results
  - Failure to pre-register studies
  - Hypothesizing after the collection of data without reporting the study as exploratory

- Challenges more specific to microbiome / bioinformatics research
  - Limitations of using mice (or other non-human organisms) as models
  - The “curse of dimensionality”
  - Sparsity
  - Compositionality
  - Uneven sampling depths
  - Methods that only provide limited resolution about the types of microbes in a sample

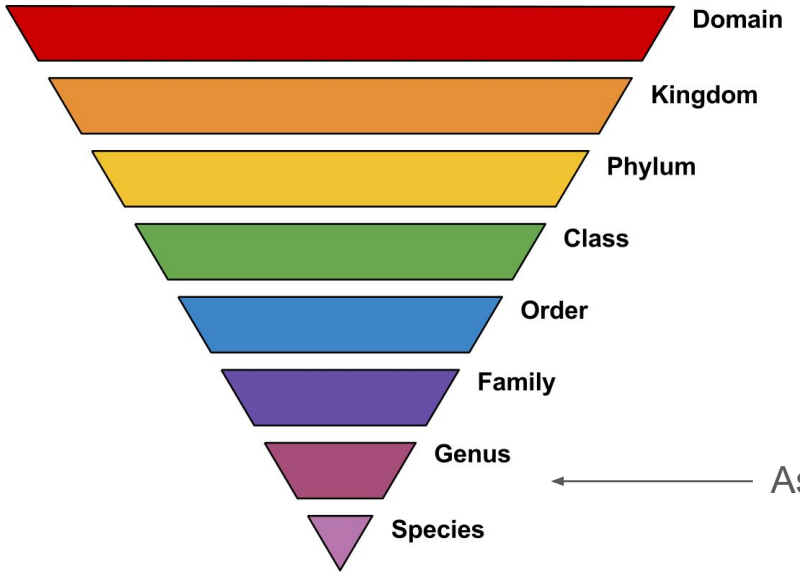
Introduction: Why is this so difficult?

- Challenges common to many areas of science
  - Limited sample sizes
  - Institutional and personal biases against the publication of null results
  - Failure to pre-register studies
  - Hypothesizing after the collection of data without reporting the study as exploratory

- Challenges more specific to microbiome / bioinformatics research
  - Limitations of using mice (or other non-human organisms) as models
  - The “curse of dimensionality”
  - Sparsity
  - Compositionality
  - Uneven sampling depths
  - **Methods that only provide limited resolution about the types of microbes in a sample**

Introduction: Our goal for this talk

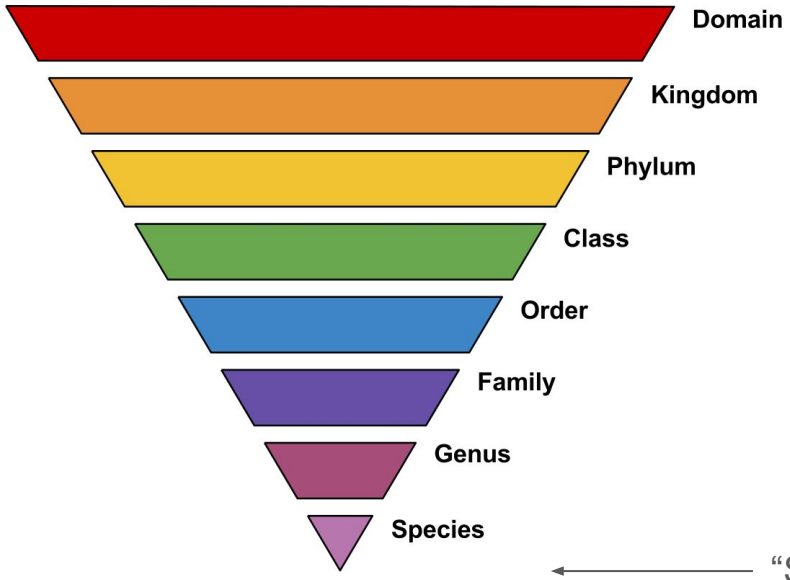
Improving the **resolution** with which we can study microbes.



As far as some technologies can go

Introduction: Our goal for this talk

Improving the **resolution** with which we can study microbes.

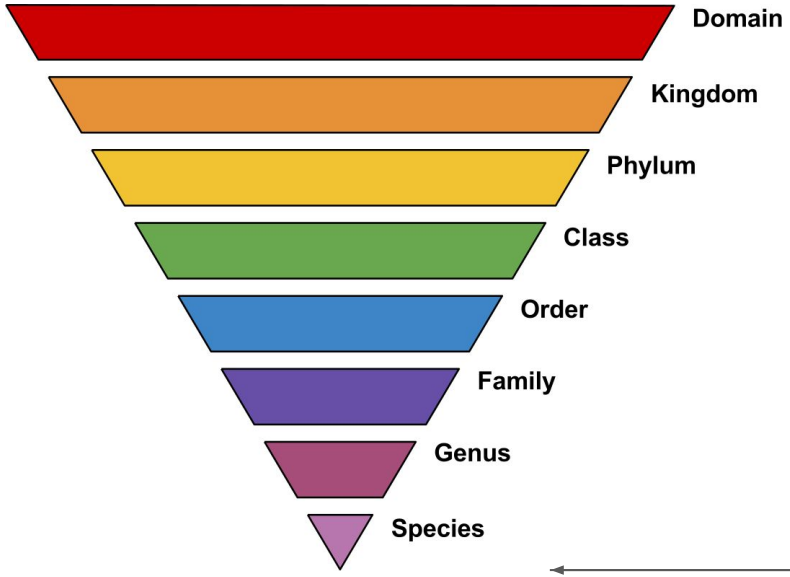


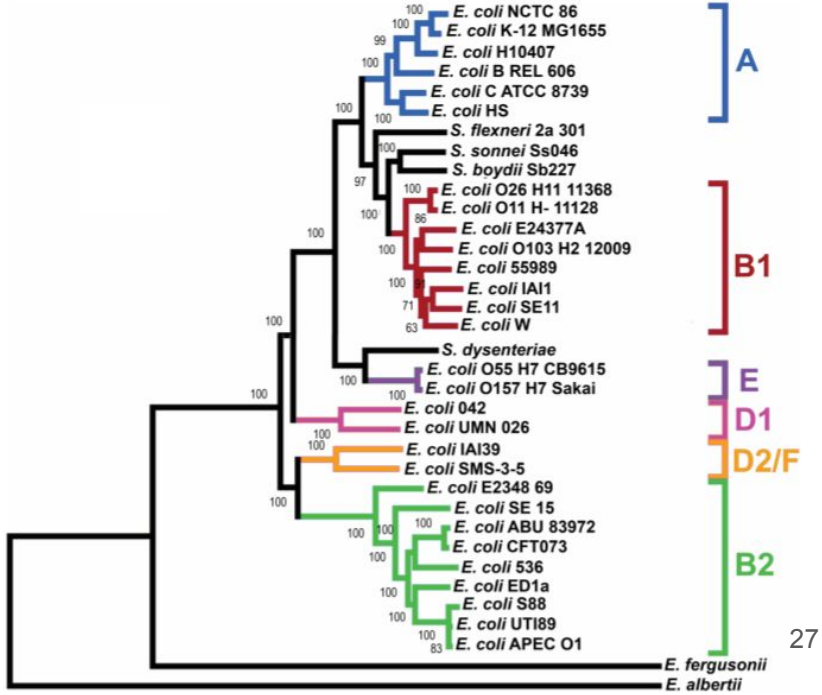
← “Strain” level: completely unique genomes

For more examples, see: Vicedomini, R., Quince, C., Darling, A. E., & Chikhi, R. (2021). Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nature Communications, 12*(1), 1-14.

Introduction: Our goal for this talk

Improving the **resolution** with which we can study microbes.





For more examples, see: Vicedomini, R., Quince, C., Darling, A. E., & Chikhi, R. (2021). Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nature Communications, 12*(1), 1-14.

Introduction: Our goal for this talk

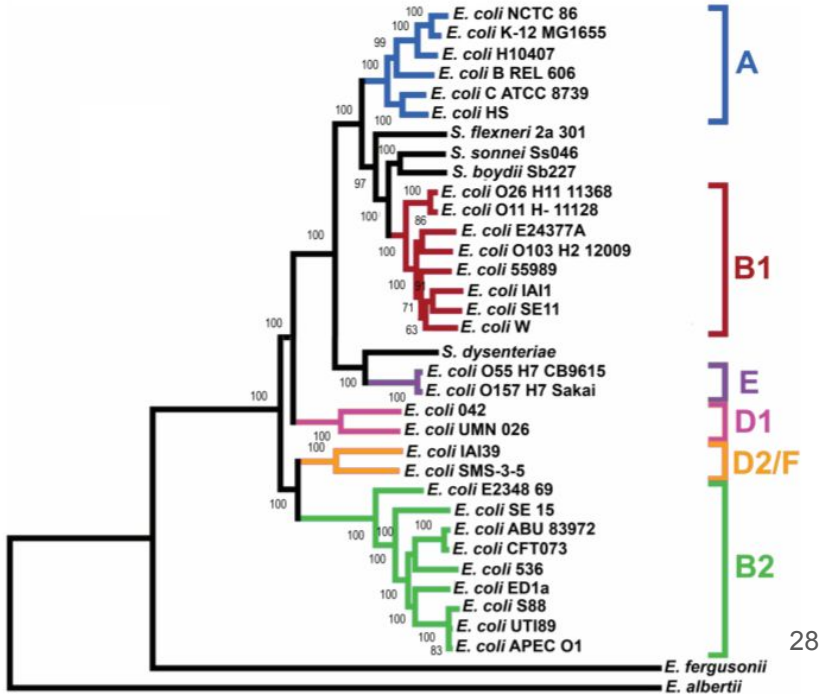
Improving the **resolution** with which we can study microbes.
Small strain-level differences can make a big difference!

Epidemic Profile of Shiga-Toxin–Producing *Escherichia coli* O104:H4 Outbreak in Germany

Christina Frank, Ph.D., Dirk Werber, D.V.M., Jakob P. Cramer, M.D., Mona Askar, M.D., Mirko Faber, M.D., Matthias an der Heiden, Ph.D., Helen Bernard, M.D., Angelika Fruth, Ph.D., Rita Prager, Ph.D., Anke Spode, M.D., Maria Wadl, D.V.M., Alexander Zoufaly, M.D., Sabine Jordan, M.D., Markus J. Kemper, M.D., Per Follin, M.D., Ph.D., Luise Müller, M.Sc., Lisa A. King, M.P.H., Bettina Rosner, Ph.D., Udo Buchholz, M.D., M.P.H., Klaus Stark, M.D., Ph.D., and Gérard Krause, M.D., Ph.D. for the HUS Investigation Team*

***Escherichia coli* induces DNA damage in vivo and triggers genomic instability in mammalian cells**

Gabriel Cuevas-Ramos[a,b], Claude R. Petit[a,b], Ingrid Marcq[a,b], Michèle Boury[a,b], Eric Oswald[a,b,c,d], and Jean-Philippe Nougayrède[a,b,1]



For more examples, see: Vicedomini, R., Quince, C., Darling, A. E., & Chikhi, R. (2021). Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nature Communications, 12*(1), 1-14.

Introduction: Our goal for this talk

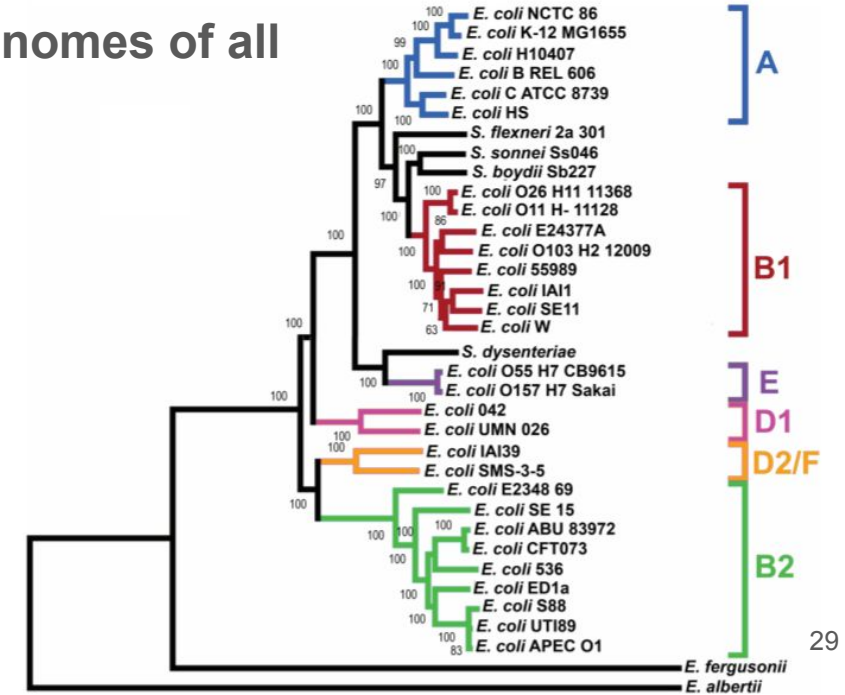
Improving the **resolution** with which we can study microbes.
Small strain-level differences can make a big difference!
Our goal, then, is **reconstructing the full genomes of all microbes in a sample**.

Epidemic Profile of Shiga-Toxin–Producing *Escherichia coli* O104:H4 Outbreak in Germany

Christina Frank, Ph.D., Dirk Werber, D.V.M., Jakob P. Cramer, M.D., Mona Askar, M.D., Mirko Faber, M.D., Matthias an der Heiden, Ph.D., Helen Bernard, M.D., Angelika Fruth, Ph.D., Rita Prager, Ph.D., Anke Spode, M.D., Maria Wadl, D.V.M., Alexander Zoufaly, M.D., Sabine Jordan, M.D., Markus J. Kemper, M.D., Per Follin, M.D., Ph.D., Luise Müller, M.Sc., Lisa A. King, M.P.H., Bettina Rosner, Ph.D., Udo Buchholz, M.D., M.P.H., Klaus Stark, M.D., Ph.D., and Gérard Krause, M.D., Ph.D. for the HUS Investigation Team*

***Escherichia coli* induces DNA damage in vivo and triggers genomic instability in mammalian cells**

Gabriel Cuevas-Ramos[a,b], Claude R. Petit[a,b], Ingrid Marcq[a,b], Michèle Boury[a,b], Eric Oswald[a,b,c,d], and Jean-Philippe Nougayrède[a,b,1]



Introduction: Our goal for this talk

Improving the **resolution** with which we can study microbes.
Small strain-level differences can make a big difference!
Our goal, then, is **reconstructing the full genomes of all microbes in a sample**.

Introduction: Our goal for this talk

Improving the **resolution** with which we can study microbes.
Small strain-level differences can make a big difference!
Our goal, then, is **reconstructing the full genomes of all microbes in a sample**.

… This isn’t really possible right now, but we’ll see what we can do!

Introduction: Our goal for this talk

Improving the **resolution** with which we can study microbes.
Small strain-level differences can make a big difference!
Our goal, then, is **reconstructing the full genomes of all microbes in a sample**.

… This isn't really possible right now, but we'll see what we can do!

One final note for the introduction, though:

Introduction: Why is this so difficult?

- Challenges common to many areas of science
  - Limited sample sizes
  - Institutional and personal biases against the publication of null results
  - Failure to pre-register studies
  - Hypothesizing after the collection of data without reporting the study as exploratory

- Challenges more specific to microbiome / bioinformatics research
  - Limitations of using mice (or other non-human organisms) as models
  - The “curse of dimensionality”
  - Sparsity
  - Compositionality
  - Uneven sampling depths
  - **Methods that only provide limited resolution about the types of microbes in a sample**

Introduction: Why is this so difficult?

- Challenges common to many areas of science
  - **Limited sample sizes**
  - **Institutional and personal biases against the publication of null results**
  - **Failure to pre-register studies**
  - **Hypothesizing after the collection of data without reporting the study as exploratory**

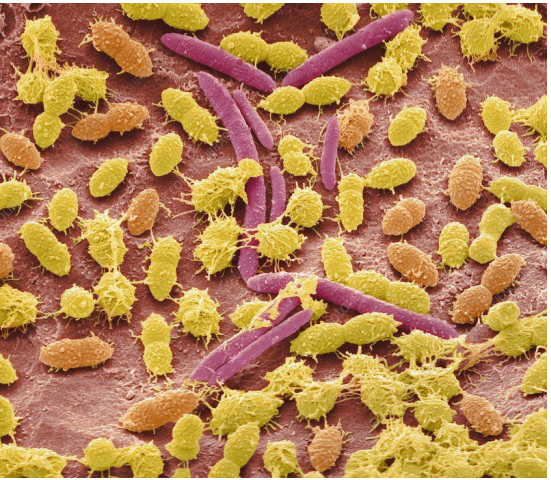
- Challenges more specific to microbiome / bioinformatics research
  - **Limitations of using mice (or other non-human organisms) as models**
  - **The “curse of dimensionality”**
  - **Sparsity**
  - **Compositionality**
  - **Uneven sampling depths**
  - Methods that only provide limited resolution about the types of microbes in a sample

This talk

1. Introduction: Studying microbiomes
    a. Why bother?
    b. Why hasn't this research been more useful?
    c. Defining a goal for this talk
2. **Culture-independent (a.k.a. sequencing-based) methods**
    a. Marker gene sequencing
    b. Metagenome sequencing
3. Metagenome assembly
    a. Input (*reads*)
    b. Outputs (*contigs*, an *assembly graph*, …)
    c. Methods (*de novo* vs. reference-based, overlap graph vs. de Bruijn graph, …)
4. Future work: Solving the *strain separation problem*

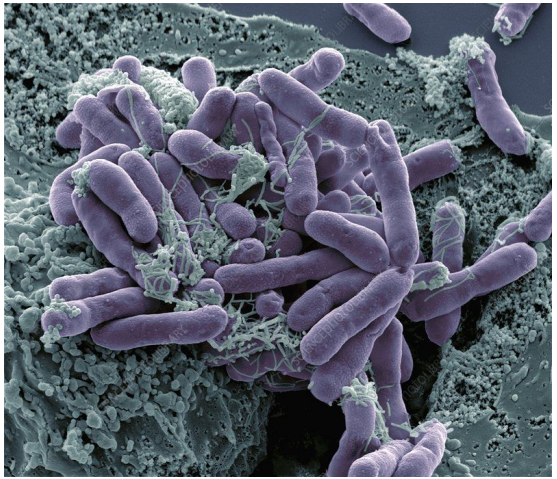
Culture-Independent Methods

Culture-Independent Methods



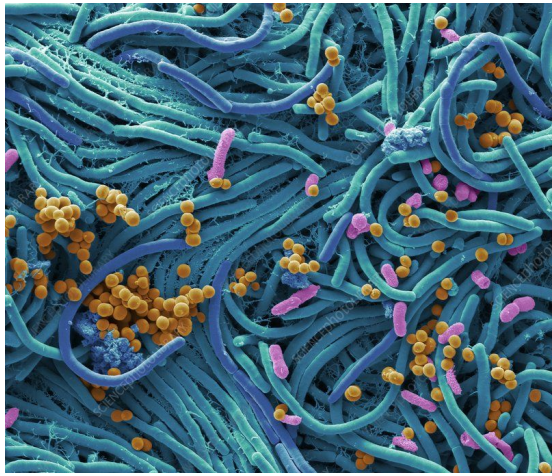
Steve Gschmeissner, Science Photo Library

“Scanning electron micrograph (SEM) of bacteria cultured from a sample of human faeces.”



Steve Gschmeissner, Science Photo Library

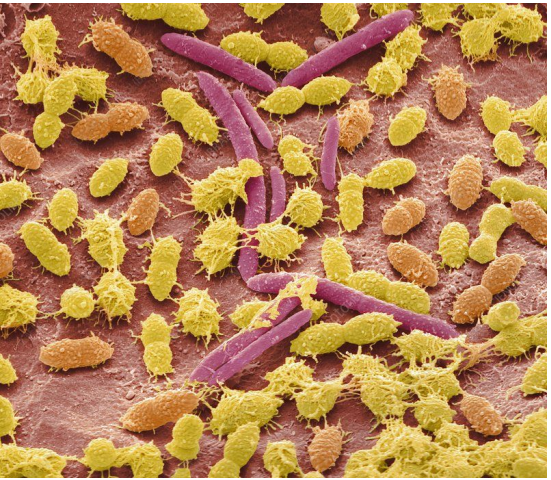
“Bacterial contamination, coloured scanning electron micrograph (SEM). *Escherichia coli* bacteria in a cell culture. This contamination has come from an unclean water source.”



Steve Gschmeissner, Science Photo Library

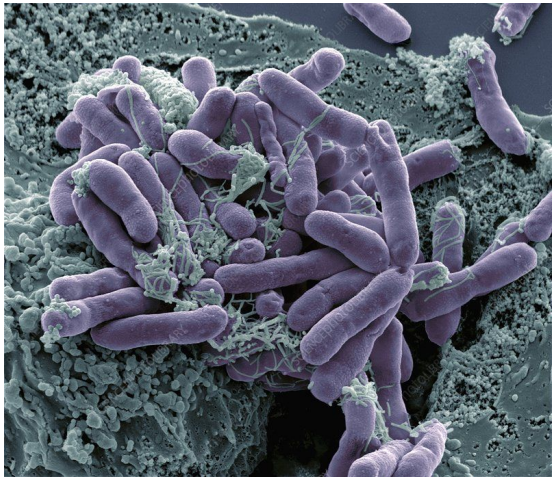
“Coloured scanning electron micrograph (SEM) of bacteria cultured from a mobile phone.”

Culture-Independent Methods



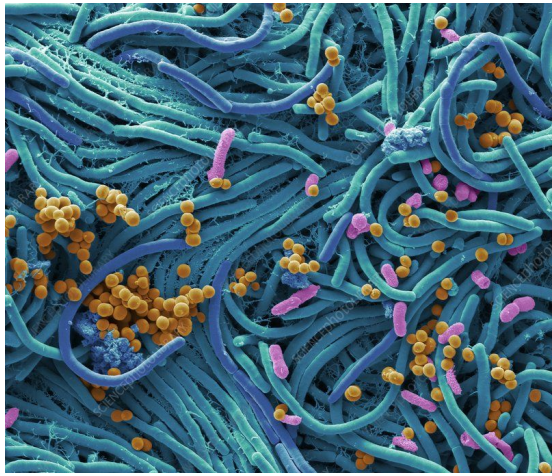
Steve Gschmeissner, Science Photo Library

"Scanning electron micrograph (SEM) of bacteria **cultured** from a sample of human faeces."



Steve Gschmeissner, Science Photo Library

"Bacterial contamination, coloured scanning electron micrograph (SEM). *Escherichia coli* bacteria in a **cell culture**. This contamination has come from an unclean water source."



Steve Gschmeissner, Science Photo Library

"Coloured scanning electron micrograph (SEM) of bacteria **cultured** from a mobile phone."

Culture-Independent Methods

"It is estimated that **>99% of microorganisms observable in nature** typically are not cultivated by using standard techniques."

Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of bacteriology*, *180*(18), 4765-4774.

Culture-Independent Methods

"It is estimated that **>99% of microorganisms observable in nature** typically are not cultivated by using standard techniques."

Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of bacteriology*, *180*(18), 4765-4774.

Although not all microbes are easily culturable, all have **genomes**.

Idea: look at the DNA in a sample and use that to study the microbes there!

For a nice history of these and other methods, see: M. H. Fraher, P. W. O'Toole, and E. M. Quigley. Techniques used to characterize the gut microbiota: a guide for the clinician. *Nature Reviews Gastroenterology & Hepatology, 9*(6):312–322, 2012.

Culture-Independent Methods

1. Terminal restriction fragment length polymorphism
   a.k.a. *T-RFLP*

2. Denaturing / temperature gradient gel electrophoresis
   a.k.a. *DGGE / TGGE*

3. Fluorescence *in situ* hybridization
   a.k.a. *FISH*

4. Marker gene sequencing
   a.k.a. *amplicon sequencing, metabarcoding, metataxonomics, ...*

5. Metagenome sequencing
   a.k.a. *metagenomics*, *shotgun metagenome sequencing*, *whole metagenome sequencing*, ...

For a nice history of these and other methods, see: M. H. Fraher, P. W. O'Toole, and E. M. Quigley. Techniques used to characterize the gut microbiota: a guide for the clinician. *Nature Reviews Gastroenterology & Hepatology, 9*(6):312–322, 2012.

Culture-Independent Methods

1. Terminal restriction fragment length polymorphism
   a.k.a. *T-RFLP*

2. Denaturing / temperature gradient gel electrophoresis
   a.k.a. *DGGE / TGGE*

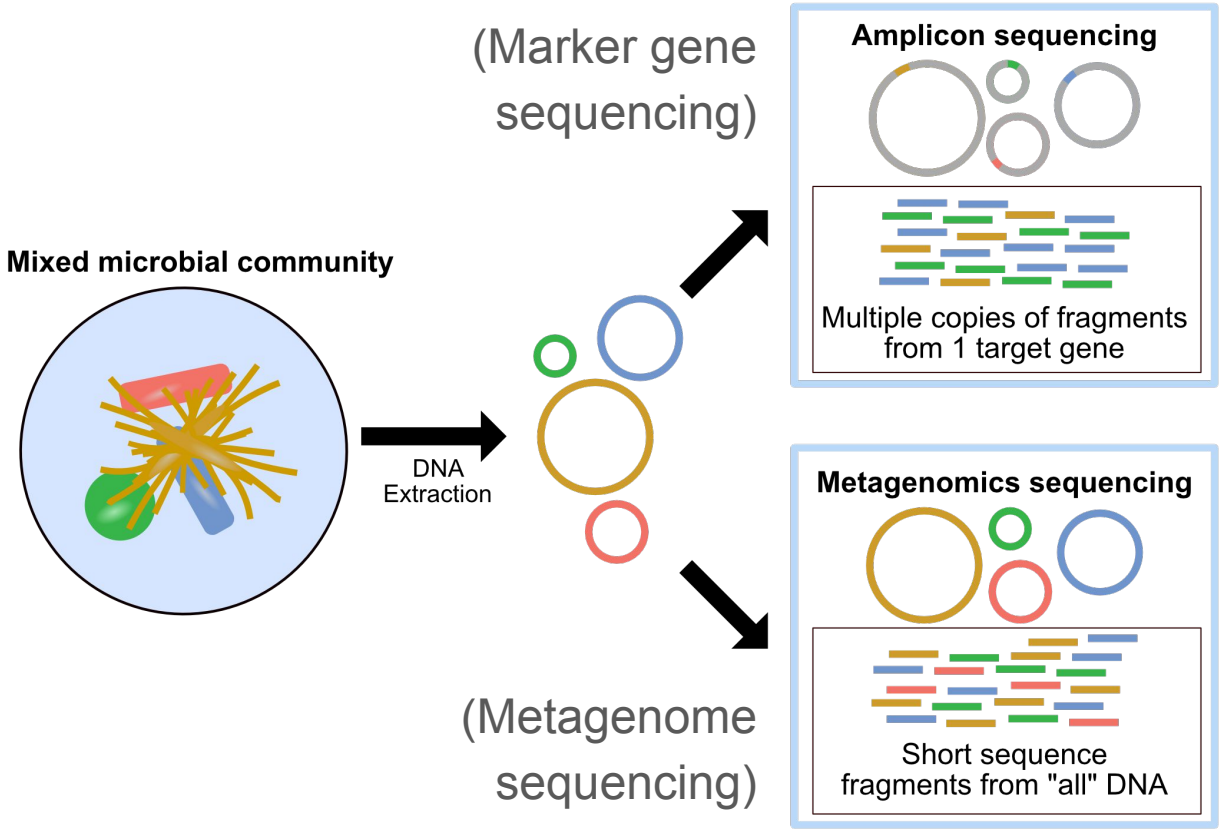
3. Fluorescence *in situ* hybridization
   a.k.a. *FISH*

4. **Marker gene sequencing**
   a.k.a. *amplicon sequencing, metabarcoding, metataxonomics, ...*

5. **Metagenome sequencing**
   a.k.a. *metagenomics*, *shotgun metagenome sequencing*, *whole metagenome sequencing*, ...

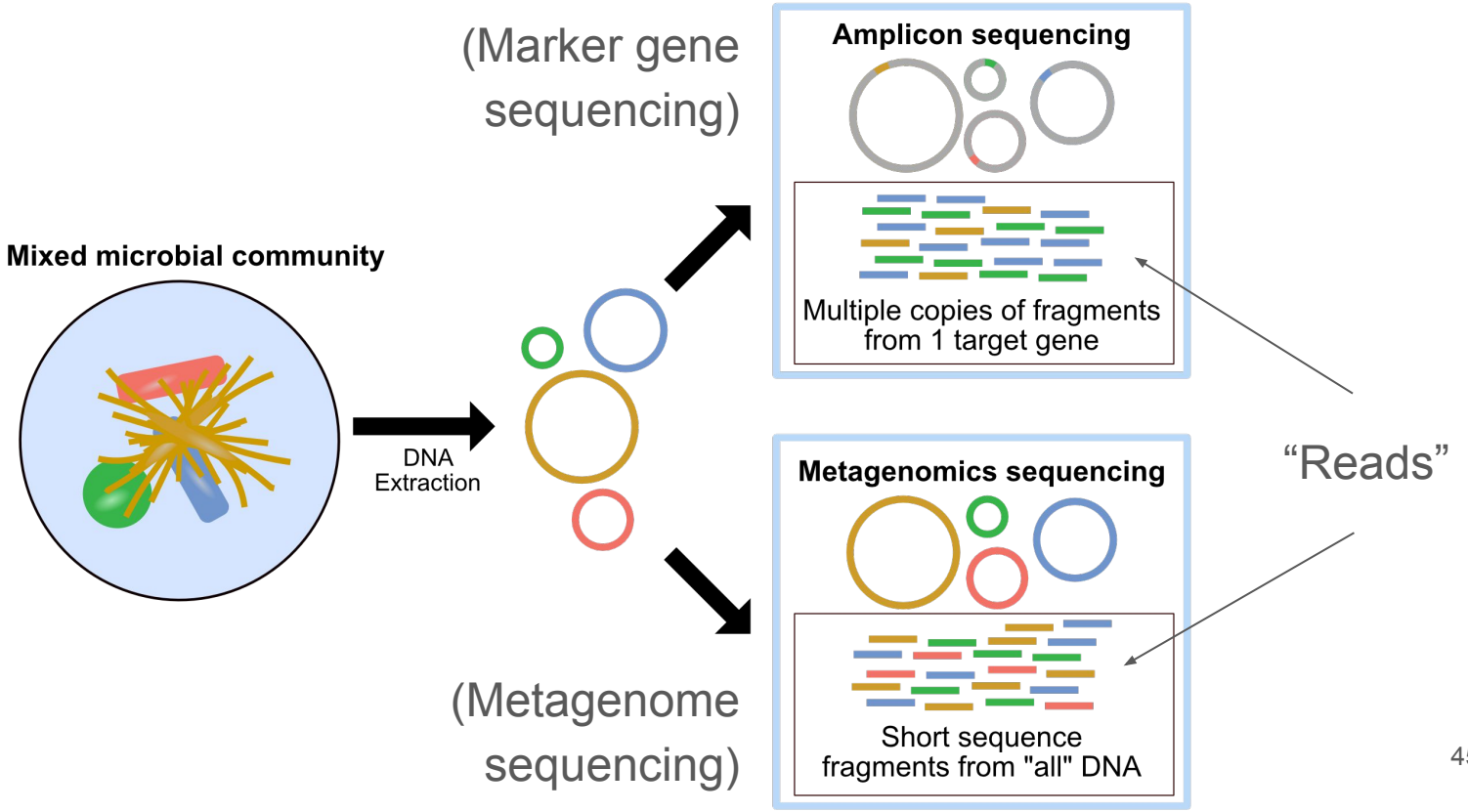
Lee, M. D. (2019). Happy Belly Bioinformatics: An open-source resource dedicated to helping biologists utilize bioinformatics. *Journal of Open Source Education*, *4*(41), 53.

Culture-Independent Methods



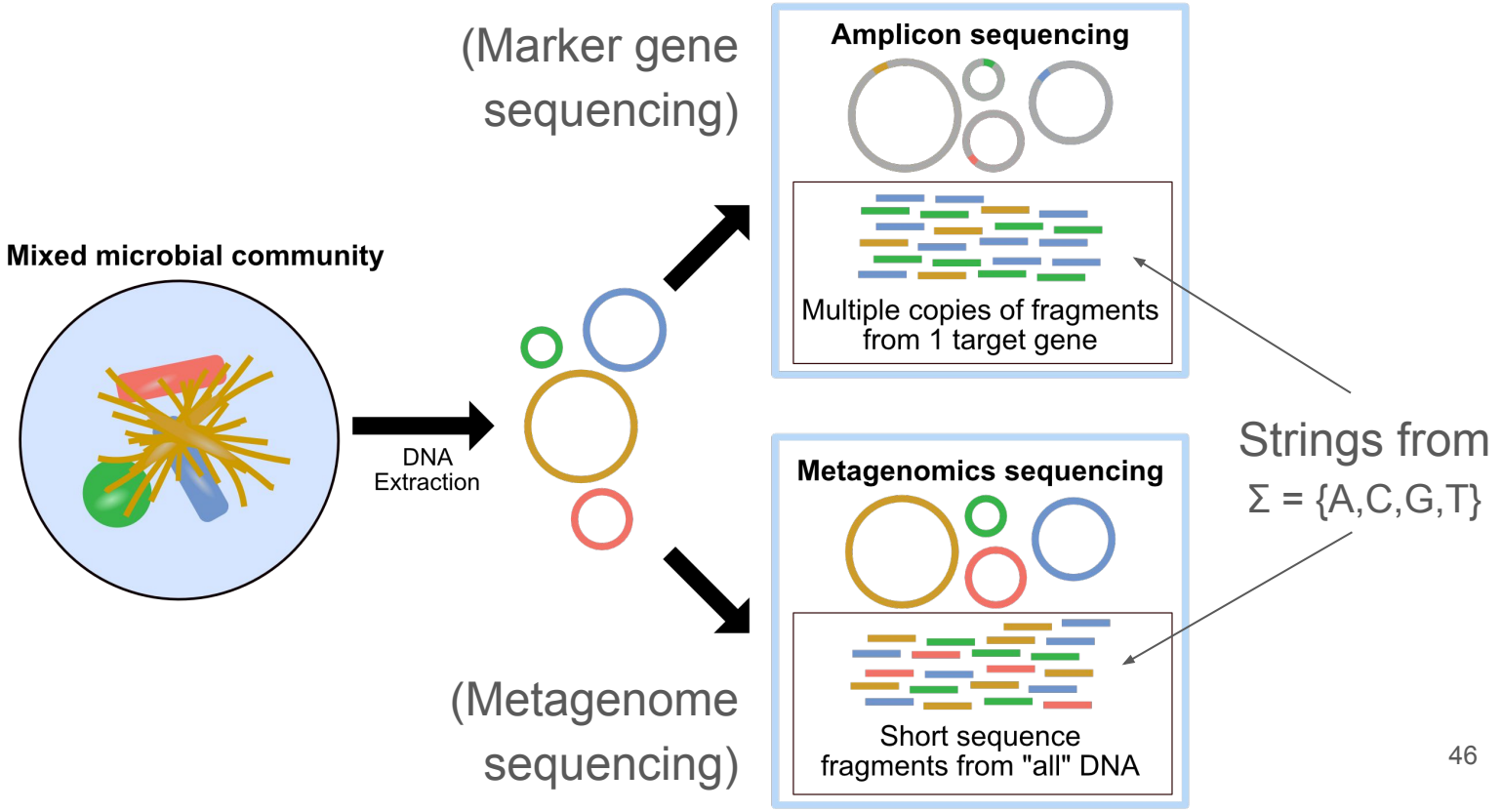
Lee, M. D. (2019). Happy Belly Bioinformatics: An open-source resource dedicated to helping biologists utilize bioinformatics. *Journal of Open Source Education*, *4*(41), 53.

