Proteomic Universal Correlate of Evolution

Erez Persi and David Horn
Tel Aviv University

http://horn.tau.ac.il

ISMB Highlight track 2014
based on

Systematic Analysis of Compositional Order of Proteins Reveals New Characteristics of Biological Functions and a Universal Correlate of Macroevolution
PLOS CB 2013 9(11): e1003346

Compositional Order (CO)

includes

- Runs of Amino-acids (frequent in human proteins). Most extreme example: ***Ataxin-8***

  MQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ
  QQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQQ

- High purity Repetitive Motives. Example: ***PRDM9,*** Zinc-finger, containing

  CRECGRGFSWKSHLLIHQRIHTGEKPYV (**x12** human; **x10** mouse)

- lower purity Repetitive Motives

- low-complexity regions and large compositional bias

Known generating mechanisms include replication slippage and recombination effects.

Levinson & Gutman (1987): **Slipped-Strand Mispairing : A major mechanism for DNA sequence evolution** *Mol. Biol. Evol.*

Longer repeats are more likely to generate even longer repeats, thus increasing repeat homogeneity.

Paques, Leung, Haber (1998): **Expansions and contractions in a tandem repeat induced by double-strand break repair** *Mol. Cell. Biol*.

Our attitude: Don't filter it out. Evolution kept it, and we should study it.

Functional Importance

- The clock gene period (*per*) in Drosophila. T-G repeat variation. The longer allele is more frequent in cold environment, such that temperature fluctuations affect less the circadian cycle.

Sawyer et al. (1997): **Natural variation in a Drosophila clock gene and temperature compensation**. *Science.*

- Repetitive elements in developmental genes of 92 breeds of dogs => Selection for elevated purity. Repeat number variation associates with limb, skull morphology

Fondon and Garner (2004): **Molecular origins of rapid and continuous morphological evolution**. *PNAS.*

- Recombination effects leading to tandem repeat number variation in cell-wall proteins correlates with phenotypic traits in Yeast => cell-cell adhesion, evasion of immune system

Verstrepen KJ et al. (2005). **Intragenic tandem repeats generate functional variability**. *Nature genetics.*

- Several Cancers & human inherited neurodegenerative diseases are related to proteins which contain long runs.
  - Glutamine runs, Alanine runs
  - Multiple runs of amino-acids, e.g., huntingtin protein containing Q_{23}, P_{11}, P_{10}, E_5, E_6

Karlin S et al. (2002). **Amino acid runs in eukaryotic proteomes and disease associations**. *PNAS*.

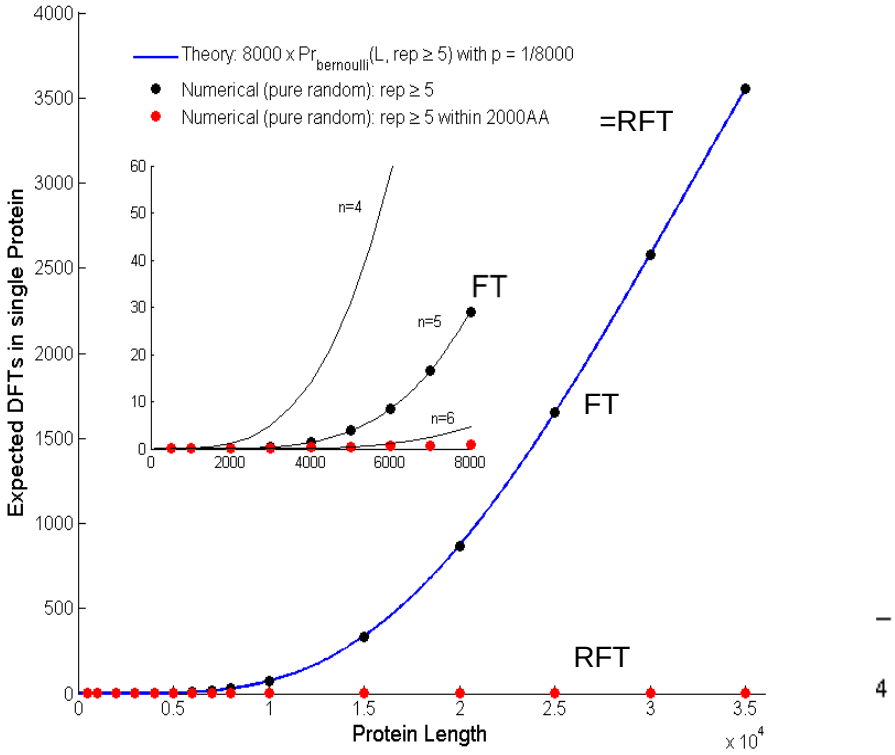
Gemayel et al. (2010). **Variable tandem repeats accelerate evolution of coding and regulatory sequences**. *Ann Rev Gen.*
Kashi & King (2006). **Simple sequence repeats as advantageous mutators in evolution**. *Trends Genet.*

