High performance computational analysis of DNA sequences from different environments

Rob Edwards

Computer Science
Biology

San Diego State University

edwards.sdsu.edu

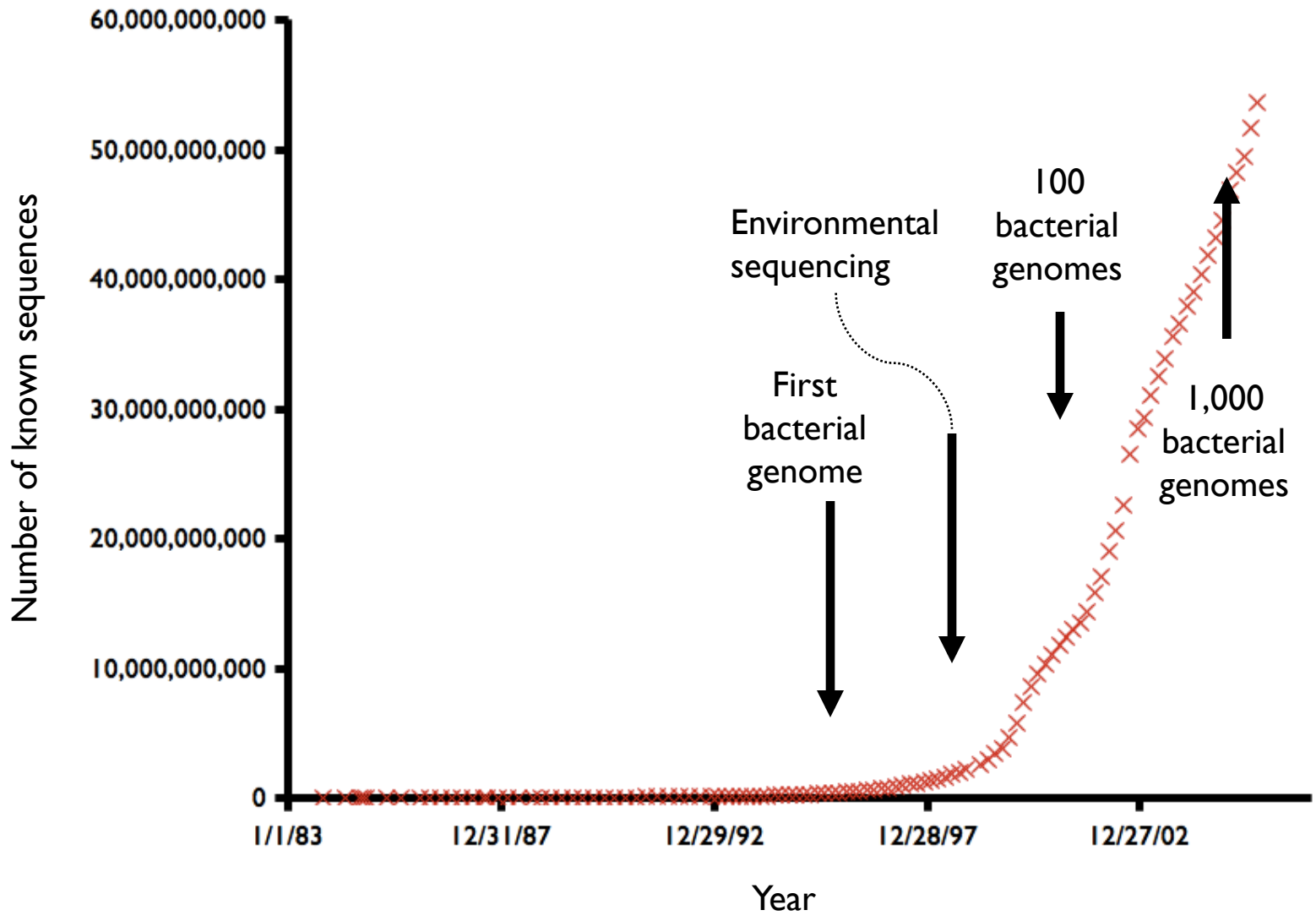
www.theseed.org

NMPDR National Microbial Pathogen Data Resource Center

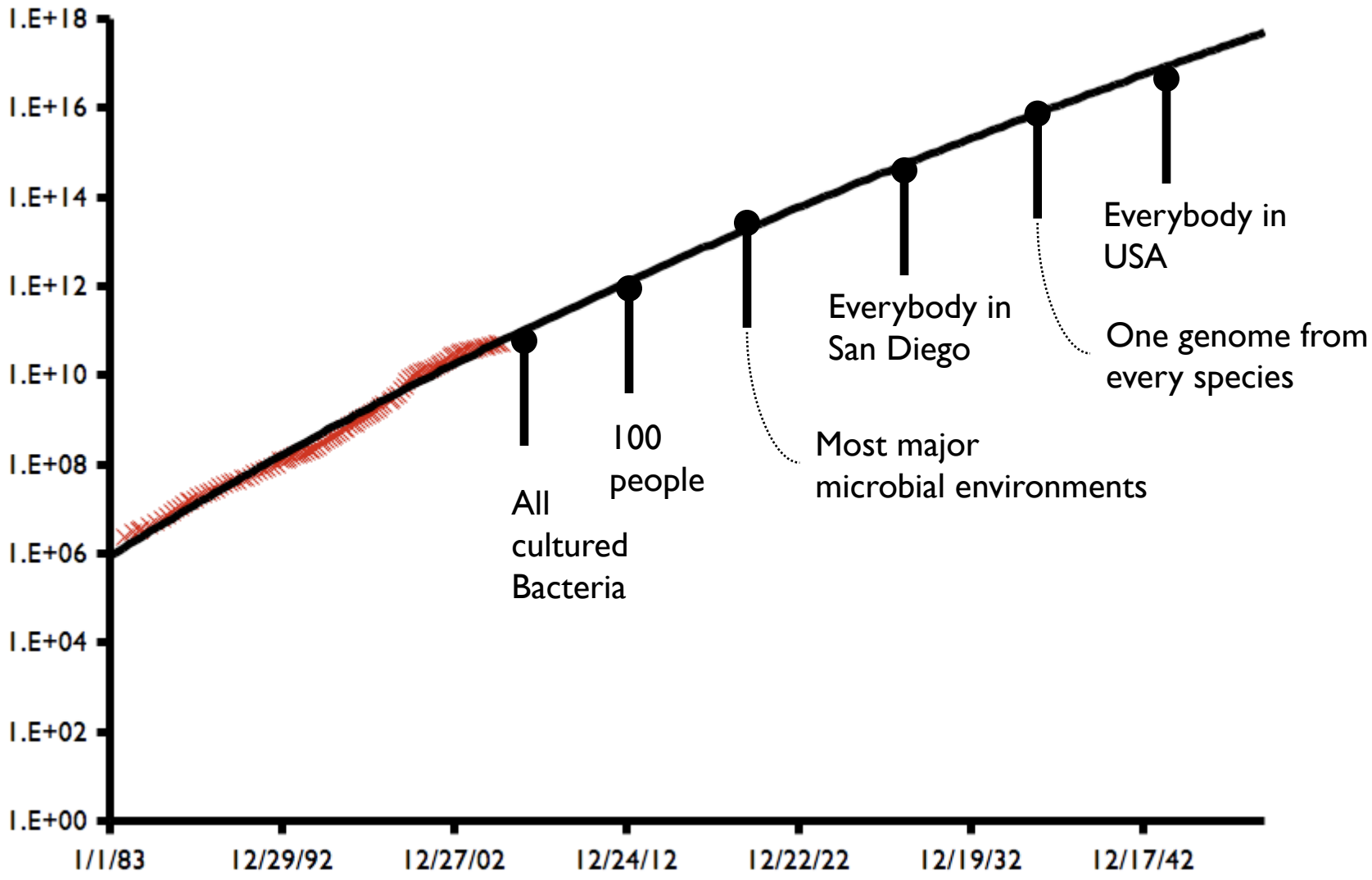
Outline

- There is a lot of sequence
- Tools for analysis
- More computers
- Can we speed analysis

How much has been sequenced?

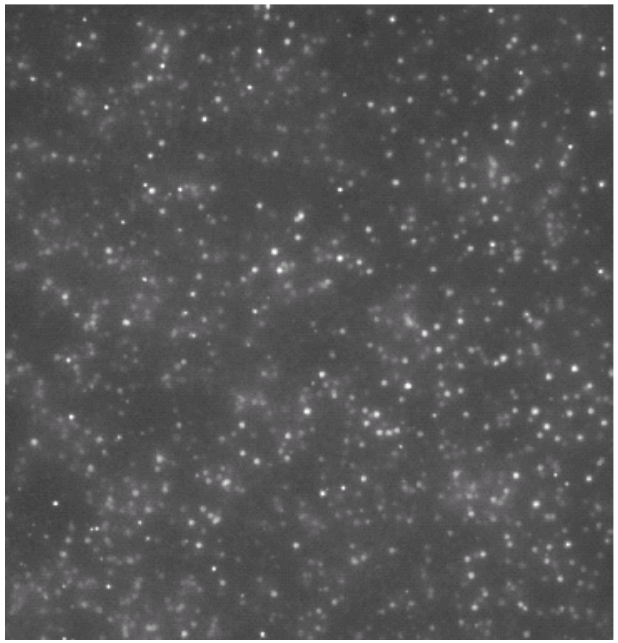


How much will be sequenced?