Culture-Independent Methods

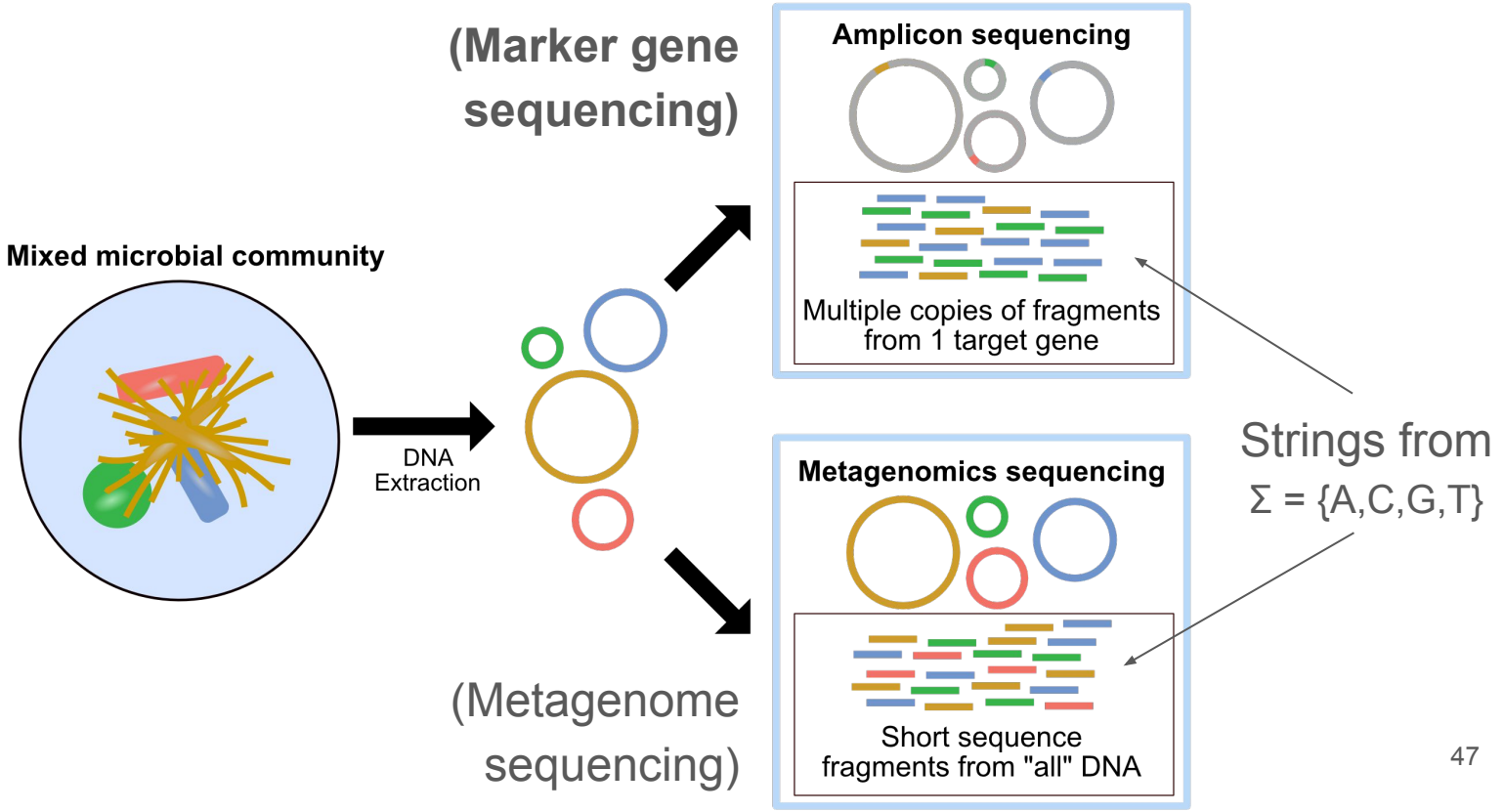


Lee, M. D. (2019). Happy Belly Bioinformatics: An open-source resource dedicated to helping biologists utilize bioinformatics. *Journal of Open Source Education*, *4*(41), 53.

Culture-Independent Methods



Culture-Independent Methods



C. I. Methods: Marker gene sequencing

For a more thorough history of Carl Woese's life and work, see:
D. Quammen. The Scientist Who Scrambled Darwin's Tree of Life. *The New York Times*, 2018.

C. I. Methods: Carl Woese and rRNA genes

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 11, pp. 5088–5090, November 1977
Evolution

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

CARL R. WOESE AND GEORGE E. FOX*

Department of Genetics and Development, University of Illinois, Urbana, Illinois 61801

ABSTRACT A phylogenetic analysis based upon ribosomal RNA sequence characterization reveals that living systems represent one of three aboriginal lines of descent:

(i) the eubacteria, comprising all typical bacteria;
(ii) the archaebacteria, containing methanogenic bacteria; and
(iii) the urkaryotes, now represented in the cytoplasmic component of eukaryotic cells.

For a more thorough history of Carl Woese's life and work, see:
D. Quammen. The Scientist Who Scrambled Darwin's Tree of Life. *The New York Times*, 2018.

C. I. Methods: Carl Woese and rRNA genes

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 11, pp. 5088–5090, November 1977
Evolution

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

CARL R. WOESE AND GEORGE E. FOX*

Department of Genetics and Development, University of Illinois, Urbana, Illinois 61801

ABSTRACT A phylogenetic analysis based upon ribosomal RNA sequence characterization reveals that living systems represent one of three aboriginal lines of descent:

(i) the eubacteria, comprising all typical bacteria;
(ii) the archaebacteria, containing methanogenic bacteria; and
(iii) the urkaryotes, now represented in the cytoplasmic component of eukaryotic cells.

For a more thorough history of Carl Woese's life and work, see:
D. Quammen. The Scientist Who Scrambled Darwin's Tree of Life. *The New York Times*, 2018.

C. I. Methods: Carl Woese and rRNA genes

ABSTRACT A phylogenetic analysis based upon ribosomal RNA sequence characterization reveals that living systems represent one of three aboriginal lines of descent:

(i) the eubacteria, comprising all typical bacteria;
(ii) the archaebacteria, containing methanogenic bacteria; and
(iii) the urkaryotes, now represented in the cytoplasmic component of eukaryotic cells.

For a more thorough history of Carl Woese's life and work, see:
D. Quammen. The Scientist Who Scrambled Darwin's Tree of Life. *The New York Times*, 2018.

C. I. Methods: Carl Woese and rRNA genes

C. I. Methods: What's in a marker gene?

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 11, pp. 5088–5090, November 1977
Evolution

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

CARL R. WOESE AND GEORGE E. FOX*

Department of Genetics and Development, University of Illinois, Urbana, Illinois 61801

ABSTRACT A phylogenetic analysis based upon ribosomal RNA sequence characterization reveals that living systems represent one of three aboriginal lines of descent:

(i) the eubacteria, comprising all typical bacteria;
(ii) the archaebacteria, containing methanogenic bacteria; and
(iii) the urkaryotes, now represented in the cytoplasmic component of eukaryotic cells.

C. I. Methods: What's in a marker gene?

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 11, pp. 5088–5090, November 1977
Evolution

Phylogenetic structure of the prokaryotic domain: The primary kingdoms

(archaebacteria/eubacteria/urkaryote/16S ribosomal RNA/molecular phylogeny)

CARL R. WOESE AND GEORGE E. FOX*

Department of Genetics and Development, University of Illinois, Urbana, Illinois 61801

ABSTRACT A phylogenetic analysis based upon **ribosomal RNA sequence** characterization reveals that living systems represent one of three aboriginal lines of descent:

(i) the eubacteria, comprising all typical bacteria;
(ii) the archaebacteria, containing methanogenic bacteria; and
(iii) the urkaryotes, now represented in the cytoplasmic component of eukaryotic cells.

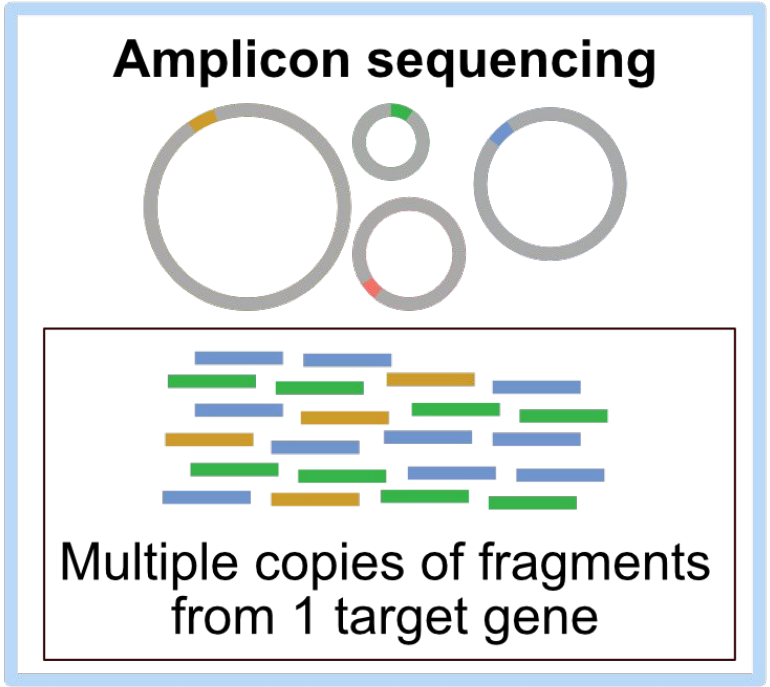
C. I. Methods: What's in a marker gene?

"To determine relationships covering the entire spectrum of extant living systems, one optimally needs a molecule of appropriately broad distribution. None of the readily characterized proteins fits this requirement. However, ribosomal RNA does. It is

- a component of all self-replicating systems;
- it is readily isolated;
- and **its sequence changes but slowly with time**—

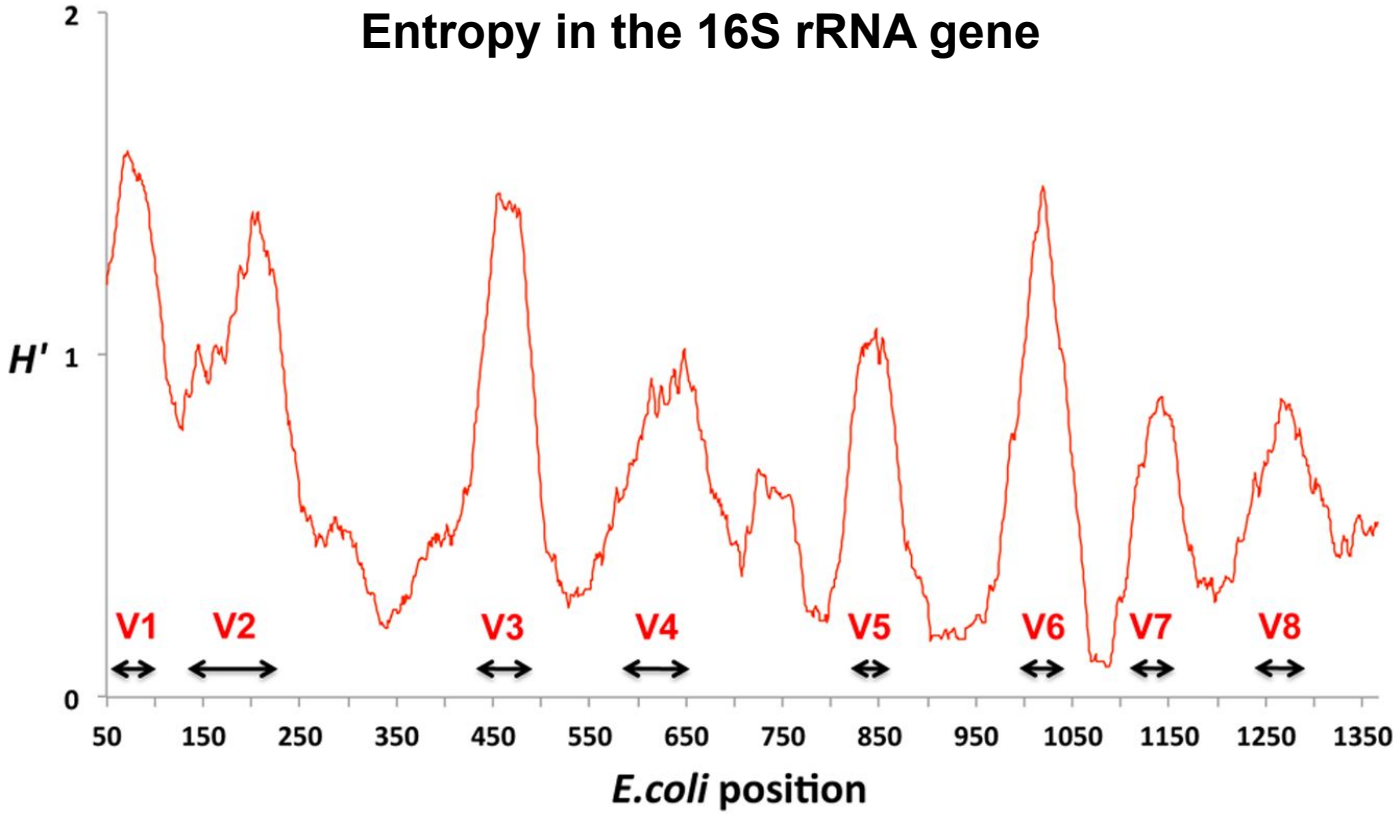
permitting the detection of relatedness among very distant species."

C. I. Methods: What's in a marker gene?



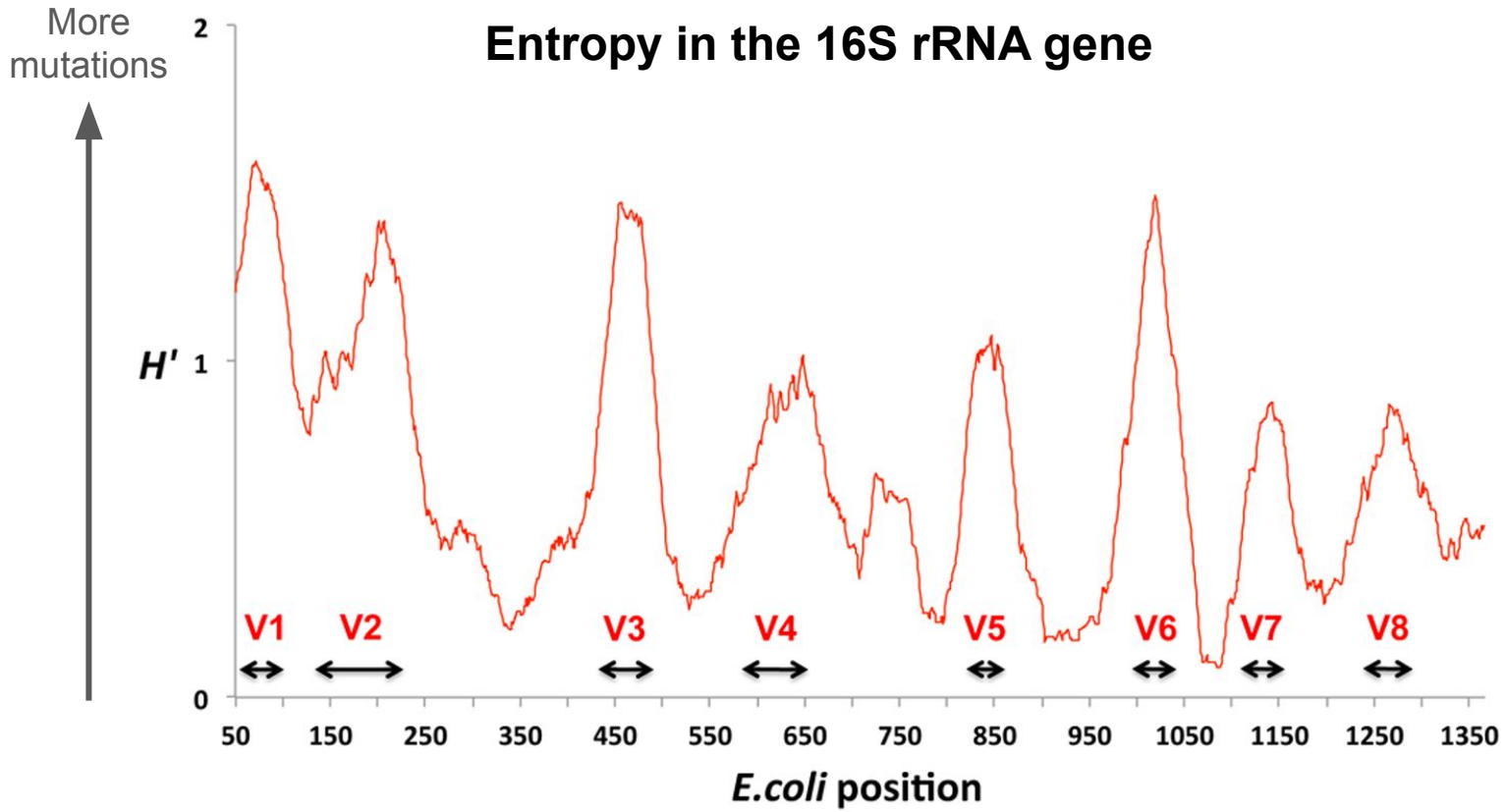
S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One, 7*(8):e42671, 2012.

C. I. Methods: What's in a marker gene?



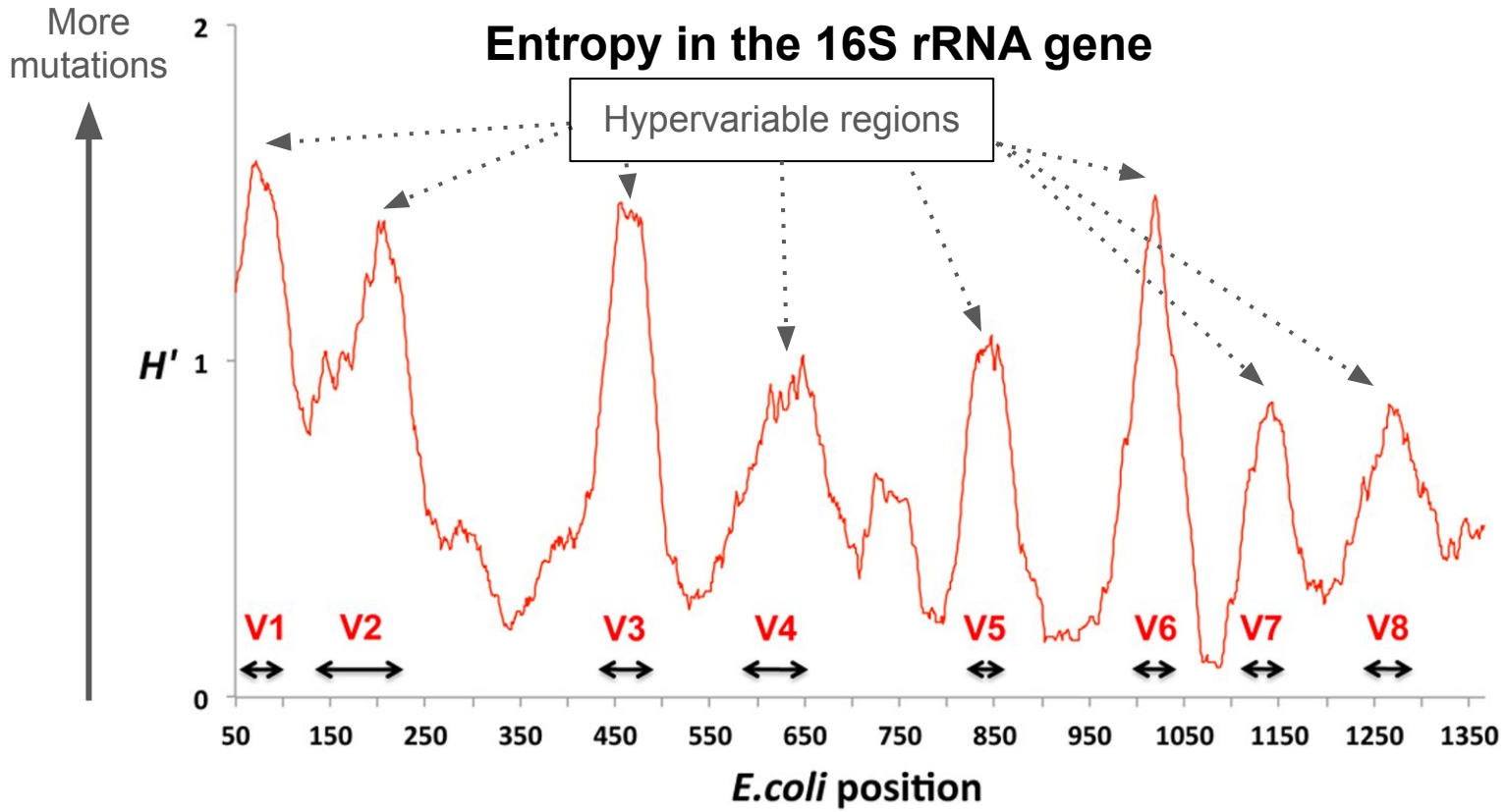
S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One, 7*(8):e42671, 2012.

C. I. Methods: What's in a marker gene?



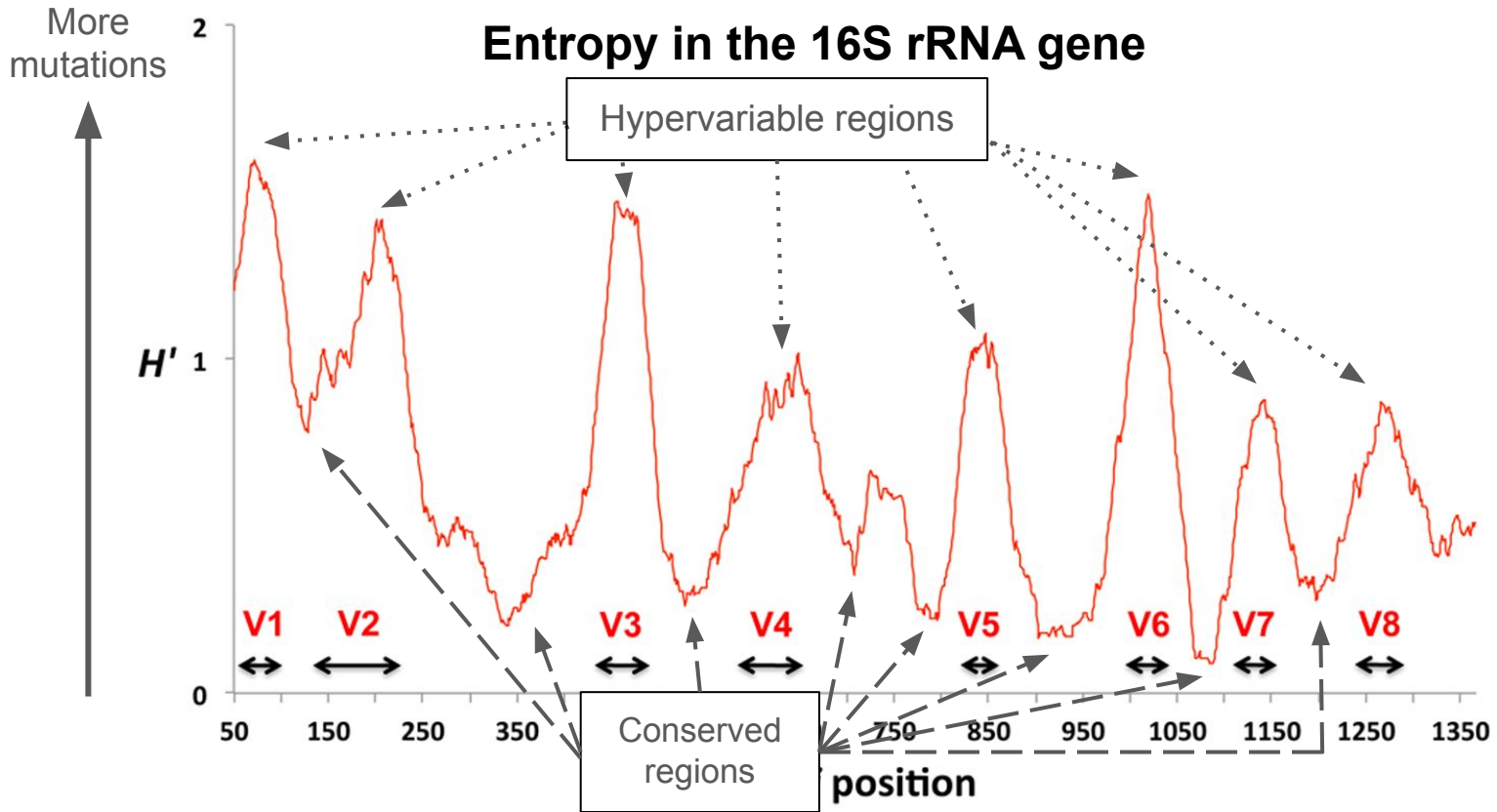
S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One, 7*(8):e42671, 2012.

C. I. Methods: What's in a marker gene?



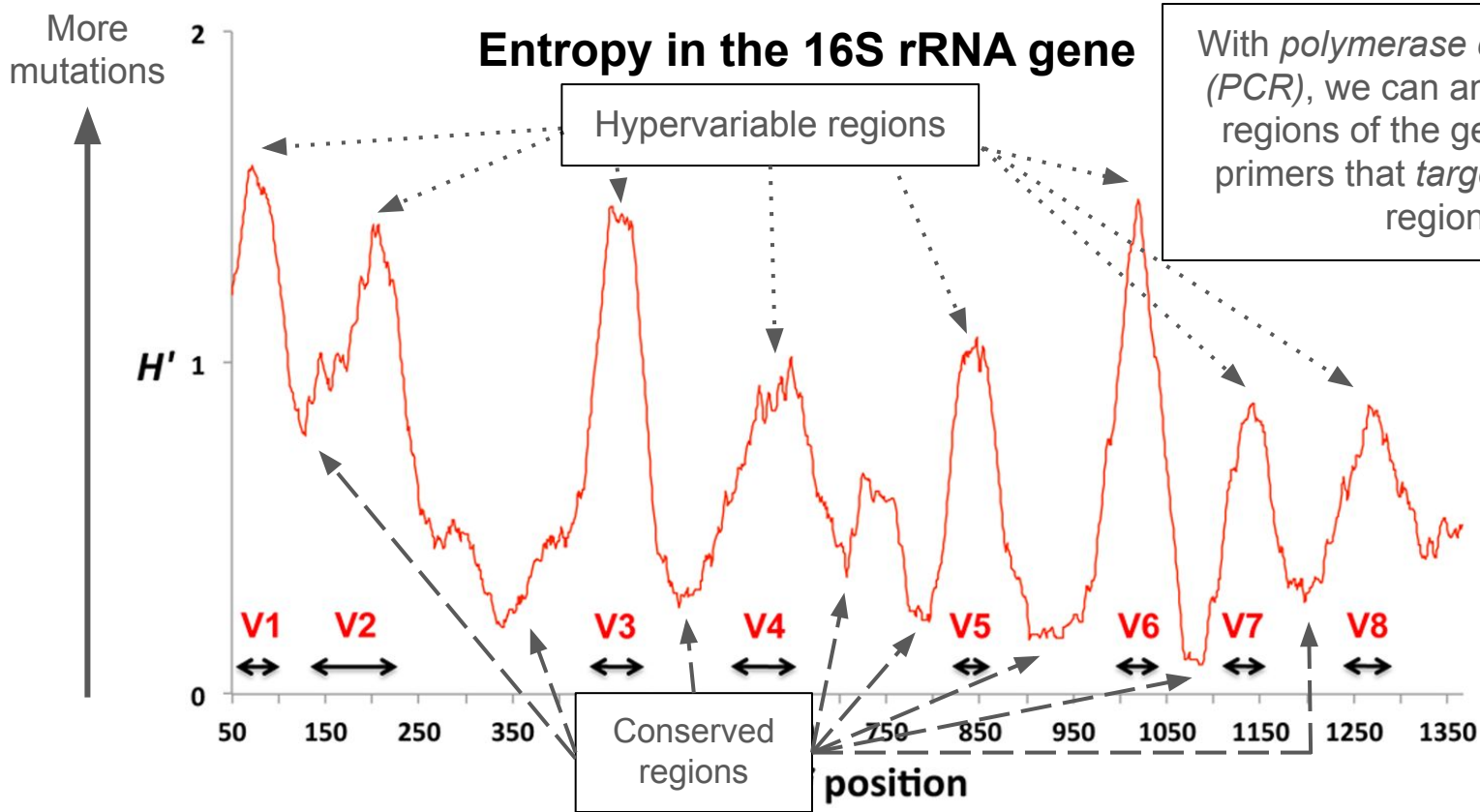
S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One, 7*(8):e42671, 2012.

C. I. Methods: What's in a marker gene?



S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One, 7*(8):e42671, 2012.

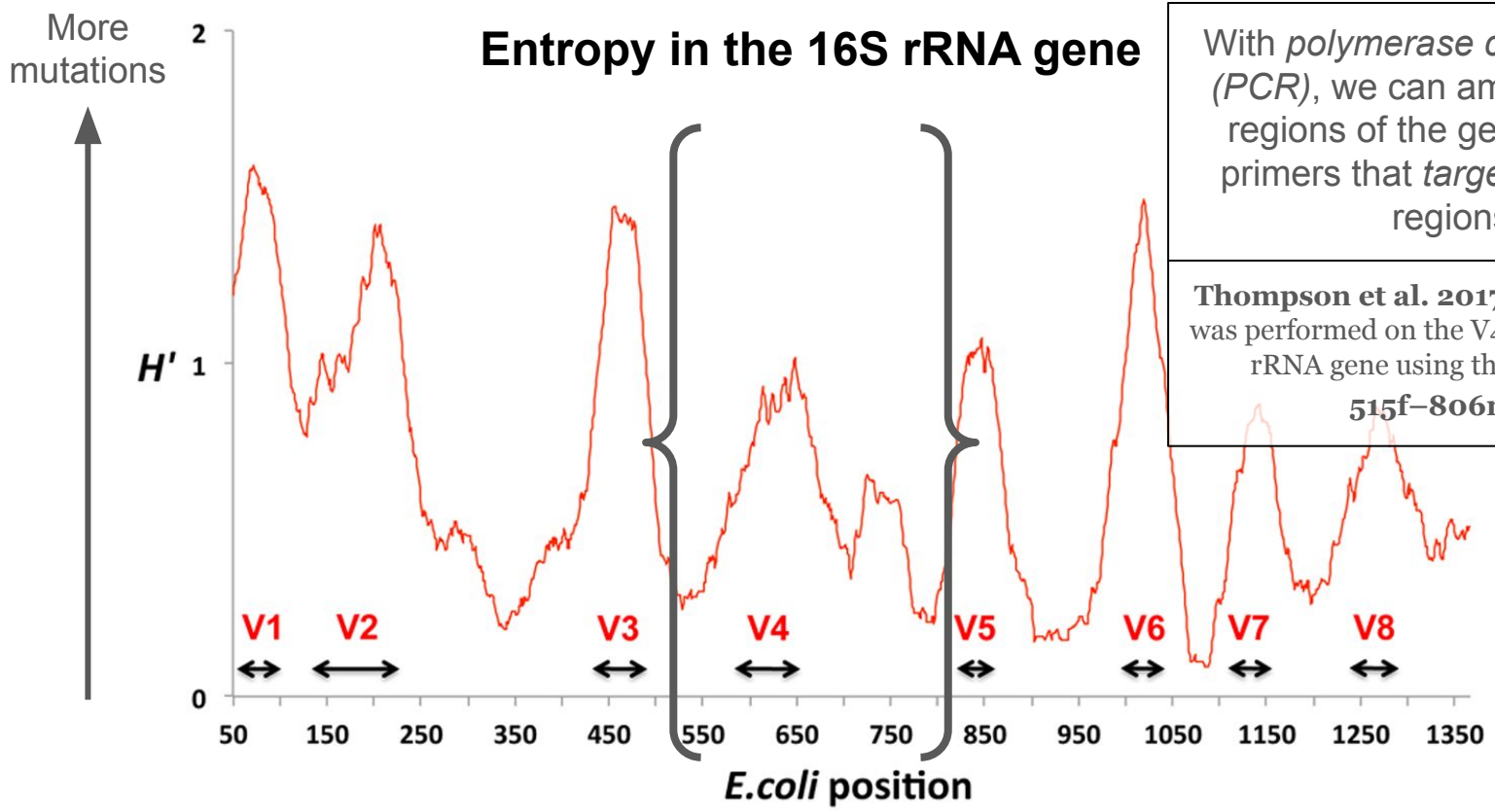
C. I. Methods: What's in a marker gene?



With *polymerase chain reaction (PCR)*, we can amplify specific regions of the genome using primers that *target* conserved regions.

S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One, 7*(8):e42671, 2012.

C. I. Methods: What's in a marker gene?

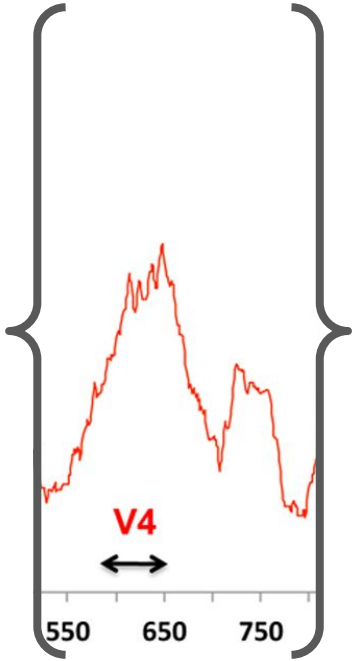


With *polymerase chain reaction (PCR)*, we can amplify specific regions of the genome using primers that *target* conserved regions.

Thompson et al. 2017: "Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair **515f–806r** […]"

S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One, 7*(8):e42671, 2012.

C. I. Methods: What's in a marker gene?

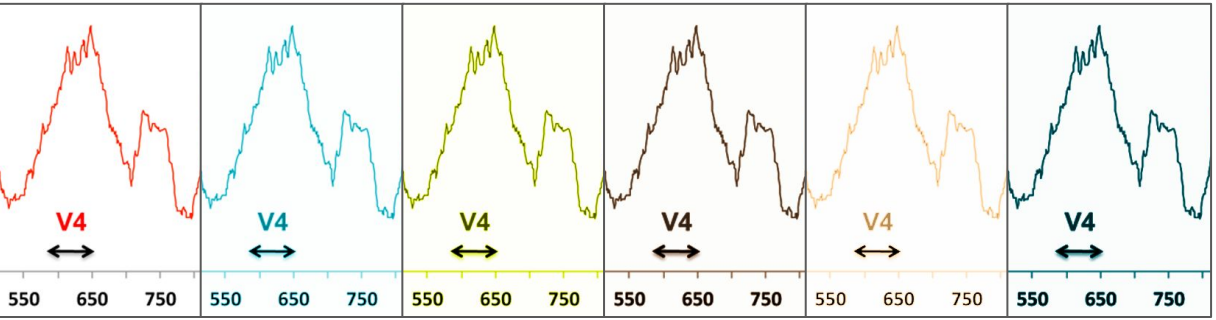


With *polymerase chain reaction (PCR)*, we can amplify specific regions of the genome using primers that *target* conserved regions.

