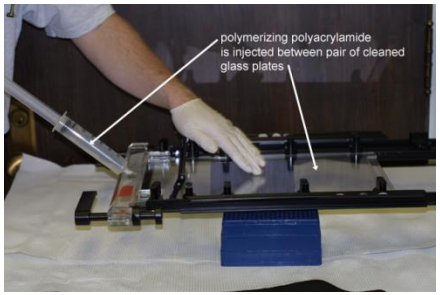
Metagenomics versus Next Generation Sequencing Technologies

Doug Rusch
Environmental Genomics Group, J. Craig Venter Institute



JCVI Rockville Campus

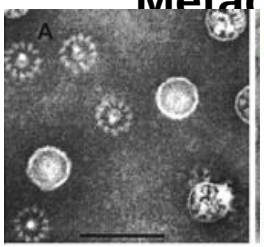
J. Craig Venter
INSTITUTE

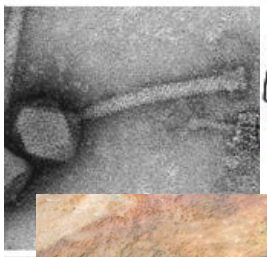
Metagenomics at the JCVI

Metac

and cann

Ecogenomics or



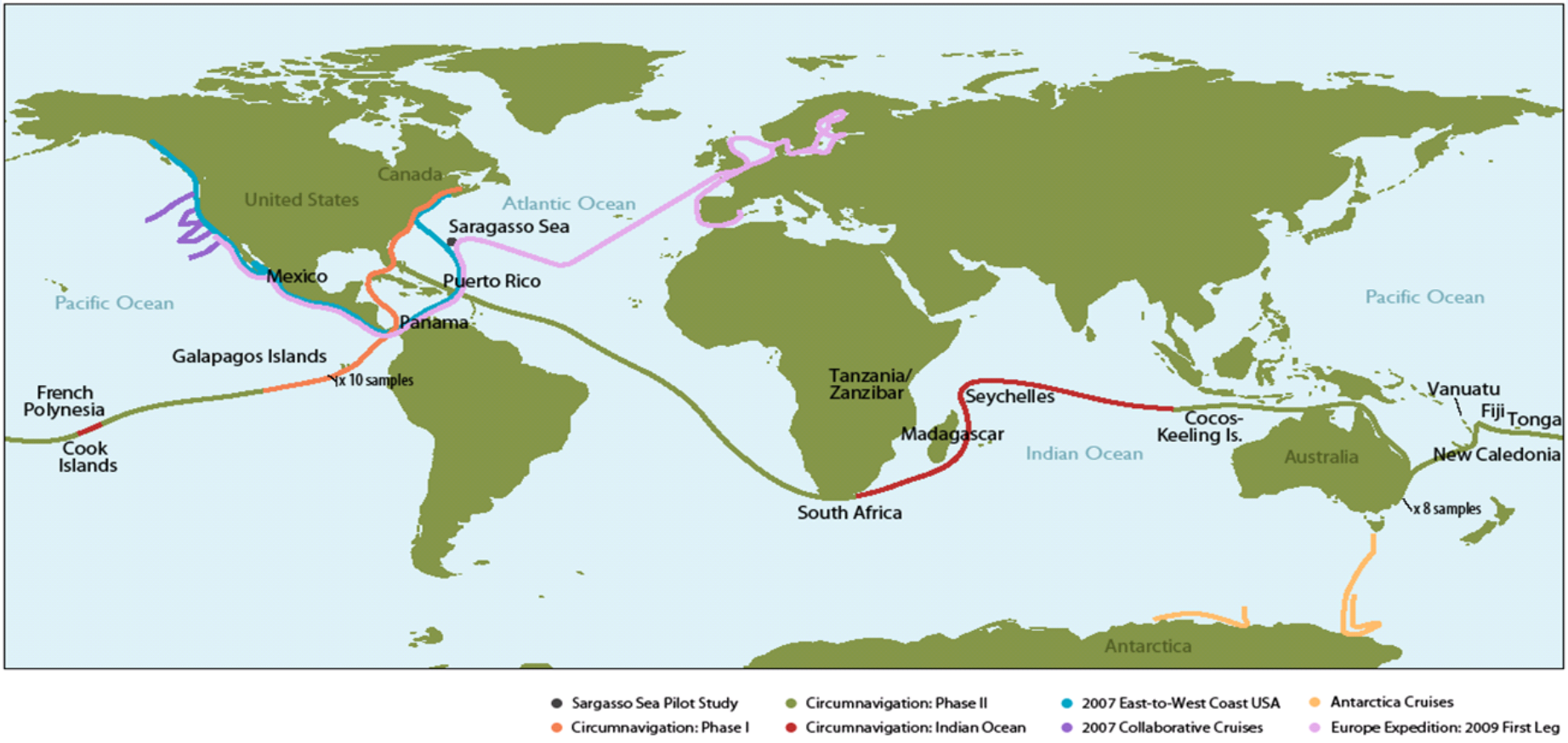


J. Craig Venter
INSTITUTE

Goal of Metagenomic Projects

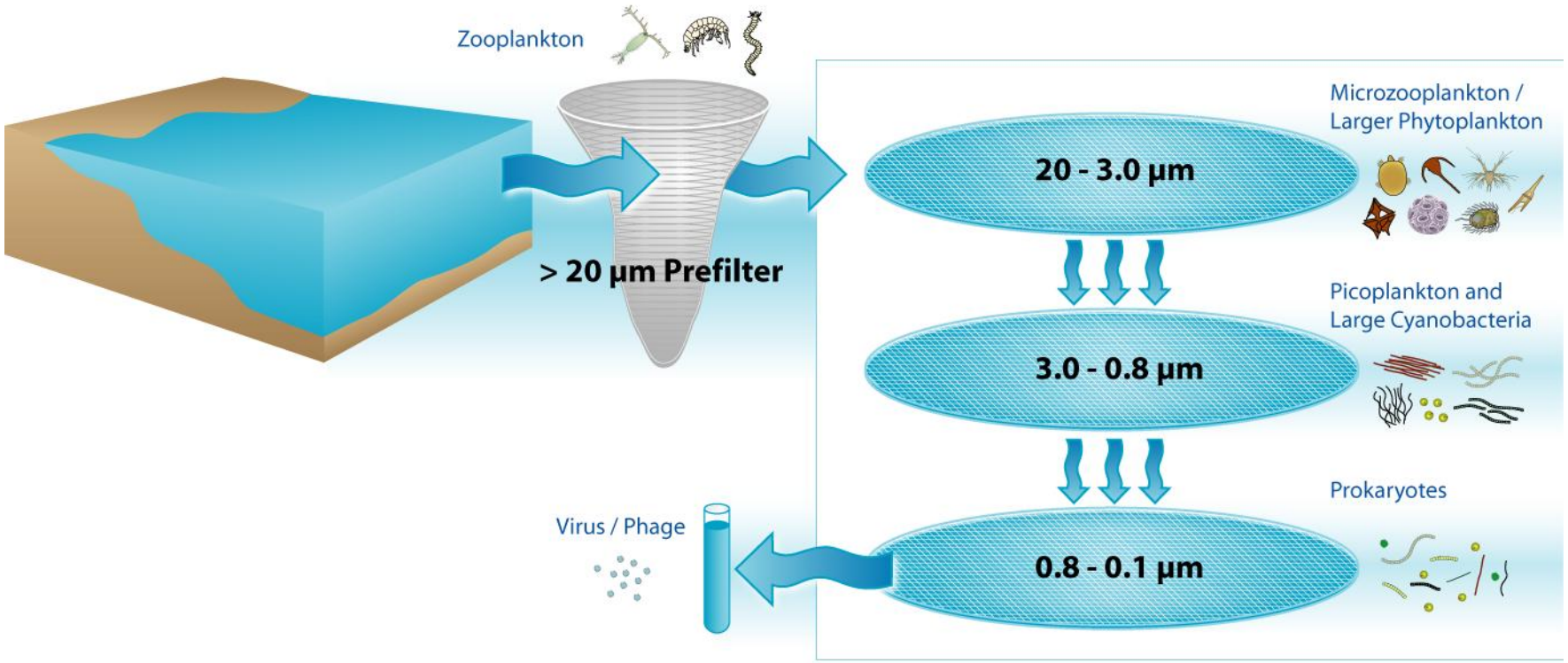
1. Gene Discovery
2. Genome Isolation
3. Variation Detection
4. Association of Variation with environment through comparison
5. Build better evolutionary models
6. Understanding of how complex communities form, are maintained, and respond to change

