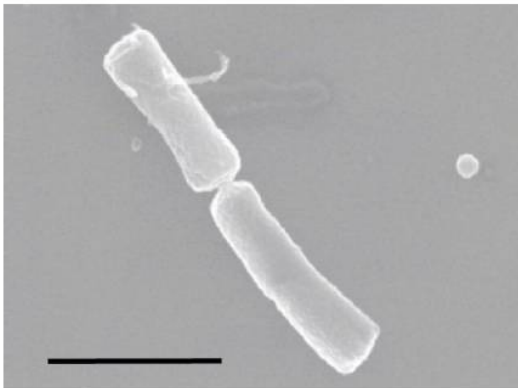
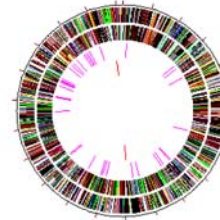
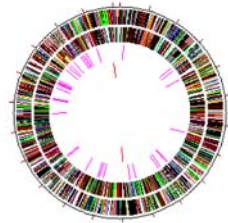
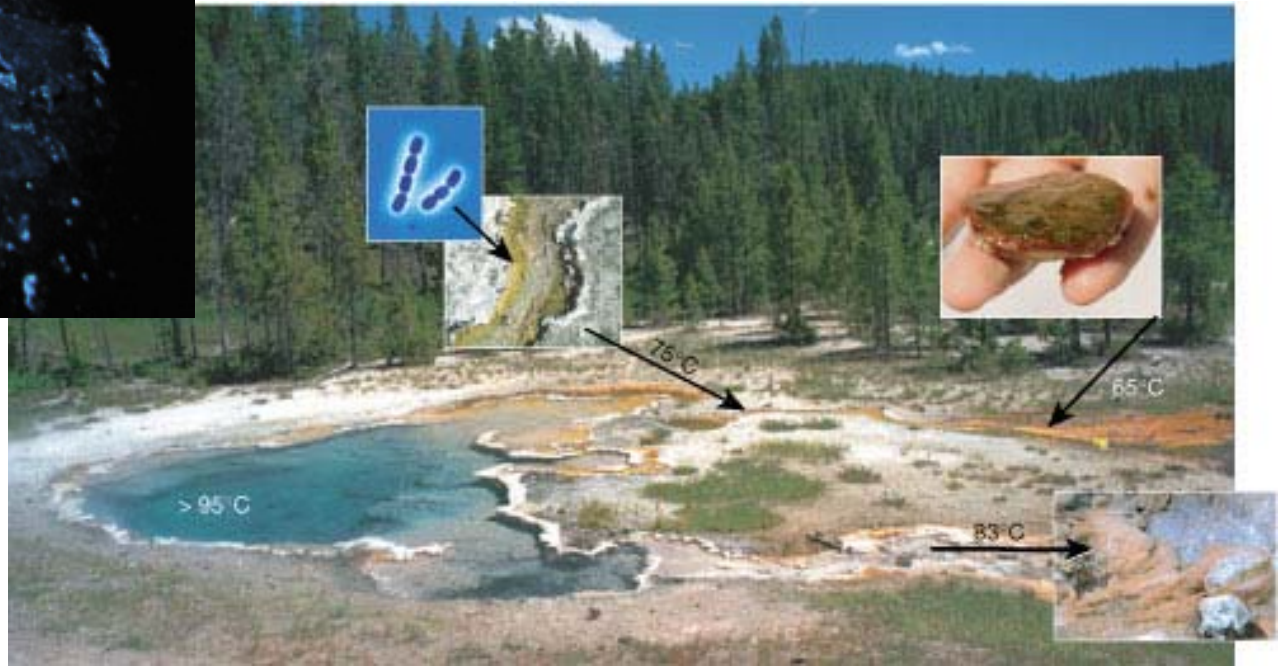
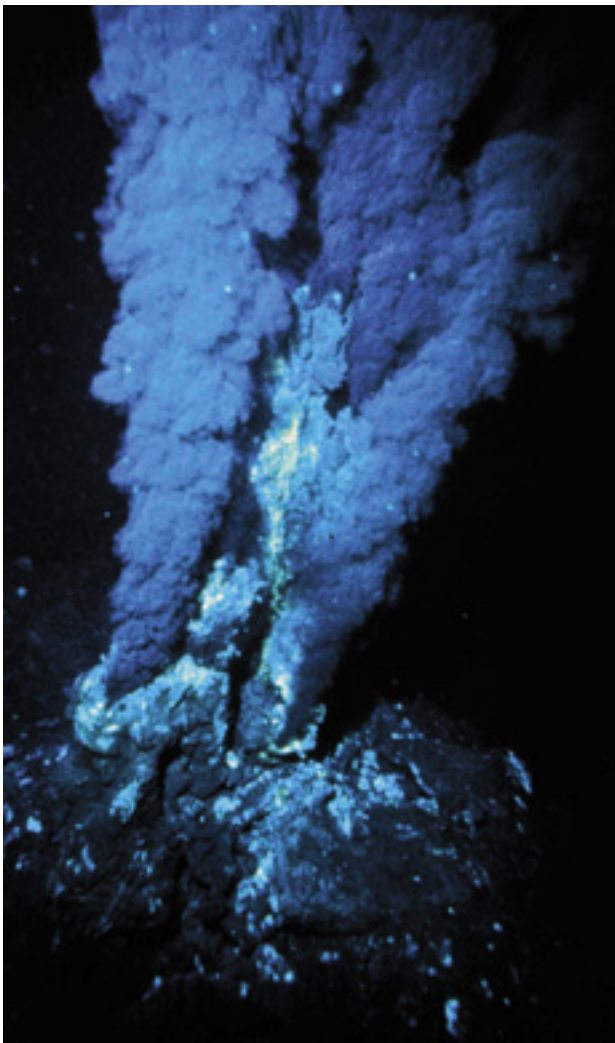


Genome Dynamics Across Five Pyrobaculum Species, and Their Novel Non-coding RNA Features



