Which is the most successful gene on Earth?

Ramy K. Aziz

SDSU, Cairo University & NMPDR (U Chicago)

Which is the most successful gene on Earth?




Which is the most ~~successful~~ prevalent gene on Earth?

Who is the winner?

- **Ubiquity:** (omnipresence/ universality → essentiality)
  - For the purpose of this study, the ubiquity of x is calculated as the number of "sets" to which x belongs
    - x= (gene, protein, function, protein-encoding gene, enzyme)
    - @sets = (genomes, metagenomes, biomes)

- **Abundance:** (profusion → 'fertility/ promiscuity')
  - The abundance of x is calculated as the (average) number of times x is represented in a particular set

Spelling out the question

- **What to count:**
  - **gene**: DNA (or RNA) that encodes a protein (through mRNA), tRNA, rRNA, tmRNA, etc…
  - **protein-encoding gene (aka gene)**: DNA (or RNA) that encodes a protein
  - **function/ functional role**: an annotation… In Genomes function = PEG! In metagenomes, a function is the annotation given to a sequence read (gene tag) **IF** there is a best blast hit to this gene tag.

Current knowledge

- What do you think is the most abundant and most ubiquitous **function** or **PEG**?

  (an enzyme? a transcription factor? a transporter? DNA metabolism? Carbohydrate metabolism?)

Current knowledge

- The most abundant protein: **RuBisCo***
- How so? It's the enzyme with the highest copy number in ecosystems (or with highest total mass).
- Is it the most ubiquitous? No! It's almost only in photosynthetic organisms.
- Is its gene the most abundant? No! Most genomes lack it.

*__r__ib__u__lose-1,5-__bis__ phosphate __c__arb__o__xylase

Methodology

What to count and how to count it?

aka

Can we even answer the question?

Counting PEGs in genomes

☺ PEG's fully sequenced

☺ one copy per function (more == paralogs)

So, just collect all genomes, extract all pegs, count functions, and get results.

☹ Sequenced genomes do not represent life, but rather human-centered interests in life.

Counting EGTs in metagenomes

Environmental gene tag (EGT) comes from one organism and represents one or more functions.

☺ Counts ∝ abundance

☹ Counts depend on:

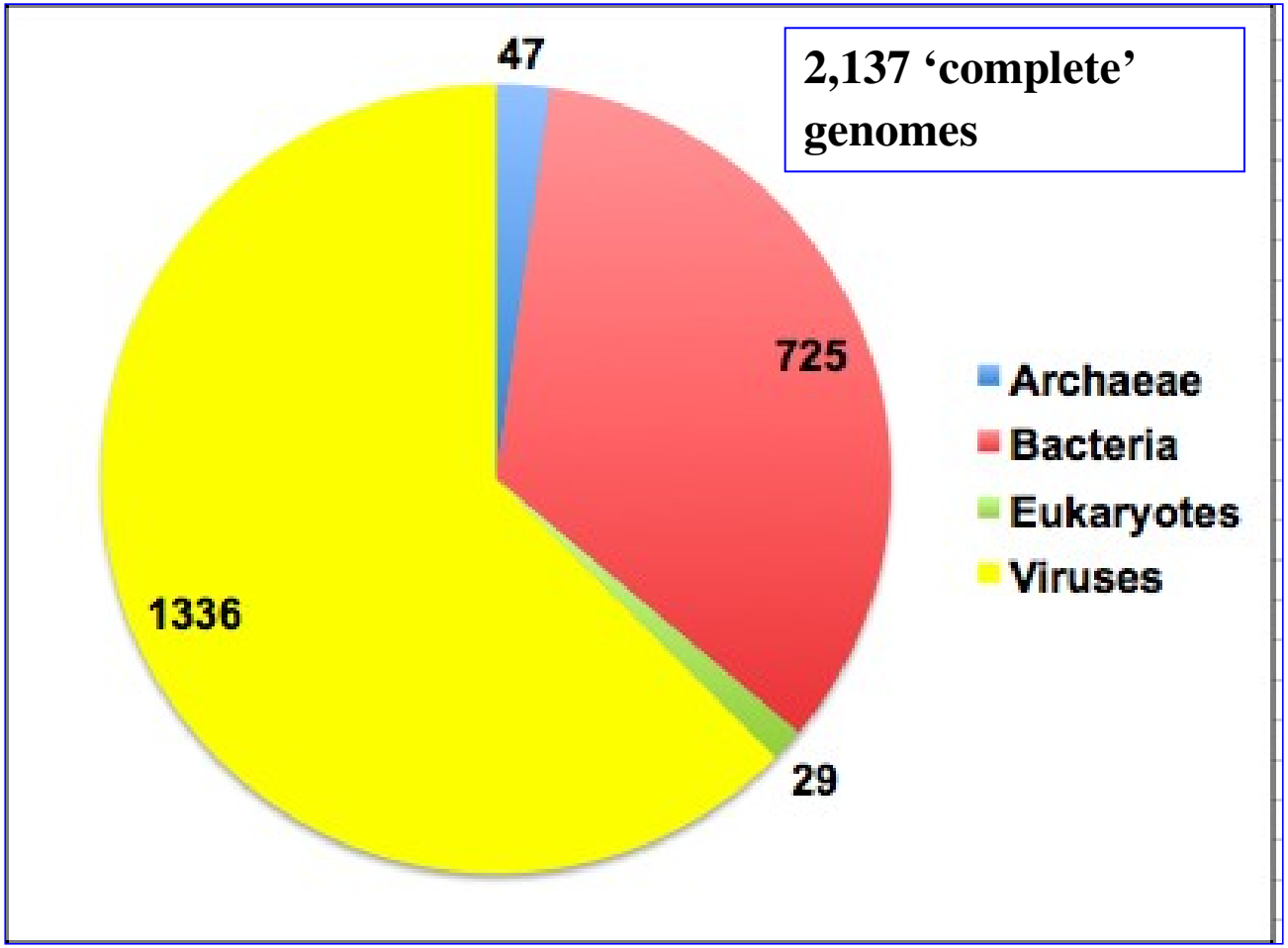
- Abundance
- Gene length
- Metagenome sample size ($$)

