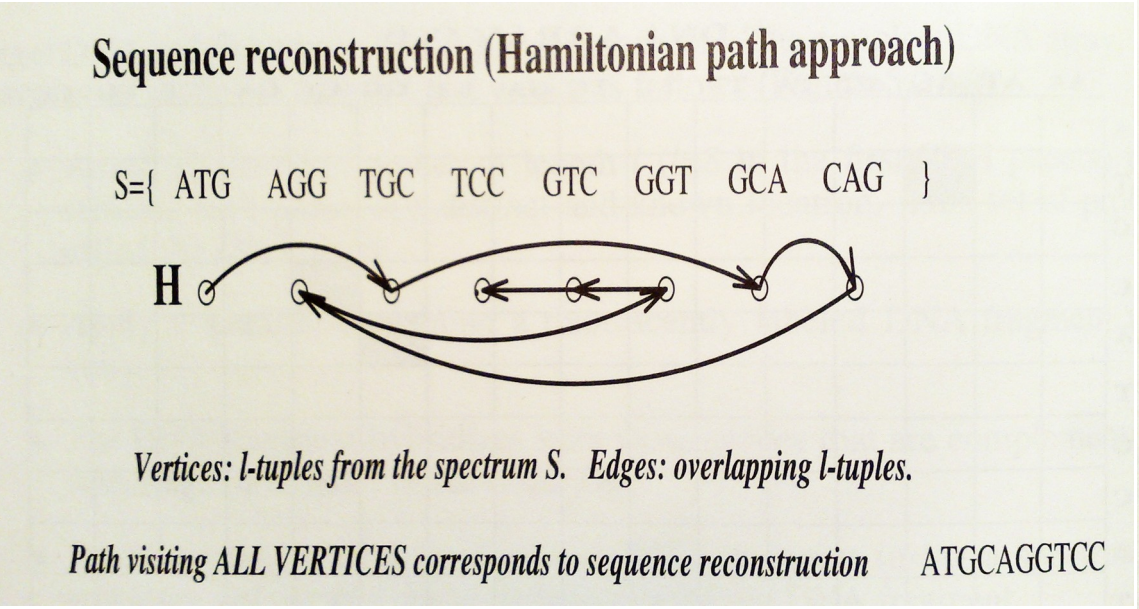
Introduction to compuational biology

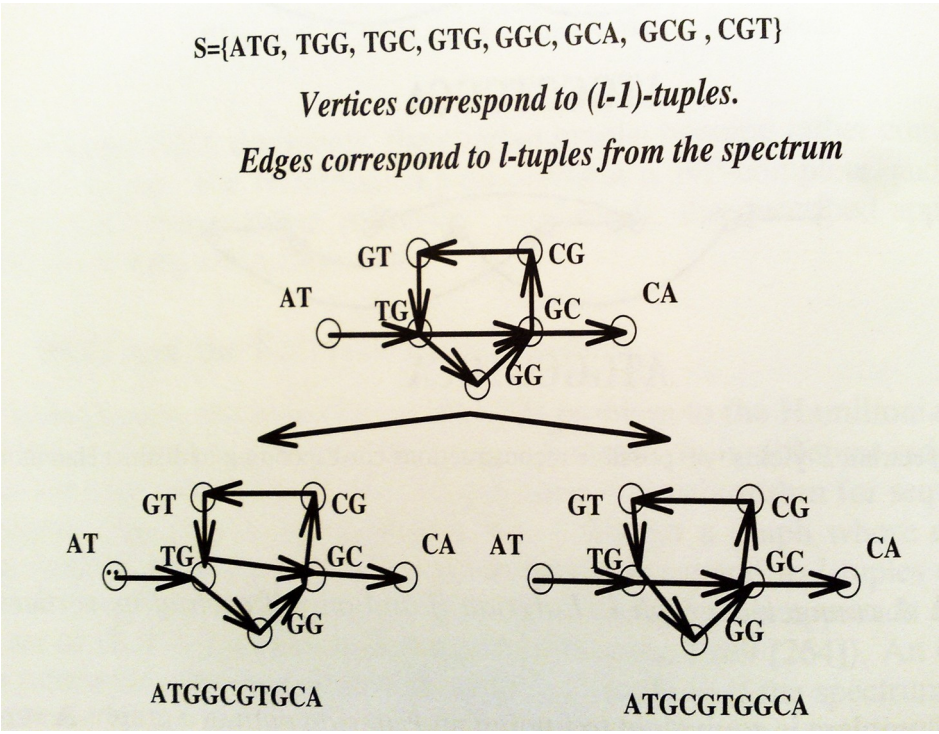
Genome assembly and metagenomics

May 26th 2020

Sequence reconstruction

- Given the spectrum of observed k-mers, we can reconstruct the sequence
- Direct approach leads to the Hamiltonian path problem (NP-Complete)
- Small change in the k-mer representation leads to Eulerian path finding (Pevzner 2000)



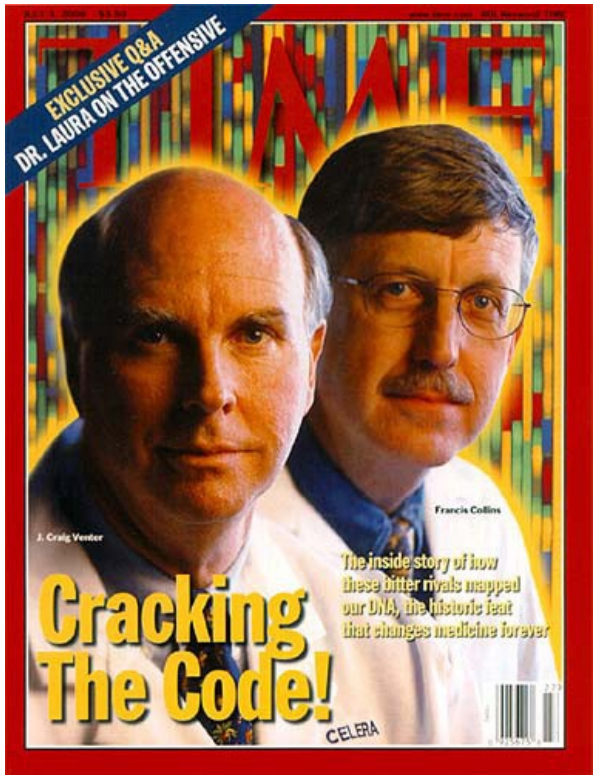


A historical digression on DNA sequence assembly

- Human Genome project
  - Started in 1984, funding since 1990, finished in 2003
  - ~$3 billion
  - Results announced in 2000 by the US president Clinton and UK prime minister Blair
- Celera genomics project
  - Started later in 1996
  - Budget ~$300 million
  - Aimed to commercialize genomic information
  - Results announced jointly with HGP

