Biosimplicity:
Engineering Simple Life

Tom Knight
MIT Computer Science and
Artificial Intelligence Laboratory

Simplicity is the ultimate sophistication
-- Leonardo da Vinci

Measuring Complexity

- Number of components
- Size of description
- Size of functional model

Have you ever thought... about whatever man builds, that all of man's industrial efforts, all his calculations and computations, all the nights spent over working draughts and blueprints, invariably culminate in the production of a thing whose sole and guiding principle is the ultimate principle of simplicity?
In any thing at all, perfection is finally attained, not when there is no longer anything to add, but when there is no longer anything to take away.

-- Antoine de Sainte Exupery, in "Wind, Sand, and Stars"

Engineered Simple Organisms

- modular
- understood
- malleable
- low complexity

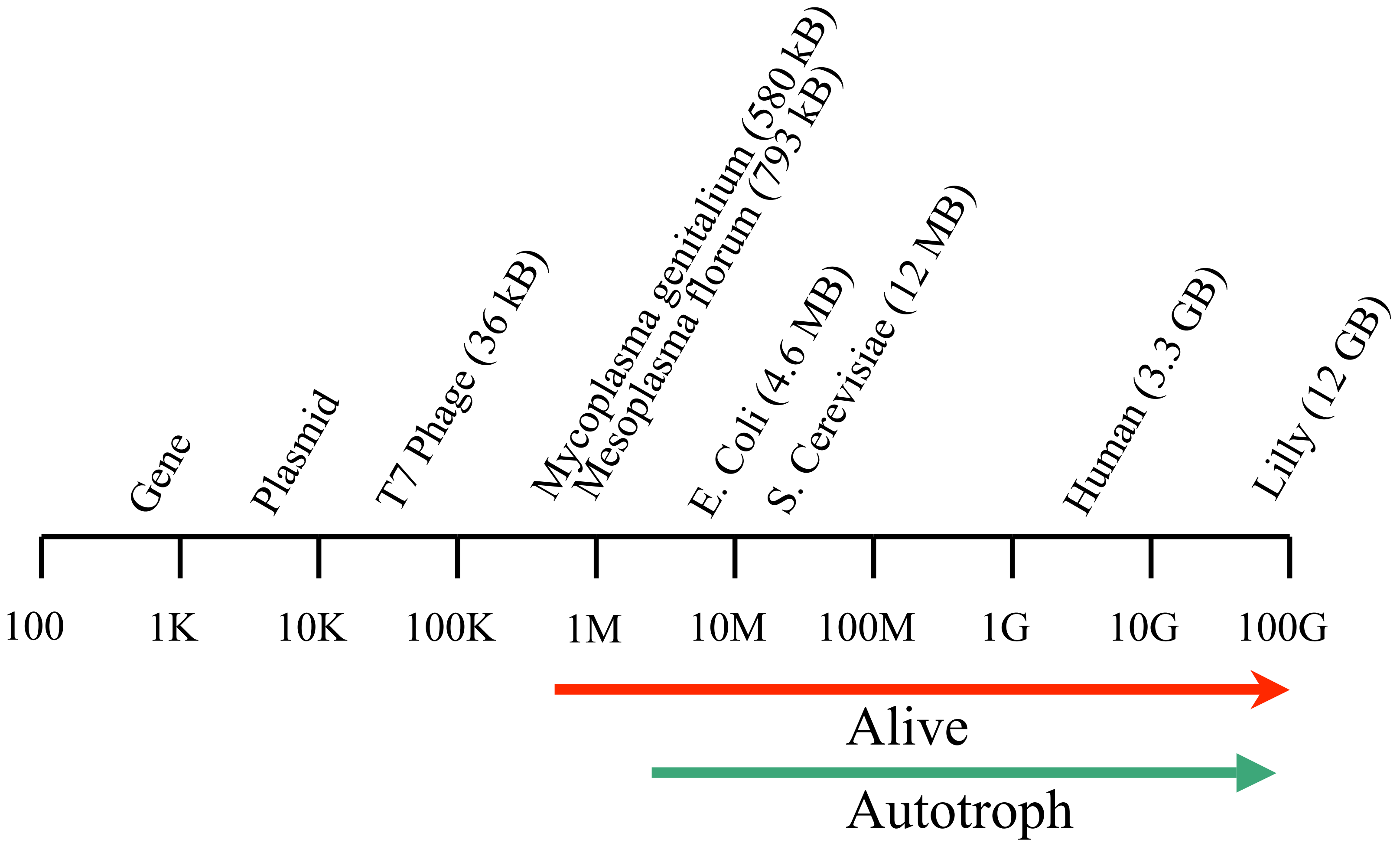
- Start with a simple existing organism
- Remove structure until failure
- Rationalize the infrastructure
- Learn new biology along the way

The chassis and power supply for our computing

Some history…

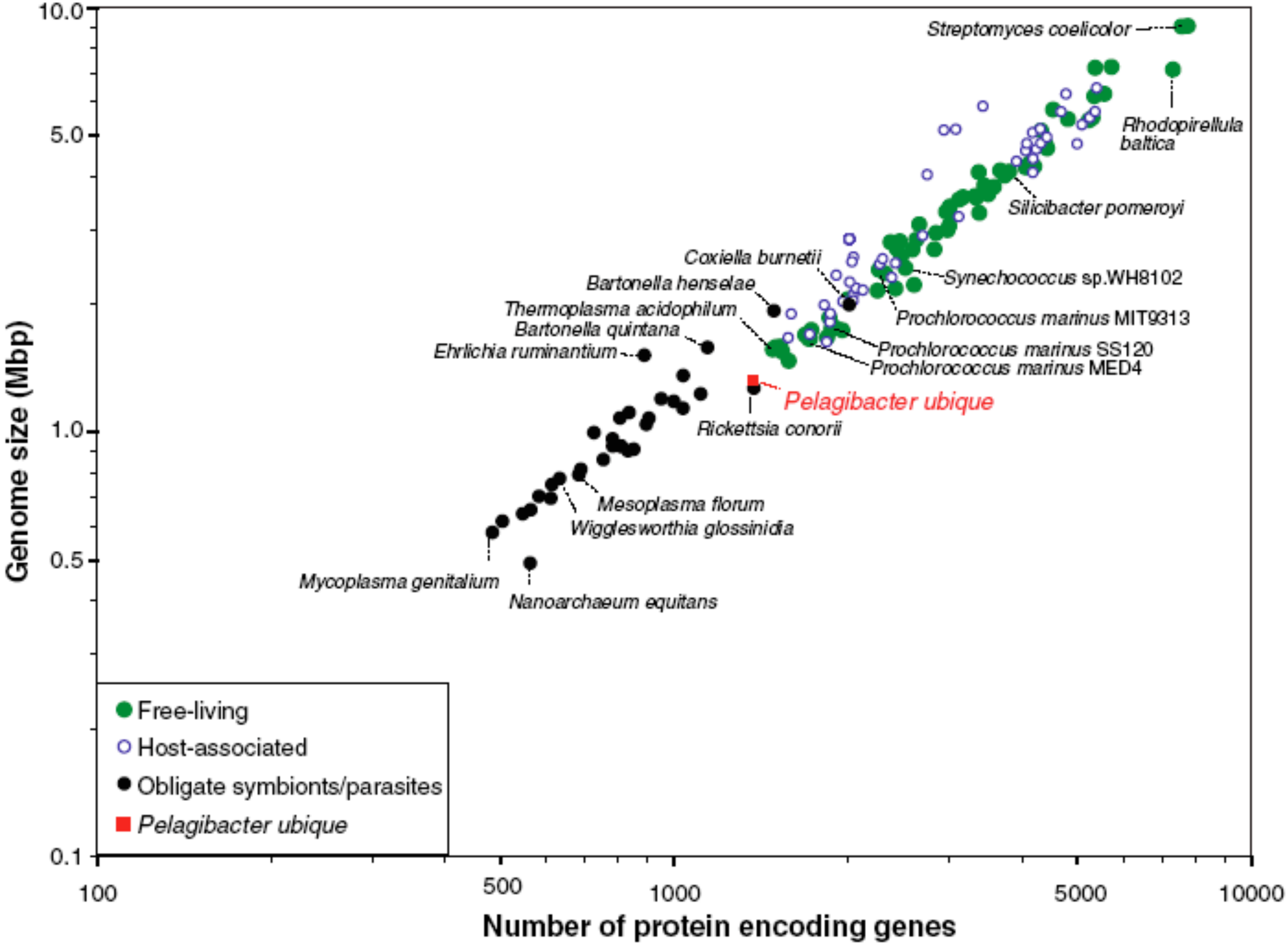
- Confusion over what PPLO/Mycoplasma were
  - “The Microbe of pleuorpneumonia” Nocard 1896
- 1932 isolation of “PPLO.” Koch postulates.
- 1958 Klieneberger-Nobel identifies them as free living bacterial species
- Morowitz 1962 SciAm: “the smallest living cell”
- 1980 Gilbert effort to sequence *M. capricolum*
- 1982 Morowitz “complete understanding of life”
- 1996 Fraser et al. *M. genitalium* sequence
- 1999 Hutchison et al. Minimal genome set for *M. genitalium*