The GOS Project

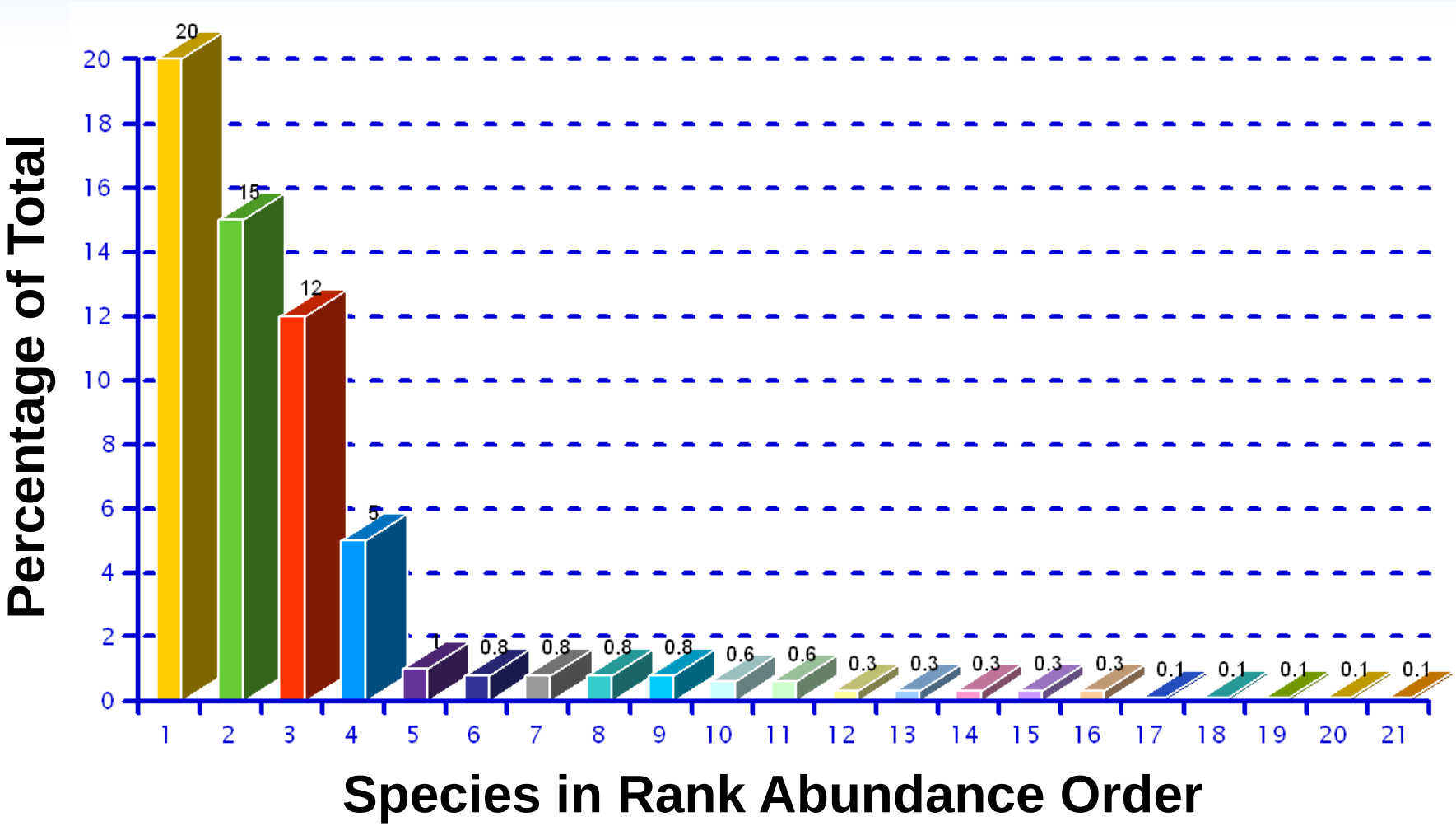


* *Sorcerer II* World Circumnavigation: 2003-2006
* *Sorcerer II* North America, and West Coast collaborative cruises: 2007-2008
* Antarctica cruises: 2007-2009
* *Sorcerer II* Europe: 2009-