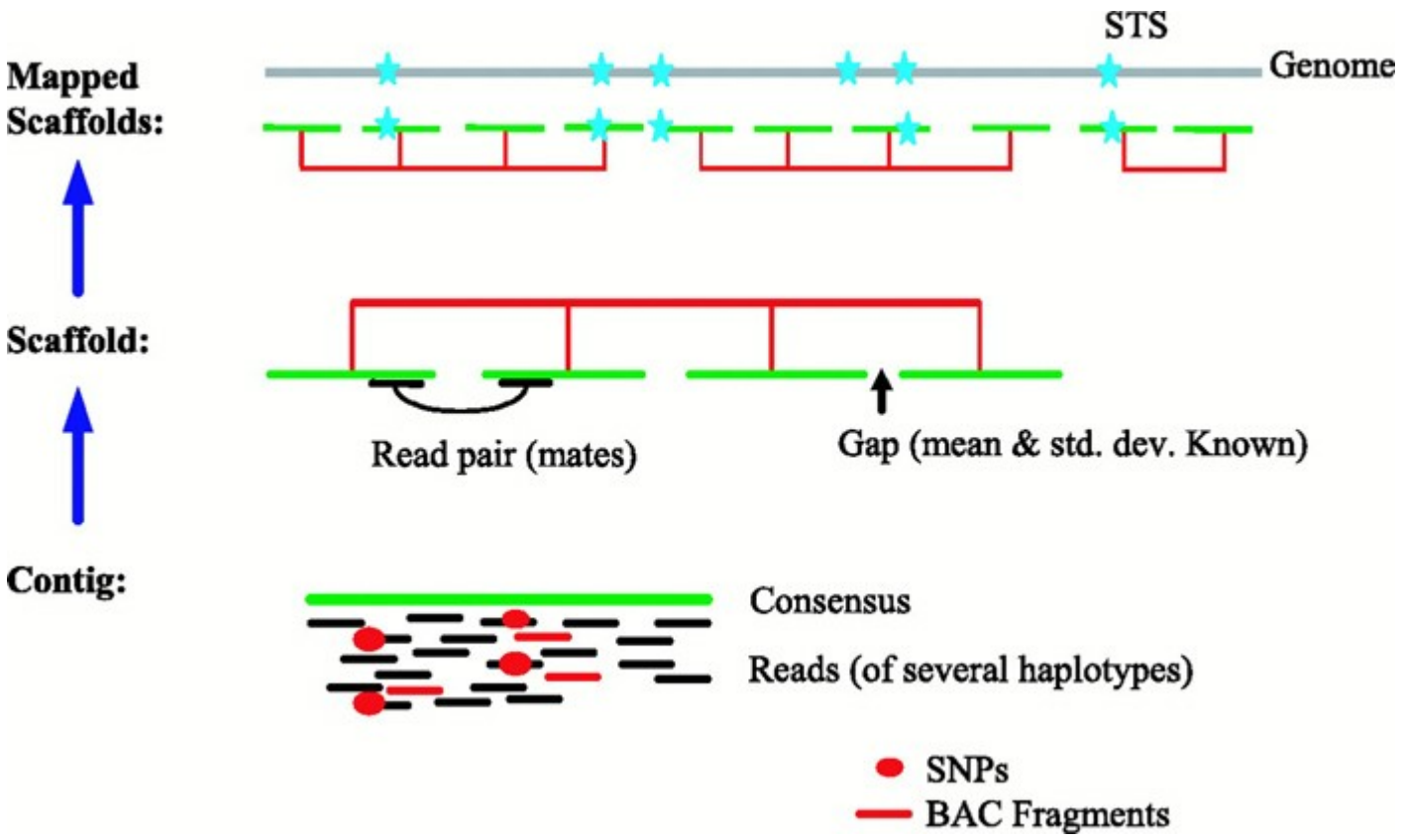
HGP announcement

- First draft announced jointly by two competing consortia
- Brought fame to Craig Venter and Francis Collins, but prevented genome commercialization