☹ Up 90% with no BLAST hits

The answer

(Show me the data!)

The genomic sample



The metagenomic sample

Metagenomes	187
Sequences	6,730,478
Sequences per Metagenome	35,991
Median size of metagenomes	15,914

And the winner is ...

- Ghost anonymous protein
- aka
  - Hypothetical protein
  - Conserved protein
  - Unknown protein
  - Protein predicted by Glimmer
  - Very hypothetical protein
  - No name

And the winner is ...
(metagenomes)

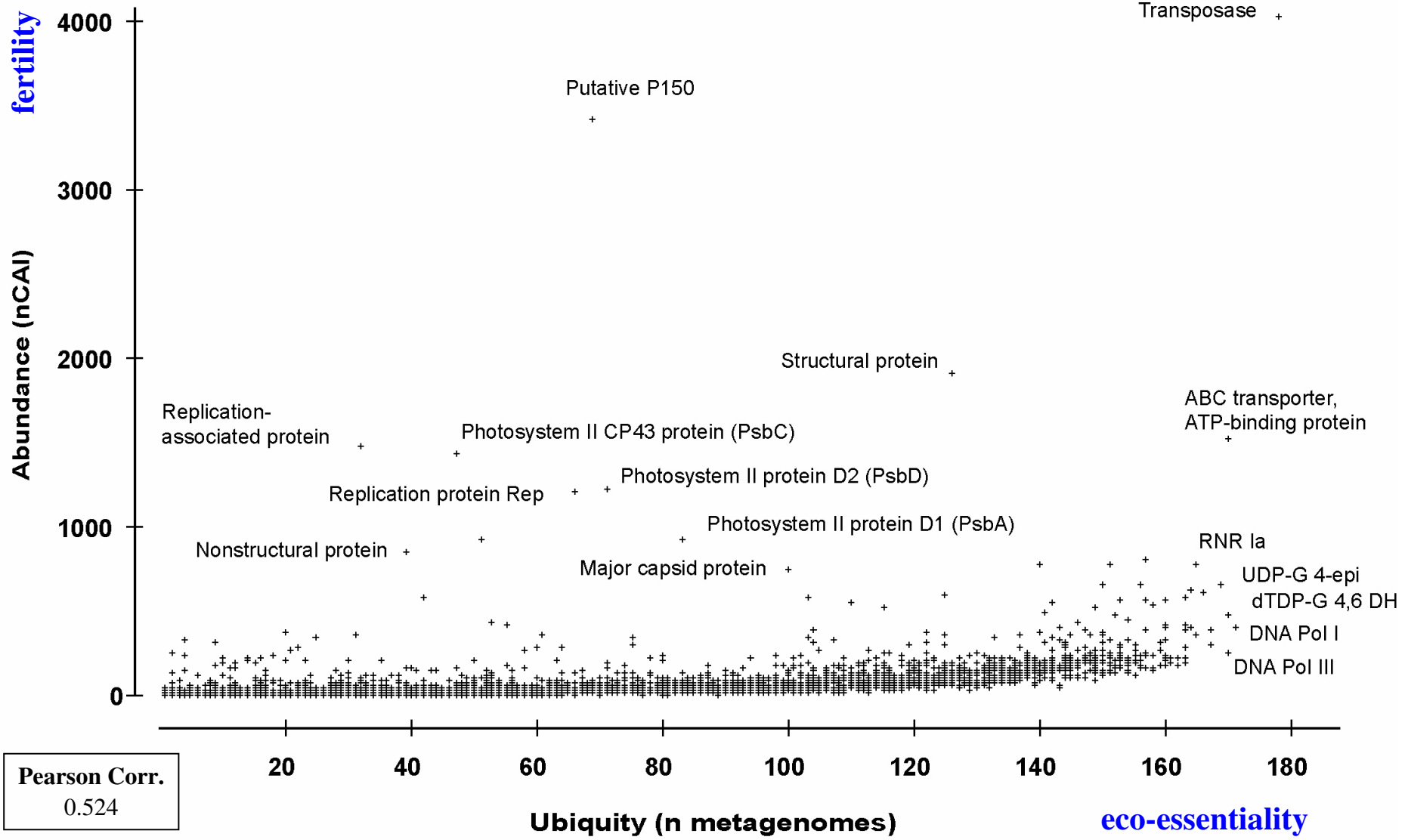
Function	#	nCAI
Transposase	178	4,026
Retrotransposon-related p150 protein	69	3,412
Viral structural protein	126	1,909
ABC transporter, ATP-binding protein	170	1,528
Replication-associated protein	32	1,481