Metagenomics

(Just sequence it)



Epifluorescent Microscopy

200 liters water
5-500 g fresh fecal matter
50 g soil

↓

Concentrate and purify bacteria, viruses, etc

↓

Extract nucleic acids

↓

Sequence

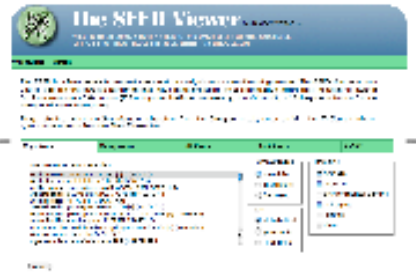
↓

Publish papers

The SEED Family

The SEED

Environmental, Viral,
Bacterial, Archaeal, and Eukaryal
Genome Interpretation



NMPDR

Display of complete genomes
Focus on pathogenesis

MG-RAST

Annotation and analysis of
metagenomes

RAST

Annotation and analysis of
complete genomes

The metagenomics RAST server

MG-RAST Meta Genome Rapid Annotation using Subsystem Technology

Metagenomics SEED Viewer version 2.0

Welcome to the Metagenomics SEED Viewer.
For more information about The SEED please visit theSEED.org.

»Navigate »Help Rob Edwards

MG-RAST is **a fully-automated service for annotating metagenome samples**.

It provides:

- **annotation** of sequence fragments,
- their **phylogenetic classification**,
- **metabolic reconstructions** and
- **comparison tools**

The service is built as a modified version of the RAST server which was originally designed to support high-quality annotation of complete or draft microbial genomes.

more

If you use our service, please cite:
The Metagenomics RAST server - A public resource for the automatic phylogenetic and functional analysis of metagenomes F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, R. Stevens, A. Wilke, J. Wilkening and R. A. Edwards
submitted



» Manage your uploaded data » Register a new account » Upload new metagenome to MG-RAST

You have access to the following metagenomes:

Number of publicly available metagenome: 158

start typing to narrow selection

Public: 5-Way (CG) Acid Mine Drainage Biofilm (4441137.3) from projec
Public: 640F6 (4440355.3) from project Cow Rumen
Public: 710F6 (4440387.3) from project Cow Rumen
Public: 80F6 (4440356.3) from project Cow Rumen
Public: ALVINELLA (4441102.3) from project Alvinella Pompejana Epibio
Public: ArcticVir2002 (4440306.3) from project Ocean Viruses
Public: Australian Phosphorus Removing (EBPR) Sludge (4441092.3) fron
Public: BBCVir96to04 (4440305.3) from project Ocean viruses

View Metagenome

Automated Processing

User uploads DNA sequences

Pre-processing:

Genome Upload has been successfully completed.

Metagenome ID - Name:	
Job:	
User:	
Date:	
Number of uploaded sequence	
Total uploaded sequence leng	

Jobs Details #570

» Browse annoted metagenome in SEED Viewer

» Available downloads for this job: Genbank export Download

» Share this metagenome with selected users

» Make this metagenome publicly accessible

» Back to the Jobs Overview

Preprocessing has been succ

The following statistics are based

Number of sequences:	23
Total sequence length:	53295251
Average read length:	227.606717772416
Longest sequence id:	>58459919
Longest sequence length:	374
Shortest sequence id:	>58696741
Shortest sequence length:	36

Similarity Computation has been successfully completed.

Similarity postprocessing has been successfully completed.

Final Assignment has been successfully completed.

Summary View

MG-RAST Meta Genome Rapid Annotation using Subsystem Technology

Metagenomics SEED Viewer version 2.0

Welcome to the Metagenomics SEED Viewer.
For more information about The SEED please visit theSEED.org.

»Navigate | »Metagenome | »Compare Metagenomes | »Management | »Help | Rob Edwards

Metagenome Overview for MG_Soybean_Field_Sample (4440698.3)

Project:	Fermi_metagenomics_samples
Metagenome	MG_Soybean_Field_Sample
Metagenome ID:	4440698.3
Description:	No description available.
Uploaded on:	Sun Mar 23 20:49:40 2008
Total no. of sequences	234,155
Total sequence size	53,295,251
Shortest sequence length	36
Longest sequence length	374
Average sequence length	227.61
Average GC content	not computed

Overview | Metabolic Analysis | Phylogenetic Analysis | Compare

The metagenome overview page provides basic information and a summary regarding the selected metagenome. Information includes project name, project description, metagenome name and unique id as well as sequence length and percent GC statistics. Histograms of sequence length and GC content is also provided. In order to provide a brief overview of the taxonomic distribution, a table is provided with domain distribution for RNA and protein based analysis.

The Overview is accessible through the menu via
» Metagenome » Overview

Summary and Statistics

The MG_Soybean_Field_Sample data set contains 234,155 contigs totaling 53,295,251 basepairs with an average fragment length of 227.61 (you can download the entire data set). A total of 61,041 sequences (26.07%) could be matched to proteins in SEED subsystems (using an e-value cut-off of 1e-5), you can explore metabolic reconstructions based on different parameters on the Metabolic Reconstruction Page. Based on 94,012 hits against the SEED protein non-redundant database (40.15 % of the fragments) and on the 127 hits against the ribosomal RNA database Greengenes (0.05%) we computed the following table (using an e-value cut-off of 1e-5 and a minimum alignment length of 50bp).

The Phylogenetic Reconstruction page will allow you to view taxonomic distributions in greater detail, change parameters and incorporate additional databases into your analysis.

The MG-RAST manual has more pointers for working with the system.

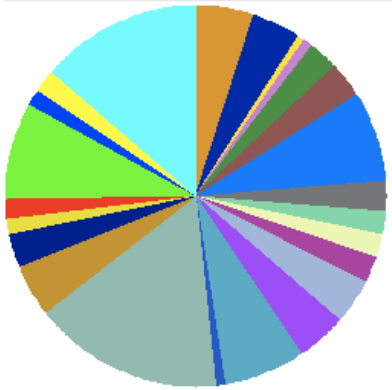
	Protein based		16s based	
Archaea	2.07%	(1946)	0.00%	(0)
Bacteria	84.76%	(79682)	90.55%	(115)
Eukaryota	1.04%	(980)	0.00%	(0)
Virus	0.00%	(0)	0.00%	(0)
Other	12.13%	(11404)	9.45%	(12)

Metagenomics Tools

Annotation & Subsystems

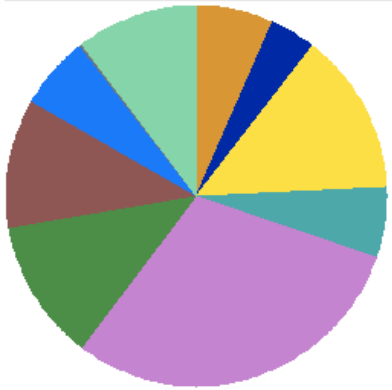
Charts | Tabular View

Group



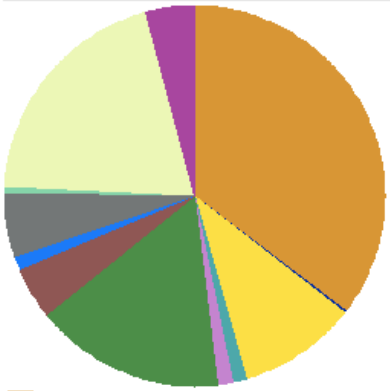
- Cofactors, Vitamins, Prosthetic Groups, Pigments 5.12% (3124)
- Cell Wall and Capsule 4.07% (2487)
- Potassium metabolism 0.74% (452)
- Photosynthesis 0.03% (19)
- Stress Response 2.62% (1597)
- Sulfur Metabolism 1.54% (942)
- Metabolism of Aromatic Compounds 1.71% (1043)
- Amino Acids and Derivatives 8.32% (5078)
- Fatty Acids and Lipids 1.43% (870)
- Phosphorus Metabolism 1.71% (1045)
- Carbohydrates 13.60% (8302)

Amino Acids and Derivatives



- Histidine Metabolism 6.74% (357)
- Osmotic stress 6.10% (323)
- Lysine, threonine, methionine, and cysteine 30.00% (1589)
- Branched-chain amino acids 11.87% (629)
- Aromatic amino acids and derivatives 10.95% (580)
- Proline and 4-hydroxyproline 6.29% (333)
- Amino Acids and Derivatives 0.19% (10)
- Alanine, serine, and glycine 10.18% (539)

Lysine, threonine, methionine, and cysteine

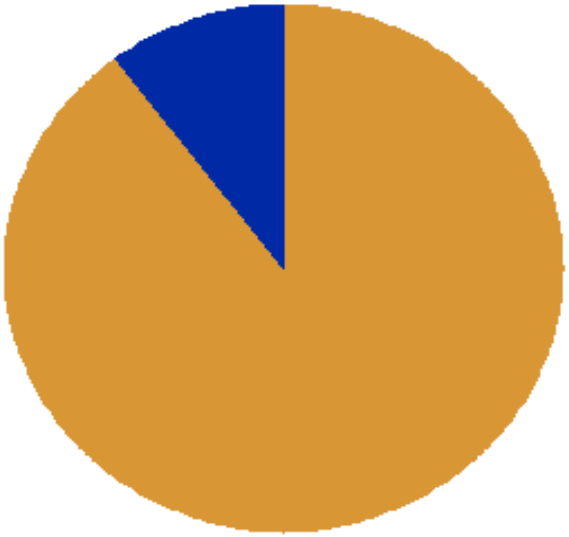


- Methionine_Biosynthesis 35.43% (563)
- CBSS-84588.1.peg.1247 0.19% (3)
- Threonine_degradation 10.20% (162)
- YeiH 0.88% (14)
- Methionine_Degradation 1.38% (22)
- Threonine_and_Homoserine_Biosynthesis 16.24% (258)
- Methionine_Salvage 4.41% (70)
- Superpathway_of_cysteine_biosynthesis 1.13% (18)
- Cysteine_Biosynthesis 5.48% (87)
- Lysine_biosynthesis_AAA_pathway_2 0.50% (8)
- Lysine_Biosynthesis_DAP_Pathway 20.14% (320)
- Lysine_degradation 4.03% (64)