Todd Lowe
Biomolecular Engineering
University of California, Santa Cruz

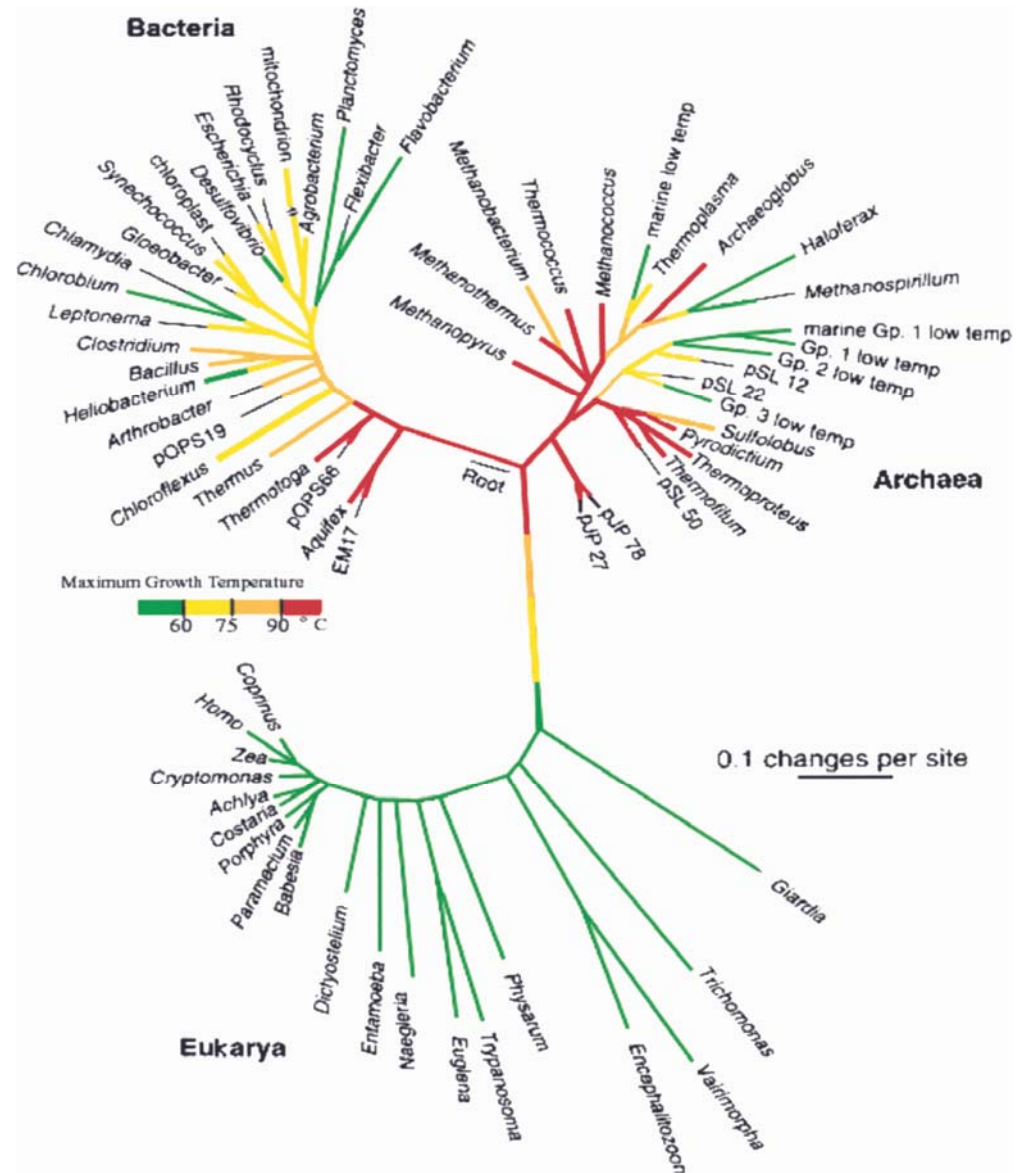


R. Cavicchioli

Thermophily by Phylogeny

Maximum observed growth temperatures

- Metazoans 50 °C
- Microbial Euk 65 °C
- Bacteria 95 °C
- Archaea 121 °C



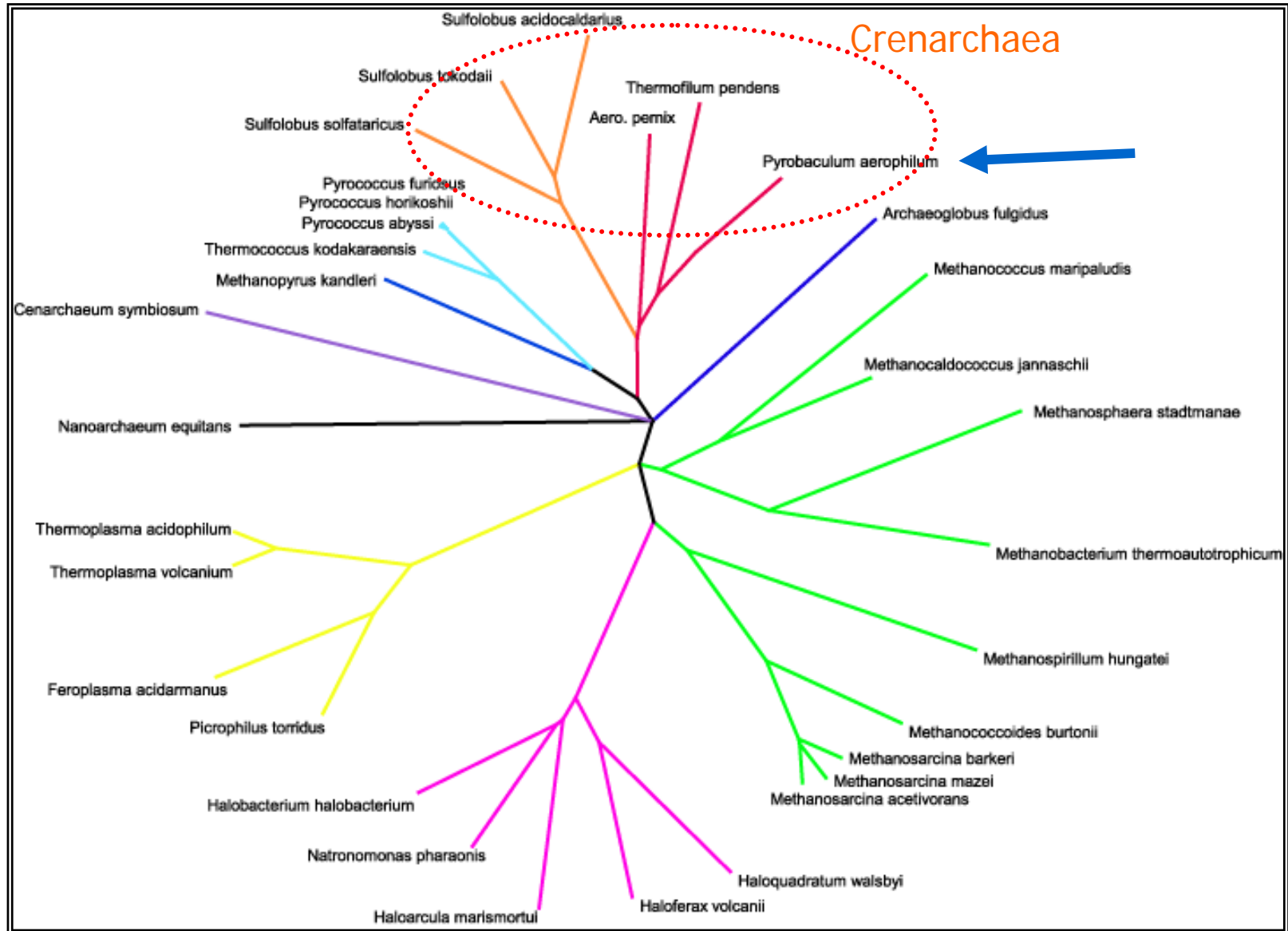
Challenges in Archaea

- Genetic systems for only a few species
- Understanding of basic processes such as transcription, translation, & DNA repair is far behind bacteria & euks
- Questionable gene predictions & functional assignments are the rule, not the exception
- Functional genomics studies are few (~20)
- Comparative genomics is limited compared to Bacteria and Eukarya

Our Approaches

- Comparative genomics
 - need more genomes
- DNA microarrays
- Traditional molecular biology
- RNA gene discovery
 - Computational
 - Small RNA and mRNA sequencing (454)

Archaeal Genome Tree (early 2006)



Pyrobaculum

P. aerophilum

- a crenarchaeote
- hyperthermophile: 70-102 °C (opt 95)
- micro-aerobic (<2.5% O₂) and anaerobic
- facultative autotroph, neutral pH
- sequence completed, 2001



Pyrobaculum genus:

- wide respiratory versatility
- some facultative aerobes, some microaerobes, some anaerobes
- widespread and abundant in geothermal systems (marine & fresh water)
- potential alternative, non-acidophile model system for crenarchaea