Thompson et al. 2017: "Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair **515f–806r** […]"

S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One, 7*(8):e42671, 2012.

C. I. Methods: What's in a marker gene?

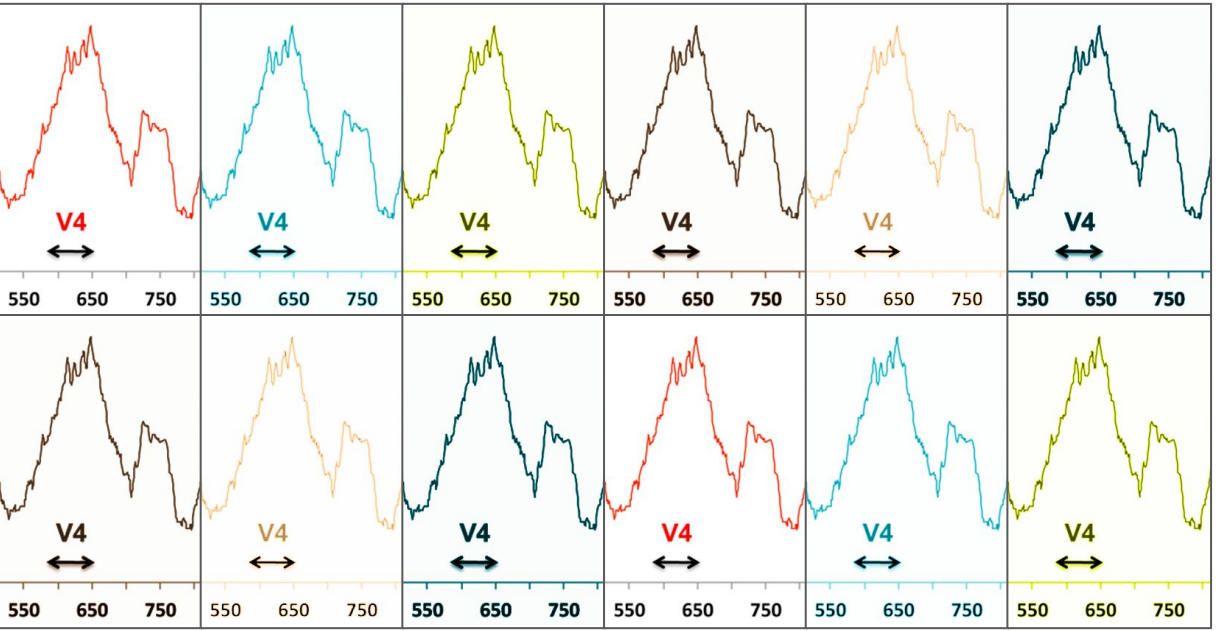


With *polymerase chain reaction (PCR)*, we can amplify specific regions of the genome using primers that *target* conserved regions.

Thompson et al. 2017: "Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair **515f–806r** […]"

S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One, 7*(8):e42671, 2012.

C. I. Methods: What's in a marker gene?

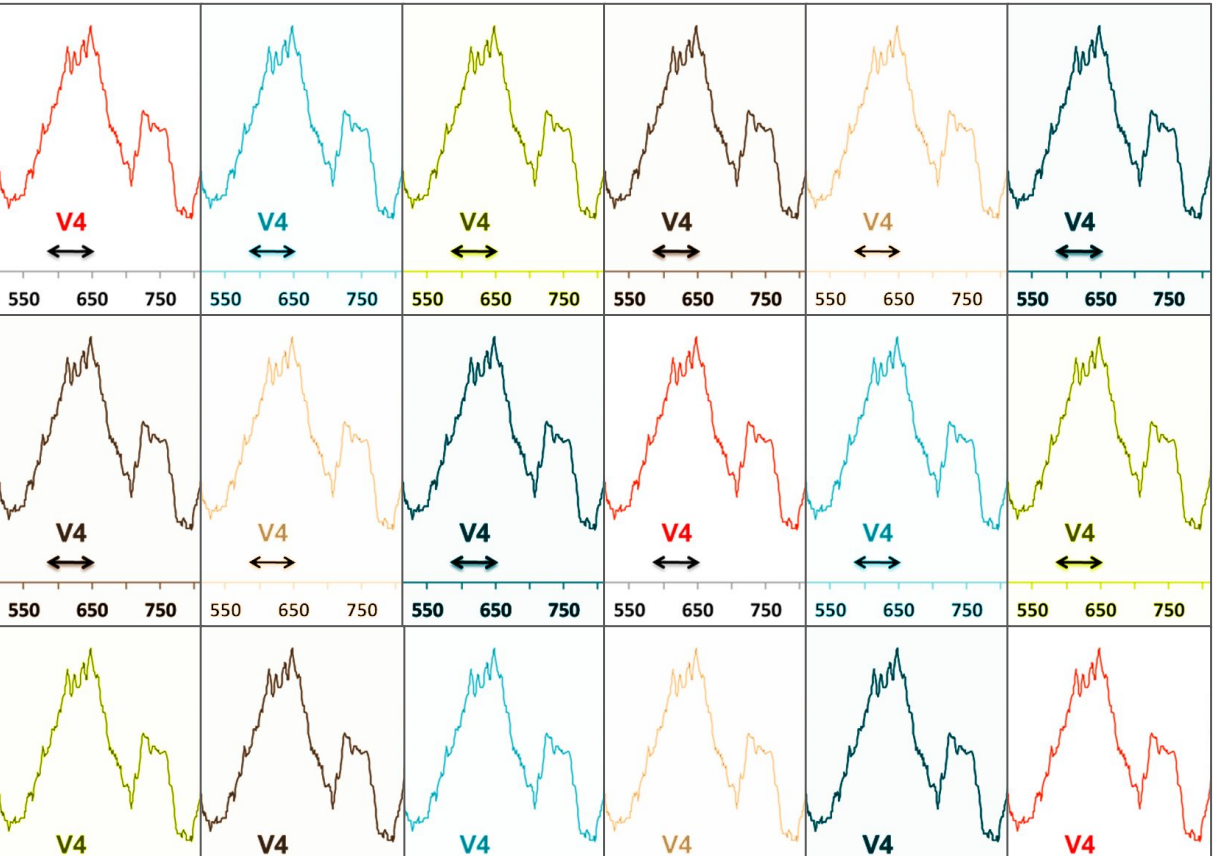


With *polymerase chain reaction (PCR)*, we can amplify specific regions of the genome using primers that *target* conserved regions.

Thompson et al. 2017: "Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair **515f–806r** […]"

S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One, 7*(8):e42671, 2012.

C. I. Methods: What's in a marker gene?

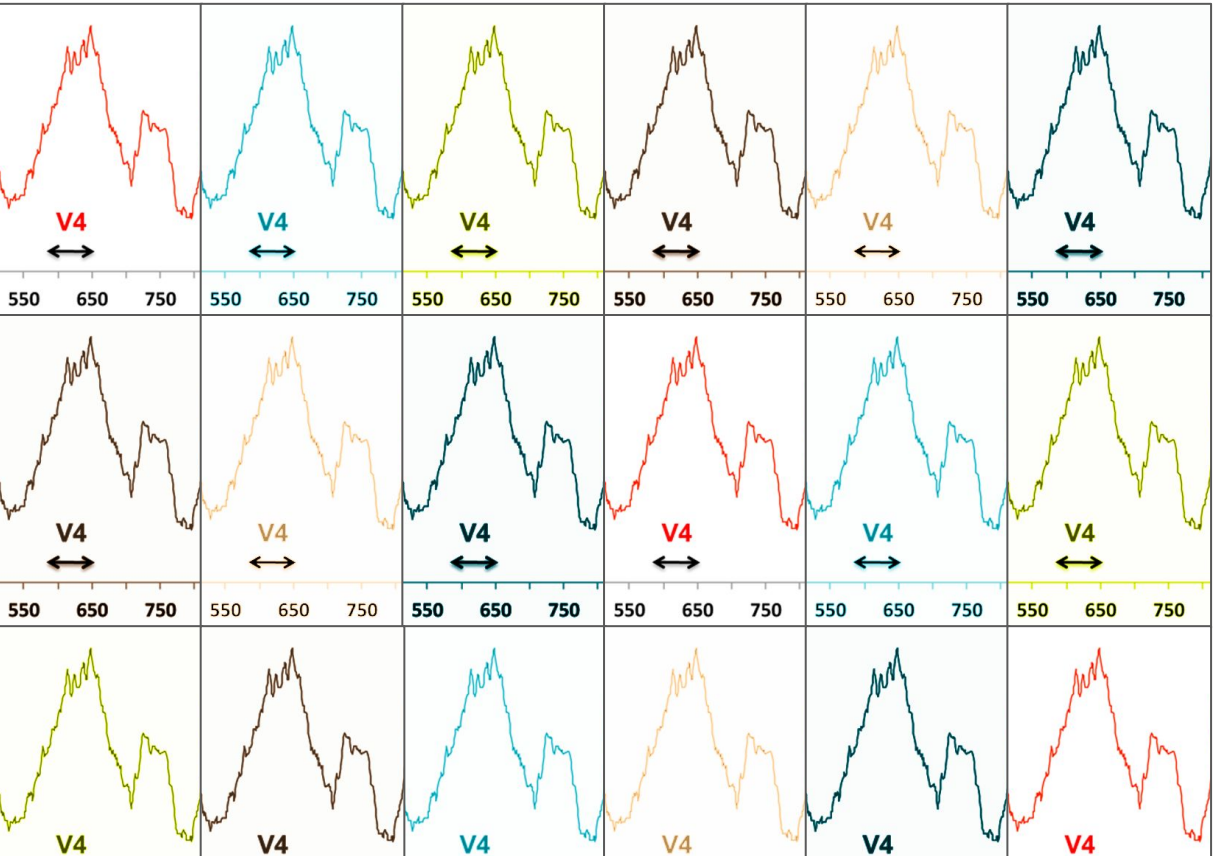


With *polymerase chain reaction (PCR)*, we can amplify specific regions of the genome using primers that *target* conserved regions.

Thompson et al. 2017: "Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair **515f–806r** […]"

S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One, 7*(8):e42671, 2012.

C. I. Methods: What's in a marker gene?



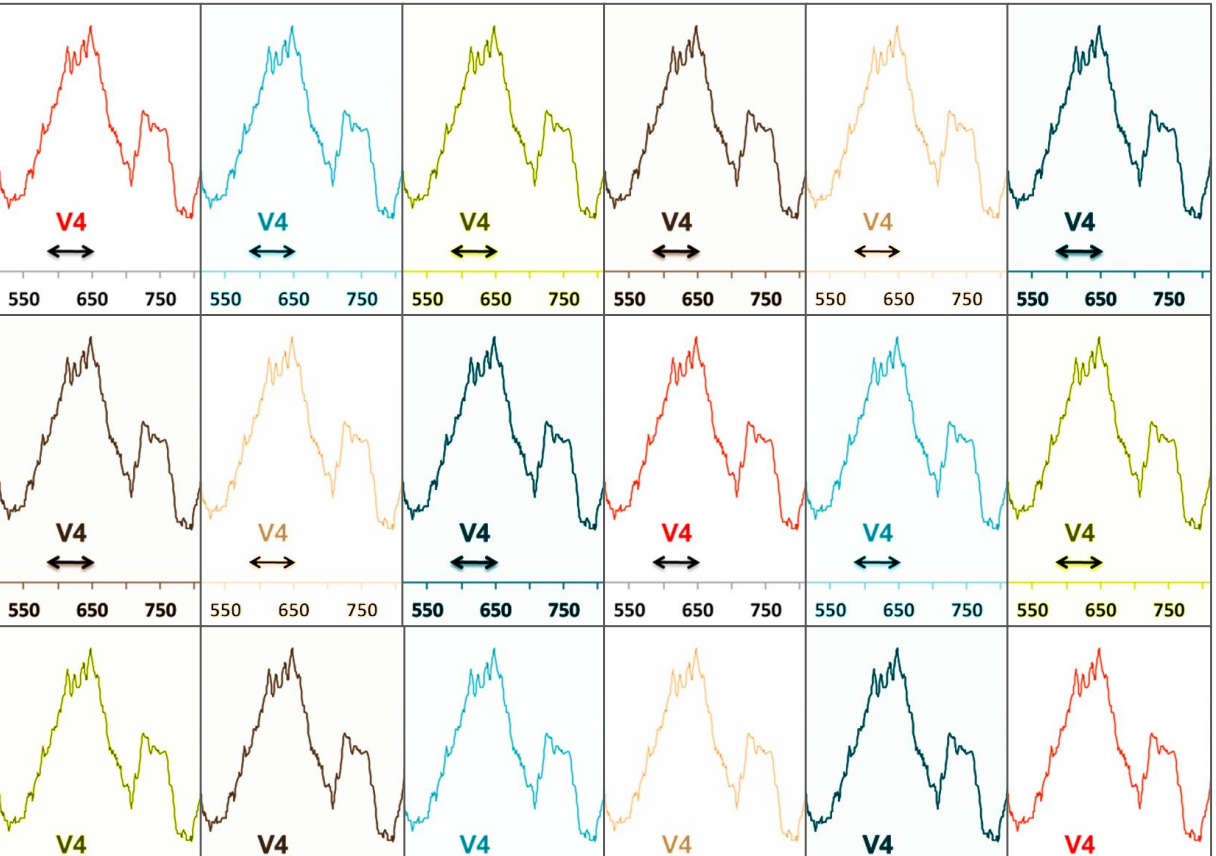
…

With *polymerase chain reaction (PCR)*, we can amplify specific regions of the genome using primers that *target* conserved regions.

Thompson et al. 2017: "Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair **515f–806r** […]"

S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One, 7*(8):e42671, 2012.

C. I. Methods: What's in a marker gene?



…

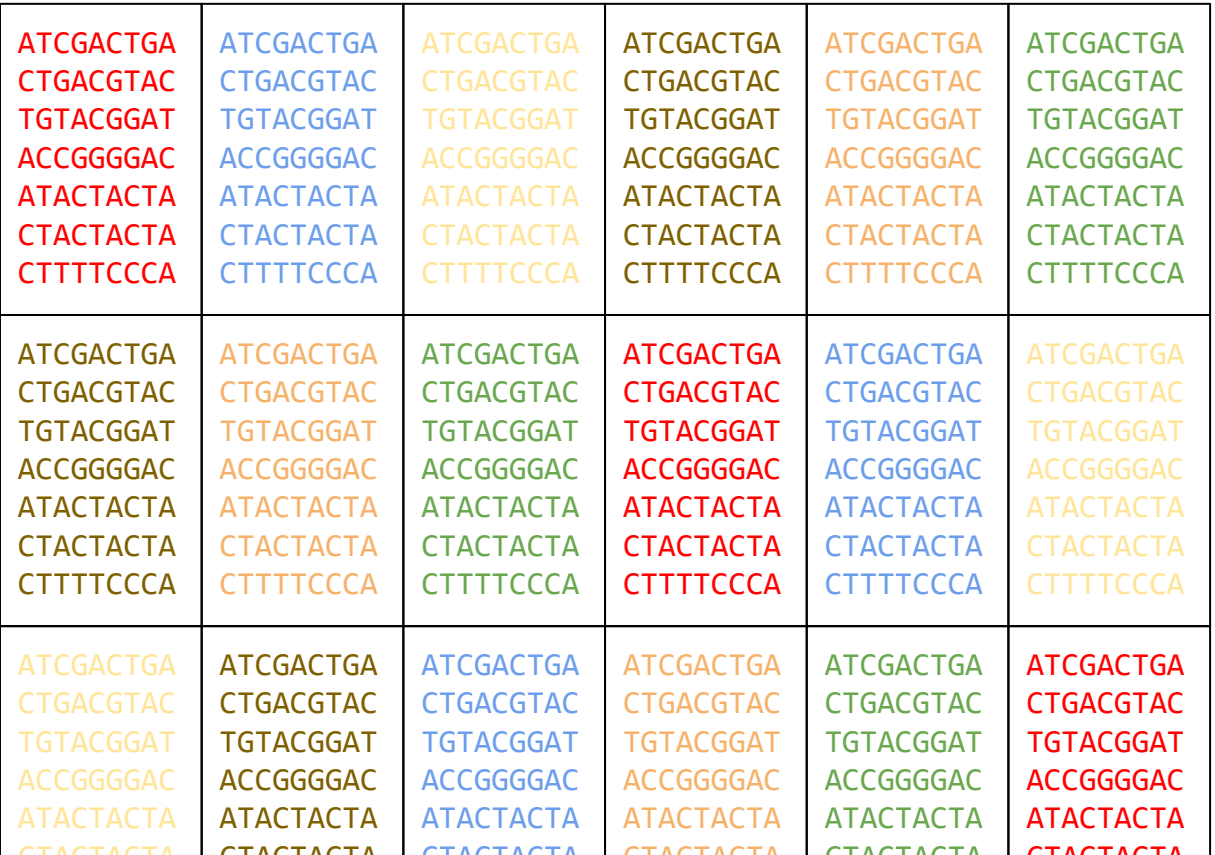
With *polymerase chain reaction (PCR)*, we can amplify specific regions of the genome using primers that *target* conserved regions.

Thompson et al. 2017: "Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair **515f–806r** […]"

Since these amplified sequences contain hypervariable region(s), these regions help us determine which sequences came from which microbe.

S. Vasileiadis, E. Puglisi, M. Arena, F. Cappa, P. S. Cocconcelli, and M. Trevisan. Soil bacterial diversity screening using single 16S rRNA gene V regions coupled with multi-million read generating sequencing technologies. *PLOS One, 7*(8):e42671, 2012.

C. I. Methods: What's in a marker gene?



…

With *polymerase chain reaction (PCR)*, we can amplify specific regions of the genome using primers that *target* conserved regions.

Thompson et al. 2017: "Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair **515f–806r** […]"

Since these amplified sequences contain hypervariable region(s), these regions help us determine which sequences came from which microbe.

C. I. Methods: What's in a marker gene?

k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__	k__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__Neisseriales; f__Neisseriaceae; g__Neisseria	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__	k__Bacteria; p__Firmicutes; c__Bacilli	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae; g__Haemophilus; s__parainfluenzae
k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__	k__Bacteria; p__Firmicutes; c__Bacilli	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae; g__Haemophilus; s__parainfluenzae	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides; s__	k__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__Neisseriales; f__Neisseriaceae; g__Neisseria	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus
k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptoco	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidac	k__Bacteria; p__Proteobacteria; c__Betaproteobacteria; o__Neisseriales; f__Neisseriace	k__Bacteria; p__Firmicutes; c__Bacilli	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae;	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidac

…

With *polymerase chain reaction (PCR)*, we can amplify specific regions of the genome using primers that *target* conserved regions.

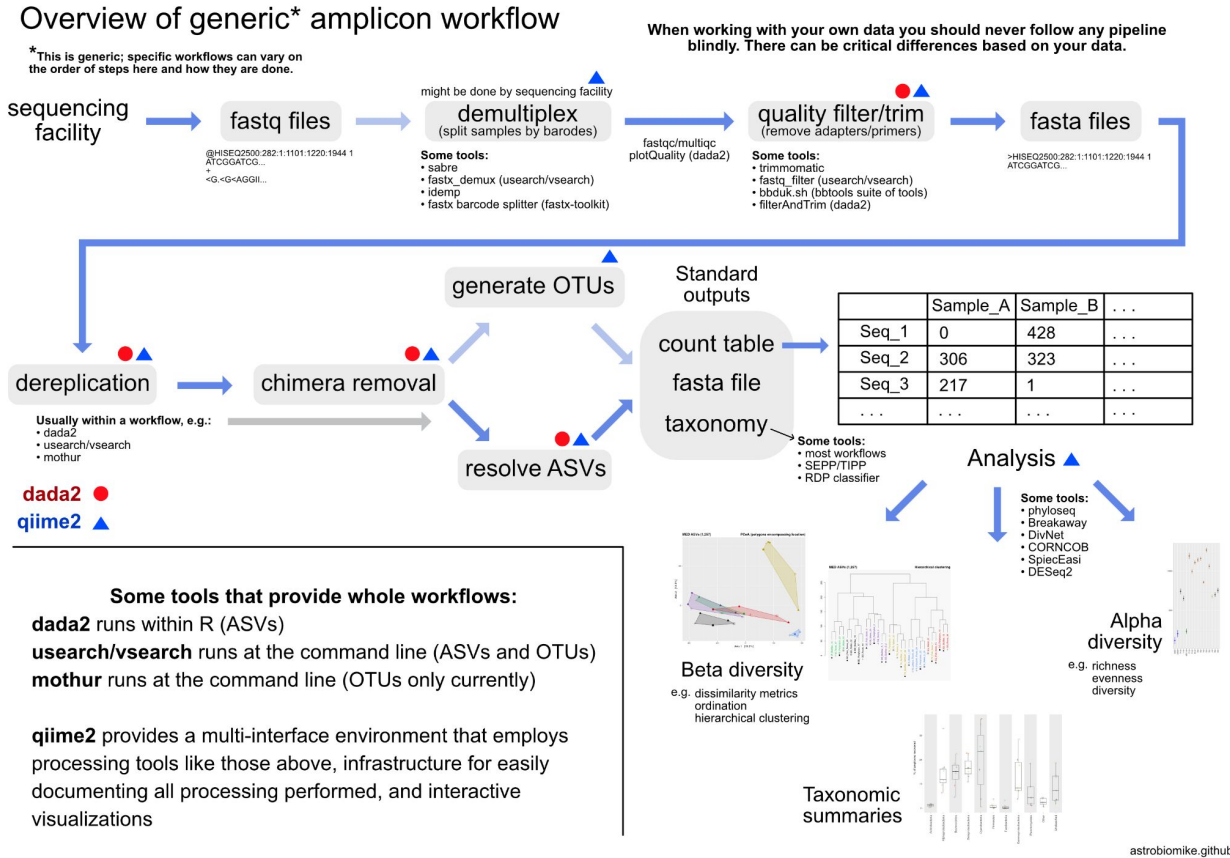
Thompson et al. 2017: "Amplicon PCR was performed on the V4 region of the 16S rRNA gene using the primer pair **515f–806r** […]"

Since these amplified sequences contain hypervariable region(s), these regions help us determine which sequences came from which microbe.

Example taxonomic annotations from the QIIME 2 "Moving Pictures" tutorial: https://docs.qiime2.org

Lee, M. D. (2019). Happy Belly Bioinformatics: An open-source resource dedicated to helping biologists utilize bioinformatics. *Journal of Open Source Education*, *4*(41), 53.

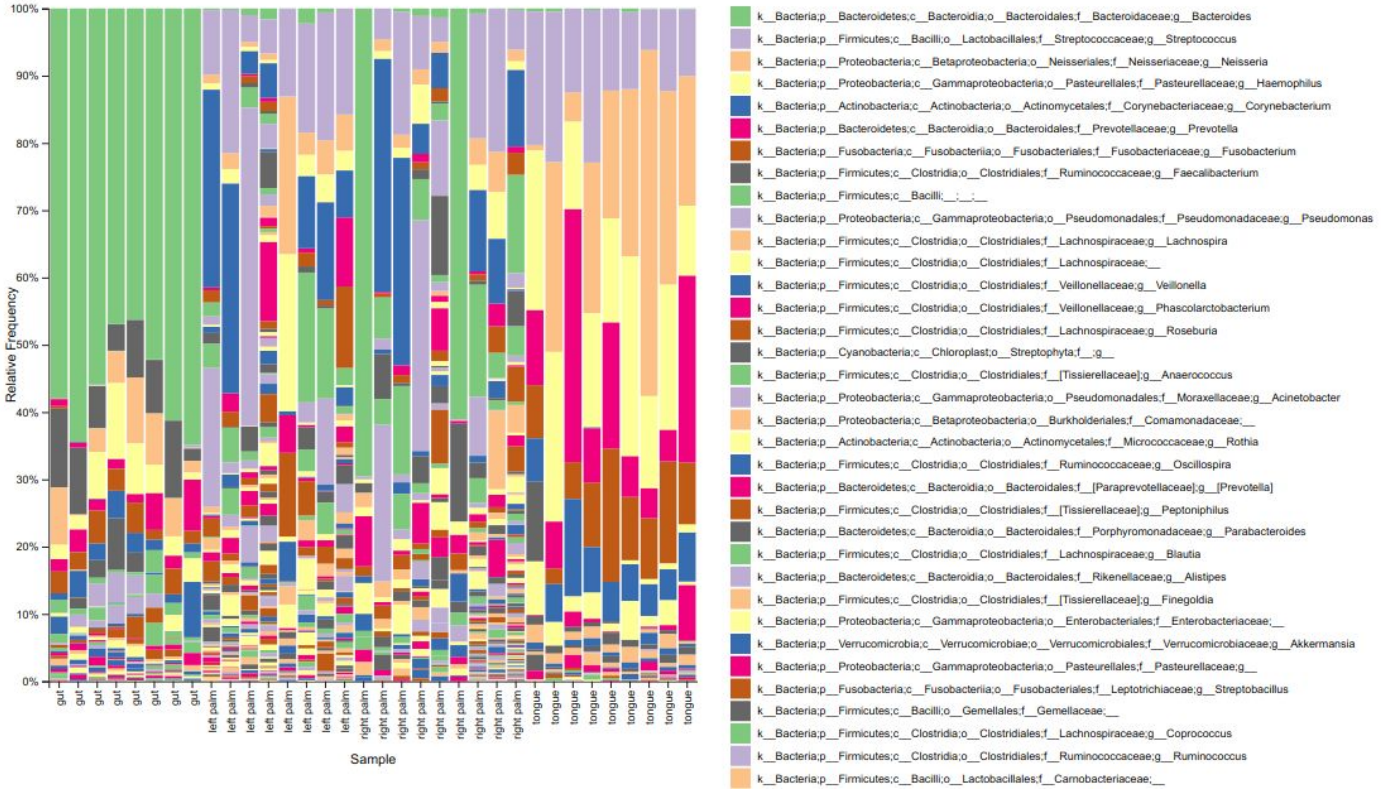
C. I. Methods: Marker gene sequencing



Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., ... & Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852-857.

C. I. Methods: Marker gene sequencing

Broad taxonomic comparisons

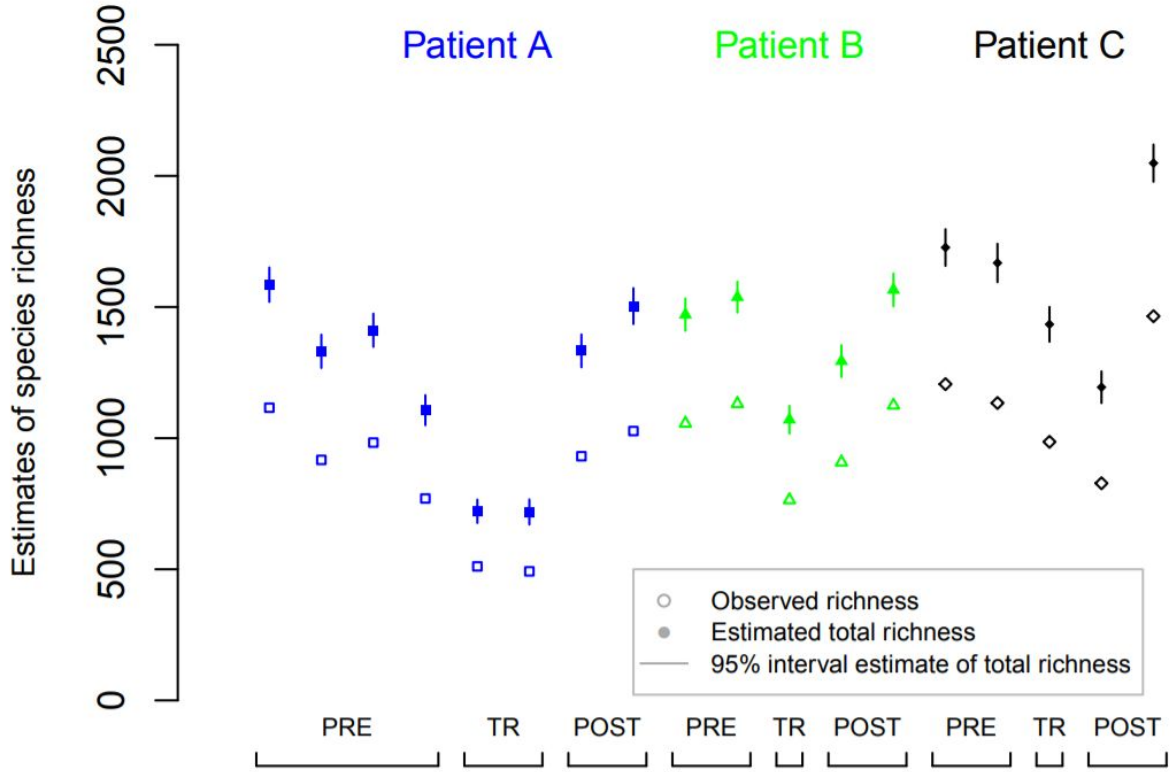


Willis, A., Bunge, J., & Whitman, T. (2017). Improved detection of changes in species richness in high diversity microbial communities. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *66*(5), 963-977. (Figure from *arXiv* manuscript version.)

C. I. Methods: Marker gene sequencing

Estimating and comparing diversity within samples

Also referred to as α-diversity ("alpha diversity")

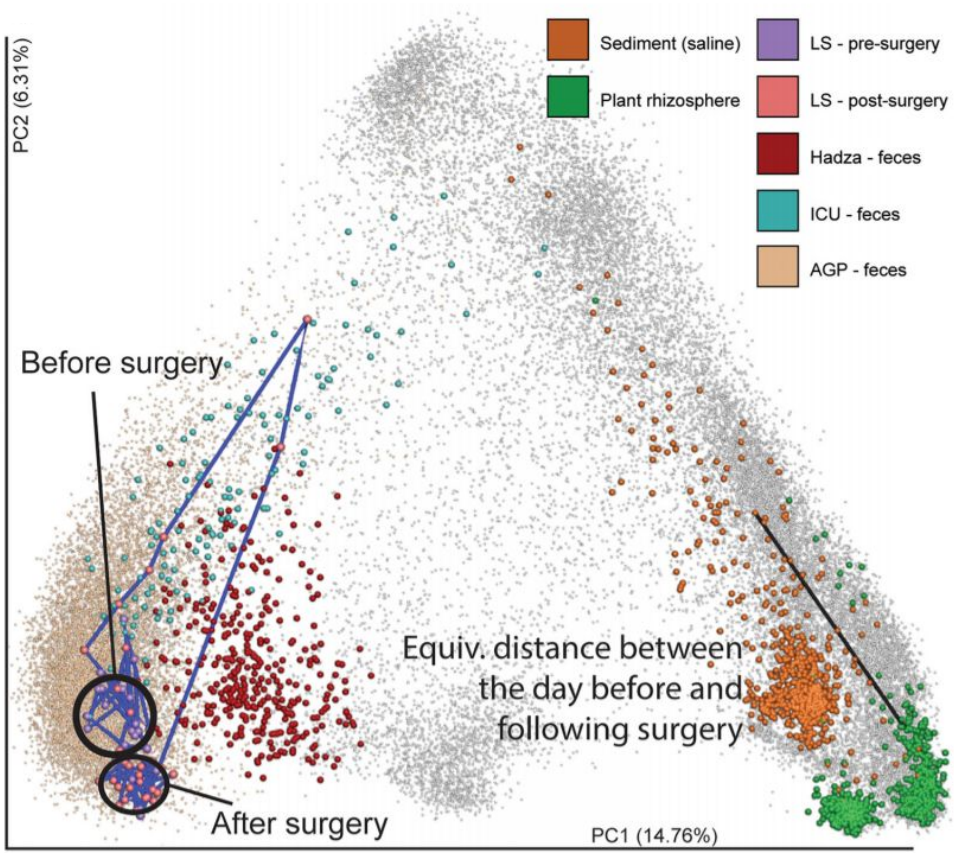


McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., ... & Knight, R. (2018). American Gut: an open platform for citizen science microbiome research. *mSystems*, *3*(3), e00031-18.

C. I. Methods: Marker gene sequencing

Unsupervised dimensionality reduction (e.g. PCA / PCoA)

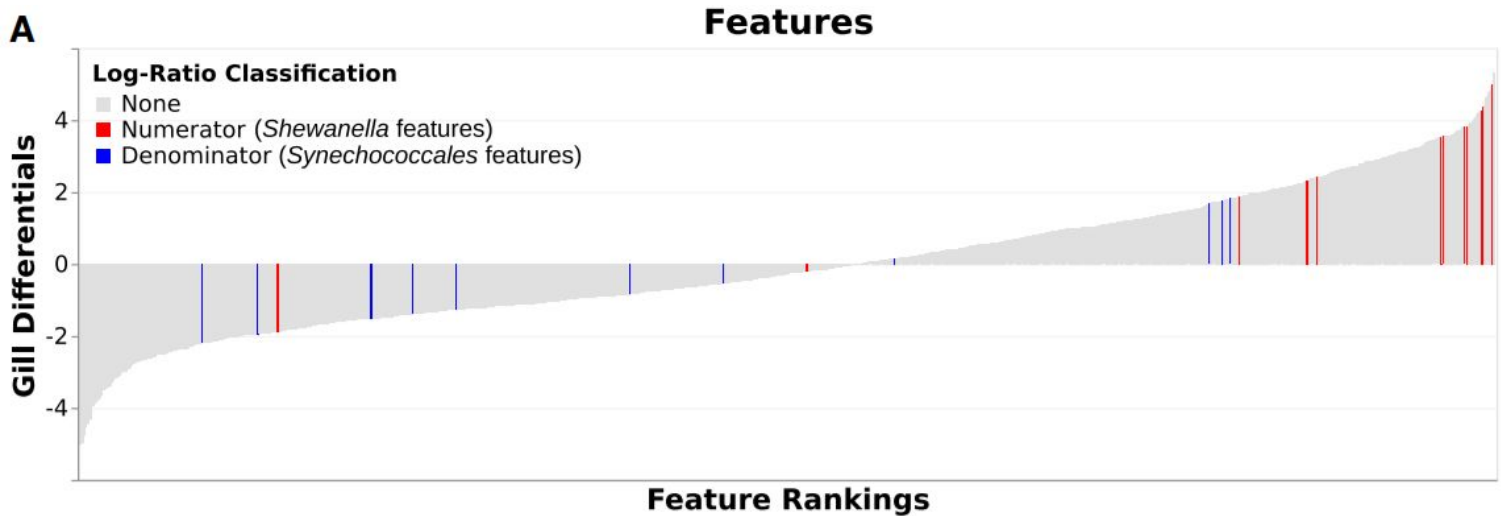
Also referred to as β-diversity ("beta diversity")

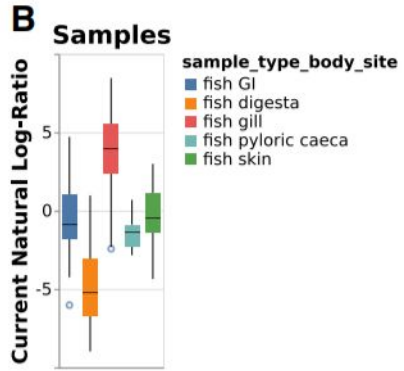


Fedarko, M. W., Martino, C., Morton, J. T., González, A., Rahman, G., Marotz, C. A., ... & Knight, R. (2020). Visualizing 'omic feature rankings and log-ratios using Qurro. *NAR Genomics and Bioinformatics*, *2*(2), lqaa023.