Metagenomics Tools

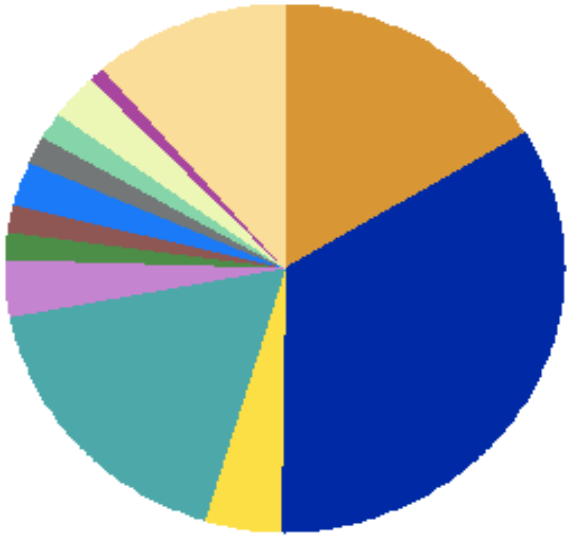
Phylogenetic Reconstruction

Domain



- Bacteria 89.84% (115)
- Unclassified 10.16% (13)

Bacteria



- Actinobacteria 16.52% (19)
- Proteobacteria 33.91% (39)
- Bacteroidetes 4.35% (5)
- Unclassified 17.39% (20)
- Planctomycetes 3.48% (4)
- Genera_incertae_sedis_OP10 1.74% (2)
- Genera_incertae_sedis_WS3 1.74% (2)
- Firmicutes 2.61% (3)
- Chloroflexi 1.74% (2)
- Acidobacteria 1.74% (2)
- Nitrospira 2.61% (3)
- Verrucomicrobia 0.87% (1)
- Gemmatimonadetes 11.30% (13)

Proteobacteria

- Alphaproteobacteria 28.21% (11)
- Gammaproteobacteria 12.82% (5)
- Betaproteobacteria 35.90% (14)
- Deltaproteobacteria 23.08% (9)

Metagenomics Tools

Comparative Tools

Subsystem Hierarchy 1 all	4440690.3	4440740.3	4440739.3	4440698.3
Carbohydrates	0.0059	0.0074	0.0084	0.0106
Clustering-based subsystems	0.0035	0.0045	0.0051	0.0066
Amino Acids and Derivatives	0.0030	0.0038	0.0042	0.0053
Virulence	0.0030	0.0036	0.0039	0.0048
Cofactors, Vitamins, Prosthetic Groups, Pigments	0.0024	0.0029	0.0032	0.0042
Respiration	0.0021	0.0025	0.0027	0.0035
Protein Metabolism	0.0020	0.0024	0.0028	0.0035
Cell Wall and Capsule	0.0018	0.0022	0.0026	0.0032
Unclassified	0.0018	0.0022	0.0024	0.0030
Metabolism of Aromatic Compounds	0.0015	0.0018	0.0020	0.0025
RNA Metabolism	0.0015	0.0018	0.0020	0.0024
Stress Response	0.0013	0.0017	0.0018	0.0022
Membrane Transport	0.0010	0.0012	0.0014	0.0017
DNA Metabolism	0.0009	0.0012	0.0013	0.0017
Regulation and Cell signaling	0.0009	0.0012	0.0011	0.0016
Nucleosides and Nucleotides	0.0008	0.0010	0.0012	0.0014

Outline

- There is a lot of sequence
- Tools for analysis
- More computers

How much data so far

986 metagenomes

79,417,238 sequences

17,306,834,870 bp (17 Gbp)

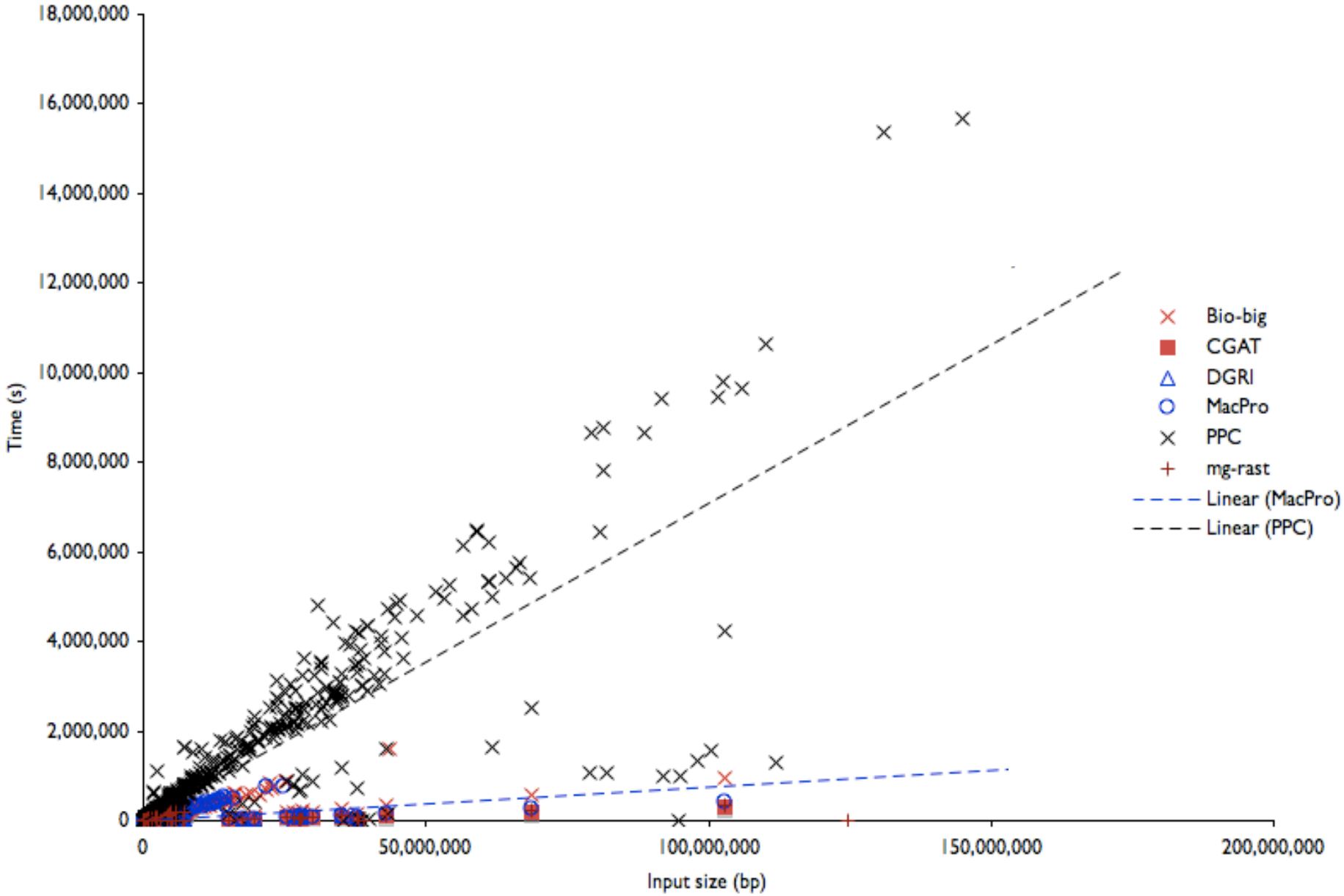
Average: ~15-20 M bp per genome

~300 GS20
~300 FLX
~300 Sanger

Computes

Processor	***Memory***	***Number of units***
16 Intel Xeon CPU X7350 @ 2.93GHz	123,823 MB	1
8 Intel Xeon CPU X5365 @ 3.00GHz	16,436 MB	1
8 Intel Xeon CPU E5335 @ 2.00GHz	16,440 MB	1
1 Intel Pentium 4 CPU 3.00GHz	3,636 MB	2
8 Intel Xeon CPU X5365 @ 3.00GHz	16,387 MB	2
8 Intel Xeon CPU E5450 @ 3.00GHz	16,436 MB	1
8 Intel Xeon CPU X5365 @ 3.00GHz	16,436 MB	1
2 PPC970FX, altivec supported	4,042 MB	45