NCAI: $\sum$ (count/mean protein length/# informative EGTs)

Gene ubiquity in metagenomes

Rank	Functional role	n MG	Length	nCAI	%
1	Transposase	178	366.21	4,026.17	95.19
2	DNA polymerase I (EC 2.7.7.7)	171	861.60	406.51	91.44
3	Glycosyltransferase (EC 2.4.1.-)	170	405.07	866.93	90.91
4	dTDP-glucose 4,6-dehydratase (EC 4.2.1.46)	170	339.88	472.59	90.91
5	DNA polymerase III alpha subunit (EC 2.7.7.7)	170	1170.13	252.21	90.91
6	ABC transporter, ATP-binding protein	170	425.31	1,589.04	90.91
7	UDP-glucose 4-epimerase (EC 5.1.3.2)	169	338.18	657.36	90.37
8	Heat shock protein 60 family chaperone GroEL	167	531.90	383.86	89.30
9	Chaperone protein DnaK	167	634.06	298.65	89.30
10	Ribonucleotide reductase of class II (coenzyme B12-dependent) (EC 1.17.4.	166	824.91	609.01	88.77
11	Ribonucleotide reductase of class Ia (aerobic), alpha subunit (EC 1.17.4.1)	165	749.73	776.57	88.24
12	Replicative DNA helicase (EC 3.6.1.-)	165	471.03	353.30	88.24
13	Integrase	164	377.89	633.18	87.70
14	Sensor histidine kinase	164	622.97	625.78	87.70
15	Long-chain-fatty-acid--CoA ligase (EC 6.2.1.3)	164	582.40	404.78	87.70
16	Phosphate starvation-inducible protein PhoH, predicted ATPase	163	352.79	584.47	87.17
17	Carbamoyl-phosphate synthase large chain (EC 6.3.5.5)	163	1041.20	222.50	87.17
18	DNA primase (EC 2.7.7.-)	163	610.06	280.77	87.17
19	Valyl-tRNA synthetase (EC 6.1.1.9)	163	895.43	196.98	87.17
20	Thymidylate synthase (EC 2.1.1.45)	163	296.55	388.95	87.17
21	ATP-dependent Clp protease ATP-binding subunit clpX	162	426.38	217.55	86.63
22	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)	162	1244.99	175.24	86.63
23	Isoleucyl-tRNA synthetase (EC 6.1.1.5)	161	996.61	188.72	86.10
24	DNA gyrase subunit A (EC 5.99.1.3)	161	860.01	220.93	86.10
25	Amidophosphoribosyltransferase (EC 2.4.2.14)	160	519.22	231.72	85.56
26	Serine hydroxymethyltransferase (EC 2.1.2.1)	160	413.78	321.93	85.56
27	Leucyl-tRNA synthetase (EC 6.1.1.4)	160	862.59	209.31	85.56
28	DNA topoisomerase I (EC 5.99.1.2)	160	832.58	177.13	85.56
29	RNA polymerase sigma factor RpoD	160	519.71	223.34	85.56
30	UDP-glucose dehydrogenase (EC 1.1.1.22)	160	429.56	216.27	85.56

