# Exploring the Microbial World with Next Generation Sequencing (NGS)

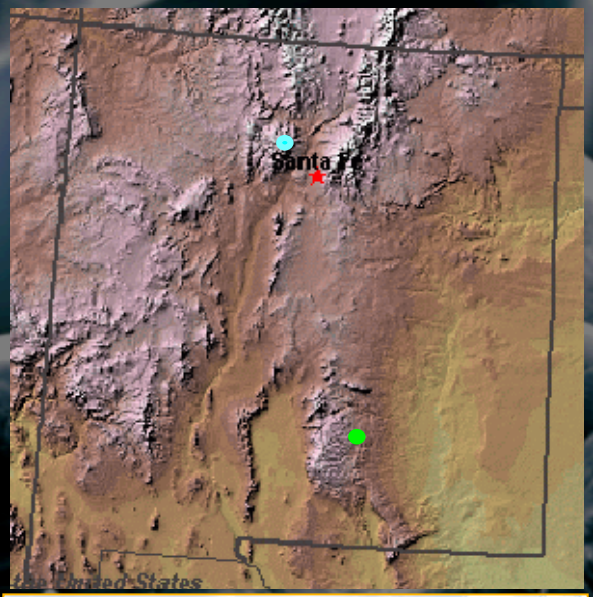
Patrick Chain (pchain@lanl.gov)
Los Alamos National Laboratory (LANL)
Joint Genome Institute (JGI)
Seoul, March 9th, 2012

# Los Alamos National Laboratory (LANL)

**Vision:** ***"We serve the nation by applying the best science and technology to make the world a better and safer place."***

## Infrastructure:

- 38 Square Miles
- Over 2,000 buildings with approx. 9 million sq. ft.
- 100 Miles of paved roads
- 30 Miles of 115 kV transmission lines
- 120 Miles of gas transmission lines
- 14 Nuclear facilities



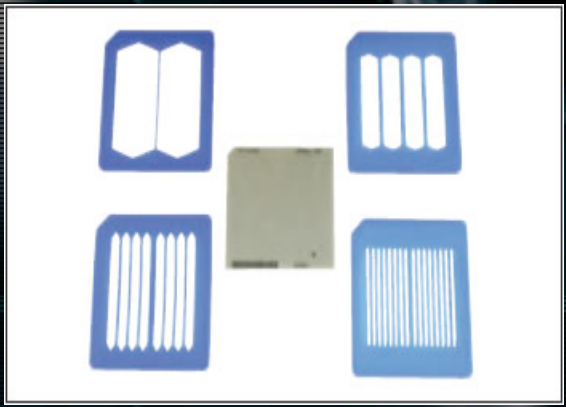

## Mission:

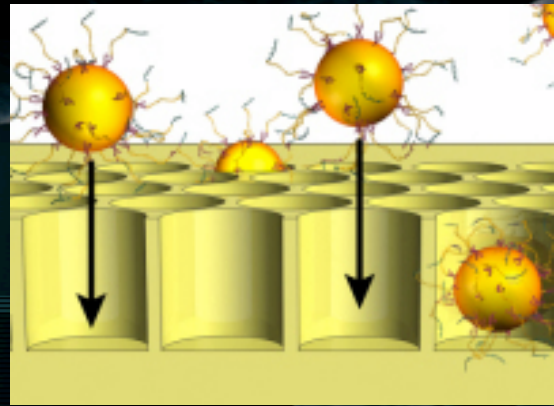
Ensure the safety and reliability of the U.S. Nuclear deterrent.

Reduce the global threat of weapons of mass destruction.

Solve national problems in energy, environment, and health security.

# Roche 454 - Overview



## Throughput

- 400 Mb
(~8 hrs runtime)

## Read-length

- up to 400 bp
- Paired ends libraries – 8kb
- 700-1000bp with upgrade

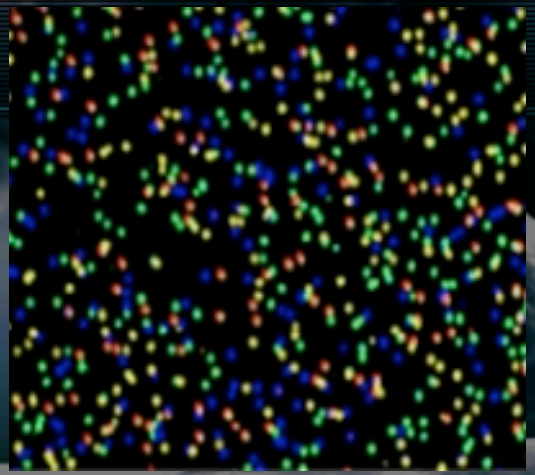
## Costs per Megabase

~$15-$25

## Applications

- *De Novo* Sequencing, Resequencing,
- Transcriptome Analysis,
- Metagenomics & Microbial Diversity

# Illumina - Overview



## Throughput

- Up to 600 Gb per run
(3 – 14 days runtime)

## Read-length

- 100 bp
- >120 bp demonstrated
- paired ends libraries – 200bp
- paired end libraries – 8 kb R&D

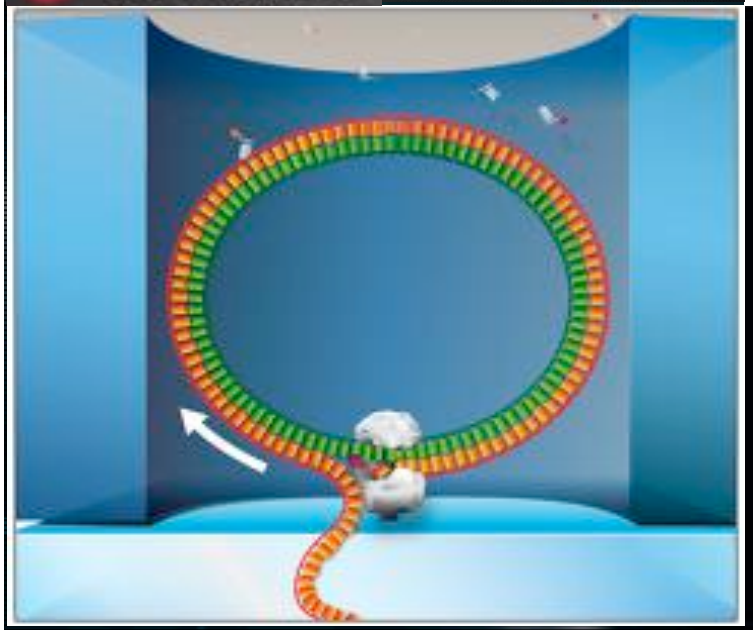
## Costs per MegaBase

~ $0.05-$0.75

## Applications

- Resequencing, SNP Discovery
- *De Novo* , Transcriptome
- Methylation Patterns

# PacBio SMRT - Overview


## Throughput

- Up to 60 Mb per chip
(30 - 45 minutes runtime)
- human genome in 30 min in FY12

## Read-length

- 1500 bp…..10000 bp
- Standard – long read
- Consensus – proof reading
- Strobe – multiple reads

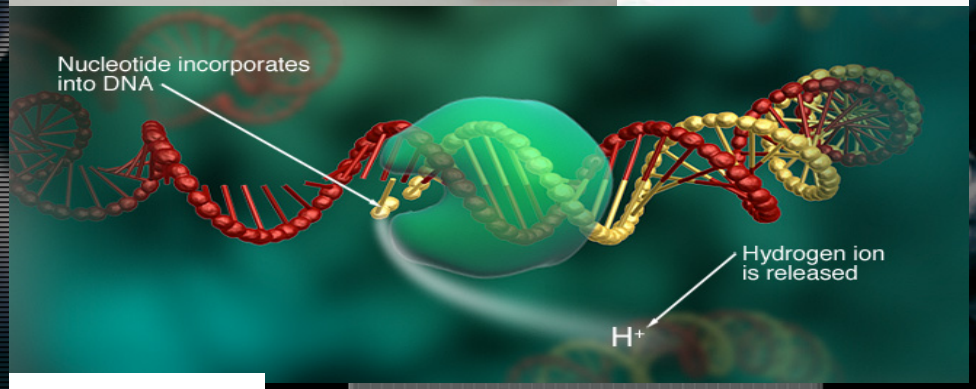
## Costs per MegaBase

$2.5-$5.0

## Applications

- *De Novo* Sequencing, Resequencing,
- Transcriptome Analysis,
- Metagenomics & Microbial Diversity

# Ion Torrent- Overview





ion torrent by life technologies

## Throughput

- Up to 100 Mb per run (2 hr run time)

## Read-length

- 200 bp

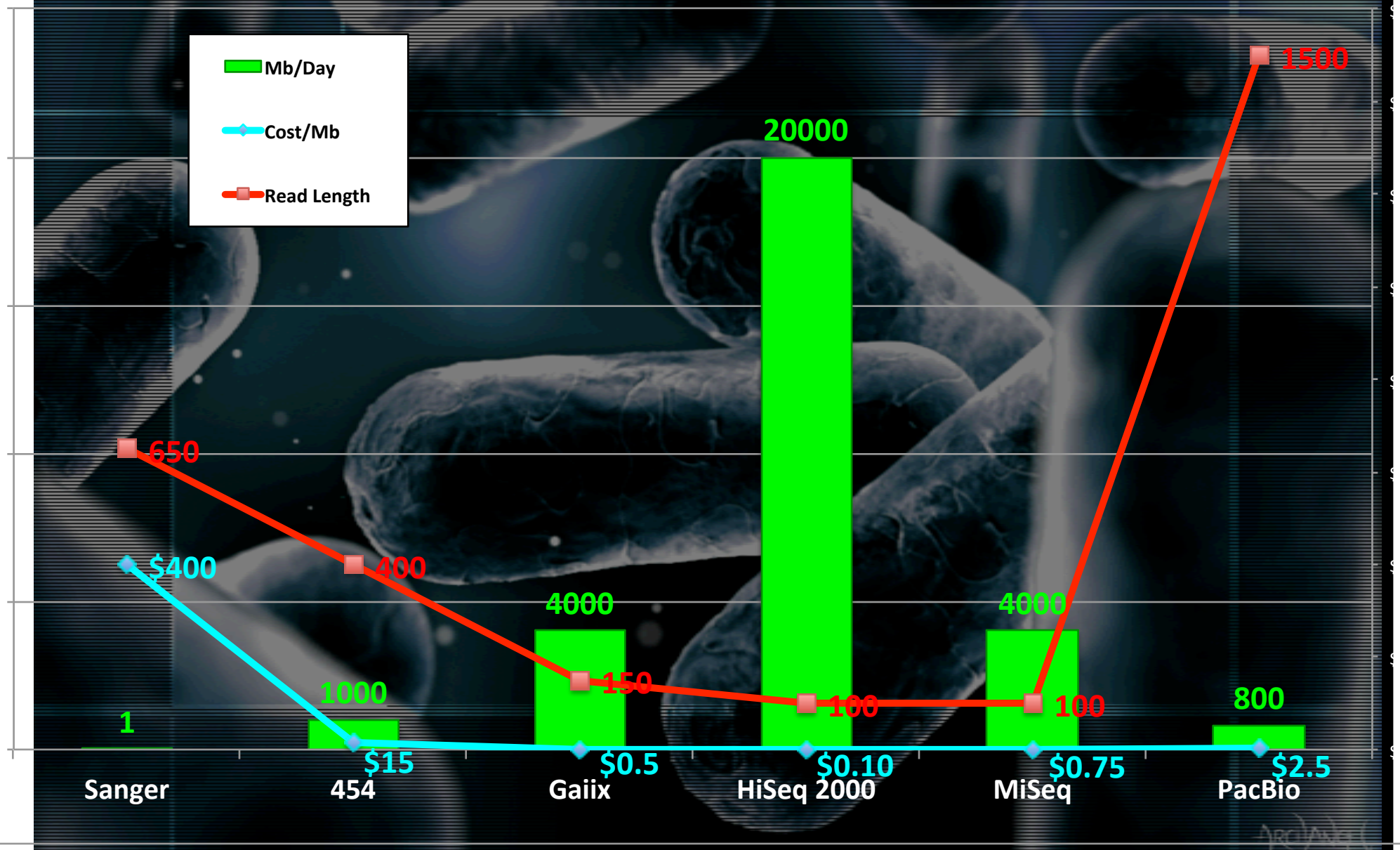
## Costs per MegaBase

?$5-$20?