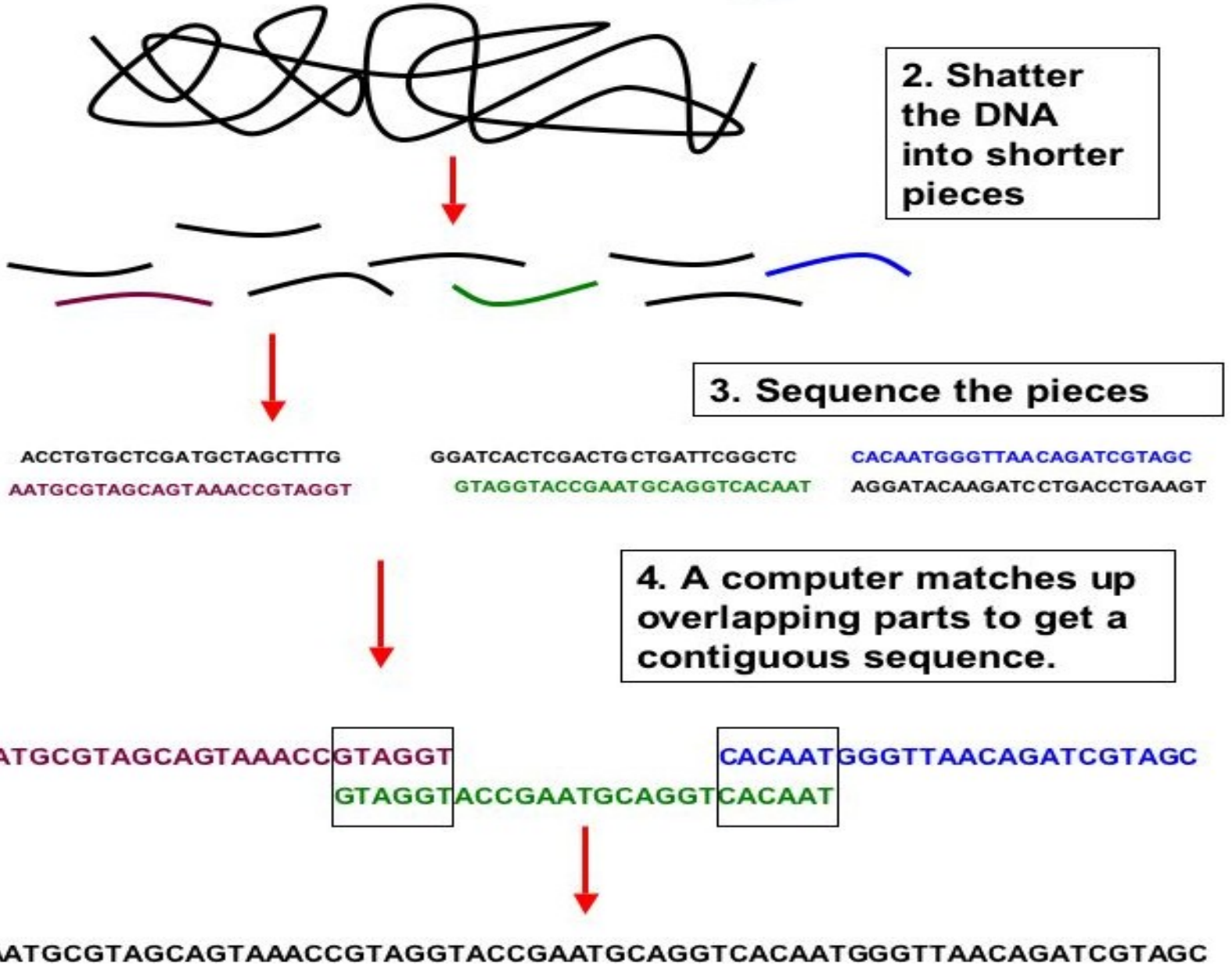


Classical genome assembly (HGP)

- Oredrly process with restriction mapped fragments and scaffold assembly



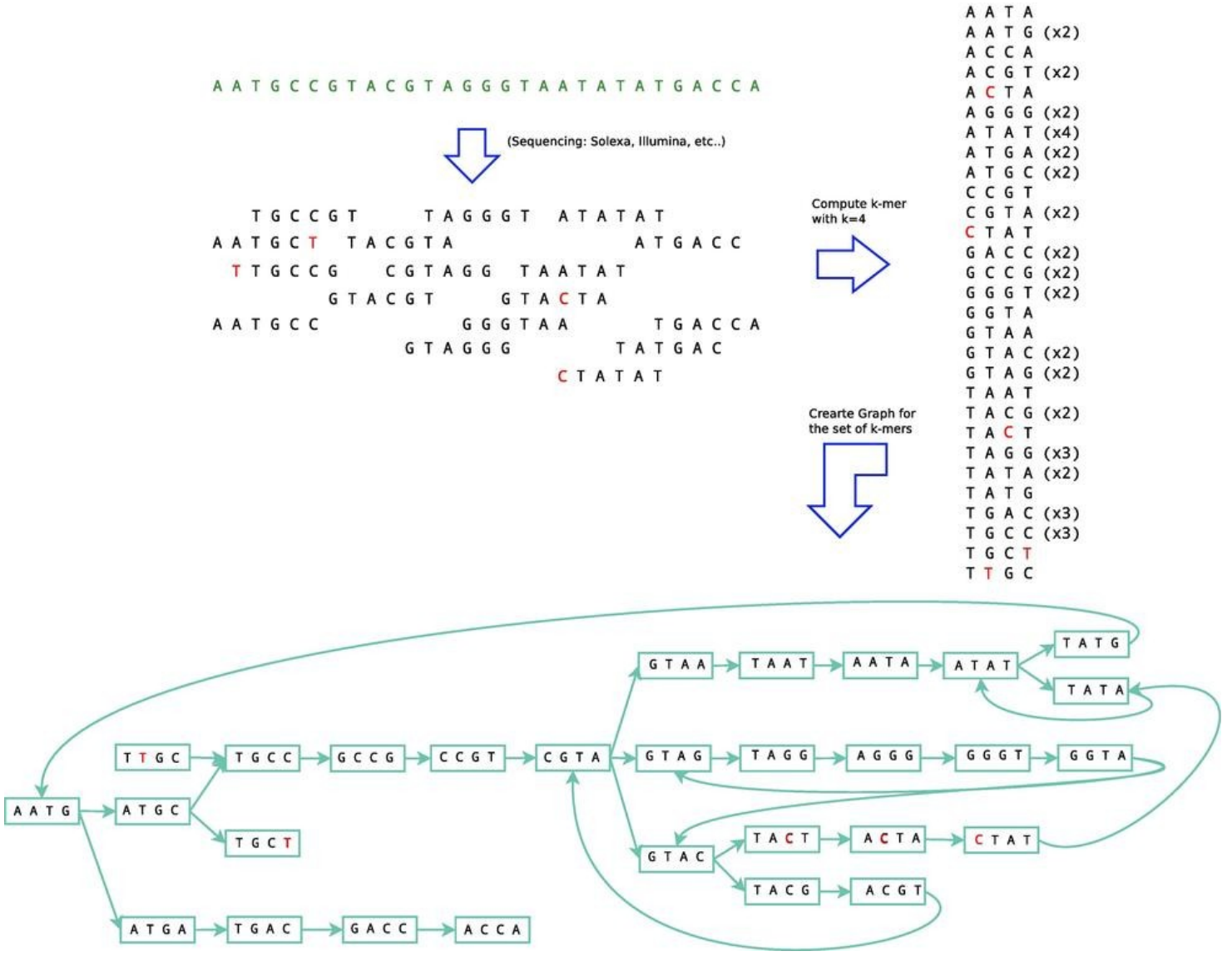
Shotgun genome sequencing (Celera, E. Myers)



Take-home message from HGP

- Celera started later and could take advantage of much cheaper computing power, therefore did not spend so much time on planning different stages of the wet-lab process
- In this case the Moore’s law and computer scientists (E. Myers in particular) allowed for speeding up the process