C. I. Methods: Marker gene sequencing

Identifying *differentially abundant* features (or ratios of features) in groups of samples



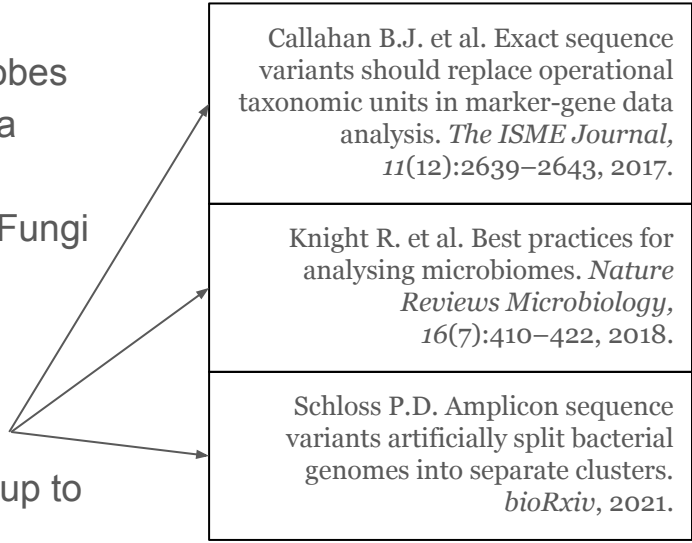


C. I. **Methods:** Marker gene sequencing

- **Pros**
  - Relatively cheap: much less sequencing needed to profile a community at a given depth than metagenome sequencing
  - Very well-studied, so many established pipelines have been created (mothur, QIIME / QIIME 2, …)
- **Cons**
  - Marker genes are usually specific to certain types of microbes
    - 16S rRNA gene: only identifies Bacteria and Archaea
    - 18S rRNA gene: only identifies certain Eukaryotes
    - Internal transcribed spacer region: mostly identifies Fungi
    - …
  - Copy number variation
  - PCR errors can result in “chimeric” gene sequences
  - Still disagreements on how to correct errors in raw reads
  - Limited resolution: marker genes are usually only reliable up to the genus level

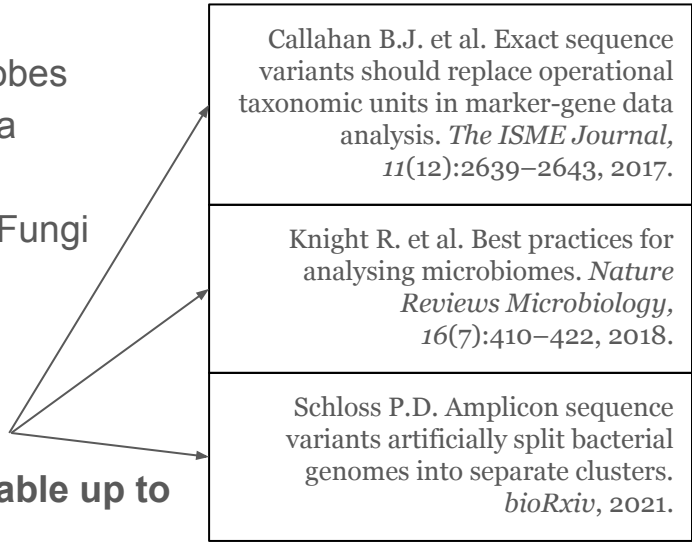
C. I. **Methods:** Marker gene sequencing

- **Pros**
  - Relatively cheap: much less sequencing needed to profile a community at a given depth than metagenome sequencing
  - Very well-studied, so many established pipelines have been created (mothur, QIIME / QIIME 2, …)
- **Cons**
  - Marker genes are usually specific to certain types of microbes
    - 16S rRNA gene: only identifies Bacteria and Archaea
    - 18S rRNA gene: only identifies certain Eukaryotes
    - Internal transcribed spacer region: mostly identifies Fungi
    - …
  - Copy number variation
  - PCR errors can result in “chimeric” gene sequences
  - Still disagreements on how to correct errors in raw reads
  - Limited resolution: marker genes are usually only reliable up to the genus level



C. I. **Methods:** Marker gene sequencing

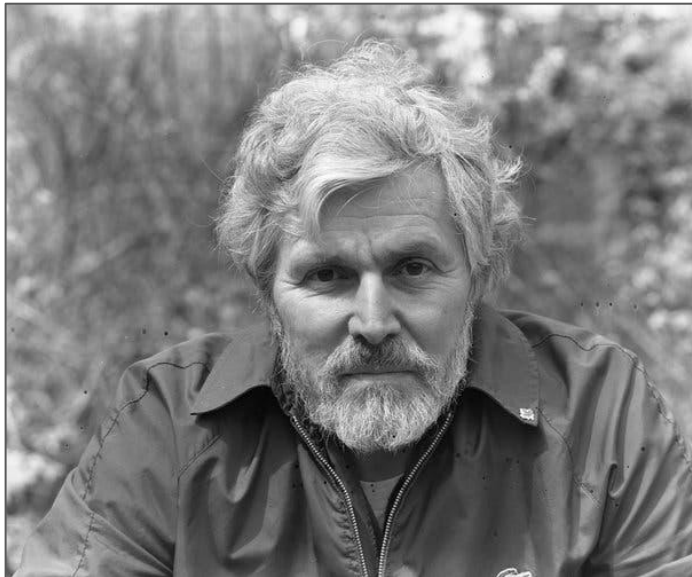
- **Pros**
  - Relatively cheap: much less sequencing needed to profile a community at a given depth than metagenome sequencing
  - Very well-studied, so many established pipelines have been created (mothur, QIIME / QIIME 2, …)
- **Cons**
  - Marker genes are usually specific to certain types of microbes
    - 16S rRNA gene: only identifies Bacteria and Archaea
    - 18S rRNA gene: only identifies certain Eukaryotes
    - Internal transcribed spacer region: mostly identifies Fungi
    - …
  - Copy number variation
  - PCR errors can result in “chimeric” gene sequences
  - Still disagreements on how to correct errors in raw reads
  - **Limited resolution: marker genes are usually only reliable up to the genus level**



For a more thorough history of Carl Woese's life and work, see:
D. Quammen. The Scientist Who Scrambled Darwin's Tree of Life. *The New York Times*, 2018.

C. I. Methods: back to Carl Woese and rRNA genes

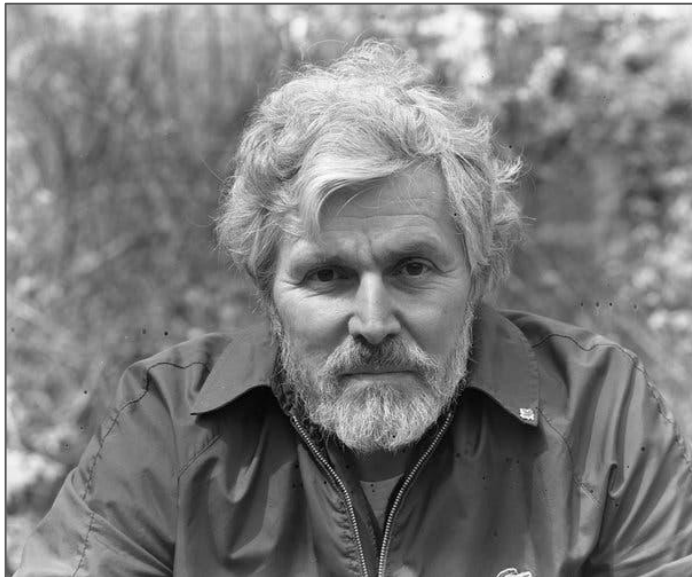
The original paper on PCR was published in **1986**.
The first automatic sequencer (the "AB370") was developed in **1987**.



For a more thorough history of Carl Woese's life and work, see:
D. Quammen. The Scientist Who Scrambled Darwin's Tree of Life. *The New York Times*, 2018.

C. I. Methods: back to Carl Woese and rRNA genes

The original paper on PCR was published in **1986**.
The first automatic sequencer (the "AB370") was developed in **1987**.
Woese and Fox's three-domain paper was published in **1977**!



For a more thorough history of Carl Woese's life and work, see:
D. Quammen. The Scientist Who Scrambled Darwin's Tree of Life. *The New York Times*, 2018.

C. I. Methods: back to Carl Woese and rRNA genes

The original paper on PCR was published in **1986**.
The first automatic sequencer (the "AB370") was developed in **1987**.
Woese and Fox's three-domain paper was published in **1977**!

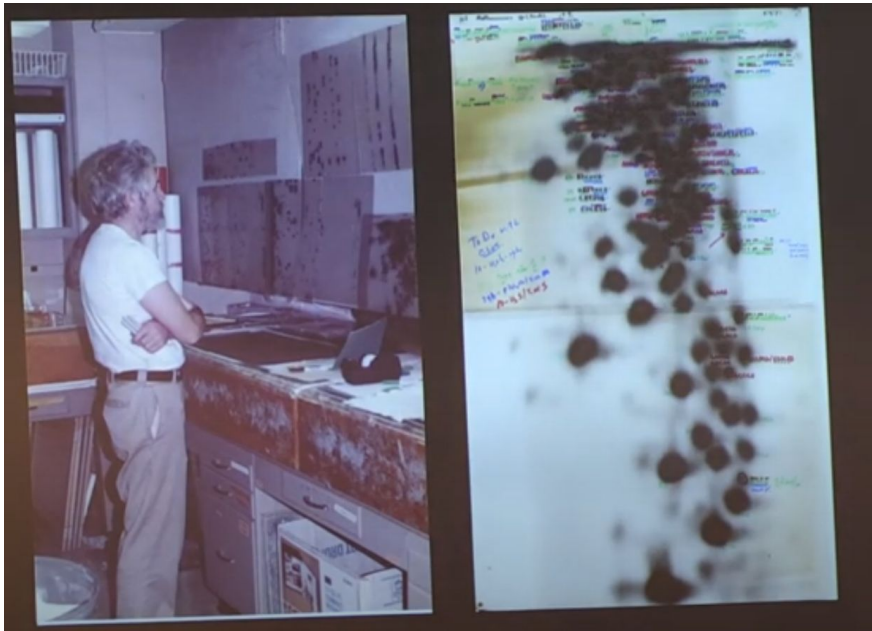


Photo credit: N. Pace. Carl Woese and the Beginnings of Metagenomics. *Looking in the Right Direction: Carl Woese and the New Biology,* 2015. https://www.youtube.com/watch?v=h3K50DD9kIM (timestamp: 2:56)

For a more thorough history of Carl Woese's life and work, see:
D. Quammen. The Scientist Who Scrambled Darwin's Tree of Life. *The New York Times*, 2018.

C. I. Methods: back to Carl Woese and rRNA genes

The original paper on PCR was published in **1986**.
The first automatic sequencer (the "AB370") was developed in **1987**.
Woese and Fox's three-domain paper was published in **1977**!

"While the grad students and technicians produced fingerprints, Woese spent his time staring at the spots. Was this effort tedious in practice as well as profound in its potential results? Yes. **'There were days,'** he wrote later, **'when I'd walk home from work saying to myself, 'Woese, you have destroyed your mind again today.' ' "**

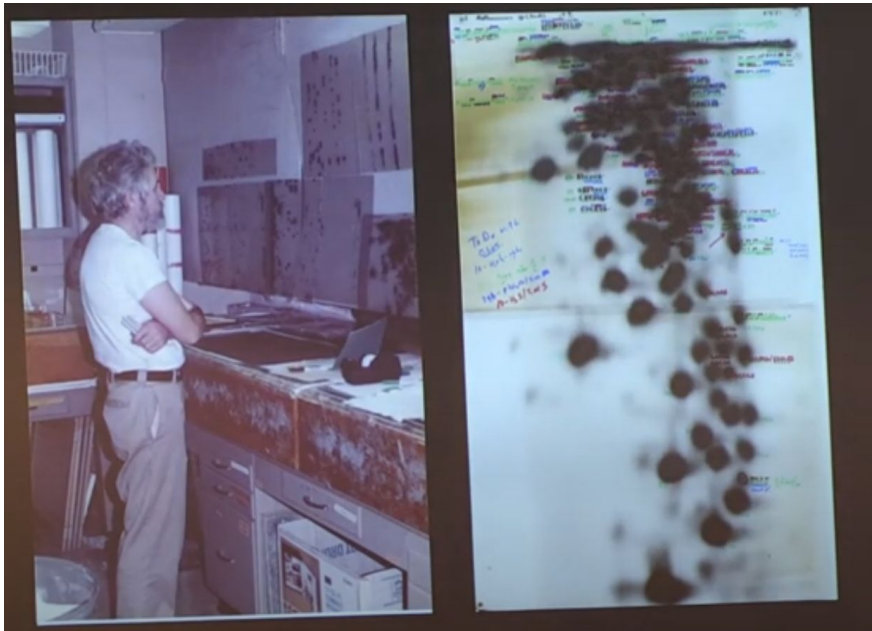
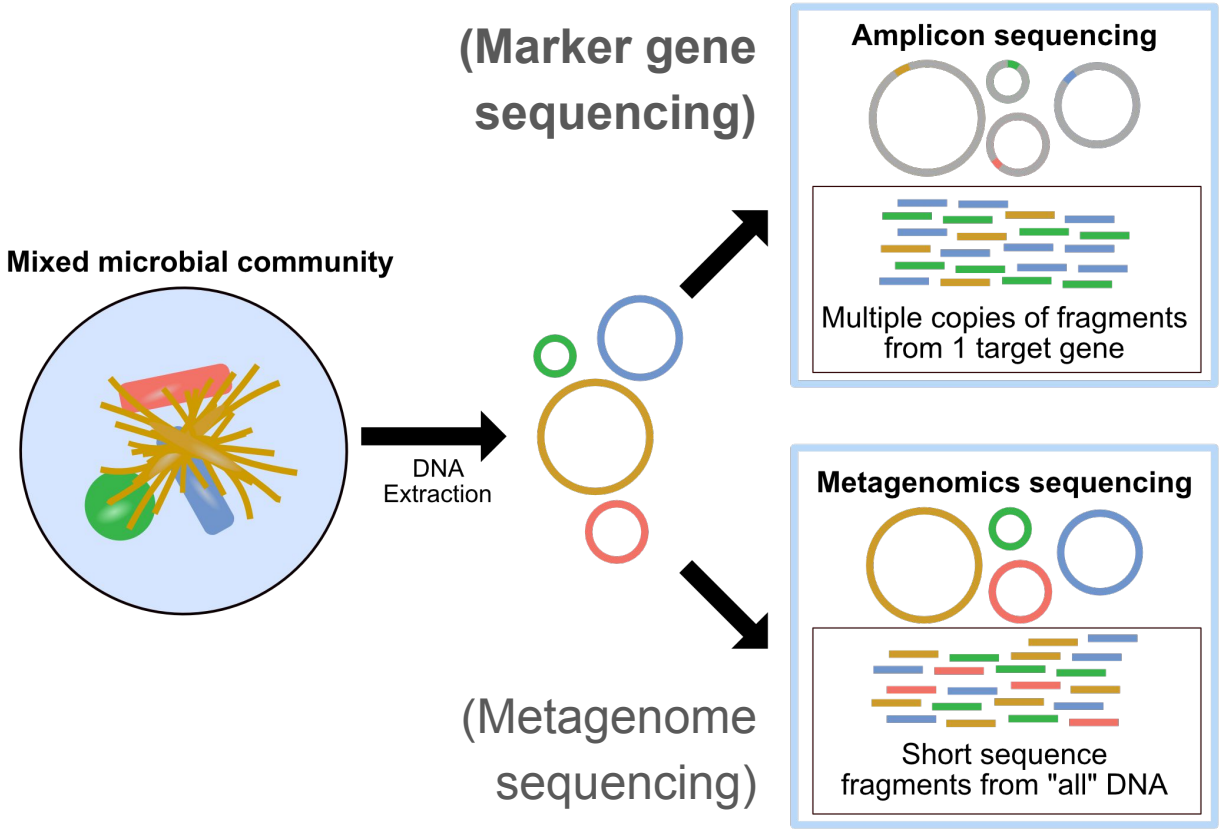


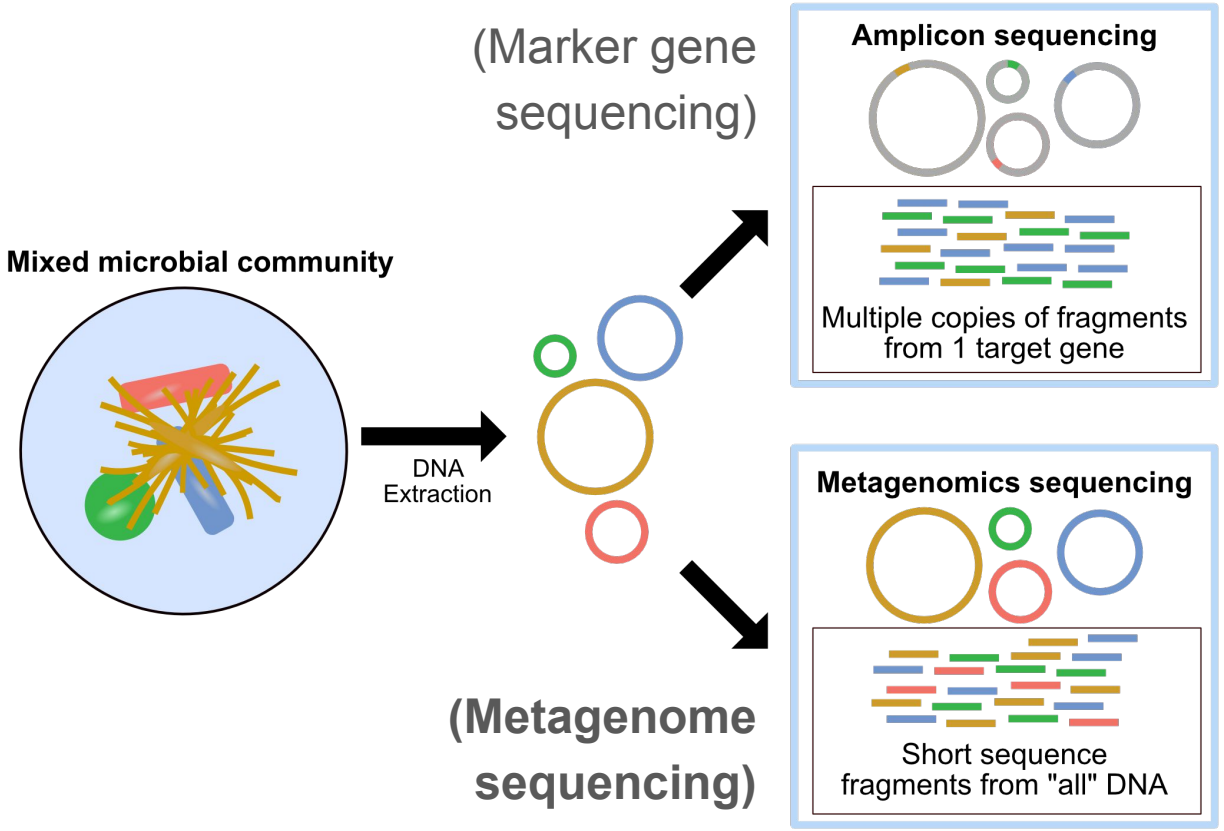
Photo credit: N. Pace. Carl Woese and the Beginnings of Metagenomics. *Looking in the Right Direction: Carl Woese and the New Biology,* 2015. https://www.youtube.com/watch?v=h3K50DD9kIM (timestamp: 2:56)

Lee, M. D. (2019). Happy Belly Bioinformatics: An open-source resource dedicated to helping biologists utilize bioinformatics. *Journal of Open Source Education*, *4*(41), 53.

C. I. Methods: Marker gene sequencing

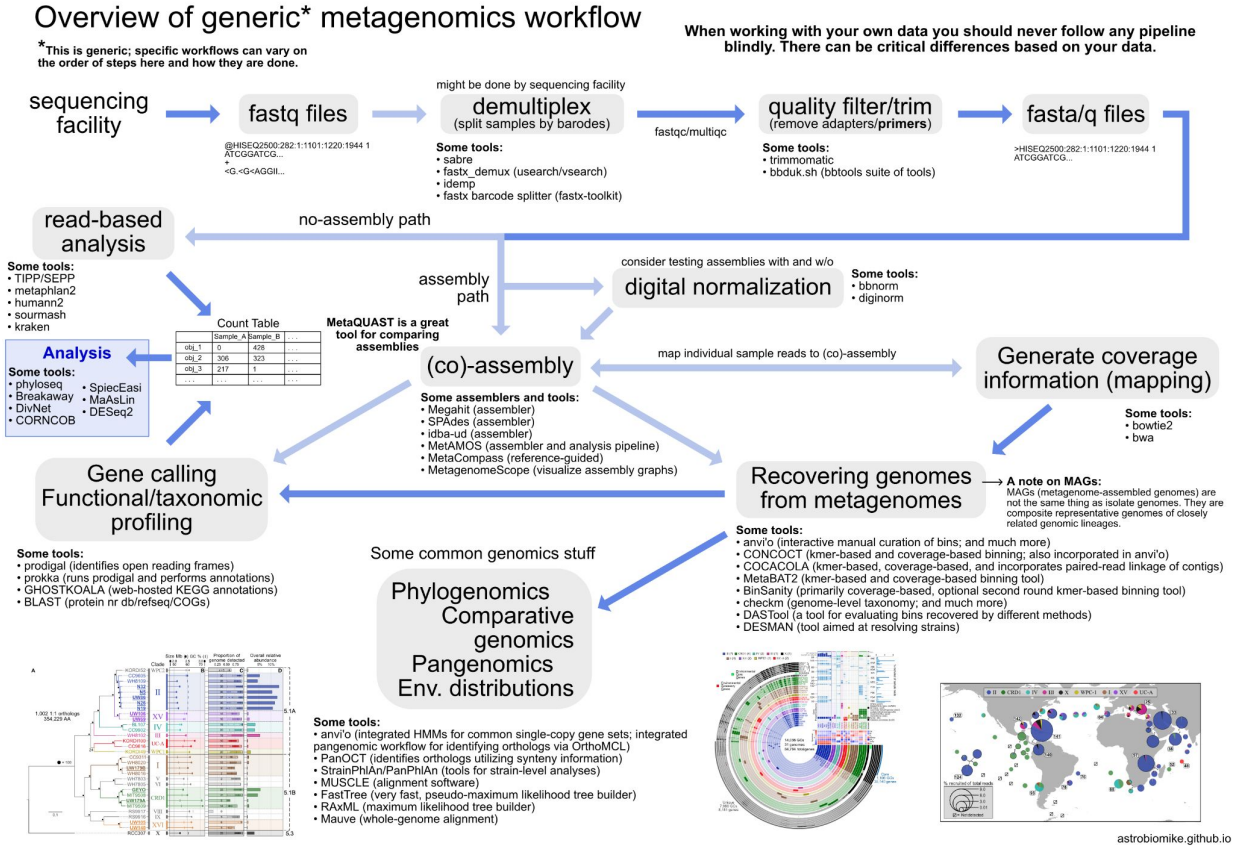


C. I. Methods: Metagenome sequencing



Lee, M. D. (2019). Happy Belly Bioinformatics: An open-source resource dedicated to helping biologists utilize bioinformatics. *Journal of Open Source Education*, *4*(41), 53.

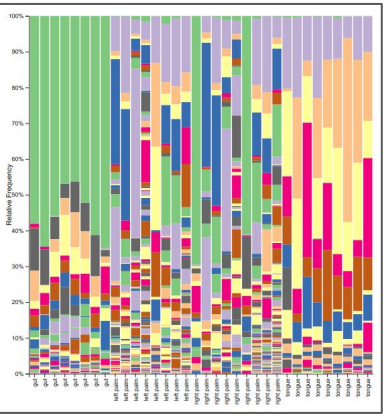
C. I. **Methods:** Metagenome sequencing



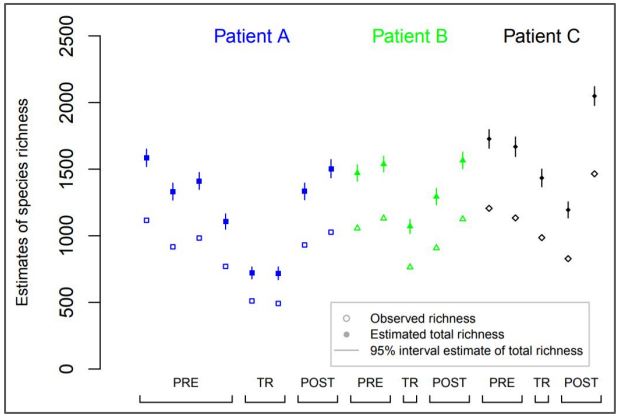
astrobiomike.github.io

C. I. **Methods:** Metagenome sequencing

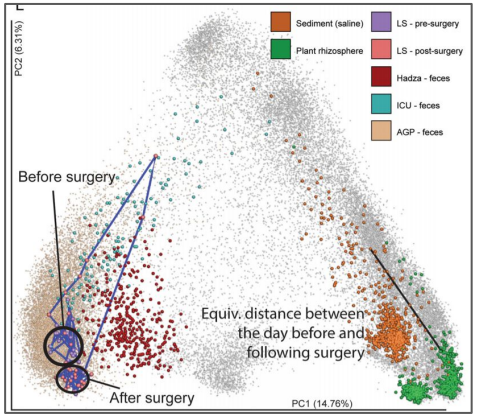
Taxonomy



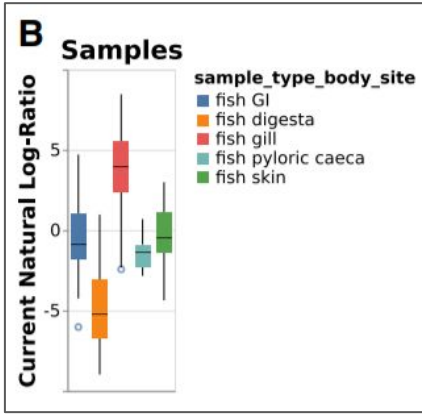
α-diversity



β-diversity



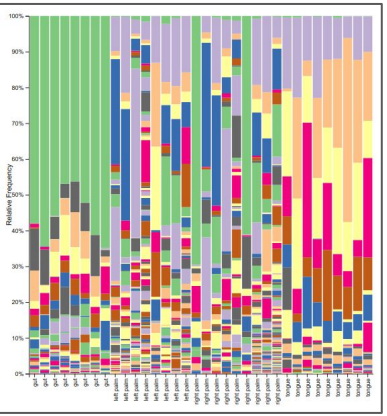
Differential abundance



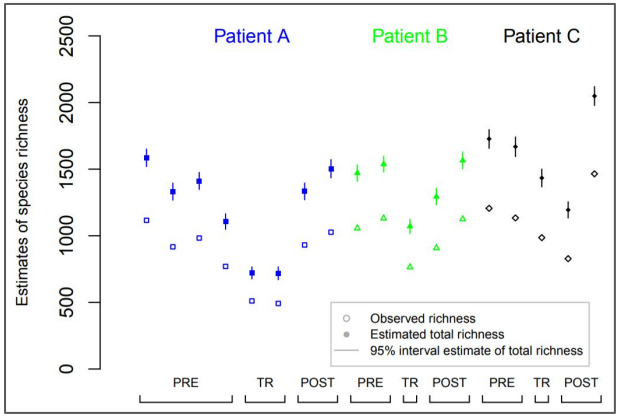
Many of the “standard analyses” for marker gene sequencing data are also applicable to metagenome sequencing data.

C. I. **Methods:** Metagenome sequencing

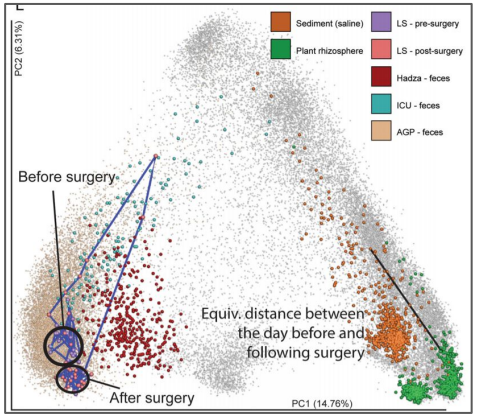
Taxonomy



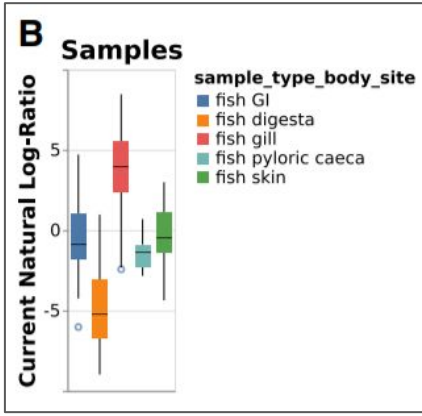
α-diversity



β-diversity



Differential abundance



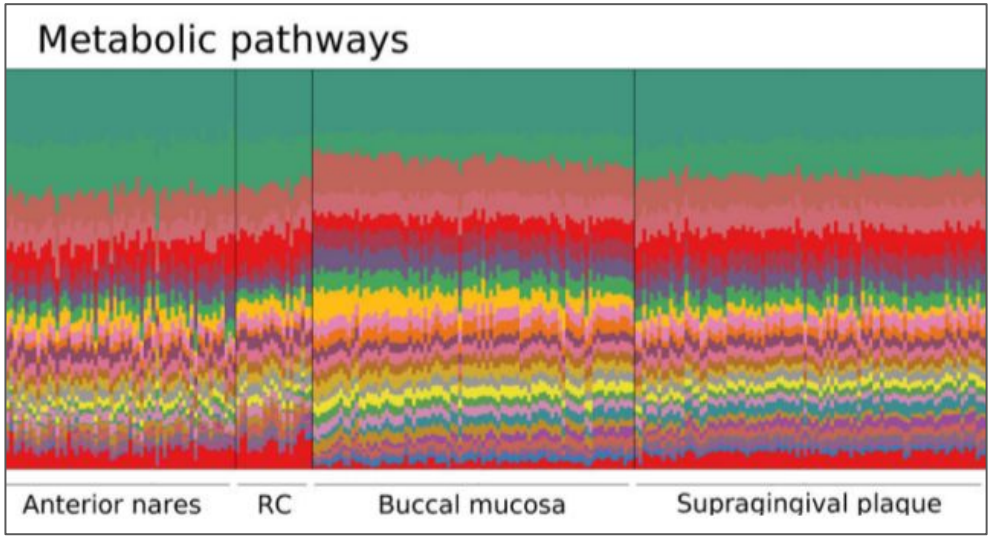
Metagenome sequencing enables two main additional types of analyses, compared to marker gene sequencing.

Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207.

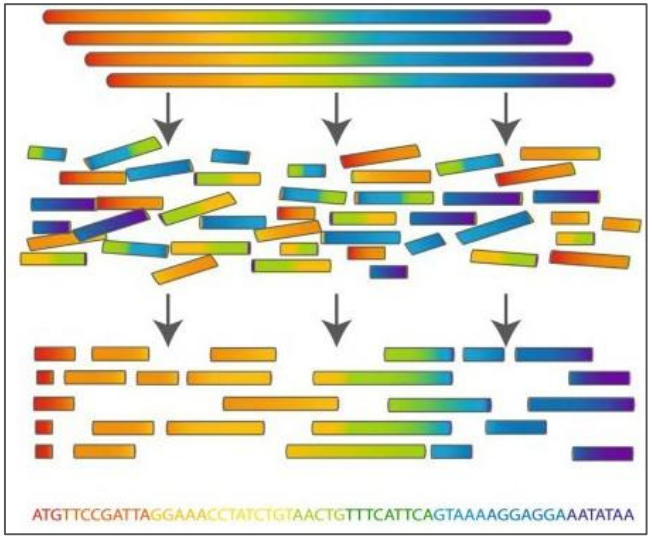
Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

C. I. Methods: Metagenome sequencing

Functional annotation



Sequence assembly



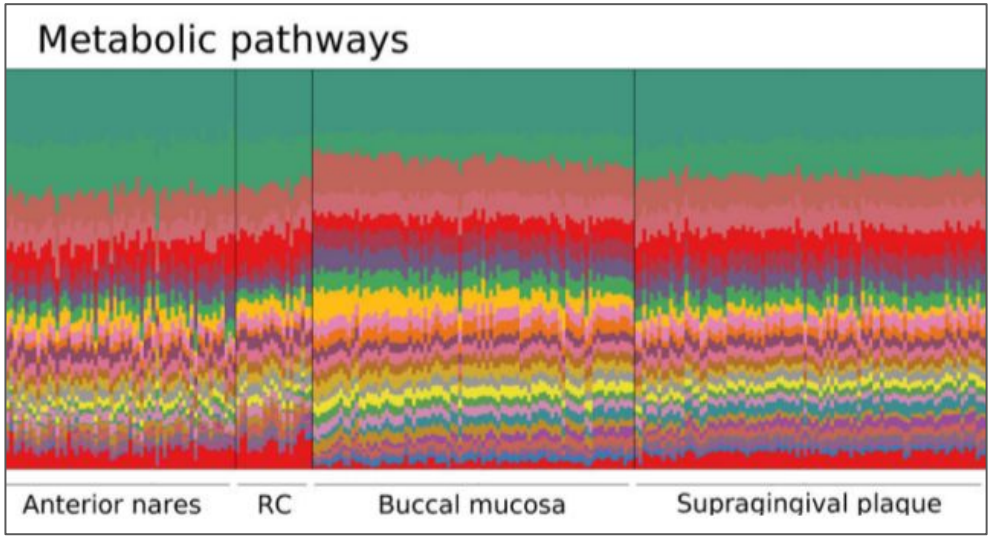
Metagenome sequencing enables two main additional types of analyses, compared to marker gene sequencing.

Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207.