## Applications

- *De Novo* Sequencing
- Resequencing
- SNP Discovery

# Technology Comparison

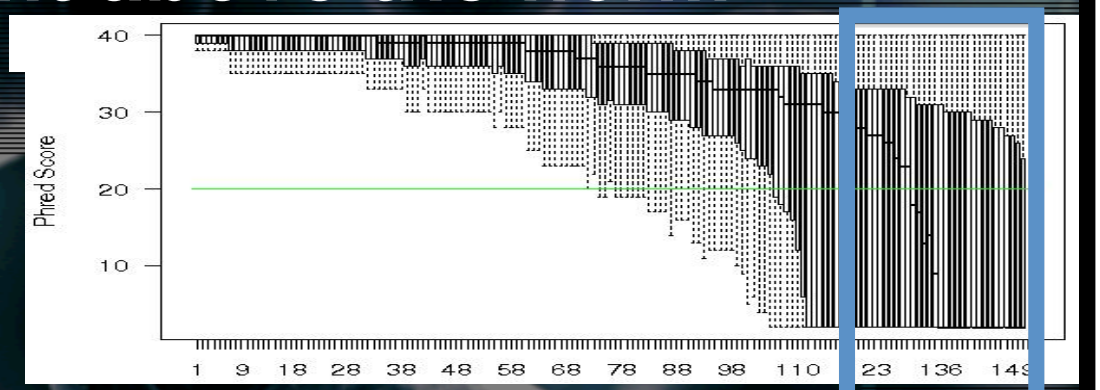



Slide courtesy of F. Chen

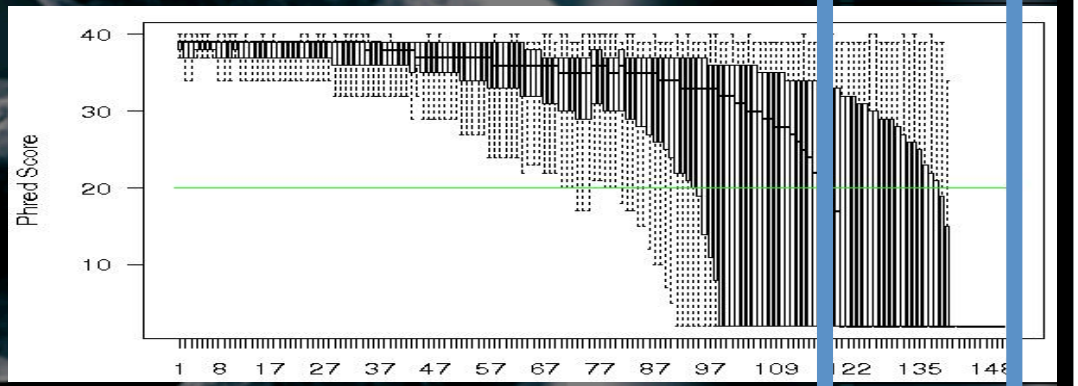
# Early access and partnering = improvement above the norm

Quality Scores

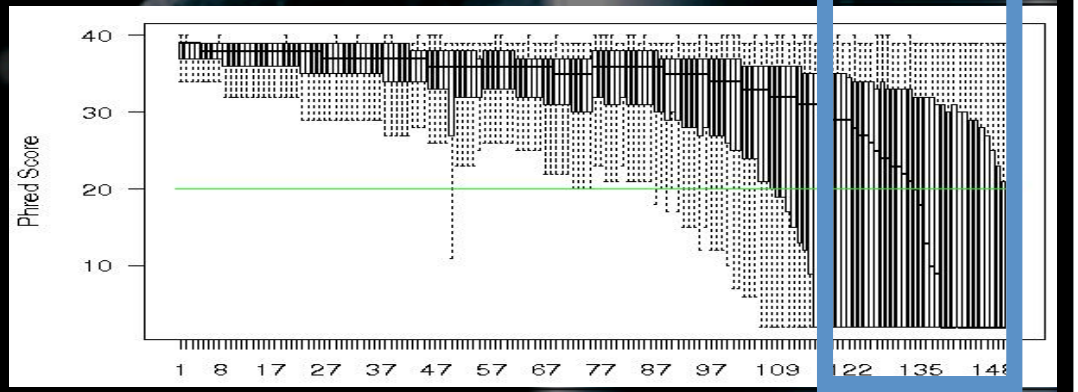
**GAII 2x150**
**(Q20 to ~130 bp)**



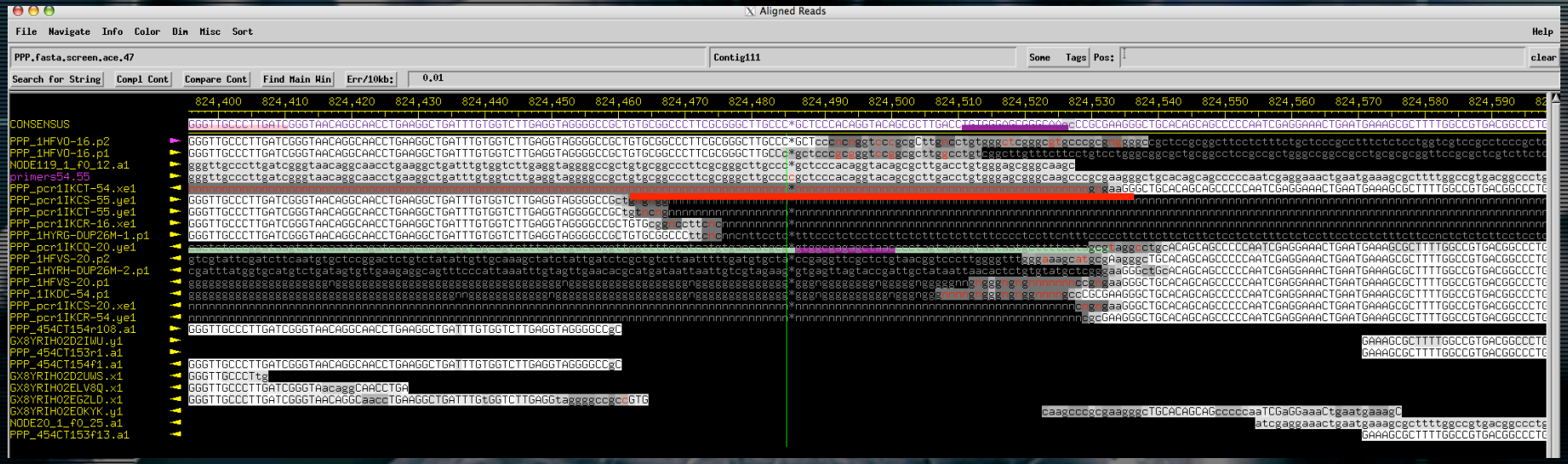

**HiSeq v2 2x150**
**(Q20 to ~113 bp)**




**HiSeq v3 2x150**
**(Q20 to ~133 bp)**



# Hard stops are no problem with PacBio

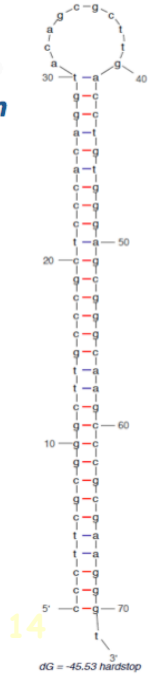



cccttcgcgggcttgcccgctcccacaggtacagcgcttgacctgtgggagcgggcaagcccgcgaaggg

30bp long

30bp long

*hairpin loop*




***An example of a sequence that creates a hairpin loop in the secondary structure which has been very difficult to complete when sequencing a genome.***

**Shown in the sequences above,**
**only the PacBio read (underlined in red) goes through the hard stop region.**
**Illumina, 454, and Sanger (PCR) reads didn't.**

# Playbook for NGS projects

- I have a sample – now what?
  - Metadata – or the sequence means little…
  - DNA extraction methods (biases), quantity considerations for library (volume of sample), spatial structure considerations, etc.
  - Library prep and sequencing – what technology? Cost and pros vs cons…
  - How much do I sequence? Cost and ability to analyze…
  - What method with the chosen platform? (paired end, long inserts)

# Playbook for NGS projects (cont'd)

- I have data – now what?
  - Analysis plan? (ie. Experimental design)
    - E.g. Reads vs contig/gene-based?
  - Much depends on previous considerations…

  - What depth is needed to achieve goal?

  → who is there, what are they doing, how?

  → phylogenetic and functional diversity?

  →population genetics, selection, lateral gene transfer?

- Can we even process the data? How fast?

# Sequencing outpacing informatics capabilities

| Quality Control of raw reads | 454 reads | Illumina reads |
|---|---|---|
| Platform | 1/4 run 454 FLX titanium | 1 lane Illumina GAIIx |
| Raw reads # | 0.18M/68.56Mbp | 15.89M/1.21Gbp |
| Human contamination | 35% | 20% |
| Trimming | lucy <20 | quality value<2 |
| Low quality % | 3.30% | 9.40% |
| input reads for assembling | 0.10M/40.3Mbp | 12.27M/0.93Gbp |

| | Computer Cluster Specs | Time | Size | Parameters |
|---|---|---|---|---|
| Illumina reads blastn vs NT | Head Node x 1: Intel Xeon L5410 / Quad Core Processors / 2.33GHz x 2, 16GB System Memory, 70 GB for OS, 1 TB RAID 1 Storage<br>Compute Node x 46: Intel Xeon E5420 / Quad Core Processors / 2.5GHz x 2, 16GB System Memory, 1 TB Raid 1 Storage (compute nodes are diskless) | 9 days | 1.6GB | blastall -C F -b 1 -F F -a 8 -p blastn -W 7 -m 8 |
| Illumina reads blastx vs NR | Head Node x 1: Intel Xeon E5520 / Quad Core Processors / 2.26GHz x 2, 32GB DDR3 1066 Mhz ECC Registered System Memory, 1 TB Raid 1 for OS, 19.8 TB RAID 6 Storage<br>Compute Node x 21: Intel Xeon E5520 / Quad Core Processors / 2.26GHz x 2, 16 GB DDR3 1066 Mhz ECC Registered System Memory, 160 GB SATA II 7200 RPM RE Rated Hard Drive | 28 days | 281GB | blastall -p blastx -d nr -W 2 -Q 11 -F "m S" -e 100 -i 10.fasta -a 2 -o 10.blastx |

**454 Ti Status**

**> 800 bp: *ab initio* gene calling**

**300-800 bp: BlastX + *ab initio***

**< 300 bp: BlastX**

**< 80 bp: Not processed**

**1 Million reads / day / 320 cores**

@Forsyth Inst.

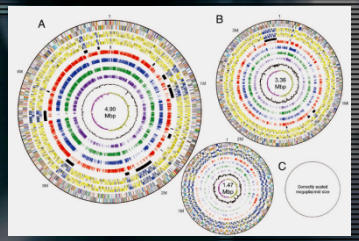
**Illumina Status:**

**20 million reads**

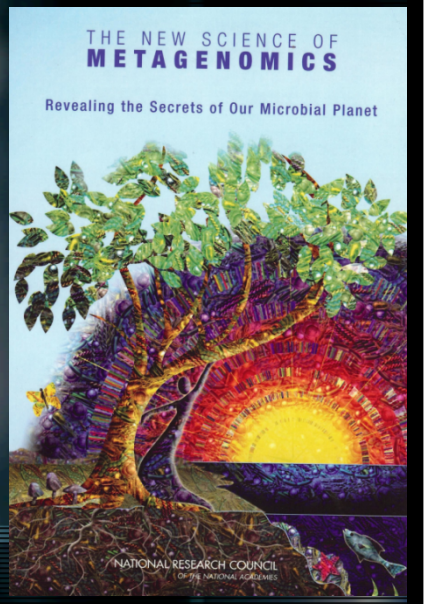
**/ 1 hr / 320 cores**

*Collaborating with hardware companies and HPC, GPUs, FPGA, etc.*

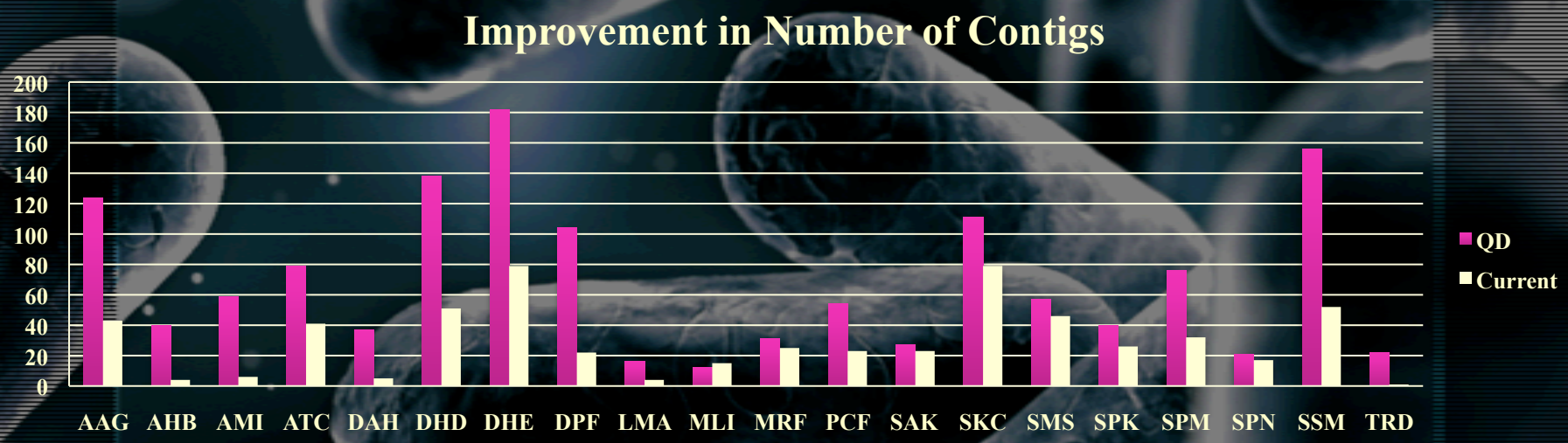
# Putting a few ~~hundred thousand~~ ~~Tens of~~ ~~Millions~~ ~~Billion~~ reads to good use!