Relative Complexity



Log Genome Size, base pairs

From Giovannoni et al. 2005



Choosing an organism

- Safe
  - BSL-1 organism -- insect commensal
- Un-regulated
  - Not a crop plant or domesticated animal pathogen
- Fast growing
  - 40 minute doubling time
  - vs. six hours for M. genitalium
- Convenient to work with
  - Facultative anaerobe
- Small genome
- Known sequence
- Complete annotation

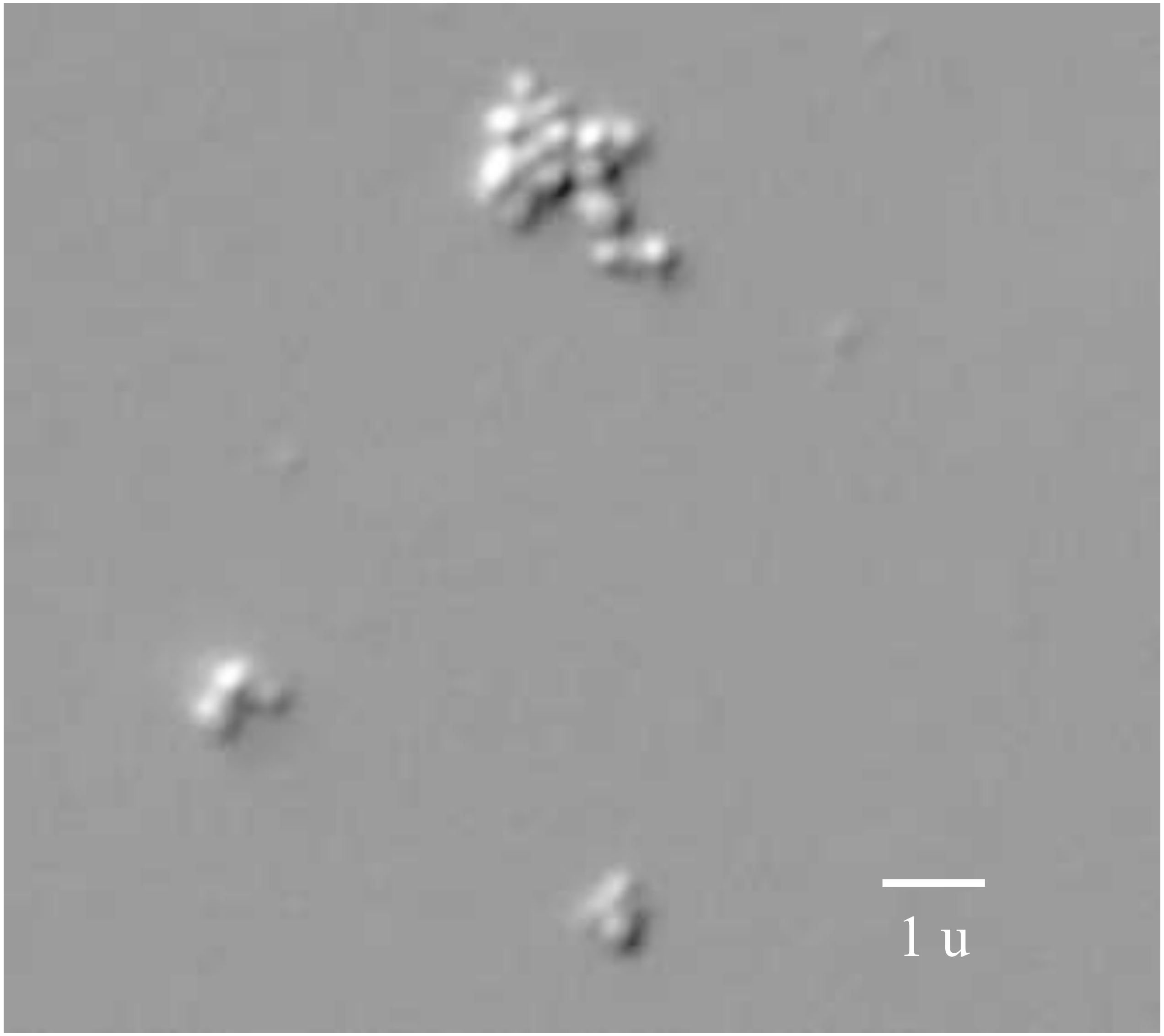
The Mollicute Bibliome

Complete collection of mycoplasma related papers:

- 6,418 and counting
- All books and book chapters also
- Endnote & Refworks
- Downloaded .pdfs for articles > 1995
- Scanned articles and books, OCR with Abbyy Finereader
- Plans for a Google appliance search engine
- Collaboration for “shallow semantic” understanding
- people.csail.mit.edu/tk/mfpapers/ user=meso, pass=meso

