• A few words about the JGI
  – who we are & what we do

• Single-cell genomics
  – the why, the how & what to expect from it

• Single-cell science vignettes
  – from symbionts to microbial dark matter

• Crystal ball
  – what the future may bring
• DOE runs 50+ national user facilities
• Incl. JGI, a high-throughput sequencing & analysis facility
  – Walnut Creek
  – 250 employees
  – Research
  – New technologies
  – Collaborative research
  – Building user communities

Serving as a genomic user facility in support of the DOE missions . . .
JGI Programs & Infrastructure

DOE Mission Areas
- Bioenergy
- Carbon Cycling
- Biogeochemistry

JGI Programs
- Metagenomes
- Plants
- Fungi
- Microbes
- Synthesis

JGI Infrastructure
- DNA Sequencing
- Genomic Technologies
- Computational Analysis
- Synthesis
JGI Sequencing Output

FY Total Bases (Gb) Sequenced

- **FY2005**: 34 Gb
- **FY2006**: 33 Gb
- **FY2007**: 39 Gb
- **FY2008**: 126 Gb
- **FY2009**: 1,004 Gb
- **FY2010**: 6,041 Gb
- **FY2011**: 29,903 Gb
- **FY2012**: 56,115 Gb
- **FY2013**: 71,000 Gb

Legend:
- **PacBio**
- **Illumina**
- **454**
- **Sanger**
Staying State of the Art

Sanger Sequencing to Next-Gen Sequencing by Synthesis

ABI 3730 Sanger reduced
MegaBase Sanger offline
454 in Production
454 Titanium
Solexa early access
Solexa in Production
SOLiD early access
Illumina GAIIx
Illumina HiSeq 2000
ABI 3730 Sanger offline
454 1K
Illumina MiSeq
454 offline
Illumina HiSeq 2500
PacBio early access
PacBio RSII


Sanger Sequencing to Next-Gen Sequencing by Synthesis

ABI 3730 Sanger reduced
MegaBase Sanger offline
454 in Production
454 Titanium
Solexa early access
Solexa in Production
SOLiD early access
Illumina GAIIx
Illumina HiSeq 2000
ABI 3730 Sanger offline
454 1K
Illumina MiSeq
454 offline
Illumina HiSeq 2500
PacBio early access
PacBio RSII

### JGI Sequencing Platforms

<table>
<thead>
<tr>
<th></th>
<th>Illumina HiSeq 2000</th>
<th>Illumina HiSeq 2500</th>
<th>Illumina MiSeq</th>
<th>Pacific Biosciences RSII</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Units</strong></td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td><strong>Reads (Single-Read/Cluster)</strong></td>
<td>&gt;1,000 Million per Flowcell</td>
<td>200 Million per Flowcell</td>
<td>&gt;10 Million per Flowcell</td>
<td>0.06 Million per SMRT Cell</td>
</tr>
<tr>
<td><strong>Readlength</strong></td>
<td>2 X 150bp Max*</td>
<td>2 X 250bp Max*</td>
<td>2 X 300bp Max</td>
<td>&gt;5,000bp Avg; 20,000bp Max</td>
</tr>
<tr>
<td><strong>Total Bases</strong></td>
<td>300 Gb per Flowcell</td>
<td>100 Gb per Flowcell</td>
<td>5-10 Gb per Flowcell</td>
<td>0.3 Gb per SMRT Cell</td>
</tr>
<tr>
<td><strong>Run Time</strong></td>
<td>16 Days for 2 X 150</td>
<td>4.5 Days for 2 X 250</td>
<td>3 Days for 2 X 300</td>
<td>0.08 Days (2 hours)</td>
</tr>
<tr>
<td><strong>Applications</strong></td>
<td>Primary Sequence Generator at JGI</td>
<td>Rapid output HiSeq</td>
<td>10S Hags; Library QC, R&amp;D</td>
<td>Assembly improvement, de novo, SynBio validation, methylation/epigenome</td>
</tr>
</tbody>
</table>

*Not supported by Illumina

Most published single cells have been sequenced on the Illumina platform: HiSeq, (MiSeq)
• A few words about the JGI
  – who we are & what we do

• Single-cell genomics
  – the why, the how & what to expect from it

• Single-cell science vignettes
  – from symbionts to microbial dark matter

• Crystal ball
  – what the future may bring
The cultivated minority

~99% uncultivated

~1% cultivated

ecological functions & metabolic capabilities largely unknown
In the context of phylogenetic diversity

