The DOE Joint Genome Institute

Dan Drell, Program Manager, JGI

GOAMAZON WORKSHOP

July 26, 2011

http://www.jgi.doe.gov

JGI: Present

Mission: The mission of the U.S. Department of Energy (DOE) Joint Genome Institute (JGI) is to serve the diverse scientific community as a user facility enabling the application of large-scale genomics and analysis of plants and microbes to address the DOE mission goals in bioenergy and the environment.

Exclusions: pathogens, biothreat agents, anything external to "mission"

DOE Scientific User Facility: The JGI

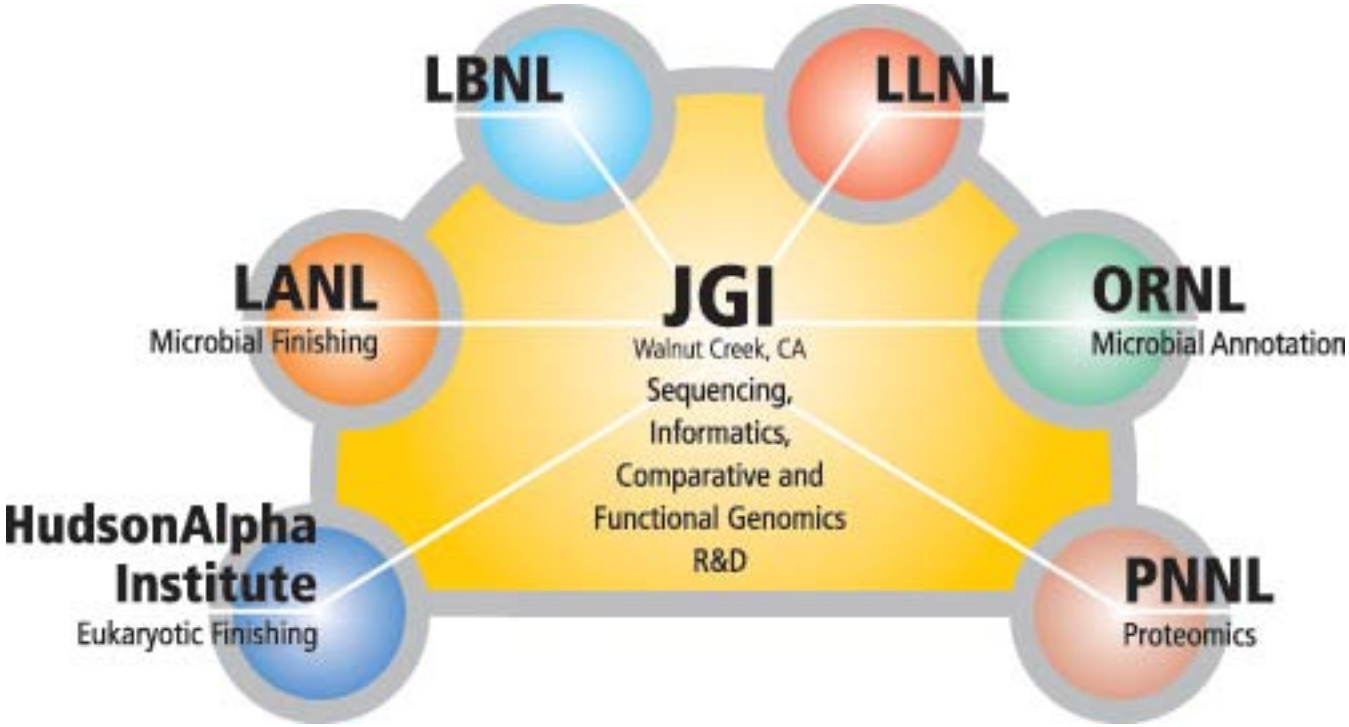
Office of Science

- **Focus on genomes and metagenomes of plants, microbes, microbial communities, fungi, and other science vital to the DOE mission**
- **Provide state-of-the-science capabilities for sequencing and analysis**
- **Maintain expert staff in a range of computing and biological research disciplines**
- **Host workshops and annotation jamborees**

