Understanding the Human Oral Microbiome with Next-Generation Sequencing Technologies

Next Generation Sequencing and Oralgen2.0

www.lanl.gov/bioscience
www.oralgen.lanl.gov

Patrick Chain
pchain@lanl.gov
Metagenomics Applications Team, LANL
Metagenomics Program, JGI

IADR, Barcelona, Spain
July 2010

Genomics and Bioinformatics at LANL

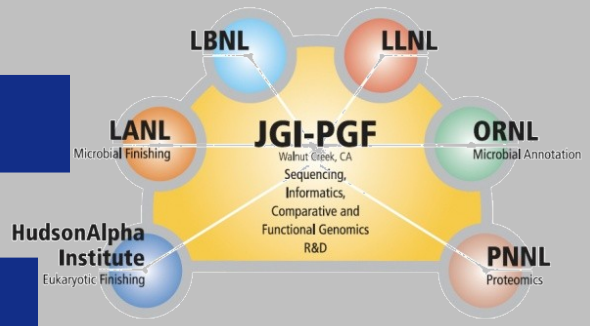
1982 GenBank

1988 The Los Alamos Center for Human Genome Studies

1990 HIV Sequence Database

1995 Influenza Sequence

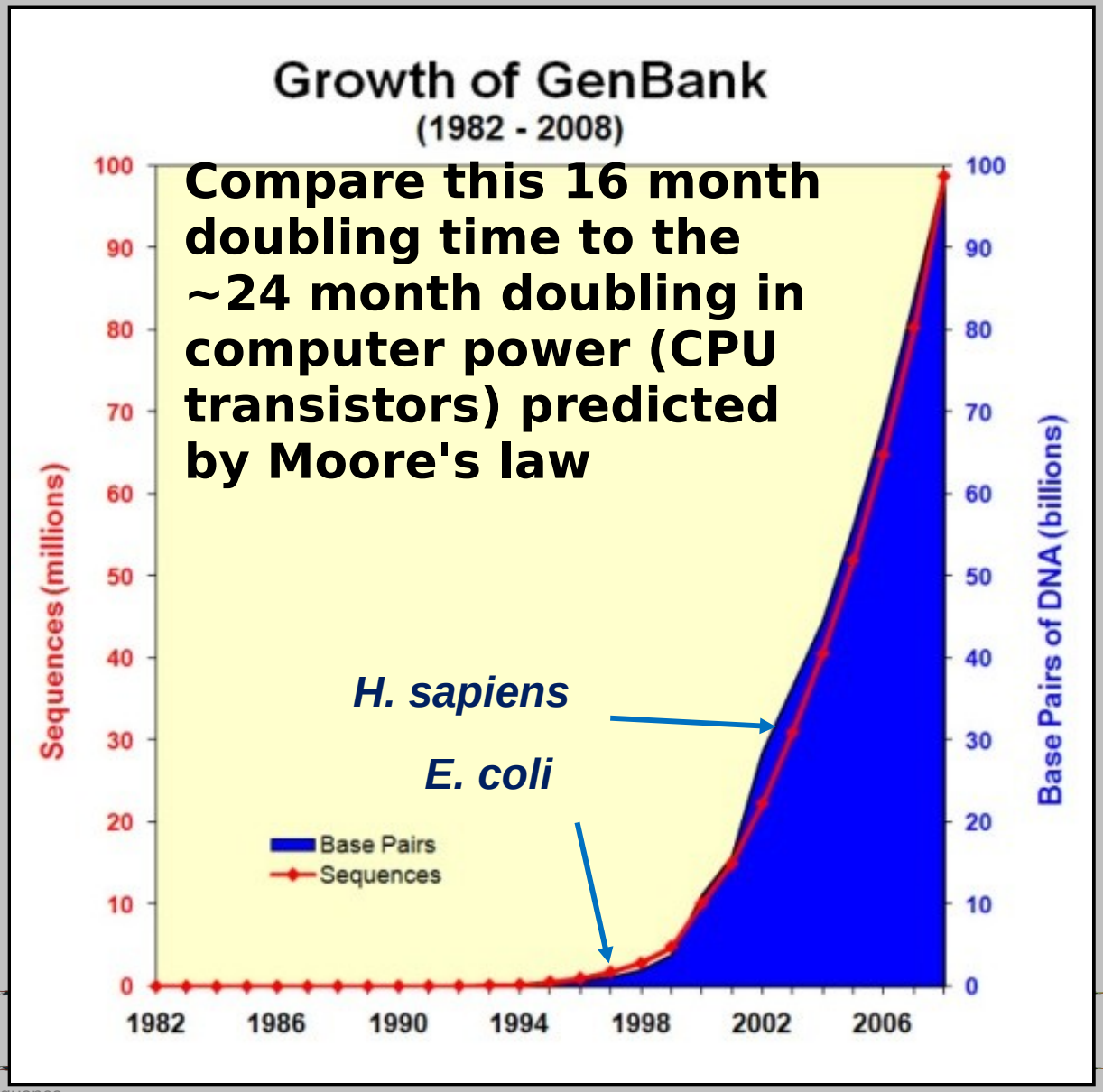
1997 Joint Genome Institute



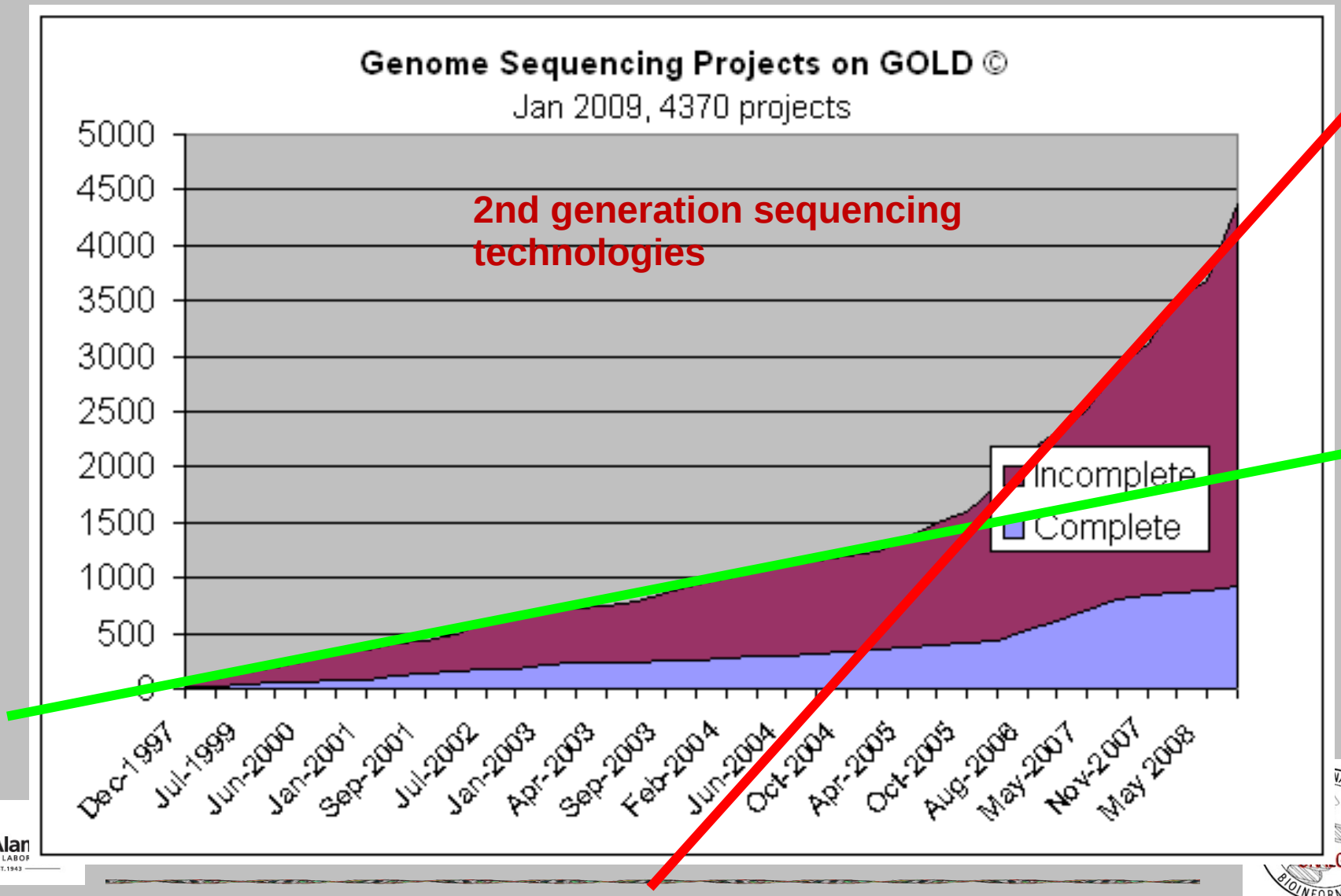
1998 STD Sequence

1999 Oral Pathogen Sequence

Growth of Sequence Databases

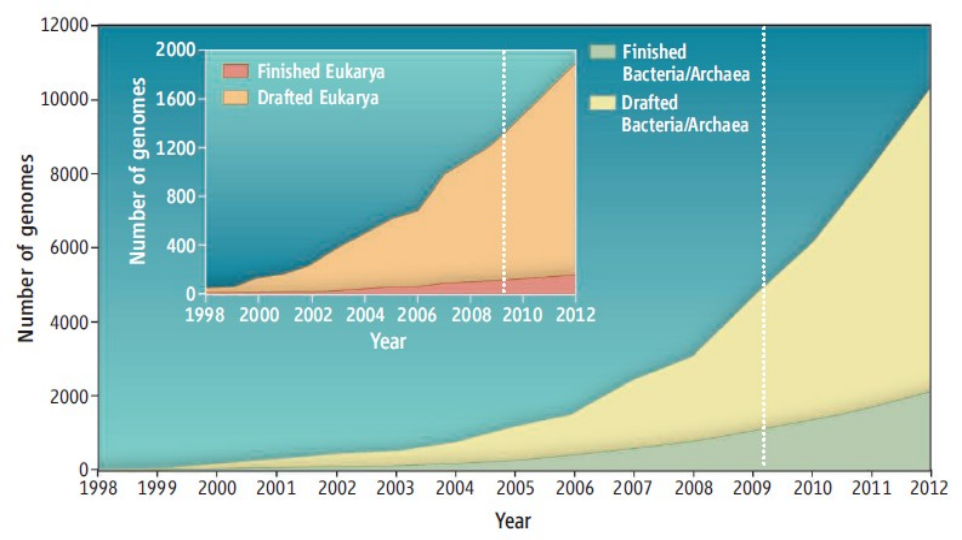


Genome projects: A growing problem

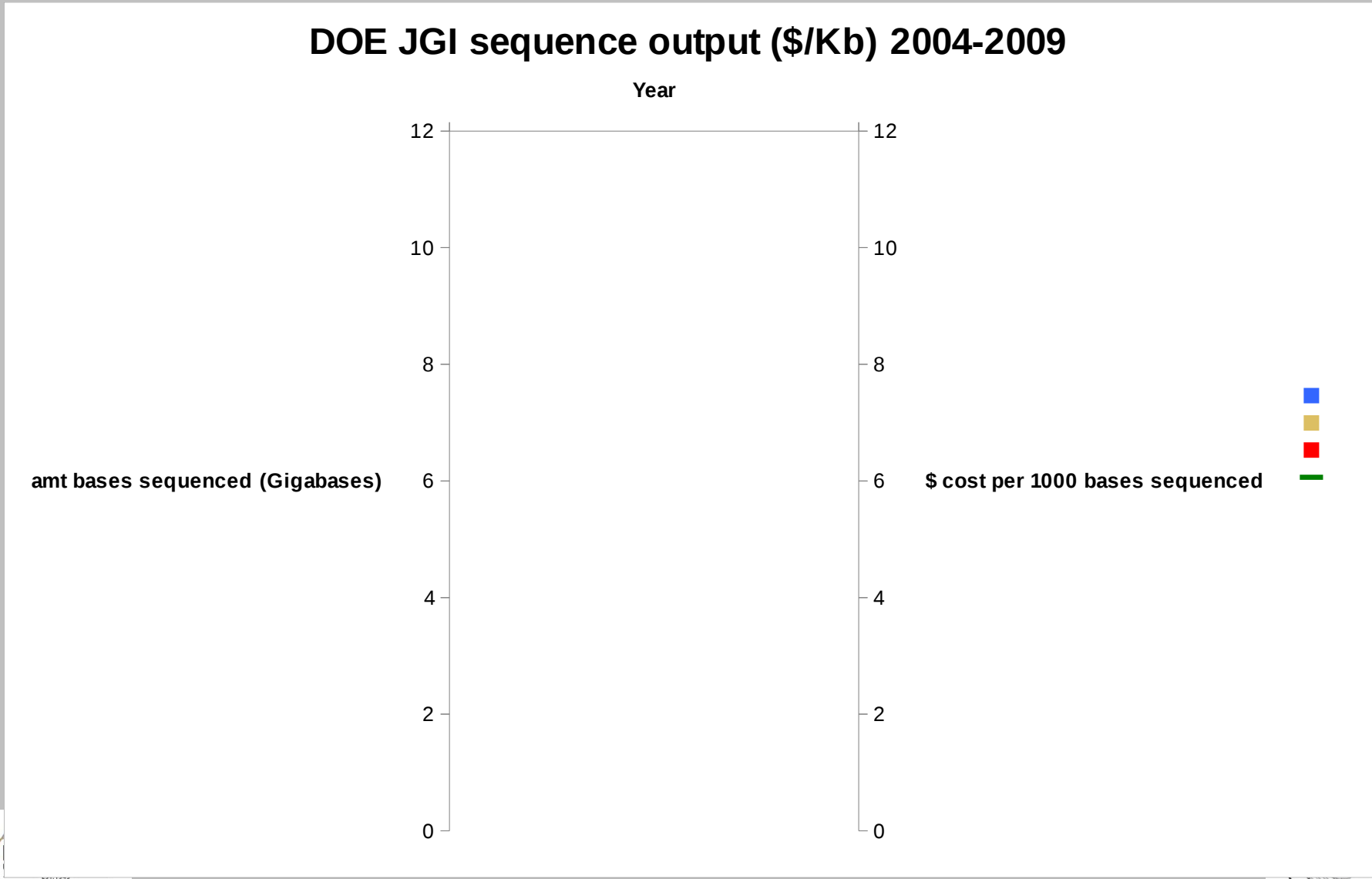


Behind the Surge: NGS Technologies

§ **The advent of 2nd generation sequencing technologies has revolutionized the genomics field!**



Driving the cost down: a new way to do business…



When “more” isn’t just more!

- Sanger - 1975
- ABI gel “automated” - 1986
- ABI Capillary 1999 – current

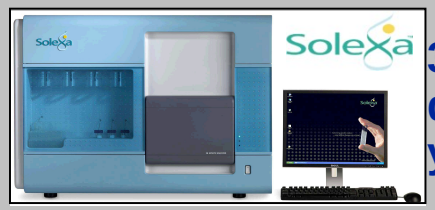
Capillary Based Sequencer, 70 kb / run

2005

½ day

30 > 100 > 400 mb / run
100bp > 250bp > 400bp
Pyrosequencing

2007



3 days

1.0 > 3 > 10 gb / run
25bp > 50bp > 75bp
Seq. by Synthesis

2008

7 days

1.0 > 10 > 20 gb / run
25bp > 35bp > 50bp
Seq. by Ligation

2010

What’s

More on this at NGS workshop!! 11:45am Fri. Room 133

Oral pathogen sequence repository

www.oralgen.lanl.gov

Oralgen Databases - Windows Internet Explorer

http://oralgen/

File Edit View Favorites Tools Help

Infranet Controller | Oralgen Databases

Oral Pathogen Sequence Databases

Home

Search

Search Oralgen Google

Oral Bacteria

- Actinobacillus actinomy
- Actinomyces naeslundii
- Fusobacterium nucleatu
- Fusobacterium nucleatu
- Porphyromonas gingiva
- Porphyromonas gingiva
- Prevotella intermedia
- Streptococcus gordonii
- Streptococcus mitis
- Streptococcus mutans
- Streptococcus sanguinis
- Tannerella forsythensis
- Treponema denticola
- Oral Plasmid Collection

Los Alamos NATIONAL LABORATORY

Bioscience Division

Oralgen Survey:
LANL_Oralgen_Questionnaire_082009-1.doc
LANL_Oralgen_Questionnaire_082009-1.pdf

Oral Pathogen Sequence Databases | Oralgen NEWS | Help Desk

These specialized databases are funded by the National Institute of Dental and Craniofacial Research (NIDCR) within the National Institutes of Health, Bethesda Maryland.

The scope of the project includes molecular information pertaining to oral pathogens, bacterial and viral. The Oralgen project was renewed in May 2006, thus we anticipate continuing expansion over the coming five years. A full statement of our goals is forthcoming. In general we strive to provide one-stop shopping in which analysis and compilation go hand in hand. Currently, the included databases comprise:

Oral Bacteria

- Actinobacillus actinomycetemcomitans
- Actinomyces naeslundii
- Fusobacterium nucleatum
- Fusobacterium nucleatum polymorphum
- Porphyromonas gingivalis W83
- Porphyromonas gingivalis ATCC 33277

Done | Local intranet | 100%

Goal to Maintain/Improve an Analytical Resource

- **Conduct Specialized Annotation and Other Analyses:**
  - ***Hypothetical protein ranking for P. gingivalis and S. mutans***
  - ***Protein cellular localization prediction***
  - ***Metabolic pathways and transport capability analysis***
  - ***Recent gene duplication, insertion sequence, conjugative transposons, and "Genomic Island" predictions***
  - ***Small, non-coding RNAs (sRNAs) prediction***
  - ***Phylogenetic fingerprints***
  - ***etc.***
- **Advanced searching capabilities**
- **Provide comparative genomic tools (like StepToto DB)**
- **Links to other oral microbiome/pathogen resources**

New site up – slow and steady…

http://hemisphere.lanl.gov/oralgen-tng/

File Edit View Favorites Tools Help

Oralgen

Go To Oralgen Classic

Oral Genomic and Metagenomic Database

Search

Search Oralgen Google

My Oralgen | Search | Tools | Community | Help

Genome Sequences

- Oral Bacteria
- Near Neighbors
- Draft Genomes
- Herpes Virus Database

Metagenome Sequences

- Oral 16s rRNA
- Shotgun Metagenome
- DACC Reference Genomes

Downloads

- Sequences
- Supplemental Files

Analysis

- Metabolic Pathways
- ABC Transporters
- Small Noncoding RNAs
- Insertion Sequence Elements
- Cellular Location Predictions
- Hypothetical Proteins Ranking
- Recent Gene Duplications

Tools

- Local BLAST Search
- PSI BLAST Search
- COGs Search
- InterProScan
- ProDom Search
- Blocks Search
- Fragment Recruitment NEW
- 16s rRNA Classifier (under construction)

External Links

- HOMD
- HMP DACC
- JGI-IMG-M
- MG-RAST

What Is New

Oralgen News
Oralgen Survey

Job Opening

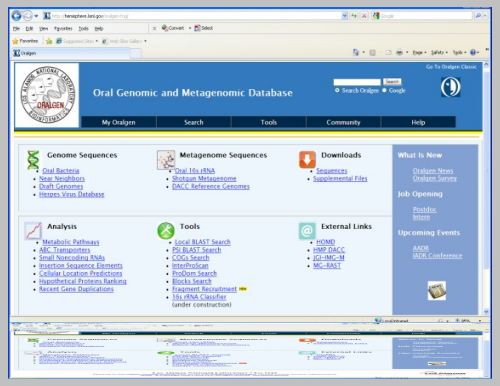
Postdoc
Intern

Upcoming Events

AADR
IADR Conference

New technologies, new challenges!!

Oralgen v2.0



§ **Keeping up!**

- How to feed the sequencing monsters
- How to handle the data (store, analyze)

§ **What to do with a few hundred thousand reads?**

tens of Million s

- ***de novo* sequencing (and finishing!)**

How quickly things change…

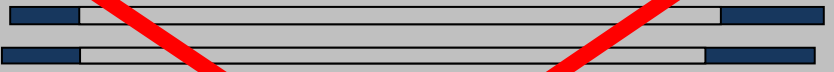
Sanger libraries

Paired ends

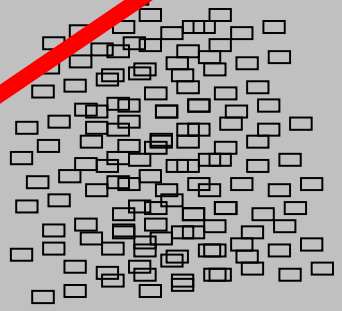
3 kb

6-8 kb

and 40kb

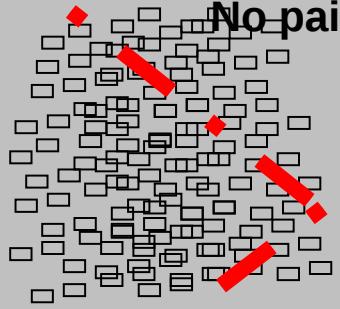


454 libraries (150-450 bp No paired ends)

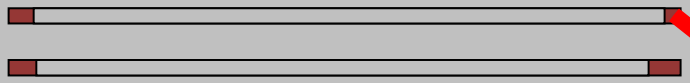


New Tech

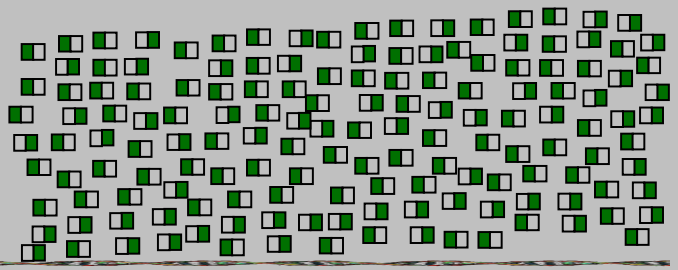
454 standard libraries (400 bases long No paired ends)