Isolates

Uncultivated taxa (16S barcodes available)

Uncultivated taxa (no 16S barcodes available)

???
Major uncultivated branches in the bacterial ToL

Modified from Baker et al 2013 (Microbe)
The situation is similar for archaea
EM3  BRC1  “bacterial rice cluster”
AC1  Marine group A = SAR406
NC10  “Nullarbor cave”
WYO  SR1  “sulphur river”
WWE1, WWE3
KSB2, KSB3
MSBL2, MSBL3, MSBL5, MSBL6

GN01-GN05, GN09, GN13, GN14  “guerrero negro”
NKB19  OD1  “OP11-derived 1”
OP1-OP4, OP6–OP9, OP11-12
EM19  “obsidian pool”
EM3  GN01- GN05, GN09, GN13, GN14  “guerrero negro”
NKB19  OD1  “OP11-derived 1”
OP1-OP4, OP6–OP9, OP11-12
EM19  “obsidian pool”
EM3  GN01- GN05, GN09, GN13, GN14  “guerrero negro”
NKB19  OD1  “OP11-derived 1”
OP1-OP4, OP6–OP9, OP11-12
EM19  “obsidian pool”

BSM2, BSM3, BSM5, BSM6

ZB3  “spring alpine meadow”
WS1-5, WS6
Sediment-1 – Sediment-4
KB1 group
TG2, TG3
How to access the coding potential?
marker genes

**culture**
- cultured
- uncultured

- culture
  - single cell
  - target cell enrichment
  - metagenomics

- draft/ complete genomes
- partial draft genomes, complete genomes (rarely)
- draft genomes, complete genomes
- unassembled data, genome bins, complete genomes
<table>
<thead>
<tr>
<th>biased: only some players can be cultured</th>
<th>biased: limited lysis &amp; cell isolation</th>
<th>biased: only some organisms susceptible to enrichments</th>
<th>biased: assemblies largely limited to abundant players</th>
</tr>
</thead>
<tbody>
<tr>
<td>applicability: +</td>
<td>applicability: +++</td>
<td>applicability: +</td>
<td>applicability: +++</td>
</tr>
<tr>
<td>culturing can introduce genotype changes</td>
<td>genome snapshot at point in time</td>
<td>genome snapshot at point in time</td>
<td>genome snapshot at point in time</td>
</tr>
<tr>
<td>axenic: no assembly challenges</td>
<td>heterogenous populations may be dissected</td>
<td>heterogenous populations: assembly challenge</td>
<td>heterogenous populations: assembly challenge</td>
</tr>
<tr>
<td>phenotypic characterization / metadata: extensive</td>
<td>phenotypic characterization / metadata: limited to non-existing</td>
<td>phenotypic characterization / metadata: limited to non-existing</td>
<td>phenotypic characterization / metadata: limited to non-existing</td>
</tr>
<tr>
<td>generally less expensive</td>
<td>requires more specialized equipment, costly, technical challenges (bias, chimera..)</td>
<td>possibly requires more specialized equipment, costly</td>
<td>generally less expensive</td>
</tr>
</tbody>
</table>
The single-cell approach: how it works

1. **Isolation**
2. **Lysis**
3. **MDA**

![Diagram of the single-cell approach](image)

- Polymerase
- Random hexamers
- dNTPs
- Template DNA
Multiple displacement amplification (MDA)

- isothermal amplification process
- requires polymerase + random primers + dNTPs
MDA on linear & circular templates

J. Rajendhran et al, 2008
(Biotechnology Advances)
Phi29

- DNA polymerase from phage phi29 (Φ29)
- Exceptional strand displacement properties
- \( \rightarrow \) 100 kb amplicons(!)
- Replication at moderate temperatures
- Extreme processivity
- \( 3' \rightarrow 5' \) exonuclease activity
- High-fidelity (error rate \( \sim 5 \times 10^{-6} \))
Key challenges

**CHALLENGE**

Sample contamination ('hitchhiker’ DNA)

No universal lysis for all taxa

Chimerism

Reagent contamination

MDA bias
Key challenges

- Sample contamination (‘hitchhiker’ DNA)
- No universal lysis for all taxa
- Chimerism
- Reagent contamination
- MDA bias
QC of single cell data is critical

- 16S rRNA gene and marker gene phylogenetic analyzes
- Tetramer analysis (Kmer Frequency Analysis) and GC contents
- BLAST analyzes and IMG’s Phylogenetic profiler
- Removal of contamination via binning and data reload
Tetramer analysis

This red scaffold has points in the main cloud but extends well out.