High throughput sequencing line at DOE-JGI in Walnut Creek, CA

Sequencing more than 16 *trillion* base pairs (**16 Terabases**) of DNA per year**! (= 5300 1x human genomes!)**

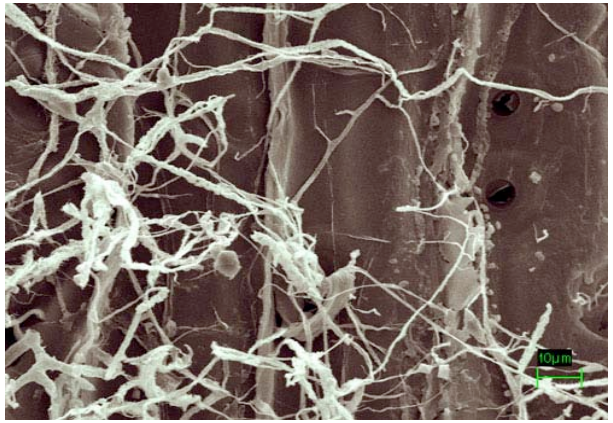
JGI Partners



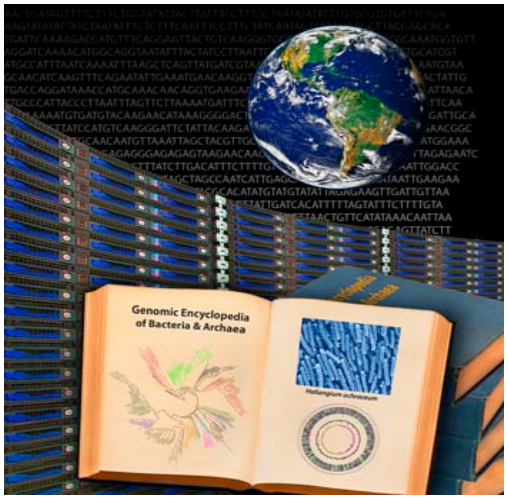
Partner management is not without its challenges!