454 Paired ends (Paired ends with reads of 150 bp - Average insert 15-20 kb)

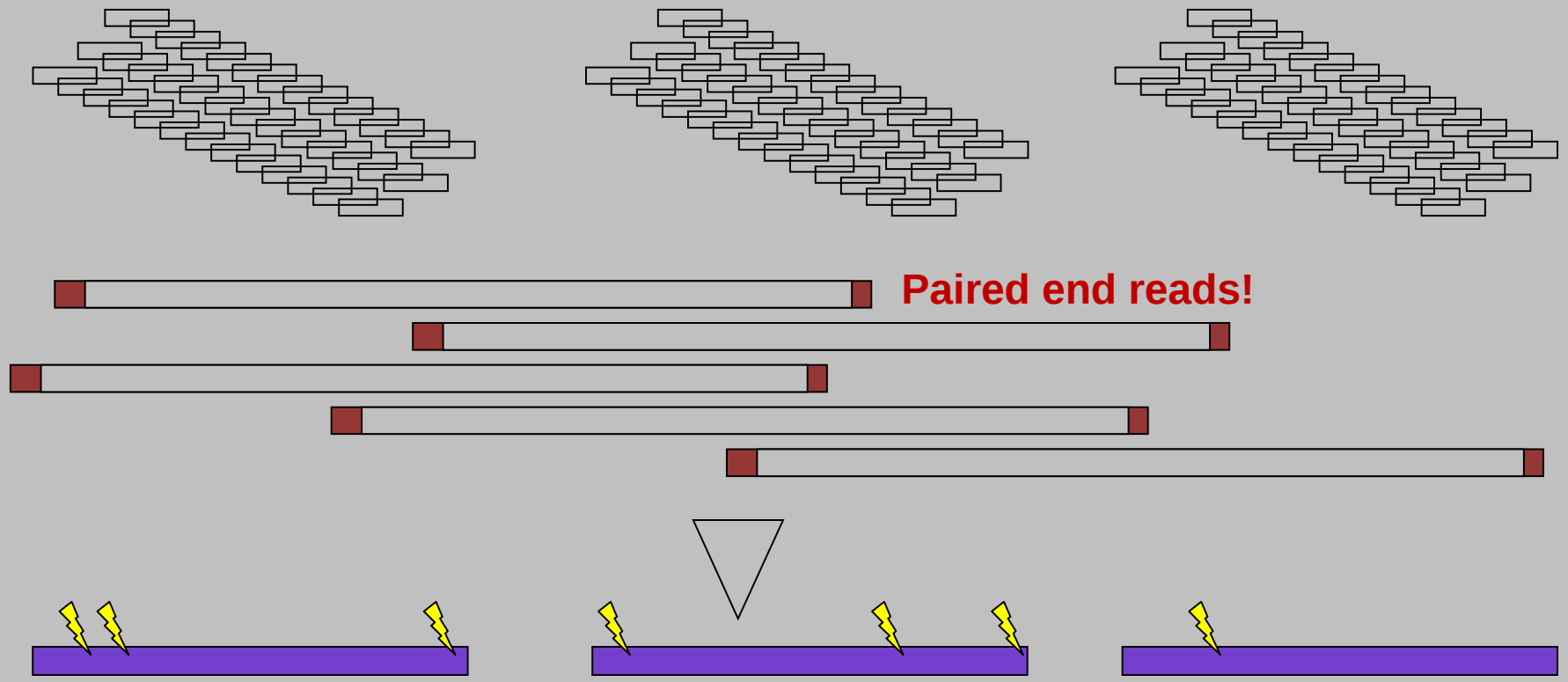


Solexa (Illumina) 150 bp inserts with reads of 36 bp (one end)

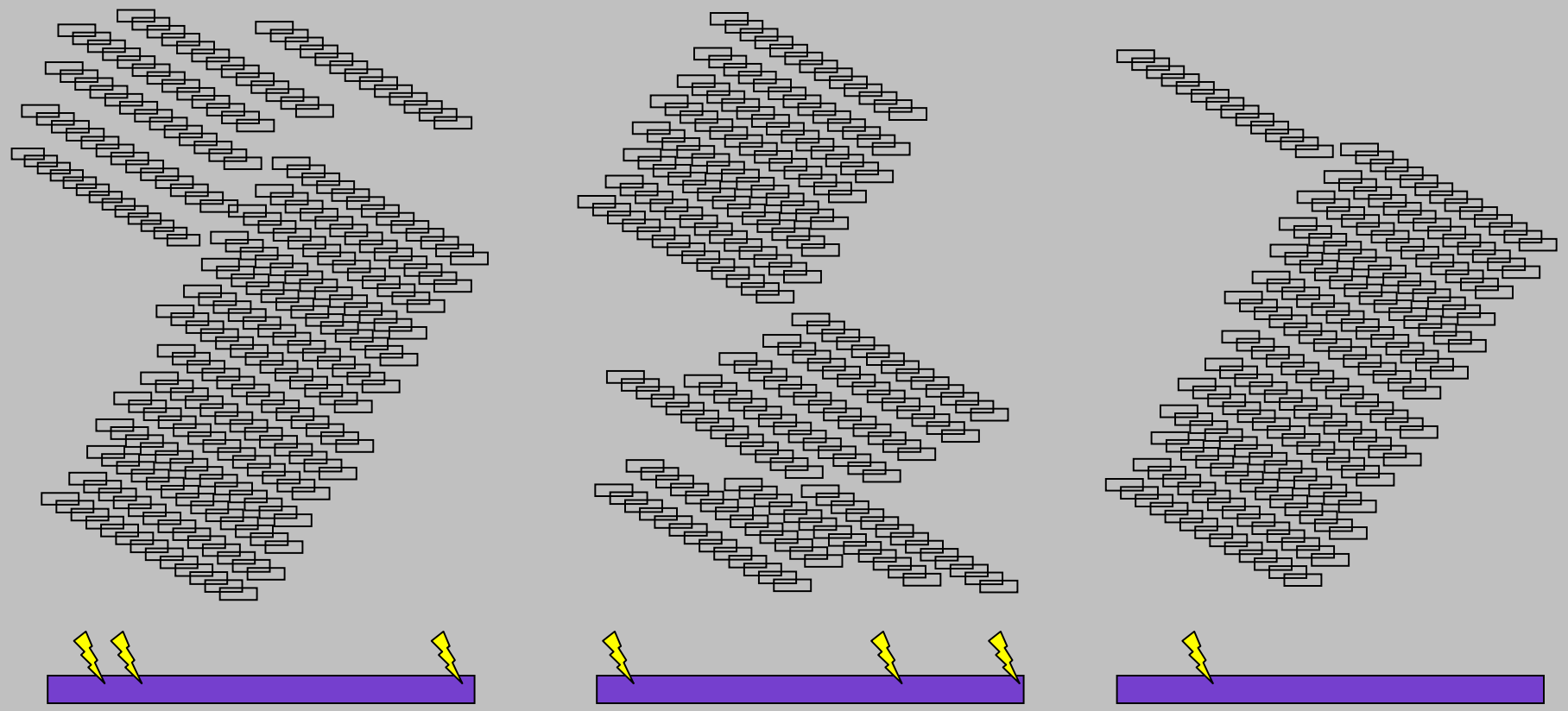


Genome Assembly with NGS technology

Short reads – but many of them!



Genome Assembly with NGS technology



Contigs (consensus sequences)

****Corrected sequences, but still many gaps generally**

Improved NGS assembly

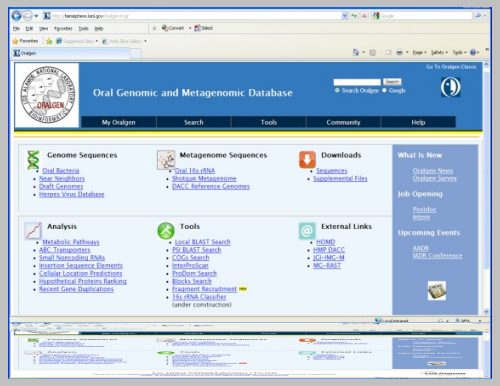
Kingella kingae	Original assembly	LANL improved assembly
Assembled_reads:		98.03%
Singleton:		2,388
Contigs_number:	180	101
N50:	21,244	54,948
N90:	6,210	11,656
Max:	76,431	188,302
Min:	103	110
Total_bases:	1,942,587	2,012,773

Input
↓
QC
↓
Filter
↓
Trim
↓
Assemble
↓
Merge data

Can also suggest primers for genome closure (ie. Finishing)

New technologies, new challenges!!

Oralgen v2.0



§ Keeping up!

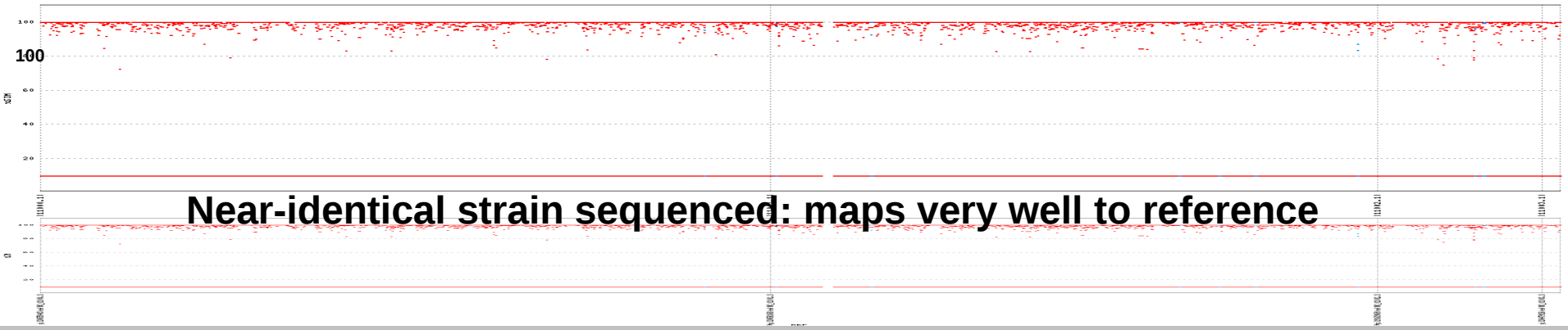
- How to feed the sequencing monsters
- How to handle the data (store, analyze)