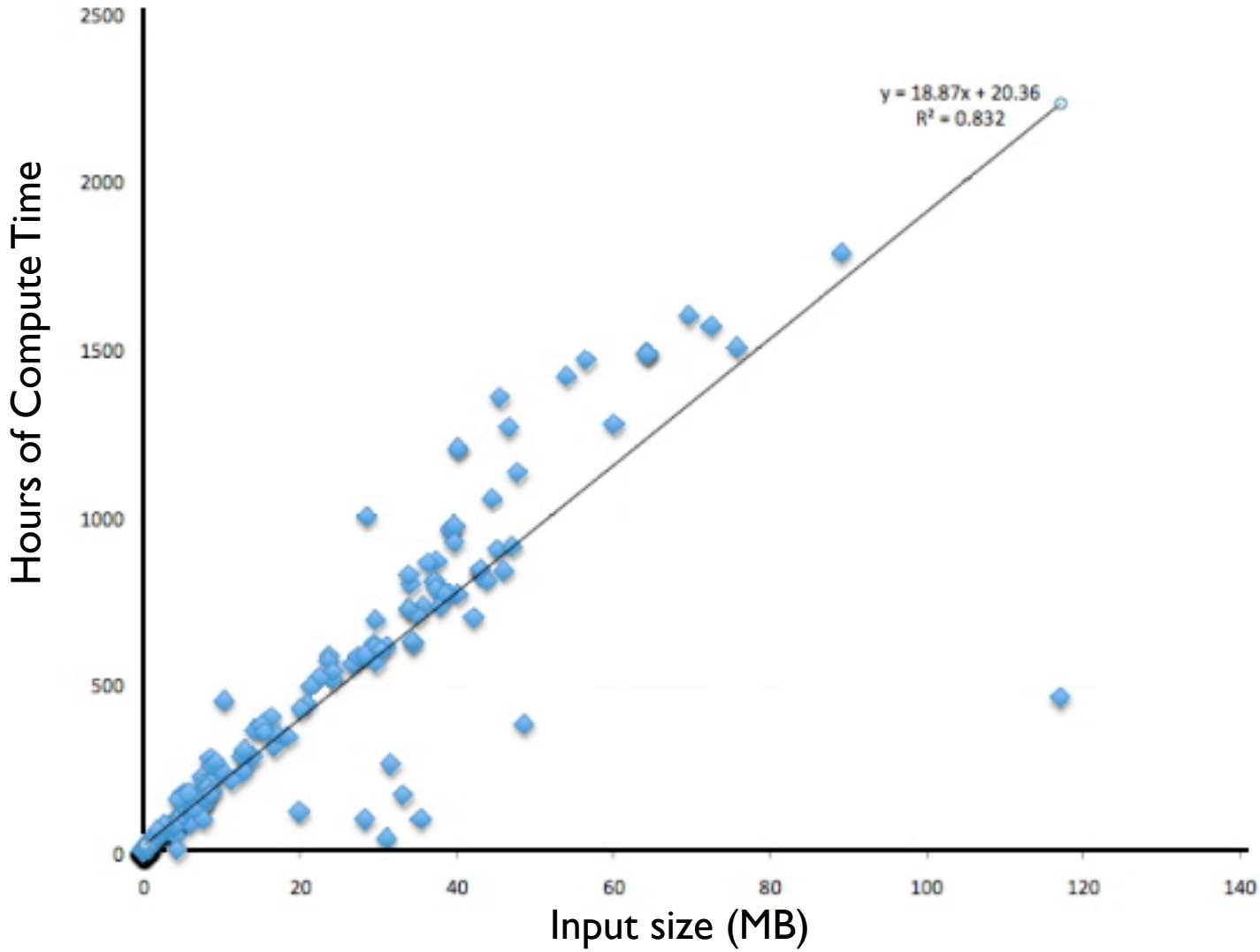
Linear compute complexity



Just waiting ...

Overall compute time

~19 hours of compute per input megabyte



How much so far

986 metagenomes

79,417,238 sequences

17,306,834,870 bp (17 Gbp)

Average: ~15-20 M bp per genome

~300 GS20
~300 FLX
~300 Sanger

Compute time (on a single CPU):

328,814 hours = 13,700 days = 38 years

Outline

- There is a lot of sequence
- Tools for analysis
- More computers
- Can we speed analysis

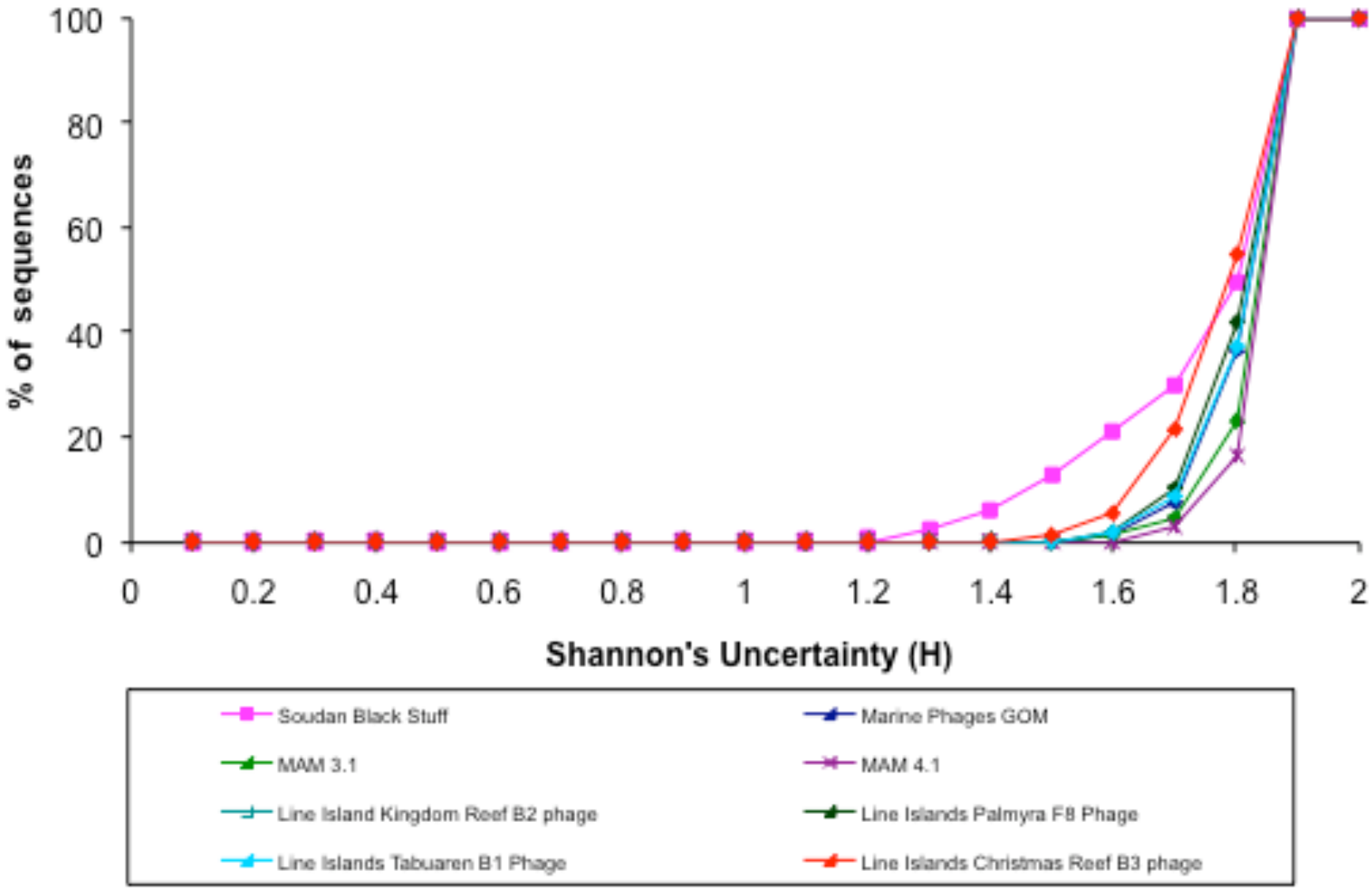
Shannon's Uncertainty

- Shannon's Uncertainty – Peter's surprisal

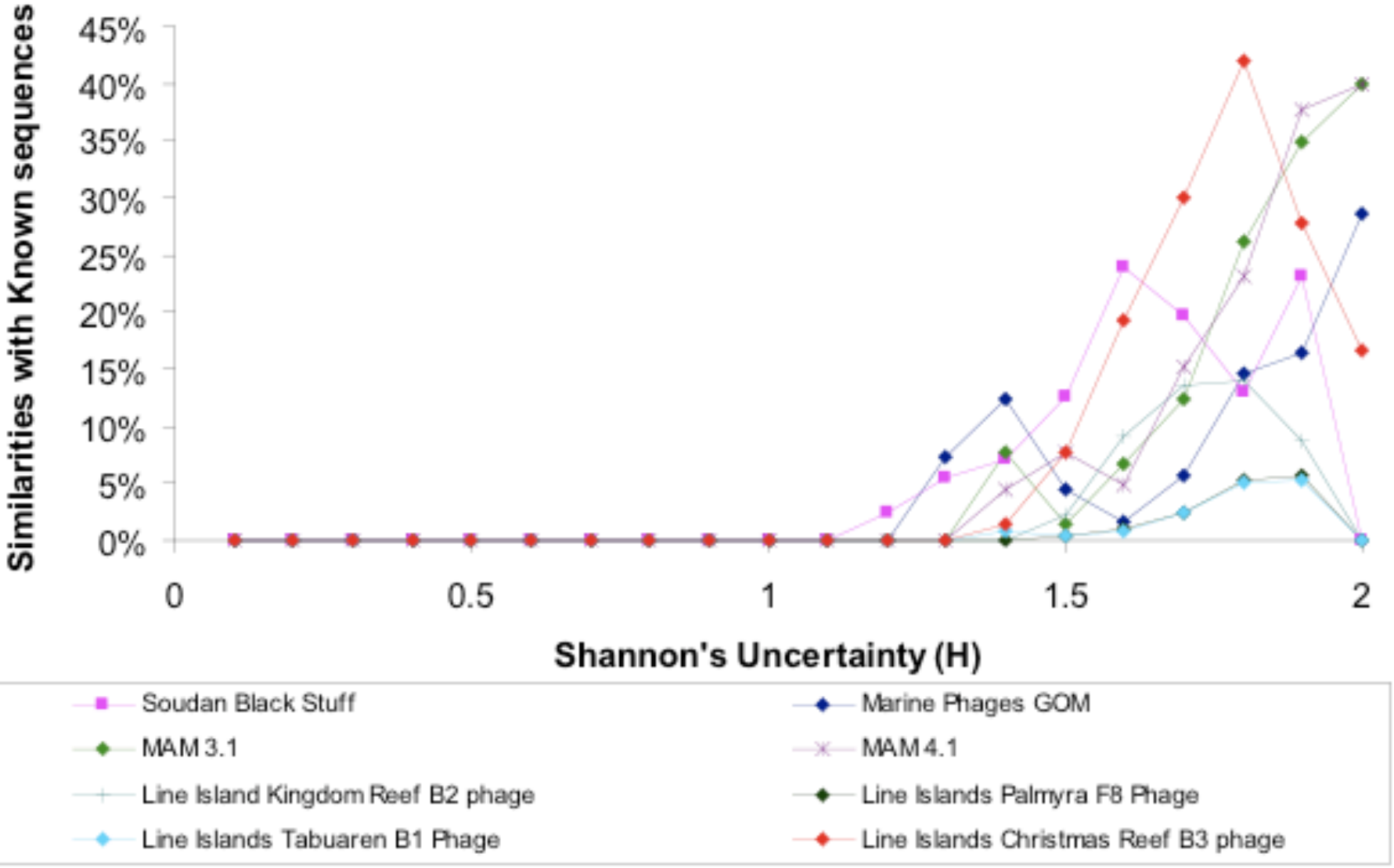
$$H(X) := -\sum_{i=1}^{n} p(x_i) \log_b p(x_i)$$