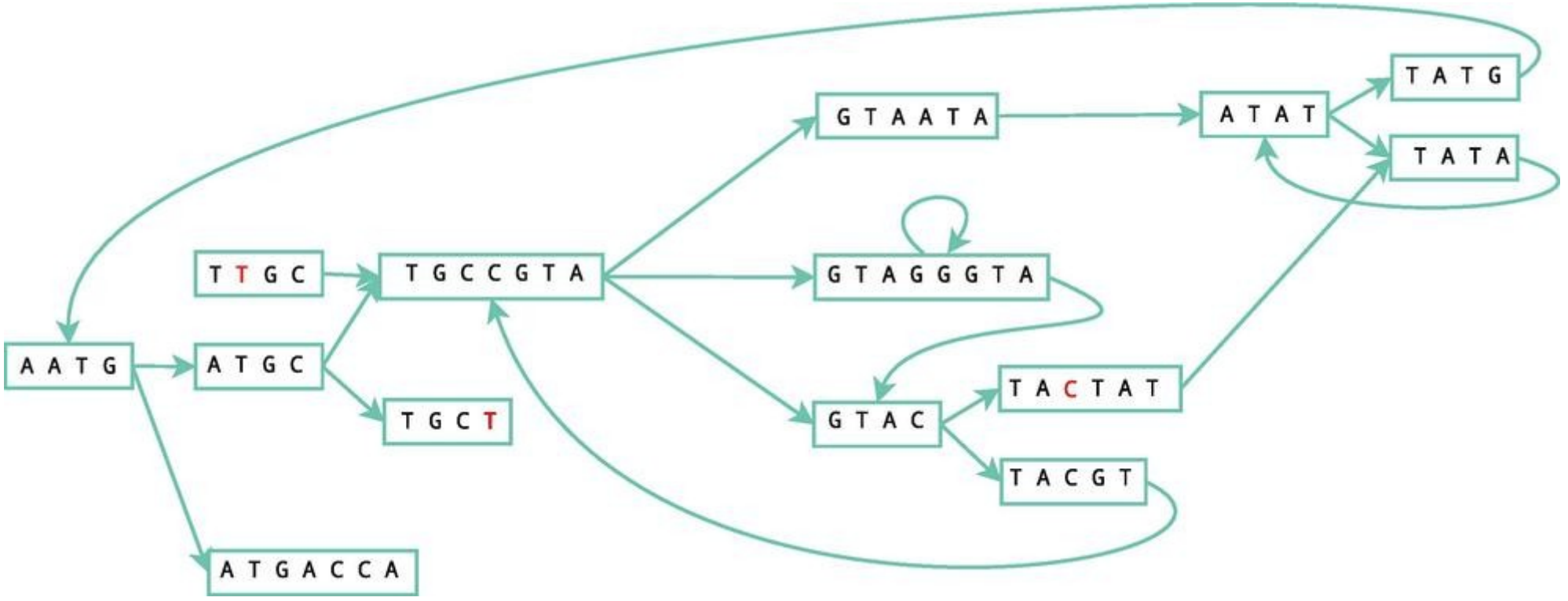
Sequence asembly from short reads



VELVET assembler, Zerbino et al. 2008

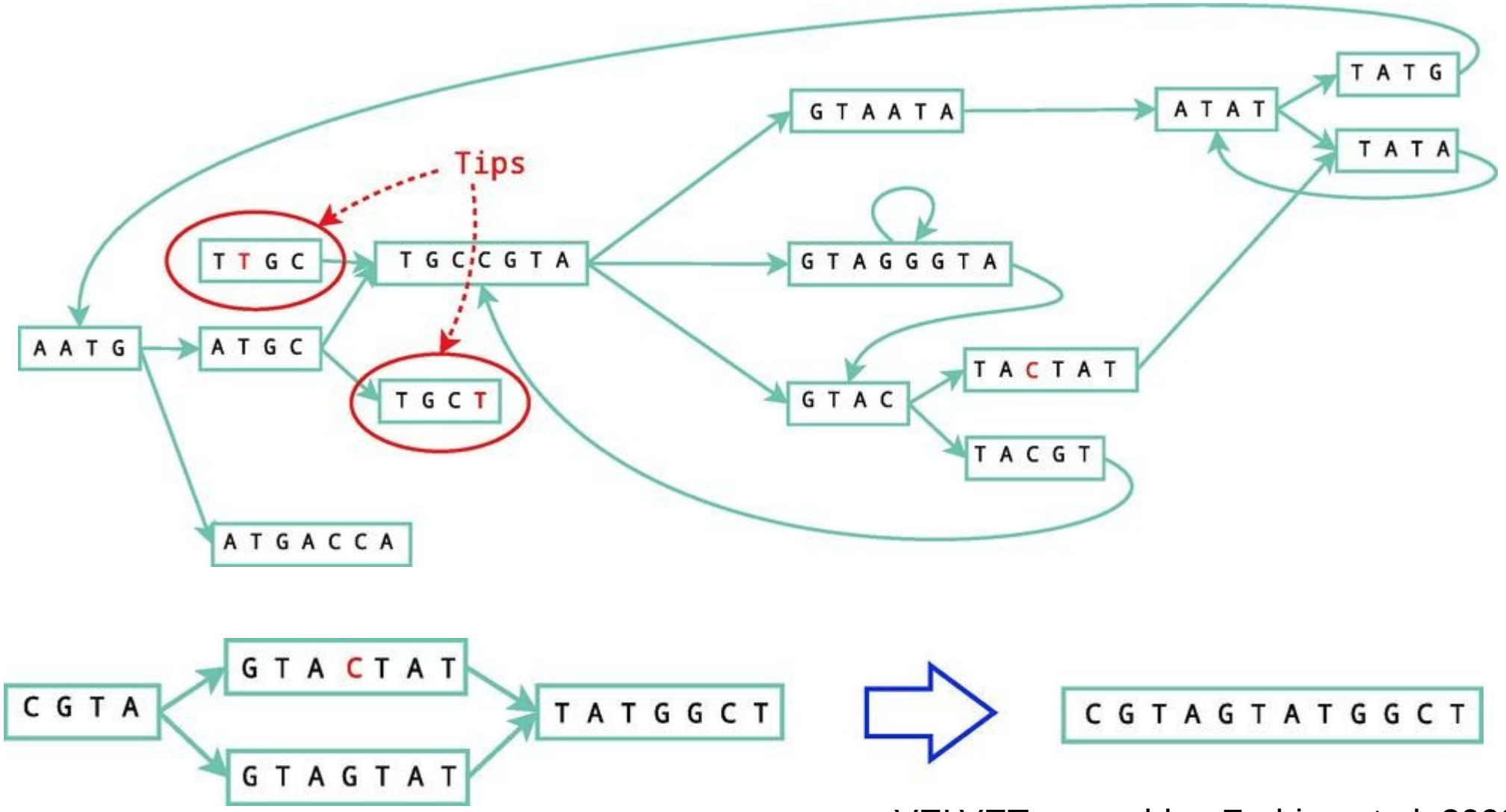
Simplification of deBruijn graph

- We can compress paths without forks



VELVET assembler, Zerbino et al. 2008

Tips and bubble removal



VELVET assembler, Zerbino et al. 2008

De novo assembly

- De novo assemblers (VELVET, Spades, etc.) are ressurecting the idea behind Sequencing by hybridization
- Even though there are limitations to their use (repetitive regions, k-mer length, memory constraints) they are very useful in contig creation from raw short reads
- Many heuristic improvements and specialized tools for specific applications