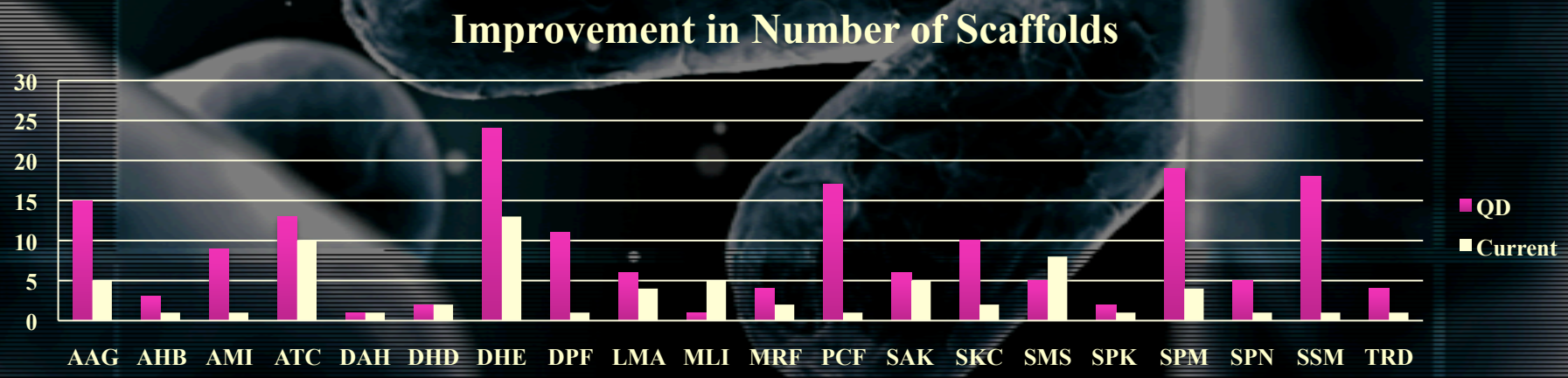


- Genomics
  - *De novo* and re-sequencing of isolates and near neighbors
- Metagenomics
  - 16S and shotgun
- Minute quantities of DNA and Single-cell genomics
- Transcriptomics (RNAseq)



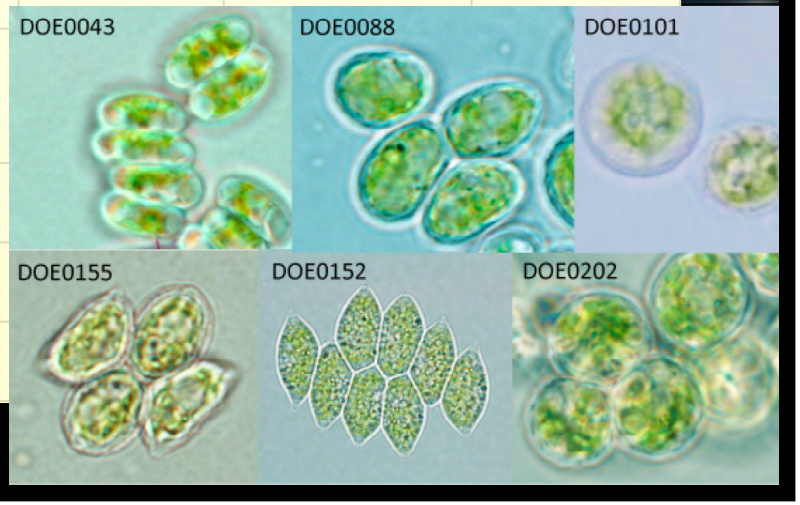

# Changes in genome statistics with finishing effort



# Algal Genome Projects Underway

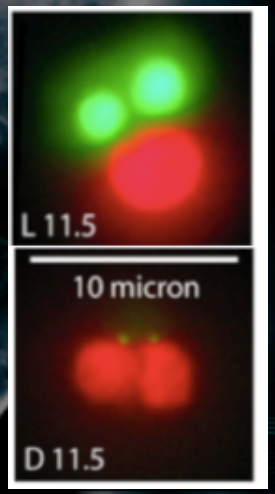
| Genome | Sequenced | Assembly Quality | Size (Mbp) | Scaffolds | Total Contigs |
|---|---|---|---|---|---|
| *Nannochloris sp.* | Yes | Improved HQ Draft | 15.2 | 39 | 58 |
| *Chlorella protothecoides* | Yes | Improved Draft | 21.4 | 149 | 1491 |
| *Chrysochromulina sp.* | Yes | Std. Draft | 75.9 | N.D. | 55838 |
| *Nannochloropsis salina* | Yes | Improved Draft | 29.4 | 225 | 1117 |
| *Tetraselmis sp. LANL1001* | In process | | | | |
| *Algae sp. DOE101* | No | | | | |
| *Algae sp. DOE1412* | No | | | | |
| *Algae sp. Phycal1228* | No | | | | |



# NAABB Projects at LANL

- **Genomes:**
  - 5 genomes
  - Sequencing:
    - Illumina (20)
    - 454 SE (4)
    - 454 PE (7)
    - Sanger (2000)
    - PacBio (10)
- **Transcriptomes (RNA-seq):**
  - 130 samples
  - Illumina only
  - Time courses, varying conditions

***Many collaborators:*** *Pete Lammers (NMSU, Solix), Jian Xu (Qingdao Institute of BioEnergy and Bioprocess Tech.), Judy Brown (U.Arizona), Tim Devarenne (TAMU), Sabeeha Merchant (UCLA), Dick Sayre (NMC/LANL)*

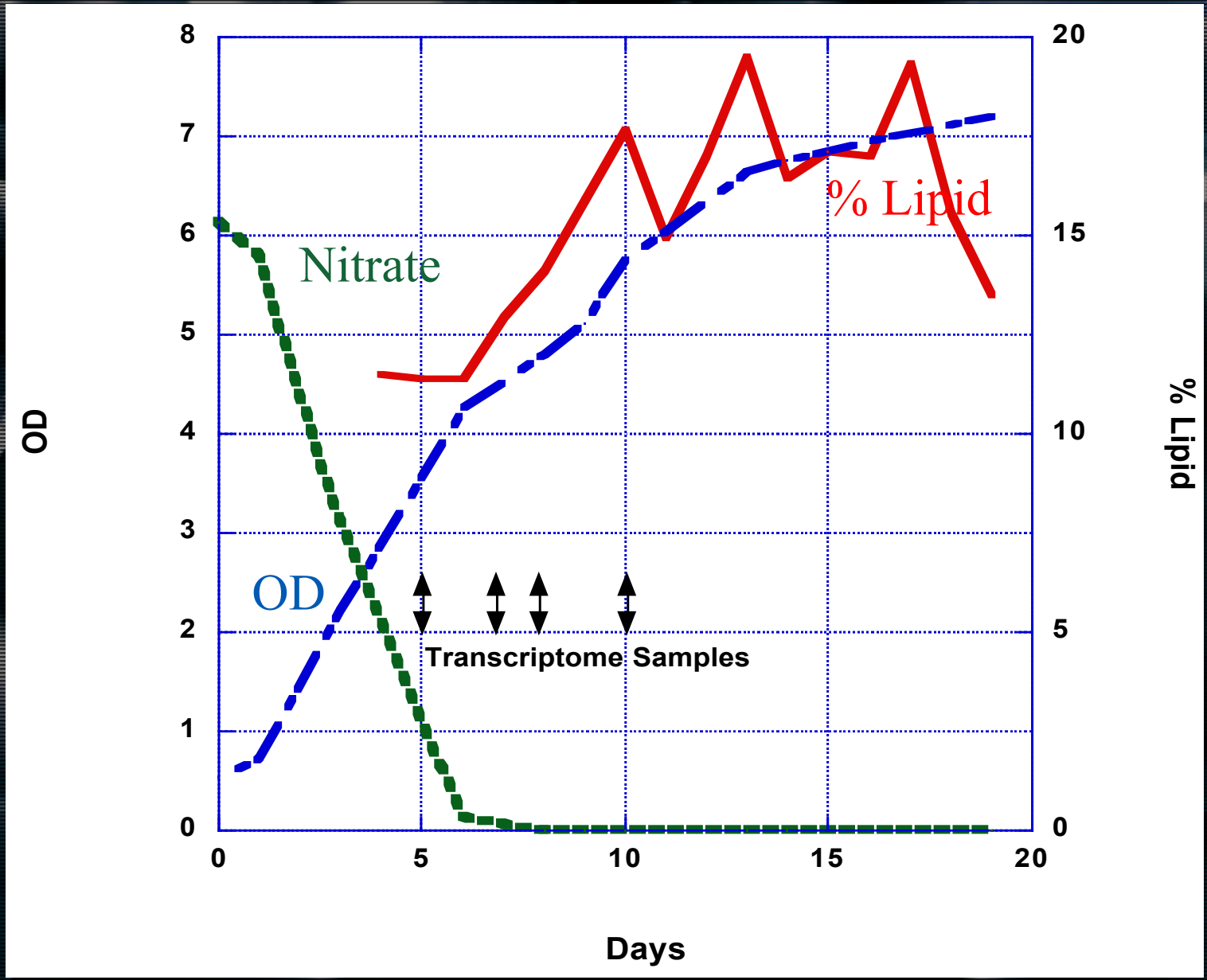



Near end of light phase

Near end of dark phase

12 hr light/12 hr dark
Green = lipid body; Red = Chloroplast

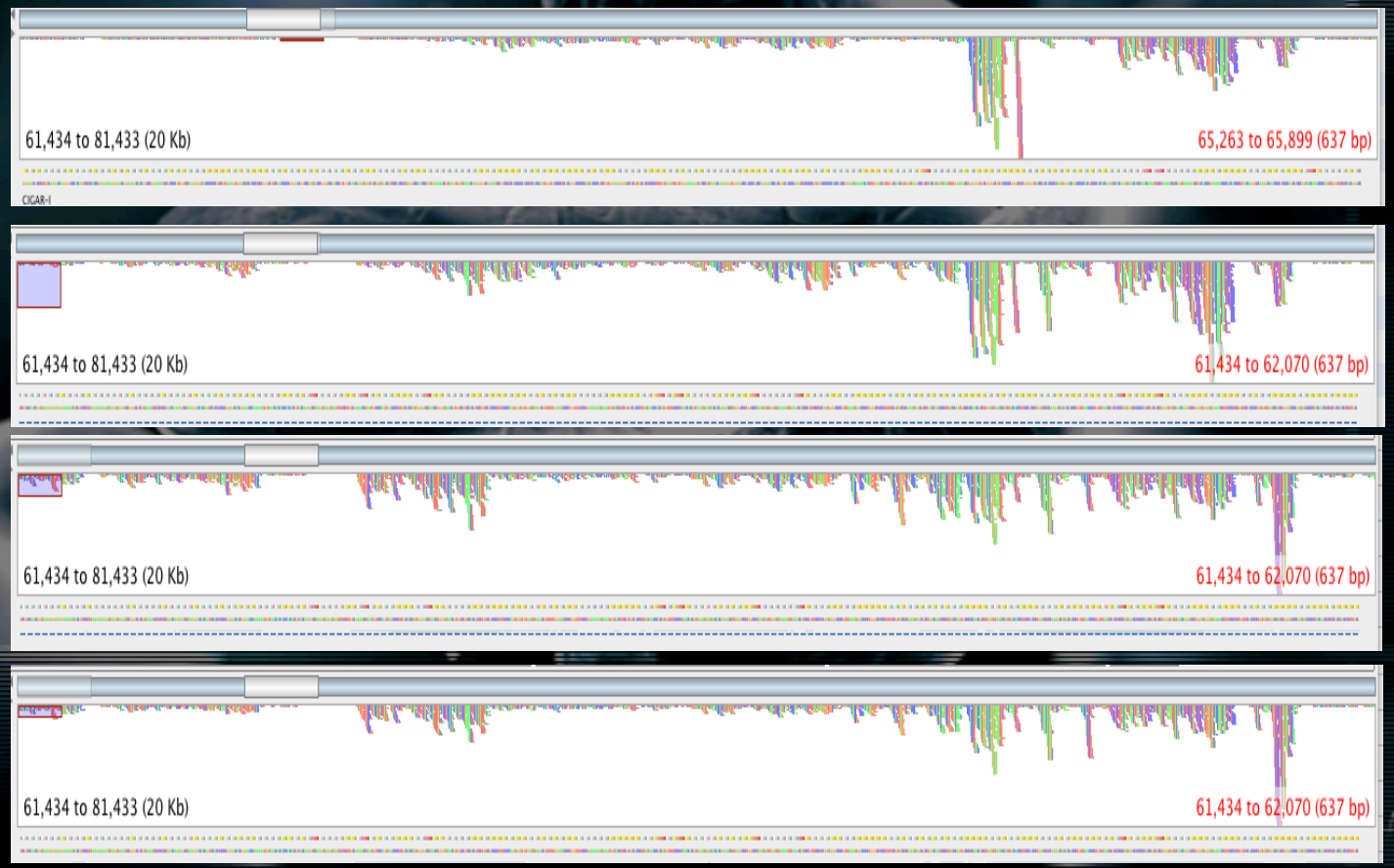
# Bioreactor Batch Culture – N deprivation



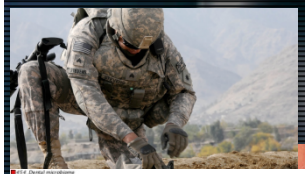
# Transcript expression during N depletion

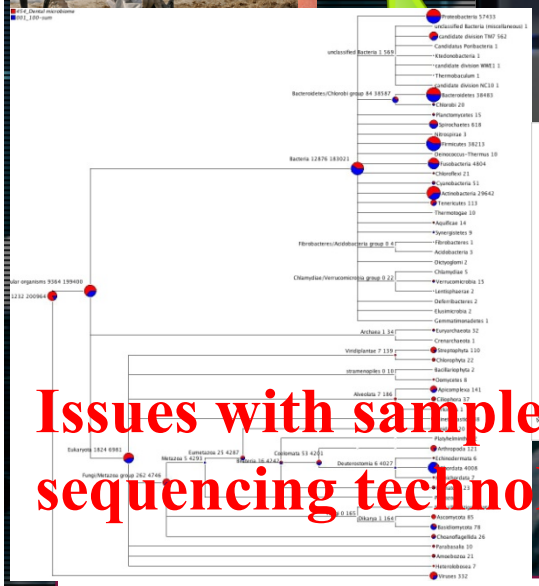
*Illumina can provide an extremely large dynamic range to be explored*