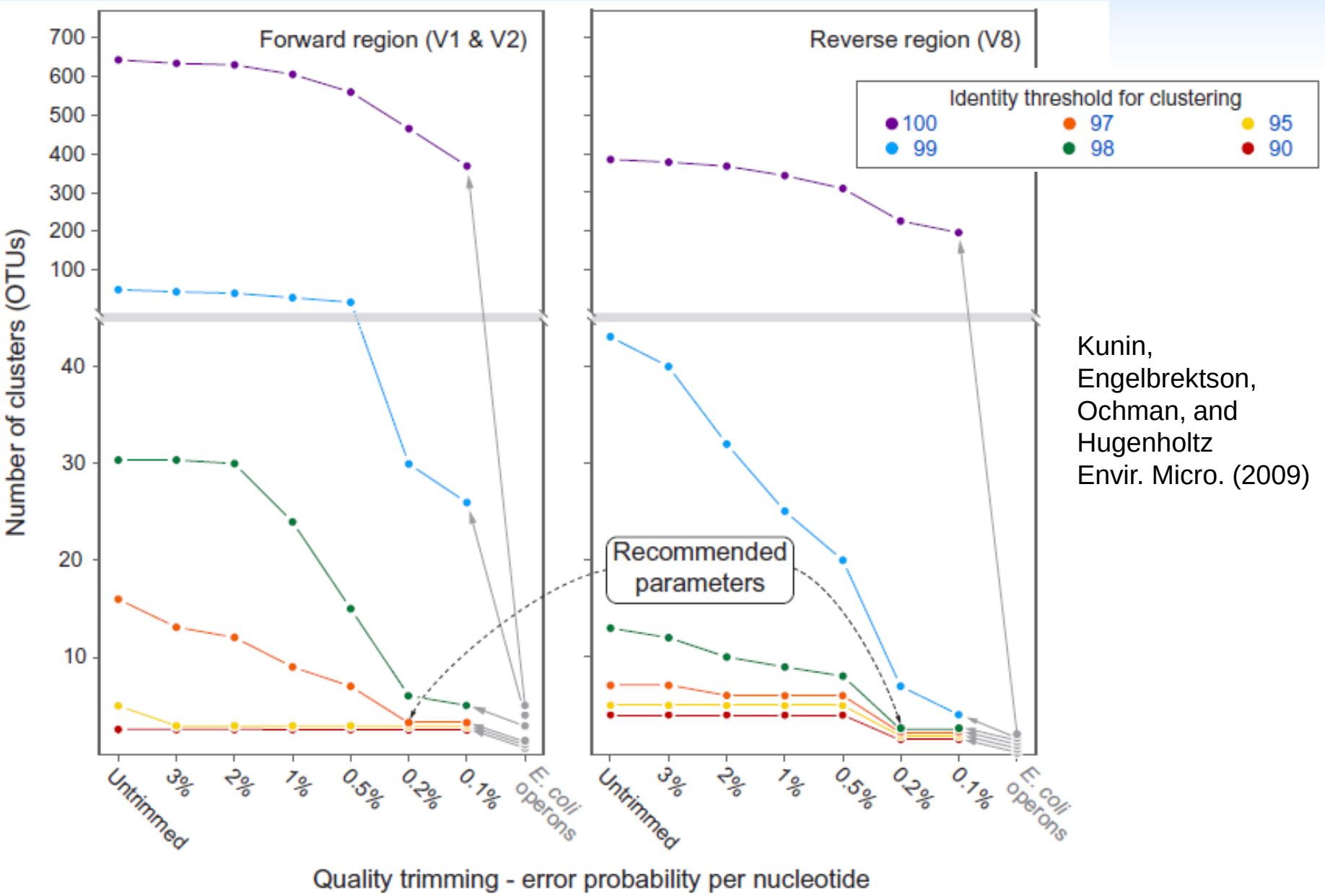
Sample Filtration



Dynamic Range of Organisms



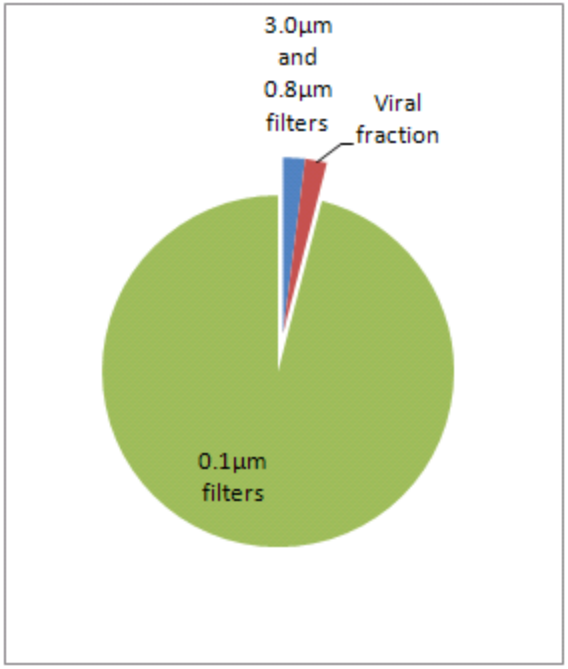
Diversity Estimates



Kunin, Engelbrektson, Ochman, and Hugenholtz Envir. Micro. (2009)

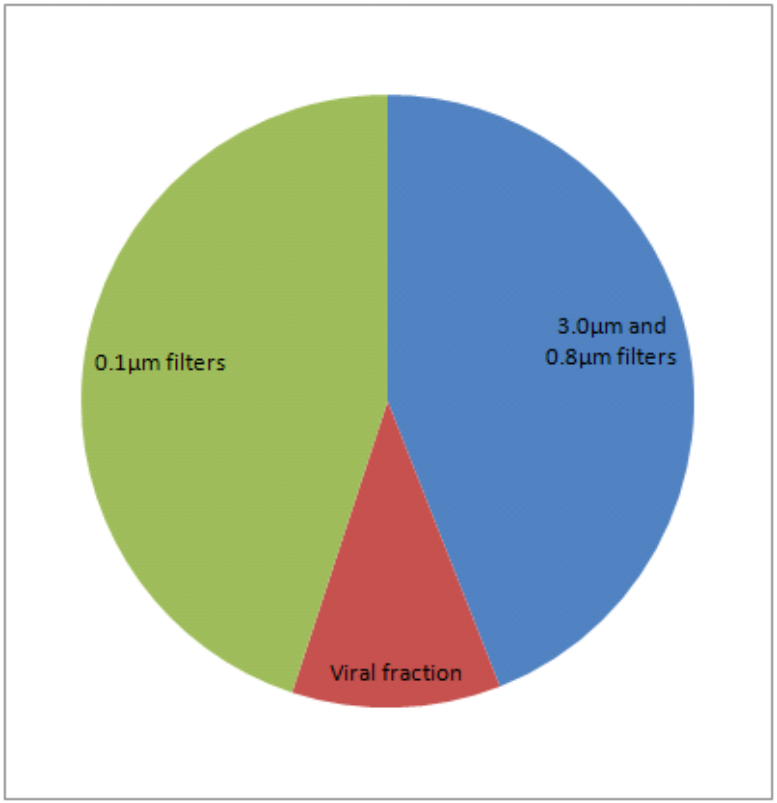
GOS Sequence Data

GOS Sanger data generated 2003-2008:
22 Million Sanger reads at ~750 bp/read.



