Supplementary Information

Table1 –Adjusted p-values for the KS and Wilcoxon Mann-Whitney U test (p-values lower than 0.05 are highlighted)

Protein feature	Wilcoxon p-value	KS Test p-value
R	4.01E-20	0.00E+00
W	1.37E-19	0.00E+00
F	2.79E-16	0.00E+00
E	8.37E-16	9.19E-13
K	6.46E-15	4.35E-11
H	7.45E-14	1.51E-11
D	8.41E-14	9.88E-10
Q	7.72E-11	0.00E+00
ProteinLength	3.96E-10	3.66E-15
ProteinMolecularWeight	3.96E-10	1.78E-14
AAPcostAverage	1.28E-09	7.04E-09
FlexibilityAverage	5.25E-09	9.48E-08
NumDisorderedResidues	4.74E-08	1.79E-08
N	3.19E-06	5.62E-07
M	3.68E-06	4.43E-08
DisorderFraction	1.16E-05	3.57E-04
S	1.16E-05	6.70E-06
LowComplexityFraction	2.07E-05	1.56E-03
P	0.000227829	0.000357349
GRAVYhydrophobicity	8.01E-04	5.11E-07
C	4.13E-03	1.62E-07
A	0.011677954	0.00019729
G	0.01559121	0.03245018
HelixFraction	0.02164104	0.06164591
Aromaticity	0.06559371	0.009073204
SheetFraction	0.077949175	0.009073204
AAHcostAverage	0.4226544	0.02778622
V	0.42778058	0.06164591
AAMolWeightAverage	0.585415717	0.000625006
Y	0.77123968	0.05819005
I	1	0.03257857
L	1	0.434619
T	1	0.03245018
InstabilityIndex	1	0.000331559
IsoElectricPoint	1	0.05819005
TurnFraction	1	0.06164591
AverageNContent		

Table 2 – Taxonomic summary of the species from which positive set proteins came.

Kingdom/Phylum		Number of Species
Plants		37
Protists		1
Bacteria		2
Animals	Birds	5
	Echinoderms	1
	Fish	3
	Amphibians	1
	Insects	17
	Mammals	3

Table 3- List of individual species in the positive set

Species name	Species type	More specifically
Fenneropenaeus merguiensis	Animal	Arthropod
Dromaius novaehollandiae	Animal	Bird
Meleagris gallopavo	Animal	Bird
Anas platyrhynchos	Animal	Bird
Gallus gallus	Animal	Bird
Coturnix coturnix	Animal	Bird
Strongylocentrotus purpuratus	Animal	Echinoderm
Ichthyomyzon unicuspus	Animal	Fish
Oncorhynchus mykiss	Animal	Fish
Fundulus heteroclitus	Animal	Fish
Xenopus laevis	Animal	Frog
Oscheius brevesophaga	Animal	Insect
Calliphora vicina	Animal	Insect
Blaberus discoidalis	Animal	Insect
Solenopsis invicta	Animal	Insect
Bombyx mori	Animal	Insect
Drosophila melanogaster	Animal	insect
Aedes aegypti	Animal	insect
Caenorhabditis elegans	Animal	Insect
Apis mellifera	Animal	Insect
Manduca sexta	Animal	Insect