p(xi) is the probability of the occurrence of each base or string

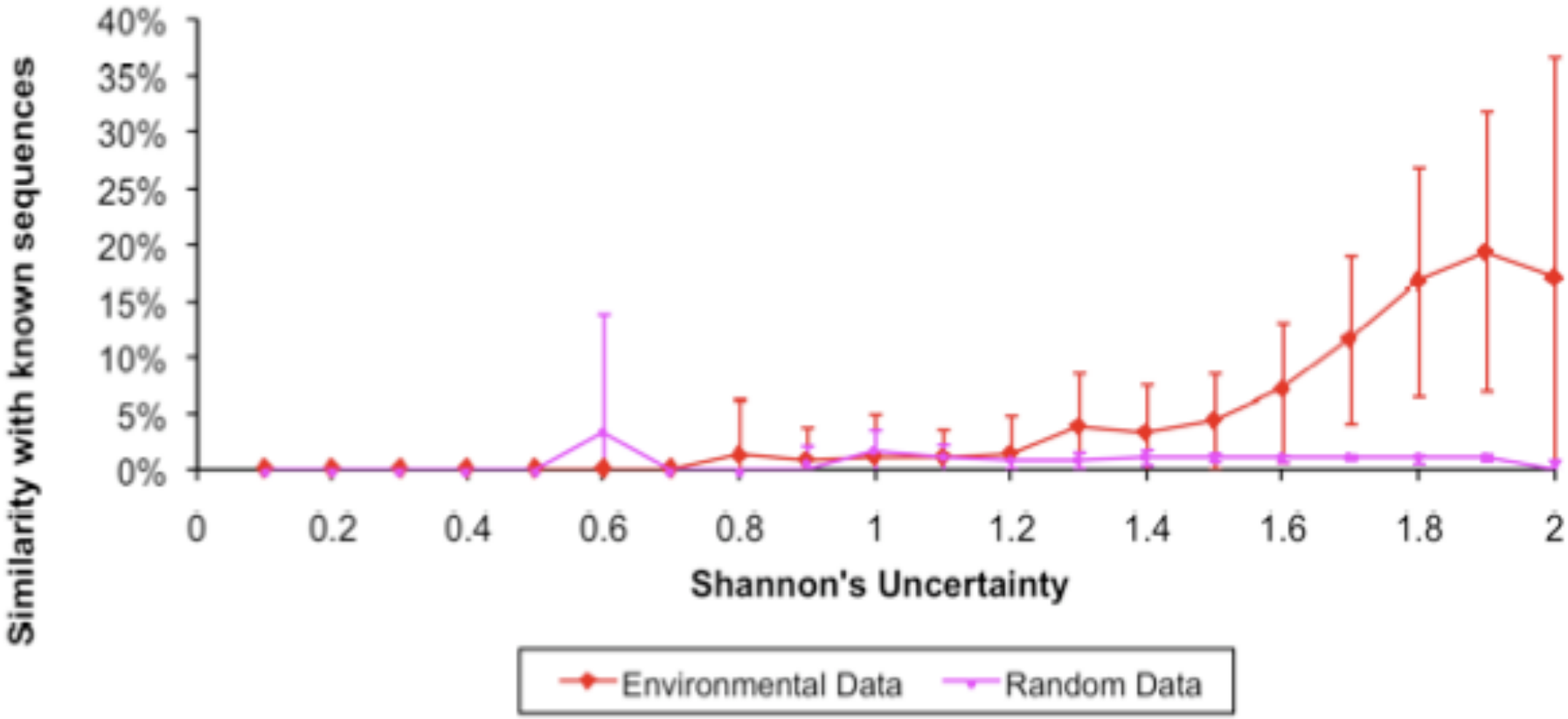
Surprisal in Sequences



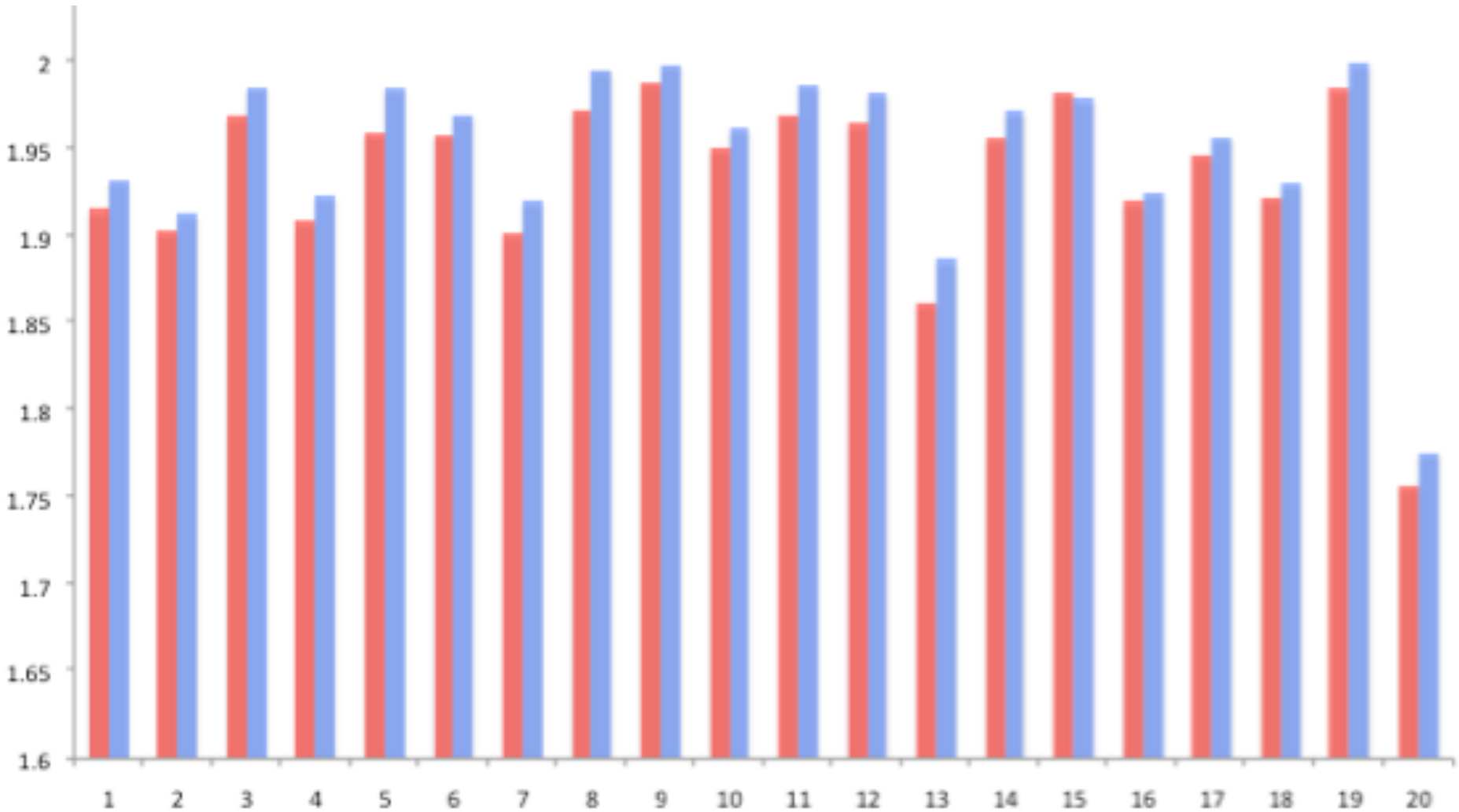
Uncertainty Correlates With Similarity



But it's not just randomness...