Clicking on the points in this scaffold opens a separate window with more detail on the scaffold shown below.
Tetramer analysis

Staphylococcus sp. JGI 0001002-I23 - Plot of PC1, PC2, and PC3

The 3D below is generated using the Kbase applet.

PC 3 explains 3.5% of variation
PC 2 explains 5.3% of variation
PC 1 explains 39.3% of variation

Potential Contaminants
Target Organism
Select GC Content from drop down menu then click Show Histogram.
Phylogenetic profiler

Guideline by Scott Clingenpeel
(https://img.jgi.doe.gov)

<table>
<thead>
<tr>
<th>Kingdom</th>
<th>Genus</th>
<th>Genomes</th>
<th>Probes</th>
<th>Percentage</th>
<th>Phyla</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bacteria</td>
<td>Acidobacteria</td>
<td>7 (1)</td>
<td>2 (1)</td>
<td>0.02%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Actinobacteria</td>
<td>280 (1)</td>
<td>1 (1)</td>
<td>0.02%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Grambacteria</td>
<td>4 (2)</td>
<td>2 (2)</td>
<td>0.17%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bacilli</td>
<td>58 (30)</td>
<td>27 (15)</td>
<td>2.23%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Chloroplasts</td>
<td>231 (2)</td>
<td>2 (2)</td>
<td>0.17%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Planctomycetes</td>
<td>11 (2)</td>
<td>2 (2)</td>
<td>0.17%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Archaea</td>
<td>255 (17)</td>
<td>13 (12)</td>
<td>1.08%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Betaproteobacteria</td>
<td>179 (2)</td>
<td>4 (2)</td>
<td>0.33%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Deltaproteobacteria</td>
<td>57 (1)</td>
<td>1 (1)</td>
<td>0.05%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gammaproteobacteria</td>
<td>636 (8)</td>
<td>5 (6)</td>
<td>0.50%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Spirochaetes</td>
<td>69 (6)</td>
<td>20 (9)</td>
<td>1.85%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Firmicutes</td>
<td>19 (1)</td>
<td>4 (1)</td>
<td>0.08%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thaumarchaeota</td>
<td>14 (1)</td>
<td>1 (1)</td>
<td>0.08%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thermus</td>
<td>339 (8)</td>
<td>2 (2)</td>
<td>0.17%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SP5 RNA virus</td>
<td>502 (3)</td>
<td>1 (1)</td>
<td>0.17%</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Unassigned</td>
<td>120</td>
<td>20</td>
<td>0.87%</td>
<td></td>
</tr>
</tbody>
</table>

- **Filter**
- **Select All**
- **Clear All**
- **Show All Phyta**
Key challenges

**CHALLENGE**

- Sample contamination (‘hitchhiker’ DNA)
- No universal lysis for all taxa
- Chimerism
- Reagent contamination
- MDA bias
A fraction of single cells can be recovered

Clingenpeel et al 2014 (Frontiers in Microbiol)
Key challenges

- **Chimerism**
- **Sample contamination** (‘hitchhiker’ DNA)
- **No universal lysis for all taxa**
- **Chimerism**
- **Reagent contamination**
- **MDA bias**
Chimeric rearrangements

85% Inverted sequences

(A)

Genomic
Chimera

15% Direct sequences

(B)

Genomic
Chimera

(C)

(D)

Lasken et al 2007 (BMC Biotechnology)
Key challenges

- Sample contamination (‘hitchhiker’ DNA)
- No universal lysis for all taxa
- Chimerism
- Reagent contamination
- MDA bias
MDA – coverage bias

- *E. coli* isolate dataset

- *E. coli* single-cell dataset

A Copeland
MDA coverage bias

Shotgun sequencing theoretical kmer distribution

Isolate kmer distribution

SAG kmer distribution

A Copeland
MDA DNA normalization (wet lab)

1. Shear + linker ligation
2. Denaturation
3. Hybridization + DSN treatment
4. PCR amplification

Biased MDA DNA pool

Normalized MDA DNA pool
Coverage normalization (computational)

~1,500x coverage

5x coverage

~50x coverage
JGI single-cell sequencing pipeline

- Whole genome amplification
- Single cell isolation
- Sample
- 16S rRNA gene identification
- Genome sequencing
- Assembly
- Data QC
- Annotation
- Data curation
1-2 ul MDA reactions possible with Echo LH
Current limitations: 100 cells ≠ 100 SAGs

- Not every cell can be isolated
- Not every cell can be lysed and WGA’d
- Not every cell can be 16S ID’d
Recovered diversity: 16S tags vs SAGs

Modified from Clingenpeel et al 2014 (Frontiers in Microbiol)
Benchmark experiment: what to expect?