Home of *P. aerophilum*

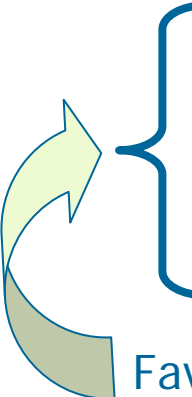


Vulcano, Italy

The *Pyrobaculum* genus: high respiratory versatility

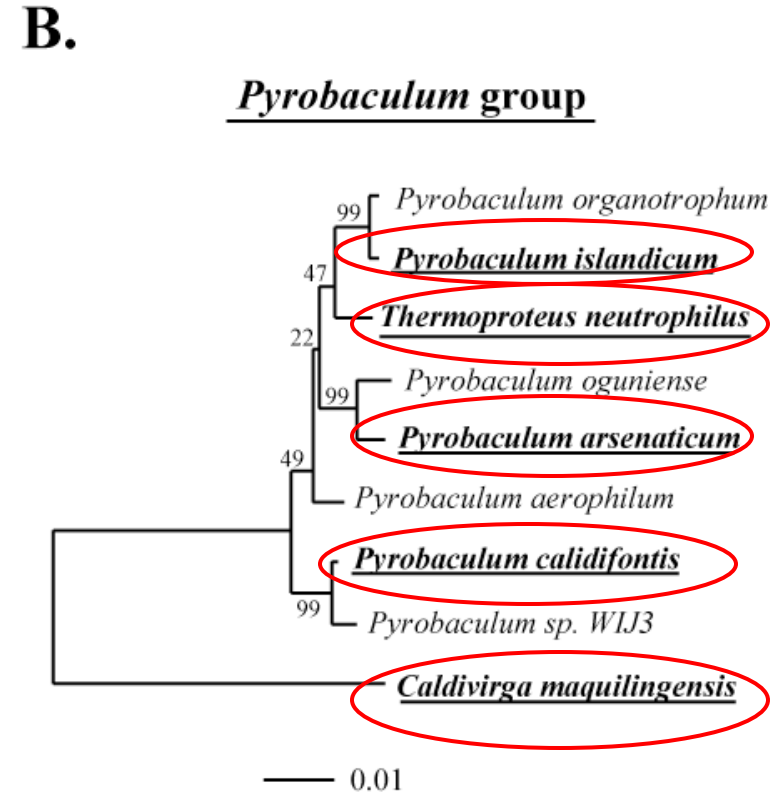
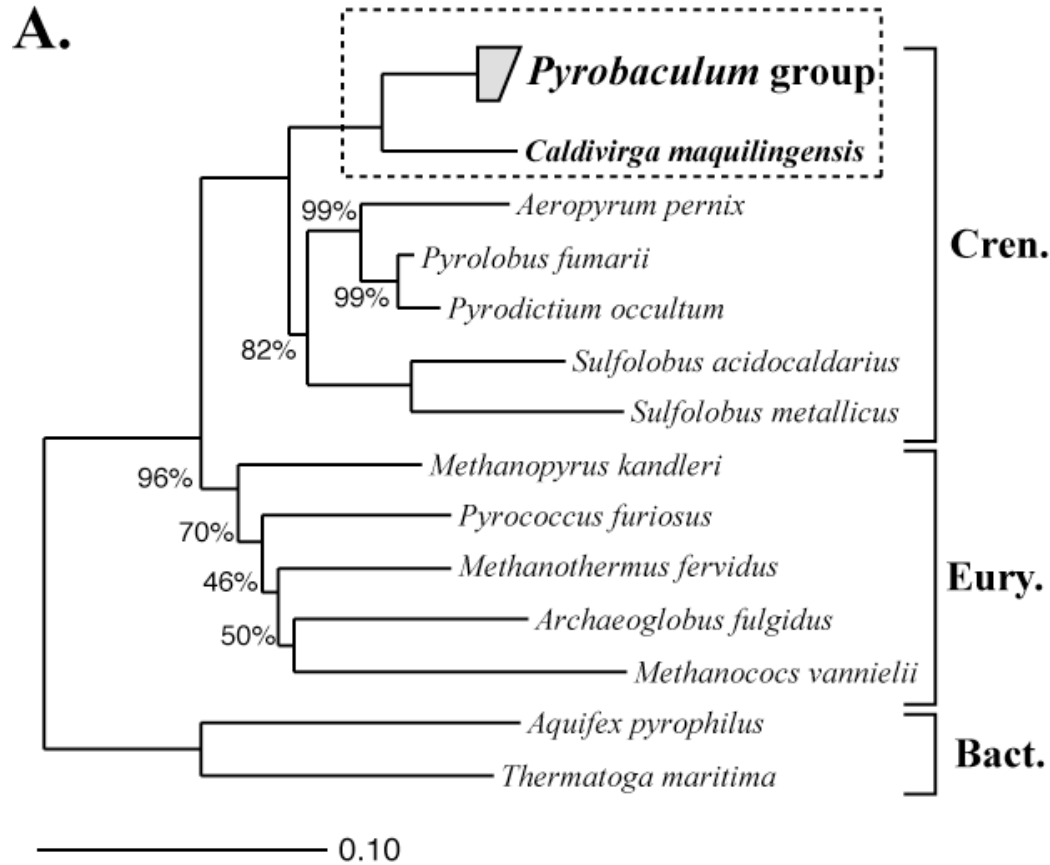
X	no growth
✓	growth
✓✓	better growth
✓✓✓	best growth

	oxygen	nitrate	nitrite	arsenate	selenate	ferric iron	thio-sulfate	sulfur
<i>Pyrobaculum aerophilum</i>	✓✓✓	✓✓✓	✓✓	✓✓	✓	✓✓✓	✓	X
<i>Pyrobaculum caldifontis</i>	✓✓✓	✓✓✓		✓		X	✓✓	X
<i>Pyrobaculum arsenaticum</i>	X	X		✓✓	✓	X	✓✓✓	✓✓✓
<i>Pyrobaculum islandicum</i>	X	X		✓		✓✓✓	✓✓✓	✓✓✓
<i>Thermoproteus tenax</i>								✓✓✓
<i>Sulfolobus solfataricus</i>	✓✓✓							
<i>Pyrococcus furiosus</i>								✓✓✓



Favorite Thermophile Models

DOE Community Sequencing Program 2006



Selection: Four *Pyrobaculum* & *Caldivirga*

Varied Lifestyles

	Optimal Growth	Oxygen Req.	Nutrition	Isolation	Electron Acceptors
<i>P. aerophilum</i>	100°C, pH 7	Facultative microaerobe	Facultative Autotroph	Shallow marine water hole, Italy	Oxygen, nitrate, nitrite, arsenate, selenate, selenite, thiosulfate, ferric iron
<i>P. arsenaticum</i>	95°C, pH 6	Strict anaerobe	Facultative autotroph	Hot spring, Italy	Sulfur, thiosulfate, arsenate , selenate
<i>P. islandicum</i>	100°C, pH 6	Strict anaerobe	Facultative autotroph	Geothermal power plant, Iceland	Sulfur, thiosulfate, sulfite, L-cystine, oxidized glutathione, ferric ion, arsenate
<i>Thermoproteus neutrophilus</i>	85°C, pH 6.5	Strict anaerobe	Obligate autotroph	Hot spring, Iceland	Sulfur only
<i>P. calidifontis</i>	90-95°C, pH 7	Facultative aerobe	Heterotroph	Hot spring, Philippines	Oxygen , nitrate
<i>Caldivirga maquilingensis</i>	85°C, pH 4.0 (2.3-6.4)	Facultative microaerobe	Heterotroph	Hot spring, Philippines	Sulfur, thiosulfate, sulfate