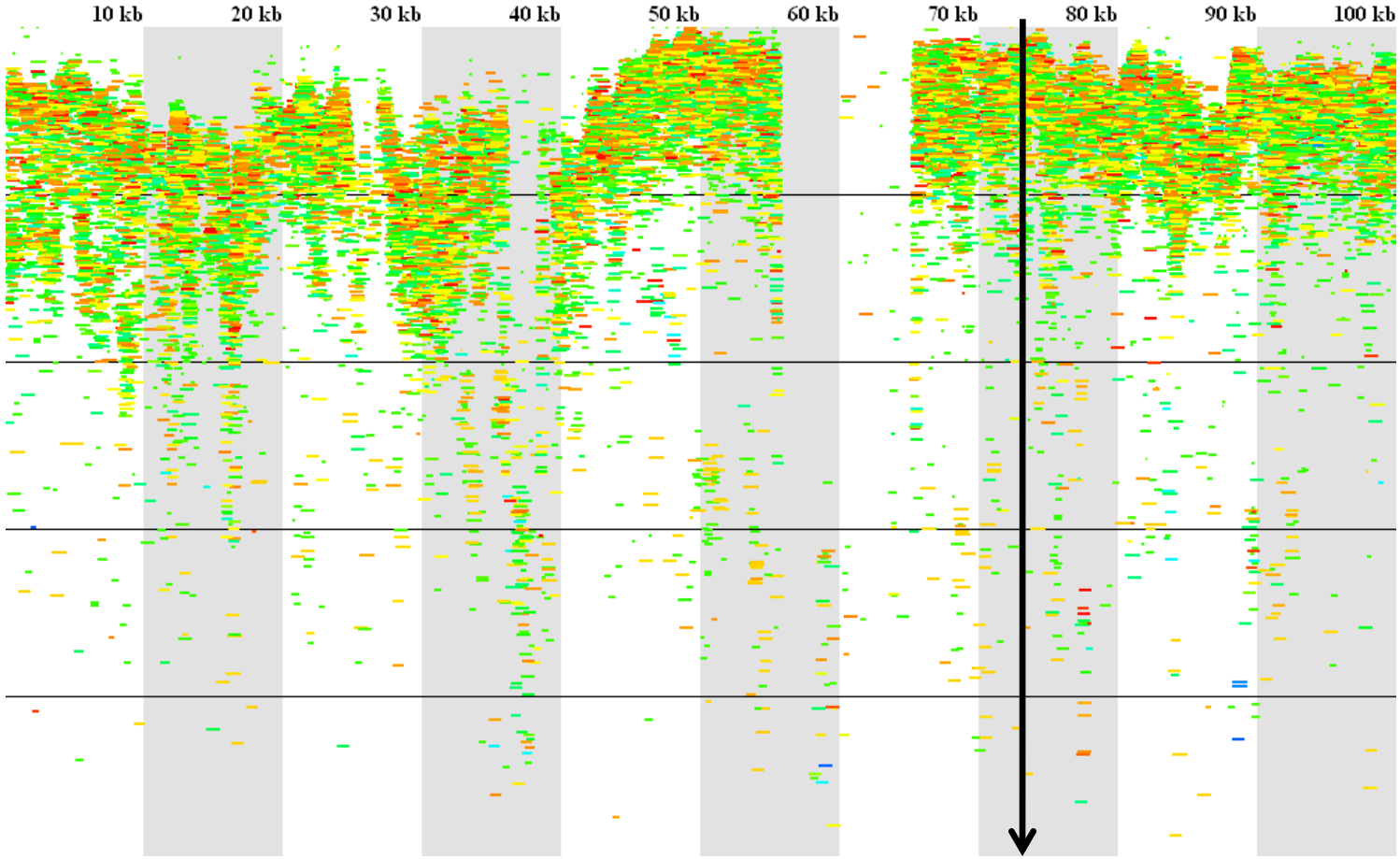
2003-2008

GOS 454-Titanium, projection for 2009:
~45 Million reads at ~400 bp/read.



2009

Abundant Diversity



~300 fold depth of coverage

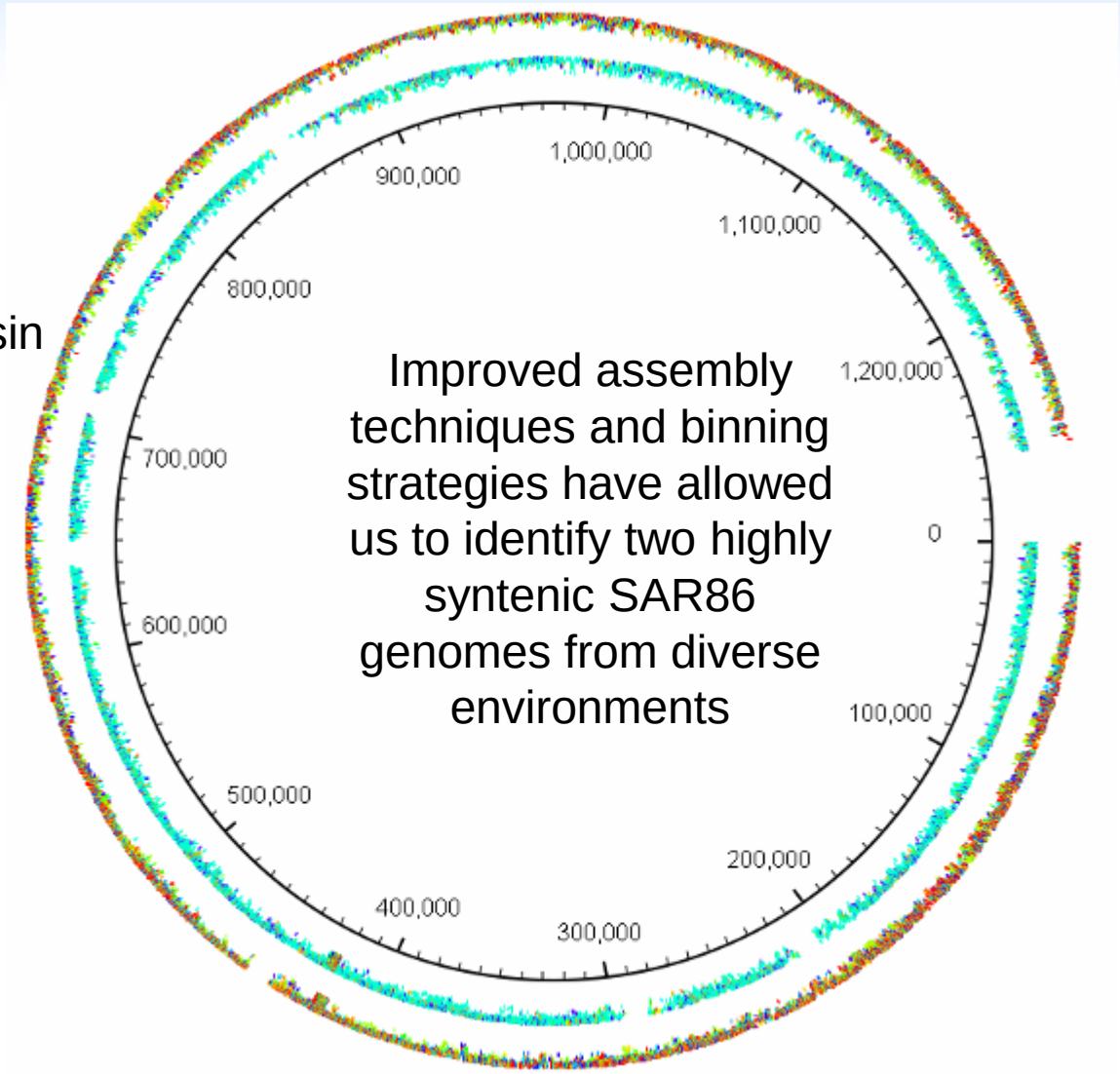
Genome Assembly

SAR86 Background:

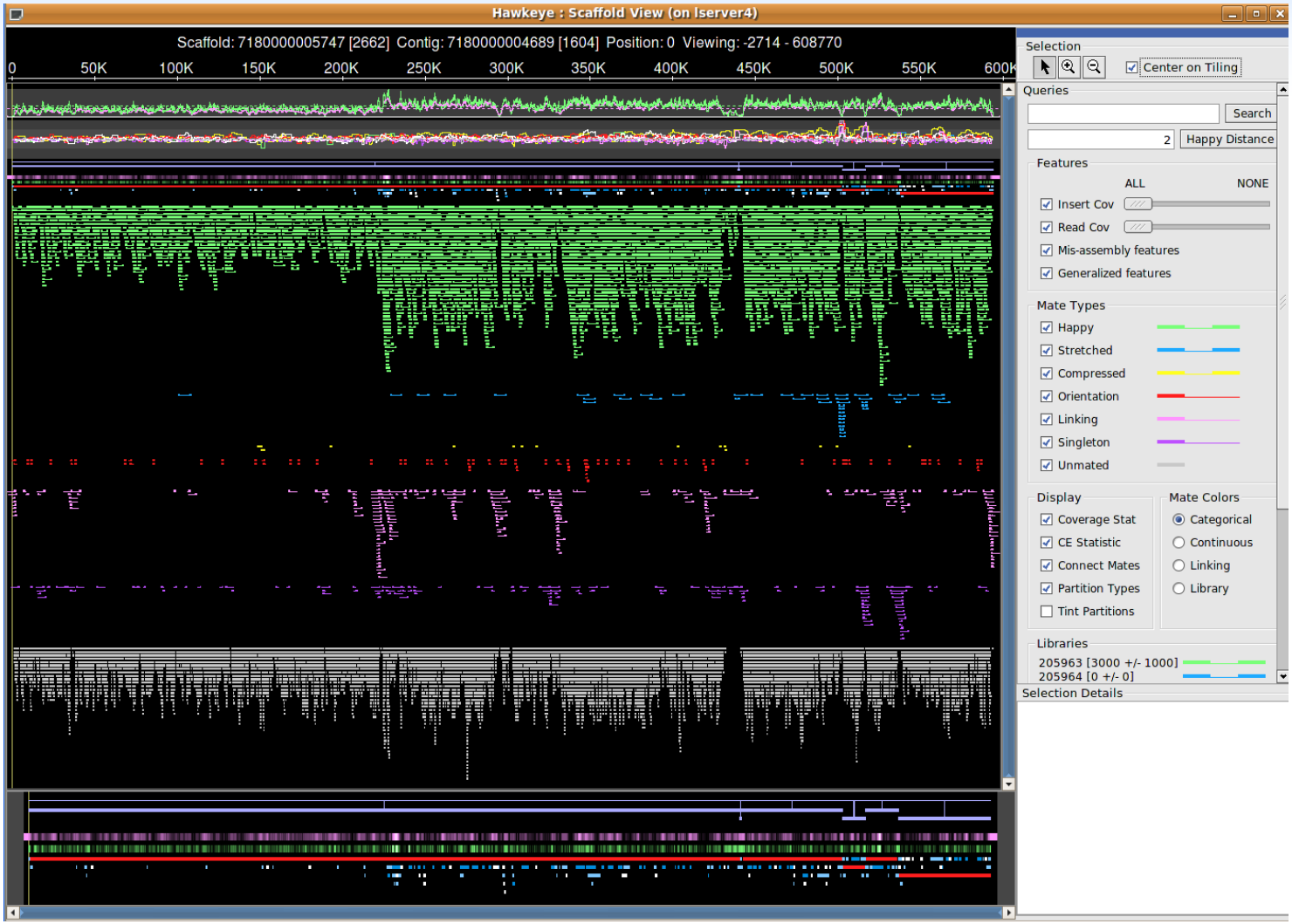
- Highly abundant
- Widely dispersed
- Uncultivated
- Source of first proteorhodopsin gene