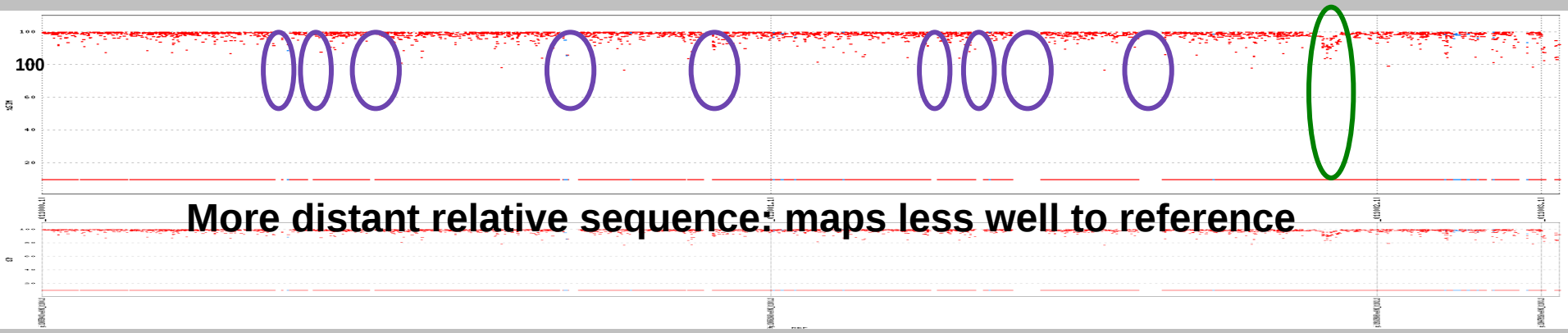
§ What to do with a few hundred thousand reads?

tens of Million s

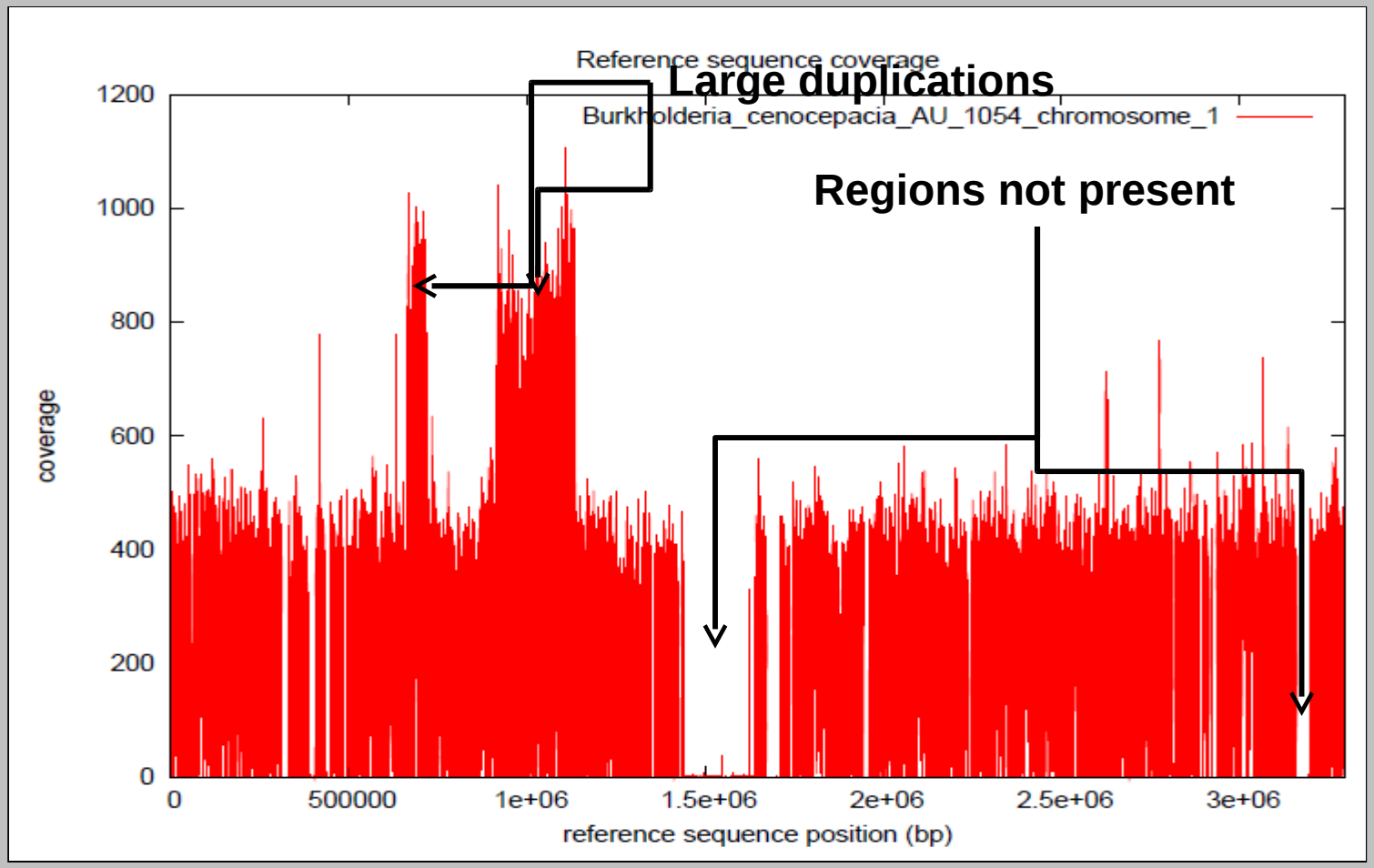
- *de novo* sequencing (and finishing!)
- **Re-sequencing and Transcriptomics (RNAseq)**

Mapping reads to a reference sequence

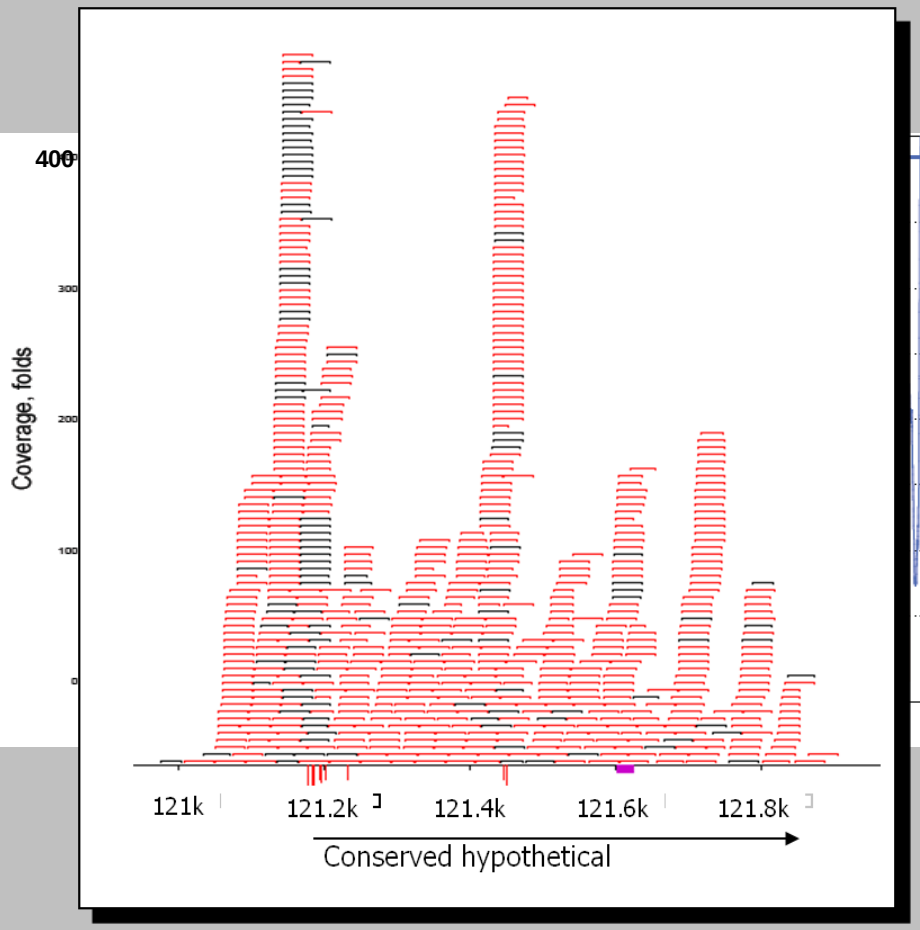


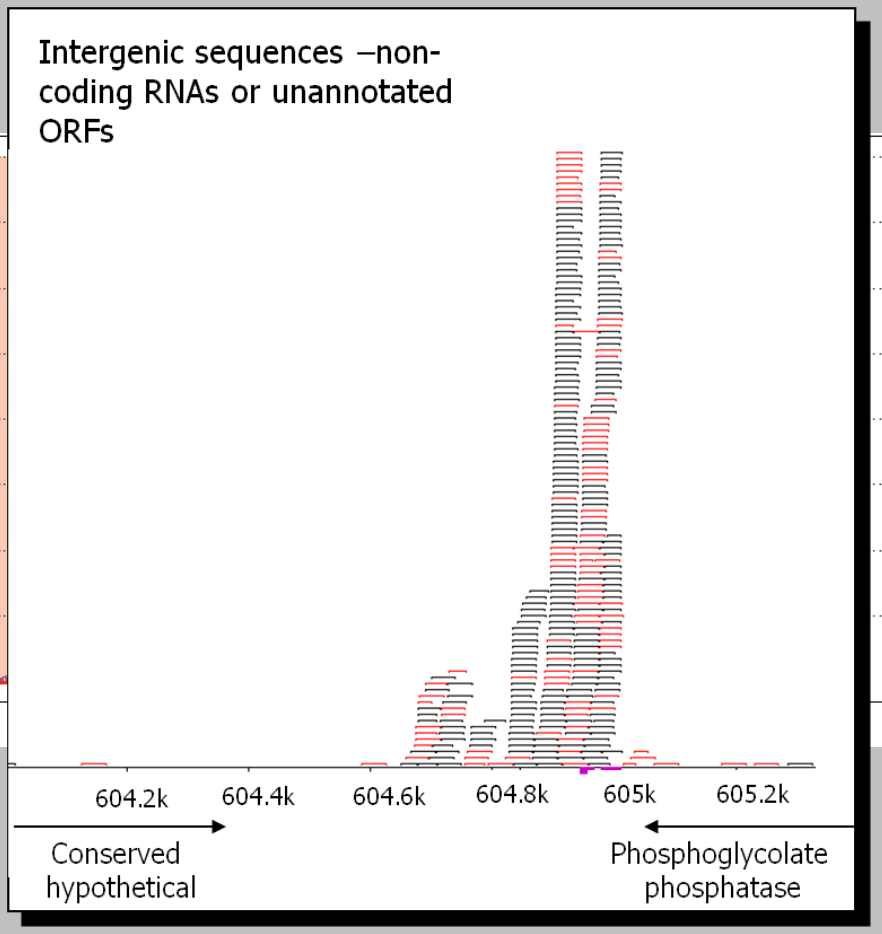


Uneven coverage revelations



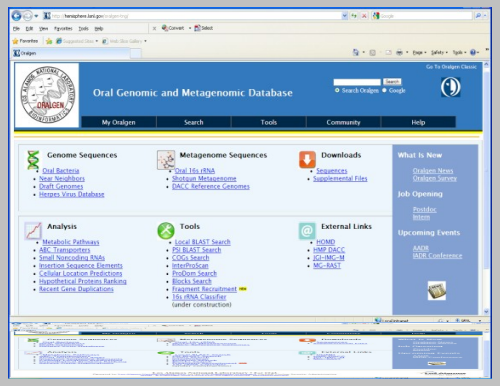
Transcriptome (RNA-sequencing)





New technologies, new challenges!!