Metagenomics

- Popularized by Craig Venter in Global Ocean Sampling expedition
- Shotgun sequencing of microbes from Sargasso sea
- Identified many novel gene sequences without attributing them to specific species
- Now very frequently done in other environments: soil, human skin, human intestine
- Helpful in finding new important enzymes (from soil around chemical waste facilities)
- Identified some microbes that are relevant for human health

RESEARCH ARTICLE

Environmental Genome Shotgun Sequencing of the Sargasso Sea

J. Craig Venter,[1]* Karin Remington,[1] John F. Heidelberg,[3] Aaron L. Halpern,[2] Doug Rusch,[2] Jonathan A. Eisen,[3] Dongying Wu,[3] Ian Paulsen,[3] Karen E. Nelson,[3] William Nelson,[3] Derrick E. Fouts,[3] Samuel Levy,[2] Anthony H. Knap,[6] Michael W. Lomas,[6] Ken Nealson,[5] Owen White,[3] Jeremy Peterson,[3] Jeff Hoffman,[1] Rachel Parsons,[6] Holly Baden-Tillson,[1] Cynthia Pfannkoch,[1] Yu-Hui Rogers,[4] Hamilton O. Smith[1]

We have applied "whole-genome shotgun sequencing" to microbial populations collected en masse on tangential flow and impact filters from seawater samples collected from the Sargasso Sea near Bermuda. A total of 1.045 billion base pairs of nonredundant sequence was generated, annotated, and analyzed to elucidate the gene content, diversity, and relative abundance of the organisms within these environmental samples. These data are estimated to derive from at least 1800 genomic species based on sequence relatedness, including 148 previously unknown bacterial phylotypes. We have identified over 1.2 million previously unknown genes represented in these samples, including more than 782 new rhodopsin-like photoreceptors. Variation in species present and stoichiometry suggests substantial oceanic microbial diversity.

Venter et al. Science, 2004

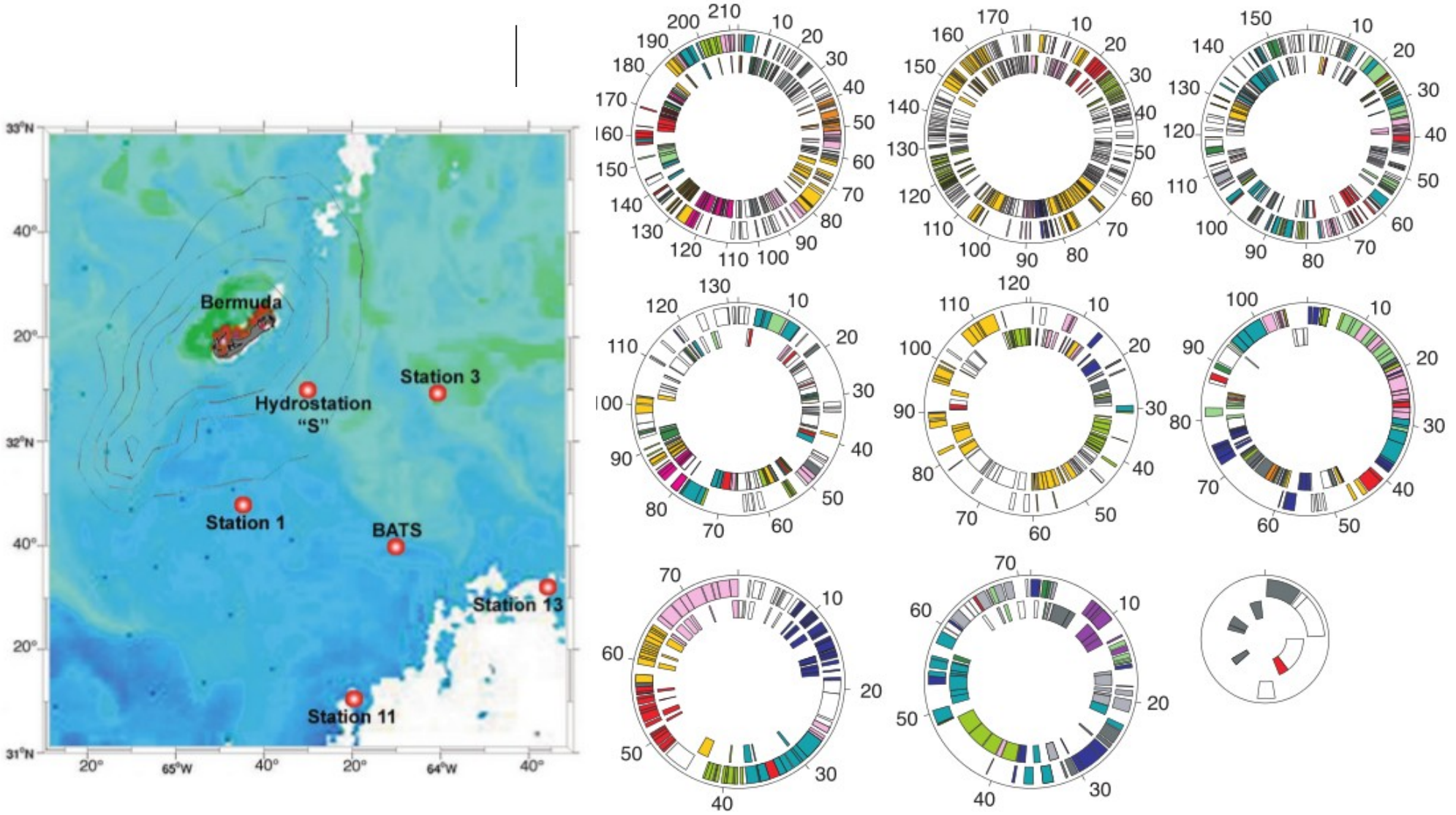


Fig. 4. Circular diagrams of nine complete megaplasmids. Genes encoded in the forward direction are shown in the outer concentric circle; reverse coding genes are shown in the inner concentric circle. The genes have been given role category assignment and colored accordingly: amino acid biosynthesis, violet; biosynthesis of cofactors, prosthetic groups, and carriers, light blue; cell envelope, light green; cellular processes, red; central intermediary metabolism, brown; DNA metabolism, gold; energy metabolism, light gray; fatty acid and phospholipid metabolism, magenta; protein fate and protein synthesis, pink; purines, pyrimidines, nucleosides, and nucleotides, orange; regulatory functions and signal transduction, olive; transcription, dark green; transport and binding proteins, blue-green; genes with no known homology to other proteins and genes with homology to genes with no known function, white; genes of unknown function, gray; Tick marks are placed on 10-kb intervals.