200 Micron

Mesoplasma florum



Tomographic TEM

Courtesy
Jensen Lab,
Caltech

Culture Medium 1161

- Beef Heart infusion
- 4% Sucrose
- Fresh yeast culture broth
- 20% horse serum
- Penicillin
- Phenol red

Defined medium

Sodium phosphate
KCl
Magnesium sulfate

Glycerol
Spermine

Nicotinic acid
Thiamine
Riboflavin
Pyridoxamine
Thioctic acid
Coenzyme A

Amino acids (minus asp, glu)
Guanine
Uracil
Thymine
Adenine

Glucose

BSA
Palmitic acid
Oleic acid

Synthesis vs. Import

- Mycoplasma import virtually all small biochemical molecules
- Each import is done with a specific membrane protein – some are capable of importing a class
- Complexity is reduced if the import is simpler than the synthesis
- Example of the opposite:
  - Glutamine → Glutamic acid
  - Asparagine → Aspartic acid

Sequencing the genome

- First sequenced small portions of the genome to test that we had the correct species
  - Compared the results to Genbank entries
  - Sequenced PTS system gene, identical to reported sequence
  - Sequenced 16S rRNA (unreported)
  - Discovered identical to Mesoplasma entomophilum 16S rRNA sequence – probably the same species
    - Genbank entry
- Measured genome size with pulse field gels
- Sequenced 12% of genome to see what we were up against