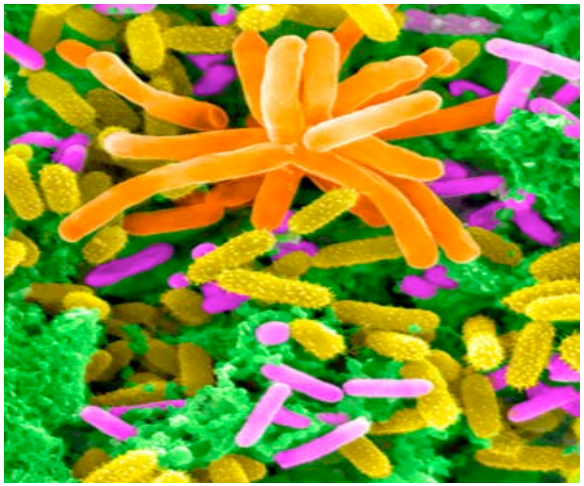
Two Programs

Plants and Fungi: Dan Rokhsar

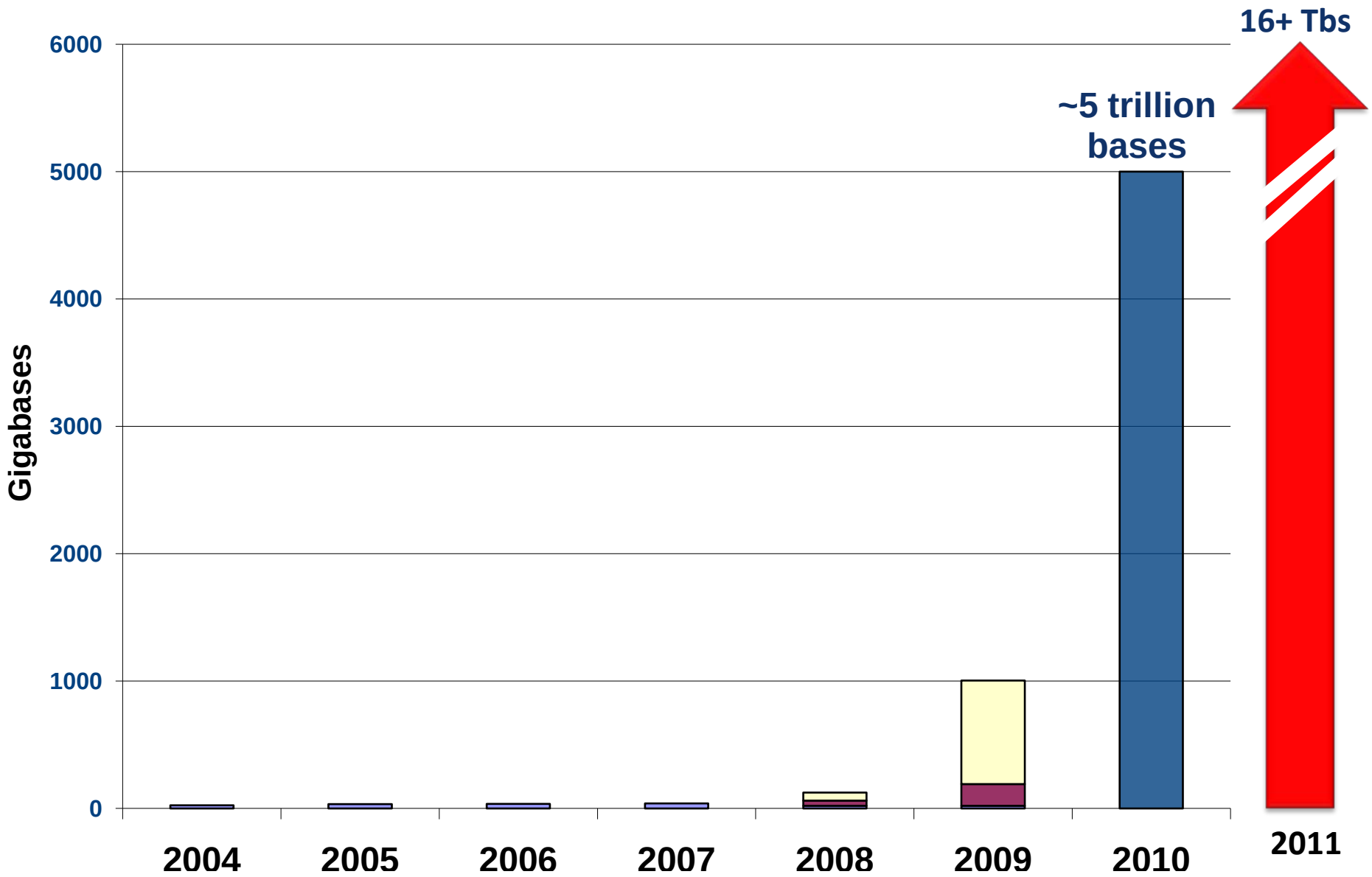


Microbes and Metagenomes: Nikos Kyrpides



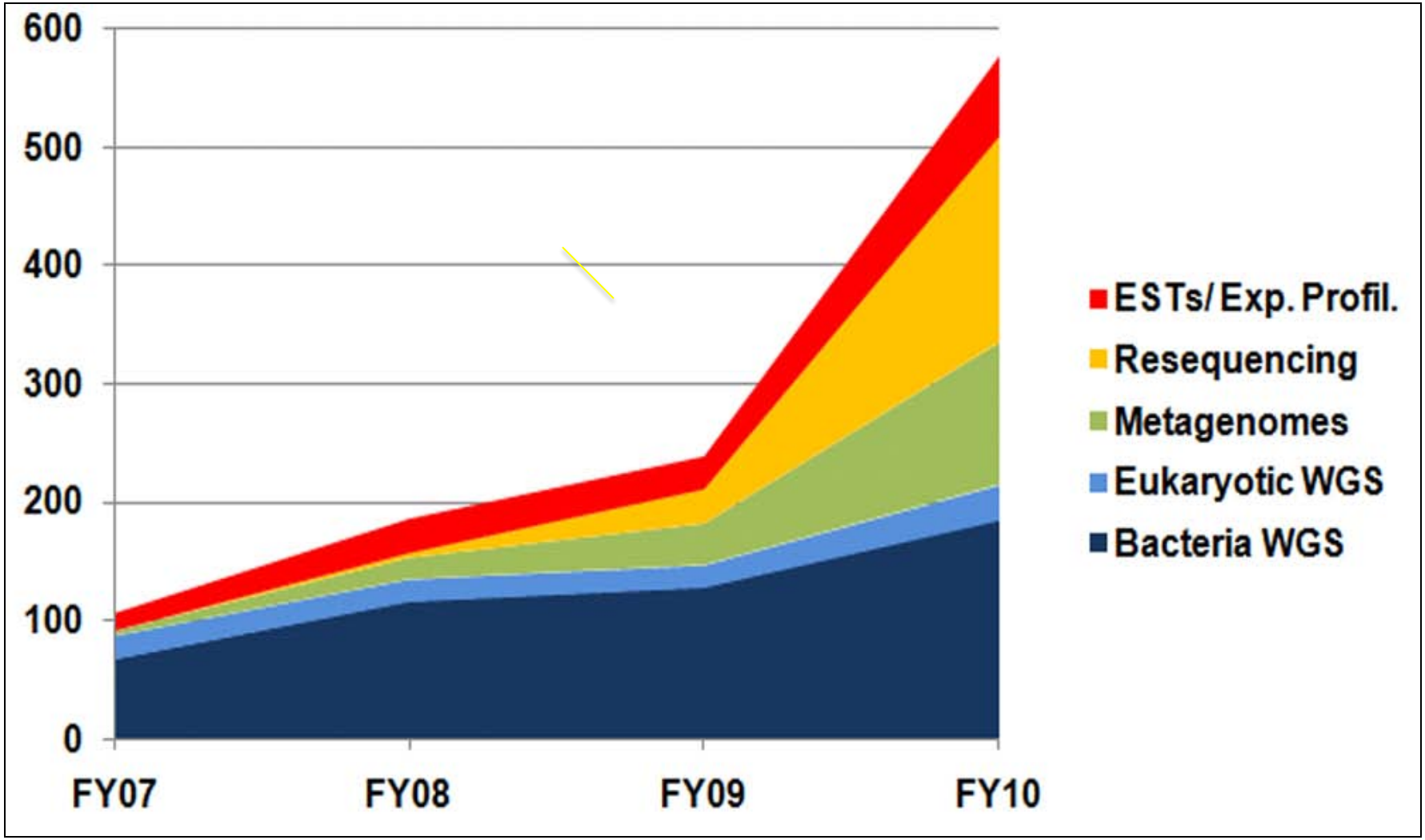


Massive Increase in Sequencing Output



600 Projects Completed FY10

Projects Types



JGI Project Programs: sequencing allocation

Community Sequencing Program:	~47%
Bioenergy Research Centers:	30%
Grand Challenges:	~9%
GEBA:	~9%
Other:	~3%
Research and Development:	~2%