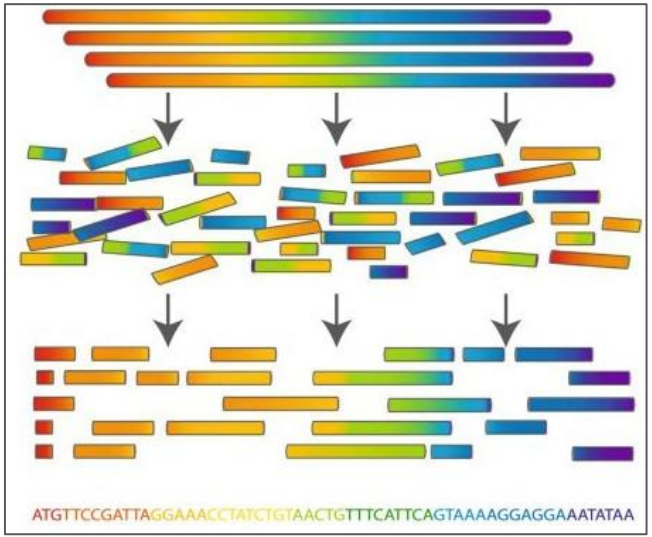
Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

C. I. Methods: Metagenome sequencing

Functional annotation

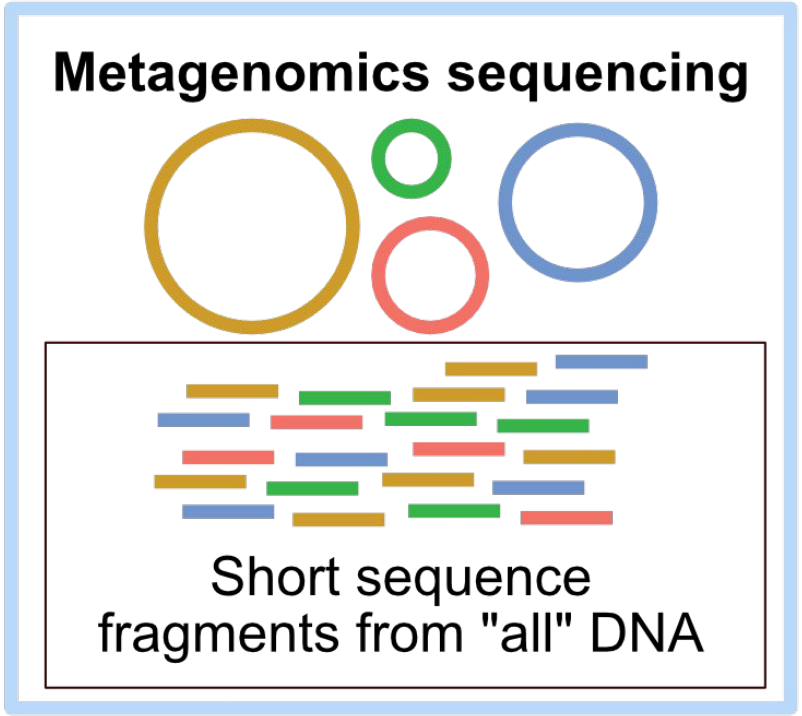


Sequence assembly

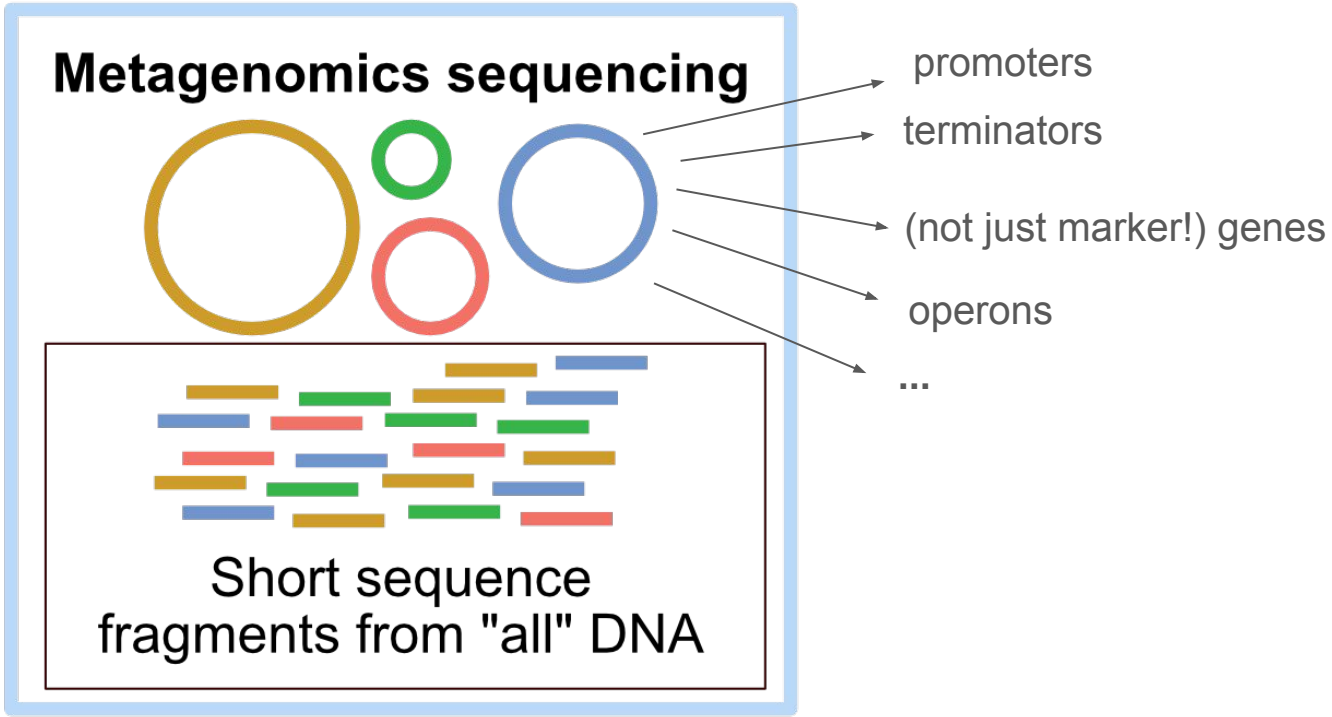


Metagenome sequencing enables two main additional types of analyses, compared to marker gene sequencing.

C. I. Methods: Functional annotation



C. I. Methods: Functional annotation

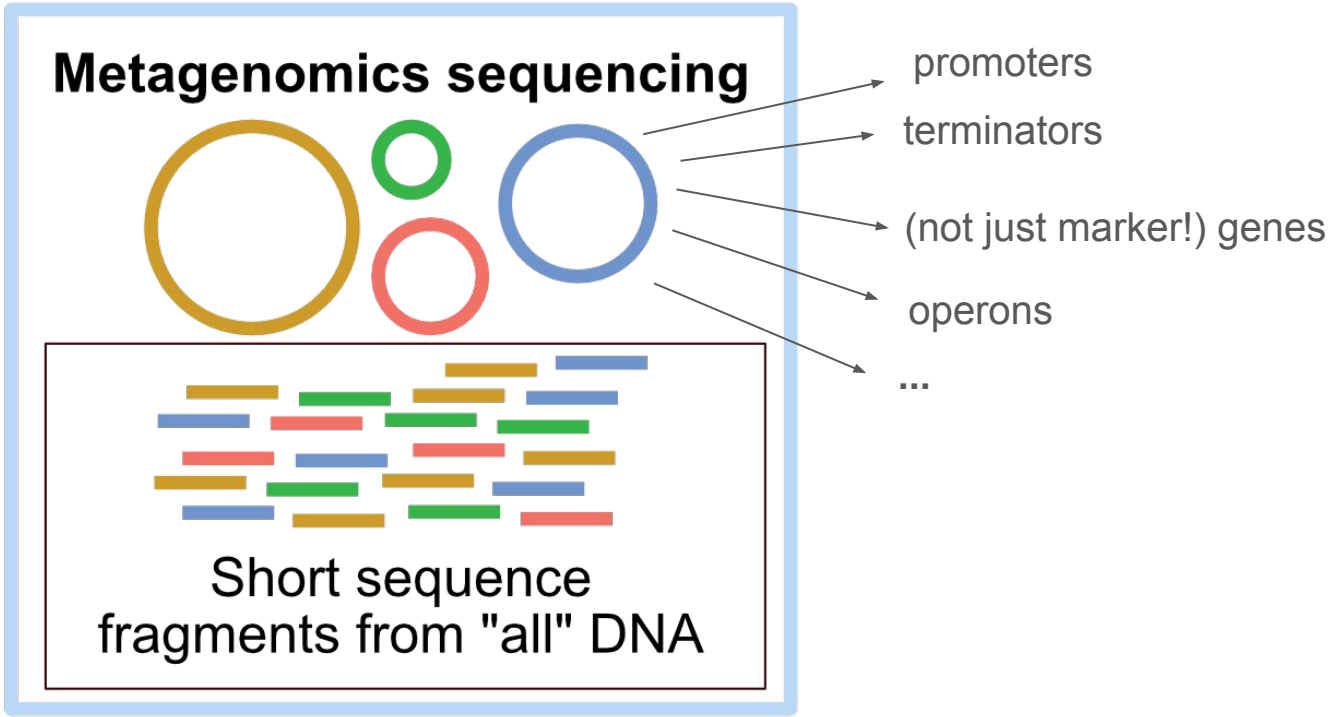


For more information on metagenomic functional annotation, see: Salamov, V. S. A., & Solovyevand, A. (2011). Automatic annotation of microbial genomes and metagenomic sequences. *Metagenomics and Its Applications in Agriculture, Biomedicine and Environmental Studies; Li, RW, Ed*, 61-78.

C. I. Methods: Functional annotation, in practice

Usually, F.A. involves aligning ("mapping") sequences to a reference database with information about "function" in well-studied organisms.

(But there are fancier approaches, e.g. metabolic modelling methods.)



For more information on metagenomic functional annotation, see: Salamov, V. S. A., & Solovyevand, A. (2011). Automatic annotation of microbial genomes and metagenomic sequences. *Metagenomics and Its Applications in Agriculture, Biomedicine and Environmental Studies; Li, RW, Ed*, 61-78.

C. I. Methods: Functional annotation, in practice?

An Integrated Encyclopedia of DNA Elements in the Human Genome

The ENCODE Project Consortium

"The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure, and histone modification. These data enabled us to **assign biochemical functions for 80% of the [human] genome**, in particular outside of the well-studied protein-coding regions."

C. I. Methods: Functional annotation, in practice??

An Integrated Encyclopedia of DNA Elements in the Human Genome

The ENCODE Project Consortium

"The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure, and histone modification. These data enabled us to **assign biochemical functions for 80% of the [human] genome**, in particular outside of the well-studied protein-coding regions."

On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE

Dan Graur[1,*], Yichen Zheng[1], Nicholas Price[1], Ricardo B.R. Azevedo[1], Rebecca A. Zufall[1], and Eran Elhaik[2]

A recent slew of ENCyclopedia Of DNA Elements (ENCODE) Consortium publications, specifically the article signed by all Consortium members, put forward the idea that more than 80% of the human genome is functional. **This claim flies in the face of current estimates** [...]

C. I. Methods: Functional annotation, in practice???

An Integrated Encyclopedia of DNA Elements in the Human Genome

The ENCODE Project Consortium

"The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure, and histone modification. These data enabled us to **assign biochemical functions for 80% of the [human] genome**, in particular outside of the well-studied protein-coding regions."

On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE

Dan Graur[1,*], Yichen Zheng[1], Nicholas Price[1], Ricardo B.R. Azevedo[1], Rebecca A. Zufall[1], and Eran Elhaik[2]

A recent slew of ENCyclopedia Of DNA Elements (ENCODE) Consortium publications, specifically the article signed by all Consortium members, put forward the idea that more than 80% of the human genome is functional. **This claim flies in the face of current estimates** [...]

Is junk DNA bunk? A critique of ENCODE

W. Ford Doolittle[1]

Junk or functional DNA? ENCODE and the function controversy

Pierre-Luc Germain · Emanuele Ratti · Federico Boem

The ENCODE project: Missteps overshadowing a success

A farewell to bioinformatics

by Fred Ross

Last updated: March 26, 2012

genome being functionally annotated. It says something of our ability to annotate genomes, that the proportion of a genome functionally annotated is often correlated to the genetic distance to the very well researched *Escherichia coli* (anecdotal observation). How-

C. I. Methods: Functional annotation, in practice???

An Integrated Encyclopedia of DNA E… the Human Genome

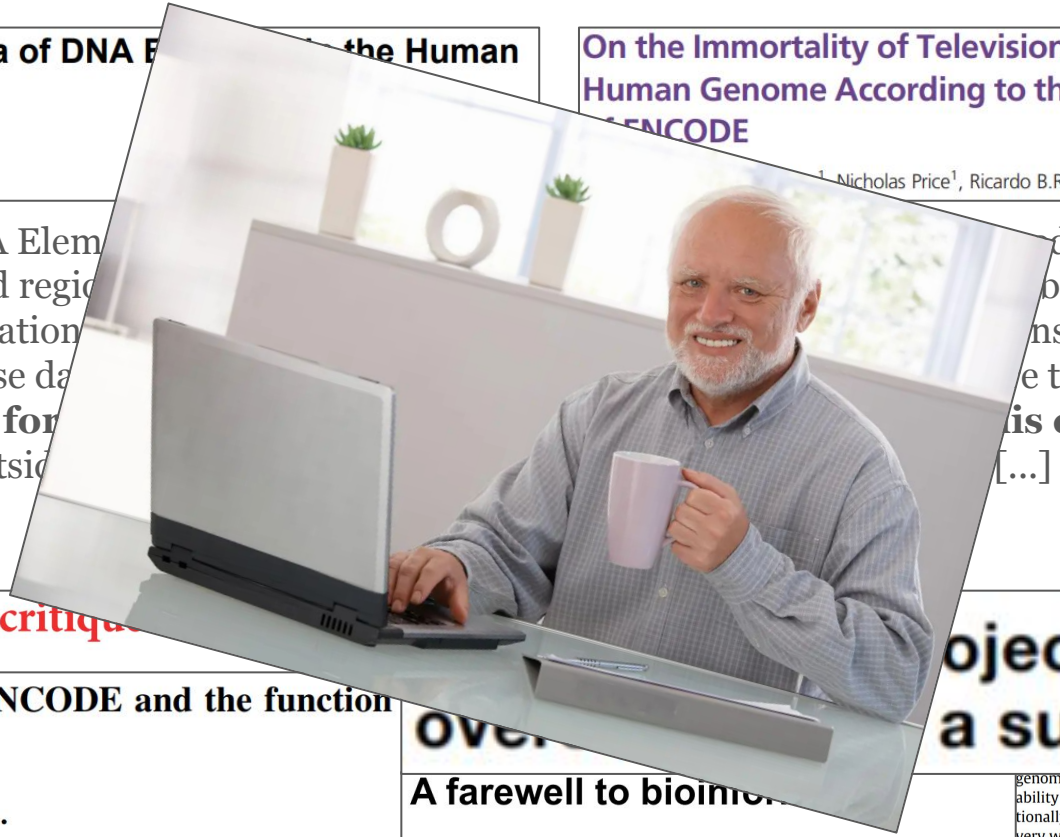
The ENCODE Project Consortium

"The Encyclopedia of DNA Elem… has systematically mapped regi… transcription factor association… histone modification. These da… **biochemical functions for… genome**, in particular outsi… protein-coding regions."

On the Immortality of Television Sets: "Function" in the Human Genome According to the Evolution-Free Gospel of ENCODE

… Nicholas Price[1], Ricardo B.R. Azevedo[1], Rebecca A. Zufall[1], and Eran Elhaik[2]

…dia Of DNA Elements …blications, specifically …nsortium members, put …e than 80% of the human **…is claim flies in the face** […]



Is junk DNA bunk? A critiqu…

W. Ford Doolittle[1]

Junk or functional DNA? ENCODE and the function… controversy

Pierre-Luc Germain · Emanuele Ratti · Federico Boem

…oject: Missteps … a success

A farewell to bioinf…

by Fred Ross Last updated: March 26, 2012

genome being functionally annotated. It says something of our ability to annotate genomes, that the proportion of a genome functionally annotated is often correlated to the genetic distance to the very well researched *Escherichia coli* (anecdotal observation). How-

Wooley, J. C., Godzik, A., & Friedberg, I. (2010). A primer on metagenomics. *PLOS computational biology*, *6*(2), e1000667.

C. I. Methods: Functional annotation, in practice

Table 2. The information contained in different lengths of genomic DNA.

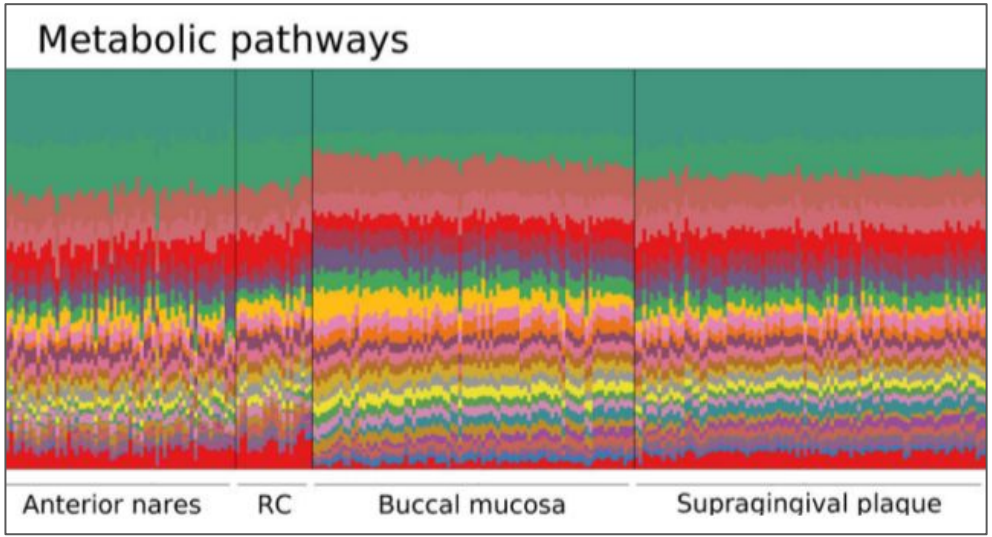
Sequence Length (bp)	Genome Element
25–75	SNPs, short frameshift mutations
100–400	Short functional signatures
500–1,000	Whole domains, single domain genes
1,000–5,000	Short operons, multidomain genes
5,000–10,000	Longer operons, some *cis*-control elements
>100,000	Prophages, pathogenicity islands, various mobile insertion elements
>1,000,000	Whole prokaryotic chromosome organization

Human Microbiome Project Consortium. (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, *486*(7402), 207.

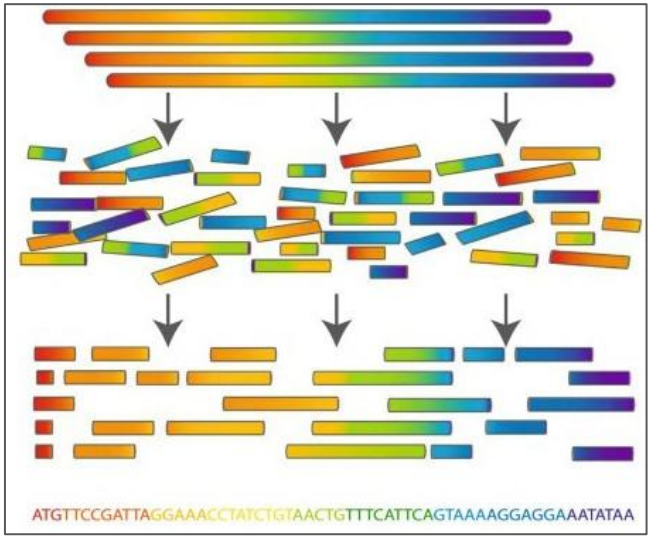
Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

C. I. Methods: Metagenome sequencing

Functional annotation



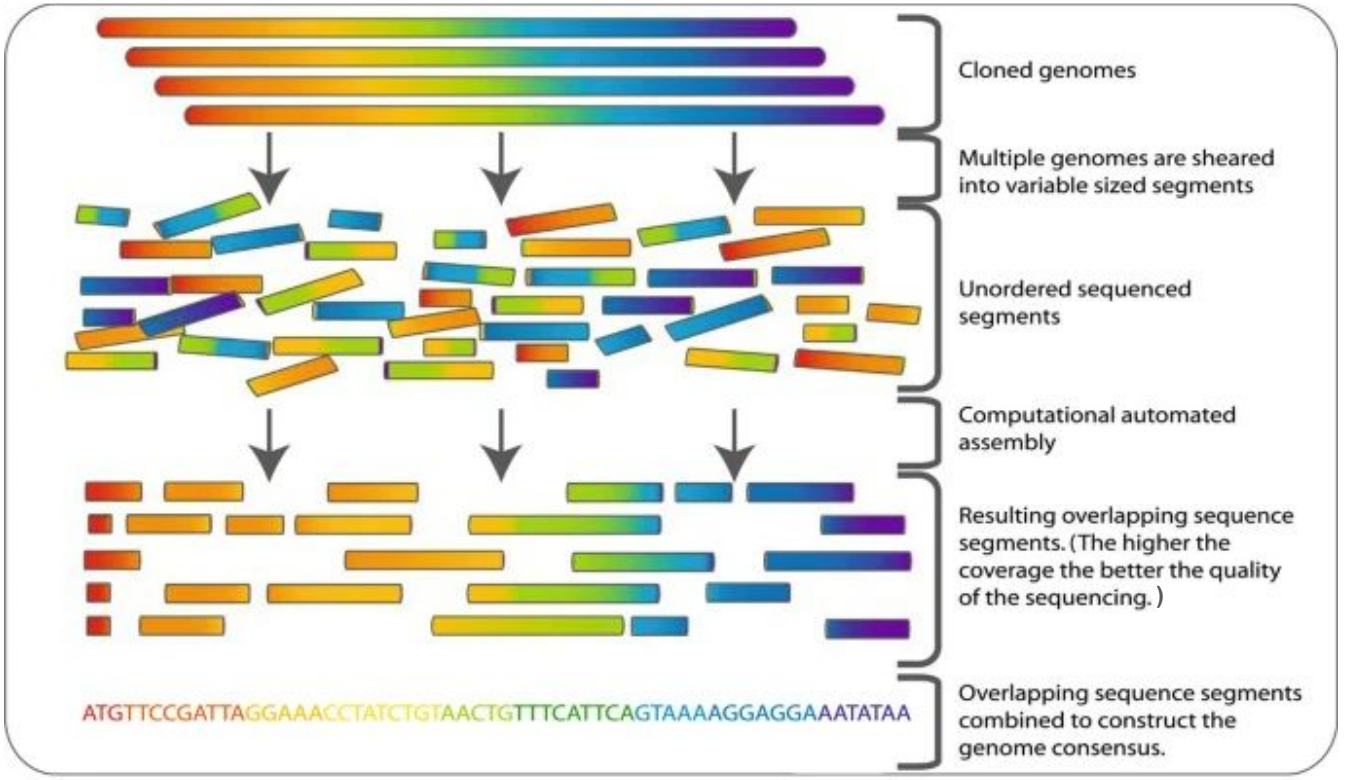
Sequence assembly



Metagenome sequencing enables two main additional types of analyses, compared to marker gene sequencing.

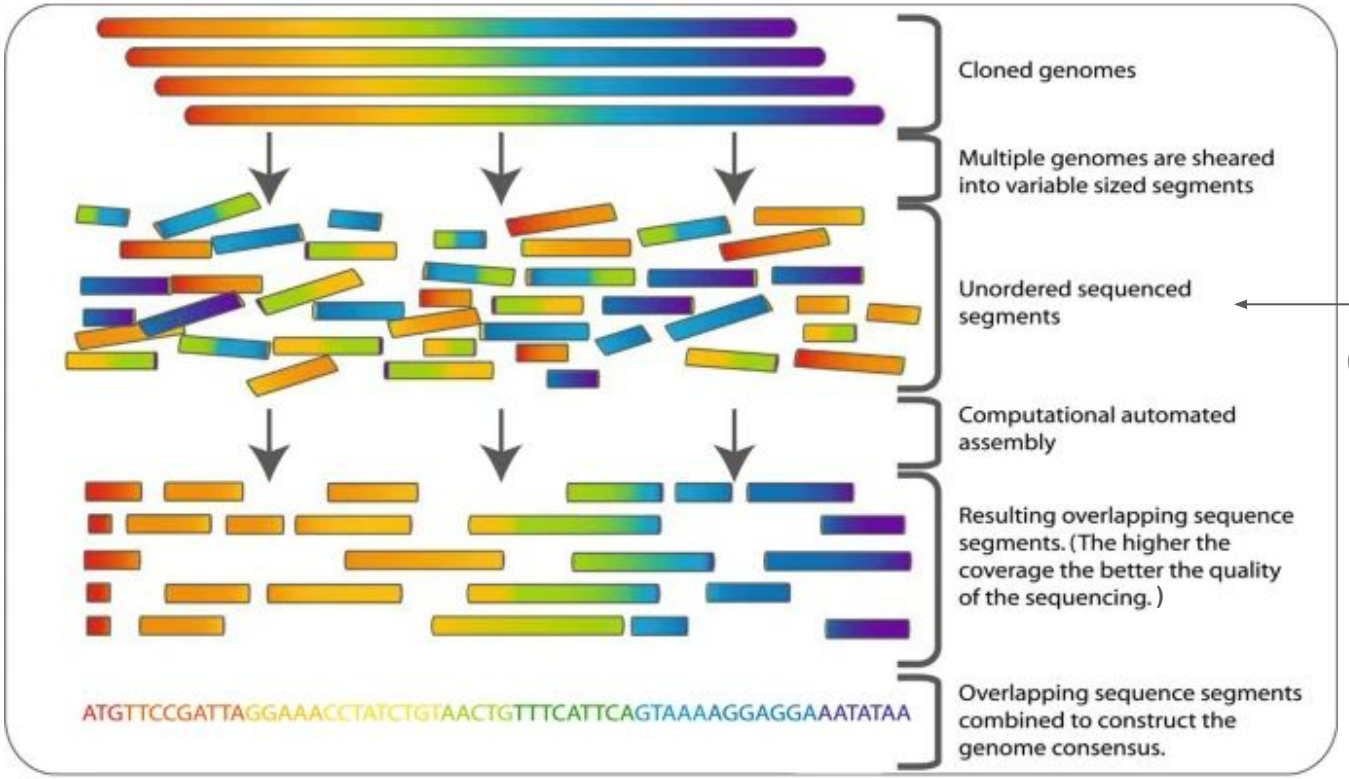
Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

C. I. Methods: Metagenome sequence assembly



Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

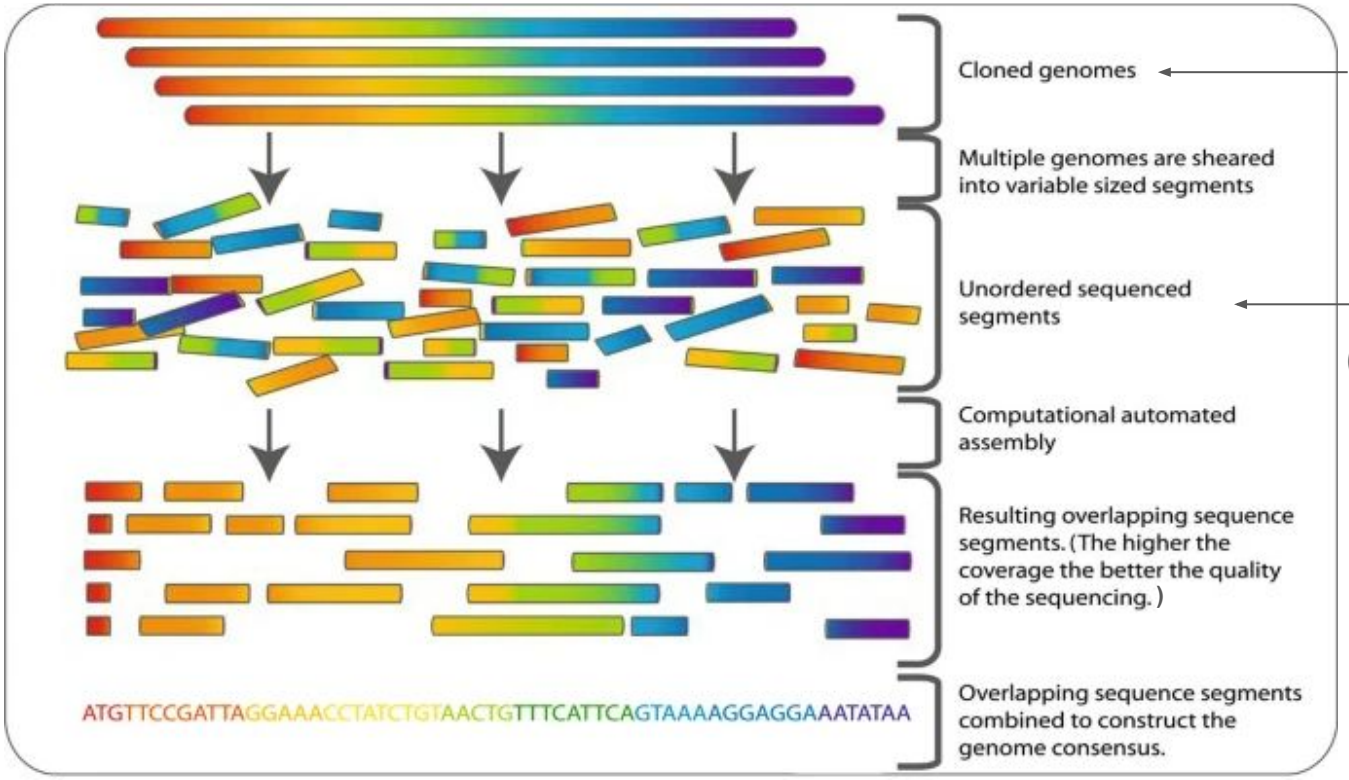
C. I. Methods: Metagenome sequence assembly



"Reads" (Strings from $\Sigma = \{A,C,G,T\}$)

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

C. I. Methods: Metagenome sequence assembly

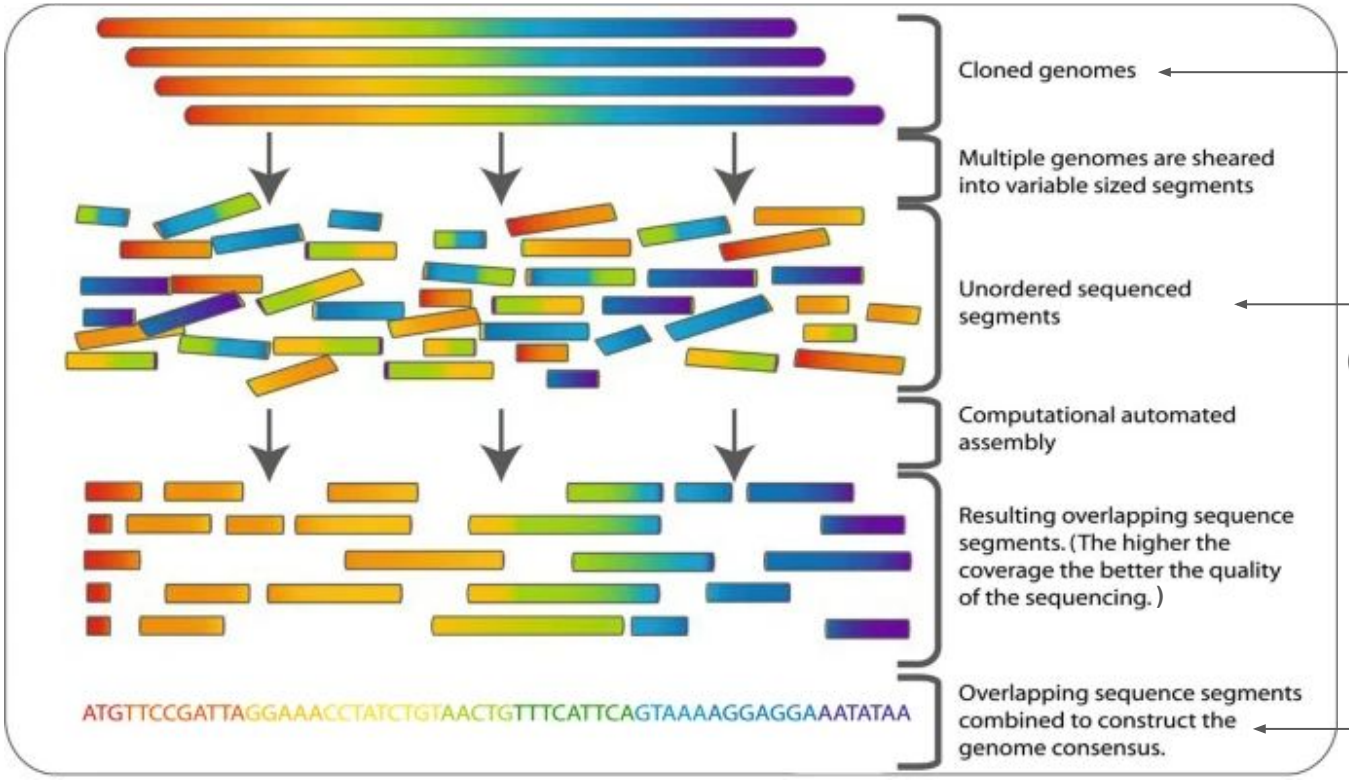


In practice there'll be many input molecules here...

"Reads" (Strings from $\Sigma = \{A,C,G,T\}$)

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

C. I. Methods: Metagenome sequence assembly



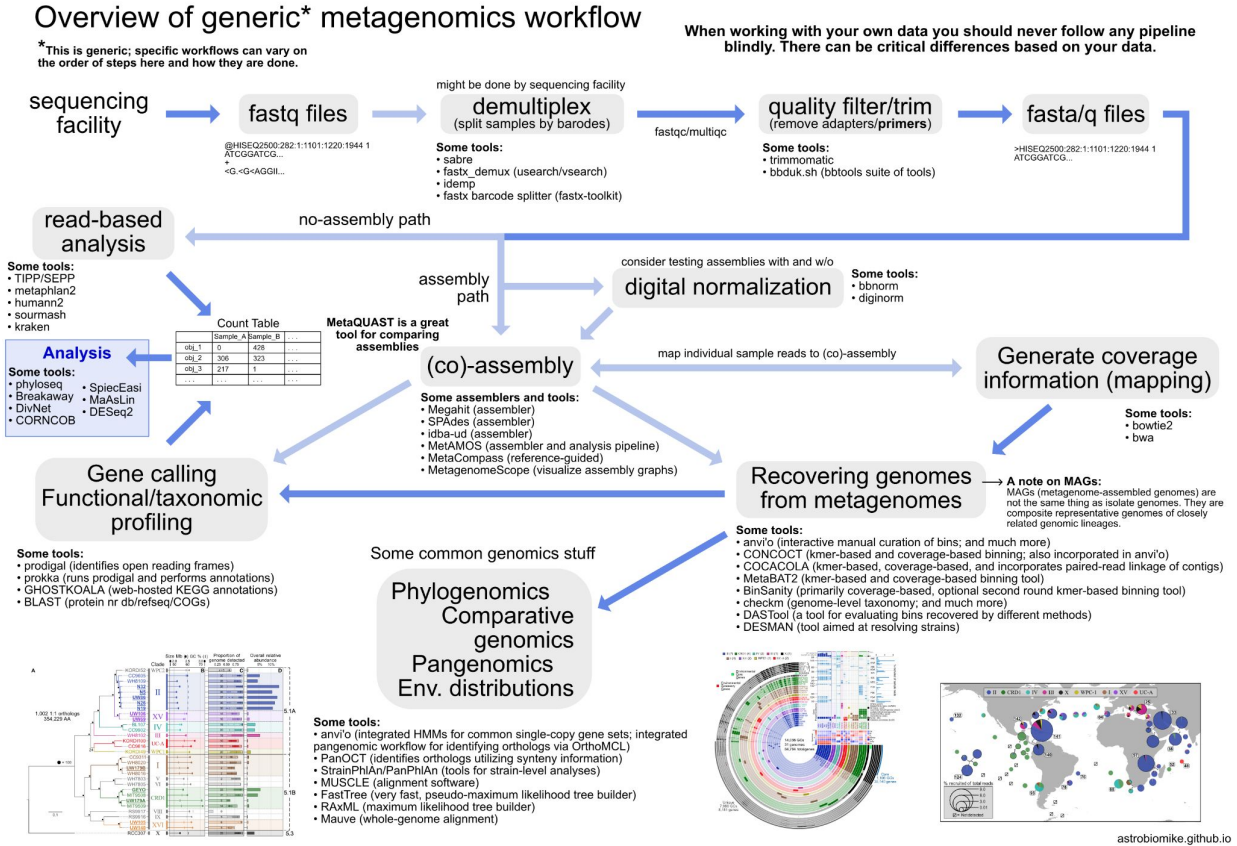
In practice there'll be many input molecules here...