PFGE of Mesoplasma genomic DNA

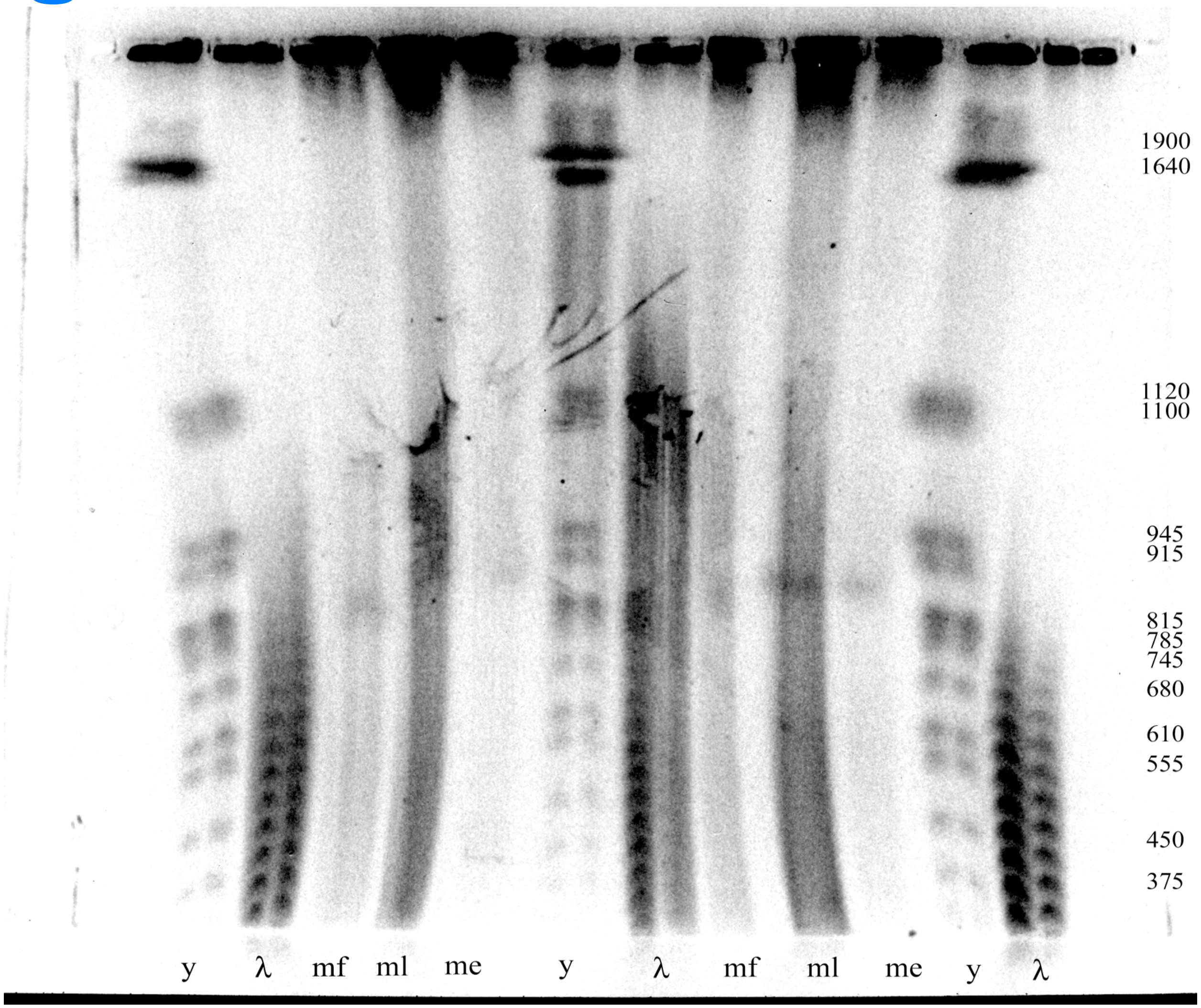
y
yeast marker

l
lambda marker

me
Mesoplasma entomophilum

mf
Mesoplasma florum

ml
Mesoplasma lactucae



1.2% agarose 9C   6V/cm Ramped 90 - 120 sec 48 hours

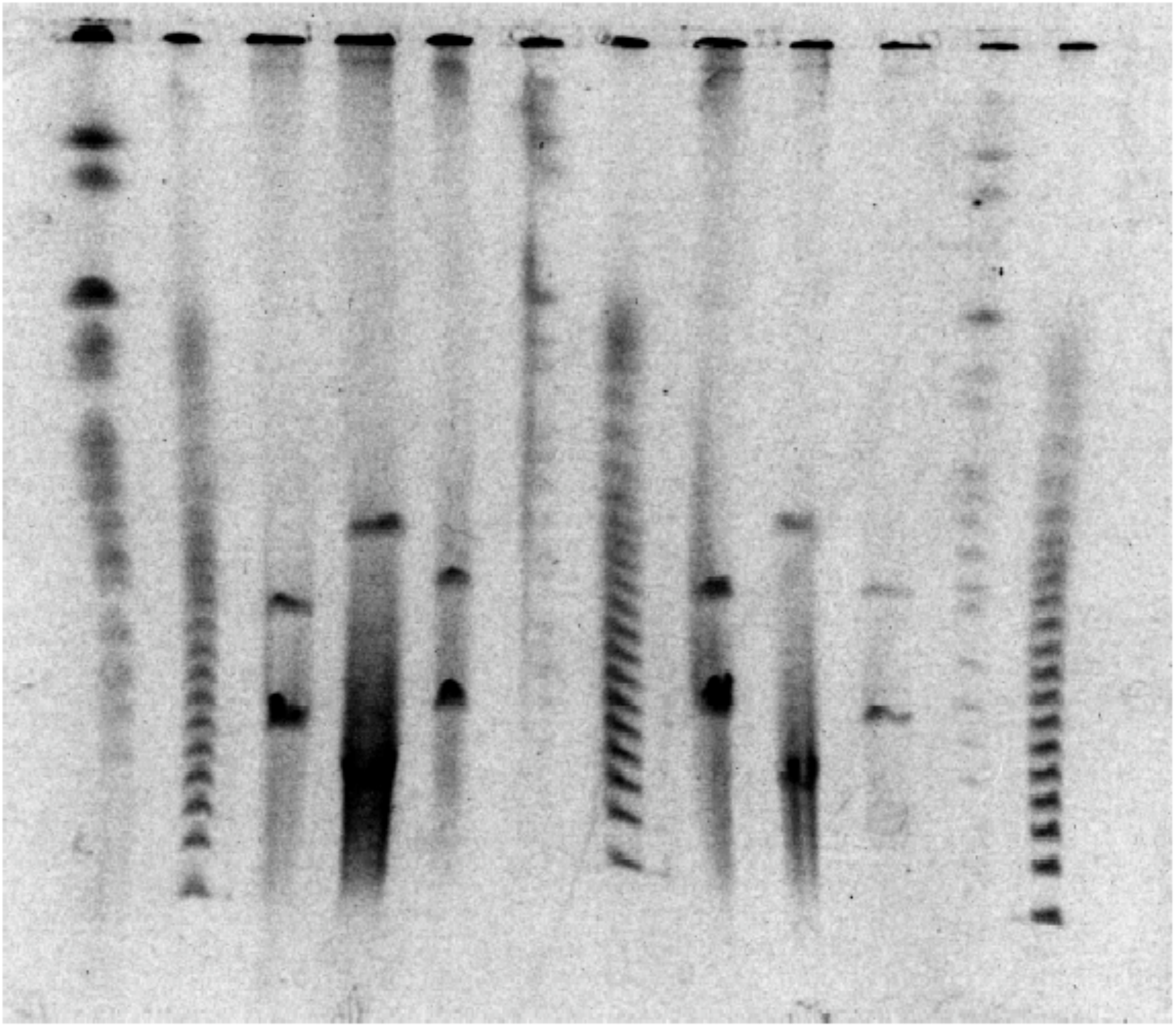
I-CeuI digestion

- Special restriction enzymes cut only at 23S rRNA sites

```
5´ ..TAACTATAACGGTC CTAA^GGTAGCGA..3´
3´ ..ATTGATATTGCCAG^GATT CCATCGCT..5´
```

Calculate the number of rRNA sites in the genome from the number of cut fragments