**P: Pedobacter heparinus**
Size: 5,167,383 bp
GC-content: 42%

**E: Escherichia coli str. K-12**
Size: 4,639,675 bp
GC-content: 51%

**M: Meiothermus ruber**
Size: 3,097,457 bp
GC-content: 63%

Sequenced 8 single-cell genomes/each

Image source: http://www.standardsingenomics.org
Assemblies are draft quality

Clingenpeel et al 2014 (Frontiers in Microbiol)
With minimal errors

Clingenpeel et al 2014 (Frontiers in Microbiol)
SAG assemblies complement each other

Clingenpeel et al 2014 (Frontiers in Microbiol)
Low CP values lead to larger assemblies

Clingenpeel et al 2014 (Frontiers in Microbiol)
A few words on sample preservation & preparation

glycerol, betaine, DMSO, ...

(https://scgc.bigelow.org/PDFs/Sample_cryopreservation_glyTE.pdf)

Rinke et al 2014 (Nature Protocols)
Does sample pre-treatment affect genome recovery?

Clingenpeel et al 2014 (ISME J)
What to do with this single-cell data?
Utility of single-cell data is nearly endless.

Anchors for biogeography studies

References to assess heterogeneity
Utility of single-cell data is nearly endless.

Metagenome assembly and binning validation

Large-scale linkage of phylogeny and function
Utility of single-cell data is nearly endless..

Inferring metabolism of uncultured microbes

Metatranscriptome data mapping
Utility of single-cell data is nearly endless..

Identifying novel genomic features

Inferring phylogenetic relationships
• A few words about the JGI
  – who we are & what we do

• Single-cell genomics
  – the why, the how & what to expect from it

• Single-cell science vignettes
  – from symbionts to microbial dark matter

• Crystal ball
  – what the future may bring
Marine Flavobacteria

N. Moran (U of Arizona)

Sharpshooter symbiont

J. Dangl (UNC)

Arabidopsis endophytes

MDM Community

R. Stepanauskas (Bigelow Lab)
Two draft single-cell genomes

Marine Flavobacteria sp.

Ramunas Stepanauskas (Bigelow Lab)

Syto-13
Side scatter
single cells

Two draft single-cell genomes

<table>
<thead>
<tr>
<th>Assembly statistic</th>
<th>MS024-2A</th>
<th>MS024-3C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly size [Mbp]</td>
<td>1.905</td>
<td>1.515</td>
</tr>
<tr>
<td>Estimated genome size [Mbp]</td>
<td>2.095</td>
<td>1.947</td>
</tr>
<tr>
<td>Estimated genome recovery [%]</td>
<td>91</td>
<td>78</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>17</td>
<td>21</td>
</tr>
<tr>
<td>Largest contig [kbp]</td>
<td>684</td>
<td>549</td>
</tr>
</tbody>
</table>

Extensive manual gap closures

Ace #3
1.44 Mbp
47 contigs
Longest 56 kb

Ace #65
1.9 Mbp
17 contigs
Longest 680 kb
SAGs are representative of their environment

Green sharpshooter symbionts

Green sharpshooter (*D. minerva*)
- sap-feeding leafhoppers
- harbors two symbionts
- involved in plant pathogen spread (*Xylella*)
- causes serious plant diseases

*Sulcia* as ideal candidate for single-cell finishing
- uncultured
- small genome size
- simple single cell isolation
- fresh sample availability
- polyploid
Single cell & metagenome approach

1 sharpshooter

25 sharpshooters

Sulcia

marker genes

single cell approach

metagenomic approach
Are the genomes identical?

1 sharpshooter

single cell approach

25 sharpshooters

metagenomic approach
Are the genomes identical?