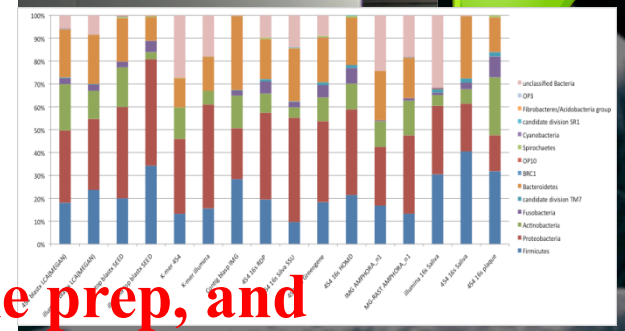
$[NO_3^-]$

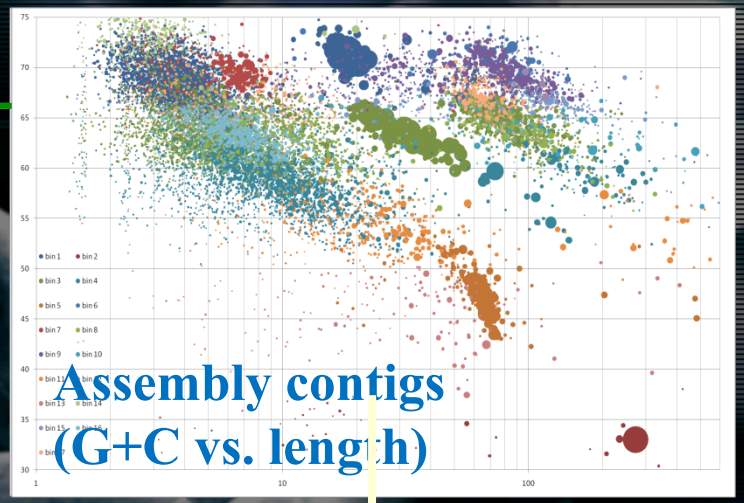


# From metagenomes to microbes

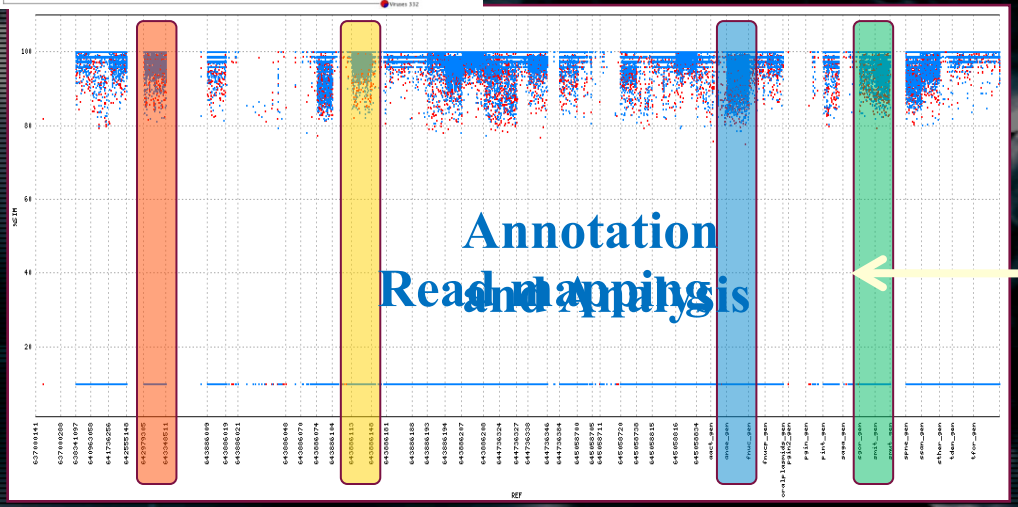


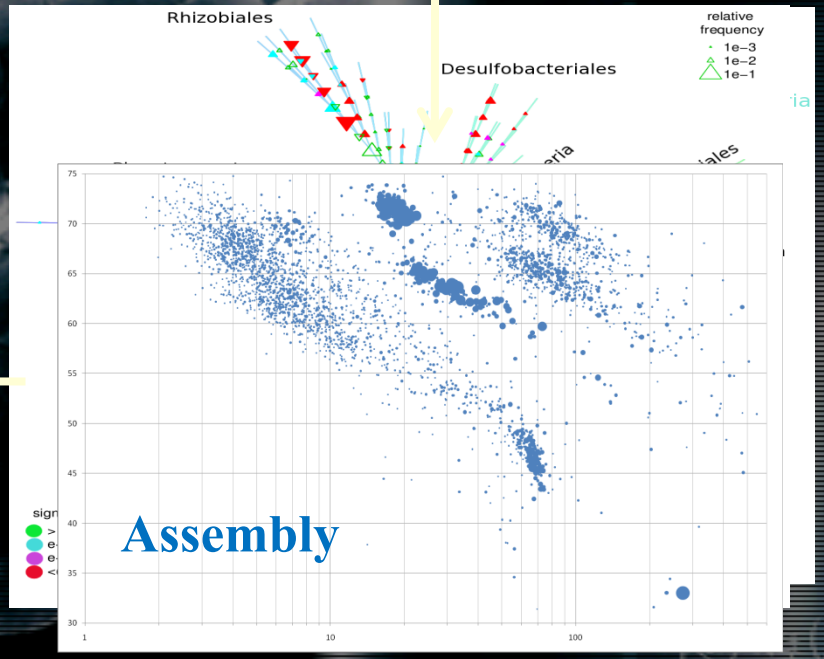




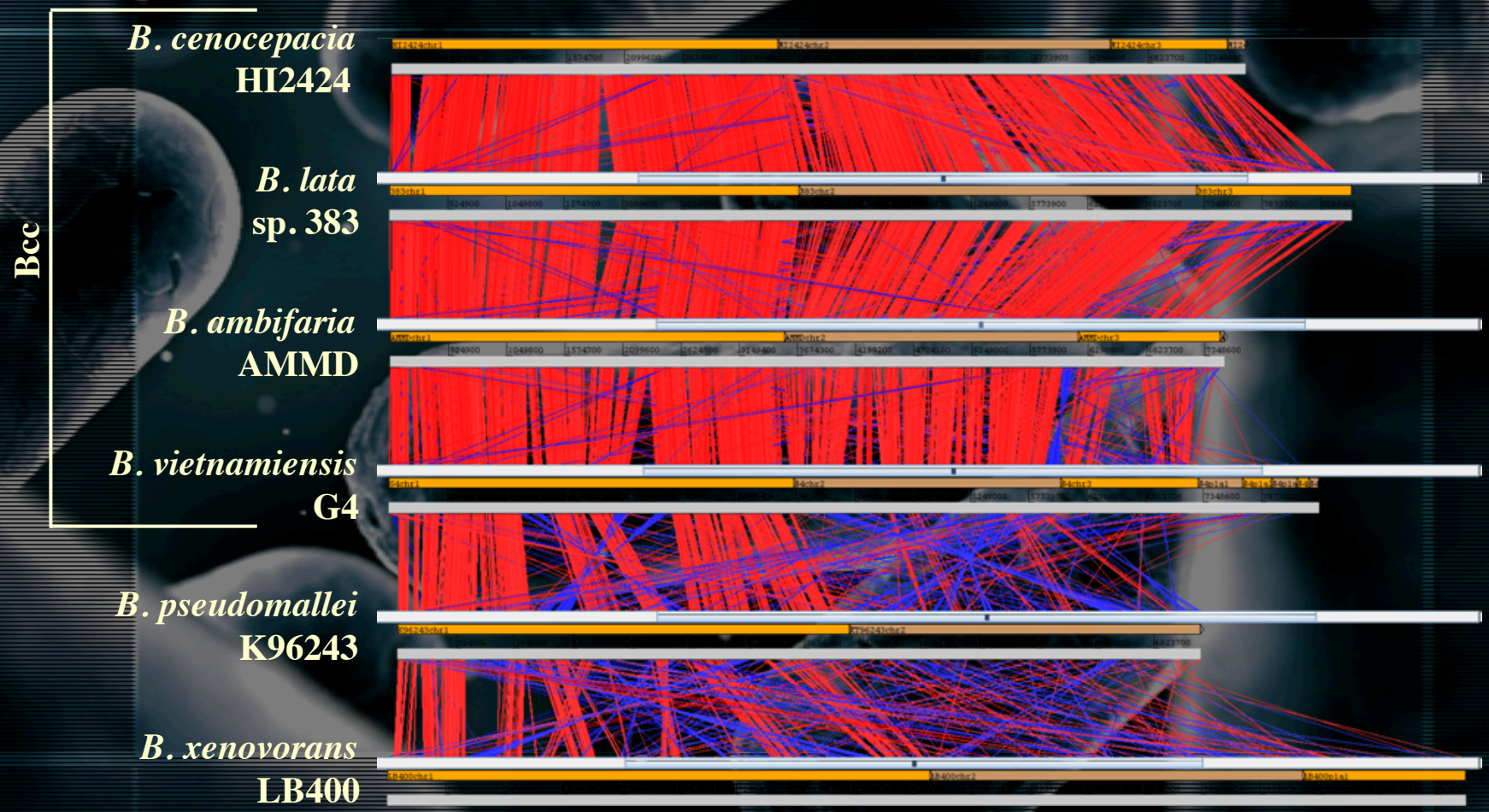



**Issues with sample prep, and sequencing technology biases!!:**

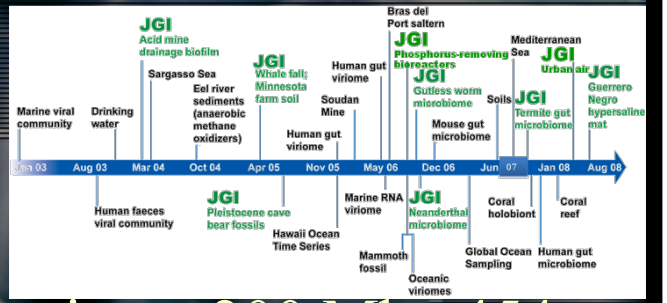


# What 16S may miss in terms of genome (microbial) diversity



# Next-Generation Sequencing has opened the metagenomic floodgates (now >>300)!




62 Termite hindgut, 62 Mbp Sanger

3,200 Avg. Metagenome project, 3Gbp Illumina + 200 Mbp 454

17,000 Cow rumen, 17 Gbp Illumina

# The next challenge: Terabase scale

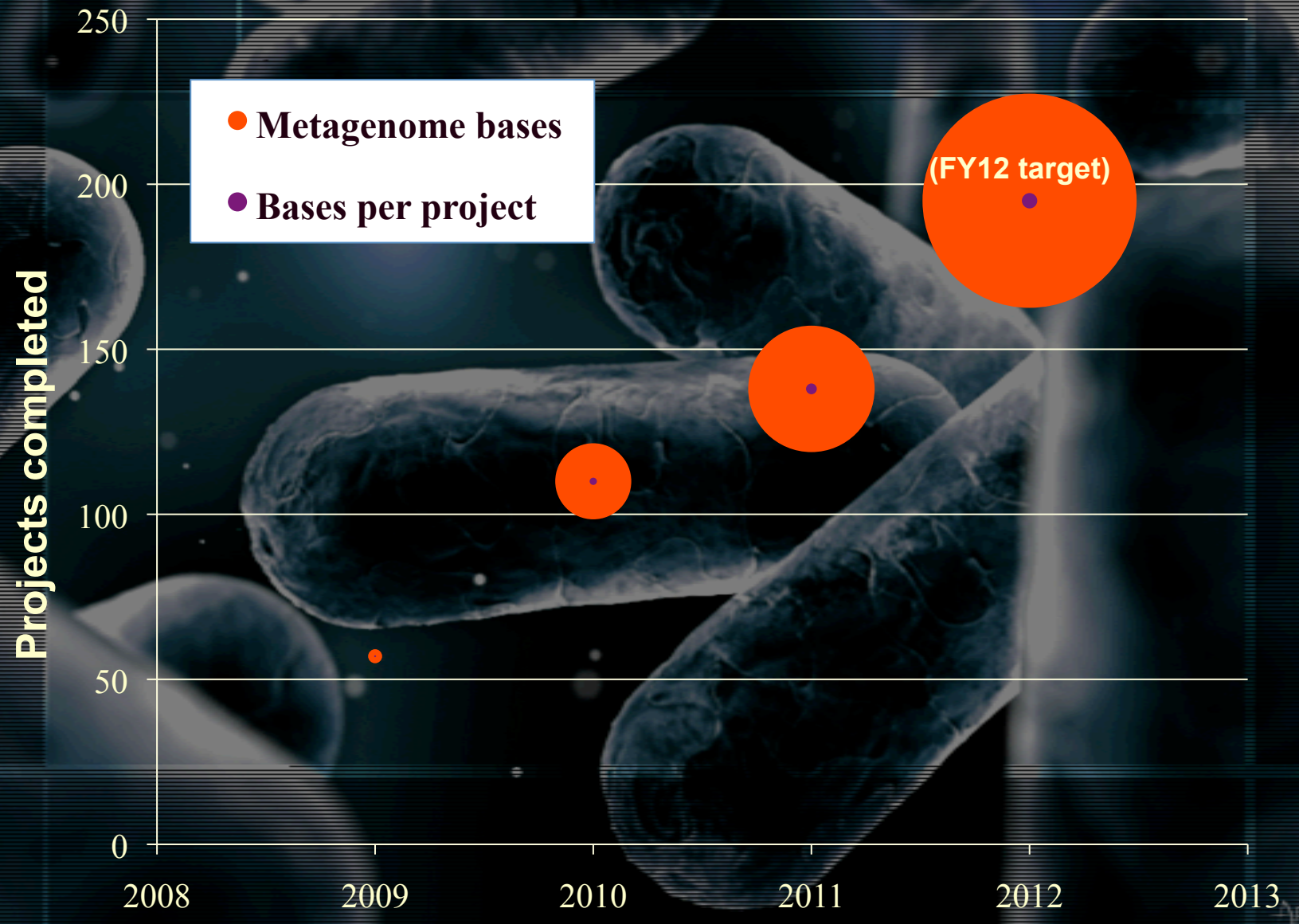
100,000 JGI Tb-challenge project pilots, ~100 Gbp