<u>Improvements to Celera Assembler:</u>

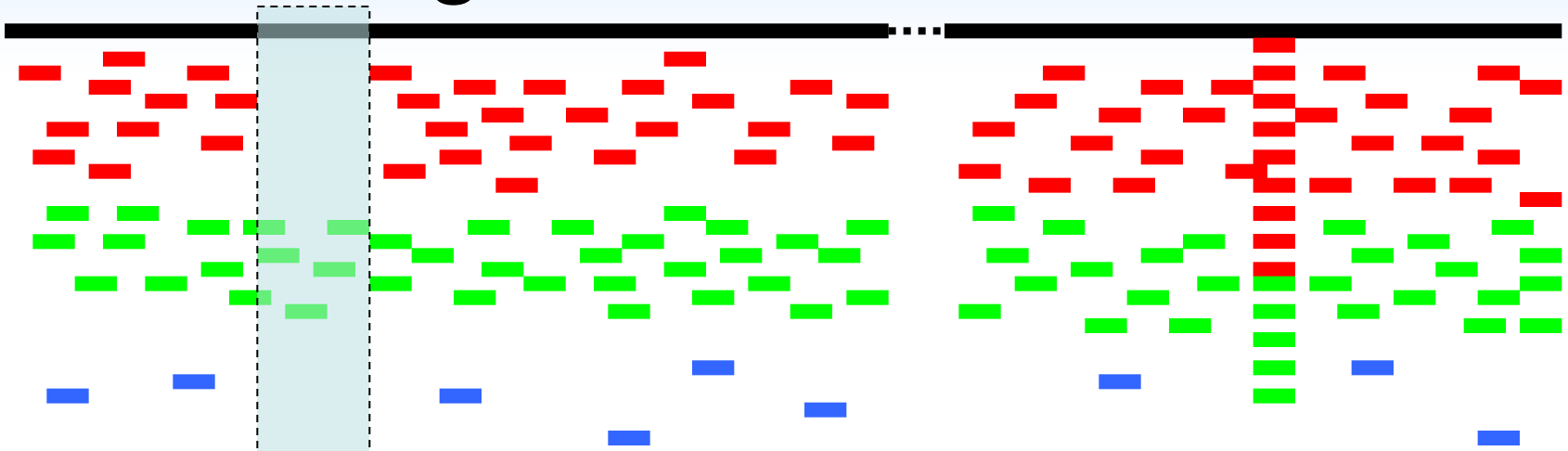
Supports hybrid assemblies of Sanger, 454, and Solexa reads (>=60 bps)



Chimeric Assembly



Identifying Region Specific Segments On Scaffolds



KEY:

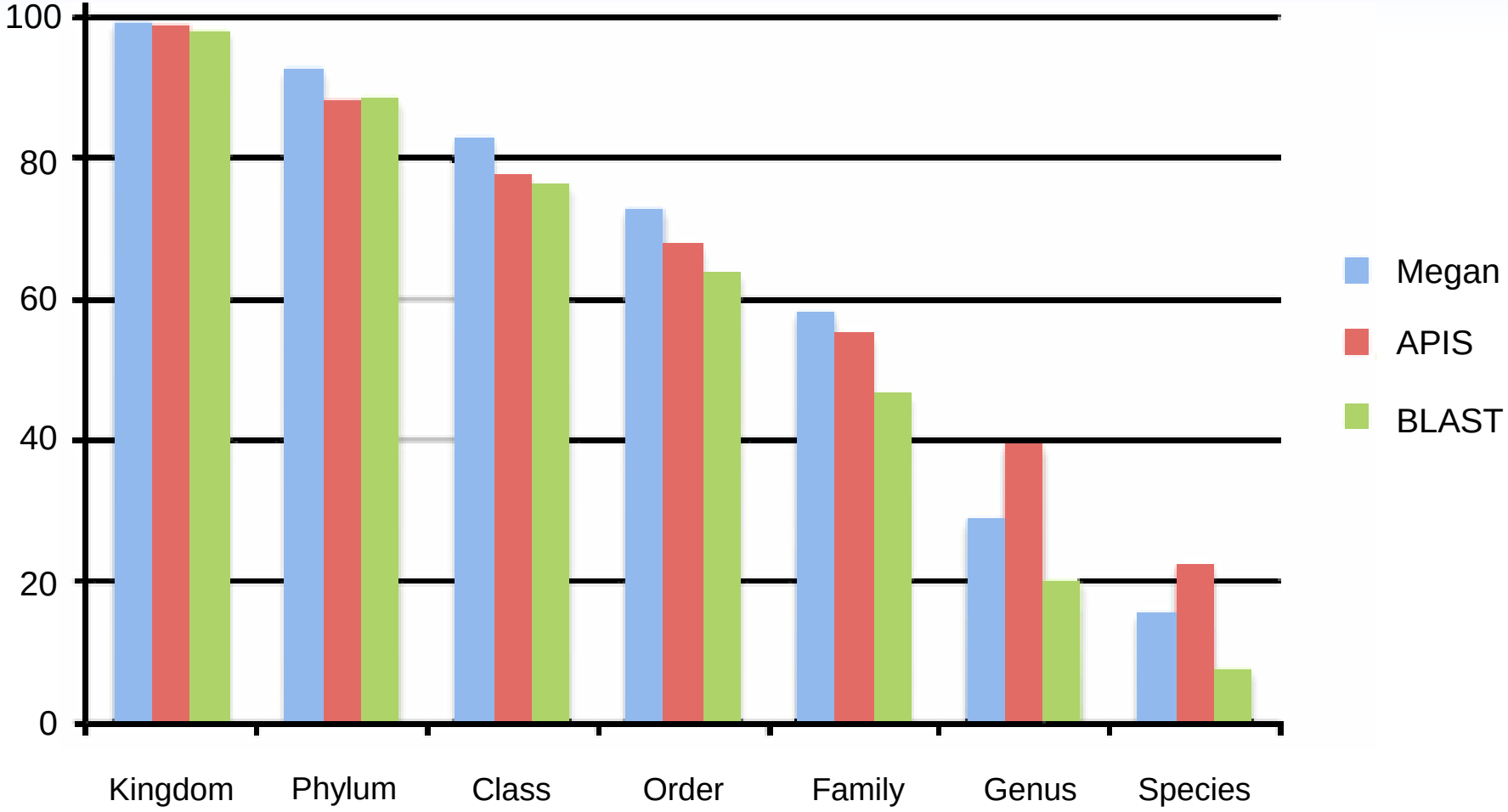
Sample A

Sample B

Sample C

Red reads abundant but missing from highlighted region. Green reads are uniformly present can cover region missing in red and allow assembly of entire scaffold to proceed. Poisson statistics provides probability that highlighted region would does not contain any red reads and can be used to rapidly and automatically identify other regions given a probability cutoff. Blue reads are present at too low a density to produce a statistically significant signal on their own.