Sequencing Complete

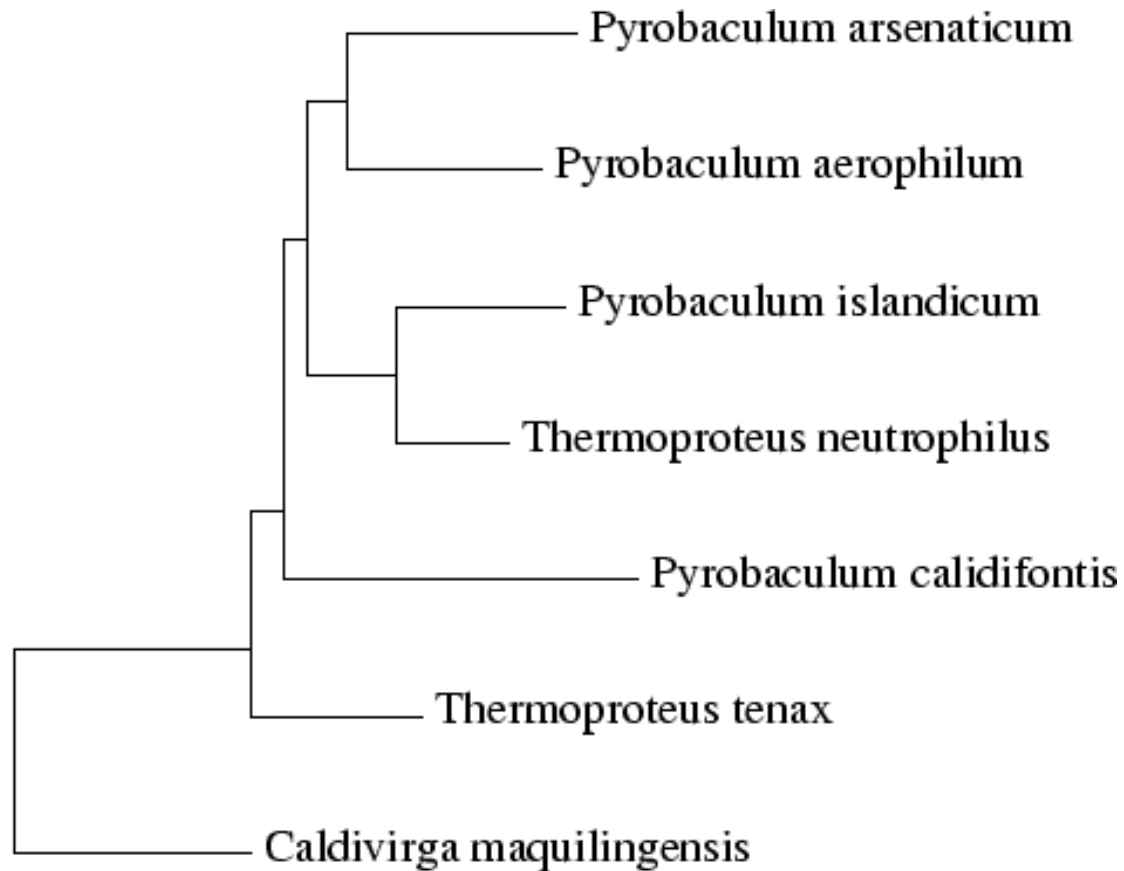
Species	Genome Size (bp)	G/C Content	Protein Genes	Genes w/ PFAM Hits	Unique Genes
<i>P. aerophilum</i>	2,222,430 (completed 2001)	51.4%	2650	1509 57%	125+
<i>P. arsenaticum</i>	2,121,076	55.0%	2407	1353 56%	50
<i>P. islandicum</i>	1,826,402	49.6%	2045	1181 58%	33
<i>Thermoproteus neutrophilus</i>	1,761,005	60%	2030	1160 57%	33
<i>P. calidifontis</i>	2,009,313	57.1%	2117	1285 61%	41
<i>Caldivirga maquilingensis</i>	2,077,575	43.1%	2147	1301 61%	80

- First-pass protein gene predictions from DOE's Oakridge National Labs annotation pipeline system (Critica, Glimmer)

Phylogenetic relationship?

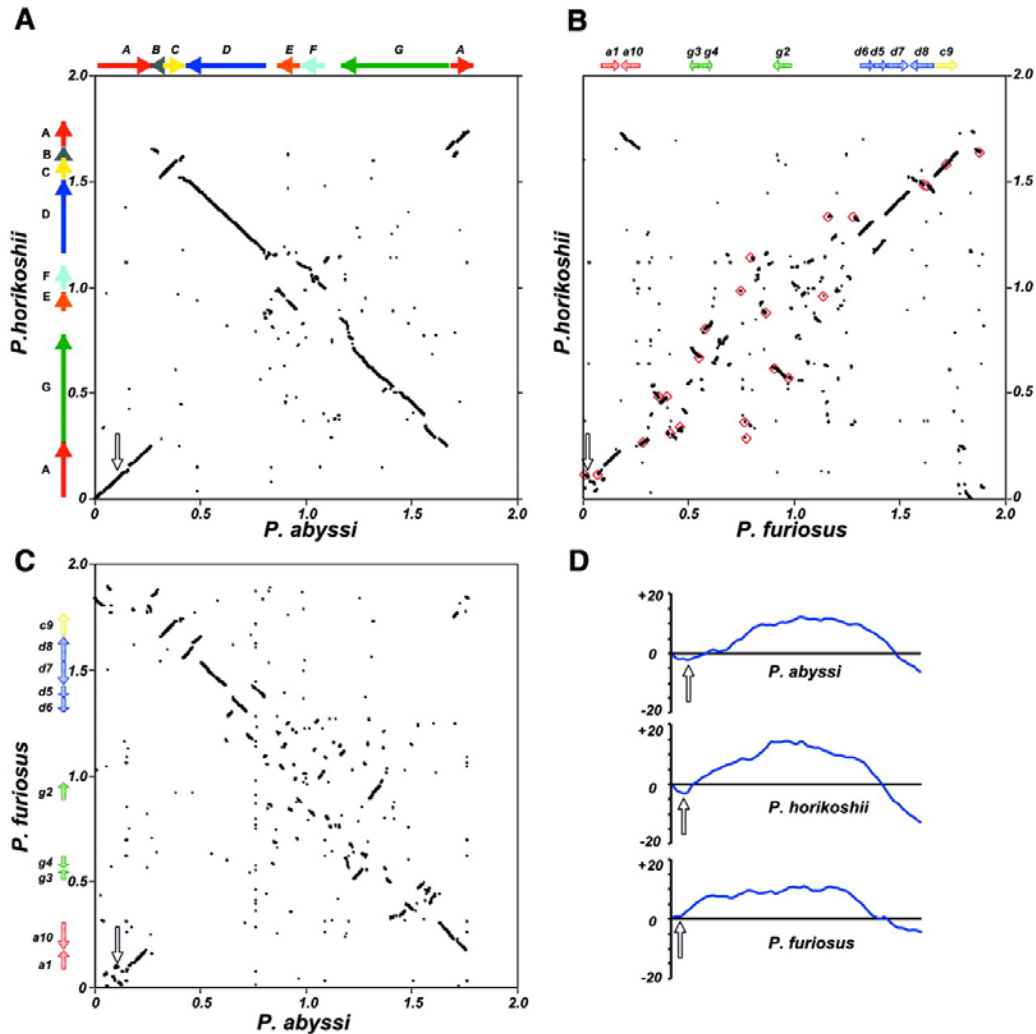
- 16S rRNA is 98-99% identical, need better resolution
- Higher-resolution methods
 - DNA-based
 - Full genome alignment similarity (phylo-HMM)
 - Dot plot / synteny
 - Protein-based
 - Average orthologous protein similarity

Phylogenetic Tree Based on 7-way Full Genome Alignment



Phylo-HMM estimates divergence based on DNA alignment
(Seipel & Haussler, 2004)