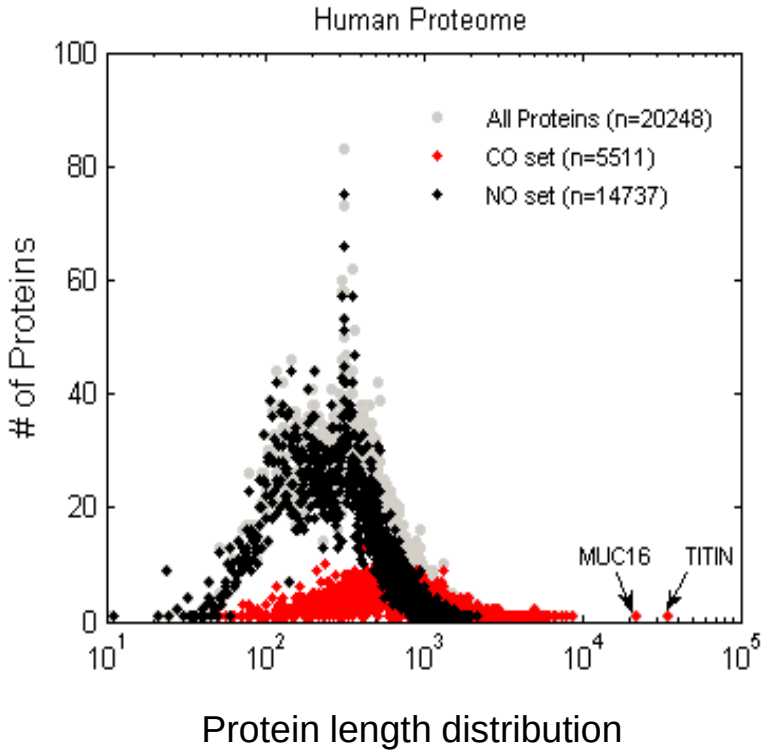
New Global Characterization

- Amino Acid Triplets - a set of $20^3 = 8000$ elements.
- Purpose: Identification of non-random patterns.

A **Frequent Triplet** (FT) = a triplet of amino-acids that occurs at least 5 times on a protein

$$Pr(L, p, i \geq n) = \sum_{i=n}^{L} \binom{L}{i} p^i (1-p)^{L-i}$$

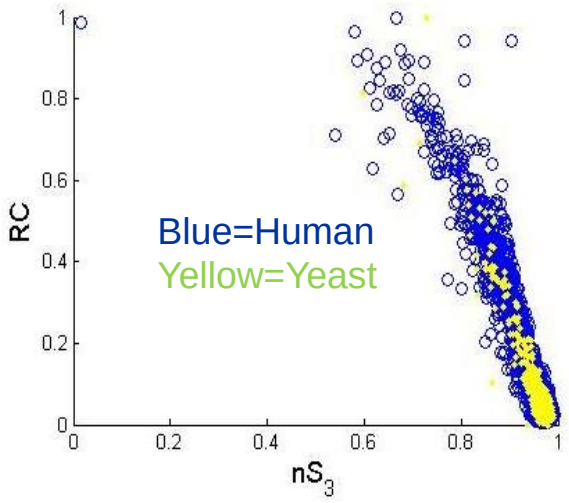




Protein length distribution

Measures of Compositional Order

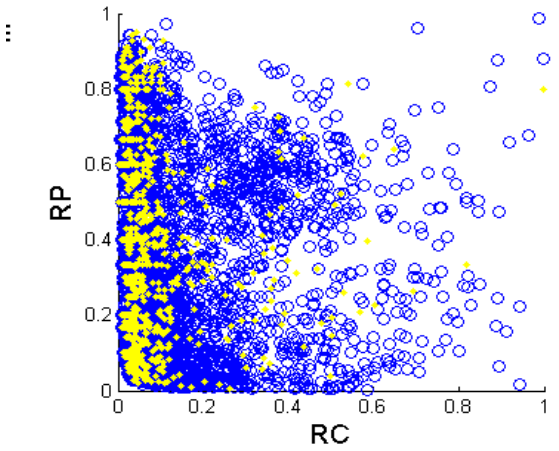
- **Regularity**
  - Entropy of single amino-acid -> entropy of triplets: $S_k = -\sum_{i=1}^{N_k} \frac{n_i}{L} \log_2(\frac{n_i}{L})$
  - **Relative coverage (RC)** = total coverage of FTs / protein length