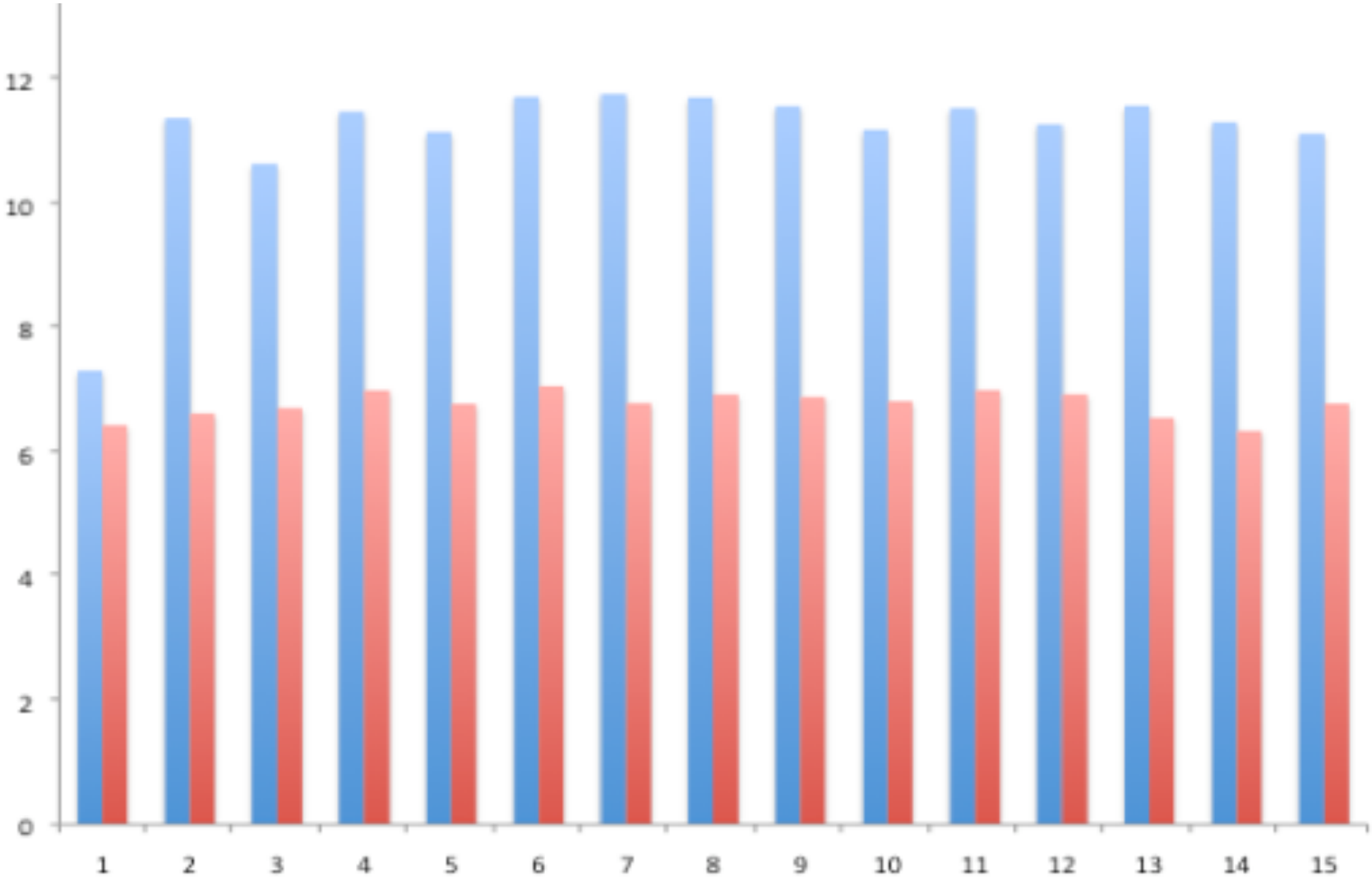
Uncertainty in complete genomes



Coding regions

Non-coding regions

More extreme differences with 6-mers



Coding regions

Non-coding regions

Can we predict proteins

- Short sequences of 100 bp
- Translate into 30-35 amino acids
- Can we predict which are real and could be doing something?
- Test with bacterial proteins

Kullback-Leibler Divergence

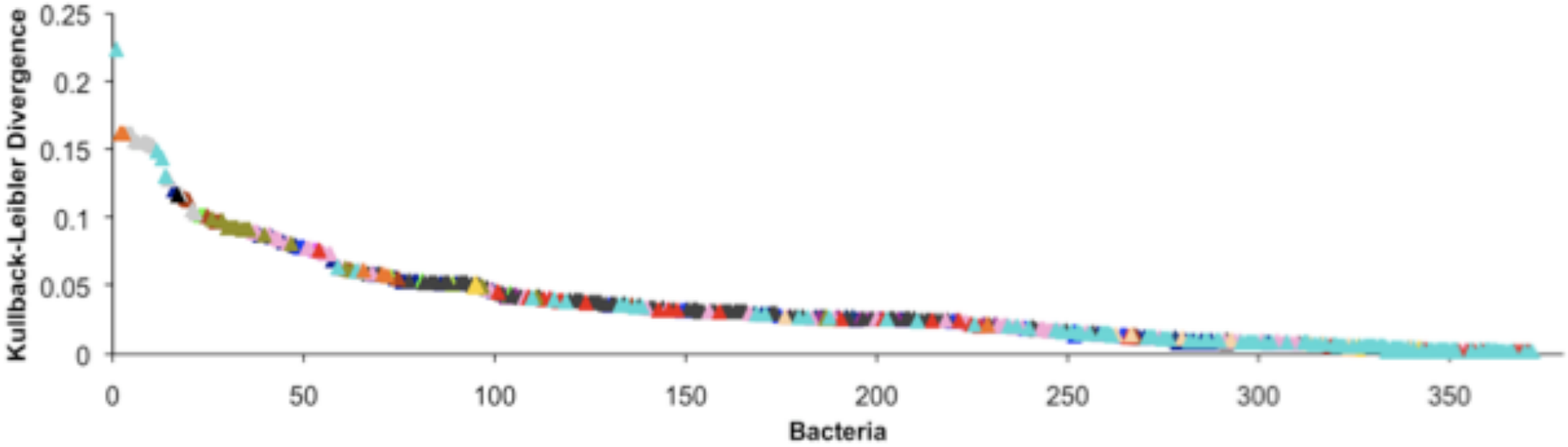
Difference between two probability distributions

$$D_{\mathrm{KL}}(P\|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Difference between amino acid composition and average amino acid composition