Oralgen v2.0



§ Keeping up!

- How to feed the sequencing monsters
- How to handle the data (store, analyze)

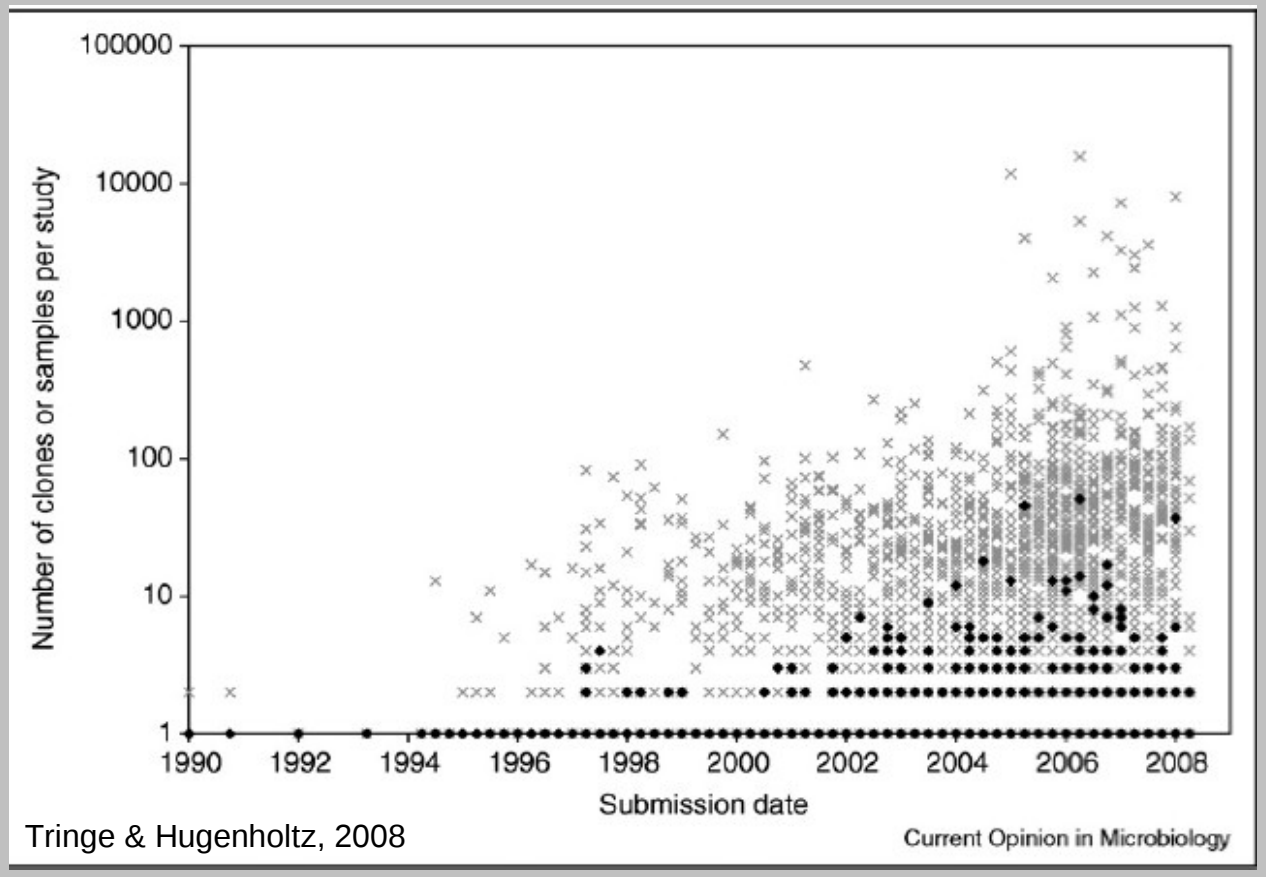
§ What to do with a few hundred thousand reads?

tens of Million s

- *de novo* sequencing (and finishing!)
- Re-sequencing and Transcriptomics
- **Metagenomics (both rRNA surveys and shotgun "environmental" sequencing)**

Metagenomic (16S) population surveys

"Pyrotags" gaining the upper hand: 200,000 to 1,000,000 reads per run

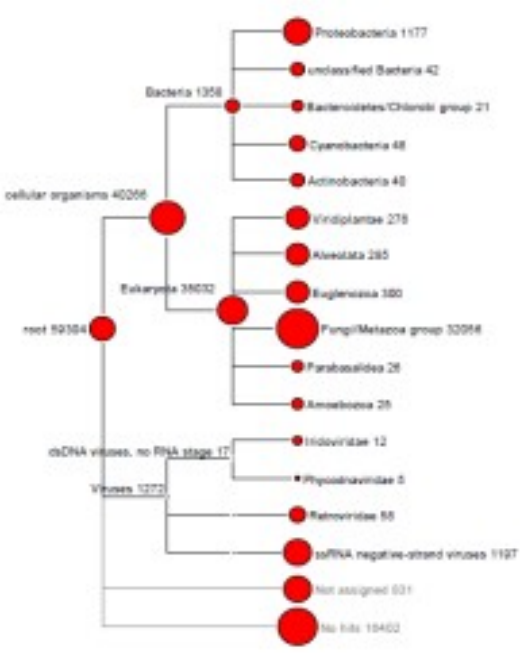


Tringe & Hugenholtz, 2008

- **Much higher resolution than Sanger 16S surveys**
- **Has highlighted "rare biosphere"**
- **May even allow saturation of diversity in a given habitat**

Improving Classification Tools

- **RDP Naïve Bayesian Classifier:** a very fast and accurate taxonomy assignment algorithm
- LANL has applied a Markov model to enhance accuracy while maintaining speed
- Also, the LANL Markov NBC will be trained on oral microbes so may be more accurate
  - Will provide on Oralgen2.0!

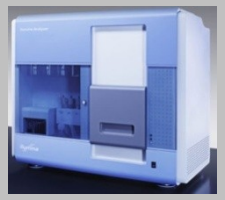


However: only tells us who is there…

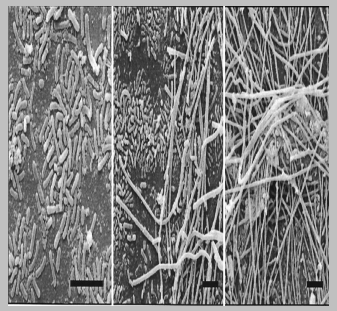
not what they are doing…

“Next Gen” metagenomics: Sequencing technology changing the landscape

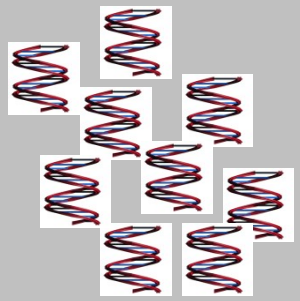
new way



Sample

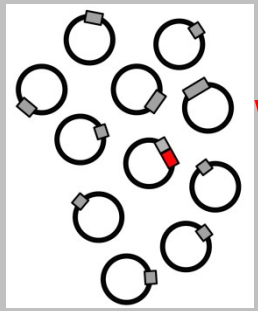


Extract DNA



old

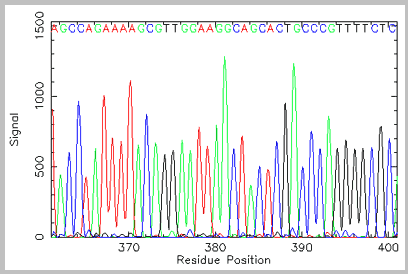
Shotgun clones



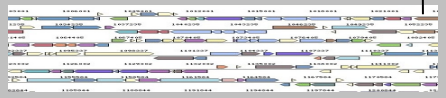
3, 8, 40 kb

way

High throughput sequences



Assemble and Map reads



Gene calls/annotation

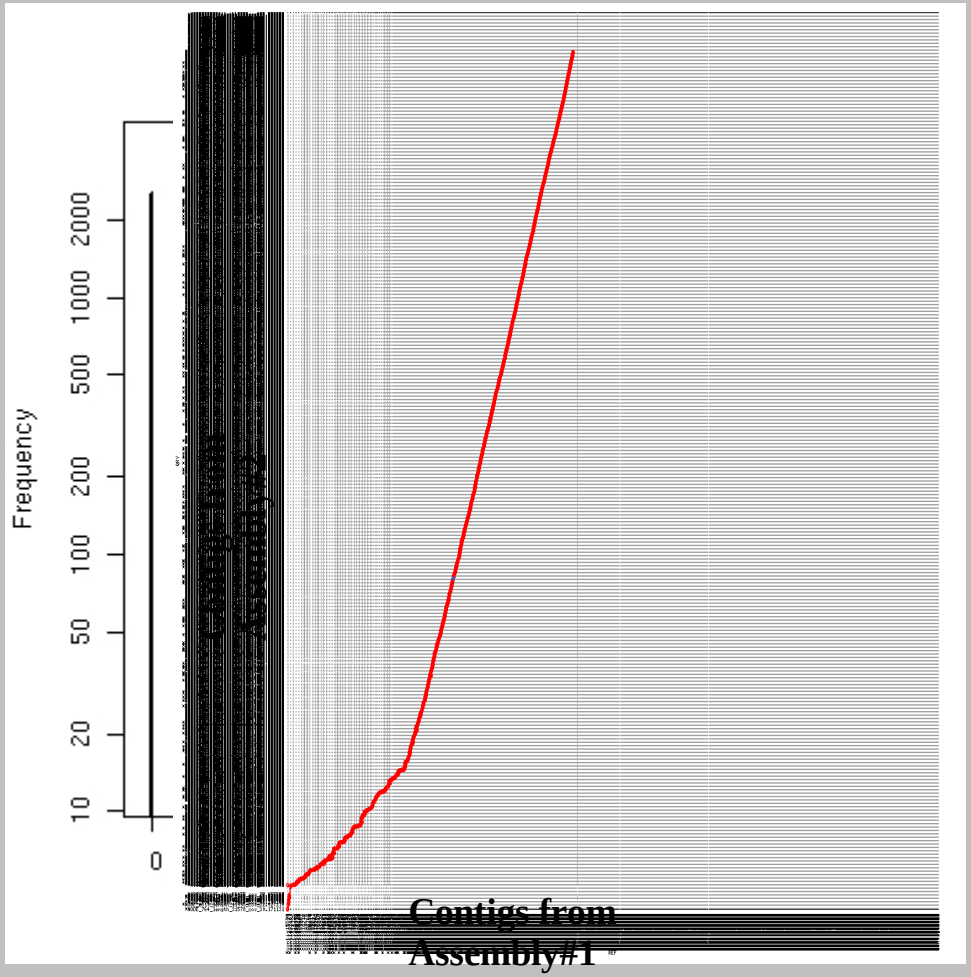
Bin fragments

Los Alamos NATIONAL LABORATORY EST. 1943

LOS ALAMOS NATIONAL LABORATORY ORALGEN BIOINFORMATICS

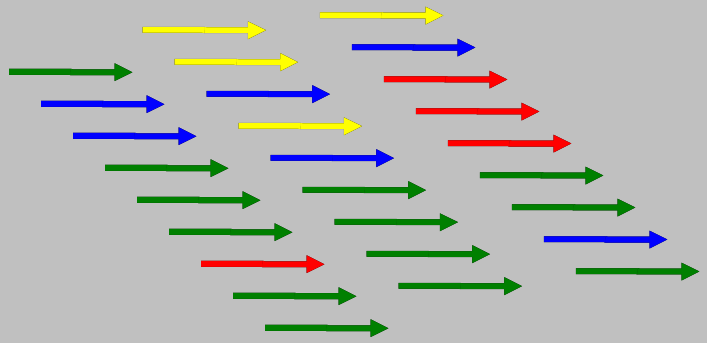
Current Impediments….