Anopheles gambiae	Animal	Insect
Drosophila simulans	Animal	Insect
Anthonomus grandis	Animal	insect
Periplaneta americana	Animal	Insect
Antheraea pernyi	Animal	insect
Trichoplusia ni	Animal	Insect
Homo sapiens	Animal	Mammal
Oryctolagus cuniculus	Animal	Mammal
Mus musculus	Animal	Mammal
Physarum polycephalum	Animal	Protist
Azotobacter vinelandii	Bacterium	
Bacillus subtilis	Bacterium	
Nicotiana langsdorffii	Plant	
Capparis masaikai	Plant	
Vicia faba	Plant	
Ipomoea batatas	Plant	
Nicotiana plumbaginifolia	Plant	
Macadamia integrifolia	Plant	
Arachis hypogaea	Plant	
Ricinus communis	Plant	
Cicer arietinum	Plant	
Raphanus sativus	Plant	
Cucurbita maxima	Plant	
Picea glauca	Plant	
Canavalia ensiformis	Plant	
Arabidopsis thaliana	Plant	
Triticum aestivum	Plant	
Hordeum vulgare	Plant	
Lathyrus sativus	Plant	
Mesembryanthemum crystallinum	Plant	
Glycine soja	Plant	
Fagopyrum esculentum	Plant	
Clitoria ternatea	Plant	
Phaseolus vulgaris	Plant	
Brassica juncea	Plant	
Helianthus annuus	Plant	
Gossypium hirsutum	Plant	
Sinapis alba	Plant	
Sesamum indicum	Plant	
Brassica napus	Plant	
Pisum sativum	Plant	
Solanum tuberosum	Plant	
Bertholletia excelsa	Plant	

Panax ginseng	Plant	
Glycine max	Plant	
Ipomoea nil	Plant	
Psophocarpus tetragonolobus	Plant	
Sorghum bicolor	Plant	
Zea mays	Plant	
Phaseolus lunatus	Plant	
Canavalia gladiata	Plant	
Avena sativa	Plant	
Sinapis arvensis	Plant	
Solanum cardiophyllum	Plant	
Prunus persica	Plant	
Populus deltoides	Plant	
Oryza sativa	Plant	
Lupinus angustifolius	Plant	
Theobroma cacao	Plant	

Table 4 Detailed descriptions of protein features.

Variable	**Description**	
Number and Percentage of disordered residues in the protein	Measures the amount of naturally unfolded residues, not predicted to form part of a particular tertiary structure, but may assume order upon binding to substrates or other proteins in vivo.	IUPred
Percentage of low complexity regions	Regions of biased amino acid composition, which can be clustered, sporadic or periodic. The often repeated or repetitive nature of such regions can result from and result in expansion of the regions. Includes tandem repeats	Seg
Amino Acid Percentages	For each of the 21 main amino acids, the percentage of each in the protein is given.	ProtParam
Aromaticity	A measure developed by [19], denoting the relative frequencies of Phenylalanine, Tryptophan and Tyrosine, which contain conjugated carbon rings.	
Instability Index	A measure described in [20], based on the composition of dipeptides in the protein. High instability indexes suggest a protein likely to be degraded rapidly in a test tube. http://web.expasy.org/protparam/protparam-doc.html	
Flexibility Average	Flexibility indices based on a rolling window of 9 amino acids, described by [21], using parameters derived from the study of existing crystal structures.	
Isoelectric point	The pH at which a particular molecule or surface carries no net electrical charge. calculated using pKa values of amino	

	acids. The pKa value of Amino acids depends on its side chain	
Secondary structure percentages (sheet, turn, helix)	The percentage of amino acids predicted to form parts of beta sheets, turns and alpha helices	
Hydrophobicity	The GRAVY value for a protein or a peptide is calculated by adding the hydropathy values [22] of each amino acid residue and dividing by the number of residues in the sequence, or length of the sequence. Increasing positive score indicates a greater hydrophobicity.	
Metabolic cost Hcost, Pcost, Ncost	Based on the cost of producing each amino acid in terms of hydrogen bond and phosphate bonds we can calculate an average for each protein overall, taken from [12]. We might expect cheap amino acids to be used if the aim is solely to store nitrogen. Average of per-residue values.	Myself
Molecular weight, molecular weight per residue	Molecular weight is another method to measure the cost of producing a protein metabolically.	Protparam

Table 5 – Matrix giving P-bond metabolic cost, H-bond metabolic cost and N content scores for each amino acid (P and H bond costs calculated from [12]