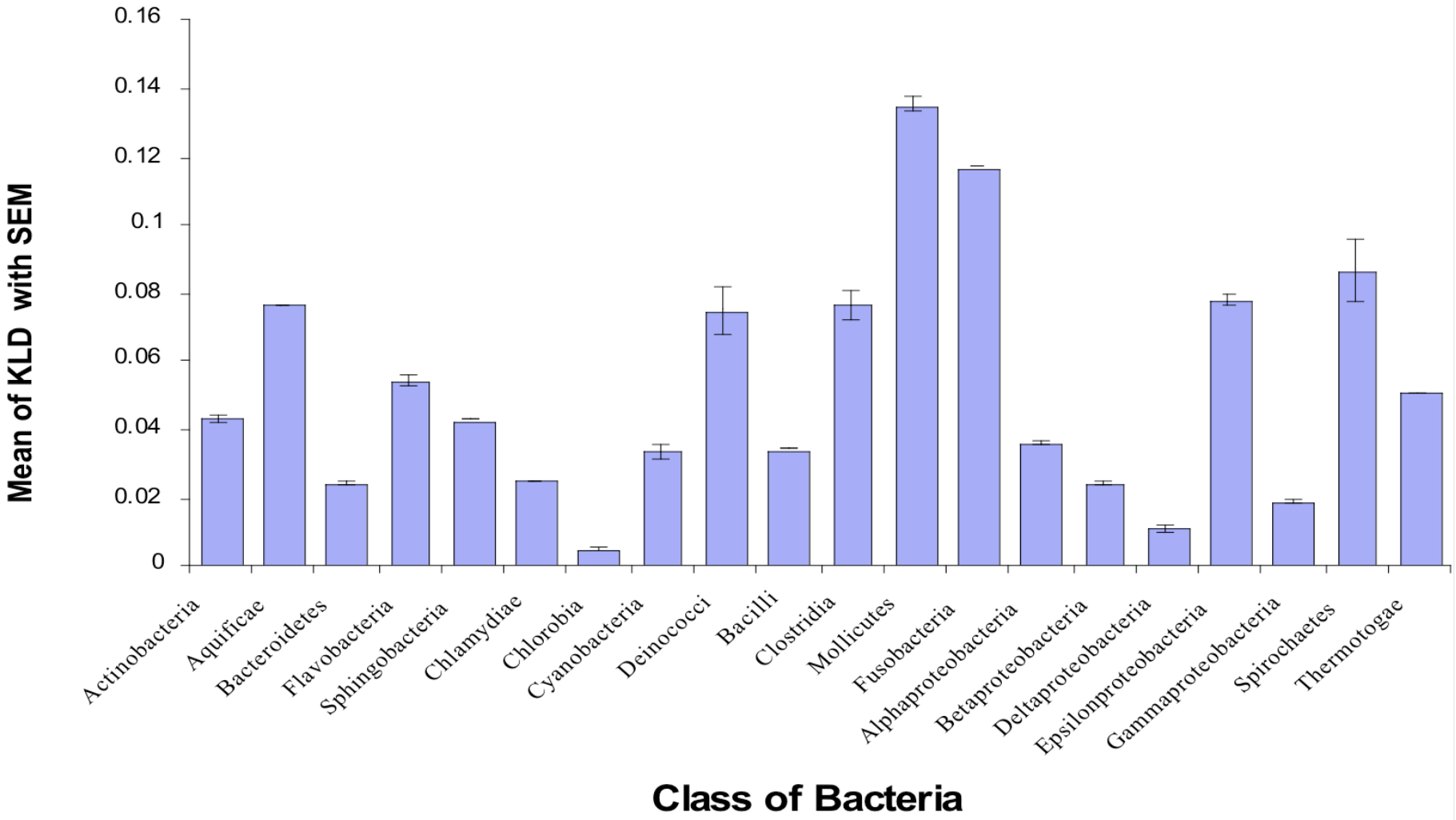
Calculate KLD for 372 bacterial genomes

KLD varies by bacteria



Colored by taxonomy of the bacteria

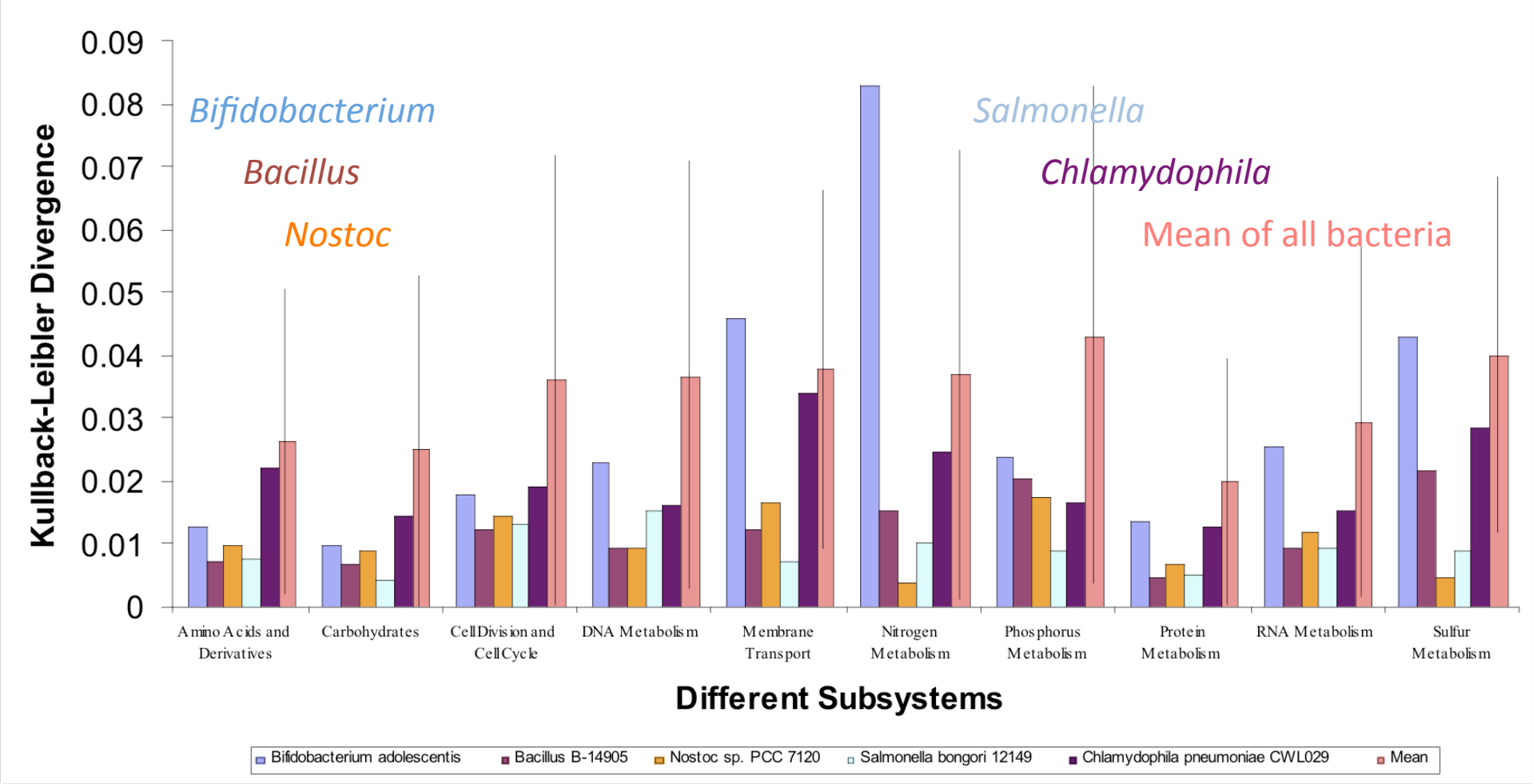
KLD varies by bacteria



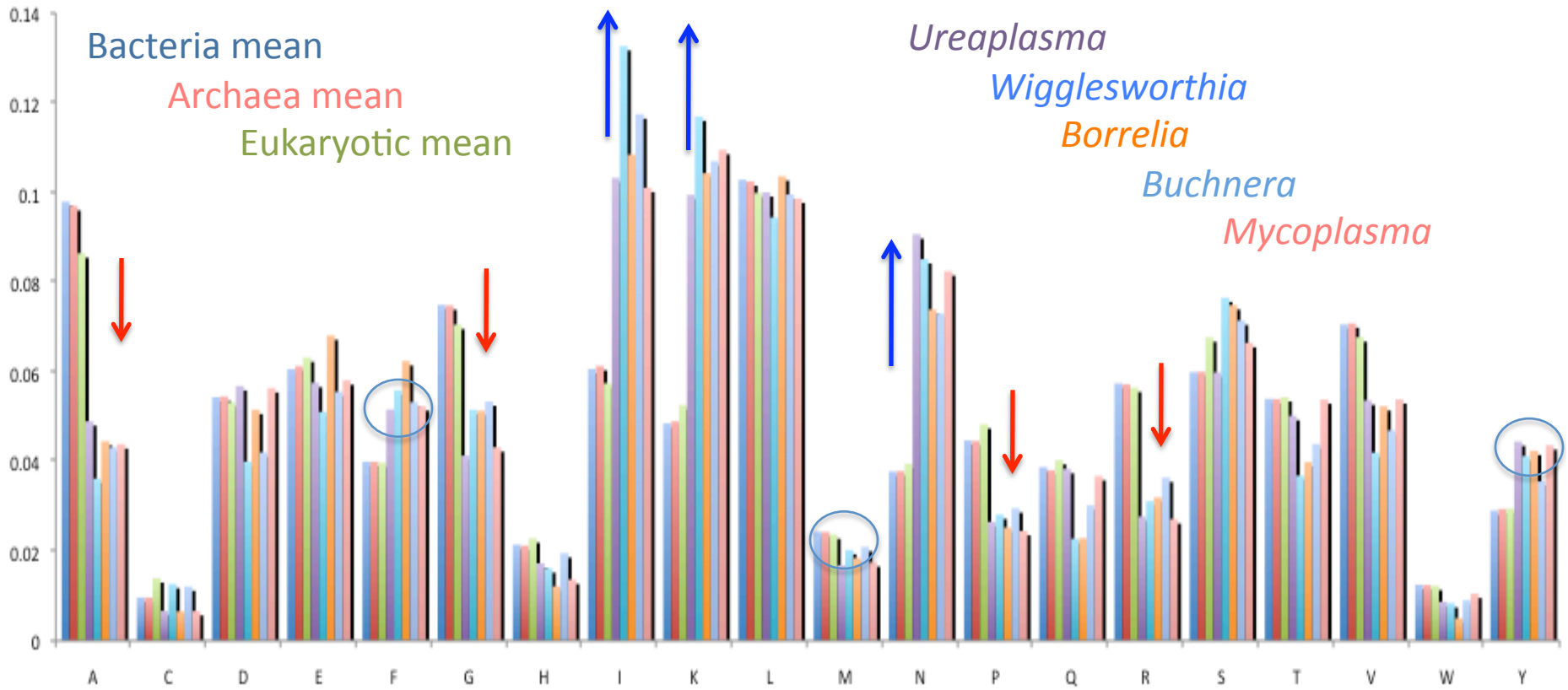
Most divergent genomes

- *Borrelia garinii* – Spirochaetes
- *Mycoplasma mycoides* – Mollicutes
- *Ureaplasma parvum* – Mollicutes
- *Buchnera aphidicola* – Gammaproteobacteria
- *Wigglesworthia glossinidia* – Gammaproteobacteria