Vol 464|4 March 2010|doi:10.1038/nature08821

nature

ARTICLES

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin[1]*, Ruiqiang Li[1]*, Jeroen Raes[2,3], Manimozhiyan Arumugam[2], Kristoffer Solvsten Burgdorf[4], Chaysavanh Manichanh[5], Trine Nielsen[4], Nicolas Pons[6], Florence Levenez[6], Takuji Yamada[2], Daniel R. Mende[2], Junhua Li[1,7], Junming Xu[1], Shaochuan Li[1], Dongfang Li[1,8], Jianjun Cao[1], Bo Wang[1], Huiqing Liang[1], Huisong Zheng[1], Yinlong Xie[1,7], Julien Tap[6], Patricia Lepage[6], Marcelo Bertalan[9], Jean-Michel Batto[6], Torben Hansen[4], Denis Le Paslier[10], Allan Linneberg[11], H. Bjørn Nielsen[9], Eric Pelletier[10], Pierre Renault[6], Thomas Sicheritz-Ponten[9], Keith Turner[12], Hongmei Zhu[1], Chang Yu[1], Shengting Li[1], Min Jian[1], Yan Zhou[1], Yingrui Li[1], Xiuqing Zhang[1], Songgang Li[1], Nan Qin[1], Huanming Yang[1], Jian Wang[1], Søren Brunak[9], Joel Doré[6], Francisco Guarner[5], Karsten Kristiansen[13], Oluf Pedersen[4,14], Julian Parkhill[12], Jean Weissenbach[10], MetaHIT Consortium†, Peer Bork[2], S. Dusko Ehrlich[6] & Jun Wang[1,13]

To understand the impact of gut microbes on human health and well-being it is crucial to assess their genetic potential. Here we describe the Illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from faecal samples of 124 European individuals. The gene set, ~150 times larger than the human gene complement, contains an overwhelming majority of the prevalent (more frequent) microbial genes of the cohort and probably includes a large proportion of the prevalent human intestinal microbial genes. The genes are largely shared among individuals of the cohort. Over 99% of the genes are bacterial, indicating that the entire cohort harbours between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species, which are also largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in terms of functions present in all individuals and most bacteria, respectively.

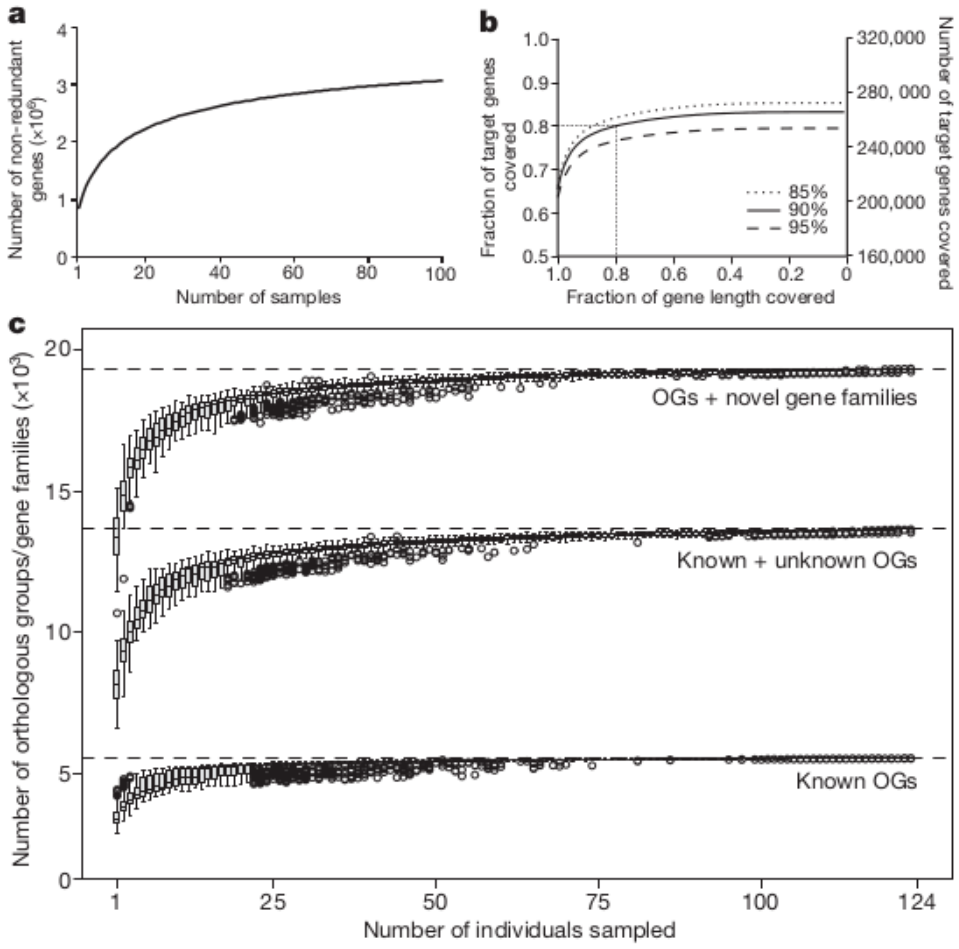


Figure 2 | Predicted ORFs in the human gut microbiome. **a**, Number of unique genes as a function of the extent of sequencing. The gene accumulation curve corresponds to the S_{obs} (Mao Tau) values (number of observed genes), calculated using EstimateS[21] (version 8.2.0) on randomly chosen 100 samples (due to memory limitation). **b**, Coverage of genes from 89 frequent gut microbial species (Supplementary Table 12). **c**, Number of functions captured by number of samples investigated, based on known (well characterized) orthologous groups (OGs; bottom), known plus unknown orthologous groups (including, for example, putative, predicted, conserved hypothetical functions; middle) and orthologous groups plus novel gene families (>20 proteins) recovered from the metagenome (top). Boxes denote the interquartile range (IQR) between the first and third quartiles (25th and 75th percentiles, respectively) and the line inside denotes the median. Whiskers denote the lowest and highest values within 1.5 times IQR from the first and third quartiles, respectively. Circles denote outliers beyond the whiskers.