Metagenomes ...



Pearson Corr.
0.524

Metagenomes ...

Pearson Corr.
0.524

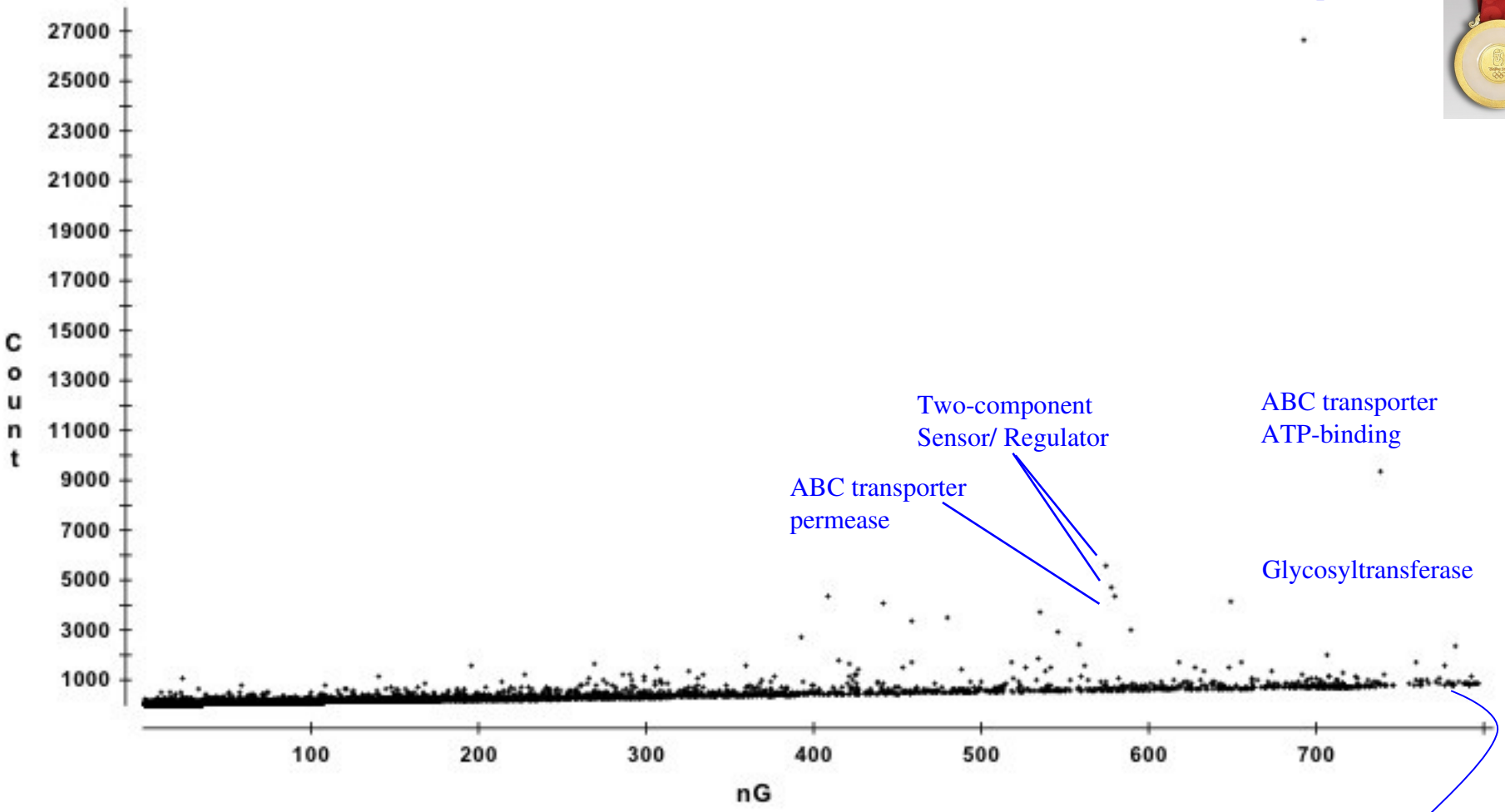
And the winner is ...
(genomes)