I-CeuI digests

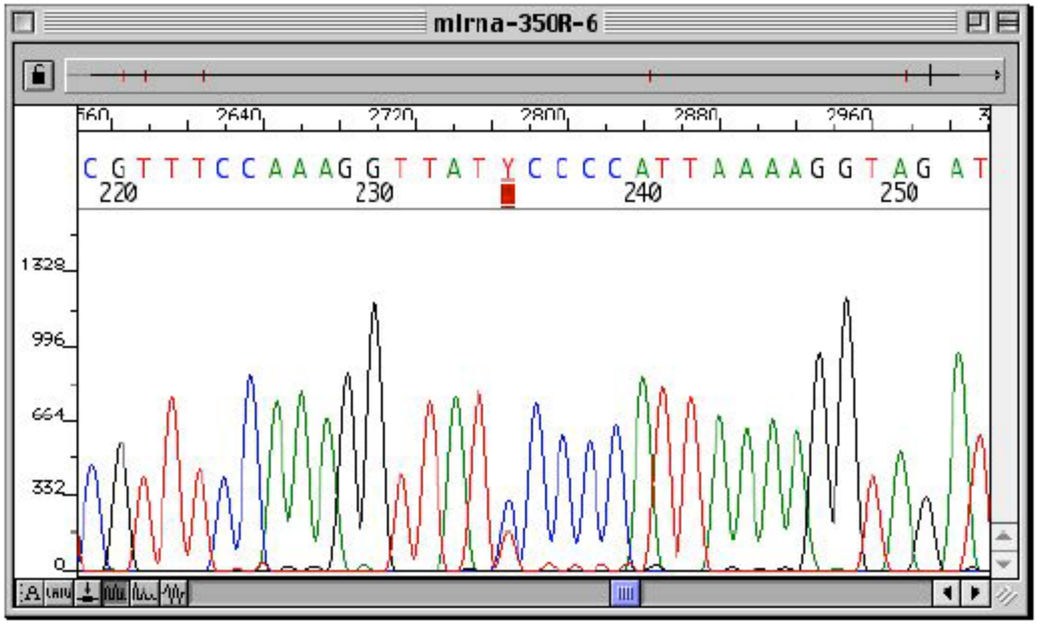


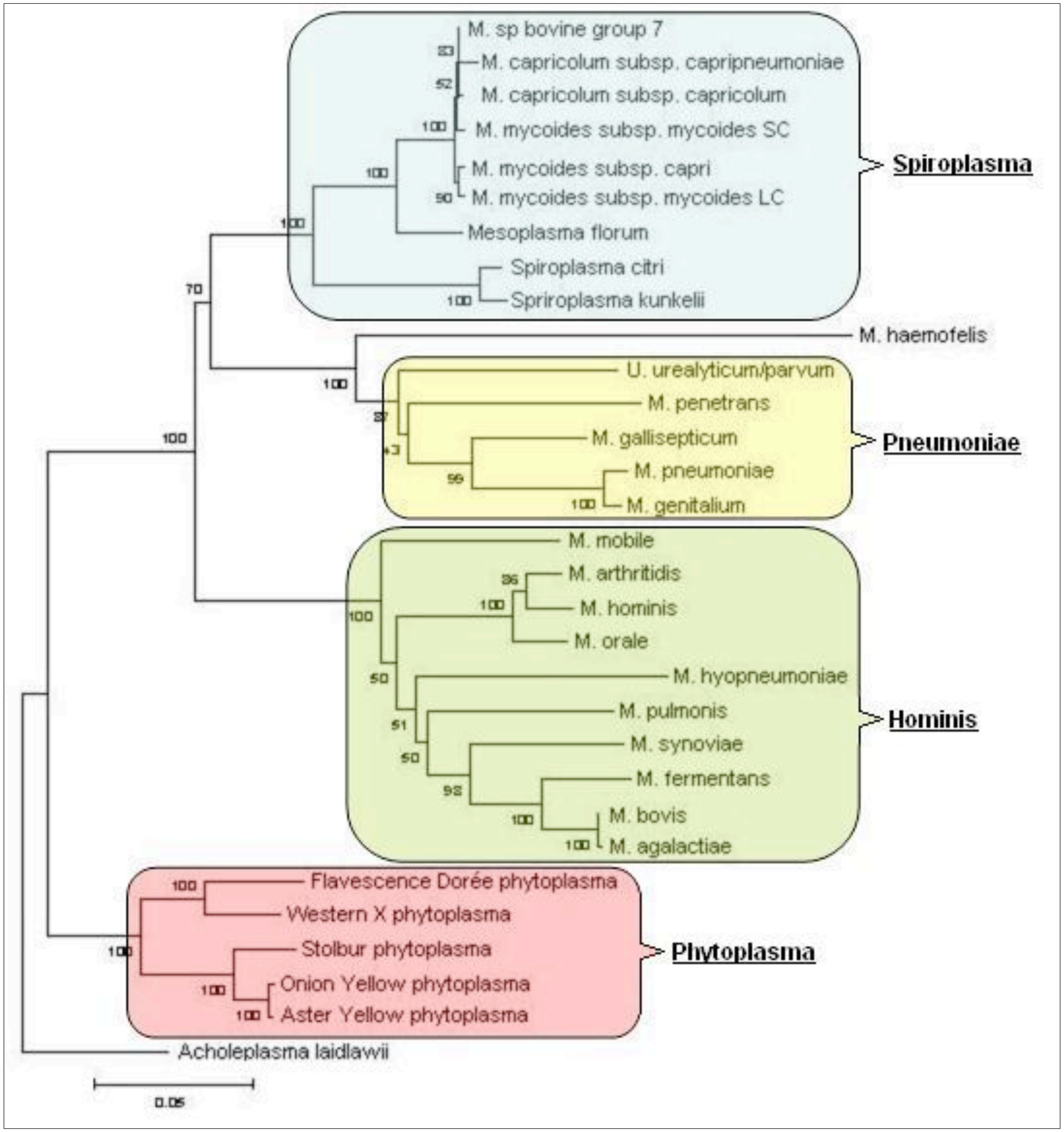
rRNA sequences

- Degenerate primers

Primer	F/R	Sequence 5' to 3'
U1	F	gtt tga tcc tgg ctc agg ayd aac g
U2	F	tca rac tcc tac ggr agg cag c
U5	R	ctt gtg cgg gyy ccc gtc aat tc
U8	R	gaa agg agg trw tcc ayc csc ac
1300	F	taa tcg cga atc agc tat gtc
350	R	tgc ttc atc aga ctt tcg tcc
1000	F	tgg agg tta aca ttg ata cag g
1150	R	cat gat gat ttg acg tca tcc

- The two rRNA sequences are different:


mlrna-350R-6
560
2640
2720
2800
2880
2960
3
C G T T T C C A A A G G T T A T Y C C C C A T T A A A A G G T A G A T
220
230
240
250
1328
996
664
332
0


M. sp bovine group 7
83
M. capricolum subsp. capripneumoniae
52
M. capricolum subsp. capricolum
100
M. mycoides subsp. mycoides SC
100
M. mycoides subsp. capri
90
M. mycoides subsp. mycoides LC
100
Mesoplasma florum
Spiroplasma citri
100
Spriroplasma kunkelii
Spiroplasma
70
M. haemofelis
100
U. urealyticum/parvum
87
M. penetrans
43
M. gallisepticum
99
M. pneumoniae
100
M. genitalium
Pneumoniae
100
M. mobile
86
M. arthritidis
100
M. hominis
100
M. orale
50
M. hyopneumoniae
51
M. pulmonis
50
M. synoviae
98
M. fermentans
100
M. bovis
100
M. agalactiae
Hominis
100
Flavescence Dorée phytoplasma
Western X phytoplasma
100
Stolbur phytoplasma
100
Onion Yellow phytoplasma
100
Aster Yellow phytoplasma
Phytoplasma
Acholeplasma laidlawii
0.05

Library creation

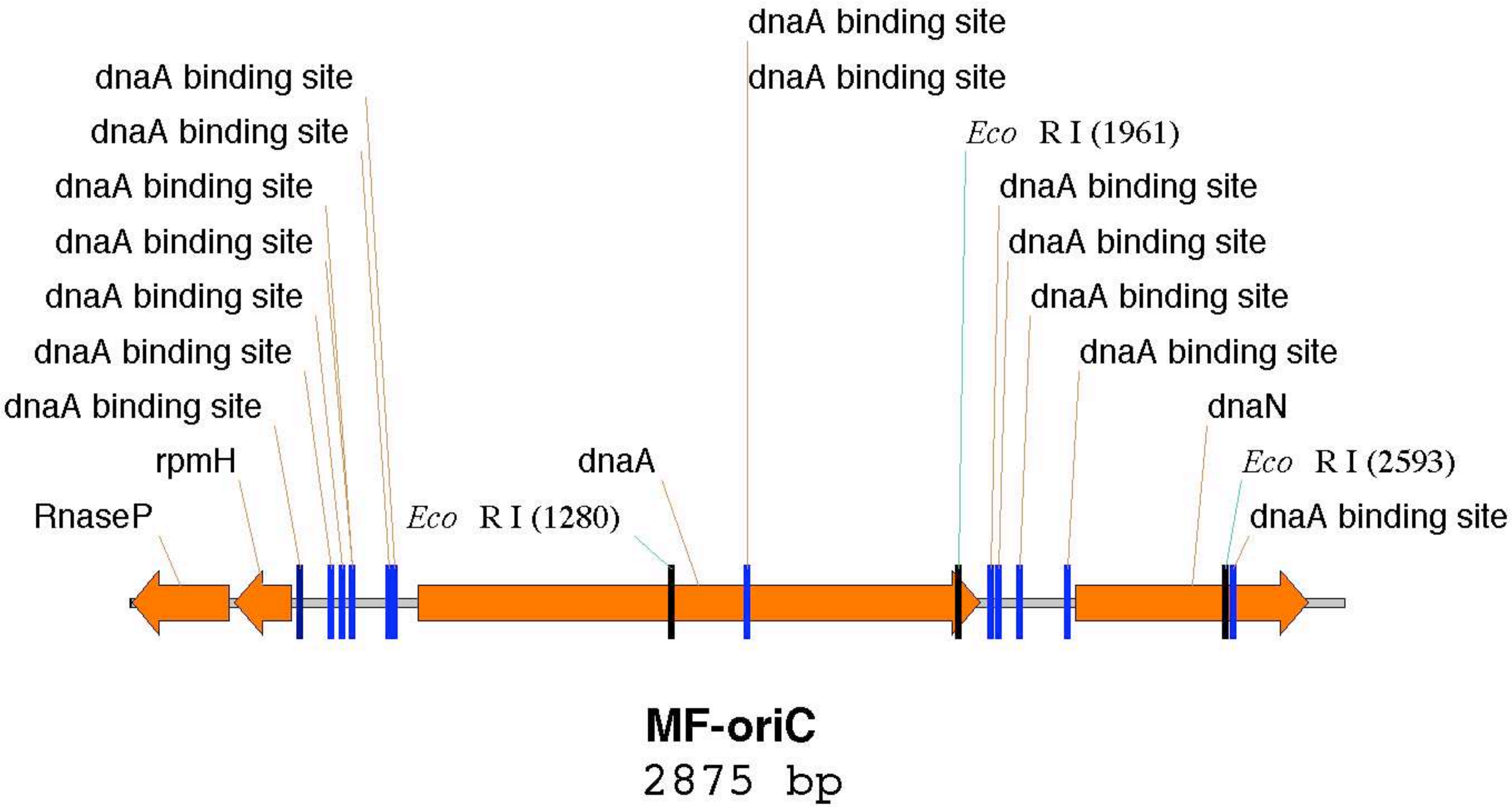
- Randomly cut genomic DNA with EcoRI
- Shotgun cloned into pUC18 vector
- Sequenced the inserts (0.1 – 8 Kb)