Full-genome Dot Plots

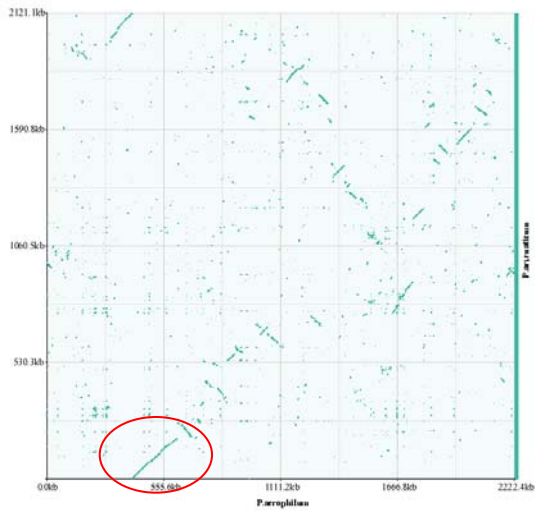


Pyrococcus
comparisons

From **Zivanovic et al.**,
NAR 30: 1902-10

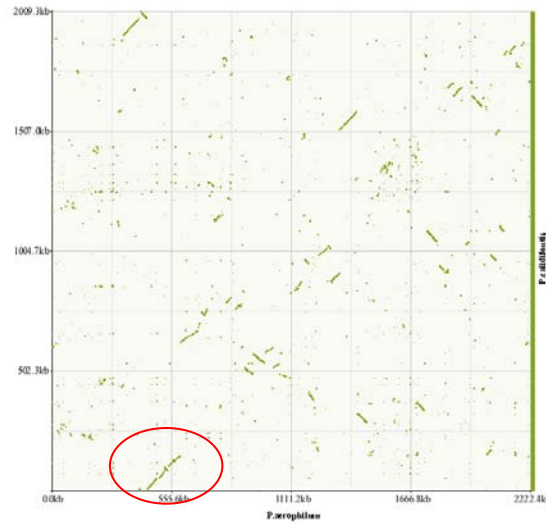
Full-Genome Alignment Dot Plots

P. arsenaticum



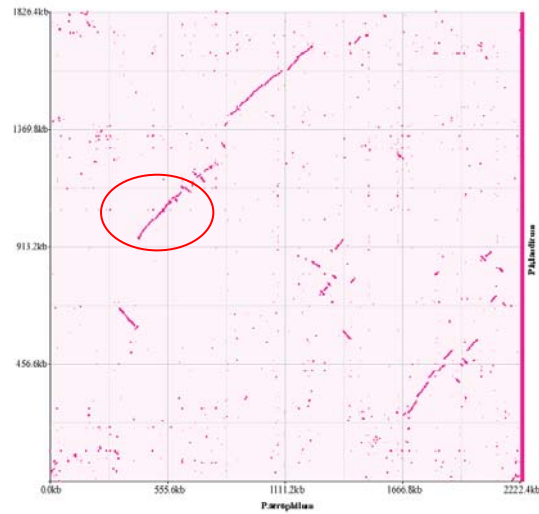
P. aerophilum

P. calidifontis



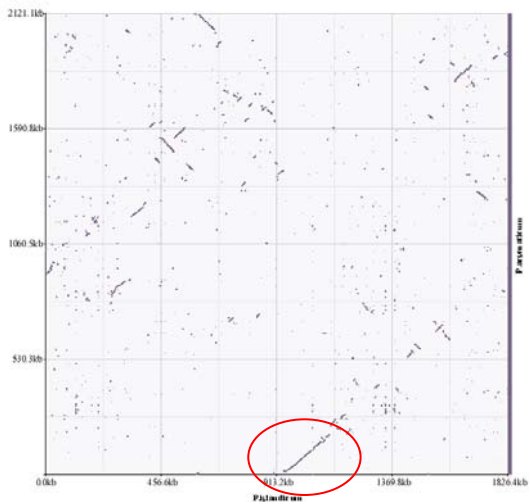
P. aerophilum

P. islandicum



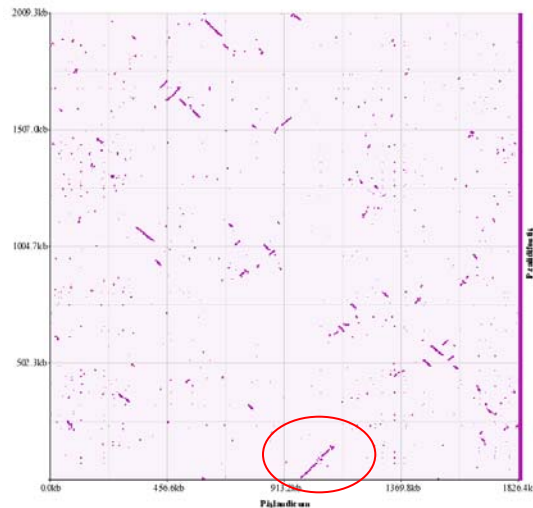
P. aerophilum

P. arsenaticum



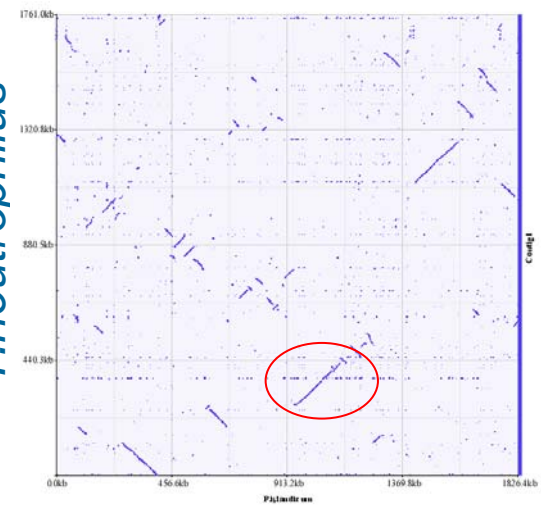
P. islandicum

P. calidifontis



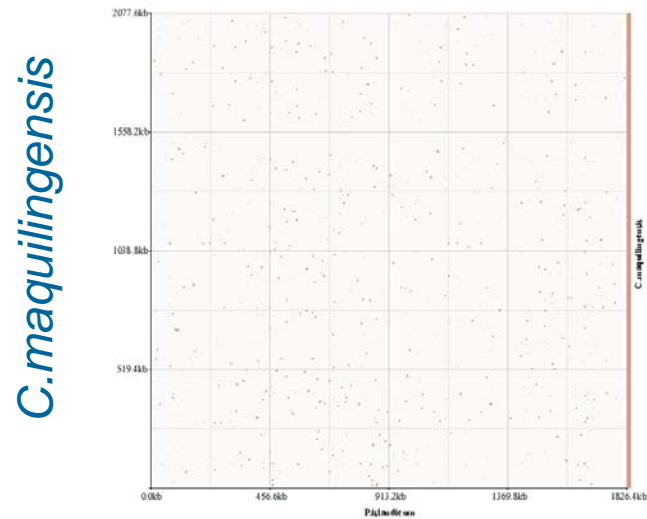
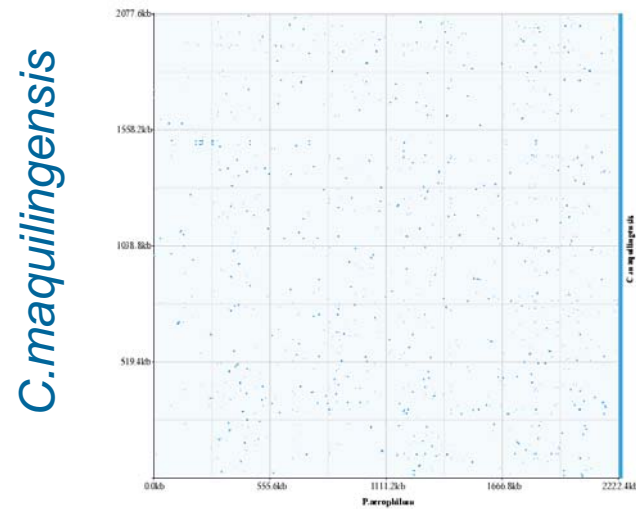
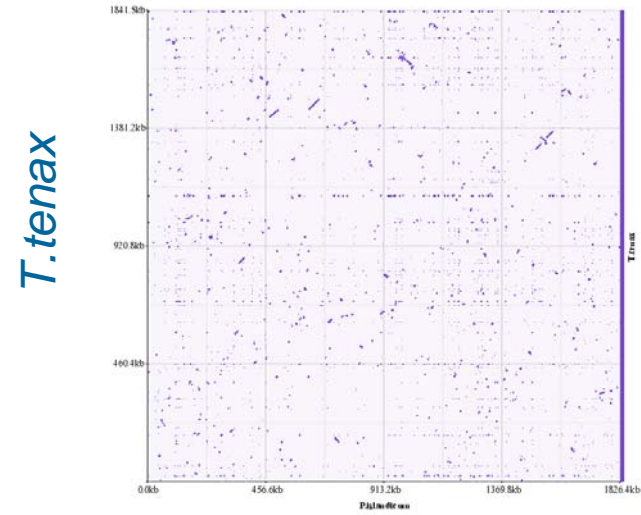
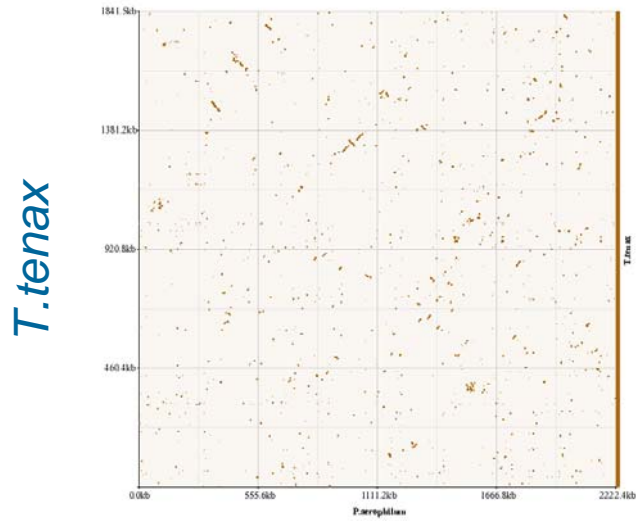
P. islandicum