1 sharpshooter

- single cell approach

- TGAA[ttatca]_n TTAA

- CTAA[tgaagttaa]_n CCAT

- TAAA[agaaattgagaagttgcaataataaa]_n TTTA

- CTTG[ctaetca]_n ATAA

25 sharpshooters

- metagenomic approach
One complete single-cell genome

Sequence data
(Sanger/454)

Assembly

31 contigs

Gap closure / polishing
(Sanger/ Illumina)

Candidatus Sulcia muelleri DMIN
243,929 bp

Region with majority of gaps

Sequence coverage

Illumina
454
Sanger

Evaluation of population variation

Single cell approach

25 sharpshooters

Metagenomic approach

2 SNPs found

Low genetic diversity within Sulcia population
Are there unique communities in each compartment? Does the plant control access?
Arabidopsis microbiome project

Identifying the major determinants of microbial community assembly

Host factors

Root-associated microbial communities

Variables investigating:
- Soil type – Mason Farm vs. Clayton
- Sample fraction – Bulk soil vs. rhizosphere vs. endophyte
- Plant age – bolting (young) vs. senescent (old)
- Genotype – 8 ecotypes
- Individual – Aim for 10 individuals per condition

Full factorial design
1117 samples
16S pyrotag profiles

J. Dangl
The Arabidopsis microbiome

The endophyte community is unique and reproducible and similar across soil types

Lundberg
Nature 2012

Rhizosphere/Soil 1
Rhizosphere/Soil 2
Endophyte

S. Tringe
The Arabidopsis Microbiome

Isolates, single cells & enriched metagenomes

Cultured isolates

“Plate scrape” metagenomes

Single cells

Flow-sorted “mini-metagenomes”

Streptomyces

Sphingobacteriales

Pseudonocardiaciae

Isolates, single cells & enriched metagenomes

S. Tringe
An endophyte genome catalog

99 isolates and 130 SAGs

fully sequenced

>50% of target OTUs
Marine Flavobacteria | Sharpshooter symbiont | Arabidopsis endophytes

N. Moran (U of Arizona) | J. Dangl (UNC)

R. Stepanauskas (Bigelow Lab)
Expansion of phylogenetic diversity

Using single-cell genomics to look outside the lamp post
Expansion of phylogenetic diversity

16S rRNA tree of known bacterial phyla
Samples from 9 sites were selected
Some sites of high underexplored diversity

Nearly 50% of tags assigned to candidate phyla
From nearly 10,000 cells to 200 draft genomes

16S identification

Rinke et al, 2013 (Nature)
A good candidate phyla representation

Rinke et al 2013 (Nature)
Assembly statistics of SAGs

- Total assembly size
- Number of contigs
- N50 contig size
- Largest contig

Individual SAG assemblies (n=203)

Rinke et al 2013 (Nature)
Estimated genome recoveries

Rinke et al. 2013 (Nature)
• Provided the first substantive genomic data for candidate bacterial phyla SAR406, OP3, OP8, WS1, WS3, BRC1, CD12, EM19, EM3, NKB19, and Oct-Spa1-106, as well Nanoarchaeota-related groups
• Resolved numerous intra-and interphylum level relationships
• Proposed new bacterial and archael super-phyla

Rinke et al 2013 (Nature)
Genomic and functional novelty

Rinke et al. 2013 (Nature)
Single cells as phylogenetic anchors

Example: Sakinaw Lake

Rinke et al 2013 (Nature)
Single cells assist in metagenome assignments

Sakinaw Lake

With dark matter single cells

**Who can do what?**

- Assigned: 12%
- Unassigned: 88%
- Assigned by dark matter single cells: 20%
### Single cells as phylogenetic anchors