Function	#	Count
Transposase	693	26,625
ABC transporter, ATP-binding protein	738	9,382
Sensor histidine kinase	574	5,575
DNA-binding response regulator	578	4,708
Methyl-accepting chemotaxis protein	408	4,389

Gene ubiquity in genomes

Rank	Function	nD	Domains	nG	%	Count	Copy #
1	Phenylalanyl-tRNA synthetase alpha chain (EC 6.1.1.20)	3	A B E	797	37.3	825	1.04
2	Leucyl-tRNA synthetase (EC 6.1.1.4)	3	A B E	796	37.25	839	1.05
3	Tyrosyl-tRNA synthetase (EC 6.1.1.1)	4	A B E V	795	37.2	871	1.10
4	Isoleucyl-tRNA synthetase (EC 6.1.1.5)	3	A B E	795	37.2	877	1.10
5	Methionyl-tRNA synthetase (EC 6.1.1.10)	4	A B E V	794	37.15	826	1.04
6	Methionine aminopeptidase (EC 3.4.11.18)	3	A B E	793	37.11	1143	1.44
7	Phenylalanyl-tRNA synthetase beta chain (EC 6.1.1.20)	3	A B E	792	37.06	817	1.03
8	Preprotein translocase secY subunit (TC 3.A.5.1.1)	3	A B E	792	37.06	841	1.06
9	Dimethyladenosine transferase (EC 2.1.1.-)	3	A B E	791	37.01	821	1.04
10	Prolyl-tRNA synthetase (EC 6.1.1.15)	3	A B E	790	36.97	848	1.07
11	Arginyl-tRNA synthetase (EC 6.1.1.19)	4	A B E V	790	36.97	850	1.08
12	Signal recognition particle receptor protein FtsY (=alpha subunit)	3	A B E	789	36.92	833	1.06
13	Alanyl-tRNA synthetase (EC 6.1.1.7)	3	A B E	789	36.92	851	1.08
14	Threonyl-tRNA synthetase (EC 6.1.1.3)	3	A B E	788	36.87	878	1.11
15	Valyl-tRNA synthetase (EC 6.1.1.9)	3	A B E	788	36.87	836	1.06
16	Sua5 YciO YrdC YwlC family protein	3	A B E	785	36.73	908	1.16
17	Cysteine desulfurase (EC 2.8.1.7)	3	A B E	783	36.64	2362	3.02
18	GTP-binding and nucleic acid-binding protein YchF	3	A B E	782	36.59	833	1.07
19	Histidyl-tRNA synthetase (EC 6.1.1.21)	3	A B E	782	36.59	820	1.05
20	Translation initiation factor 2	3	A B E	782	36.59	820	1.05
21	Seryl-tRNA synthetase (EC 6.1.1.11)	3	A B E	781	36.55	817	1.05
22	Cysteinyl-tRNA synthetase (EC 6.1.1.16)	4	A B E V	781	36.55	839	1.07
23	Enolase (EC 4.2.1.11)	3	A B E	779	36.45	903	1.16
24	CTP synthase (EC 6.3.4.2)	3	A B E	779	36.45	864	1.11
25	Thioredoxin reductase (EC 1.8.1.9)	3	A B E	777	36.36	1594	2.05
26	Signal recognition particle, subunit Ffh SRP54 (TC 3.A.5.1.1)	3	A B E	777	36.36	805	1.04
27	Triosephosphate isomerase (EC 5.3.1.1)	3	A B E	773	36.17	845	1.09
28	Ribose-phosphate pyrophosphokinase (EC 2.7.6.1)	3	A B E	773	36.17	1041	1.35
29	Chaperone protein DnaK	3	A B E	771	36.08	955	1.24
30	NAD kinase (EC 2.7.1.23)	3	A B E	767	35.89	886	1.16

Genomes ...



Pearson Corr.
0.645

Some take-home messages

- Current annotations suck – we need substantial improvement to really understand metagenomes.
- Transposases are not just junk hypothetical proteins; their quorum dictates some more attention!
- The ‘selfish’ transposase genes must be offering their hosts some advantage.