- Different methods (assembly algorithms) and different technologies result in different contigs:
  - 12,500 total (5kb max)
  - 287,000 total (30kb max)
  - For either: only ~50% of reads get assembled still!!
- Very different answers when varying parameters…



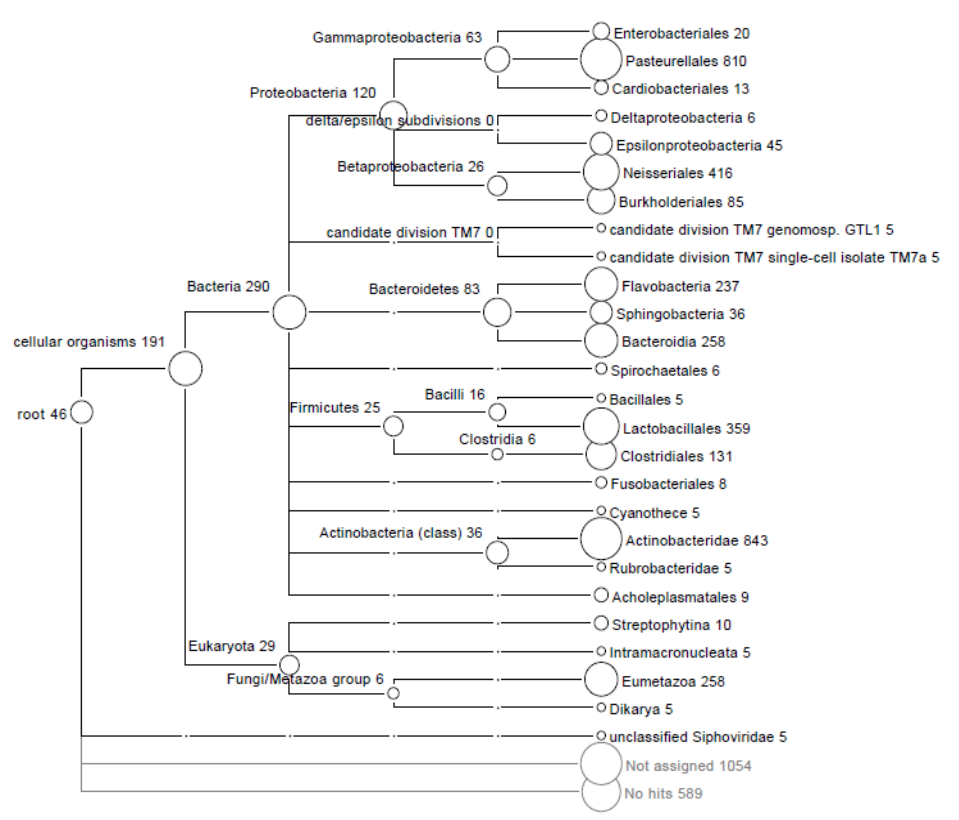
Binning and other issues….

- Binning methods do not yet work on short reads



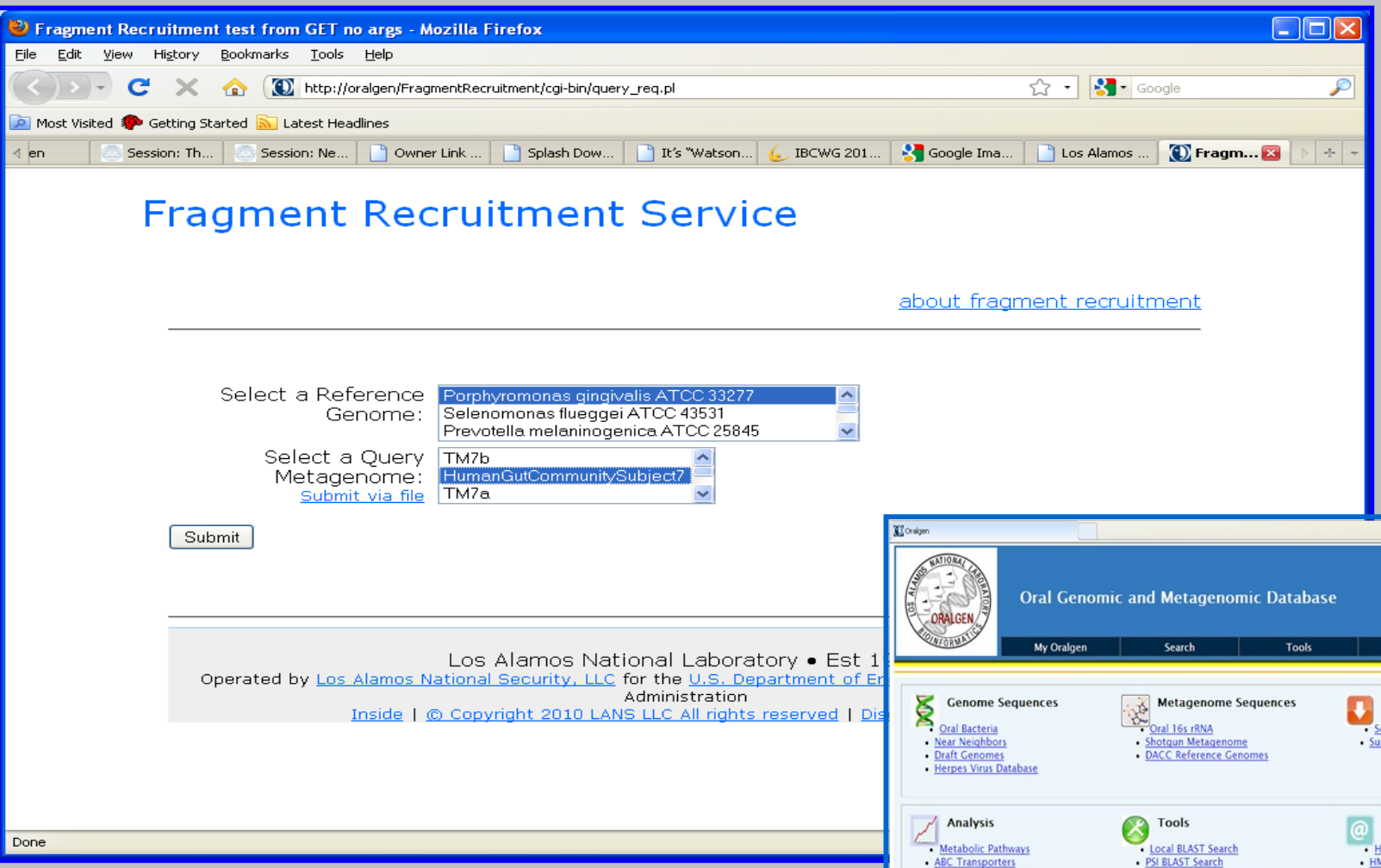
Binning and other issues….

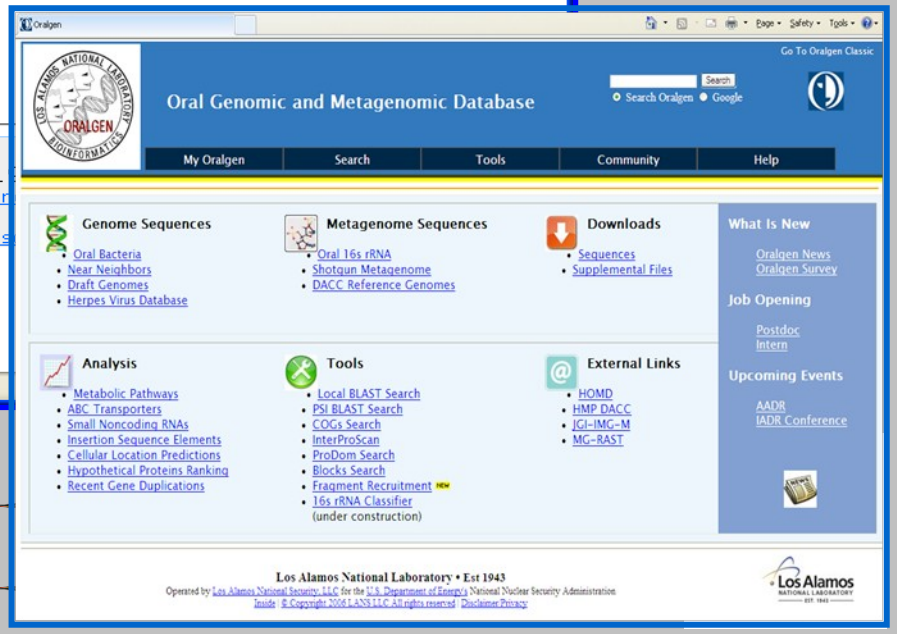
- Binning methods do not yet work on short reads
- Annotation of metagenomic contigs takes time!
- Read-based analyses take many CPU hours; so often look only at assembled contigs (ie. ~50% of the data)