T. neutrophilus



P. islandicum

Outside-Genus Dot Plots



Unusually Short Operons in Pyrobaculum

- Average operon is relatively short
- Even highly conserved long operons are short in Pyrobaculum

Average Length of Predicted Operons (Arkin Lab)

	Operons	Genes/Operon
<i>P. aerophilum</i>	414	2.72
<i>Halobacterium</i> <i>sp.</i>	306	2.76
<i>M. acetivorans</i>	760	2.85
<i>S. sulfolobus</i>	554	2.90
<i>P. abyssi</i>	344	3.25
<i>P. furiosus</i>	426	3.17
<i>E. coli</i>	834	3.34

Ribosomal Protein Gene Cluster #1 (table from KEGG)

Halobacterium

Species	Genes										<i>Halobacterium</i>		
H mja [P G T]	MJ0324 MJ0495	MJ0322	MJ0176	MJ0177	MJ0178	MJ0179	MJ0180	MJ0460	MJ0461		MJ0462	MJ0463	MJ0464
mac [P G T]	MA1256	MA1255	MA1072	MA1073	MA1074	MA1075	MA1076	MA1077	MA1078		MA1079	MA4059	MA1080
T mth [P G T]	MTH1058 MTH1185	MTH1059	MTH2	MTH3	MTH4	MTH5	MTH6	MTH7	MTH8		MTH9	MTH10	MTH11
H mka [P G T]	MK0247 MK0148	MK0246	MK0415	MK0414	MK0413	MK0412	MK0990	MK0841	MK0842		MK0843	MK0844	MK0846 MK1216
H afu [P G T]	AF0937 AF0744	AF0938	AF1925	AF1924	AF1923	AF1922	AF1921	AF1920	AF1919		AF1918	AF0914	AF1917
hal [P G T]	VNG2649G VNG1676G	VNG2648G	VNG1689G	VNG1690G	VNG1691G	VNG1692G	VNG1693G	VNG1695G	VNG1697G		VNG1698G	VNG2584C	VNG1699C
T tac [P G T]	Ta0444	Ta0445	Ta1271	Ta1270	Ta1269	Ta1268	Ta1267	Ta1266	Ta1265				
H pho [P G T]	PH1484	PH1483	PH1777	PH1776	PHS049	PH1775	PH1774	PH1773	PH1772		PHS048		PH1771
H pab [P G T]	PAB0465	PAB0466	PAB2120	PAB2121	PAB7083	PAB2122	PAB2123	PAB2396	PAB2125		PAB8082	PAB7082	PAB2126
H ape [P G T]	APE_1844	APE_1842	APE_0227	APE_0228.1	APE_0228a.1	APE_0218	APE_0367.1	APE_0365	APE_0363		APE_0362a	APE_1629	APE_0362
H sso [P G T]	SSO0216 SSO2429	SSO0215	SSO0719	SSO0718	SSO6401	SSO0716	SSO0715	SSO0713	SSO0712		SSO6397	SSO5866	
pai [P G T]	PAE1764 PAE3000	PAE2907	PAE1970	PAE1971	PAE1972	PAE0803	PAE1729	PAE1782	PAE1779		PAE1778	PAE3340	

P. aerophilum

- Scattered to 7 loci

H = Hyperthermophiles
T = Thermophiles

Ribosomal Protein Gene Cluster #2

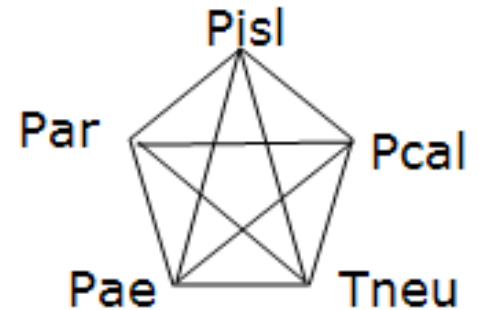
Halobacterium

aae [P G T]	aq_020	aq_1654	aq_1653		aq_1652	aq_1651a	aq_1651	aq_1649		aq_1648	aq_1645	aq_1644	aq_1642	aq_079	
mja [P G T]	MJ0465	MJ0466	MJ0467	MJ0468	MJ0469	MJ0469.1	MJ0470	MJ0471	MJ0472	MJ0473	MJ0474	MJ0475	MJ0476	MJ0477	MJ0478
mac [P G T]	MA1081 MA2024	MA1082	MA1083	MA1084	MA1085	MA1086	MA1087	MA1088	MA1089	MA1090	MA1091	MA1092	MA1093	MA1094	MA1095
mth [P G T]	MTH12	MTH13	MTH14	MTH15	MTH16	MTH17	MTH18	MTH19	MTH20	MTH21	MTH22	MTH23	MTH24	MTH25	MTH26
mka [P G T]	MK1217	MK1218	MK1219	MK1220	MK1221	MK1222	MK1223	MK1224	MK1225	MK0031	MK0030	MK0029	MK0028	MK0027	MK0026
afu [P G T]	AF1916	AF1915	AF1914	AF1913	AF1912	AF1911	AF1910	AF1909	AF1908	AF1907	AF1906	AF1905	AF1904	AF1903	AF1902
hal [P G T]	VNG1700G	VNG1701G	VNG1702G	VNG1703G	VNG1705G	VNG1706G	VNG1707G	VNG1709G	VNG1711G	VNG1713G	VNG1714G	VNG1715G	VNG1716G	VNG1718G	VNG1719G
tac [P G T]	Ta1262	Ta1261		Ta1259	Ta1258	Ta1257	Ta1256	Ta1255	Ta1254	Ta1253	Ta1252m	Ta1251	Ta1250m	Ta1249	Ta1248
pho [P G T]	PH1770	PH1768	PH1767	PH1766	PH1765	PHS047	PH1764	PH1763	PH1761	PH1759	PH1758	PH1757	PH1756	PH1755	PH1754
pab [P G T]	PAB2127	PAB2436	PAB2128	PAB2397	PAB2130	PAB7080	PAB2131	PAB2132	PAB2133	PAB2134	PAB2135	PAB2136	PAB2137	PAB2138	PAB2139
ape [P G T]	APE_0360.1	APE_0359	APE_0358	APE_0356.1	APE_0354.1	APE_0352a	APE_0352.1	APE_0350	APE_0349	APE_0348	APE_0347	APE_0346	APE_0345.1	APE_0343	APE_0983.1
ssu [P G T]	SSO0709	SSO0708	SSO0707	SSO0705	SSO0704	SSO6391	SSO0703	SSO0702	SSO0701	SSO0700	SSO0699	SSO0698	SSO0697	SSO0696	SSO0695
pai [P G T]	PAE1730	PAE3136	PAE3312	PAE3313	PAE3575	PAE2097	PAE2098	PAE2377	PAE2099	PAE2100	PAE2101	PAE1182	PAE1183	PAE3436	PAE3435