Some take-home messages

- If rRNA is used to track genomes' vertical history, transposases are good to track 'horizontal' history
- Cheaters (always) win...
- Transposases shall inherit the earth...

Acknowledgments

- This study could not have been possible without

Science

the good habits of talking science while drinking, sleeping, etc.

BACK!

Raw data

Cumulative Frequency

Non-normalized

Rank	Function	n biomes	Mean Ln	Frequency
1		187	587.6476	682,720
2	hypothetical protein	187	424.3936	524,138
3	Putative p150	69	456.6075	96,016
4	unnamed protein product	144	490.9272	50,584
5	Protein sequence is in conflict with the conceptual translation	2	676.41	48,870
6	Transposase	169	359.8556	24,748
7	cys repeats; this loci is open in all six reading frames, part of IE-A	110	674.3229	19,705
8	putative outer membrane protein, probably involved in nutrient binding	141	888.3307	19,395
9	Alu subfamily SX sequence contamination warning entry	41	164.4349	18,397
10	Thioredoxin reductase (EC 1.8.1.9)	157	336.8662	16,925
11	ABC transporter ATP-binding protein	165	430.3743	15,832
12	Alu subfamily SQ sequence contamination warning entry	43	148.8028	14,443
13	Beta-galactosidase (EC 3.2.1.23)	138	845.4113	12,248
14	TonB-dependent receptor	148	819.081	11,593
15	Transcriptional regulator	165	281.3998	11,524
16	DNA polymerase III alpha subunit (EC 2.7.7.7)	170	1170.127	11,302
17	ABC transporter, ATP-binding protein	158	412.0492	11,209
18	structural protein	63	517.6257	10,664
19	putative membrane protein	167	429.3326	10,346
20	Long-chain-fatty-acid--CoA ligase (EC 6.2.1.3)	164	582.3957	10,098
21	Ribonucleotide reductase of class Ia (aerobic), alpha subunit (EC 1.17.4.1)	165	749.731	9,323
22	unknown	173	296.0219	9,125
23	Carbamoyl-phosphate synthase large chain (EC 6.3.5.5)	163	1041.195	9,094
24	replication-associated protein	32	337.6856	9,027
25	Glutamate synthase [NADPH] large chain (EC 1.4.1.13)	158	1511.417	8,911
26	Rep	20	402.878	8,254
27	DNA-directed RNA polymerase beta' subunit (EC 2.7.7.6)	158	1383.479	8,138
28	Ribonucleotide reductase of class II (coenzyme B12-dependent) (EC 1.17.4.1)	166	824.9139	8,123
29	membrane protein, putative	162	480.7901	8,103
30	Excinuclease ABC subunit A	155	964.0463	8,092
31	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)	162	1244.993	8,066
32	Isoleucyl-tRNA synthetase (EC 6.1.1.5)	161	996.6065	7,848
33	unknown protein	161	428.846	7,665
34	DNA polymerase I (EC 2.7.7.7)	171	861.603	7,618
35	DNA gyrase subunit A (EC 5.99.1.3)	161	860.0085	7,524