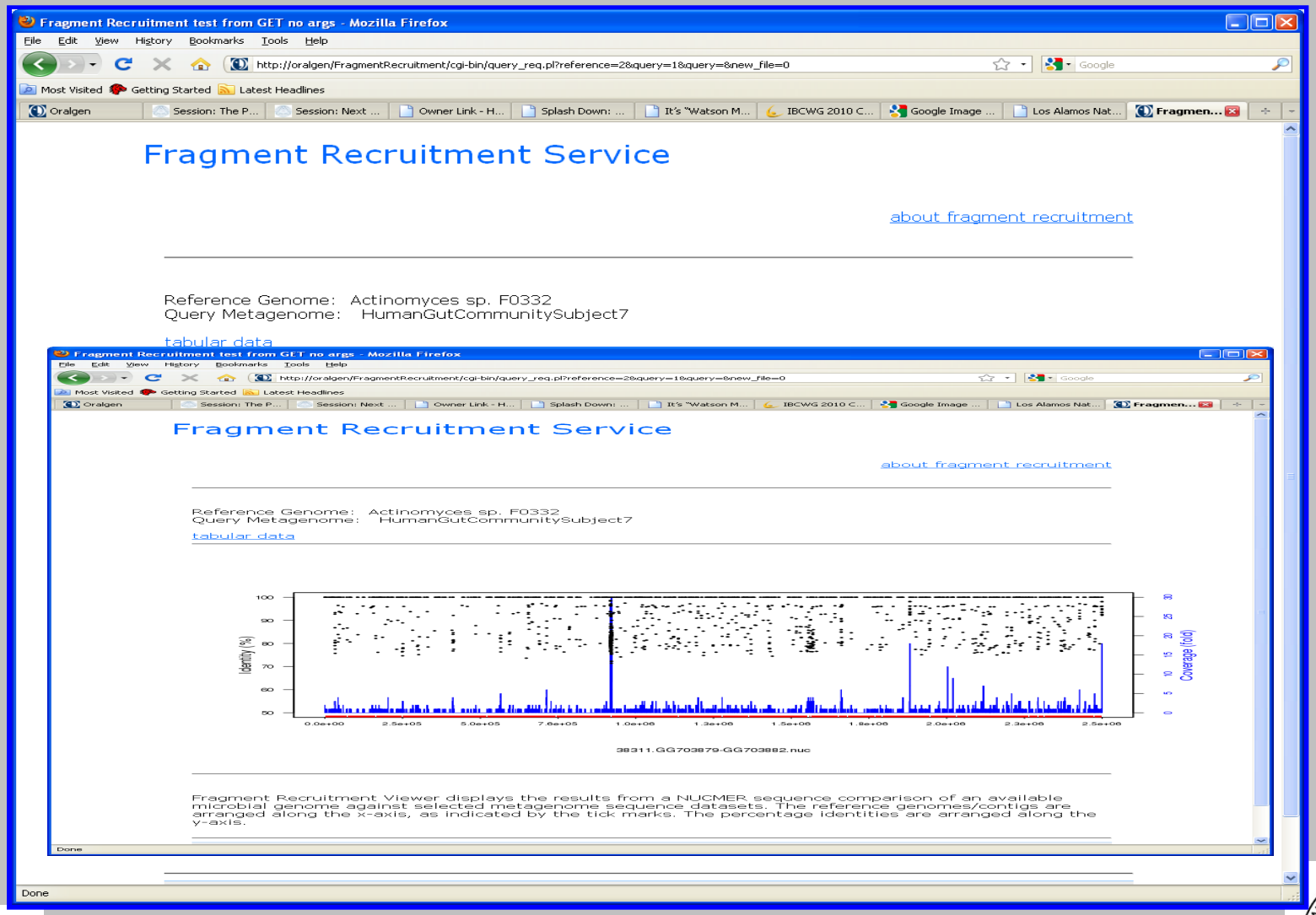
Can one reduce complexity of the data/sample?