AMINO ACID	MOLECULAR WEIGHT	P BOND COST	H BOND COST	N content
Alanine - Ala - A	89	3	16	1
Arginine - Arg - R	174	32	25	4
Asparagine - Asn - N	132	10	17	2
Aspartate - Asp- D	133	4	17	1
Cysteine - Cys - C	121	22	26	1
Glutamate - Glu - E	147	8	19	1
Glutamine - Gln - Q	146	11	19	2
Glycine - Gly - G	75	7	14	1
Histidine - His - H	155	61	27	2
Isoleucine - Ile - I	131	13	42	1
Leucine - Leu -L	131	8	37	1
Lysine - Lys - K	146	13	39	2
Methionine - Met - M	149	29	37	1

Phenylalanine - Phe -F	165	40	58	1
Proline - Pro - P	115	11	25	1
Serine - Ser - S	105	7	14	1
Threonine - Thr - T	119	10	23	1
Tryptophan - Trp - W	204	83	70	2
Tyrosine -Tyr - Y	181	40	55	1
Valine - Val - V	117	6	32	1

Table 6 - *M.smegmatis* top storage protein candidates from the classifier

ID	P(negative)	P(positive)	Annotations
MSMEG_6354	0.1672	0.8328	Serine esterase, cutinase family protein http://www.uniprot.org/uniprot/A0R5Y1
MSMEG_2016	0.1848	0.8152	Molybdate ABC transporter, periplasmic molybdate-binding protein http://www.uniprot.org/uniprot/A0QTZ2
MSMEG_3479	0.2102	0.7898	**Probable thiol peroxidase** http://www.uniprot.org/uniprot/A0QXZ5
MSMEG_1136	0.2164	0.7836	**Uncharacterized protein** http://www.uniprot.org/uniprot/A0QRJ2
MSMEG_4330	0.2224	0.7776	**Short chain dehydrogenase** http://www.uniprot.org/uniprot/A0R0B7
MSMEG_6133	0.223	0.777	**5-dehydro-4-deoxyglucarate dehydratase** http://www.uniprot.org/uniprot/A0R5B6
MSMEG_5518	0.2338	0.7662	**Antigen 34 kDa** http://www.uniprot.org/uniprot/A0R3L7
MSMEG_4804	0.2484	0.7516	Uncharacterized protein http://www.uniprot.org/uniprot/A0R1M2
MSMEG_5275	0.2544	0.7456	**Conserved transmembrane protein** http://www.uniprot.org/uniprot/A0R2Y3
MSMEG_2474	0.255	0.745	**Probable cutinase Cut3**
MSMEG_6180	0.2588	0.7412	http://www.uniprot.org/uniprot/A0QV75
MSMEG_4465	0.2632	0.7368	**Cutinase** http://www.uniprot.org/uniprot/A0R0Q0
MSMEG_5268	0.2646	0.7354	**Uncharacterized protein http://www.uniprot.org/uniprot/A0R2X6**

MSMEG_1528	0.3016	0.6984	**Cutinase** http://www.uniprot.org/uniprot/A0QSM2
MSMEG_4010	0.3036	0.6964	**Glyoxalase family protein http://www.uniprot.org/uniprot/A0QZF8**
MSMEG_5878	0.3096	0.6904	**Serine esterase, cutinase family protein** http://www.uniprot.org/uniprot/A0R4L6
MSMEG_0242	0.3126	0.6874	**Putative exported protein** http://www.uniprot.org/uniprot/A0QP19
MSMEG_0344	0.3128	0.6872	**Uncharacterized protein** http://www.uniprot.org/uniprot/A0QPB9
MSMEG_2742	0.3172	0.6828	**DNA-damage-inducible protein** http://www.uniprot.org/uniprot/A0QVY7

Table 6b-Pinney transcriptomics-based candidates in *M.smegmatis*

Record ID	P(negative)	P(positive)	
MSMEG_0945	0.6056	0.3944	negative
MSMEG_0845	0.5522	0.4478	negative
MSMEG_0879	0.4456	0.5544	positive
MSMEG_1999	0.4318	0.5682	positive
MSMEG_2654	0.5616	0.4384	negative
MSMEG_1473	0.5392	0.4608	negative
MSMEG_1398	0.672	0.328	negative
MSMEG_1520	0.5908	0.4092	negative
MSMEG_1521	0.7782	0.2218	negative
MSMEG_3840	0.3792	0.6208	positive
MSMEG_1339	0.6624	0.3376	negative
MSMEG_5489	0.6168	0.3832	negative
MSMEG_6895	0.639	0.361	
MSMEG_6946	0.5704	0.4296	negative

First 4 rows are the top candidates from the transcriptomic approach, while the other 10 are thought to be ribosomal false positives.