<table>
<thead>
<tr>
<th>Metagenomes</th>
<th>Anchored metagenome reads [%]</th>
<th>Read assignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>[SAK] Sakinaw Lake</td>
<td>19.56</td>
<td>0.1%</td>
</tr>
<tr>
<td>TA mother reactor</td>
<td>14.39</td>
<td>1%</td>
</tr>
<tr>
<td>[TA] TA reactor biofilm</td>
<td>10.82</td>
<td>2%</td>
</tr>
<tr>
<td>[GBS] CS 85C sediment</td>
<td>9.95</td>
<td>3%</td>
</tr>
<tr>
<td>[GBS] 77S water</td>
<td>9.27</td>
<td>≥3%</td>
</tr>
<tr>
<td>Saanich Inlet pooled fosmids</td>
<td>8.83</td>
<td></td>
</tr>
<tr>
<td>GOS Mangrove on Isabella Island</td>
<td>6.24</td>
<td></td>
</tr>
<tr>
<td>Yellowstone Bison Hot Spring</td>
<td>5.18</td>
<td></td>
</tr>
<tr>
<td>[GBS] CS 77C sediment</td>
<td>4.63</td>
<td></td>
</tr>
<tr>
<td>GOS Sargasso Stations 3</td>
<td>4.38</td>
<td></td>
</tr>
<tr>
<td>[TA] TA reactor sludge</td>
<td>4.30</td>
<td></td>
</tr>
<tr>
<td>Line P P20-A09-1000m</td>
<td>3.57</td>
<td></td>
</tr>
<tr>
<td>Rifle 2008 BG 13ft</td>
<td>3.41</td>
<td></td>
</tr>
<tr>
<td>Peru Margin</td>
<td>3.28</td>
<td></td>
</tr>
<tr>
<td>Marine Sediments sample SCG71</td>
<td>2.91</td>
<td></td>
</tr>
<tr>
<td>[ETL] Etoliko lagoon sediment</td>
<td>2.63</td>
<td></td>
</tr>
<tr>
<td>[GBS] 77S sediment</td>
<td>2.42</td>
<td></td>
</tr>
<tr>
<td>Biofilm from redox zones LI09 3</td>
<td>2.24</td>
<td></td>
</tr>
<tr>
<td>Line P P20-J09-500m</td>
<td>2.17</td>
<td></td>
</tr>
</tbody>
</table>

**Bacteria**
- Acetothermia
- Aerophobetes
- Aminocanantes (OP8)
- Armatimonadetes
- Atribacteria (OP9)
- Caldesiica
- Calescimantetes (WWE1)
- Cloacimonetes
- Eubacteriia
- Hydrogenetutes
- Lateclibacteria
- Marinimicrobia
- Microgenomates
- Omnitrophica
- Parcubacteria
- Synergistes

**Archaea**
- Aenigmarchea
- Aigarchaeota
- Alphaproteobacteria
- Euryarchaeota
- Nanoarchaeota
- Thaumarchaeota

*Note: Read assignment percentages and phylogenetic anchoring information derived from Rinke et al 2013 (Nature)*
Single cells as phylogenetic anchors

Modified from Rinke et al 2013 (Nature)
Novel stop codon reassignment

UAG ("amber")
UAA ("ochre")
UGA ("opal") \[\rightarrow\] glycine

new translation table:
“Candidate Division SR1 and Gracilibacteria Code”

UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota

James H. Campbell*,1, Patrick O’Donoghue*1, Alisha G. Campbell*1, Patrick Schwientek*, Alexander Sczyrba*, Tanja Woyke*, Dieter Söll*1,2, and Mircea Podar*1,2

*Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831; 1Department of Molecular Biophysics and Biochemistry and 2Department of Chemistry, Yale University, New Haven, CT 06520; 4Genome Science and Technology Program and 5Department of Microbiology, University of Tennessee, Knoxville, TN 37998; 6Department of Energy Joint Genome Institute, Walnut Creek, CA 94598; and 7Center for Biotechnology, Bielefeld University, 33501 Bielefeld, Germany

25. Candidate Division SR1 and Gracilibacteria Code (transl_table=25)

Click here to change format

Differences from the Standard Code:

<table>
<thead>
<tr>
<th>Code 25</th>
<th>Standard</th>
</tr>
</thead>
<tbody>
<tr>
<td>UGA</td>
<td>Gly</td>
</tr>
</tbody>
</table>
Stop codon reassignments in the wild

Eddy Rubin

Natalia Ivanova

Nikos Kyrpides

Patrick Schwientek

Ivanova et al 2014 (Science)
Sites of stop codon reassignments

Ivanova et al 2014 (Science)
Bacterial opal recoding in freshwater and human

Ivanova et al 2014 (Science)
Two stop codon assignment evens in bacteria

Sample site:
- Freshwater
- Human
- Marine
- Cow rumen
- Multiple

Collapsed branches:
- Recoded clades

Sample type:
- Isolate genomes
- Single cell genome
- Metagenome contigs

Ivanova et al 2014 (Science)
Microbial Dark Matter phase II: A community effort
Microbial Dark Matter phase II: A community effort