Binning via sequence recruitment…

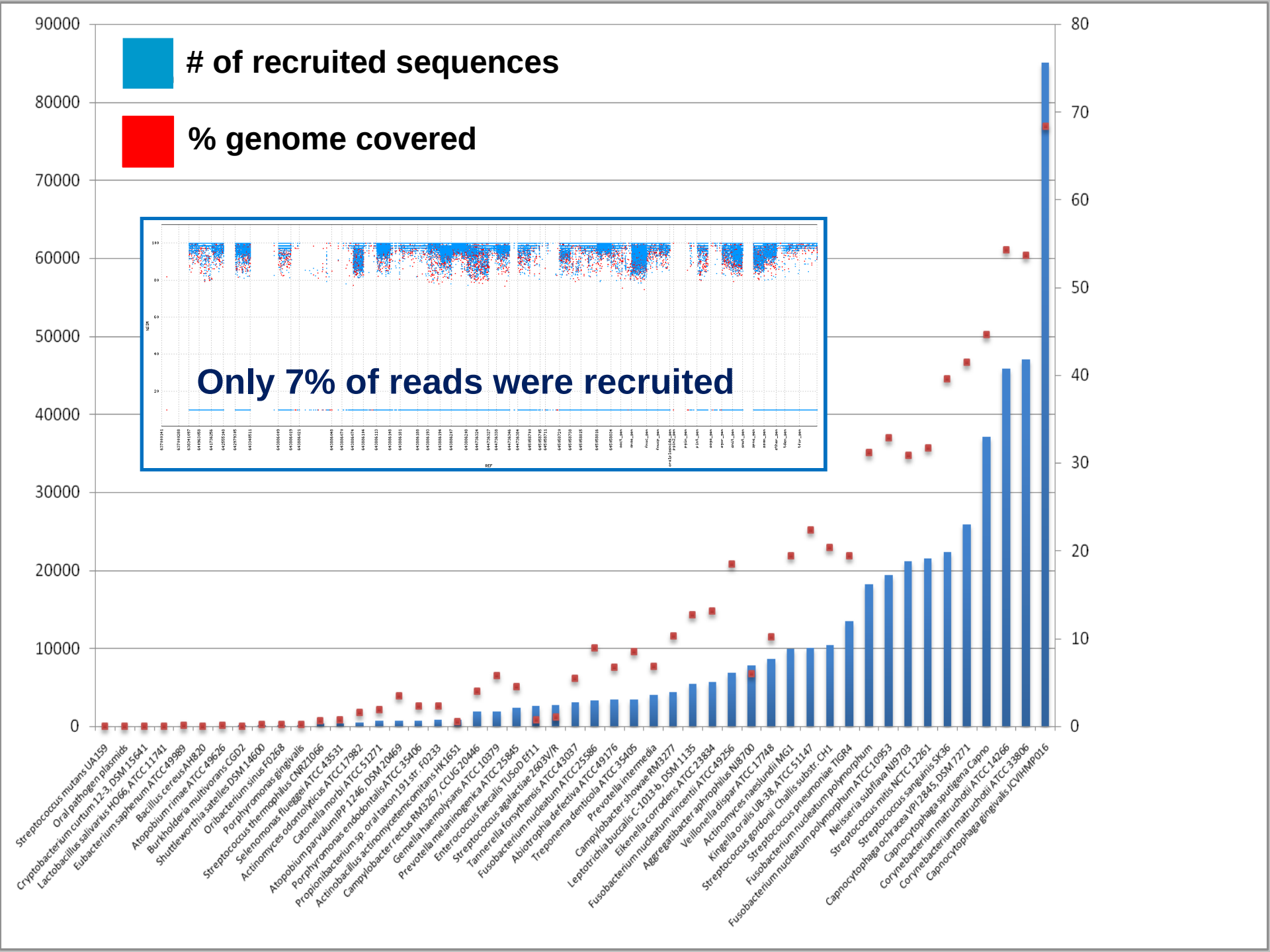




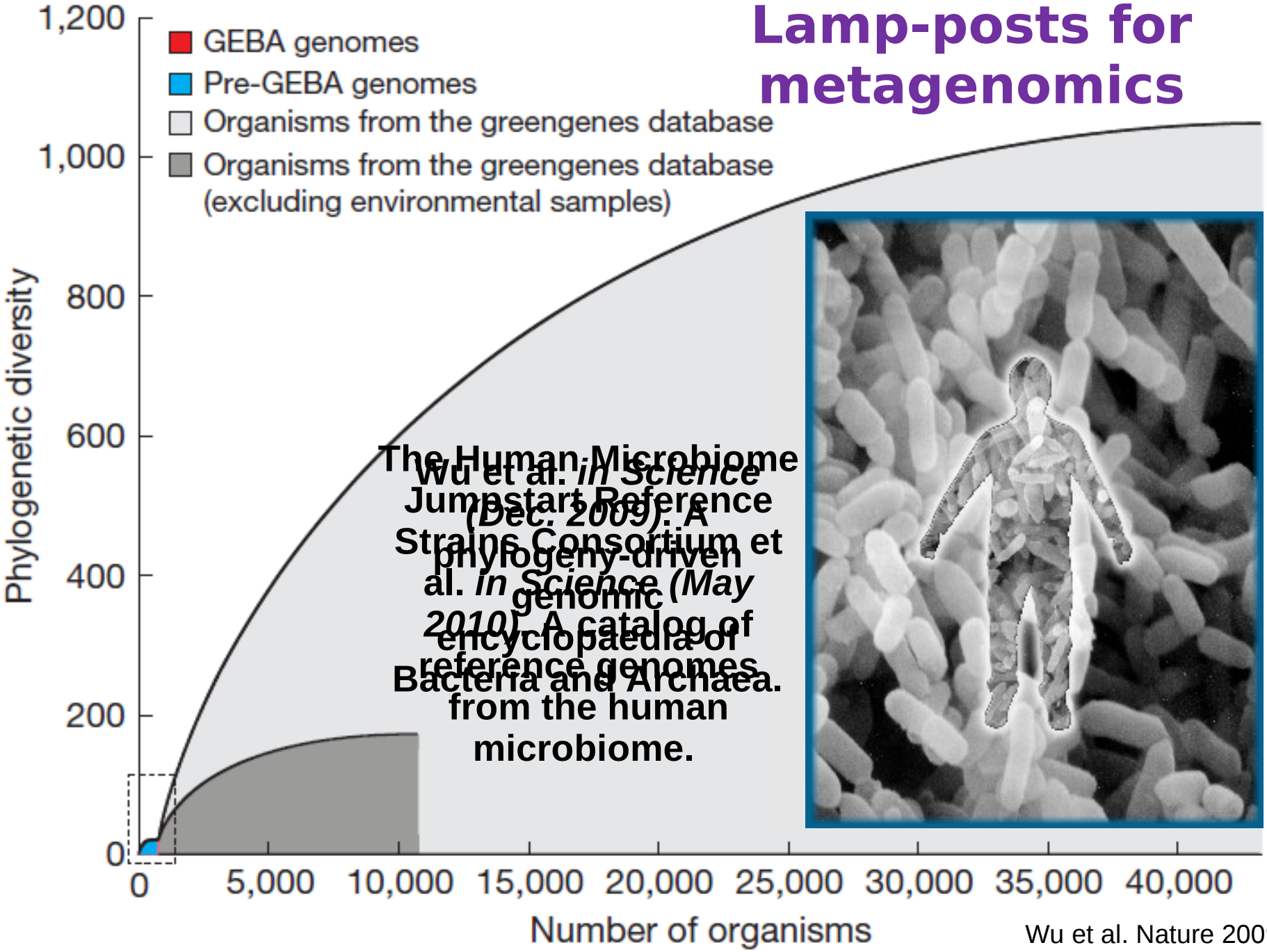
What, of who, is present…



Can “recruit” reads and contigs to many reference genomes


# of recruited sequences
% genome covered
Only 7% of reads were recruited
90000
80000
70000
60000
50000
40000
30000
20000
10000
0
80
70
60
50
40
30
20
10
0
Streptococcus mutans UA159
Oral pathogen plasmids
Cryptobacterium curtum 12-3, DSM 15641
Lactobacillus salivarius HO66, ATCC 11741
Eubacterium saphenum ATCC 49989
Bacillus cereus AH820
Atopobium rimae ATCC 49626
Burkholderia multivorans CGD2
Shuttleworthia satelles DSM 14600
Oribacterium sinus F0268
Porphyromonas gingivalis
Streptococcus thermophilus CNRZ1066
Selenomonas flueggei ATCC 43531
Actinomyces odontolyticus ATCC 17982
Catonella morbi ATCC 51271
Atopobium parvulum PP 1246, DSM 20469
Porphyromonas endodontalis ATCC 35406
Propionibacterium sp. oral taxon 191 str. F0233
Actinobacillus actinomycetemcomitans HK1651
Campylobacter rectus RM3267, CCUG 20446
Gemella haemolysans ATCC 10379
Prevotella melaninogenica ATCC 25845
Enterococcus faecalis TUSoD Ef11
Streptococcus agalactiae 2603V/R
Tannerella forsythensis ATCC 43037
Fusobacterium nucleatum ATCC 25586
Abiotrophia defectiva ATCC 49176
Treponema denticola ATCC 35405
Prevotella intermedia
Campylobacter showae RM3277
Leptotrichia buccalis C-1013-b, DSM 1135
Eikenella corrodens ATCC 23834
Fusobacterium nucleatum vincentii ATCC 49256
Aggregatibacter aphrophilus NJ8700
Veillonella dispar ATCC 17748
Actinomyces naeslundii MG1
Kingella oralis UB-38, ATCC 51147
Streptococcus gordonii Challis substr. CH1
Streptococcus pneumoniae TIGR4
Fusobacterium nucleatum polymorphum ATCC 10953
Fusobacterium nucleatum polymorphum ATCC 14266
Neisseria subflava NJ9703
Streptococcus mitis NCTC 12261
Streptococcus sanguinis SK36
Capnocytophaga ochracea VPI 2845, DSM 7271
Capnocytophaga sputigena Capno
Corynebacterium matruchotii ATCC 33806
Corynebacterium matruchotii ATCC 14266
Capnocytophaga gingivalis JCVIHMP016

Lamp-posts for metagenomics



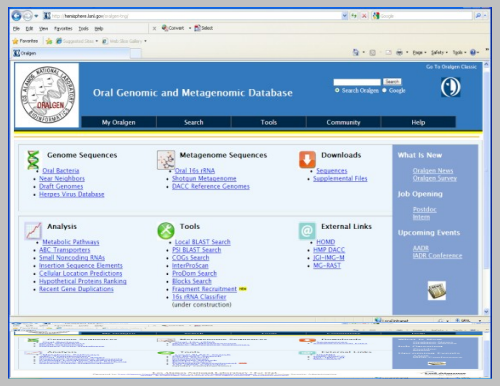
Wu et al. *in Science (Dec. 2009)*: A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea.

The Human Microbiome Jumpstart Reference Strains Consortium et al. *In Science (May 2010)*: A catalog of reference genomes from the human microbiome.

Wu et al. Nature 2009

New technologies, new challenges!!

Oralgen v2.0



§ Keeping up!

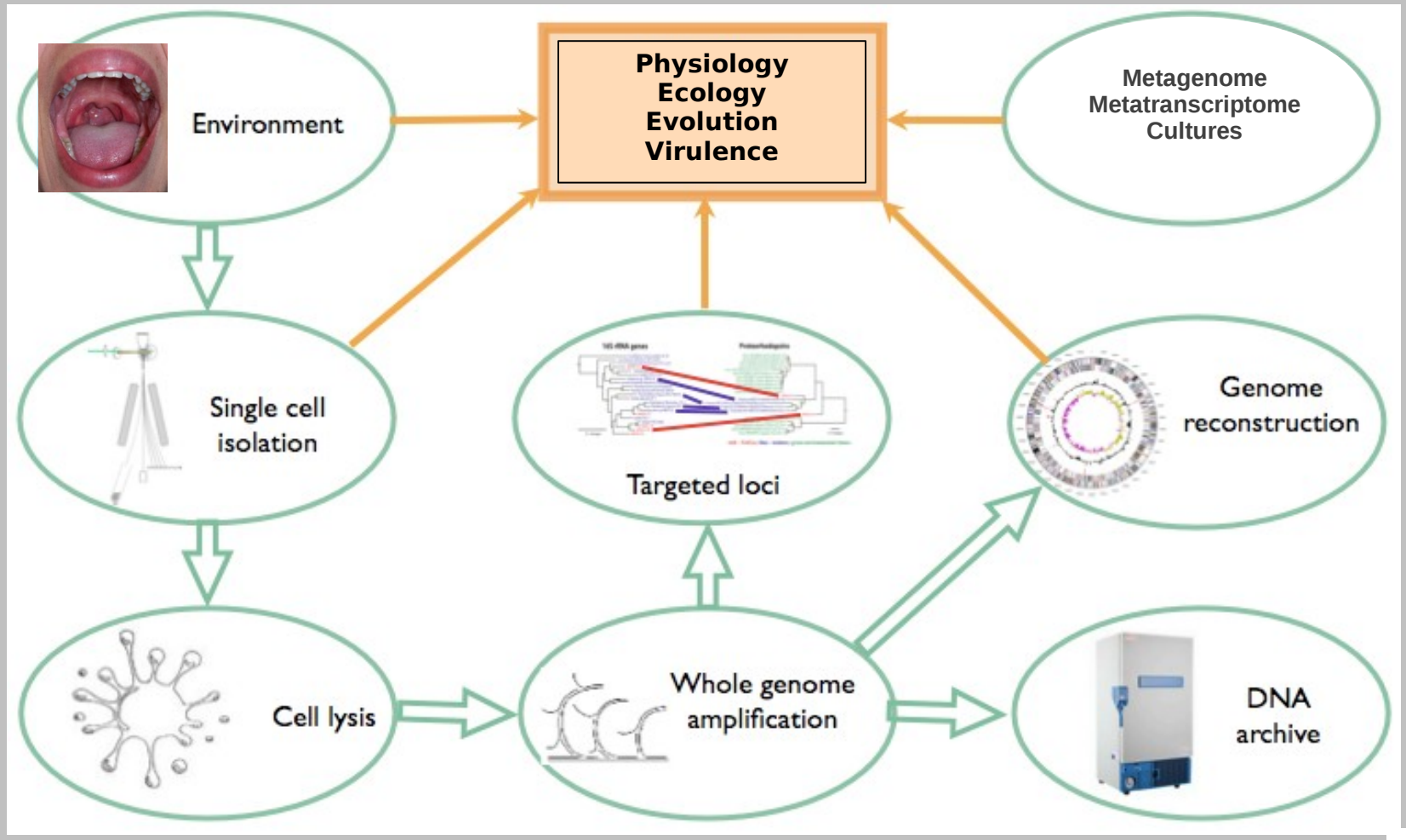
- How to feed the sequencing monsters
- How to handle the data (store, analyze)

§ What to do with a few hundred thousand reads?

tens of Million s

- *de novo* sequencing (and finishing!)
- Re-sequencing and Transcriptomics
- Metagenomics (both rRNA surveys and shotgun "environmental" sequencing)
- **Single cell genomics**

Single cell genomics



Modified from http://www.bigelow.org/files/8412/5850/8665/Pipeline.png

Dealing with NGS Technology

1. Make *de novo* and re-sequencing of oral microbiome isolates and single cells a trivial task for investigators

Assembly suite

1. Provide a tailored metagenomic population survey resource and analysis suite

Markov NB Classifier

1. Develop and provide automated and semi-automated (meta)genomics analysis tools (coupled with manual-curation/enhancement)

Refined annotations
Fragment recruitment

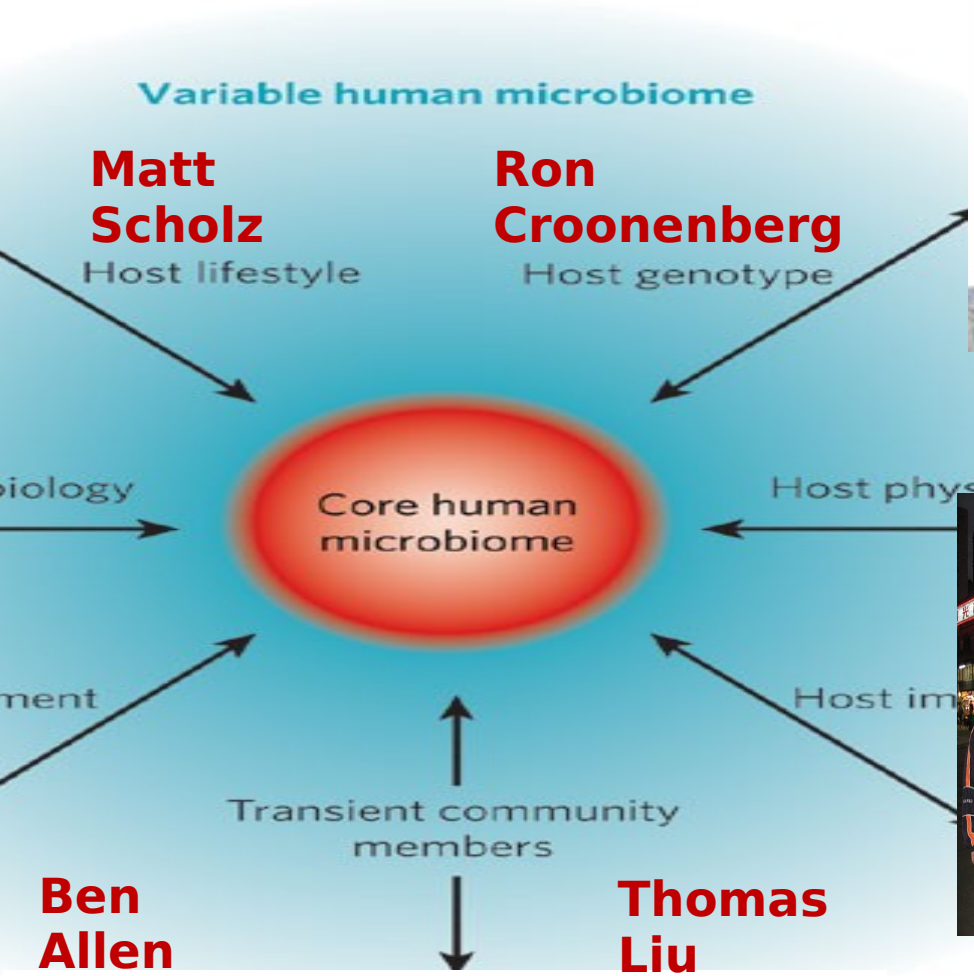
1. Integrate the above, and work with experts to provide the community with a comprehensive genomics work environment

Comprehensive database
Tailored analyses

Acknowledgements

Chien-Chi Lo

Andrew Liu



Matt Scholz

Ron Croonenberg

Pavel Senin

Ben Allen

Thomas Liu

Gary Xie

NIDCR National Institute of Dental and Craniofacial Research National Institutes of Health

Los Alamos NATIONAL LABORATORY EST. 1943

LOS ALAMOS NATIONAL LABORATORY ORALGEN BIOINFORMATICS