1,000,000 JGI Tb-challenge projects, ~1 Tbp

**High probability of computational bottlenecks**
**all vs all will NOT scale!**
**New approaches needed…**

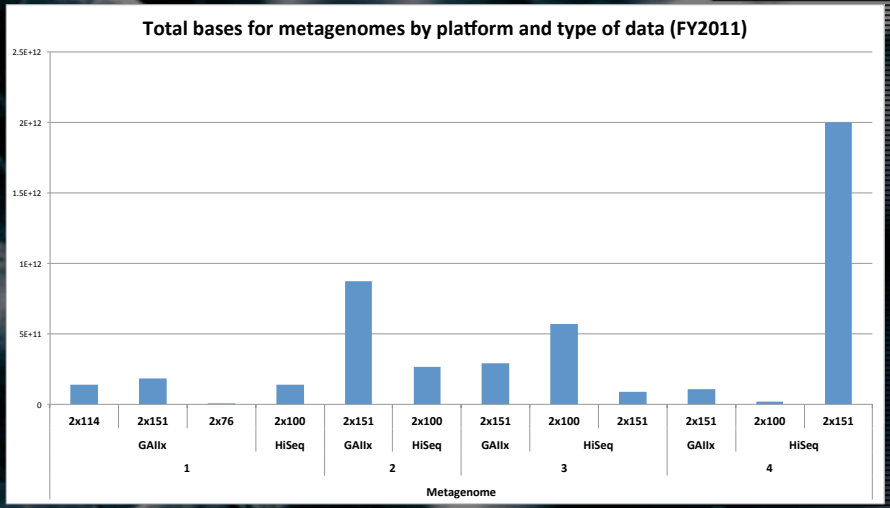
Slide courtesy of P. Hugenholtz

# Shotgun metagenomics now common…

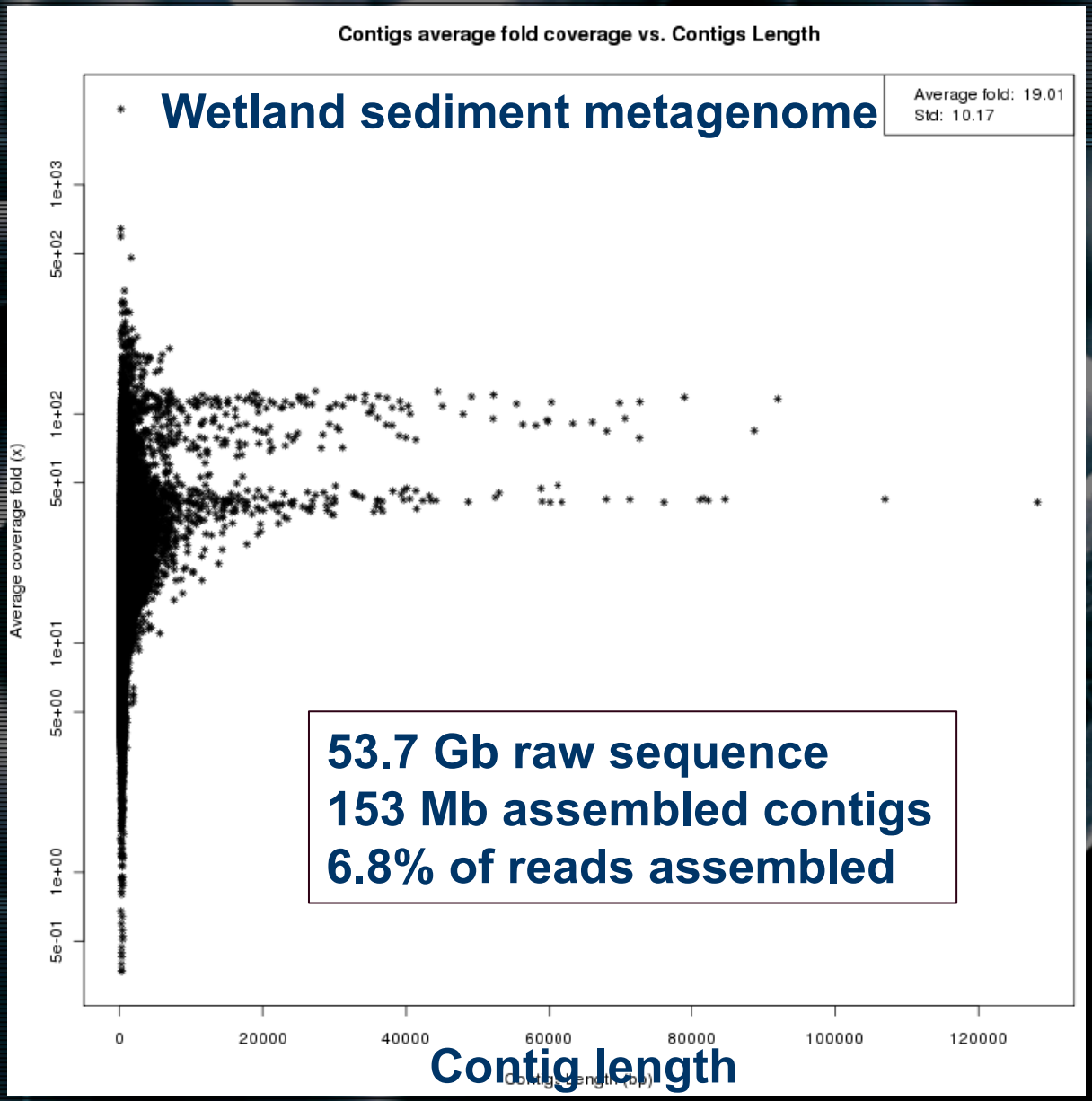


# Metagenomes tackled

- Projects ranging from 1 lane of Illumina 1x36bp to 454+many lanes of 2x150bp or 2x100bp
- ~150 JGI metagenomes
- ~250 HMP metagenomes
- Many others…
- Current:
  - 7.1TB of RAM, 1000 cores, 300TB of usable shared storage
  - Additional 1TB of memory and 128 cores for the Single System Image cluster
  - Hadoop "DISC" cluster, 420 cores, ~420GB of memory, ~105TB of storage
  - Applying for an additional single system: 8TB, 512 core



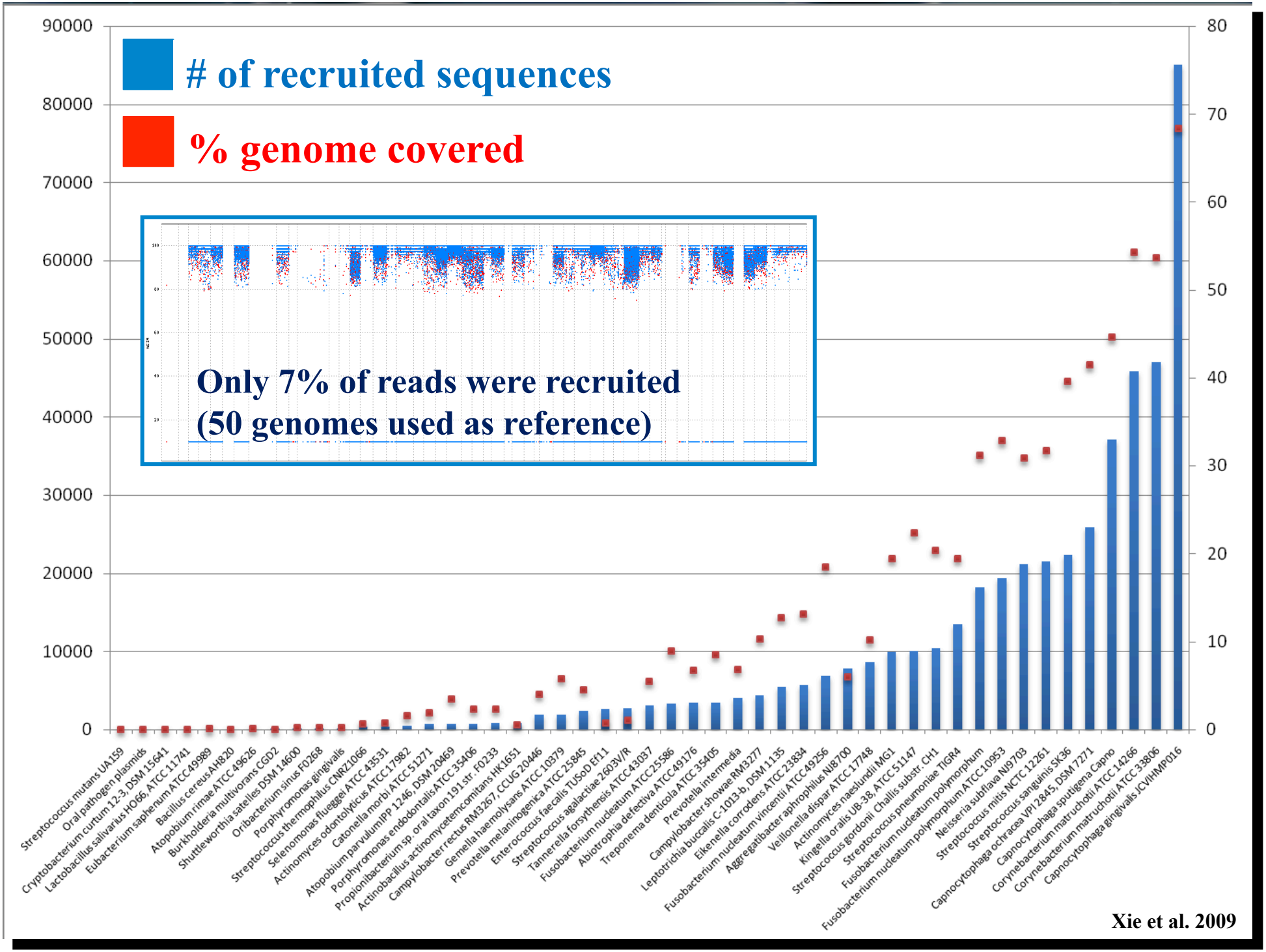
# Do we need to sequence so deep?




Contig depth of coverage

**Issue:**
**2 lanes provide minimal assembly – need more data**

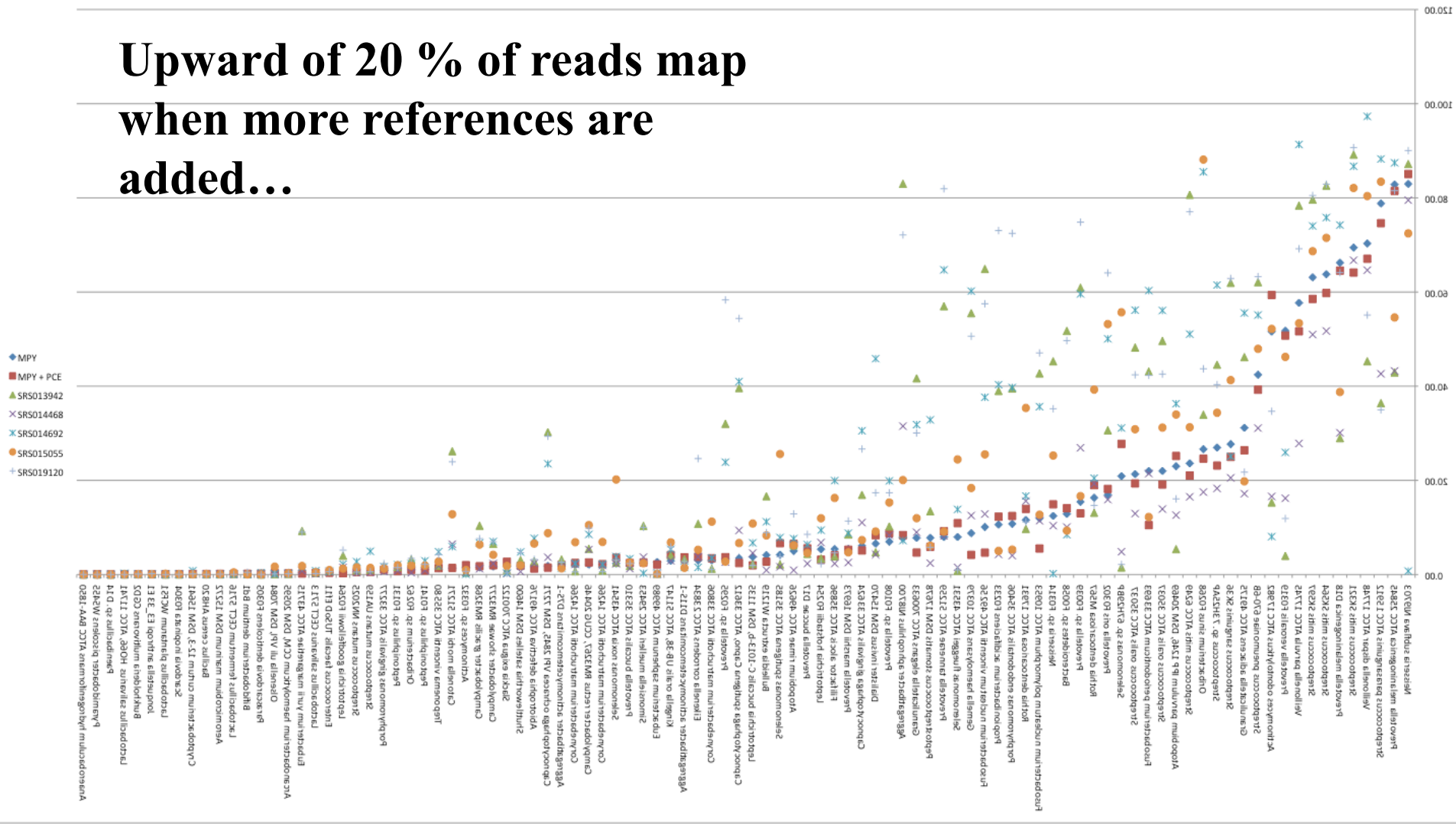
**2 lanes can barely be assembled – need less data (or new algorithms)**