JGI Sequencing Platforms

Major platforms

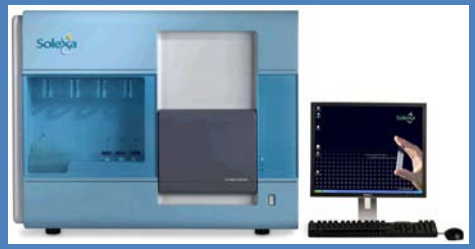
Illumina HiSeq2000 x 8

PacBio RS x 1

Supplement platforms

ABI 3730xl x 2

454 x 2



Illumina GAiix x 5

Partner Support (LANL & Hudson Alpha):

ABI 3730xl x 16 454 x 2 Illumina Hiseq x 1

Illumina HiSeq

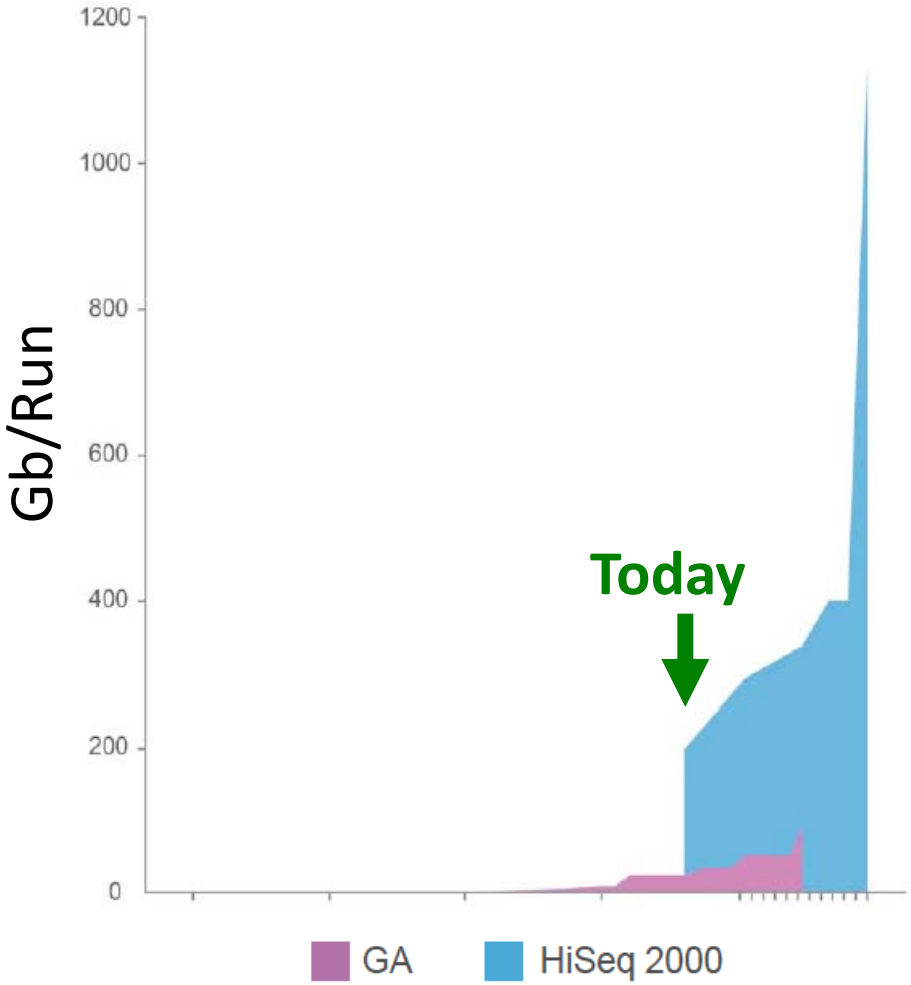
Pluses:

- Total Output: 600 Gb/Run
- Accuracy: >99%
- $0.1 /Million Bp

Minuses:

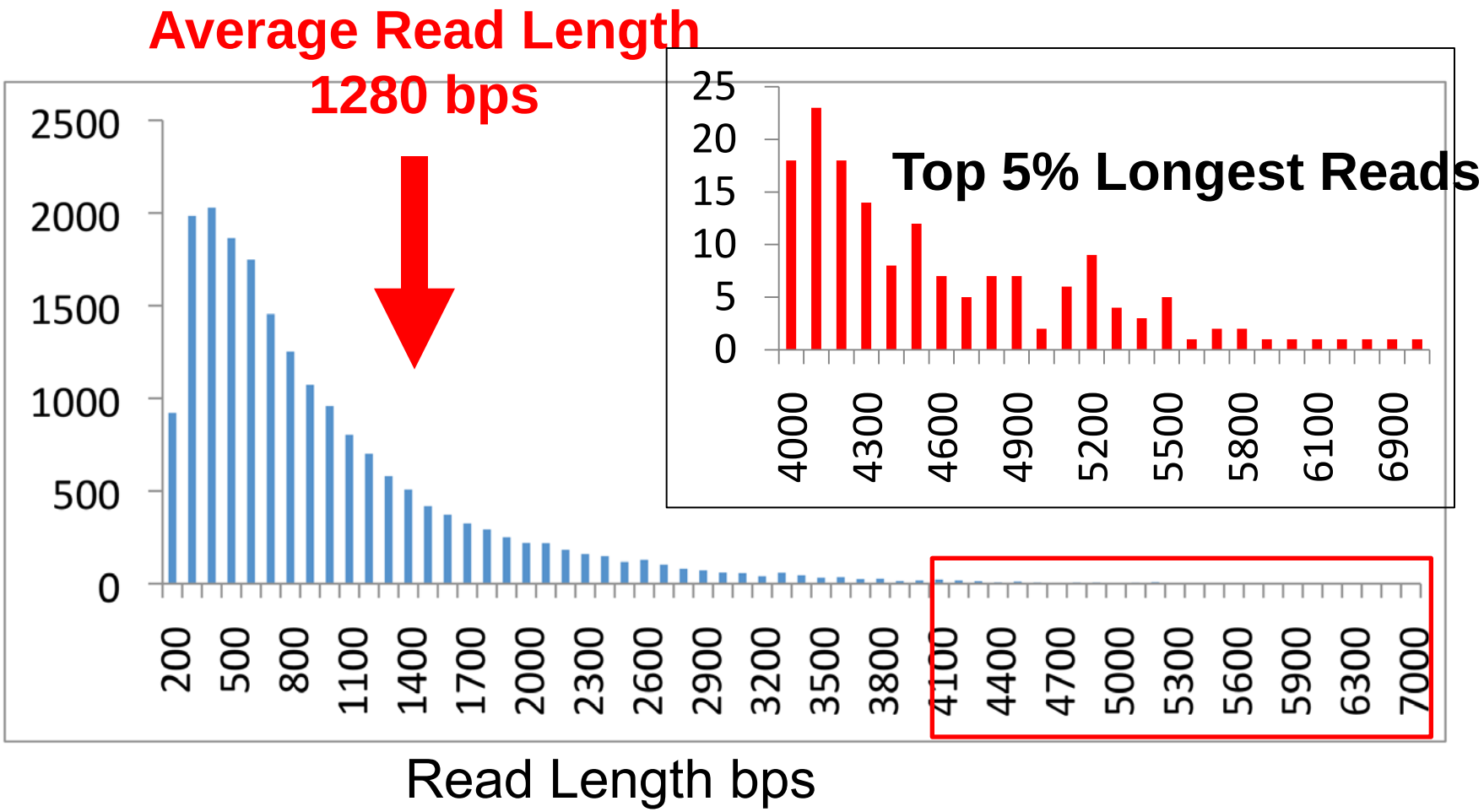
- Read length: 100-150 bp
- GC biases

Output Growth Prediction



PacBio

Pluses: Read length avg: 1500bp, no GC bias
Minuses: Total output 800 Mb/run:
Accuracy ~ 85%, cost: $3/Mb