P. aerophilum

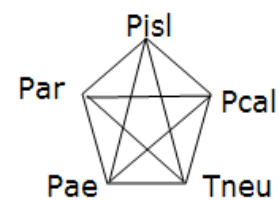
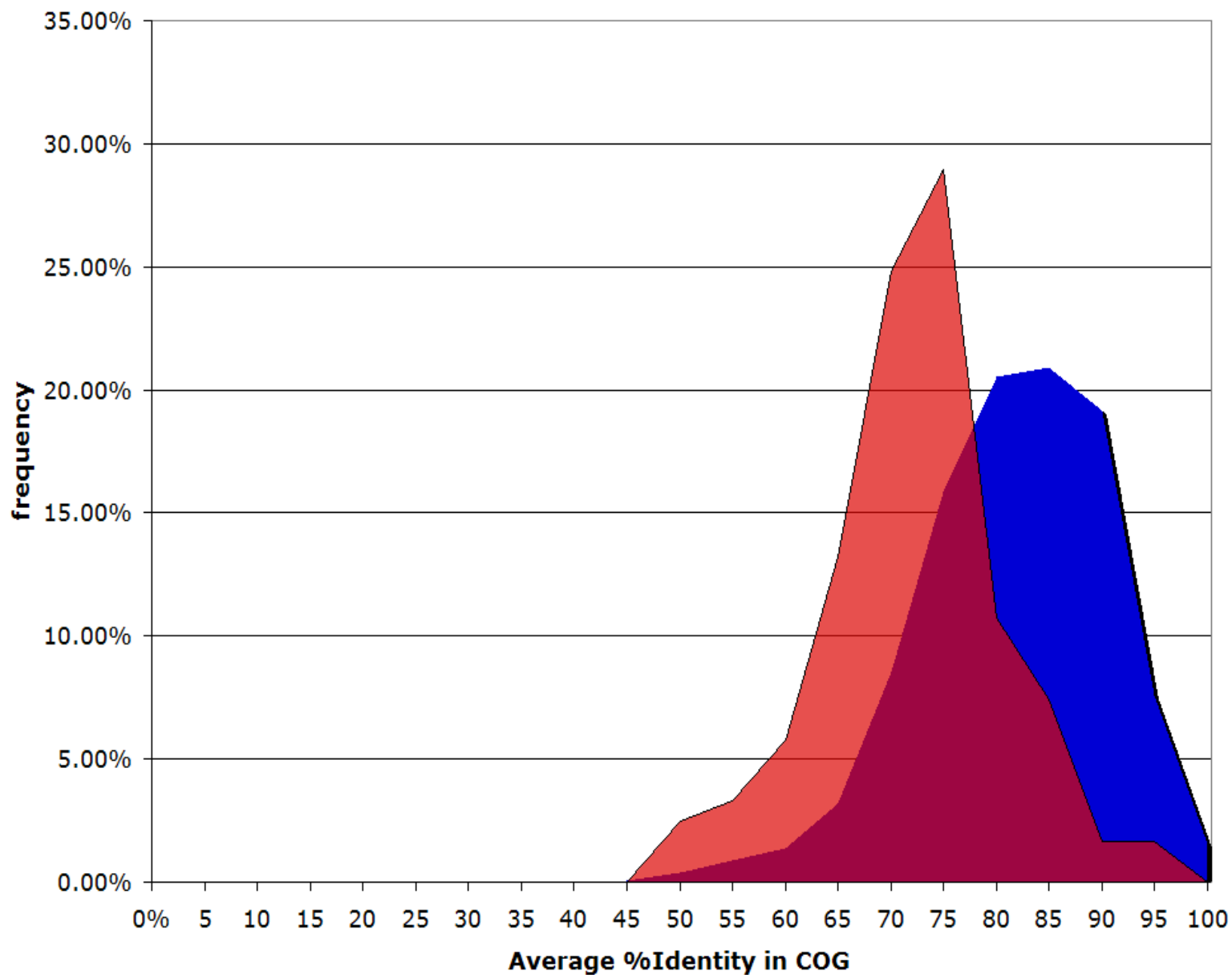
- Scattered to 9 loci

Pyrobaculum Core Set



- 1369 Reverse Best-Blast (COG-like) ortholog groups found in all Pyrobaculum species (Core)
- Pyrobaculum-specific Core
 - 121 of these are found *only* in Pyrobaculum
 - Zero hits to Pfam protein domains
 - All but 3 labelled as “hypothetical”

Mean Protein %identity between *Pyrobaculum* core genes



■ Pyrobaculum Core
■ Only in Pyrobaculum

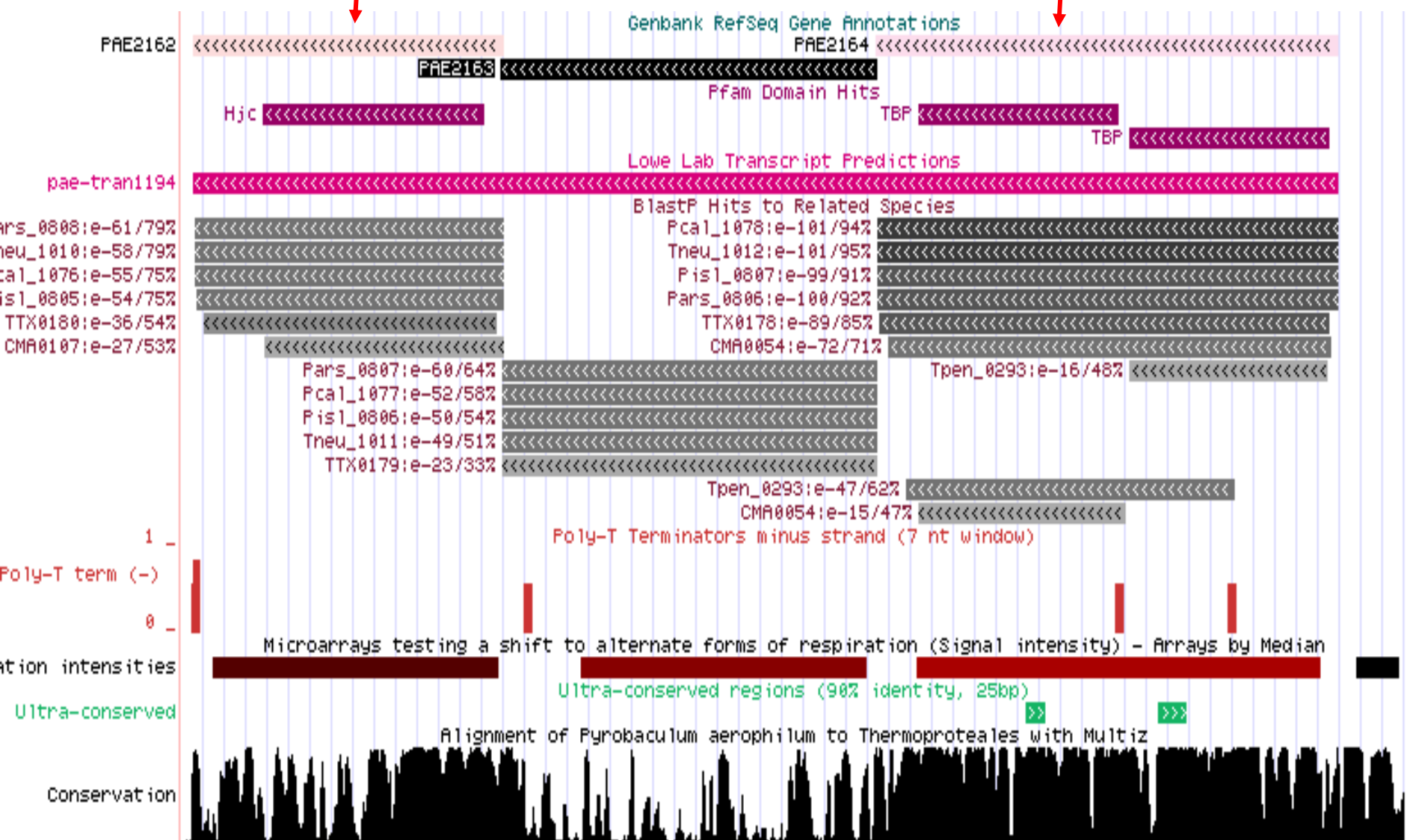
Protein % Identity against *P. aerophilum*

Species	Core Genes (N=1369)	Pyrobaculum-specific Genes (N=121)
<i>P. arsenaticum</i>	76.6%	65.4%
<i>P. islandicum</i>	73.7%	62.5%
<i>Thermoproteus neutrophilus</i>	73.2%	61.4%
<i>P. calidifontis</i>	72.1%	59.2%
<i>Thermoproteus tenax</i>	58.2% (N=1208)	41.8% (N=85)
<i>Caldivirga Maquilingensis</i>	46.7%	N/A

New Pyrobaculum Elaboration on an Old System?