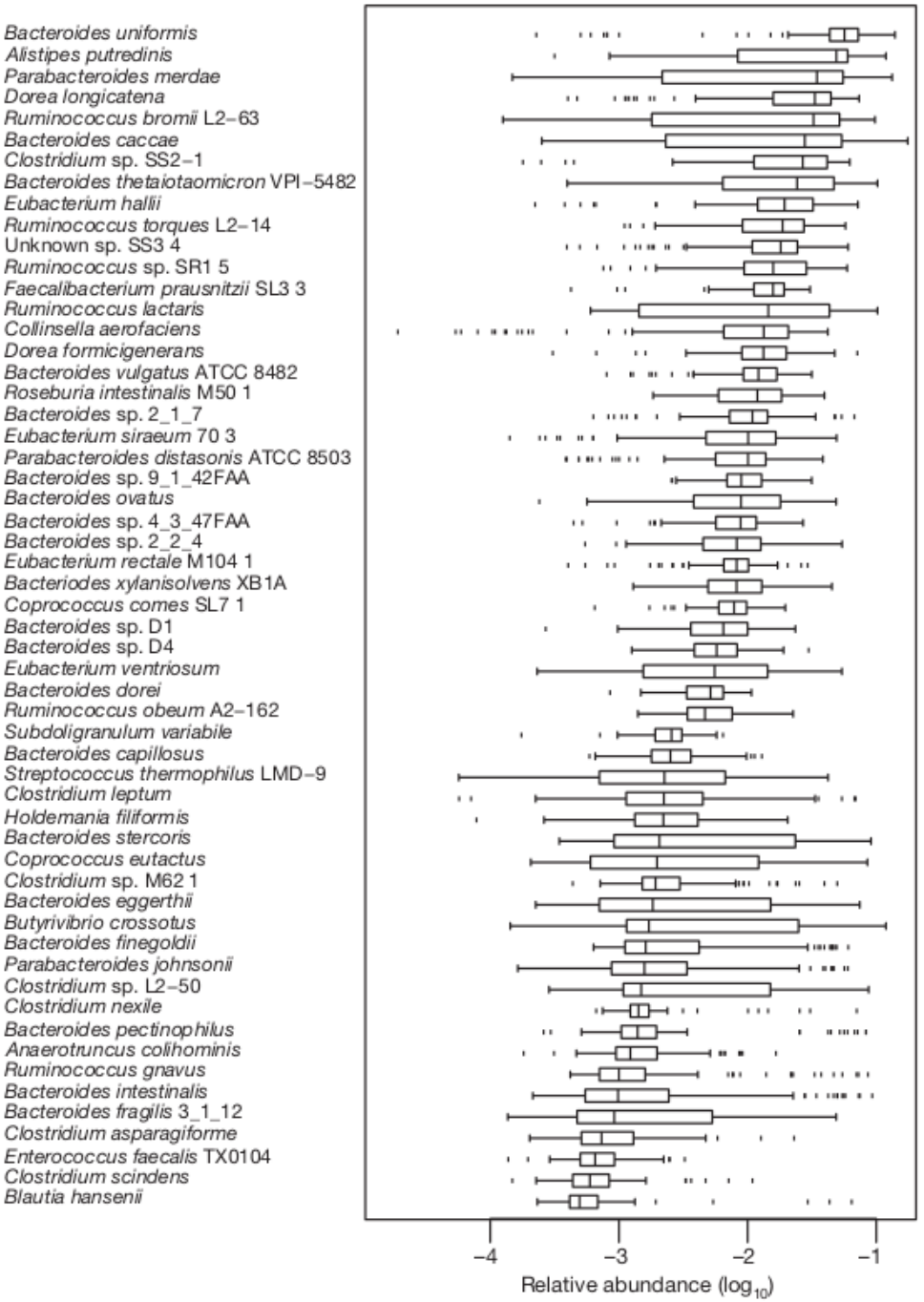


Figure 3 | Relative abundance of 57 frequent microbial genomes among individuals of the cohort. See Fig. 2c for definition of box and whisker plot. See Methods for computation.

Clostridium difficile

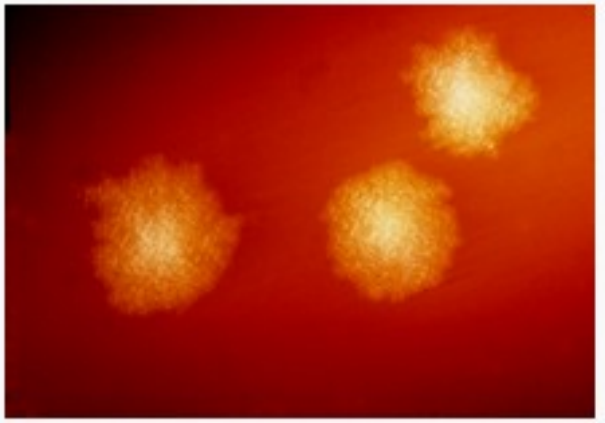
From Wikipedia, the free encyclopedia

Clostridium difficile (pronunciation below) (from the Greek *kloster* (κλωστήρ), 'spindle',[*citation needed*] and Latin *difficile*, 'difficult, obstinate'),[1] also known as "CDF/cdf", or "*C. diff*", is a species of Gram-positive bacteria of the genus *Clostridium* that causes severe diarrhea and other intestinal disease when competing bacteria in the gut flora have been wiped out by antibiotics.

Clostridia are anaerobic, spore-forming rods (bacilli).[2] *C. difficile* is the most serious cause of antibiotic-associated diarrhea (AAD) and can lead to pseudomembranous colitis, a severe inflammation of the colon, often resulting from eradication of the normal gut flora by antibiotics.[3]

In a very small percentage of the adult population, *C. difficile* bacteria naturally reside in the gut. Other people accidentally ingest spores of the bacteria while they are patients in a hospital (where 14,000 people a year in America alone die as a result),[4] nursing home, or similar facility. When the bacteria are in a colon in which the normal gut flora has been destroyed (usually after a broad-spectrum antibiotic such as clindamycin has been used), the gut becomes overrun with *C. difficile*. This overpopulation is harmful because the bacteria release toxins that can cause bloating and diarrhea, with abdominal pain, which may become severe. *C. difficile* infections are the most common cause of pseudomembranous colitis, and in rare cases this can progress to toxic megacolon, which can be life-threatening.