- Sheared the genomic DNA with a needle
- End repaired
- Cloned into defective lambda phage vector
- Packaged vector into phage heads
- Infected E. coli cells with phage
  - ~ 40 Kb inserts

What we learned from partial sequence

- Almost all “old friends”
- Little or no extra junk
- Inter-gene sequences small (-4 to 30 bp)
- Little transcriptional control
- High AT vs. GC content – 27% average GC
  - 20% in typical genes, even less in control regions
  - 40% in rRNA regions

Origin of Replication



Whitehead Agreement

- Whitehead agreed in January 2002 to sequence the organism
- Estimated to take about two hours of time on their sequencers
  - “Sure, we can do it Tom, but what do we do with the rest of the day after the coffee break?”
- “How many other organisms like this are there?”
  - 300
- “Why don’t we sequence them all?”
- Good draft available November 2002
- Gaps closed July 2003 -- final November 2003

Gap closure

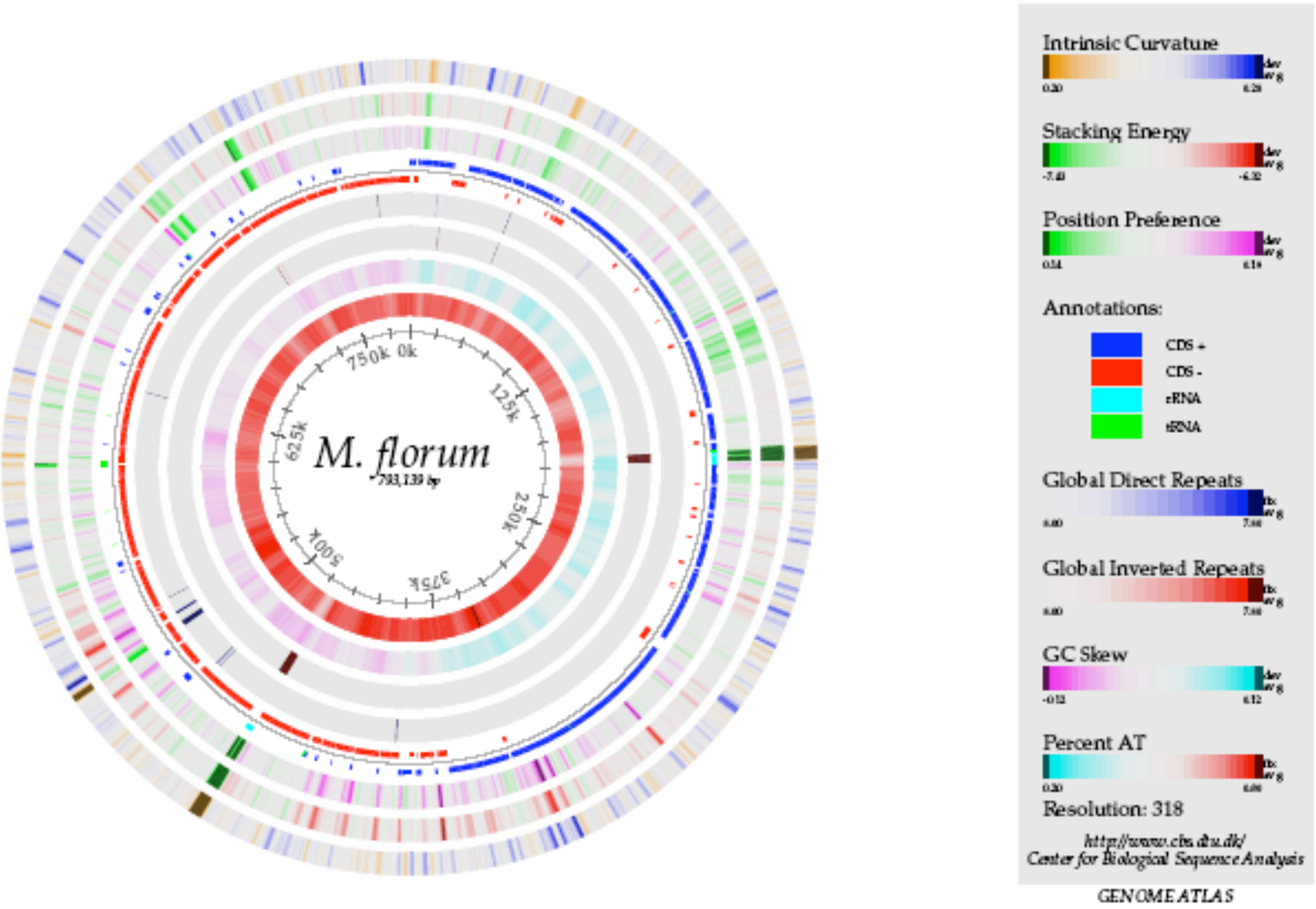
- 9 gaps remained
- Long range PCR
- Primer walking
- One difficult sequence
  - Poly A region 16-17 bp long
  - Sequencing stuttered
  - Reprime with aaaaaaaaaaaaaaaag
- Repeat region 186 bp in surface lipoprotein
  - Give up on accurate sequence, PCR for length
- Final assembly verification

Genome characteristics

- 793281 base pairs
- 26.52% G + C
- 682 protein coding regions
  - UGA for tryptophan
  - No CGG codon or corresponding tRNA
  - Classic circular genome
  - oriC, terminator region, gene orientation
- 39 stable RNAs
  - 29 tRNAs
  - 2x 16S, 23S, 5S
  - RNAse-P, tmRNA, SRP