Holiday junction resolvase

TATA Binding Protein



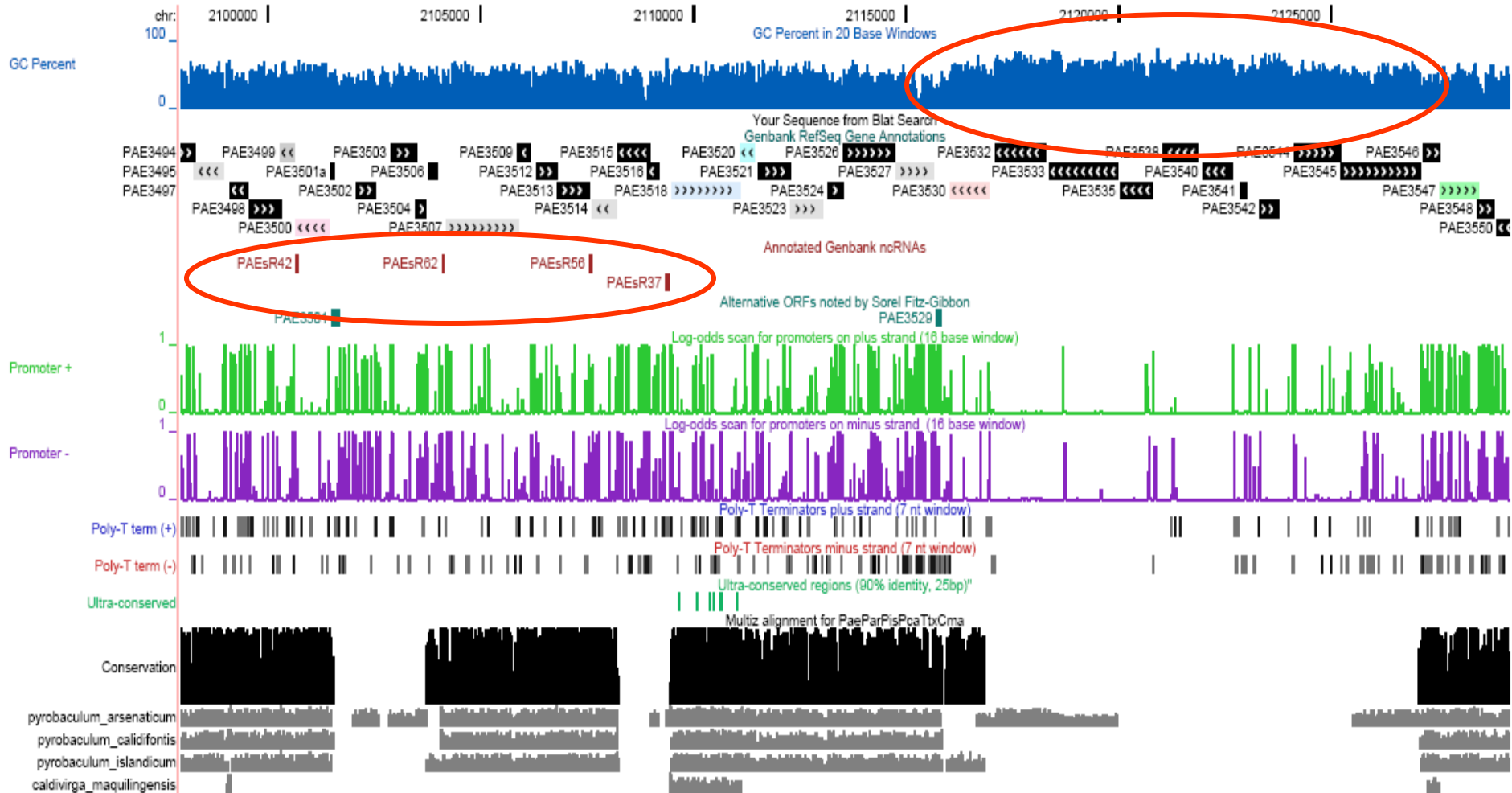
Conserved Gene Neighborhoods

- Definition: Genes are in same neighborhood if:
 - they are on same strand
 - separated by maximum of 2 genes
 - like “operons” but not necessarily driven by same promoter
- **285** conserved neighborhoods
 - 772 predicted promoters / transcripts
 - **2.7 promoters / neighborhood**

911/1369 (68%) of core genes found in a conserved neighborhood

Therefore => Strong suppression of shuffling within core gene neighborhoods

“Viral” / Foreign DNA Insertions?



- Often flanked by tRNAs OR C/D box sRNA genes
- Insertion regions extremely numerous, roughly 1 / 10kb
- Region circled upper right, high G/C content

Large Insertions Across Genome v. Sudden Changes in G/C Content



How Fluid Are these Genomes?

Observations:

1. Operons are relatively short
2. Many insertions in genomes, many with non-native G/C content
3. Excessive number of “young” introns in tRNAs

Many many odd-ball introns

Non-canonical tRNA introns in each new *Pyrobaculum* genome are more numerous than those found in all other sequenced species

Species	tRNAs w/ Non-canonical introns	Total non-canonical introns
<i>P. aerophilum</i>	19	19
<i>P. calidifontis</i>	33	45
<i>P. arsenaticum</i>	15	15
<i>P. islandicum</i>	26	32
<i>Caldivirga maquilingensis</i>	2	2
<i>A. pernix</i>	5	5

Non-canonical tRNA introns change frequently among different species

```
P.aero  ggcgggcccggtgggattc-----gaaccacgacctacggcttaggaggctcggcggaccgaggtgggtccaccgcgctct-atcctgactgagctacgggc
P.arse  ggcgggcccggtgggattc-----gaaccacgacctacggcttaggagg-----cgccgctct-atcctggctgagctacgggc
P.cali  ggcgggcccggtgggattcggtaggeggtttgcctgaaccacgacctacggcttaggagg-----ccgccgctct-atcctgactgagctacgggc
P.isla  ggcgggcccggtgggattc-----gaaccacgacctacgggttaaagcctcggcagacctagccgggtcct-cgccgctct-aaccgggtgagctacgggc
C.maqu  ggcgggcccggtgggattt-----gaaccacgacctacggcttaggagg-----cgccgctctaatacctgctgagctacgggc
```

```
P.aero  ccccggtagctcagtcggcagag-----cggcgg-----gctgcag-----accgtaggtcccgggttcaa
P.arse  ccccggtagctcagtcggcagagtgctgcagagttgcaccgggagagcggcgg-----gctgcag-----accgtaggtcccgggttcaa
P.isla  ccccggtagctcagtcggcagag-----cggcggaggaccgcccagtcggcggcagcgcgaggttacggcagccaccggtaggtcccgggttcaa
C.maqu  ccccggtagctcagcacggtagag-----cggcgg-----gctgcag-----accgtaggtcccgggttcaa
```

- Mechanism of intron insertion is not known
- All tRNAs are single copy, so removal is absolutely required to be functional

In Summary

- *Pyrobaculum* & *Caldivirga* genomes are complete, all are publicly available
- Unusually short operons but highly conserved neighborhoods
- Extensive tRNA intron insertions and “foreign DNA” inserted into all *Pyrobaculum* species
 - Possible adaptive advantage?
- Many new non-coding RNA gene candidates, and other repeat families

Future Directions

- Complete comparative study of metabolic genes/pathways among 5 species
- Identification of all non-coding RNA genes in known *Pyrobaculum*
- Development of a genetic system for *Pyrobaculum calidifontis*
- Identification of conserved *Pyrobaculum* transcription binding sites genome-wide using existing microarray data

Acknowledgements

Joint Genome Institute
Sequencing Team

Oak Ridge National Labs
Annotation Team

Pyrobaculum Sequencing
Consortium

Sorel T. Fitz-Gibbon
University of California, Los Angeles

Christopher H. House
Pennsylvania State University

Chad Saltikov
University of California, Santa Cruz

Lowe Lab

David Bernick
Patricia Chan
Aaron Cozen
Matt Weirauch
Matthias Hoechsman

Andrew Holmes
Shawn Yost