Clostridium difficile



C. difficile colonies on a blood agar plate.

Micrograph of *Clostridium difficile*

Scientific classification

Kingdom: Bacteria

Clin Gastroenterol Hepatol. Author manuscript; available in PMC 2012 December 1.
PMCID: PMC3223289
Published in final edited form as:
NIHMSID: NIHMS320727
Clin Gastroenterol Hepatol. 2011 December; 9(12): 1044–1049.
Published online 2011 August 24. doi: 10.1016/j.cgh.2011.08.014

Treating Clostridium difficile Infection with Fecal Microbiota Transplantation

Johan S. Bakken, MD, PhD, Thomas Borody, MD, PhD, Lawrence J. Brandt, MD, Joel V. Brill, MD, Daniel C. Demarco, MD, Marc Alaric Franzos, MD, MPH, Colleen Kelly, MD, Alexander Khoruts, MD, Thomas Louie, MD, Lawrence P. Martinelli, MD, Thomas A. Moore, MD, George Russell, MD, MS, and Christina Surawicz, MD

Author information ► Copyright and License information ►

Abstract

Go to:

Clostridium difficile infection is increasing in incidence, severity, and mortality. Treatment options are limited and appear to be losing efficacy. Recurrent disease is especially challenging; extended treatment with oral vancomycin is becoming increasingly common but is expensive. Fecal microbiota transplantation (FMT) is safe, inexpensive, and effective; according to case and small series reports, about 90% of patients are cured. We discuss the rationale, methods, and use of FMT.

Keywords: *Clostridium difficile*, transplantation, microbiota, fecal enema, recurrent infection, diarrhea

Terragenome

International Soil Metagenome Sequencing Consortium

HOME

Leonardo da Vinci wrote, “We know more about the movement of celestial bodies than about the soil underfoot.” This statement is just as true today, 500 years after it was written, and is particularly applicable to the microorganisms that inhabit the soil as the identities and roles of most of these microorganisms remain a mystery; even the relevance of microbial diversity to the functioning of soils remains obscure. The complete sequencing of a soil metagenome (i.e., the genomes of all microorganisms inhabiting the soil environment) is now an achievable objective that requires a strong international and interdisciplinary collaboration. The purpose of the TerraGenome network is to facilitate activities that will increase our knowledge and understanding of the soil metagenome.

Announcements

- Analysis Tools
- Links
- Project Directory
- Workshops

Contact

For Terragenome info, contact David Myrold

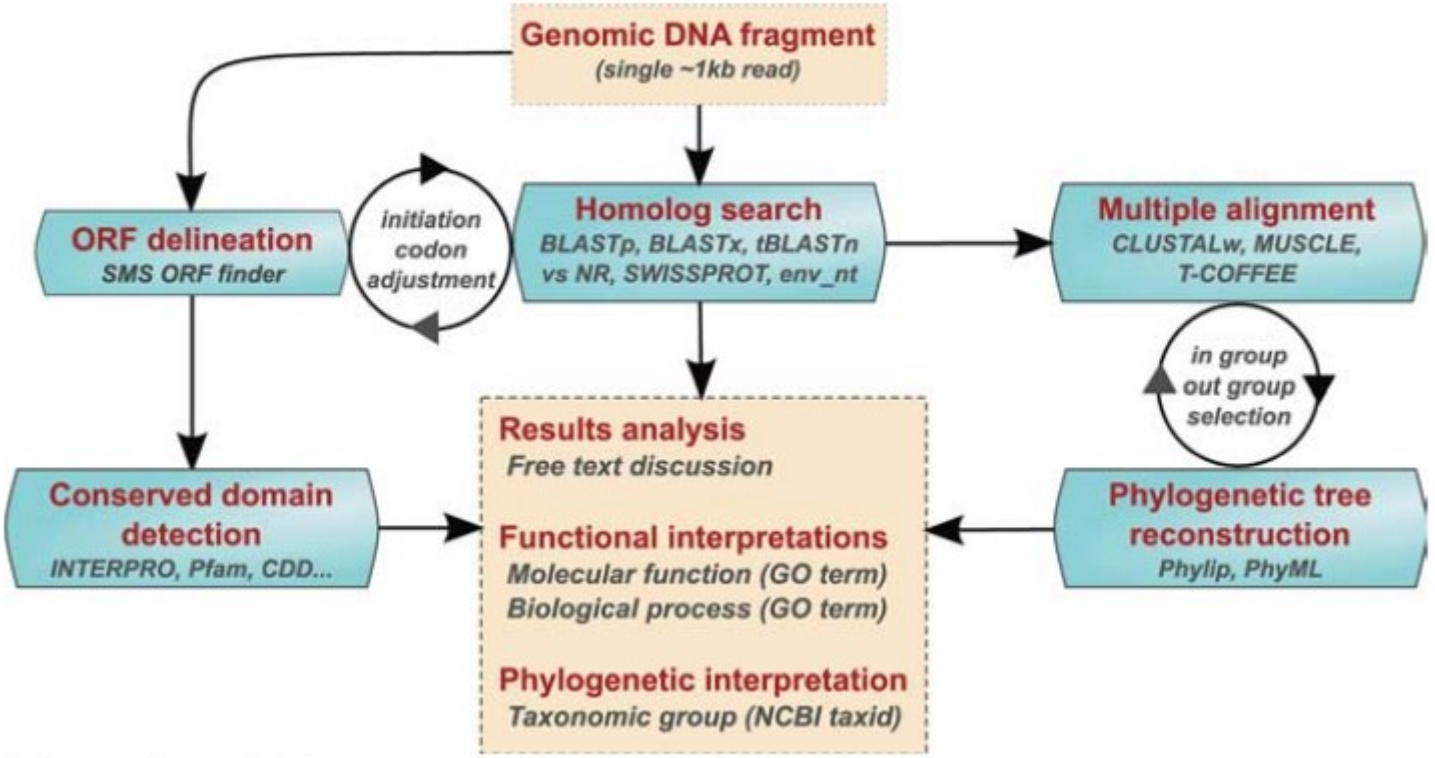
OPEN ACCESS Freely available online

PLoS BIOLOGY

Community Page

Metagenome Annotation Using a Distributed Grid of Undergraduate Students

Pascal Hingamp*, Céline Brochier, Emmanuel Talla, Daniel Gautheret, Denis Thieffry, Carl Herrmann



doi:10.1371/journal.pbio.0060296.g001