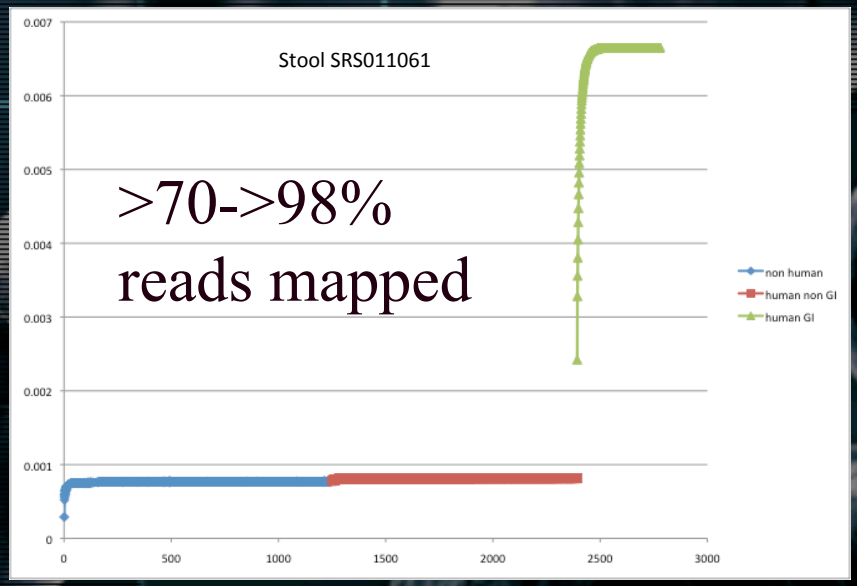

Xie et al. 2009

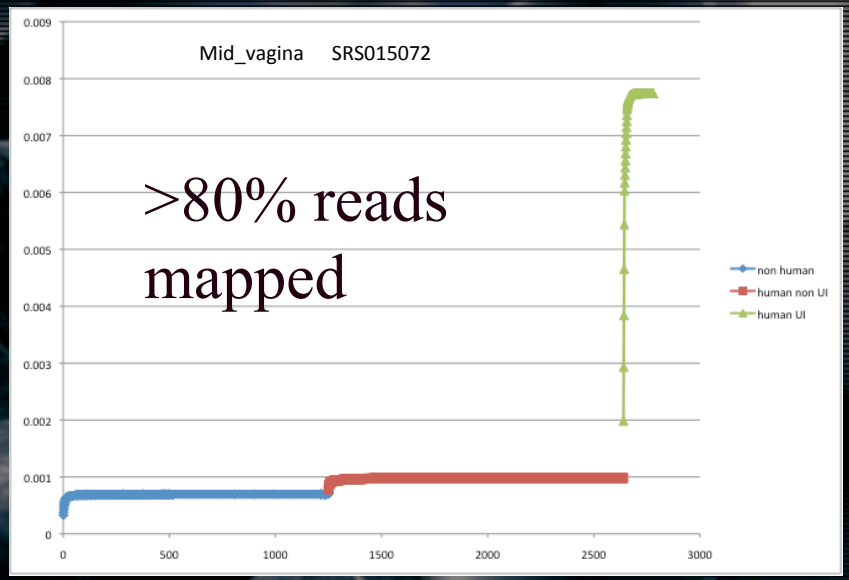
# Metagenome studies are much aided by reference genomes

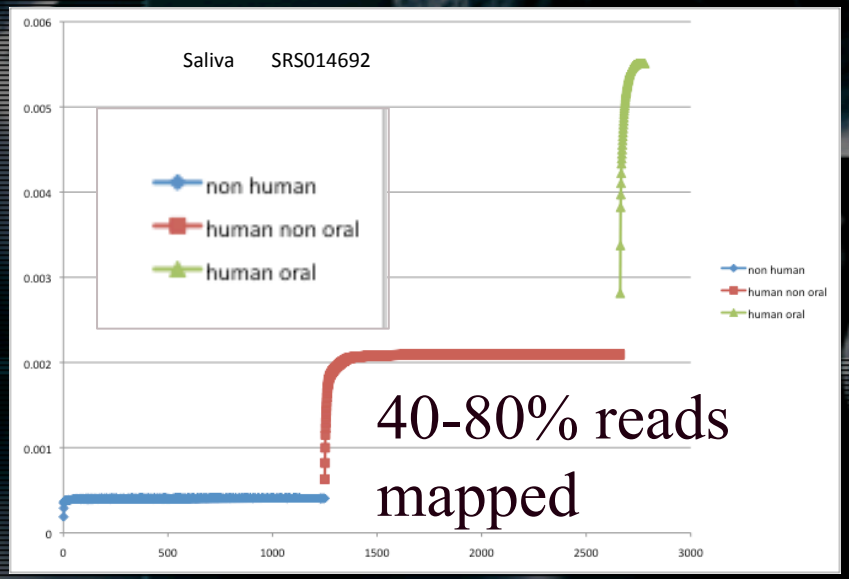
**Upward of 20 % of reads map when more references are added…**

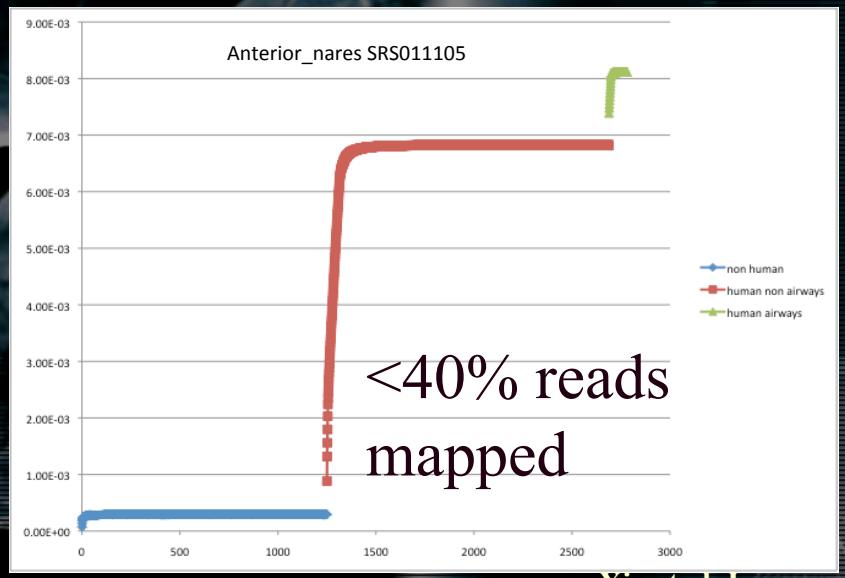


# Percent of reads that map to now >2500 HMP reference genomes




Xie et al. In prep

# A test for a rapid response scenario…

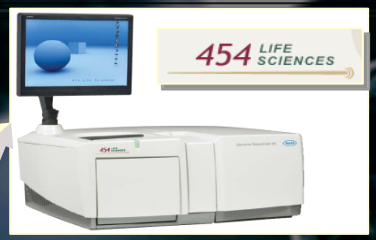
Spiked sample
30ng-2ug DNA

6-8hrs

MDA

+18hrs

10-36hrs




0.25-10hrs
Analysis

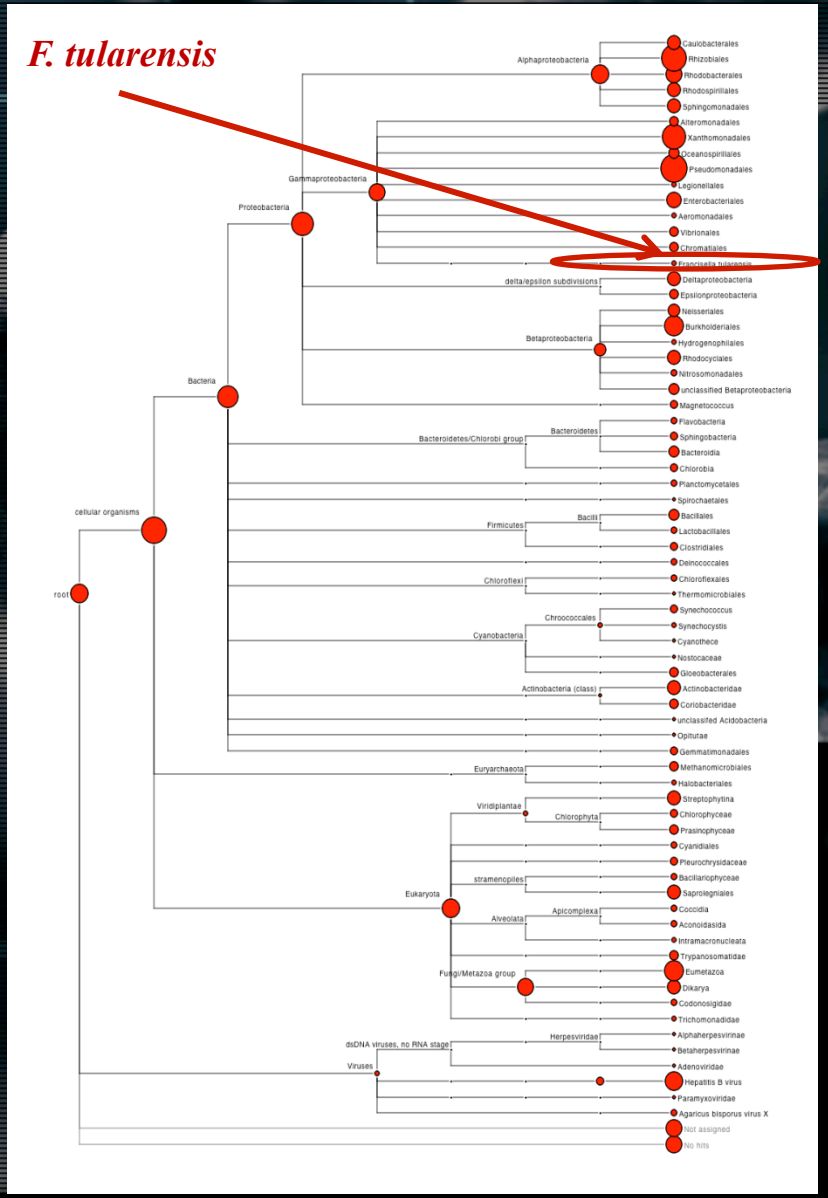
Read mapping

and

Assembly and Blast

VS.
human
viral genomes
bacterial genomes

# Detection and characterization by sequencing complex samples




**Unamplified:**

**Illumina 1,202,614 reads mapped (0.3% of total)**

**Amplified DNA:**

**454: 92 reads mapped (0.005% of total )**

**PacBio: 39 reads mapped (0.059% of total)**

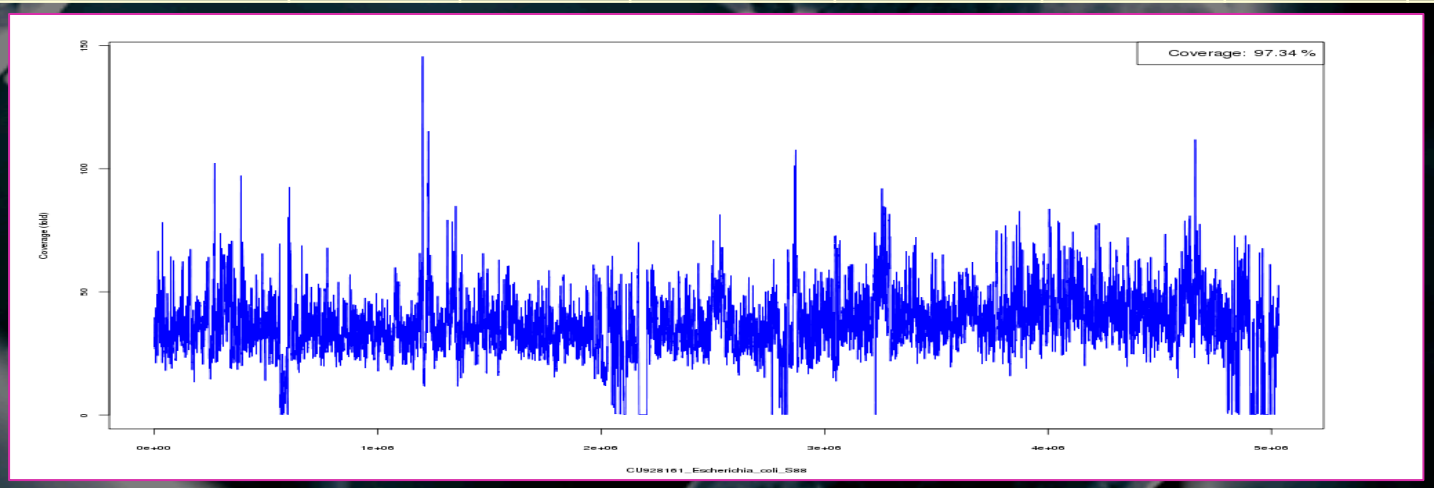
**=18X coverage – strain ID based on SNP analysis**

***Francisella tularensis* subsp *tularensis* FSC198**

# Analysis of 2 Clinical Metagenomes:

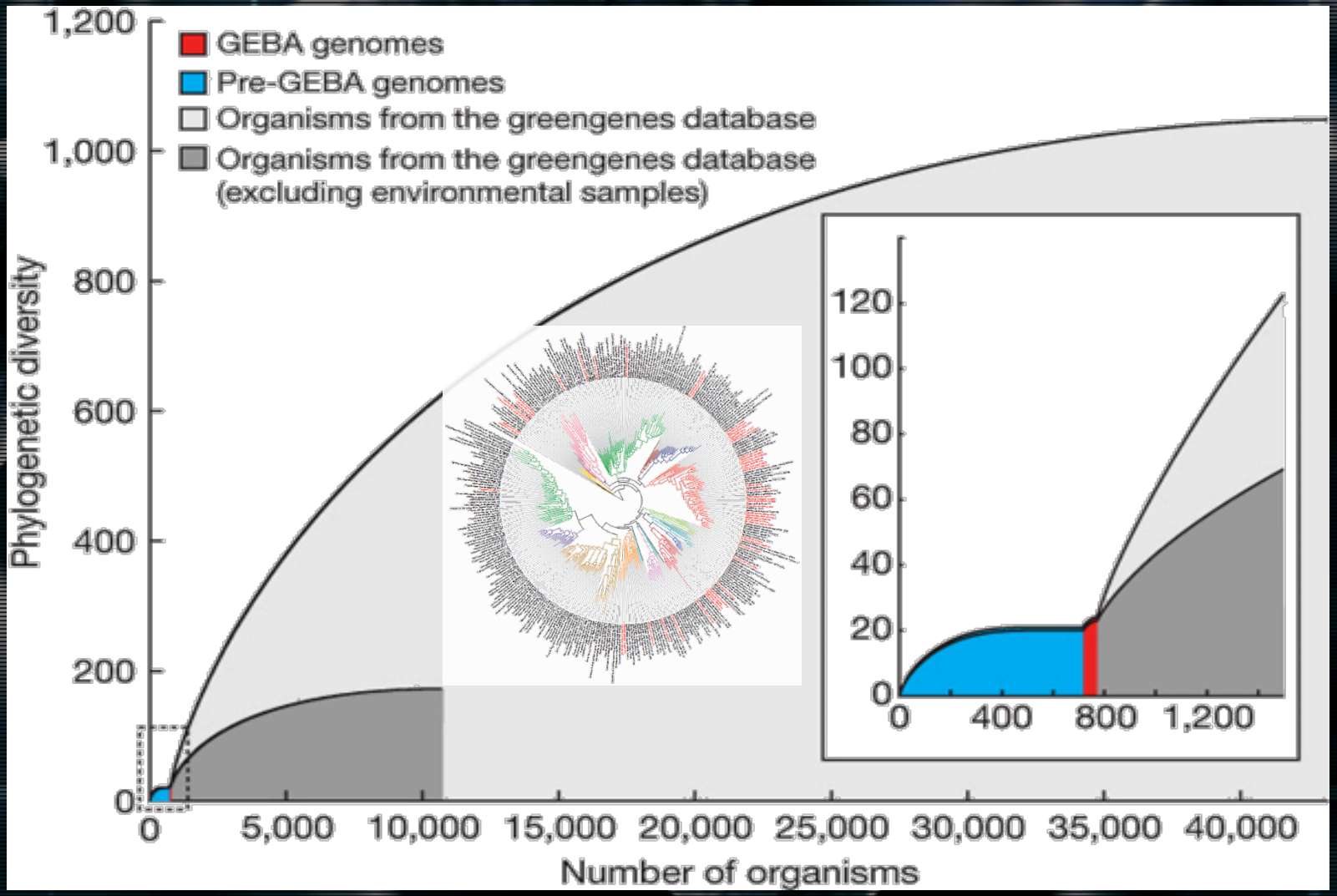
- 2 fecal sample suspensions prepped by U. Columbia (Lipkin)
- Unamplified metagenome mostly Human with some *E. coli* (0.8%)

| ID | Length | GC | Avg_fold | Base_Coverage | #Gap | Gap bases | # SNPs/ INDELs |
|---|---|---|---|---|---|---|---|
| E. coli S88_chromosome (NC_011742) | 5032268 | 0.51 | 37.13 | 97.34 | 229 | 133817 | 10941 |
| S88 plasmid pECOS88 | 133853 | 0.49 | 60.32 | 98.04 | 3 | 2620 | 51 |
| Salmonella enterica serovar Enteritidis plasmid pB(NC_005002) | 1983 | 0.57 | 517.4 | 100 | 0 | 0 | 50 |
| TY-2482_pTY3 | 1549 | 0.51 | 14.96 | 98.84 | 1 | 18 | 0 |