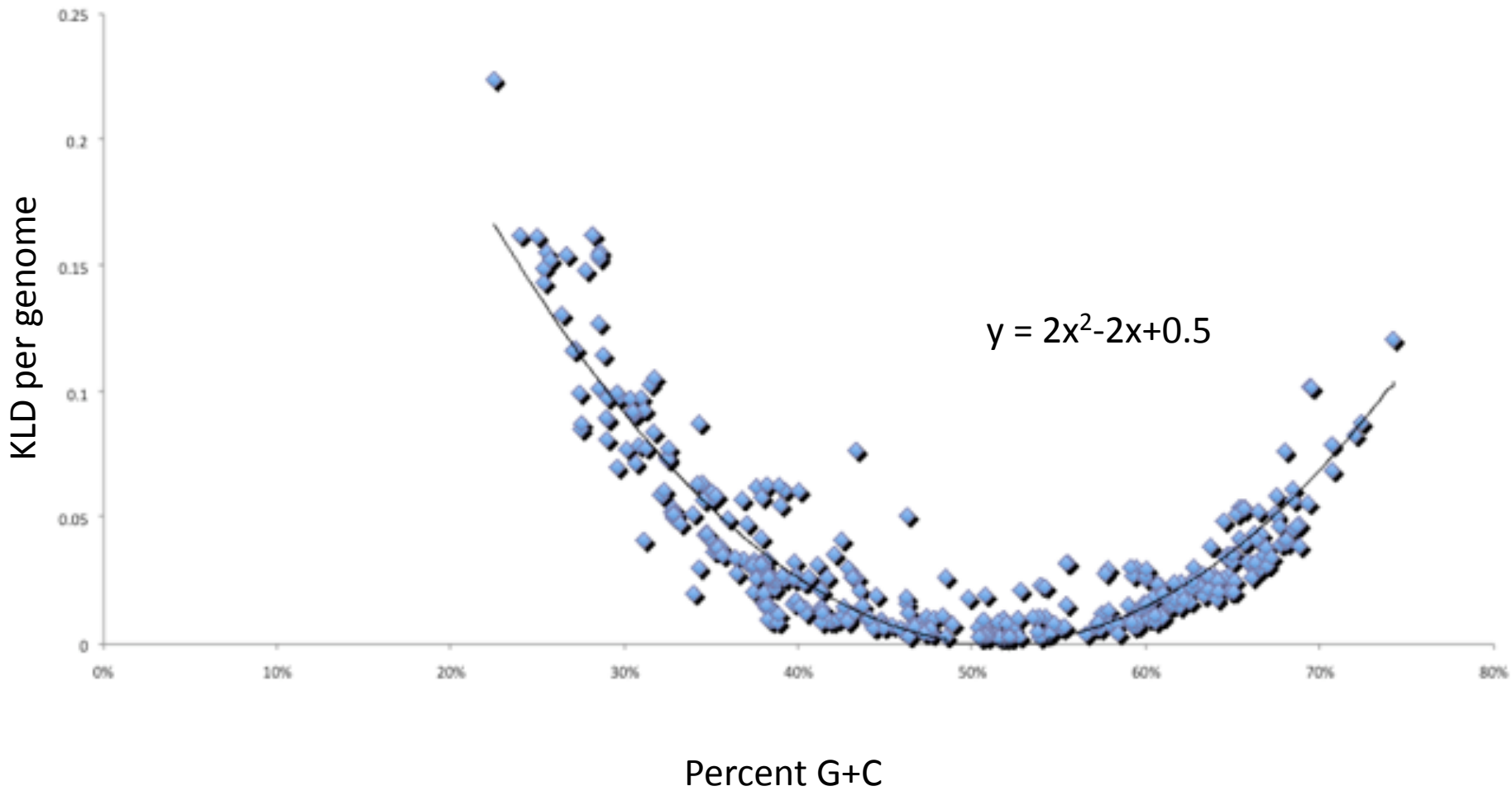
Divergence and metabolism



Divergence and amino acids



Predicting KLD

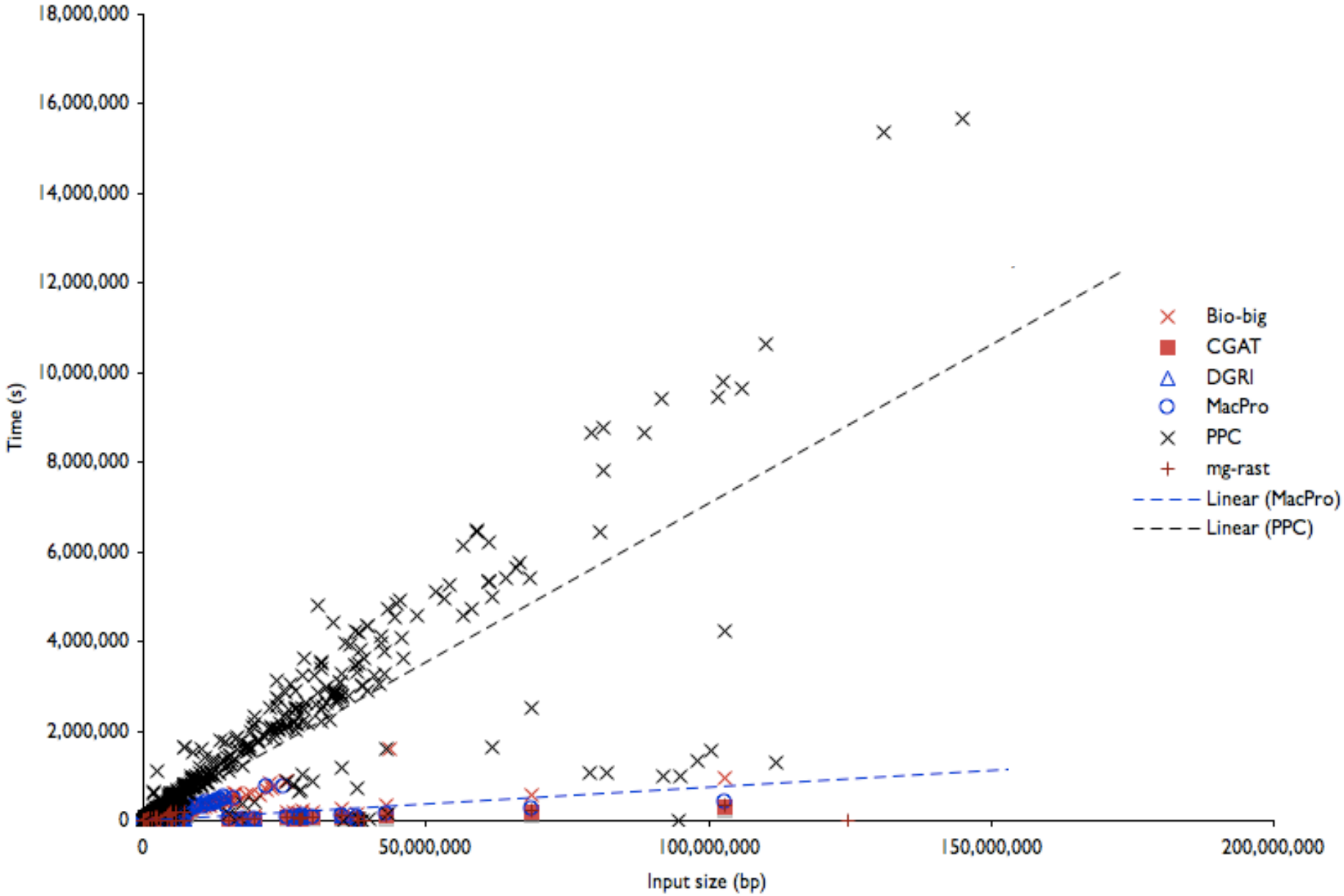


Summary

- Shannon's uncertainty could predict useful sequences

- KLD varies too much to be useful and is driven by %G+C content

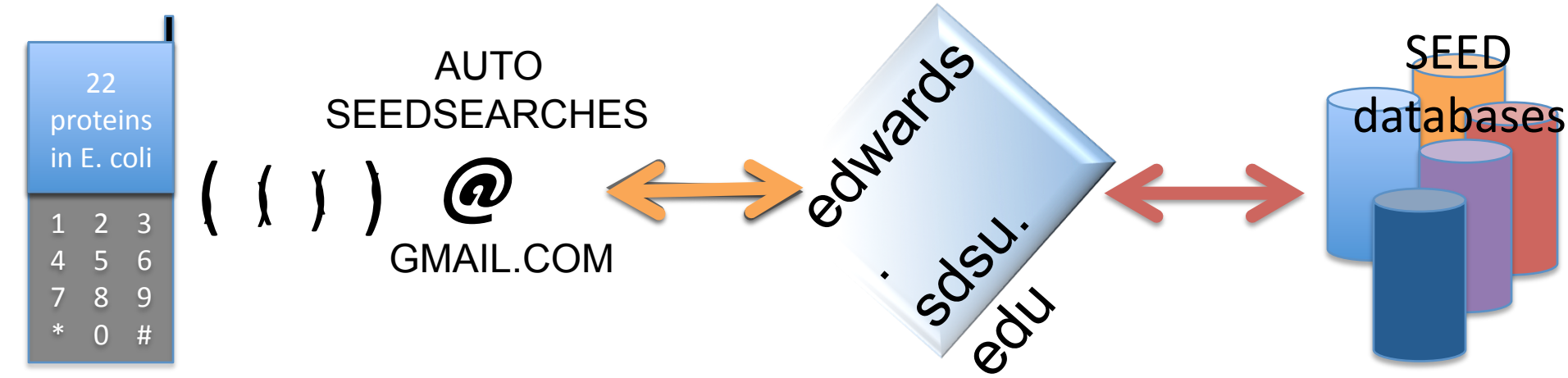
New solutions for old problems?



Xen and the art of imagery

The cell phone problem

Searching the seed by SMS



Anywhere *Idaho* *GMCS429* *Argonne*

Challenges

- Too much data
- Not easy to prioritize
- New models for HPC needed
- New interfaces to look at data

Acknowledgements

- Sajia Akhter
- Rob Schmieder

- Nick Celms
- Sheridan Wright

- Ramy Aziz

- FIG
- The mg-RAST team

- Rick Stevens

- Peter Salamon
- Barb Bailey
- Forest Rohwer
- Anca Segall