Standard Motifs

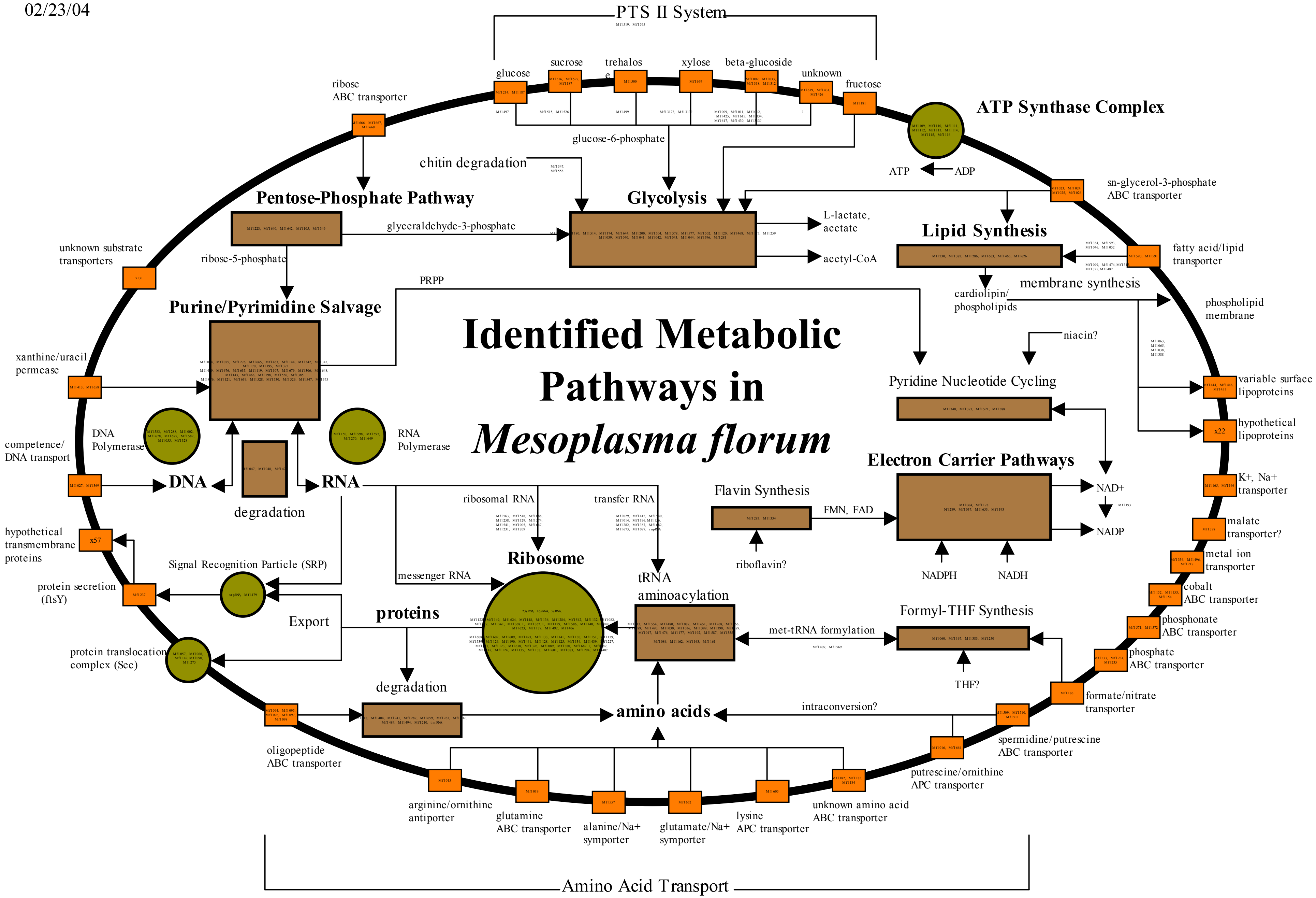
- -10 present usually very highly conserved
  - Often preceded by a “TG” 1-2 bp upstream
- Seldom a conserved -35 region
- RBS is standard Shine-Dalgarno
- Alternate RBS matches complementary region of 16S rRNA UAACAACAU (Loechel 91)
- Standard stem-loop terminators with loop TTAA
  - 6-8 poly T tail in forward direction
- dnaA box TTATCCACA
- Four ribo-box sequences (Thi, Ile, Val, Guanine)


M. florum
793,139 bp
0k
125k
250k
375k
500k
625k
750k
Intrinsic Curvature
0.20
0.28
dev
avg
Stacking Energy
-7.40
-6.32
dev
avg
Position Preference
0.14
0.18
dev
avg
Annotations:
CDS +
CDS -
rRNA
tRNA
Global Direct Repeats
0.00
7.00
fix
avg
Global Inverted Repeats
0.00
7.00
fix
avg
GC Skew
-0.12
0.12
dev
avg
Percent AT
0.20
0.80
fix
avg
Resolution: 318
http://www.cbs.dtu.dk/
Center for Biological Sequence Analysis
GENOME ATLAS

Understand the metabolism

- Identify major metabolic pathways by finding critical genes coding for known enzymes
- Predict necessary enzymes which may not have been found
- Evaluate the list of unknown function genes for candidates
- Build the major metabolic pathway map of the organism
- Consider elimination of entire pathways

G. Fournier
02/23/04



Identified Metabolic Pathways in *Mesoplasma florum*

PTS II System
glucose
sucrose
trehalose
xylose
beta-glucoside
unknown
fructose
glucose-6-phosphate
ribose ABC transporter
chitin degradation
Pentose-Phosphate Pathway
Glycolysis
glyceraldehyde-3-phosphate
L-lactate, acetate
acetyl-CoA
ATP Synthase Complex
ATP
ADP
sn-glycerol-3-phosphate ABC transporter
Lipid Synthesis
fatty acid/lipid transporter
cardiolipin/ phospholipids
membrane synthesis
phospholipid membrane
unknown substrate transporters
ribose-5-phosphate
PRPP
Purine/Pyrimidine Salvage
xanthine/uracil permease
niacin?
Pyridine Nucleotide Cycling
variable surface lipoproteins
hypothetical lipoproteins
DNA Polymerase
RNA Polymerase
competence/ DNA transport
DNA
RNA
degradation
Electron Carrier Pathways
Flavin Synthesis
FMN, FAD
NAD+
NADP
K+, Na+ transporter
malate transporter?
riboflavin?
NADPH
NADH
metal ion transporter
cobalt ABC transporter
ribosomal RNA
transfer RNA
hypothetical transmembrane proteins
x57
Signal Recognition Particle (SRP)
Ribosome
messenger RNA
tRNA aminoacylation
protein secretion (ftsY)
Export
proteins
met-tRNA formylation
Formyl-THF Synthesis
phosphonate ABC transporter
phosphate ABC transporter
protein translocation complex (Sec)
degradation
THF?
formate/nitrate transporter
amino acids
intraconversion?
oligopeptide ABC transporter
spermidine/putrescine ABC transporter
putrescine/ornithine APC transporter
arginine/ornithine antiporter
glutamine ABC transporter
alanine/Na+ symporter
glutamate/Na+ symporter
lysine APC transporter
unknown amino acid ABC transporter
Amino Acid Transport
x22

How Simple is this?