- *Escherichia coli* clonal group, O45:K1, belonging to the highly virulent subgroup $B2_1$
  - pS88 is a major virulence determinant circulating in human urosepsis and avian pathogenic strains

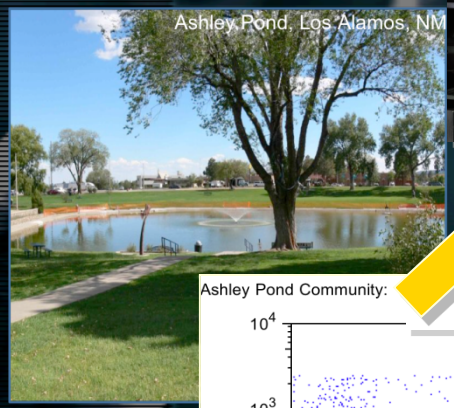
# Harnessing the power of sequencing to explore the unsequenced majority!

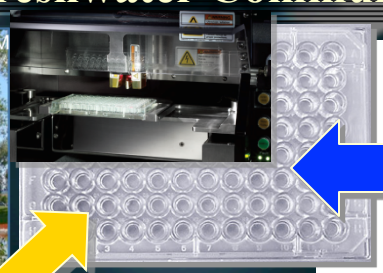


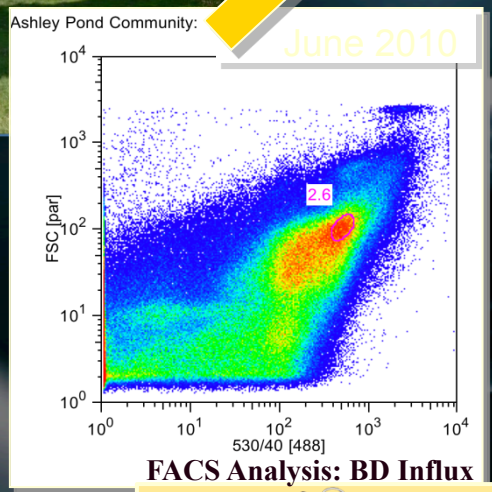

Wu et al. 2009 Science

# How to get more references from the environment?

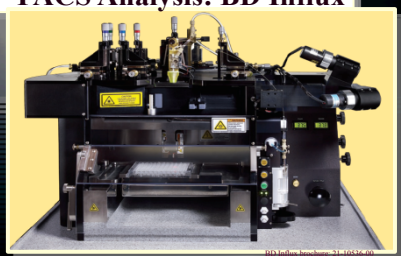
- 16S rRNA Freshwater Community Phylotyping

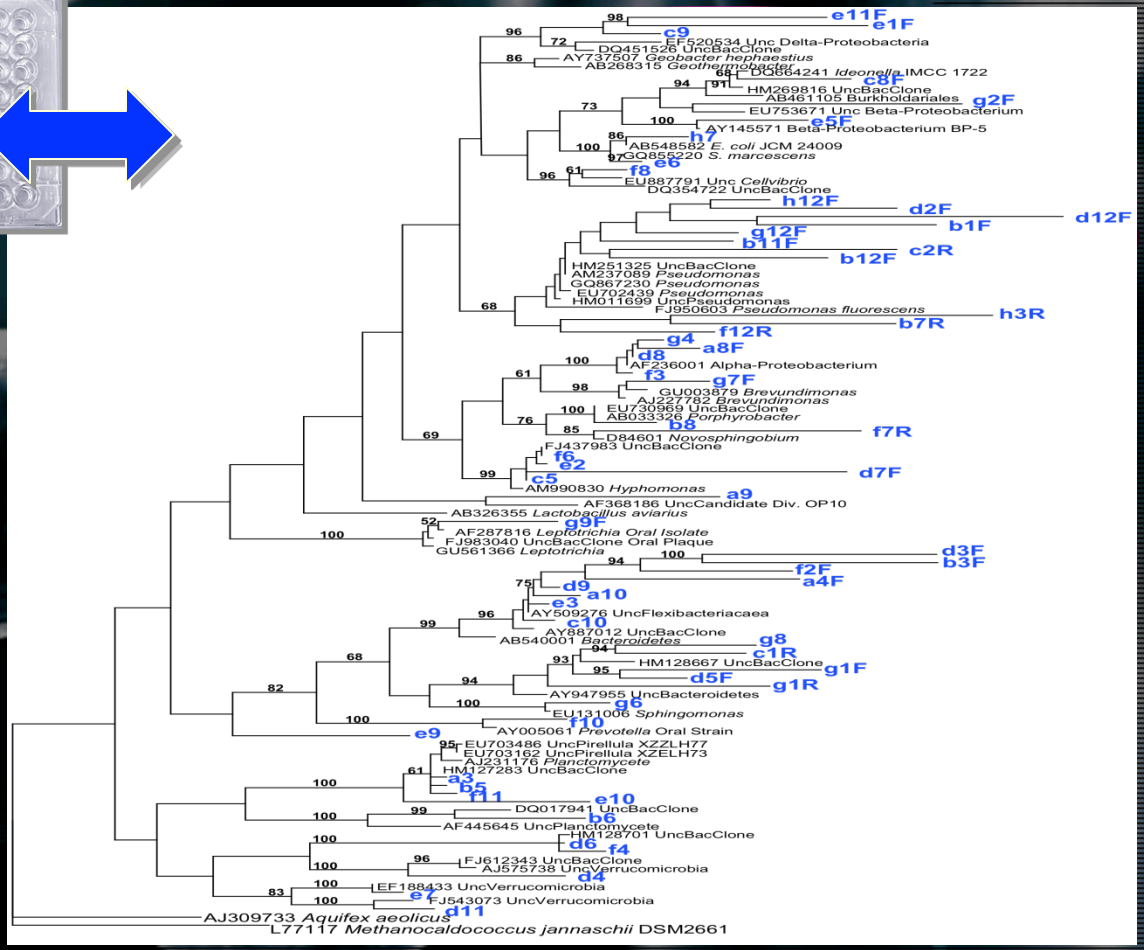




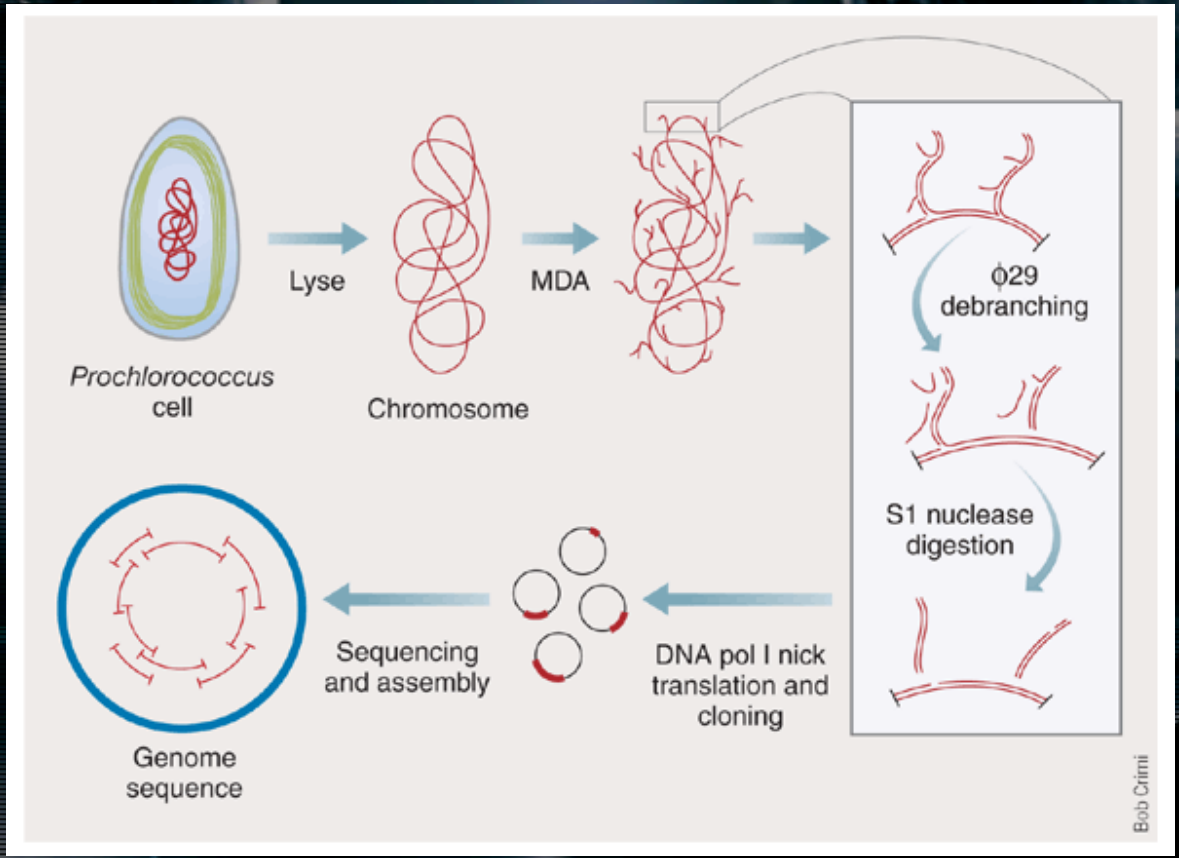

FACS Analysis: BD Influx





# Genome recovery from single cells

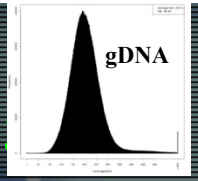
- Amplification results in "random" bias
- Affects recovery of genomic DNA – typically 30-70% of genome can be recovered from a single cell