Next Gen Capabilities

	2011	2012	2013	2014	2015
Gb/run	200	400	800	1600	3200
Read length	2x100	2x125	2x150	2x150	2x200
Cost/run	$20,000	$20,000	$20,000	$15,000	$15,000
Runs/year	25	25	30	30	35
Cost/machine/year	$500 K	$500 K	$600 K	$450 K	$525 K
Tb/machine/year	5	10	24	48	112
Cost/Gb	$100	$50	$25	$10	$5

CSP 2011

Large-scale sequence based genomic science projects that address questions of relevance to DOE missions in alternative fuels, global carbon cycling and biogeochemistry.

1) **Plant and plant-microbe interactions: interaction of plants with their rhizosphere communities and other microbes or fungi that affect bioenergy-relevant plant phenotypes. Plant resequencing or transcriptomic projects**
2) **Microbial emission and capture of greenhouse gases: global carbon, nitrogen, and methane cycles**
3) **Metagenomics: couple metagenomic analyses with measures of the active component of microbial populations and associated environmental biogeochemistry to explore dynamic changes in the active community composition and expressed metabolism of microbial communities in DOE mission-relevant areas**

CSP 2011

Proposals are encouraged involving combinations of the following features:

- a significant scale, large DNA/RNA sequencing need (100's of gigabases to terabases), engaging a large group of collaborators,

- involving more than a single species

- requiring JGI capabilities in addition to genome sequencing.

As much as 10 Tb may be available for this CSP. Each proposal must carefully justify the amount of sequencing requested but no single proposal should expect more than a maximum of 1 Tb.

CSP 2011

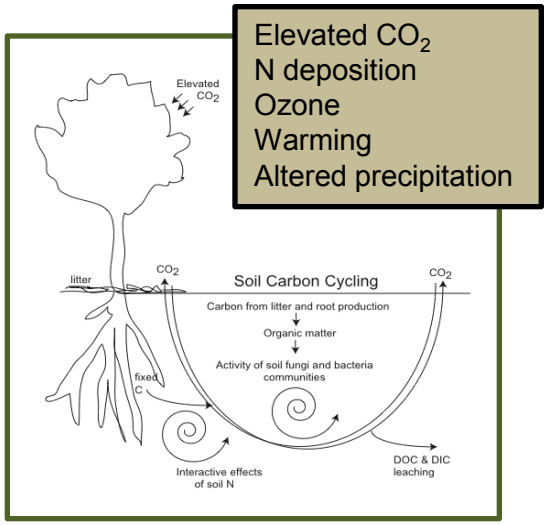
Core Products Include:

- ***De novo* sequencing of fungal, bacterial, and archaeal genomes**
- **Resequencing for variation detection**
- **Metagenomic DNA/RNA sequencing**
- **RNA-Seq for genome annotation**
- **RNA-Seq for reference gene sets**
- **RNA-Seq for gene counting**
- **Analysis pipelines for the datasets above**

JGI also has limited capacity for the following developing capabilities:

- **3rd generation single-molecule sequencing**
- **Fluorescent Activated Cell Sorting**
- **Micro-manipulator isolation of single cells**
- **Single cell DNA sequencing**
- **Multiplex emulsion PCR amplicon sequencing**
- **Limited DNA/gene synthesis**
- **Custom genome analysis of generated datasets**