Raw data

Cumulative Frequency **Non-normalized**

Rank	Function	n biomes	Mean Ln	Frequency
1		187	587.6476	682,720
2	hypothetical protein	187	424.3936	524,138
3	Putative p150	69	456.6075	96,016
4	unnamed protein product	144	490.9272	50,584
5	Protein sequence is in conflict with the conceptual translation	2	676.41	48,870
6	Transposase	169	359.8556	24,748
7	cys repeats; this loci is open in all six reading frames, part of IE-A	110	674.3229	19,705
8	putative outer membrane protein, probably involved in nutrient binding	141	888.3307	19,395
9	Alu subfamily SX sequence contamination warning entry	41	164.4349	18,397
10	Thioredoxin reductase (EC 1.8.1.9)	157	336.8662	16,925
11	ABC transporter ATP-binding protein	165	430.3743	15,832
12	Alu subfamily SQ sequence contamination warning entry	43	148.8028	14,443
13	Beta-galactosidase (EC 3.2.1.23)	138	845.4113	12,248
14	TonB-dependent receptor	148	819.081	11,593
15	Transcriptional regulator	165	281.3998	11,524
16	DNA polymerase III alpha subunit (EC 2.7.7.7)	170	1170.127	11,302
319	DNA polymerase III beta subunit (EC 2.7.7.7)	138	369.55	2,602
19	putative membrane protein	167	429.3326	10,346
20	Long-chain-fatty-acid--CoA ligase (EC 6.2.1.3)	164	582.3957	10,098
21	Ribonucleotide reductase of class Ia (aerobic), alpha subunit (EC 1.17.4.1)	165	749.731	9,323
22	unknown	173	296.0219	9,125
23	Carbamoyl-phosphate synthase large chain (EC 6.3.5.5)	163	1041.195	9,094
24	replication-associated protein	32	337.6856	9,027
25	Glutamate synthase [NADPH] large chain (EC 1.4.1.13)	158	1511.417	8,911
26	Rep	20	402.878	8,254
27	DNA-directed RNA polymerase beta' subunit (EC 2.7.7.6)	158	1383.479	8,138
28	Ribonucleotide reductase of class II (coenzyme B12-dependent) (EC 1.17.4.1)	166	824.9139	8,123
86	Ribonucleotide reductase of class Ia (aerobic), beta subunit (EC 1.17.4.1)	150	365.235933	5,021
31	DNA-directed RNA polymerase beta subunit (EC 2.7.7.6)	162	1244.993	8,066
32	Isoleucyl-tRNA synthetase (EC 6.1.1.5)	161	996.6065	7,848
33	unknown protein	161	428.846	7,665
34	DNA polymerase I (EC 2.7.7.7)	171	861.603	7,618
35	DNA gyrase subunit A (EC 5.99.1.3)	161	860.0085	7,524

Raw data

Cumulative Abundance **Mean (mean len)**

Rank	Function	n biomes	Mean Ln	Abundance
1		187	587.6476	48105.51
2	hypothetical protein	187	424.3936	37012.95
3	unnamed protein product	144	490.9272	4716.34
4	Putative p150	69	456.6075	3412.12
5	Transposase	169	359.8556	1834.79
6	structural protein	63	517.6257	1764.06
7	Alu subfamily SX sequence contamination warning entry	41	164.4349	1635.17
8	replication-associated protein	32	337.6856	1481.34
9	Photosystem II CP43 protein (PsbC)	47	466.8228	1429.44
10	Alu subfamily SQ sequence contamination warning entry	43	148.8028	1371.35
11	photosystem II protein D2 (PsbD)	71	337.4417	1224.89
12	cys repeats; this loci is open in all six reading frames, part o	110	674.3229	1216.08
13	unknown	173	296.0219	1117.67
14	Transcriptional regulator	165	281.3998	979.05
15	photosystem II protein D1 (PsbA)	83	357.5047	930.2
16	Cytochrome b6-f complex subunit, cytochrome b6	51	242.3929	925.32
17	ABC transporter ATP-binding protein	165	430.3743	864.43
18	ATP synthase alpha chain (EC 3.6.3.14)	157	514.327	804.47
19	nonstructural protein	27	219.0252	787.31
20	Ribonucleotide reductase of class Ia (aerobic), alpha subuni	165	749.731	776.57
21	Thymidylate synthase thyX (EC 2.1.1.-)	140	268.3299	771.16
22	Single-stranded DNA-binding protein	151	179.3024	769.41
23	Rep; 36.3 kDa	20	303.8045	767.09
24	putative membrane protein	167	429.3326	696.05
25	ABC transporter, ATP-binding protein	158	412.0492	677.86
26	ATP synthase beta chain (EC 3.6.3.14)	156	479.6914	661.21
27	UDP-glucose 4-epimerase (EC 5.1.3.2)	169	338.1789	657.36
28	Ribonucleotide reductase of class Ia (aerobic), beta subunit	150	365.2359	652.32
29	major viral coat protein	28	533.4586	632.83
30	Ribonucleotide reductase of class II (coenzyme B12-depend	166	824.9139	609.01