- Ramunas Stepanauskas, Bigelow Lab
- Steven Hallam, UCB
- John Spear, Colorado
- Mostafa Elshaed, OK State
- Sean Crowe, UBC
- Brian Hedlund, U Nevada
- Peter Dunfield, U of Calgary
- Andreas Teske, UNC
- Hailiang Dong, North China
- Paul van de Wielen, KWR Water
- Sari Peura, Uppsala U.
- Konstaninos Kormas, U. Thessaly
- Nikolai Ravin, Russian Acad. Sci.
- Matt Stott, GNC Science
- Chuanlun Zhang, Tongji University
- Karin Rengefors, Lund University
- Steve Lindemann, PNNL
- Nils-Kare Birkeland, U of Bergen
Expected outcome

- More novel discoveries
- Improved binning of metagenome data
- Understanding of DOE–relevant systems of JGI Users
- Microbial ecology & evolution
  - Functional roles of candidate phyla in the environment
  - Phylogenetic distribution of key metabolic functions
  - Co-occurrences of candidate phyla
  - Early evolution of bacterial & archaeal domains
What else is out there?

The Search for new major branches
Search for new major branches

Deep analysis of massive metagenome data
Kyrpides team
Woyke et al 2014 (Science)

Two JGI efforts:
- Massive single-cell genomics
  Single cells/ GenTech groups
  - Large-scale assemblies of
    - environmental nucleic acid sequence
    - single cells lacking amplifiable rRNA genes
  - Advanced analyzes
    - K-mer
    - tRNA structure
    - codon usage
    - phylogeny
    - detection of non-canonical bases

Leeuwenhoek
discovery of
1st bacterium

Woese
discovery of
1st archaeon

1990
Woese
3 domains

Is there a
4th domain?
One approach to look for “new major branches”

- Isolation of single cells
- Whole genome amplification
- 16S identification
- Genome sequence and assembly
Amplified sorted single cells without 16S amplicon:

~4,000

Where else should one look?
Where else should one look?

Environments mimicking that of early Earth

- Earliest stromatolites (3.4 billion years ago)
- Biomarker evidence for cyanobacteria (2.8 billion years ago)

→ Deep subsurface “oxygen free” environments

Sanford Underground Research Facility (D Moser)

Songliao basin (H Dong)
• A few words about the JGI
  – who we are & what we do

• Single-cell genomics
  – the why, the how & what to expect from it

• Single-cell science vignettes
  – from symbionts to microbial dark matter

• Crystal ball
  – what the future may bring
fnx-driven single-cell genomics

Modified from imgkid.com
Function-driven single-cell genomics

- “Next generation single-cell sequencing”:
- Identification and pre-enrichment of uncultivated environmental microbes that are involved in biogeochemical processes of interest:
  - targeted/ function-driven
  - prior to sequencing
  - without relying on any known genetic markers
  - without cultivation biases

“finding the needles in the haystack”
Why function-driven genomics?

• Need to move beyond sequencing yet another genome and narrow our focus on studying microbes that are involved in biogeochemical processes of interest

• Function-driven genomics is part of JGI’s strategic future

• For single cells, adding a functional component is technologically highly novel (“Next-gen SCG”)

What is needed:

**Metagenomics**: high throughput, but genome context of functional genes often missing

**Single-cell genomics**: enables genome context, but low throughput

Neither methods is generally targeted
Function-driven single-cell genomics

- Sample
- Pre-screen
- Single cell isolation
- Whole genome amplification
- 16S rRNA gene identification
- Genome sequencing

Total community

Functionally-targeted
Function-driven single-cell genomics

Two examples:

**Nitrifier diversity**
ETOP: Michi Wagner (Vienna)/ Roman Stocker (MIT)

**Plant carbon decomposition**
LDRD

---

Emerging Technologies Opportunity Program
ETOP

Massachusetts Institute of Technology

JGI
Joint Genome Institute

jbei
Joint BioEnergy Institute

---

Plant in field, 20 μm scale bar.
Nitrospira diversity

Wastewater treatment plant

20 μm

16S rRNA

Cluster 2.1
- Vetmed WWTP clone 5 (02/2009)
- Vetmed WWTP clone 212 (02/2009)
- denuhll clone 1
- iron–oxidation biofilm clone 9
  - activated sludge clone SBR2046
  - Vetmed WWTP clone 124 (10/2010)
  - Vetmed WWTP clone 18 (02/2009)
  - Vetmed WWTP clone 27 (02/2009)
  - Vetmed WWTP clone 13 (02/2009)
  - Vetmed WWTP clone 25 (02/2009)
  - Vetmed WWTP clone 2 (02/2009)
  - Vetmed WWTP clone 203 (02/2009)
  - Vetmed WWTP clone 1 (02/2009)
  - activated sludge clone SBR1015