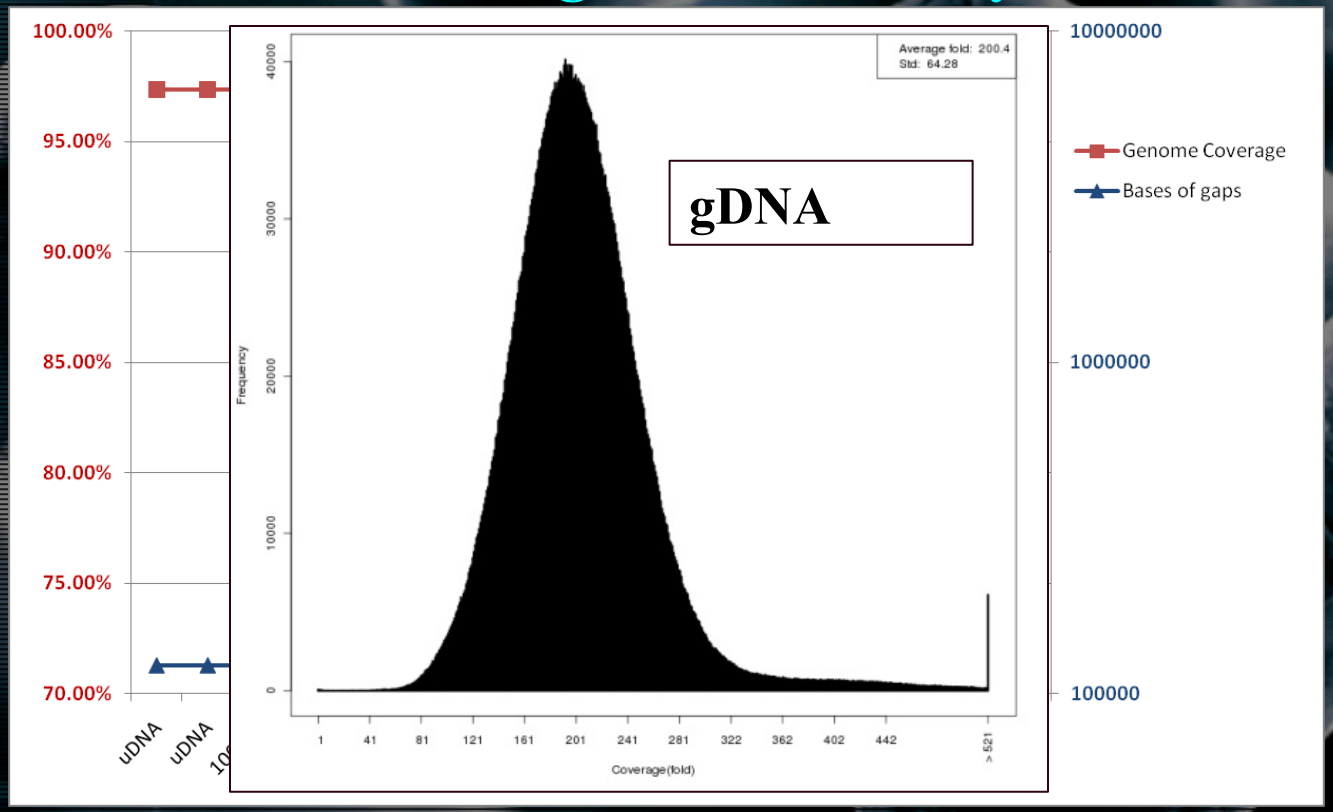

*Nature Biotechnology* **24**, 657 - 658 (2006)
doi:10.1038/nbt0606-657

Single-cell genomics
Clyde A Hutchison III & J Craig Venter

# Genome recovery from manipulated cell



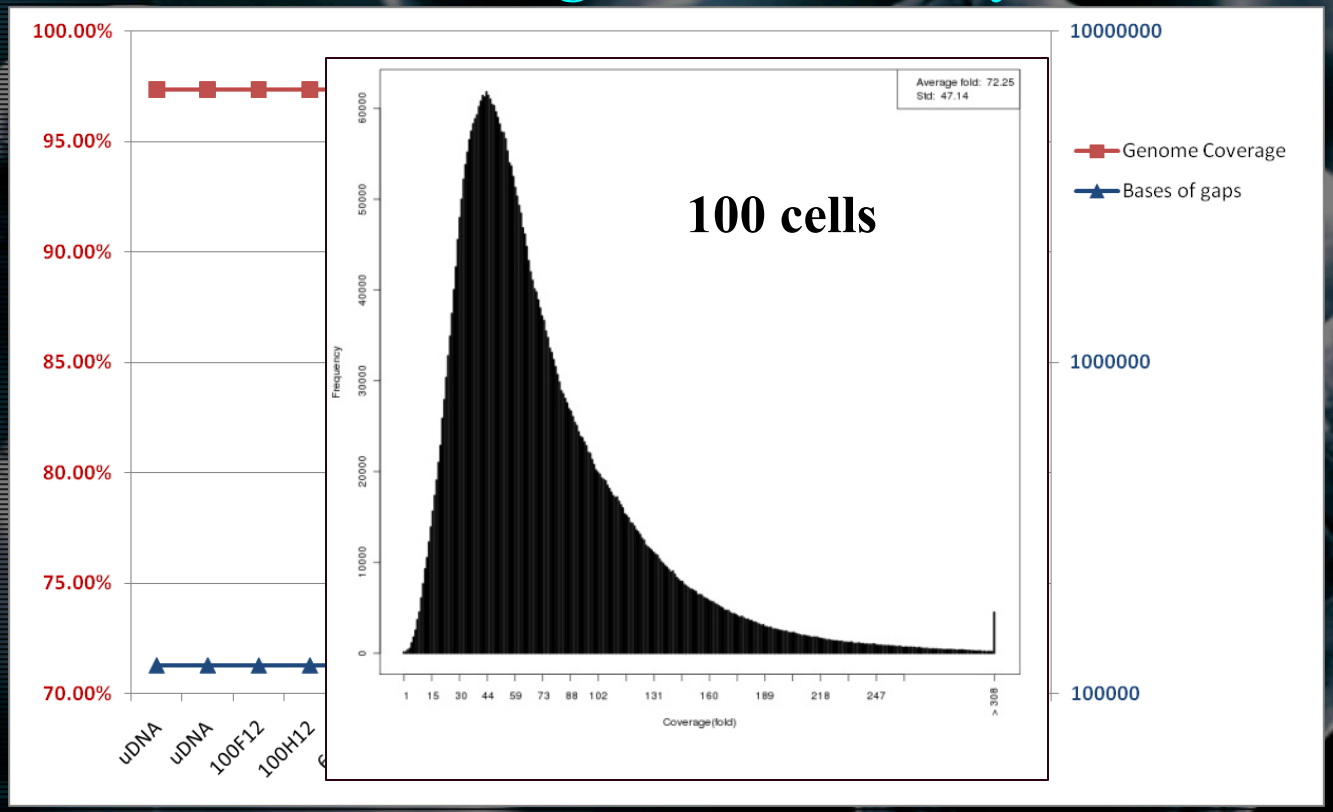

- First, evaluated genome recovery from 1-100 cells

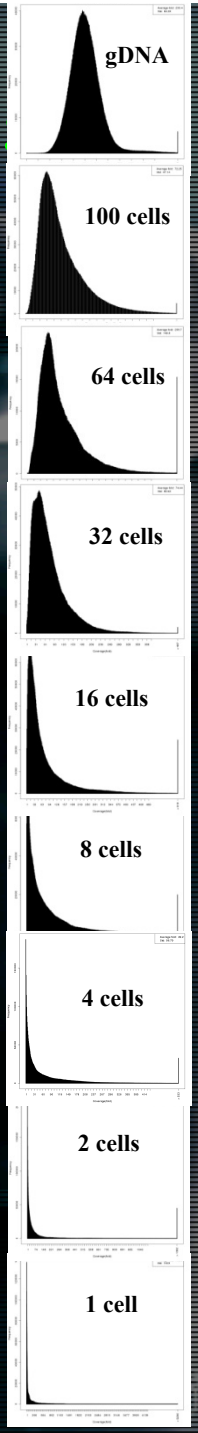



Lo et al. In prep.

# Genome recovery from manipulated cell

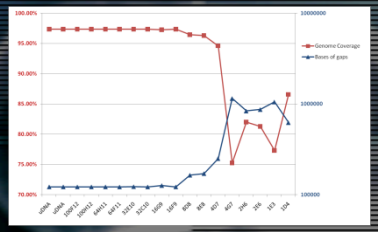
- First, evaluated genome recovery from 1-100 cells



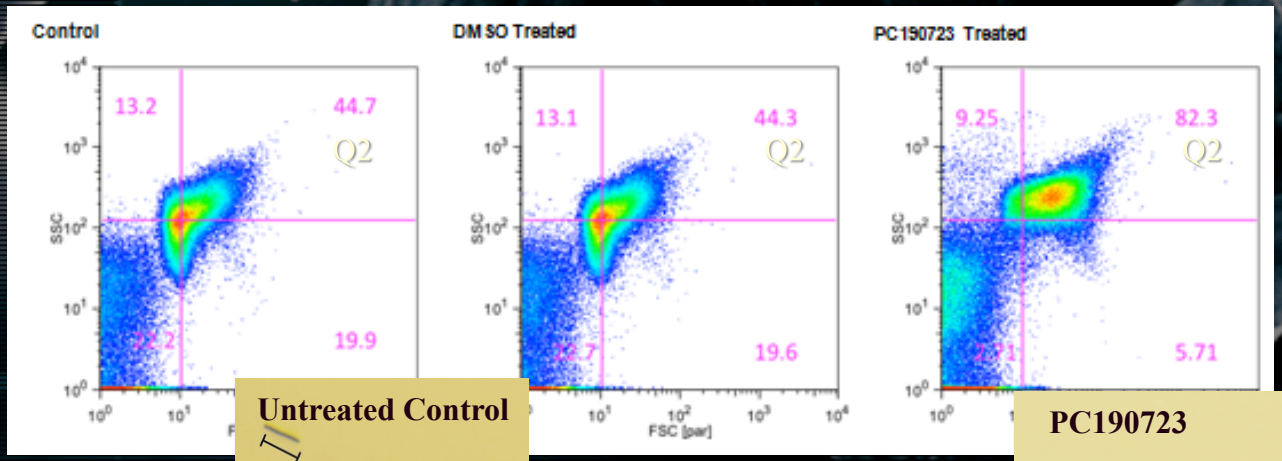



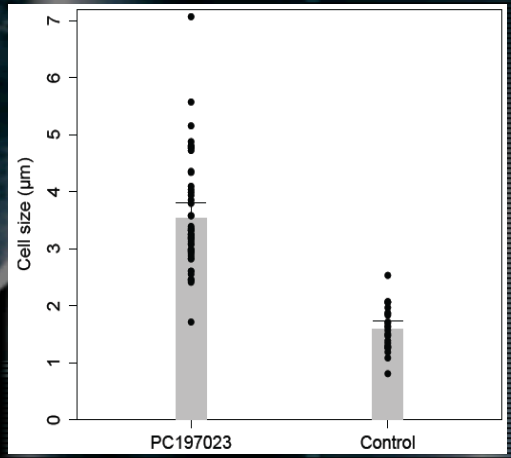

Lo et al. In prep.

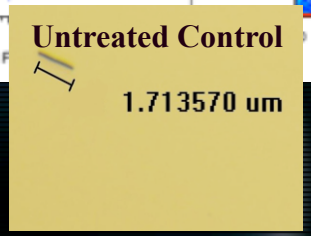
# Genome recovery from manipulated cells

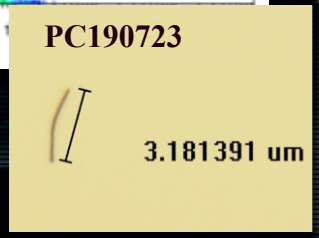


- First, evaluated genome recovery from 1-100 cells
  - Sort, MDA, sequence
  - More copies of genome = better coverage
- Inducing artificial polyploidy
  - Tested FtsZ inhibitor PC190723 on *Bacillus subtilis*



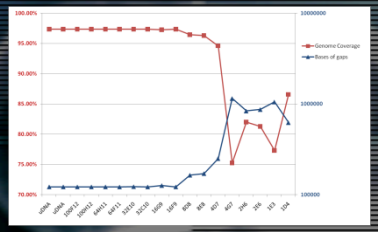



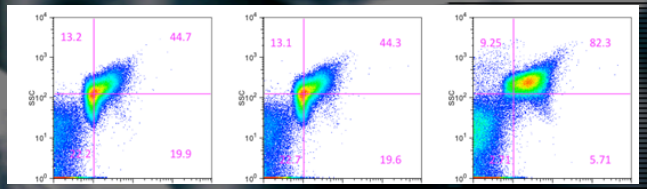




Dichosa, et al. Submitted

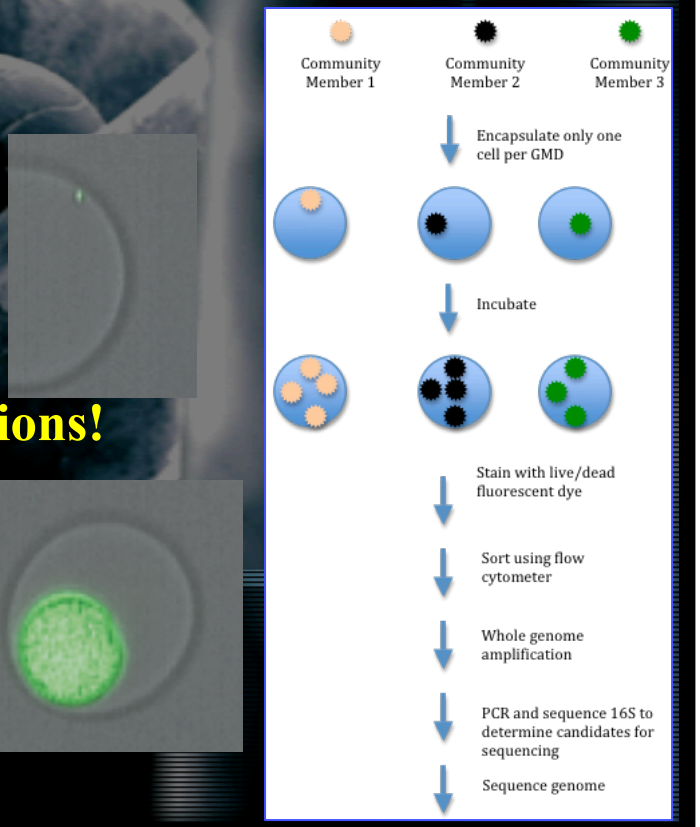
# Genome recovery from manipulated cells



- First, evaluated genome recovery from 1-100 cells
  - Sort, MDA, sequence
  - More copies of genome = better coverage




- Inducing artificial polyploidy
  - Tested FtsZ inhibitor on *Bacillus subtilis*




- Gel microdroplets for microcolonies
  - Dilute, encapsulate, incubate, stain, sort…
  - Can also be used to study microbial interactions!



# Tackling Metagenomes and NGS…

***Burkholderia* work**
- James Tiedje
- John LiPuma
- Erick Cardenas

***E. coli* work**
- David Hirshberg
- Ian Lipkin
- Sandy Gibbons
- Nicole Rosenzweig
- Shanmuga Sozhamannan
- Kim Bishop-Lilly
- David Norwood
- Tim Minogue
- Nancy Strockbine
- Many others…

**Metagenome work**
- Jim Tiedje
- Titus Brown
- Adina Howe
- HMP consortium
- Mihai Pop
- Joe Zhou
- Kostas Konstantinidis

**Metagenomics and Data Analysis Team**
- Nick Beckloff
- Tracey Freitas
- Ron Croonenberg
- Bin Hu
- Chien-Chi Lo
- Kuan-Liang Liu
- Matt Scholz
- Shawn Starkenburg
- Gary Xie
- Others…

**Informatics Team**
- Ben Allen
- Andy Seirp
- Roxanne Tapia
- Yan Xu
- Todd Yilk

**Single cell work**
- Roger Lasken
- Ramunas Stepanaskus
- Steve Hallam

**Wet-lab Team**
- Cheryl Gleasner
- Kim McMurry
- Krista Reitenga
- Xiaohong Shen
- Others…

**Project Management**
- Shannon Johnson
- Lynne Goodwin
- Others…

**Kmer team**
- Joel Berendzen
- Nick Hengartner
- Ben McMahon
- Judith Cohn

**Finishing and SCG**
- Olga Chertov
- Karen Davenport
- Armand Dichosa
- Michael Fitzsimons
- Ahmet Zeytun
- Others…

**Management Team**
- Chris Detter
- David Bruce
- Tracy Erkkila
- Lance Green
- Shunsheng Han

***And many others…***