"Reads" (Strings from $\Sigma = \{A,C,G,T\}$)

...and, ideally, one output sequence per input molecule

Lee, M. D. (2019). Happy Belly Bioinformatics: An open-source resource dedicated to helping biologists utilize bioinformatics. *Journal of Open Source Education*, *4*(41), 53.

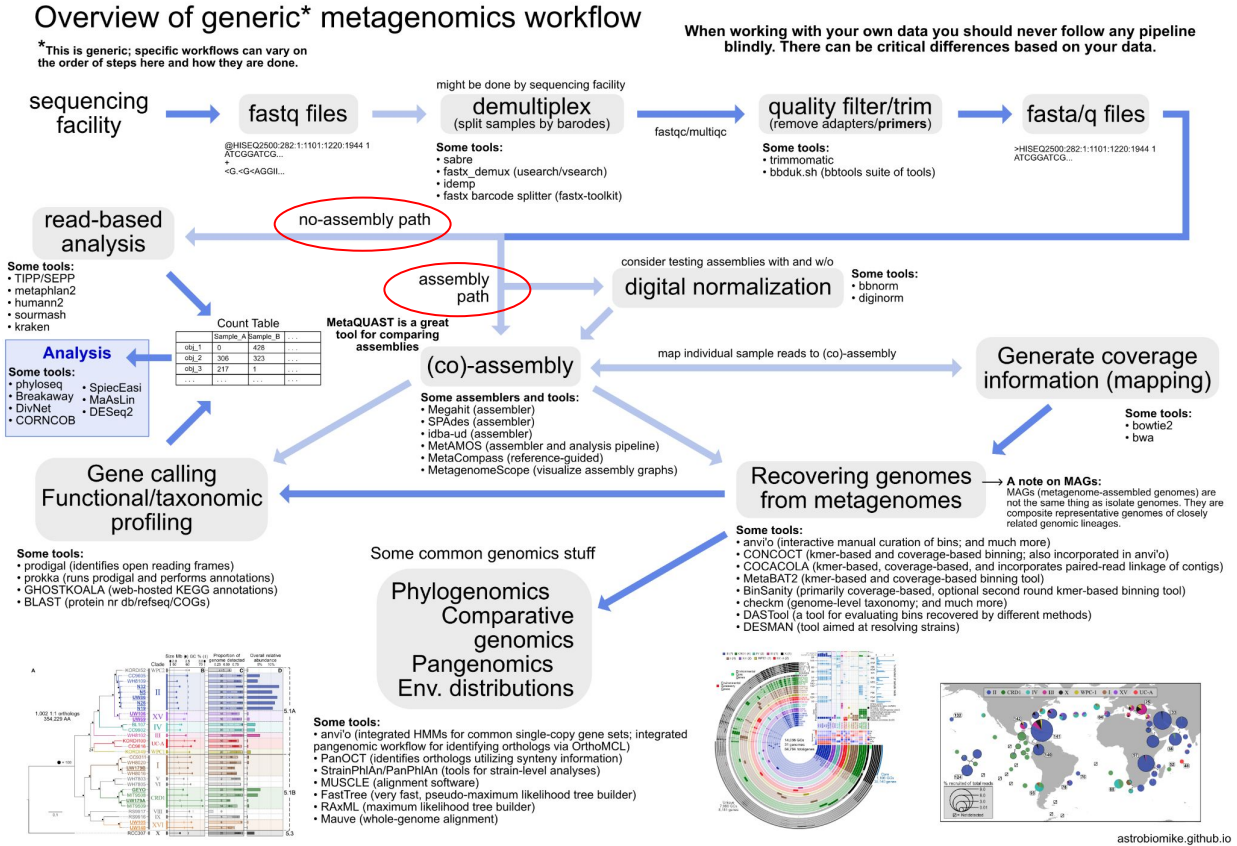
C. I. Methods: Metagenome sequence assembly



astrobiomike.github.io

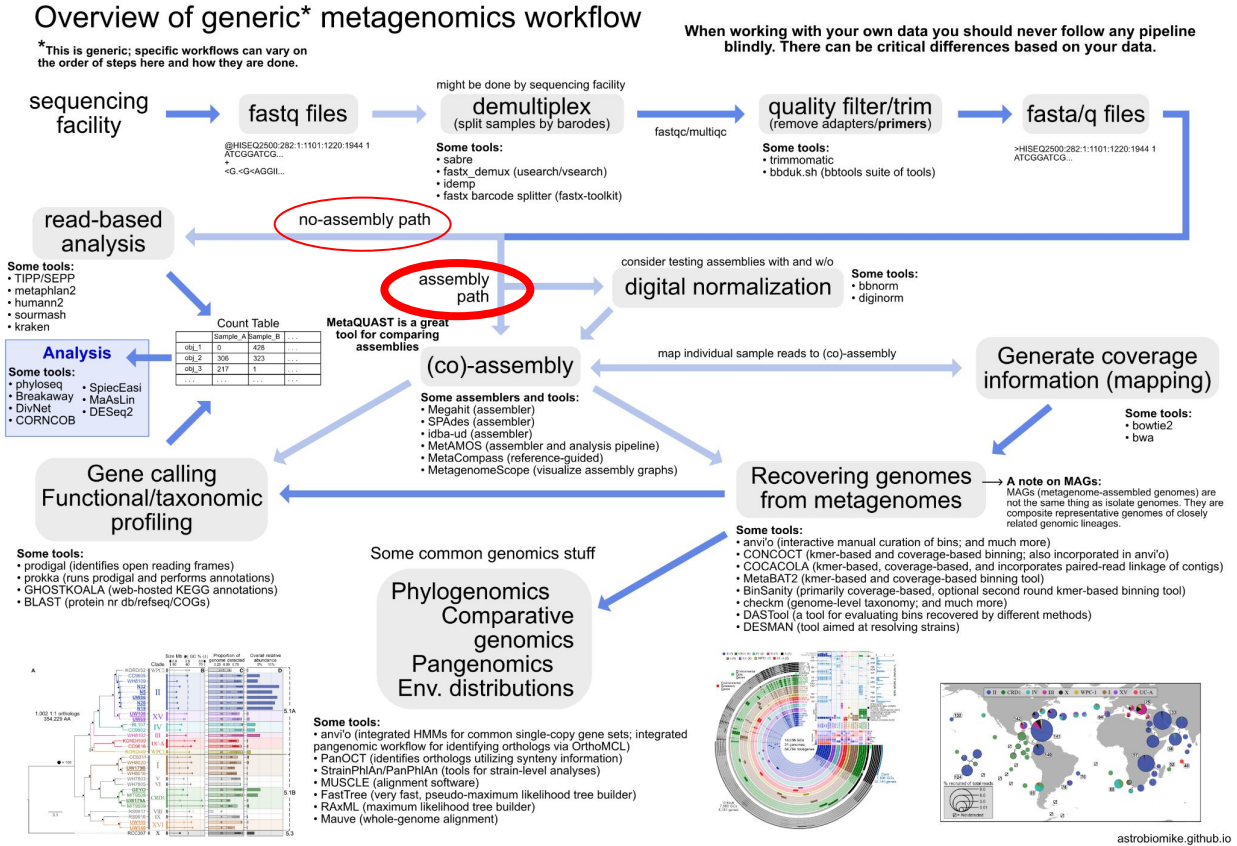
Lee, M. D. (2019). Happy Belly Bioinformatics: An open-source resource dedicated to helping biologists utilize bioinformatics. *Journal of Open Source Education*, *4*(41), 53.

C. I. **Methods:** Metagenome sequence assembly



Lee, M. D. (2019). Happy Belly Bioinformatics: An open-source resource dedicated to helping biologists utilize bioinformatics. *Journal of Open Source Education*, *4*(41), 53.

C. I. **Methods:** Metagenome sequence assembly



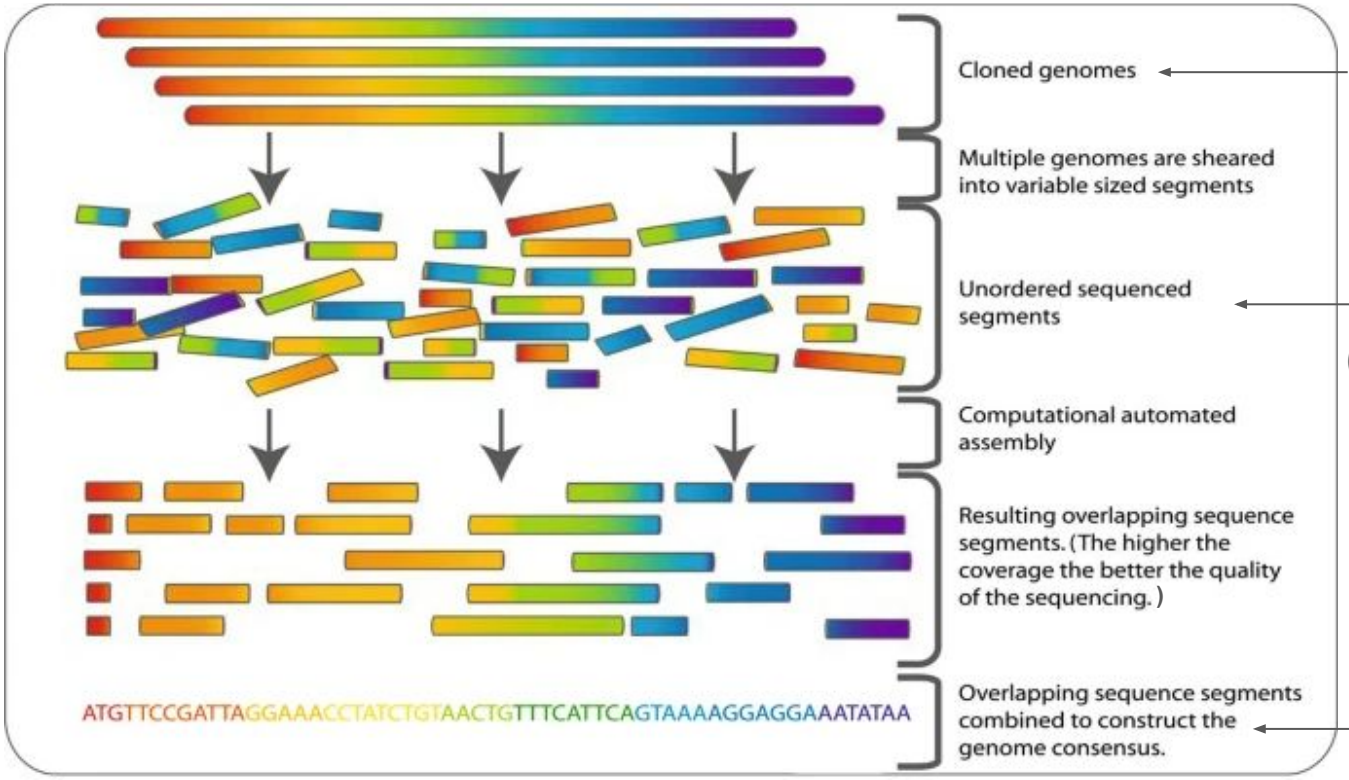
astrobiomike.github.io

This talk

1. Introduction: Studying microbiomes
   a. Why bother?
   b. Why hasn't this research been more useful?
   c. Defining a goal for this talk
2. Culture-independent (a.k.a. sequencing-based) methods
   a. Marker gene sequencing
   b. Metagenome sequencing
3. **Metagenome assembly**
   a. Input (*reads*)
   b. Outputs (*contigs*, an *assembly graph*, …)
   c. Methods (*de novo* vs. reference-based, overlap graph vs. de Bruijn graph, …)
4. Future work: Solving the *strain separation problem*

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly (in the context of a metagenome)



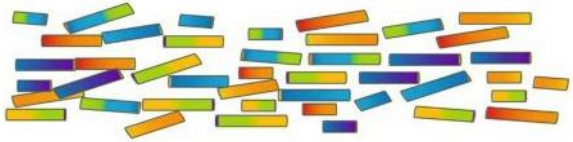
In practice there'll be many input molecules here...

"Reads" (Strings from $\Sigma = \{A,C,G,T\}$)

...and, ideally, one output sequence per input molecule

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: inputs

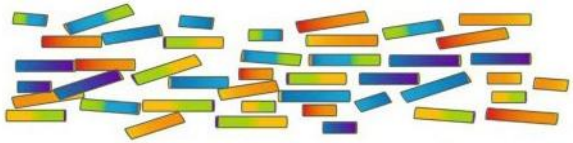


Reads: strings from the alphabet Σ = {**A**, **C**, **G**, **T**}

Occasionally this alphabet is extended if the base at a position is ambiguous:
e.g. **W** = (**A** or **T**), **S** = (**C** or **G**), **N** = (**A** or **C** or **G** or **T**), …

For more detail on this notation, see: https://en.wikipedia.org/wiki/Nucleic_acid_notation.

Assembly: inputs



Reads: strings from the alphabet Σ = {**A**, **C**, **G**, **T**}

Occasionally this alphabet is extended if the base at a position is ambiguous:
e.g. **W** = (**A** or **T**), **S** = (**C** or **G**), **N** = (**A** or **C** or **G** or **T**), …

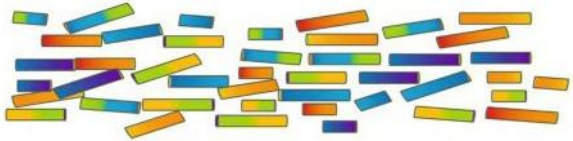
Things that can vary based on the **sequencing technology** being used:

- Read length
- Read error rate
- Number of reads
- Read structure (e.g. single vs. *paired-end* reads)

For more detail on this notation, see: https://en.wikipedia.org/wiki/Nucleic_acid_notation.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: inputs



Reads: strings from the alphabet Σ = {**A**, **C**, **G**, **T**}

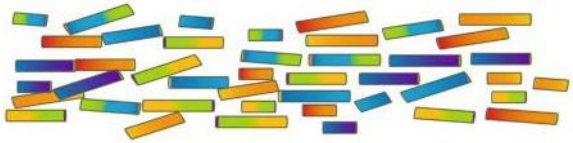
Occasionally this alphabet is extended if the base at a position is ambiguous:
e.g. **W** = (**A** or **T**), **S** = (**C** or **G**), **N** = (**A** or **C** or **G** or **T**), …

Things that can vary based on the **sequencing technology** being used:

- **Read length**
- **Read error rate**
- Number of reads
- Read structure (e.g. single vs. *paired-end* reads)

For more detail on this notation, see: https://en.wikipedia.org/wiki/Nucleic_acid_notation.

Assembly: inputs



Reads: strings from the alphabet Σ = {**A**, **C**, **G**, **T**}

Occasionally this alphabet is extended if the base at a position is ambiguous:
e.g. **W** = (**A** or **T**), **S** = (**C** or **G**), **N** = (**A** or **C** or **G** or **T**), …

Things that can vary based on the **sequencing technology** being used:

- **Read length**
- **Read error rate**
- Number of reads
- Read structure (e.g. single vs. *paired-end* reads)

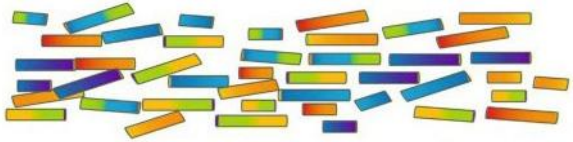
Three (modern) sequencing technologies we'll focus on:

(1) short-read, (2) long, error-prone read, (3) HiFi

For more detail on this notation, see: https://en.wikipedia.org/wiki/Nucleic_acid_notation.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

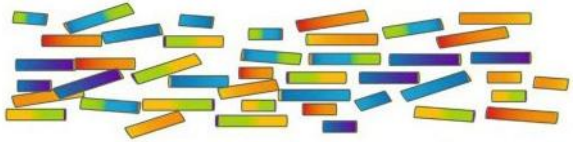
Assembly: sequencing technologies



- Short-read technologies
  - e.g. Illumina
  - Read lengths: up to 150 nt
  - Error rates: less than 1%
- Long, error-prone read technologies
  - e.g. Oxford Nanopore, Pacific Biosciences
  - Read length: over 10,000 nt
  - Error rates: 10–25%
- Long, accurate read technologies (“HiFi”)
  - Pacific Biosciences circular consensus sequencing
  - Read length: over 10,000 nt, but generally shorter than long, error-prone reads
  - Error rates:
    - less than 1% for “point” mutations
    - somewhat higher (~5%) for insertions/deletions

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

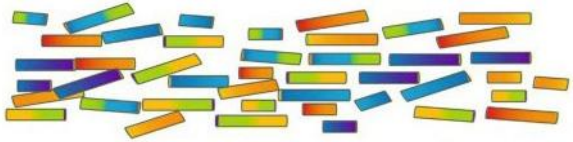
Assembly: sequencing technologies



- Short-read technologies
  - e.g. Illumina
  - Read lengths: up to 150 nt
  - Error rates: **less than 1%**
- Long, error-prone read technologies
  - e.g. Oxford Nanopore, Pacific Biosciences
  - Read length: **over 10,000 nt**
  - Error rates: 10–25%
- Long, accurate read technologies (“HiFi”)
  - Pacific Biosciences circular consensus sequencing
  - Read length: over 10,000 nt, but generally shorter than long, error-prone reads
  - Error rates:
    - less than 1% for “point” mutations
    - somewhat higher (~5%) for insertions/deletions

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: sequencing technologies



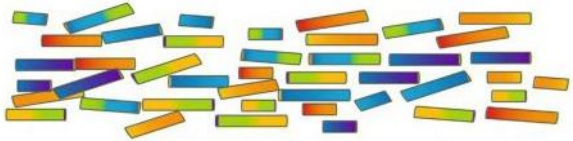
- Short-read technologies
  - e.g. Illumina
  - Read lengths: up to 150 nt
  - Error rates: **less than 1%**
- Long, error-prone read technologies
  - e.g. Oxford Nanopore, Pacific Biosciences
  - Read length: **over 10,000 nt**
  - Error rates: 10–25%
- Long, accurate read technologies (“HiFi”)
  - Pacific Biosciences circular consensus sequencing
  - Read length: over 10,000 nt, but generally shorter than long, error-prone reads
  - Error rates:
    - less than 1% for “point” mutations
    - somewhat higher (~5%) for insertions/deletions

This simple categorization ignores *a lot* of finicky details!

But this should be enough to understand how these technologies can complicate or simplify assembly.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: impacts of technologies

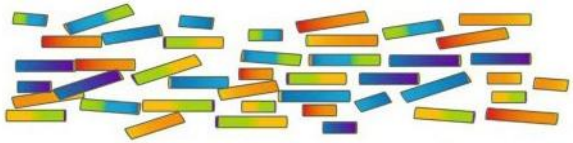


- Read lengths
  - Longer reads can span **repeats** (“identical, or nearly identical, stretches of DNA”).
  - Repeats that are longer than reads are especially challenging.
  - Accounting for repeats is the most important challenge in (metagenome) assembly!
  - So longer reads are generally better than short ones.

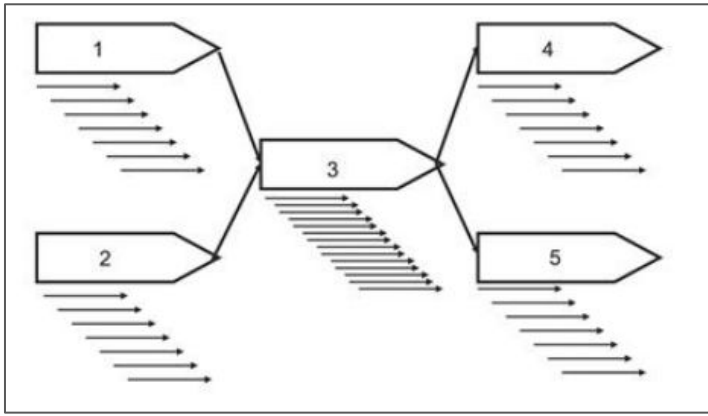
Medvedev, P., Pham, S., Chaisson, M., Tesler, G., & Pevzner, P. (2011). Paired de bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers. *Journal of Computational Biology*, *18*(11), 1625-1634.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: impacts of technologies



- Read lengths
  - Longer reads can span **repeats** (“identical, or nearly identical, stretches of DNA”).
  - Repeats that are longer than reads are especially challenging.
  - Accounting for repeats is the most important challenge in (metagenome) assembly!
  - So longer reads are generally better than short ones.

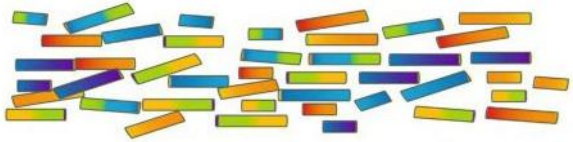


Koren, S., Treangen, T. J., & Pop, M. (2011). Bambus 2: scaffolding metagenomes. *Bioinformatics, 27*(21), 2964-2971.

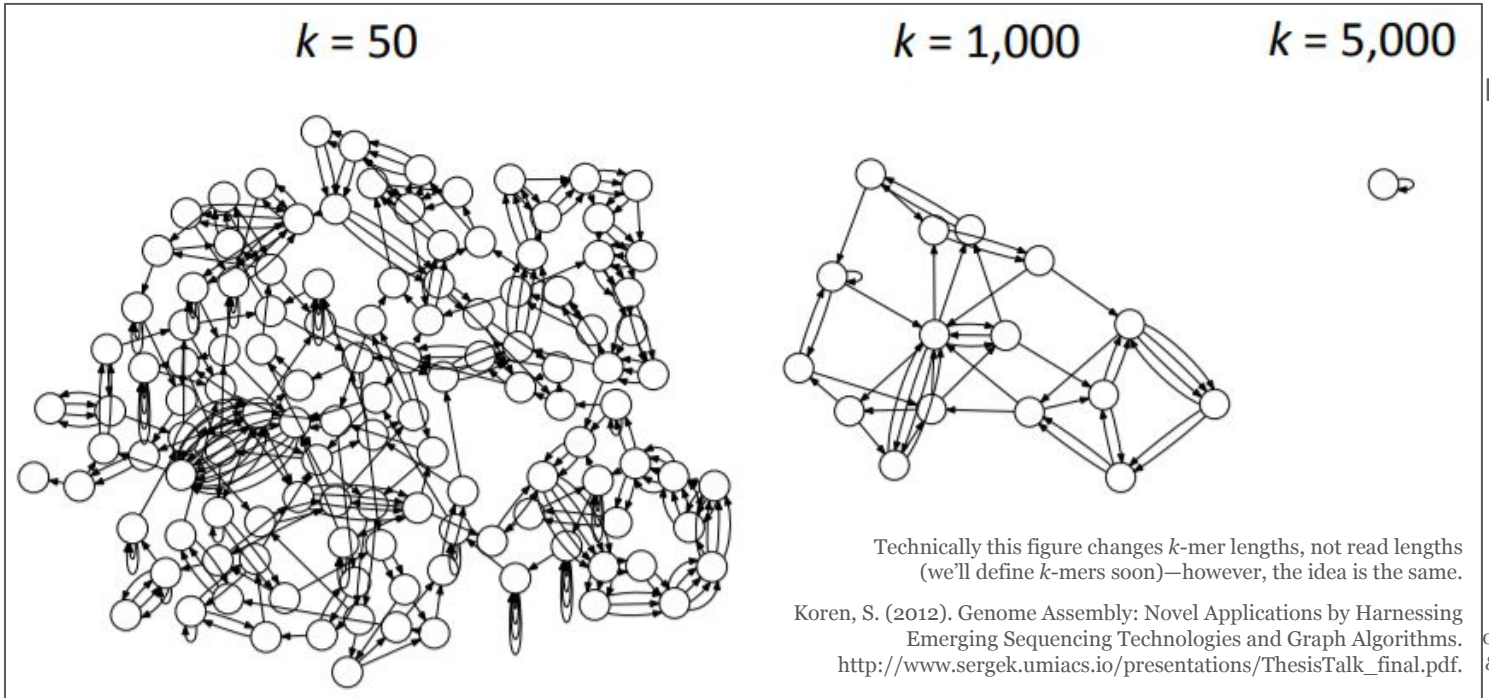
Medvedev, P., Pham, S., Chaisson, M., Tesler, G., & Pevzner, P. (2011). Paired de bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers. *Journal of Computational Biology*, *18*(11), 1625-1634.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: impacts of technologies



- Read lengths
  - Longer reads can span **repeats** ("identical, or nearly identical, stretches of DNA").
  -
  -
  -



mbly!

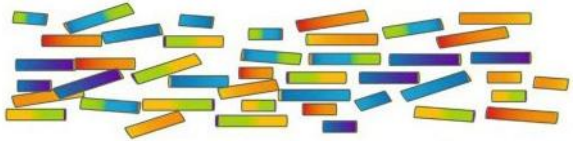
Technically this figure changes *k*-mer lengths, not read lengths (we'll define *k*-mers soon)—however, the idea is the same.

Koren, S. (2012). Genome Assembly: Novel Applications by Harnessing Emerging Sequencing Technologies and Graph Algorithms. http://www.sergek.umiacs.io/presentations/ThesisTalk_final.pdf.

orating mate pair *8*(11), 1625-1634.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: impacts of technologies

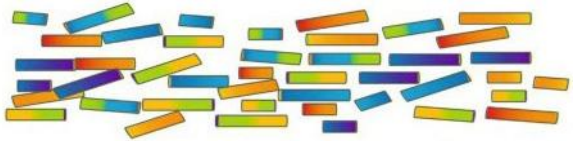


- Read lengths
  - Longer reads can span **repeats** (“identical, or nearly identical, stretches of DNA”).
  - Repeats that are longer than reads are especially challenging.
  - Accounting for repeats is the most important challenge in (metagenome) assembly!
  - So longer reads are generally better than short ones.

Medvedev, P., Pham, S., Chaisson, M., Tesler, G., & Pevzner, P. (2011). Paired de bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers. *Journal of Computational Biology*, *18*(11), 1625-1634.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: impacts of technologies



- Read lengths
  - Longer reads can span **repeats** (“identical, or nearly identical, stretches of DNA”).
  - Repeats that are longer than reads are especially challenging.
  - Accounting for repeats is the most important challenge in (metagenome) assembly!
  - So longer reads are generally better than short ones.

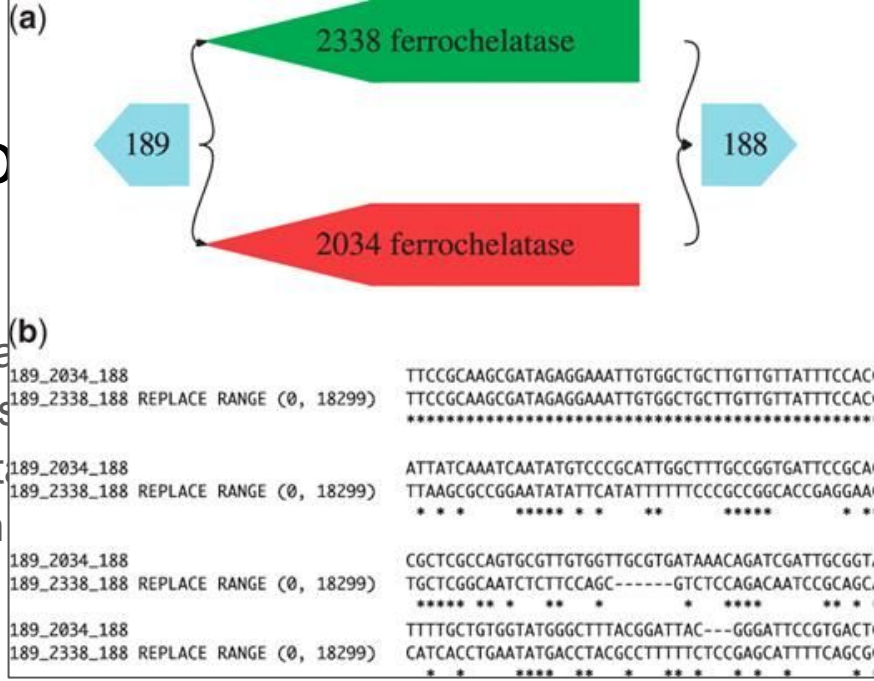
- Error rates
  - Lower error rates simplify **variant calling**, in which we attempt to distinguish real variations in the data from sequencing errors.
  - This is especially common for metagenome assembly: is a variation at some position the result of error, or is it indicative of a rare strain with this variation in its genome?

Medvedev, P., Pham, S., Chaisson, M., Tesler, G., & Pevzner, P. (2011). Paired de bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers. *Journal of Computational Biology*, *18*(11), 1625-1634.

Koren, S., Treangen, T. J., & Pop, M. (2011). Bambus 2: scaffolding metagenomes. *Bioinformatics, 27*(21), 2964-2971.

Assembly: impacts of techno

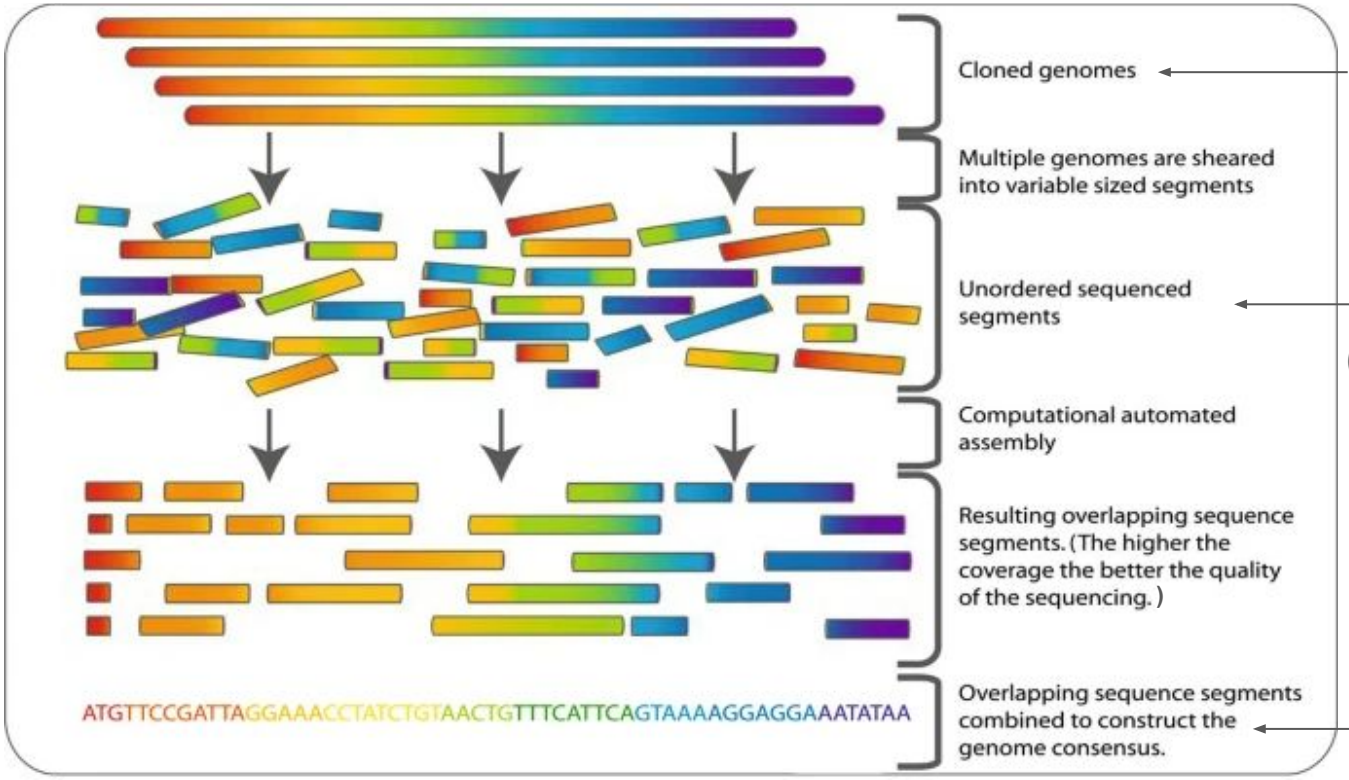
- Read lengths
  - Longer reads can span **repeats** (“identica
  - Repeats that are longer than reads are es
  - Accounting for repeats is the most import
  - So longer reads are generally better than
- Error rates
  - Lower error rates simplify **variant calling**, in which we attempt to distinguish real variations in the data from sequencing errors.
  - This is especially common for metagenome assembly: is a variation at some position the result of error, or is it indicative of a rare strain with this variation in its genome?



Medvedev, P., Pham, S., Chaisson, M., Tesler, G., & Pevzner, P. (2011). Paired de bruijn graphs: a novel approach for incorporating mate pair information into genome assemblers. *Journal of Computational Biology, 18*(11), 1625-1634.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: outputs



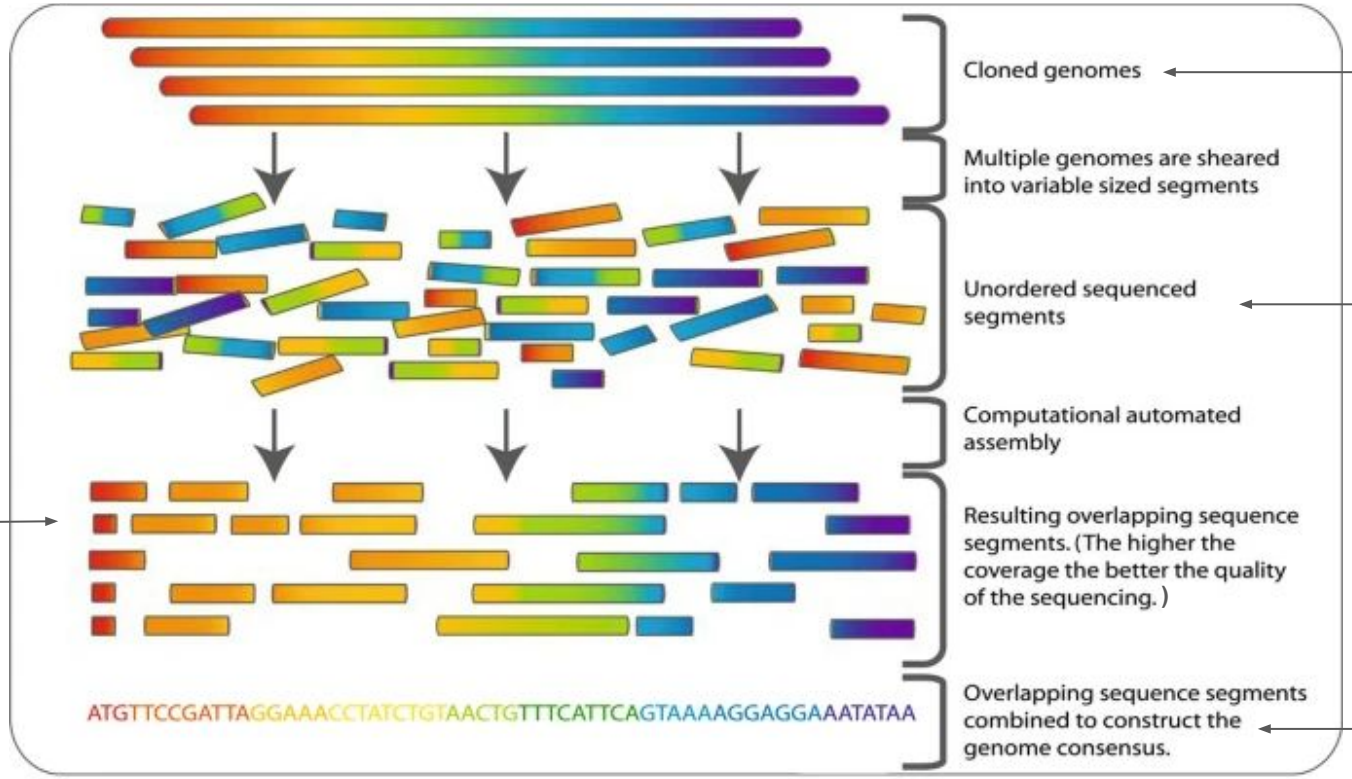
In practice there'll be many input molecules here...