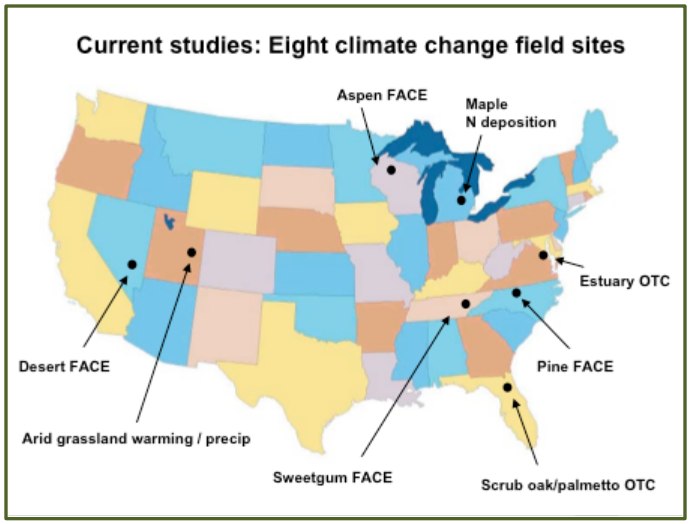
Terrestrial Ecosystem Responses to Interactive Climate Change Parameters – the Soil Component

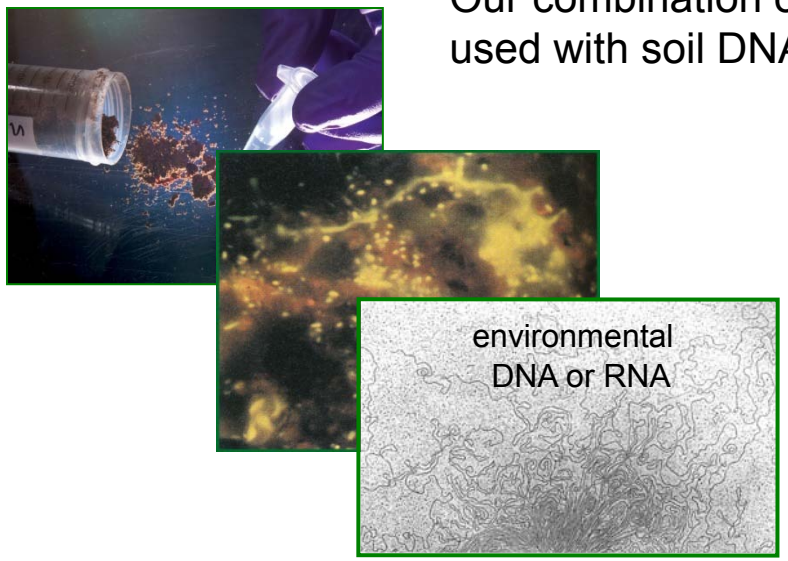


Across multiple, highly replicated, large field studies –

Look for patterns of changes in:

- Total soil biomass
- Relative abundance of major lineages
- Richness or diversity
- Community species composition
- Community functional composition
- Activity of certain populations





Our combination of TARGETED GENE and SHOTGUN METAGENOMICS, used with soil DNA and RNA - provides information on:

- Presence & activity of major species & functional groups
- Spatial & temporal heterogeneity / experimental design
- Ecosystem & response differences and similarities across different biomes
- Complexity and comparative metrics for shotgun metagenomes
- Patterns of response to climate change parameters

Observations in second year of survey & analysis

(1) Soil bacterial and fungal communities are very different across the eight field sites, and are stratified by depth and with plant association.
(2) Responses have been detected in soil components at each field experiment. In many cases the patterns of response differ between sites.
(3) Even so, certain broadly responsive groups w/ potential usefulness as harbingers of change are being identified, and tools for their use being developed.
(4) Spatial variability within many field sites is high but not insurmountable – replication (sometimes extensive) and careful experimental design is essential.
(5) The state of the art of 'omics – interpretation of results can differ dramatically with approach. Improvement of underlying databases and use of combined approaches are important.
(6) Current analysis methods only survey a small portion of the shotgun metagenome sequences.
(7) New sequence database resources are essential and they must be carefully curated.

Conclusion

Ultra deep metagenomic sequencing, even with short reads, likely to increasingly become a method of choice to identify genes and characterize the genomes of uncultured organisms