Table 7- Banana Top Storage Protein Candidates

ID	P(negative)	P(positive)
GSMUA_Achr5P18440_001	0.0078	0.9922
GSMUA_AchrUn_randomP00050_001	0.0168	0.9832

GSMUA_Achr4P01070_001	0.0186	0.9814
GSMUA_Achr8P18750_001	0.0204	0.9796
GSMUA_Achr8P33370_001	0.0324	0.9676
GSMUA_Achr5P26690_001	0.0354	0.9646
GSMUA_Achr8P33390_001	0.0354	0.9646
GSMUA_Achr1P25160_001	0.0398	0.9602
GSMUA_Achr8P29820_001	0.044	0.956
GSMUA_Achr9P30640_001	0.0468	0.9532
GSMUA_AchrUn_randomP28090_001	0.0548	0.9452
GSMUA_Achr9P16110_001	0.0752	0.9248
GSMUA_Achr5P26700_001	0.0816	0.9184
GSMUA_Achr5P26760_001	0.086	0.914
GSMUA_Achr5P26750_001	0.0906	0.9094
GSMUA_Achr5P26730_001	0.0916	0.9084
GSMUA_Achr2P18710_001	0.0978	0.9022
GSMUA_Achr8P22430_001	0.1054	0.8946
GSMUA_Achr8P22440_001	0.1114	0.8886
GSMUA_Achr5P26710_001	0.117	0.883
GSMUA_Achr4P18320_001	0.1312	0.8688
GSMUA_Achr8P09920_001	0.1326	0.8674
GSMUA_Achr4P17250_001	0.1376	0.8624
GSMUA_Achr7P02840_001	0.1392	0.8608

Table 8 - Platypus Top Storage Candidates

ID	Negative	Positive
ENSOANP00000026802	0.1002	0.8998
ENSOANP00000030793	0.1144	0.8856
ENSOANP00000020379	0.1448	0.8552
ENSOANP00000032691	0.1728	0.8272
ENSOANP00000032612	0.1778	0.8222
ENSOANP00000025341	0.1802	0.8198
ENSOANP00000024660	0.1838	0.8162
ENSOANP00000022600	0.1948	0.8052
ENSOANP00000014737	0.1954	0.8046
ENSOANP00000001565	0.1966	0.8034
ENSOANP00000025893	0.1976	0.8024
ENSOANP00000006813	0.202	0.798
ENSOANP00000008771	0.2022	0.7978
ENSOANP00000014293	0.2036	0.7964
ENSOANP00000030628	0.204	0.796

ENSOANP00000010585	0.2068	0.7932
ENSOANP00000008316	0.2082	0.7918
ENSOANP00000028489	0.2082	0.7918
ENSOANP00000023658	0.213	0.787
ENSOANP00000009038	0.2192	0.7808
ENSOANP00000007572	0.2256	0.7744
ENSOANP00000022547	0.2326	0.7674
ENSOANP00000009711	0.2384	0.7616
ENSOANP00000010521	0.243	0.757
ENSOANP00000014722	0.2448	0.7552
ENSOANP00000016820	0.2474	0.7526
ENSOANP00000009156	0.2474	0.7526
ENSOANP00000016845	0.2478	0.7522
ENSOANP00000027333	0.2504	0.7496
ENSOANP00000017325	0.2512	0.7488
ENSOANP00000014714	0.2512	0.7488
ENSOANP00000022710	0.2534	0.7466
ENSOANP00000016819	0.2558	0.7442
ENSOANP00000029754	0.256	0.744
ENSOANP00000015157	0.26	0.74
ENSOANP00000028044	0.2602	0.7398