- Missing cell wall, outer membrane
- Missing TCA cycle
- Missing amino acid synthesis
- Missing fatty acid synthesis
- One sigma factor
- Small number of dna binding proteins
- One insertion sequence, probably not active
- One restriction system (Sau3AI-like)
- CTG/CAG methylation (function?)
- Evidence for shared protein function
  - MDH/LDH (Pollack 97 Crit rev microbiol 23:269)

Minimal is not always simple

- Shared function of parts
- Overlapped genes
- Tradeoff of import vs. synthesis

- Example:
  - Television set design
  - Shared deflection coil, high voltage power supply, isolated filament supply
  - Three functions, a single circuit, a difficult engineering, modeling, debugging, and repair task

- *how many genes have multiple functions*

DNA Methylation

- Bisulfite conversion of genomic DNA
- Sequencing of converted DNA to identify methylated C positions

- Results: GATC sequences, as expected
- Unexpected: CAG and CTG sequences

Current work

- Array experiments
  - Close species
  - Transcriptional units, pseudo-genes
  - TRASH
- Protein species by LC/LC/MS/MS
- Elimination of the restriction system
- Plasmid system
  - pBG7AU based
- Recombination system
  - Positive/negative selection
- Yeast chromosome transfer
- Genome edits to reduce size
- Genome edits to modularize
- Genome edits to eliminate complexity
- Use as a construction chassis

Reconstruct the Genome

- Use recombination techniques to edit the genome
- Eliminate unnecessary genes
- Remove overlaps
- Standardize promoters, ribosomal binding sites
- Identify transcriptional and translational regulators
- Recode proteins to use a reduced portion of the coding space

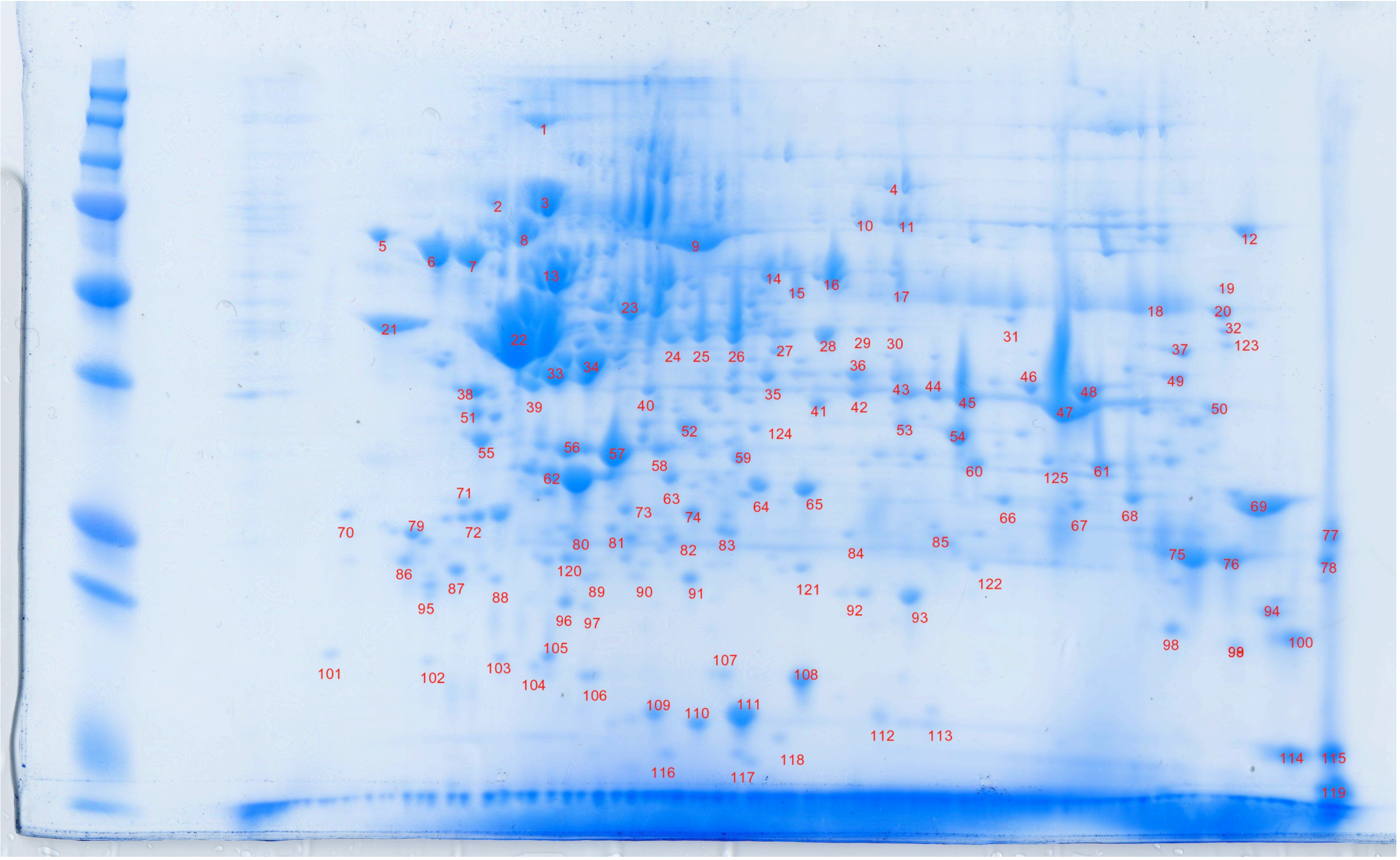
The code is 4 billion years old,
it's time for a rewrite

YAC mutagenesis

- Bring up YAC technology
  - Spheroplasts, old YAC plasmid sequencing
- Triple transform with MF chromosome and
  - pRML1 (Spencer 92)
  - pRML2
  - Genome inactive except for ARS, telomeres, selection markers
- Use yeast recombination systems for genome editing
- Isolate and recircularize YAC to form a new genome
- Lipid encapsulate genome into vesicles
- Fuse vesicles with genome-killed wild type cells

Proteome

- Collaboration with Steve Tannenbaum / Yingwu Wang
- 2-D gels + MS spot ID
- LC/LC/MS/MS ID of trypsin digests



Riboswitch Analysis

- Collaboration with Ron Breaker / Adam Roth
- Discovery of unique riboswitches specific for GTP rather than dGTP
- Found in no other sequenced genomes
- Analysis of close relatives under way

Engineer plasmids

- No known plasmids for this class of organism
- Renaudin has made plasmids using the chromosomal OriC as the replicative element
  - Lartigue 03
- M. mycoides has pADB201 and pBG7AU rolling circle plasmids similar to pE194 (1082 bp)
- We know the antibiotic sensitivities and have working resistance genes

Kit Part the genome

- Make Biobrick parts from each gene, tRNA, promoter, other part-like genome element
- Attempt to develop techniques for recombining parts into coherent modules
- Develop techniques for assembling and modeling the resulting structures

Thanks to...

- Harold Morowitz
- Greg Fournier
- Gail Gasparich
- Bob Whitcomb
- Eric Lander
- Bruce Birren
- Nicole Stange-Thomann
- George Church
- Roger Brent
- Grant Jensen
- Yingwu Wang
- Nick Papadakis
- Ron Weiss
- Drew Endy
- Randy Rettberg
- Austin Che
- Reshma Shetty
- MIT Synthetic biology working group
- DARPA, NTT, NSF, Microsoft

Synthetic Biology

- An alternative to understanding complexity is to remove it
- This complements rather than replaces standard approaches
- Engineering synthetic constructs will be easier
  - ➢ Enabling quicker more facile experiments
  - ➢ Enabling deeper understanding of the basic mechanisms
  - ➢ Enabling applications in nanotechnology, medicine and agriculture

Simplicity is the ultimate sophistication
-- Leonardo da Vinci