"Reads" (Strings from $\Sigma = \{A,C,G,T\}$)

...and, ideally, one output sequence per input molecule

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: outputs



In practice there'll be many input molecules here...

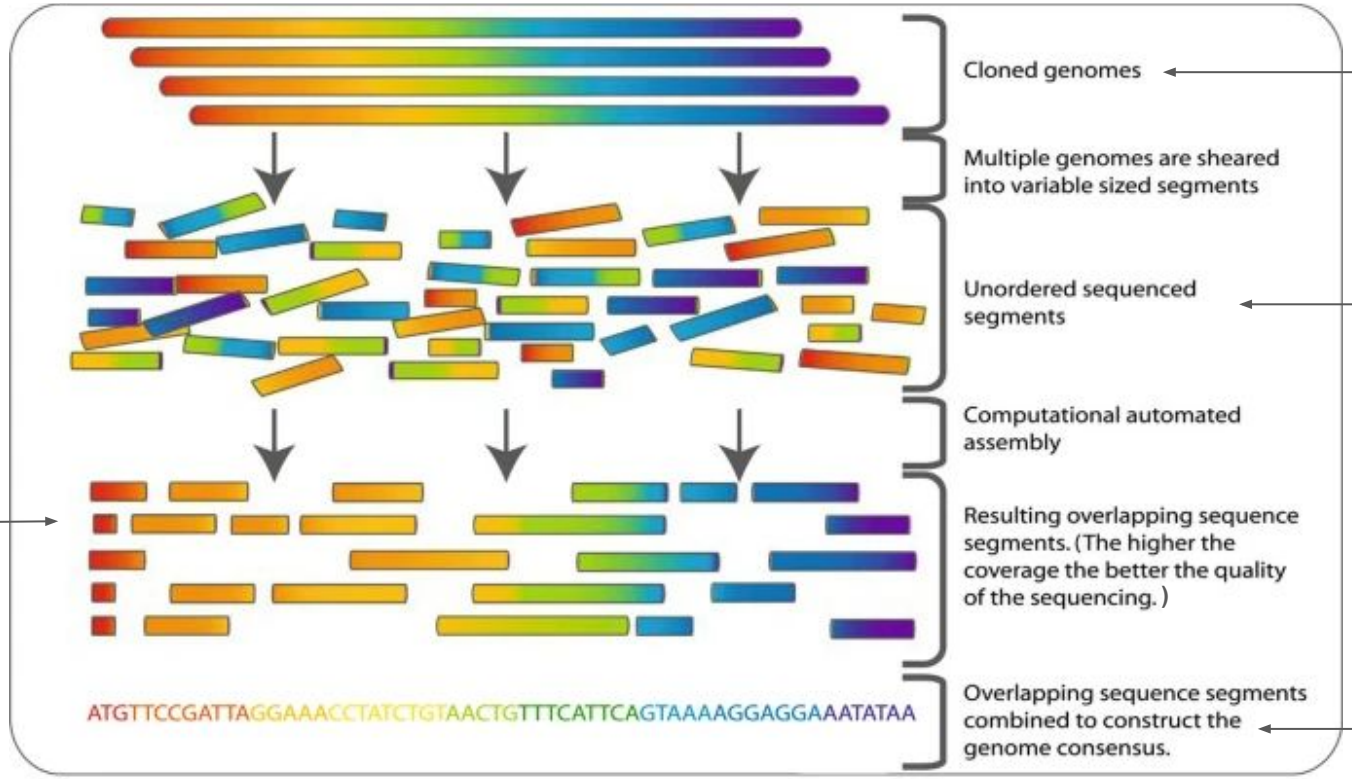
"Reads" (Strings from $\Sigma = \{A,C,G,T\}$)

"Contigs"

...and, ideally, one output sequence per input molecule

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: outputs



In practice there'll be many input molecules here...

"Reads" (Strings from $\Sigma = \{A,C,G,T\}$)

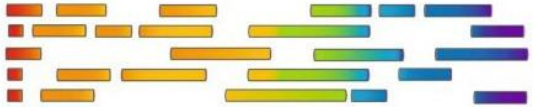
"Contigs"

(Hopefully longer) strings from $\Sigma = \{A,C,G,T\}$

...and, ideally, one output sequence per input molecule

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: outputs (contigs)

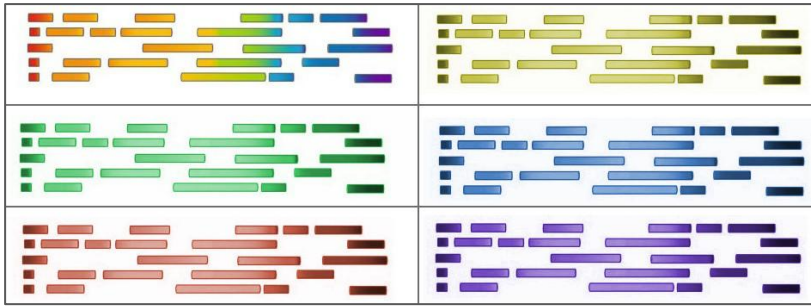


Ideally: one contig per input molecule of DNA
(e.g. each chromosome, plasmid, …)

In practice: usually more contigs than that

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: outputs (contigs)



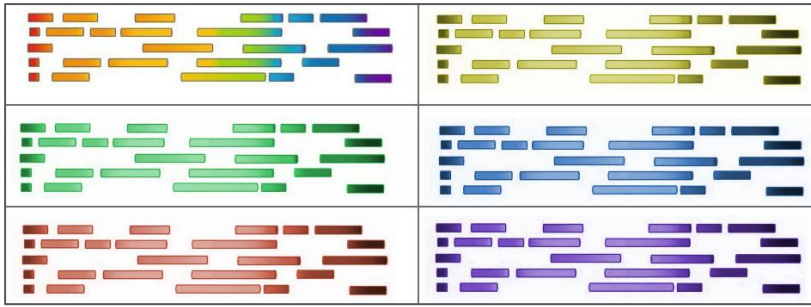
Ideally: one contig per input molecule of DNA
(e.g. each chromosome, plasmid, …)
In practice: usually more contigs than that

Some projects attempt to group contigs together into **bins** that likely originate from the same “genome”.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., ... & Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*, *35*(8), 725-731.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: outputs (contigs)



Ideally: one contig per input molecule of DNA
(e.g. each chromosome, plasmid, …)
In practice: usually more contigs than that

Some projects attempt to group contigs together into **bins** that likely originate from the same “genome”.

Bins of contigs, or especially high-quality individual contigs, are referred to as **metagenome-assembled genomes (MAGs)**.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., ... & Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*, *35*(8), 725-731.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: outputs (contigs)

"We present a metagenomic HiFi assembly of a complex microbial community from sheep fecal material that resulted in **428 high-quality MAGs** from a single sample, the highest resolution achieved with metagenomic deconvolution to date."

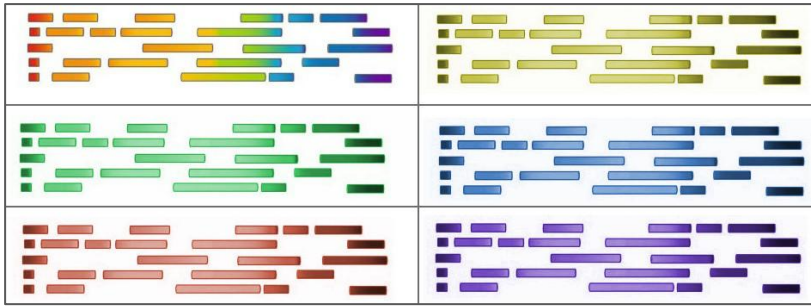
Bickhart, D. M., Kolmogorov, M., Tseng, E., Portik, D., Korobeynikov, A., Tolstoganov, I., ... & Smith, T. P. (2021). Generation of lineage-resolved complete metagenome-assembled genomes by precision phasing. *bioRxiv*.

Bins of contigs, or especially high-quality individual contigs, are referred to as **metagenome-assembled genomes (MAGs)**.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., ... & Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*, *35*(8), 725-731.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: outputs (contigs)



Ideally: one contig per input molecule of DNA
(e.g. each chromosome, plasmid, …)
In practice: usually more contigs than that

Some projects attempt to group contigs together into **bins** that likely originate from the same “genome”.

Bins of contigs, or especially high-quality individual contigs, are referred to as **metagenome-assembled genomes (MAGs)**.

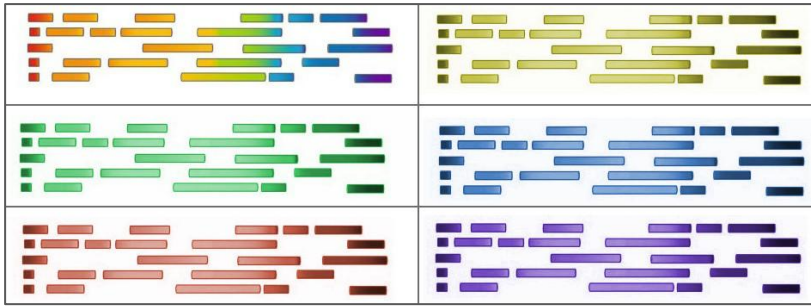
Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T. B. K., ... & Woyke, T. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*, *35*(8), 725-731.

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. *Biological Procedures Online*, *11*(1), 52-78.

Assembly: outputs (contigs)

Ideally: one contig per input molecule of DNA
(e.g. each chromosome, plasmid, …)

In practice: usually more contigs than that



Assembly: outputs (assembly graph)

Most assembly algorithms model the problem as some sort of **graph traversal**.

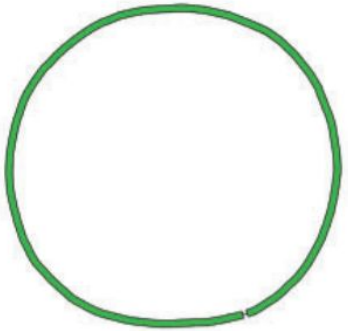
Assemblers usually output **assembly graphs**, which (generally) show overlaps between contigs.

Assembly: outputs (assembly graph)

Most assembly algorithms model the problem as some sort of **graph traversal**.

Assemblers usually output **assembly graphs**, which (generally) show overlaps between contigs.

Ideally: one connected component per input molecule of DNA



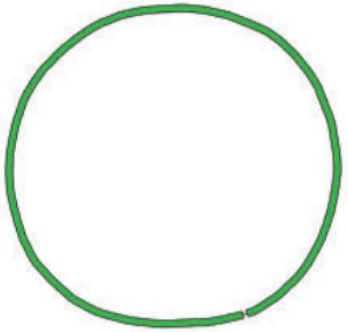
Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, *31*(20), 3350-3352.

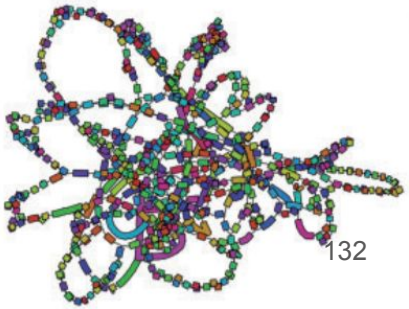
Assembly: outputs (assembly graph)

Most assembly algorithms model the problem as some sort of **graph traversal**.

Assemblers usually output **assembly graphs**, which (generally) show overlaps between contigs.

Ideally: one connected component per input molecule of DNA
In practice: the graph is usually tangled, fragmented, ...





Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, *31*(20), 3350-3352.

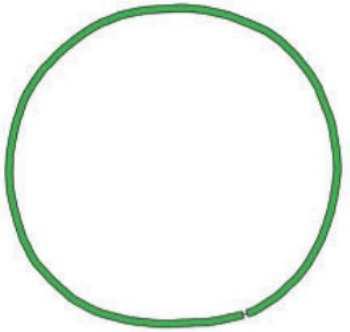
Assembly: outputs (assembly graph)

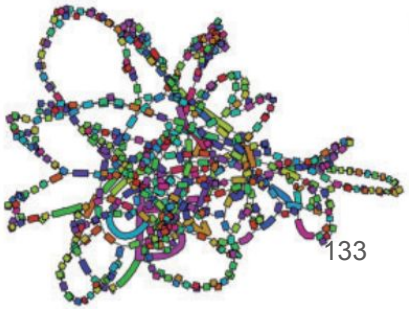
Most assembly algorithms model the problem as some sort of **graph traversal**.

Assemblers usually output **assembly graphs**, which (generally) show overlaps between contigs.

Ideally: one connected component per input molecule of DNA
In practice: the graph is usually tangled, fragmented, ...

These can be useful when “finishing” assemblies, or looking at subtle variations.





Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*, *31*(20), 3350-3352.

Assembly: outputs (assembly graph)

Most assembly algorithms model the problem as some sort of **graph traversal**.

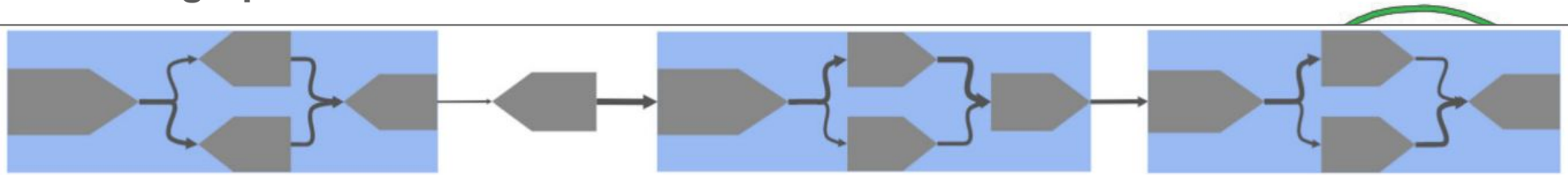
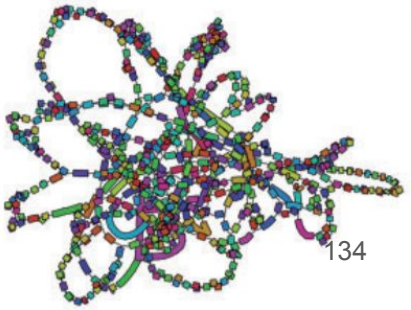


Fig. 2 Variants detected in one of the components of *Acinetobacter baumanii* scaffold graph. In this component, we find all the non-terminal nodes in a bubble are more than 97% identical to each other and originate from two different strains of *Acinetobacter baumannii* genome

Ghurye, J., Treangen, T., Fedarko, M., Hervey, W. J., & Pop, M. (2019). MetaCarvel: linking assembly graph motifs to biological variants. *Genome Biology*, *20*(1), 1-14.

These can be useful when “finishing” assemblies, or looking at subtle variations.



This talk

1. Introduction: Studying microbiomes
   a. Why bother?
   b. Why hasn't this research been more useful?
   c. Defining a goal for this talk
2. Culture-independent (a.k.a. sequencing-based) methods
   a. Marker gene sequencing
   b. Metagenome sequencing
3. **Metagenome assembly**
   a. Input (*reads*)
   b. Outputs (*contigs*, an *assembly graph*, …)
   c. **Methods** (*de novo* vs. reference-based, overlap graph vs. de Bruijn graph, …)
4. Future work: Solving the *strain separation problem*

Assembly: methods

"A good genome assembler is like a good sausage:
you would rather not know what is inside."

Apocryphal quote attributed to Sante Gnerre: http://rayan.chikhi.name/pdf/2019-july-19-cgsi.pdf

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics, 95*(6), 315-327.

Assembly: methods (*de novo* vs. reference-based)

***de novo* assembly:** Use only the read data available

"[...] reconstruction in its pure form, without consultation to previously resolved sequence including from genomes, transcripts, and proteins."

Reference-based assembly: Also use available reference sequence(s)

"For some applications, sufficient information can be extracted from the mapping of reads to a reference sequence, such as a finished genome from a related individual."

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315-327.

Assembly: methods (*de novo* vs. reference-based)

***de novo* assembly:** Use only the read data available

Far more commonly used when working with metagenome sequencing data.

Reference-based assembly: Also use available reference sequence(s)

Some reference-based assemblers have been developed for metagenome sequencing data, but they have not (yet) seen widespread use in the field.

Cepeda, V., Liu, B., Almeida, M., Hill, C. M., Koren, S., Treangen, T. J., & Pop, M. (2017). MetaCompass: reference-guided assembly of metagenomes. *bioRxiv*, 212506.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics, 95*(6), 315-327.

Assembly: methods (***de novo*** vs. reference-based)

***de novo* assembly:** Use only the read data available

Far more commonly used when working with metagenome sequencing data.

Reference-based assembly: Also use available reference sequence(s)

Some reference-based assemblers have been developed for metagenome sequencing data, but they have not (yet) seen widespread use in the field.

Cepeda, V., Liu, B., Almeida, M., Hill, C. M., Koren, S., Treangen, T. J., & Pop, M. (2017). MetaCompass: reference-guided assembly of metagenomes. *bioRxiv*, 212506.

Idury, R. M., & Waterman, M. S. (1995). A new algorithm for DNA sequence assembly. *Journal of Computational Biology, 2*(2), 291-306.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics, 95*(6), 315-327.

Assembly: methods (overlap vs. de Bruijn graphs)

OLC DBG

Overlap graph (directed graph)

Nodes: input reads

Edges: an edge is created from $n_1 \rightarrow n_2$ if read n_1 overlaps with read n_2

de Bruijn graph (also a directed graph; takes a parameter *k*)

Nodes: unique (*k* - *1*)-mers (strings of length *k* - *1*) in the reads

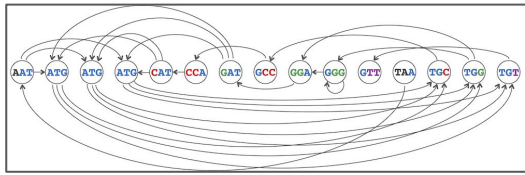
Edges: an edge is created from $n_1 \rightarrow n_2$ if there is a *k*-mer whose prefix is n_1 and whose suffix is n_2

These descriptions are very simplistic: in practice, the graph structures used have many optimizations, error correction methods, etc. applied. See the Miller et al. 2010 paper referenced above for more details.

Idury, R. M., & Waterman, M. S. (1995). A new algorithm for DNA sequence assembly. *Journal of Computational Biology, 2*(2), 291-306.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics, 95*(6), 315-327.

Assembly: methods (overlap vs. de Bruijn graphs) OLC



Overlap graph (directed graph)

Nodes: input reads

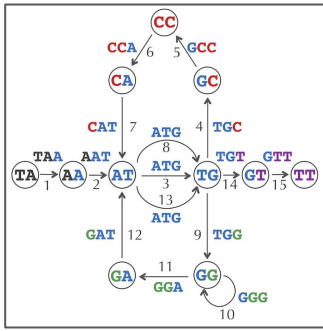
Edges: an edge is created from $n_1 \rightarrow n_2$ if read n_1 overlaps with read n_2

DBG

de Bruijn graph (also a directed graph; takes a parameter *k*)

Nodes: unique (*k* - *1*)-mers (strings of length *k* - *1*) in the reads

Edges: an edge is created from $n_1 \rightarrow n_2$ if there is a *k*-mer whose prefix is n_1 and whose suffix is n_2



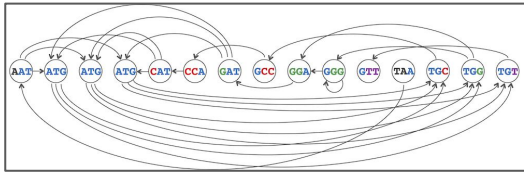
Compeau, P., & Pevzner, P. (2021). *Bioinformatics Algorithms: an active learning approach*. https://www.bioinformaticsalgorithms.org/bioinformatics-chapter-3/.

These descriptions are very simplistic: in practice, the graph structures used have many optimizations, error correction methods, etc. applied. See the Miller et al. 2010 paper referenced above for more details.

Idury, R. M., & Waterman, M. S. (1995). A new algorithm for DNA sequence assembly. *Journal of Computational Biology, 2*(2), 291-306.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics, 95*(6), 315-327.

Assembly: methods (overlap vs. de Bruijn graphs) OLC



Overlap graph (directed graph)

Nodes: input reads

Edges: an edge is created from $n_1 \rightarrow n_2$ if read n_1 overlaps with read n_2

Goal: Find **Hamiltonian** Paths (or cycles) in this graph **NP-Complete!**

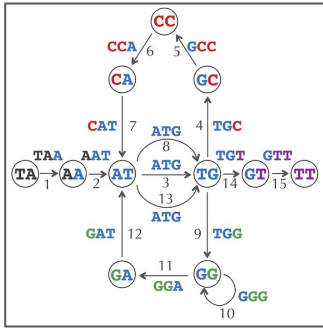
DBG

de Bruijn graph (also a directed graph; takes a parameter *k*)

Nodes: unique (*k* - *1*)-mers (strings of length *k* - *1*) in the reads

Edges: an edge is created from $n_1 \rightarrow n_2$ if there is a *k*-mer whose prefix is n_1 and whose suffix is n_2

Goal: Find **Eulerian** Paths (or cycles) in this graph **Polynomial time!**



Compeau, P., & Pevzner, P. (2021). *Bioinformatics Algorithms: an active learning approach*. https://www.bioinformaticsalgorithms.org/bioinformatics-chapter-3/.

These descriptions are very simplistic: in practice, the graph structures used have many optimizations, error correction methods, etc. applied. See the Miller et al. 2010 paper referenced above for more details.

Dida, F., & Yi, G. (2021). Empirical evaluation of methods for de novo genome assembly. *PeerJ Computer Science*, *7*, e636.

Wang, A., Wang, Z., Li, Z., & Li, L. M. (2018). BAUM: improving genome assembly by adaptive unique mapping and local overlap-layout-consensus approach. *Bioinformatics*, *34*(12), 2019-2028.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315-327.

Assembly: methods (overlap vs. de Bruijn graphs)

OLC

- Newbler
- Celera
- ARACHNE
- Edena
- SHORTY
- HINGE
- BAUM
- Canu
- hifiasm
- ...

DBG

- EULER
- AllPaths
- SOAP
- Velvet
- ABySS
- E + V-SC
- SPAdes
- ABruijn
- Flye
- ...

Neither of these lists are comprehensive! Many assemblers (of both the OLC or DBG varieties) are still actively being developed and in use today. Also, most assemblers implement their own "twist" on how they use these graph structures, so cleanly categorizing assemblers in this way ignores many details.

Dida, F., & Yi, G. (2021). Empirical evaluation of methods for de novo genome assembly. *PeerJ Computer Science*, *7*, e636.

Wang, A., Wang, Z., Li, Z., & Li, L. M. (2018). BAUM: improving genome assembly by adaptive unique mapping and local overlap-layout-consensus approach. *Bioinformatics*, *34*(12), 2019-2028.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315-327.

Assembly: methods (overlap vs. de Bruijn graphs)

OLC	DBG
What do Eulerian and Hamiltonian cycles have to do with genome assembly?	
Paul Medvedev[1,2,3]*, Mihai Pop[4,5]	
• BAUM • Canu • hifiasm • ...	• SPAdes • ABruijn • Flye • ...

Neither of these lists are comprehensive! Many assemblers (of both the OLC or DBG varieties) are still actively being developed and in use today. Also, most assemblers implement their own "twist" on how they use these graph structures, so cleanly categorizing assemblers in this way ignores many details.

Dida, F., & Yi, G. (2021). Empirical evaluation of methods for de novo genome assembly. *PeerJ Computer Science*, *7*, e636.

Wang, A., Wang, Z., Li, Z., & Li, L. M. (2018). BAUM: improving genome assembly by adaptive unique mapping and local overlap-layout-consensus approach. *Bioinformatics*, *34*(12), 2019-2028.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315-327.

Assembly: methods (overlap vs. de Bruijn graphs)

OLC

What do Euler

to do with gen

an cycles have

Paul Medvedev [1,2,3]*, Miha

- BAUM
- Canu
- hifiasm
- ...

DBG

Neither of these lists are comprehensive! Many assemblers (of both the OLC or DBG varieties) are still actively being developed and in use today. Also, most assemblers implement their own "twist" on how they use these graph structures, so cleanly categorizing assemblers in this way ignores many details.

Dida, F., & Yi, G. (2021). Empirical evaluation of methods for de novo genome assembly. *PeerJ Computer Science*, *7*, e636.

Wang, A., Wang, Z., Li, Z., & Li, L. M. (2018). BAUM: improving genome assembly by adaptive unique mapping and local overlap-layout-consensus approach. *Bioinformatics*, *34*(12), 2019-2028.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315-327.

Assembly: methods (overlap vs. de Bruijn graphs)

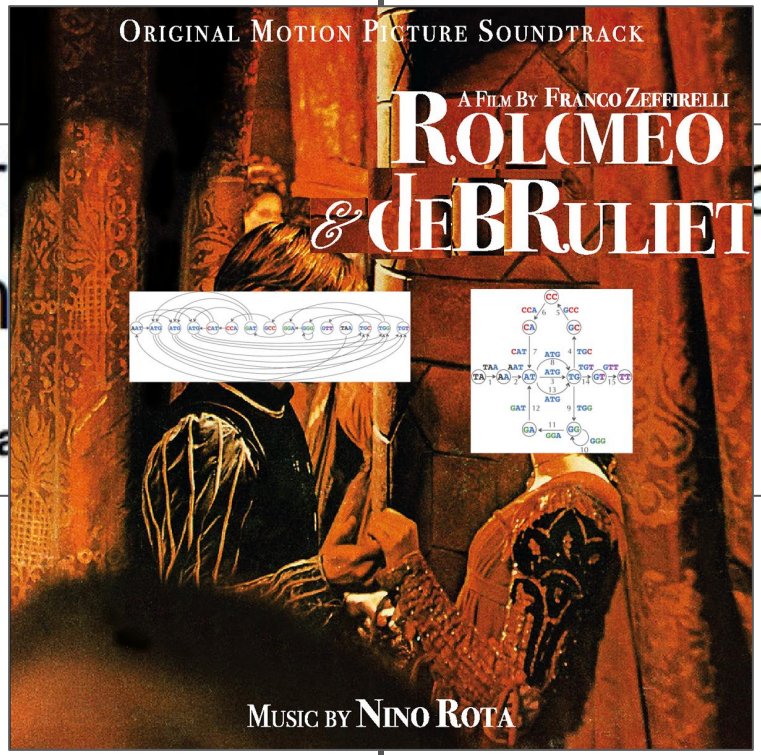
OLC

What do Euler

to do with gen

Paul Medvedev[1,2,3]*, Miha

- BAUM
- Canu
- hifiasm
- ...



DBG

an cycles have

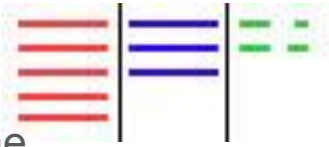
Neither of these lists are comprehensive! Many assemblers (of both the OLC or DBG varieties) are still actively being developed and in use today. Also, most assemblers implement their own "twist" on how they use these graph structures, so cleanly categorizing assemblers in this way ignores many details.

Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., ... & Pevzner, P. A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, *17*(11), 1103-1110.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824-834.

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Research*, *40*(20), e155-e155.

Assembly: methods (single-genome vs. metagenome)



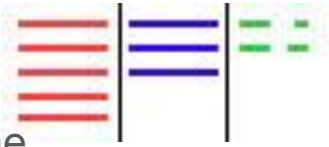
- **Problem:** Uneven coverage
  - Different microbes are present at different abundances in a microbiome.
  - The *coverage*, or number of reads “supporting” a position in a genome, varies a lot!
  - Worst case scenario: some genome(s) are only partially covered in the reads.

Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., ... & Pevzner, P. A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, *17*(11), 1103-1110.

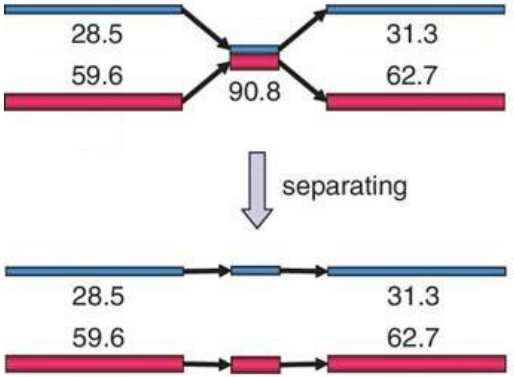
Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824-834.

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Research*, *40*(20), e155-e155.

Assembly: methods (single-genome vs. metagenome)



- **Problem:** Uneven coverage
  - Different microbes are present at different abundances in a microbiome.
  - The *coverage*, or number of reads "supporting" a position in a genome, varies a lot!
  - Worst case scenario: some genome(s) are only partially covered in the reads.
- **Problem:** Intergenomic repeats
  - Stretches of DNA shared by many organisms are repeats!
  - For example, marker genes (e.g. rRNA genes).
  - Some approaches attempt to get around this by *exploiting* uneven coverages across genomes.

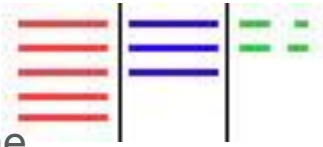


Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., ... & Pevzner, P. A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, *17*(11), 1103-1110.

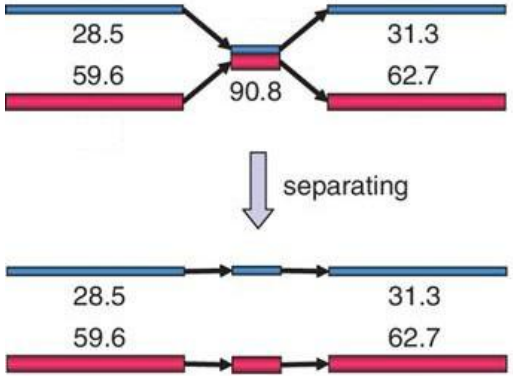
Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824-834.

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Research*, *40*(20), e155-e155.

Assembly: methods (single-genome vs. metagenome)



- **Problem:** Uneven coverage
  - Different microbes are present at different abundances in a microbiome.
  - The *coverage*, or number of reads "supporting" a position in a genome, varies a lot!
  - Worst case scenario: some genome(s) are only partially covered in the reads.
- **Problem:** Intergenomic repeats
  - Stretches of DNA shared by many organisms are repeats!
  - For example, marker genes (e.g. rRNA genes).
  - Some approaches attempt to get around this by *exploiting* uneven coverages across genomes.



- **Problem:** Strain mixtures
  - Similar to intergenomic repeats: a microbiome can contain many strains of a microbe.
  - How can we distinguish between "real" variations and sequencing errors?
  - *Separating* these strains' genomes is very challenging, especially with short reads.

Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., ... & Pevzner, P. A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, *17*(11), 1103-1110.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824-834.

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Research*, *40*(20), e155-e155.

Assembly: methods (single-genome vs. metagenome)

- **Problem:** Strain mixtures
  - Similar to intergenomic repeats: a microbiome can contain many strains of a microbe.
  - How can we distinguish between “real” variations and sequencing errors?
  - *Separating* these strains’ genomes is very challenging, especially with short reads.

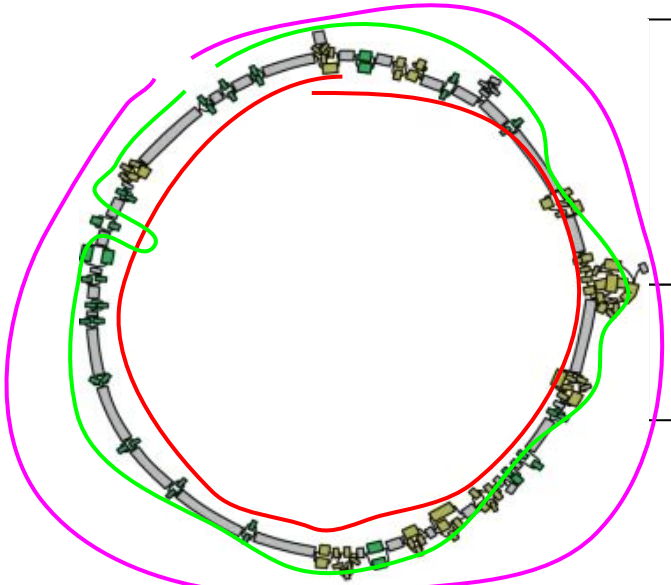
Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., ... & Pevzner, P. A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, *17*(11), 1103-1110.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824-834.

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Research*, *40*(20), e155-e155.

Assembly: methods (single-genome vs. metagenome)

- **Problem:** Strain mixtures
  - Similar to intergenomic repeats: a microbiome can contain many strains of a microbe.
  - How can we distinguish between "real" variations and sequencing errors?
  - *Separating* these strains' genomes is very challenging, especially with short reads.



"An assembly graph of a single connected component in the sheep microbiome dataset before strain collapsing [...] The component represents a bacterial genome of the *Clostridia* class [...] There are 20 simple bubbles (shown in **green**) and 10 superbubbles (shown in **yellow**) that account for 1.2 Mbp out of 2.4 Mbp long genome."

Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., ... & Pevzner, P. A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, *17*(11), 1103-1110.

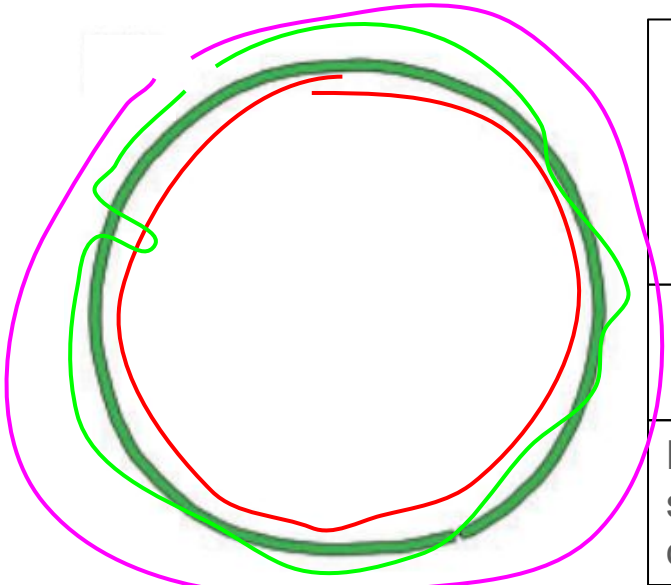
Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., ... & Pevzner, P. A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, *17*(11), 1103-1110.