Raw data

Cumulative Abundance **Mean of (mean len)**

Rank	Function	n biomes	Mean Ln	Abundance
1		187	587.6476	48105.51
2	hypothetical protein	187	424.3936	37012.95
3	unnamed protein product	144	490.9272	4716.34
4	Putative p150	69	456.6075	3412.12
5	Transposase	169	359.8556	1834.79
6	structural protein	63	517.6257	1764.06
7	Alu subfamily SX sequence contamination warning entry	41	164.4349	1635.17
8	replication-associated protein	32	337.6856	1481.34
9	Photosystem II CP43 protein (PsbC)	47	466.8228	1429.44
10	Alu subfamily SQ sequence contamination warning entry	43	148.8028	1371.35
11	photosystem II protein D2 (PsbD)	71	337.4417	1224.89
12	cys repeats; this loci is open in all six reading frames, part o	110	674.3229	1216.08
13	unknown	173	296.0219	1117.67
14	Transcriptional regulator	165	281.3998	979.05
15	photosystem II protein D1 (PsbA)	83	357.5047	930.2
16	Cytochrome b6-f complex subunit, cytochrome b6	51	242.3929	925.32
17	ABC transporter ATP-binding protein	165	430.3743	864.43
18	ATP synthase alpha chain (EC 3.6.3.14)	157	514.327	804.47
19	nonstructural protein	27	219.0252	787.31
20	Ribonucleotide reductase of class Ia (aerobic), alpha subuni	165	749.731	776.57
21	Thymidylate synthase thyX (EC 2.1.1.-)	140	268.3299	771.16
22	Single-stranded DNA-binding protein	151	179.3024	769.41
23	Rep; 36.3 kDa	20	303.8045	767.09
24	putative membrane protein	167	429.3326	696.05
25	ABC transporter, ATP-binding protein	158	412.0492	677.86
26	ATP synthase beta chain (EC 3.6.3.14)	156	479.6914	661.21
27	UDP-glucose 4-epimerase (EC 5.1.3.2)	169	338.1789	657.36
28	Ribonucleotide reductase of class Ia (aerobic), beta subunit	150	365.2359	652.32
140	DNA polymerase III alpha subunit (EC 2.7.7.7)	170	1170.127	252.21
262	DNA polymerase III beta subunit (EC 2.7.7.7)	138	369.55	197.61

Gene abundance in

Rank	Function	nD	Domains	nG	Count	C/nG	%
1	Transposase	4	A B E V	693	26,625	38.42	0.8
2	ABC transporter, ATP-binding protein	4	A B E V	738	9,382	12.71	0.3
3	Sensor histidine kinase	3	A B E	574	5,575	9.71	0.2
4	DNA-binding response regulator	3	A B E	578	4,708	8.15	0.2
5	Methyl-accepting chemotaxis protein	4	A B E V	408	4,389	10.76	0.1
6	ABC transporter, permease protein	3	A B E	580	4,377	7.55	0.1
7	Glycosyltransferase (EC 2.4.1.-)	3	A B E	649	4,172	6.43	0.1
8	Transcriptional regulator, LysR family	3	A B E	441	4,037	9.15	0.1
9	Transcriptional regulator, TetR family	2	A B	535	3,709	6.93	0.1
10	Acetyltransferase, GNAT family	3	A B E	480	3,516	7.33	0.1
11	Transcriptional regulator, AraC family	4	A B E V	459	3,382	7.37	0.1
12	Long-chain-fatty-acid--CoA ligase (EC 6.2.1.3)	3	A B E	589	2,995	5.08	0.1
13	Transcriptional regulator, MarR family	3	A B E	546	2,905	5.32	0.1
14	Permeases of the major facilitator superfamily	4	A B E V	393	2,733	6.95	0.1
15	Acetyltransferase (EC 2.3.1.-)	3	A B E	559	2,436	4.36	0.1
16	Cysteine desulfurase (EC 2.8.1.7)	3	A B E	783	2,362	3.02	0.1
17	3-oxoacyl-[acyl-carrier protein] reductase (EC 1.1.1.100)	3	A B E	706	1,975	2.80	0.1
18	Integrase	4	A B E V	534	1,829	3.43	0.1
19	Outer membrane protein	4	A B E V	415	1,803	4.34	0.1
20	Permease of the drug/metabolite transporter (DMT) superfamily	3	A B E	518	1,746	3.37	0.1
21	D-alanyl-D-alanine carboxypeptidase (EC 3.4.16.4)	2	A B	618	1,716	2.78	0.1
22	UDP-glucose 4-epimerase (EC 5.1.3.2)	3	A B E	655	1,710	2.61	0.1
23	Acyl-CoA dehydrogenase, short-chain specific (EC 1.3.99.2)	3	A B E	459	1,699	3.70	0.1
24	Thioredoxin	4	A B E V	760	1,695	2.23	0.1
25	Transcriptional regulator, GntR family	3	A B E	421	1,639	3.89	0.1