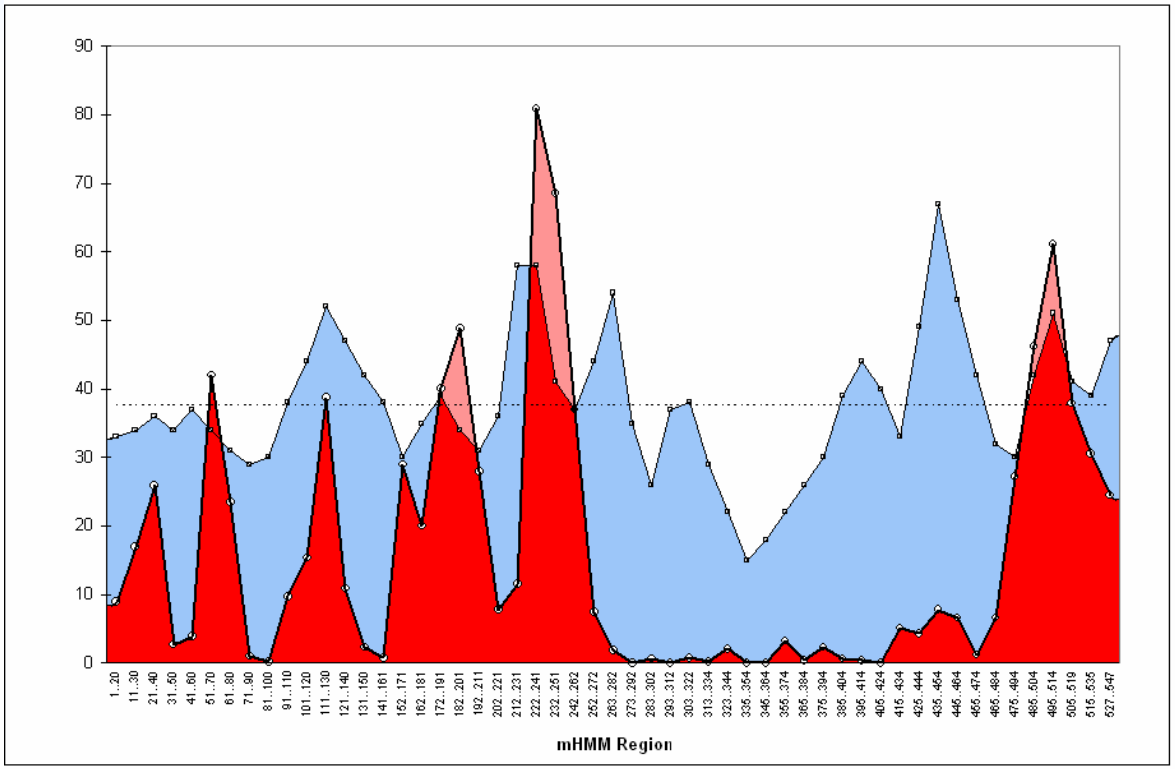
A Phylogenetic Inference System (APIS)



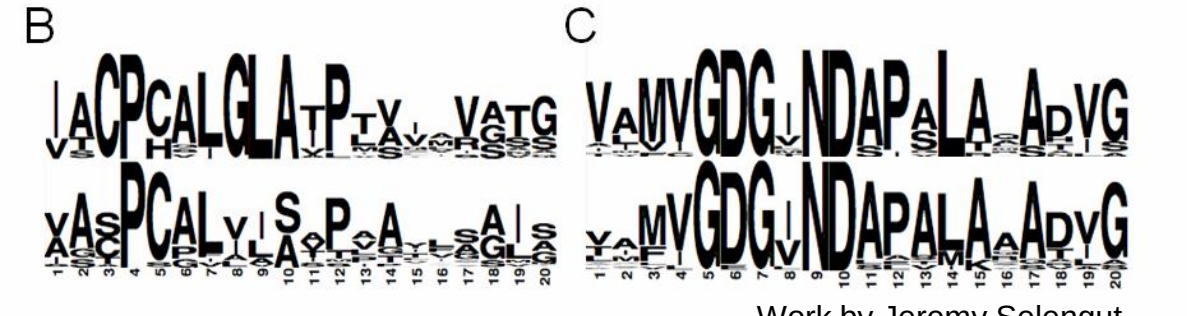
J. Craig Venter INSTITUTE

APIS is being developed by Jonathan Badger

TIGRFAMs to Mini-HMMs



Conservation as well as distinction from closely homologous outgroups determine mHMM performance.



Work by Jeremy Selengut

Conclusions

1. Sequencing is going great – more data than we can easily analyze or interpret
2. Making headway towards isolating references directly from a complex metagenome
3. Have processes in place to identify variation between samples; this is much improved by having next generation sequencing
4. Improve techniques for annotating “partial” genes produced by short next generation sequencers
5. Personal Bacterial Genomics: Would love to have single cell techniques to make sense of all the diversity

Acknowledgments

Jonathan Badger
Jeremy Selengut
Dan Haft
Matt Lewis
Saul Kravitz
Laura Sheahan
Scott Peterson
Karen Nelson
Barbara Methe
Jason Miller
David Tanenbaum

Sean Murphy
Johannes Goll
Shibu Yooseph
Bob Friedman
Craig Venter
Granger Sutton
Sergey Koren
Todd Safford

JCVI Assembly Group

The captain and crew of the Sorcerer II

454, Solexa and SOLiD sequencing teams at JCVI

J. Craig Venter INSTITUTE