Cluster 2.2
- Vetmed WWTP clone 4 (02/2009)
- Vetmed WWTP clone 101 (10/2010)
- Vetmed WWTP clone 205 (10/2010)
- Vetmed WWTP clone 119 (10/2010)
- Vetmed WWTP clone 118 (10/2010)
- Vetmed WWTP clone 122 (10/2010)
- Vetmed WWTP clone 218 (02/2009)
- Vetmed WWTP clone 28 (02/2009)
- Vetmed WWTP clone 20 (02/2009)
- Vetmed WWTP clone 32 (02/2009)
- Vetmed WWTP clone 11 (02/2009)
- Vetmed WWTP clone 30 (02/2009)
- Vetmed WWTP clone 17 (02/2009)
- Vetmed WWTP clone 211 (02/2009)
- Vetmed WWTP clone 111 (10/2010)
- Vetmed WWTP clone 104 (10/2010)
- Vetmed WWTP clone 211 (10/2010)
- Vetmed WWTP clone 105 (10/2010)
- Vetmed WWTP clone 112 (10/2010)
- Vetmed WWTP clone 107 (10/2010)
- Vetmed WWTP clone 114 (10/2010)
- Vetmed WWTP clone 109 (10/2010)
- Vetmed WWTP clone 210 (10/2010)
- Vetmed WWTP clone 202 (10/2010)
- estrogen–degrading membrane bioreactor clone M1–5

Cluster 2.3
- Vetmed WWTP clone 7 (02/2009)
- Vetmed WWTP clone 103 (10/2010)
- Vetmed WWTP clone 207 (02/2009)
- Vetmed WWTP clone 3 (02/2009)
- stream bank clone 9
- large lignite mine lake sediment clone 12/10
- fluidized bed reactor clone 09

Cluster 2.4
- Nitrospira moscowiensis
- Cretan margin sediment clone HC3MC80_8B_FL
- Vetmed WWTP clone 204 (02/2009)
- Vetmed WWTP clone 24 (02/2009)
- drinking water distribution system simulator clone DSSD62

Cluster 2.5
- Nitrospira lineage II clones (14)
- Nitrospira lineage II clones (5)

M Wagner, H Daims, TK Lee
Raman-activated colony sorting

M Wagner, H Daims, TK Lee
Plant-carbon decomposition project

Native grassland

<table>
<thead>
<tr>
<th>SFREC: Taeniatherum caput-medusae</th>
<th>Wick Ranch: Avena barbata</th>
</tr>
</thead>
<tbody>
<tr>
<td>soil</td>
<td>soil</td>
</tr>
<tr>
<td>rhizosphere</td>
<td>rhizosphere</td>
</tr>
</tbody>
</table>

BULK

SUBSTRATE-ENRICHED

Bulk community

substrate-enriched community

Total community

16S/ITS profiling, MG

C substrate enriched community

16S/ITS profiling Mini-MG, SCG
Mix of cellulose-degrading and non-degrading bacteria

Fluorescent-labeled substrate (i.e., cellulose) isolation of degrading bacteria

DNA extr.

Mini-metagenomics (MG)

WGA

Single-cell genomics (SCG)
Acknowledgements

Team of MDM collaborators
Phil Hugenholtz (U Queensland)
Ramunas Stepanauskas (Bigelow Lab)
Jonathan Eisen (UC Davis)
Alex Sczyrba (U Bielefeld, GER)
Aaron Darling (UC Davis)
Dongying Wu (UC Davis)
Nikos Kyrpides (DOE-JGI)
Steve Hallam (U British Columbia)
Stefan Sievert (WHOI)
George Tsiamis (U Ioannina)
Wen-Tso Liu (U Illinois)
Brian Hedlund (U Nevada)
Phase II MDM collaborators

DOE-JGI
Chris Rinke
Patrick Schwientek
Natalia Ivanova
Iain Anderson
Jan-Fang Cheng
Nikos Kyrpides
Rex Malmstrom
Stephanie Malfatti
Susannah Tringe
Scott Clingenpeel
Janey Lee
Alicia Clum
Alex Copeland
Rubin, Bristow, Pennacchio Labs
Kyrpides & Markovitz Groups (IMG team)

Department of Energy
Dan Drell
Thank you!

Questions?