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, *27*(5), 824-834.

Namiki, T., Hachiya, T., Tanaka, H., & Sakakibara, Y. (2012). MetaVelvet: an extension of Velvet assembler to *de novo* metagenome assembly from short sequence reads. *Nucleic Acids Research*, *40*(20), e155-e155.

Assembly: methods (single-genome vs. metagenome)

- **Problem:** Strain mixtures
  - Similar to intergenomic repeats: a microbiome can contain many strains of a microbe.
  - How can we distinguish between "real" variations and sequencing errors?
  - *Separating* these strains' genomes is very challenging, especially with short reads.



"An assembly graph of a single connected component in the sheep microbiome dataset before strain collapsing [...] The component represents a bacterial genome of the *Clostridia* class [...] There are 20 simple bubbles (shown in **green**) and 10 superbubbles (shown in **yellow**) that account for 1.2 Mbp out of 2.4 Mbp long genome."

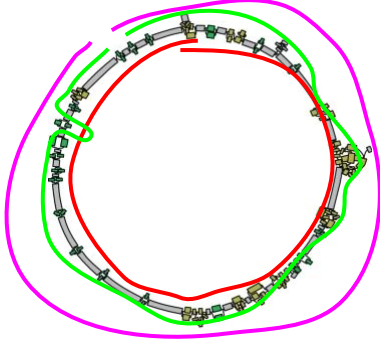
Kolmogorov, M., Bickhart, D. M., Behsaz, B., Gurevich, A., Rayko, M., Shin, S. B., ... & Pevzner, P. A. (2020). metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods*, *17*(11), 1103-1110.

Different assemblers will do different things to deal with these sorts of subtle variations: even "smooth" contigs can conceal a lot of variation. Sometimes this is desirable—sometimes not!

This talk

1. Introduction: Studying microbiomes
   a. Why bother?
   b. Why hasn't this research been more useful?
   c. Defining a goal for this talk
2. Culture-independent (a.k.a. sequencing-based) methods
   a. Marker gene sequencing
   b. Metagenome sequencing
3. Metagenome assembly
   a. Input (*reads*)
   b. Outputs (*contigs*, an *assembly graph*, …)
   c. Methods (*de novo* vs. reference-based, overlap graph vs. de Bruijn graph, …)
4. **Future work: Solving the *strain separation problem***

(Vicedomini et al.'s definition of "strain" matches the one I defined ~45 minutes ago, i.e. a completely unique genome.)



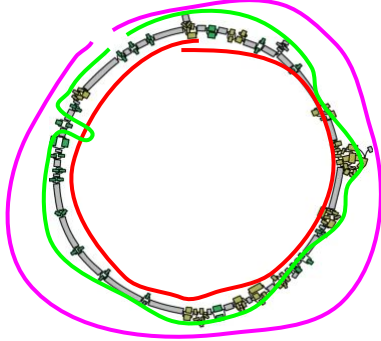
Future work

Strain separation problem (Vicedomini et al., 2021)

"The reconstruction of partial or complete DNA sequences corresponding to strains, at the base level."

Vicedomini, R., Quince, C., Darling, A. E., & Chikhi, R. (2021). Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nature Communications*, *12*(1), 1-14.

(Vicedomini et al.'s definition of "strain" matches the one I defined ~45 minutes ago, i.e. a completely unique genome.)



Future work

Strain separation problem (Vicedomini et al., 2021)

"The reconstruction of partial or complete DNA sequences corresponding to strains, at the base level."

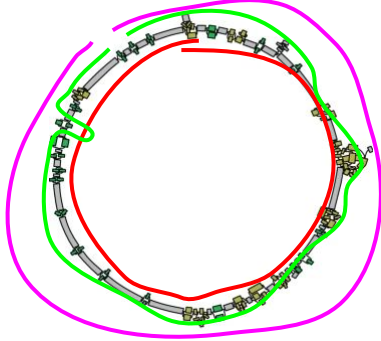
(Local) Metagenome individual haplotyping problem (Nicholls et al., 2021)

"The ideal output [...] is the collection of whole-genome sequences representing all the individual organisms in a microbial community."

Vicedomini, R., Quince, C., Darling, A. E., & Chikhi, R. (2021). Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nature Communications*, *12*(1), 1-14.

Nicholls, S. M., Aubrey, W., De Grave, K., Schietgat, L., Creevey, C. J., & Clare, A. (2021). On the complexity of haplotyping a microbial community. *Bioinformatics*, *37*(10), 1360-1366.

(Vicedomini et al.'s definition of "strain" matches the one I defined ~45 minutes ago, i.e. a completely unique genome.)



Future work

Strain separation problem (Vicedomini et al., 2021)

"The reconstruction of partial or complete DNA sequences corresponding to strains, at the base level."

(Local) Metagenome individual haplotyping problem (Nicholls et al., 2021)

"The ideal output [...] is the collection of whole-genome sequences representing all the individual organisms in a microbial community."

Haplotype assembly problem (Lancia et al., 2001)

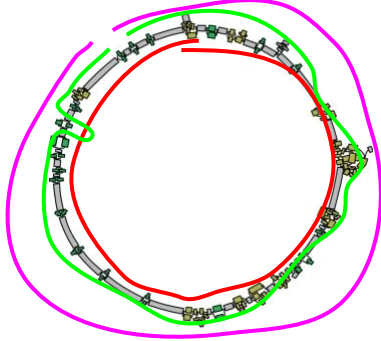
"Given a set of fragments obtained by DNA sequencing from the two copies of a chromosome, reconstruct two haplotypes that would be compatible with all the fragments observed."

Vicedomini, R., Quince, C., Darling, A. E., & Chikhi, R. (2021). Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nature Communications*, *12*(1), 1-14.

Nicholls, S. M., Aubrey, W., De Grave, K., Schietgat, L., Creevey, C. J., & Clare, A. (2021). On the complexity of haplotyping a microbial community. *Bioinformatics*, *37*(10), 1360-1366.

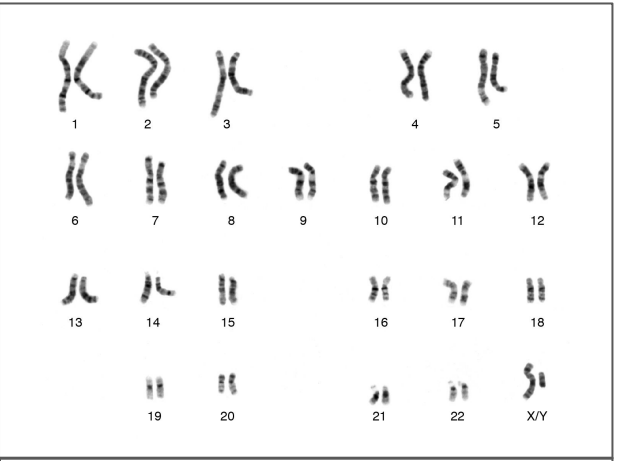
Lancia, G., Bafna, V., Istrail, S., Lippert, R., & Schwartz, R. (2001, August). SNPs problems, complexity, and algorithms. In *European Symposium on Algorithms* (pp. 182-193). Springer, Berlin, Heidelberg.

Future work: “haplotype”?



Humans (or other **diploid** organisms) usually have two copies of each chromosome.

Ordinary assemblers often “smooth out” differences between the chromosomes, creating contigs that are **chimeras** of both chromosomes’ sequences.

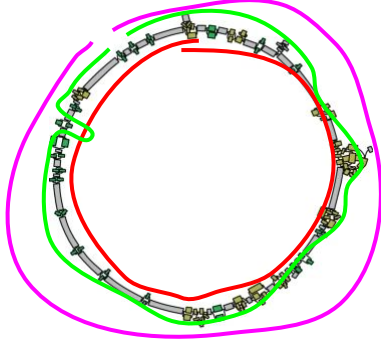


https://www.genome.gov/genetics-glossary/Karyotype

In **haplotype phasing**, we attempt to fix this. Usually, this involves looking for variations which occur on the same reads.

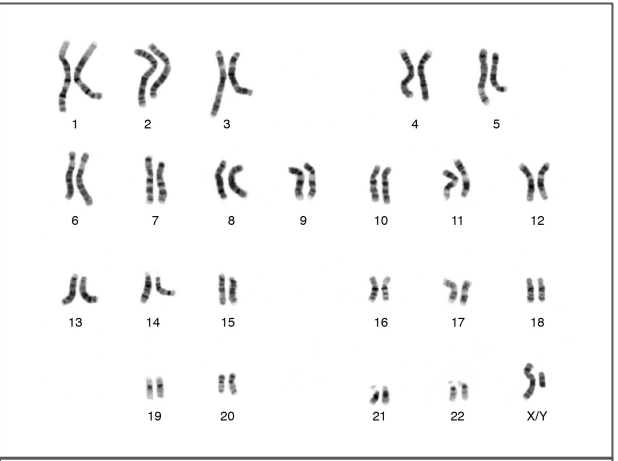
Lancia, G., Bafna, V., Istrail, S., Lippert, R., & Schwartz, R. (2001, August). SNPs problems, complexity, and algorithms. In *European Symposium on Algorithms* (pp. 182-193). Springer, Berlin, Heidelberg.

Future work: “haplotype”?



Humans (or other **diploid** organisms) usually have two copies of each chromosome.

Ordinary assemblers often “smooth out” differences between the chromosomes, creating contigs that are **chimeras** of both chromosomes’ sequences.



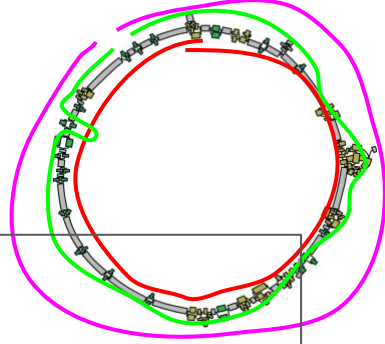
https://www.genome.gov/genetics-glossary/Karyotype

In **haplotype phasing**, we attempt to fix this. Usually, this involves looking for variations which occur on the same reads.

Even in the case of a single human genome, this problem is **NP-Hard**. It gets worse for metagenomes!

Lancia, G., Bafna, V., Istrail, S., Lippert, R., & Schwartz, R. (2001, August). SNPs problems, complexity, and algorithms. In *European Symposium on Algorithms* (pp. 182-193). Springer, Berlin, Heidelberg.

Future work: “metagenomic haplotyping”



A solution to [metagenomic haplotyping] is confounded by five problems:

(i) DNA from every genome needs to be extracted and sequenced to a depth sufficient for recovery,

(ii) genomes share homologous regions that require disambiguation,

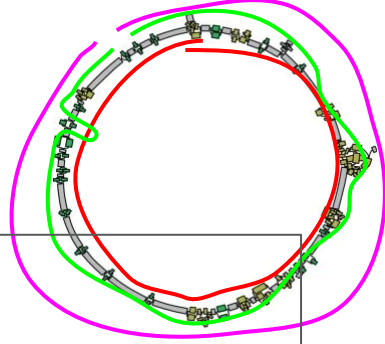
(iii) reads may be of an insufficient length to disambiguate repeats or resolve bridges between variants,

(iv) sequencing error can be indistinguishable from rare haplotypes and

(v) the presence of an unknown number of haplotypes complicates the already computationally difficult (**NP-hard**) (Cilibrasi et al., 2005) problem of haplotyping.

Nicholls, S. M., Aubrey, W., De Grave, K., Schietgat, L., Creevey, C. J., & Clare, A. (2021). On the complexity of haplotyping a microbial community. *Bioinformatics*, *37*(10), 1360-1366.

Future work: “metagenomic haplotyping”



A solution to [metagenomic haplotyping] is confounded by five problems:

(i) DNA from every genome needs to be extracted and sequenced to a depth sufficient for recovery,

So increase sequencing depth!

(ii) genomes share homologous regions that require disambiguation,

So increase read lengths!

(iii) reads may be of an insufficient length to disambiguate repeats or resolve bridges between variants,

So increase read lengths (again)!

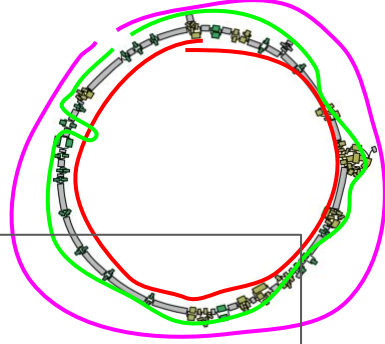
(iv) sequencing error can be indistinguishable from rare haplotypes and

So decrease error rates!

(v) the presence of an unknown number of haplotypes complicates the already computationally difficult (**NP-hard**) (Cilibrasi et al., 2005) problem of haplotyping.

So use better algorithms?

Future work: “metagenomic haplotyping”



A solution to [metagenomic haplotyping] is confounded by five problems:

(i) DNA from [all] [species] must be extracted and sequenced to a depth sufficient for recovery,

So increase sequencing depth!

(ii) genome [assembly] must resolve ambiguation,

So increase read lengths!

(iii) read [alignment] must resolve [intermediate] repeats or resolve bridges between [contigs],

So increase read lengths (again)!

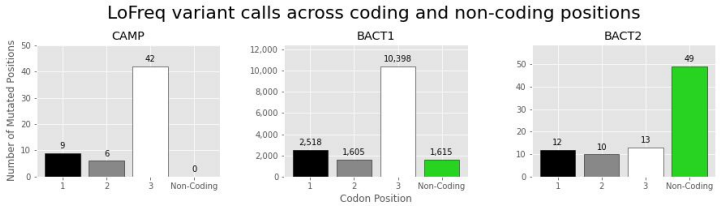
(iv) sequencing [errors] must not obscure rare haplotypes and

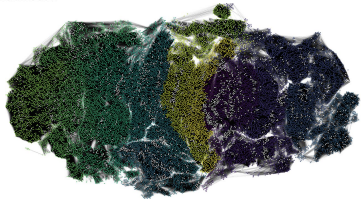
So decrease error rates!

(v) the presence of an unknown number of haplotypes complicates the already computationally difficult (**NP-hard**) (Cilibrasi et al., 2005) problem of haplotyping.

So use better algorithms?

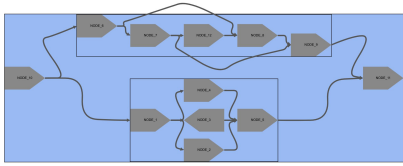
Future work: next steps forward



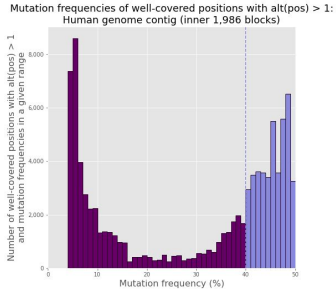


- Improving the detection (and phasing!) of rare mutations in HiFi metagenome sequencing data
  - with Misha Kolmogorov (UCSC) and Pavel Pevzner (UCSD)

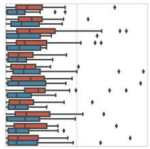
- Improved methods for visualizing metagenome assembly graphs
  - with Jay Ghurye (Verily Life Sciences), Todd Treangen (Rice), Misha Kolmogorov (UCSC), Pavel Pevzner (UCSD), Jacquelyn Michaelis + Harihara Muralidharan + Mihai Pop (Maryland)



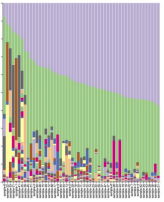
- Improved classification of prokaryotic/eukaryotic contigs
  - with Misha Kolmogorov (UCSC) and Pavel Pevzner (UCSD)



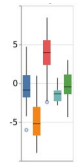
Some things I've been up to at UCSD



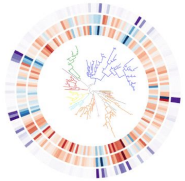
- Contributing to various analyses using marker gene / metagenome sequencing
  - **"Co-assemblies" of metagenome sequencing data:** Sanders JG, Nurk S, Salido RA, Minich J, Xu ZZ, Martino C, **Fedarko M**, Arthur TD, Chen F, Boland BS, Humphrey GC, Brennan C, Sanders K, Gaffney J, Jepsen K, Khosroheidari M, Green C, Liyange M, Dang JW, Phelan VV, Quinn RA, Bankevich A, Chang JT, Rana TM, Conrad DJ, Sandborn WJ, Smarr L, Dorrestein PC, Pevzner PA, and Knight R (2019). "Optimizing sequencing protocols for leaderboard metagenomics by combining long and short reads." *Genome Biology, 20*(1):226.



  - **Analyzing marker gene sequencing data:** Huey SL, Jiang L, **Fedarko MW**, McDonald D, Martino C, Ali F, Russell DG, Udipi SA, Thorat A, Thakker V, Ghugre P, Potdar RD, Chopra H, Rajagopalan K, Haas JD, Finkelstein JL, Knight R, and Mehta S (2020). "Nutrition and the Gut Microbiota in 10- to 18-Month-Old Children Living in Urban Slums of Mumbai, India." *mSphere, 5*(5):e00731-20.



- Developing visualization tools for microbiome sequencing data
  - **Differential abundance:** **Fedarko MW**, Martino C, Morton JT, González A, Rahman G, Marotz CA, Minich JJ, Allen EA, and Knight R (2020). "Visualizing 'omic feature rankings and log-ratios using Qurro." *NAR Genomics and Bioinformatics, 2*(2):lqaa023.



  - **Phylogenetic trees (and associated data):** Cantrell K*, **Fedarko MW***, Rahman G, McDonald D, Yang Y, Zaw T, Gonzalez A, Janssen S, Estaki M, Haiminen N, Beck KL, Zhu Q, Sayyari E, Morton JT, Armstrong G, Tripathi A, Gauglitz JM, Marotz C, Matteson NL, Martino C, Sanders JG, Carrieri AP, Song SJ, Swafford AD, Dorrestein PC, Andersen KG, Parida L, Kim H-C, Vázquez-Baeza Y, and Knight R (2021). "EMPress Enables Tree-Guided, Interactive, and Exploratory Analyses of Multi-omic Data Sets." *mSystems, 6*(2):e01216-20. (* = contributed equally)

Research Exam Committee and Logistics

Thank you!

Gary Cottrell, Vineet Bafna, Melissa Gymrek, Julie Conner

Advice/support over the years

Pavel Pevzner
Mikhail Kolmogorov
Andrey Bzikadze
Vikram Sirupurapu
Rob Knight
Yoshiki Vázquez-Baeza
Lisa Marotz
Cameron Martino
Jamie Morton
Antonio González
Gibraan Rahman
Jake Minich
Eric Allen
Dan Hakim

Kalen Cantrell
Daniel McDonald
Yimeng Yang
Thant Zaw
Stefan Janssen
Mehrbod Estaki
Niina Haiminen
Kristen Beck
Qiyun Zhu
Erfan Sayyari
George Armstrong
Priya Tripathi
Julia Gauglitz
Nate Matteson

Jon Sanders
Anna Paola Carrieri
Se Jin Song
Austin Swafford
Pieter Dorrestein
Kristian Andersen
Laxmi Parida
Ho-Cheol Kim
Larry Smarr
Gail Ackermann
Jeff DeReus
Michiko Souza
Justin Shaffer
Pedro Belda-Ferre

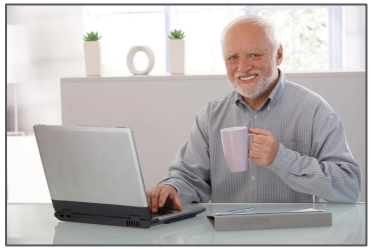
Greg Humphrey
Celeste Allaband
Rodolfo Salido
Greg Poore
Victor Cantu
Jeffrey Chiu
Franck Lejzerowicz
Shi Huang
Sarah Adams
Tomasz Kosciolek
Zech Xu
Charles Cowart
Farhana Ali
Robert Mills

Alison Vrbanc
Bryn Taylor
Jerry Kennedy
Yna Villanueva
Justine Debelius
Evan Bolyen
Matthew Dillon
Jay Ghurye
Jacquelyn Michaelis
Harihara Muralidharan
Nidhi Shah
Brian Brubach
Todd Treangen
Mihai Pop

Research Exam Committee and Logistics

Thank you!

Gary Cottrell, Vineet Bafna, Melissa Gymrek, Julie Conner



Advice/support over the years

Pavel Pevzner
Mikhail Kolmogorov
Andrey Bzikadze
Vikram Sirupurapu
Rob Knight
Yoshiki Vázquez-Baeza
Lisa Marotz
Cameron Martino
Jamie Morton
Antonio González
Gibraan Rahman
Jake Minich
Eric Allen
Dan Hakim

Kalen Cantrell
Daniel McDonald
Yimeng Yang
Thant Zaw
Stefan Janssen
Mehrbod Estaki
Niina Haiminen
Kristen Beck
Qiyun Zhu
Erfan Sayyari
George Armstrong
Priya Tripathi
Julia Gauglitz
Nate Matteson

Jon Sanders
Anna Paola Carrieri
Se Jin Song
Austin Swafford
Pieter Dorrestein
Kristian Andersen
Laxmi Parida
Ho-Cheol Kim
Larry Smarr
Gail Ackermann
Jeff DeReus
Michiko Souza
Justin Shaffer
Pedro Belda-Ferre

Greg Humphrey
Celeste Allaband
Rodolfo Salido
Greg Poore
Victor Cantu
Jeffrey Chiu
Franck Lejzerowicz
Shi Huang
Sarah Adams
Tomasz Kosciolek
Zech Xu
Charles Cowart
Farhana Ali
Robert Mills

Alison Vrbanc
Bryn Taylor
Jerry Kennedy
Yna Villanueva
Justine Debelius
Evan Bolyen
Matthew Dillon
Jay Ghurye
Jacquelyn Michaelis
Harihara Muralidharan
Nidhi Shah
Brian Brubach
Todd Treangen
Mihai Pop

Misc. Acknowledgements

Emojis: Google emojis from emojipedia.org: https://emojipedia.org/pile-of-poo/ (removed the emoji eyes manually in GIMP), https://emojipedia.org/non-potable-water/, https://emojipedia.org/potable-water/, https://emojipedia.org/mobile-phone/
Citation of Blaser 1992 in the context of H. pylori based on the Strainberry paper's introduction: Vicedomini, R., Quince, C., Darling, A. E., & Chikhi, R. (2021). Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nature Communications*, *12*(1), 1-14.
Taxonomic ranks figure modified from https://en.wikipedia.org/wiki/Taxonomic_rank#/media/File:Taxonomic_Rank_Graph.svg, % Annina Breen.
E. coli phylogeny from Dunne, K. A., Chaudhuri, R. R., Rossiter, A. E., Beriotto, I., Browning, D. F., Squire, D., ... & Henderson, I. R. (2017). Sequencing a piece of history: complete genome sequence of the original Escherichia coli strain. *Microbial Genomics*, *3*(3).
Stock photos of someone in a suit kicking a can down the road all from Shutterstock.com % Jim Barber (all images marked as royalty-free). Why were there five separate images of this? That's a great question. I wish my job was literally just putting on a suit and kicking cans down a road. That would be so sick. I bet it pays better than grad school. Wait, why are you reading this? Seriously, there's nothing important here. It's just links. And this text.
https://www.shutterstock.com/search/kick+the+can,
https://www.shutterstock.com/image-photo/politicians-shoe-stops-dented-can-rolling-85554979,
https://www.shutterstock.com/image-photo/close-politicians-shoe-kicking-dented-shiny-85554970,
https://www.shutterstock.com/image-photo/close-shiny-dentedl-can-sitting-on-85554964,
https://www.shutterstock.com/image-photo/close-politicians-shoe-kicking-dented-shiny-85554973,
https://www.shutterstock.com/image-photo/politician-kicks-shiny-dented-can-down-85554967
Harold's face:
https://www.independent.co.uk/arts-entertainment/interviews/hide-pain-harold-meme-gif-interview-model-real-name-arato-andras-thumbs-stock-photo-a7835076.html
Quote about functional annotation and E. coli distance is from C. Frioux, D. Singh, T. Korcsmaros, and F. Hildebrand. From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. Computational and Structural Biotechnology Journal, 18:1722–1734, 2020.
Jigsaw puzzle photo: From VisitIndiana.com. Also, I acknowledge that I used this figure (and the Commins figure for assembly) in a talk I gave last December.
Original PCR paper: Mullis et al. 1986. Specific Enzymatic Amplification of DNA In Vitro: The Polymerase Chain Reaction.
Source about the AB370 being the first sequencer: https://www.hindawi.com/journals/bmri/2012/251364/
Use of the Bambus 2 "variant" figure in the context of variant calling based on Serge's PhD defense from 9 years ago: Koren, S. (2012). Genome Assembly: Novel Applications by Harnessing Emerging Sequencing Technologies and Graph Algorithms. http://www.sergek.umiacs.io/presentations/ThesisTalk_final.pdf. I already cited this when using the "read lengths help" figure a few slides beforehand, but I figure I might as well make that clear here. Serge's a cool dude.

Funding

Fall 2018–Winter 2019
Standard first-year CSE department fellowship

Spring 2019–Summer 2019
Joint University Microelectronics Program (JUMP)'s
Center for Research on Intelligent Storage and Processing-in-memory (CRISP)

Fall 2019–Fall 2020
IBM Research AI, via the AI Horizons Network and
the UCSD Center for Microbiome Innovation (CMI)

Winter 2021
Teaching assistantship (CSE 282)

Spring 2021–
Pevzner Lab grants

Bonus: So how many microbes *are* there?

Turnbaugh 2007: “The vast majority of the **10–100 trillion microbes** in the human gastrointestinal tract live in the colon.”

Locey and Lennon 2016: “[...] we predict that Earth is home to upward of **1 trillion microbial species**.”

Willis 2016: The method used by L&L 2016 isn’t statistically admissible!

∴ Maybe the only answer right now that won’t anger any statistician or biologist: “a lot, I guess”

An obesity-associated gut microbiome with increased capacity for energy harvest

Peter J. Turnbaugh[1], Ruth E. Ley[1], Michael A. Mahowald[1], Vincent Magrini[2], Elaine R. Mardis[1,2] & Jeffrey I. Gordon[1]

"We performed microbiota transplantation experiments to test directly the notion that the ob/ob microbiota has an increased capacity to harvest energy from the diet and to determine whether increased adiposity is a transmissible trait. **Adult germ-free C57BL/ 6J mice were colonized (by gavage) with a microbiota harvested from the caecum of obese (ob/ob) or lean (1/1) donors (1 donor and 4–5 germ-free recipients per treatment group per experiment; two independent experiments).** 16S-rRNA-gene-sequence-based surveys confirmed that the ob/ob donor microbiota had a greater relative abundance of Firmicutes compared with the lean donor microbiota (Supplementary Fig. 4 and Supplementary Table 7). Furthermore, the ob/ob recipient microbiota had a significantly higher relative abundance of Firmicutes compared with the lean recipient microbiota ($P < 0.05$, two-tailed Student's t-test). UniFrac analysis of 16S rRNA gene sequences obtained from the recipients' caecal microbiotas revealed that they cluster according to the input donor community (Supplementary Fig. 4): that is, the initial colonizing community structure did not exhibit marked changes by the end of the two-week experiment. **There was no statistically significant difference in (1) chow consumption over the 14-day period (55.4 6 2.5 g (ob/ob) versus 54.0 6 1.2 g (1/1); caloric density of chow, 3.7 kcal g21), (2) initial body fat (2.7 6 0.2 g for both groups as measured by dual-energy X-ray absorptiometry), or (3) initial weight between the recipients of lean and obese microbiotas. Strikingly, mice colonized with an ob/ob microbiota exhibited a significantly greater percentage increase in body fat over two weeks than mice colonized with a 1/1 microbiota** (Fig. 3c; 47 6 8.3 versus 27 6 3.6 percentage increase or 1.3 6 0.2 versus 0.86 6 0.1 g fat (dual-energy X-ray absorptiometry): at 9.3 kcal g21 fat, this corresponds to a difference of 4 kcal or 2% of total calories consumed)."

Bonus: Culture-Independent methods and dark matter

"It is estimated that **>99% of microorganisms observable in nature** typically are not cultivated by using standard techniques."

Hugenholtz, P., Goebel, B. M., & Pace, N. R. (1998). Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. *Journal of bacteriology*, *180*(18), 4765-4774.

Some folks have used the term "dark matter" to refer to these so-far-uncultured microbes, but it's not a great analogy...

Microbial Dark Matter: The mullet of microbial ecology

A. Murat Eren (Meren)

Web Email Twitter Github

The Dark Matter Metaphor in Biology

By Iddo on November 27th, 2015

"Tempting as it may be, perhaps we should calm down on the use of the term dark matter in biology. Biology is confusing, complicated, and mysterious enough without it."

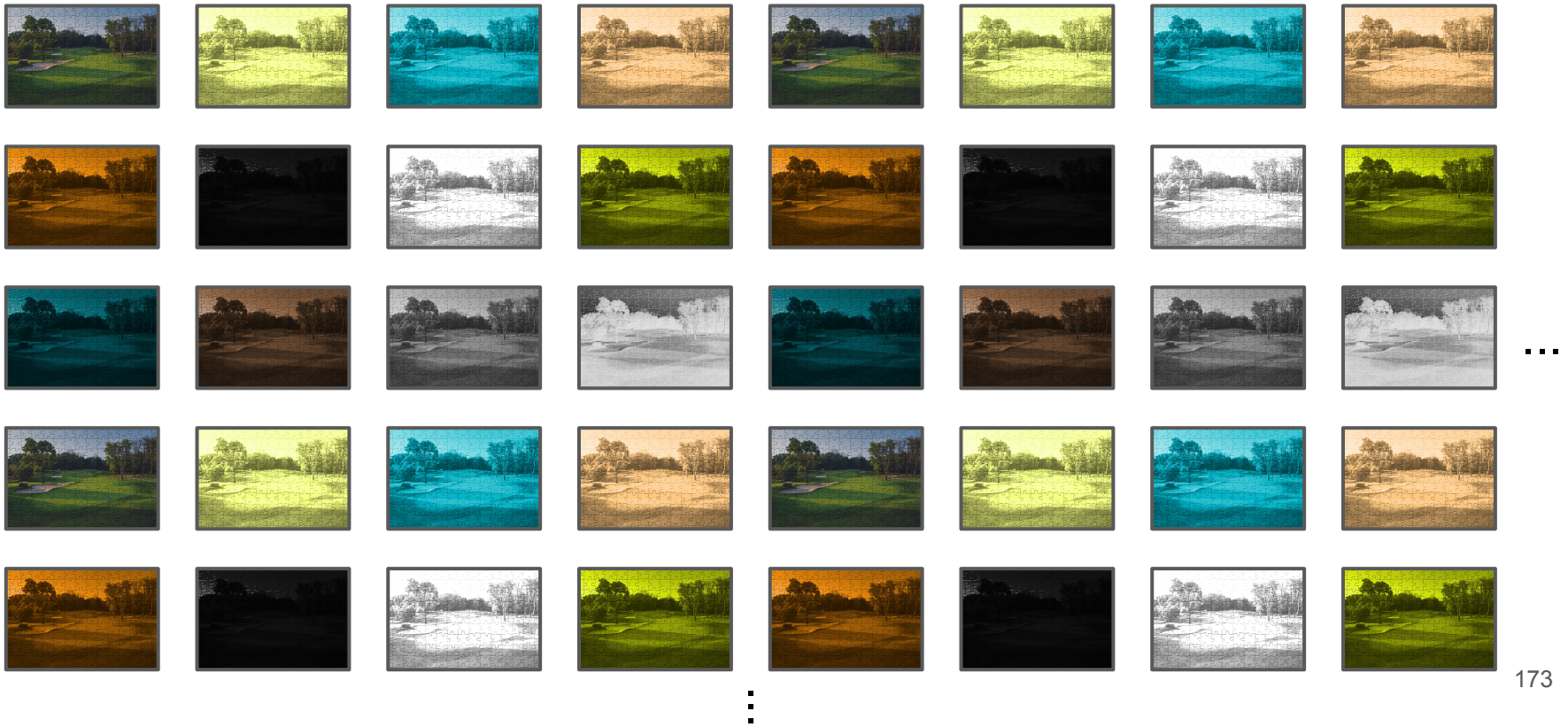
Bonus: About functional annotation...

"It says something of our ability to annotate genomes, that the proportion of a genome functionally annotated is often correlated to the genetic distance to the very well researched *Escherichia coli* (anecdotal observation)."

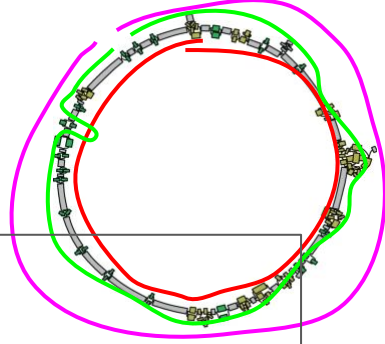
C. Frioux, D. Singh, T. Korcsmaros, and F. Hildebrand. From bag-of-genes to bag-of-genomes: metabolic modelling of communities in the era of metagenome-assembled genomes. *Computational and Structural Biotechnology Journal*, 18:1722–1734, 2020.

Bonus: Assembly (single genome)

Bonus: Assembly (metagenome)



Bonus: “metagenomic haplotyping”??????



A solution to [metagenomic ve problems:

(i) DNA from every genome ced to a depth sufficient for recovery,

So increase sequencing depth!

(ii) genomes share homolog guation,

So increase read lengths!

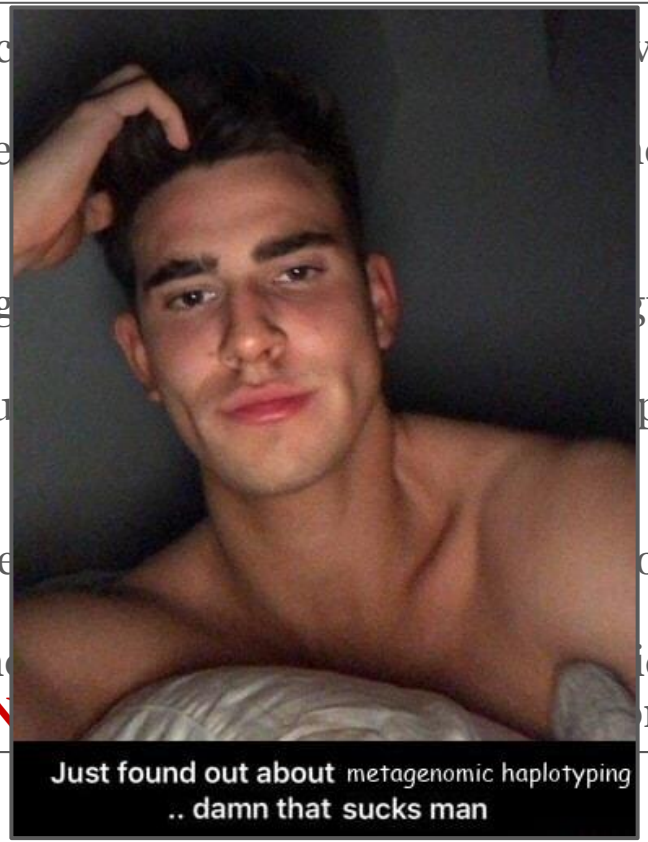
(iii) reads may be of an insu peats or resolve bridges between variants,

So increase read lengths (again)!

(iv) sequencing error can be otypes and

So decrease error rates!

(v) the presence of an unkn icates the already computationally difficult (N roblem of haplotyping.



So use better algorithms?