RC *vs* normalized entropy of amino-acid triplets

- **Periodicity -**
  - Interval = distance between two consecutive appearances of a FT on a protein.
  - **Most Frequent Interval (MFI)** = empirical, from all intervals (of all FTs).
  - **Relative periodicity (RP)** = number of FT occurrences at MFI / total number of FT occurrences

MFI may be used to define a period, e.g. by requiring existence of 4 equal intervals.
Several different periods on same protein can be observed.



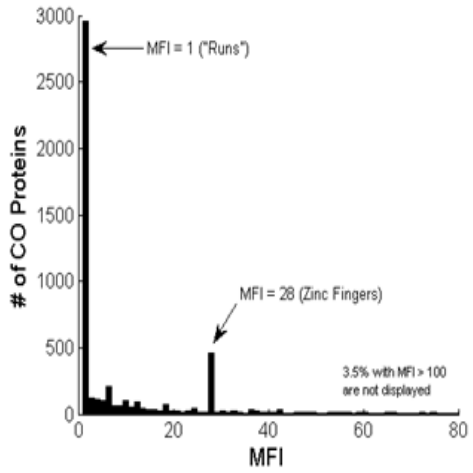
Note independence of RP and RC

- **Vocabulary**
  - **Number of different FTs (DFT)** = in a protein (or in a proteome).
  - Insensitive to redundancy

Protein	# AA	# DFT	Leading FT	RC, RP (MFI)	Amino-acid sequence (leading FTs highlighted)
CAMKV ATP binding	501	5	TPA PAT ATD	0.1, 0.69 (8)	MPFGCVTLGDKKNYNQPSEVTDRYDLGQVIKTEEFCEIFRAKDKTTGKLHT CKKFQKRDGRKVRKAAKNEIGILKMVKHPNILQLVDVFVTRKEYFIFLELA TGREVFDWILDQGYYSERDTSNVVRQVLEAVAYLHSLKIVHRNLKLENLVY YNRLKNSKIVISDFHLAKLENGLIKEPCGTPEYLAPEVVGRQRYGRPVDCW AIGVIMYILLSGNPPFYEEVEEDDYENHDKNLFRKILAGDYEFDSPYWDDI SQAAKDLVTRLMEVEQDQRITAEEAISHEWISGNAASDKNIKDGVCAQIEK NFARAKWKKAVRVTTLMKRLRAPEQSSTAAAQSASATDTATPGAAGGATAA AASGATSAPEGDAARAAKSDNVAPADRSATPATDGSATPATDGSVTPATDG SITPATDGSVTPATDRSATPATDGRATPATEESTVPTTQSSAMLATKAAAT PEPAMAQPDSTAPEGATGQAPPSSKGEEAAGYAQESQREEAS
ASPX Acroso-mal protein	265	6	SGE	0.24, 0.35 (5)	MNRFLLLMSLYLLGSARGTSSQPNELSGSIDHQTSVQQLPGEFFSLENPSD AEALYETSSGLNTLSEHGSSEHGSSKHTVAEHTSGEHAESEHASGEPAATE HAEGEHTVGEQPSGEQPSGEHLSGEQPLSELESGEQPSDEQPSGEHGSGEQ PSGEQASGEQPSGEHASGEQASGAPISSTSTGTILNCYTCAYMNDQGKCLR GEGTCITQNSQQCMLKKIFEGGKLQFMVQGCENMCPSMNLFSHGTRMQIIC CRNQSFCNKI
PRDM9 Zinc finger	894	28	HQR HTG TGE GEK YVC VCR CRE ECG	0.36, 0.84 (28)	MSPEKSQEESPEEDTERTERKPMVKDAFKDISIYFTKEEWAEMGDWEKTRY RNVKRNYNALITIGLRATRPAFMCHRRQAIKLQVDDTEDSDEEWTPRQQVK PPWMALRVEQRKHQKGMPKASFSNESSLKELSRTANLLNASGSEQAQKPVS PSGEASTSGQHSRLKLELRKKETERKMYSLRERKGHAYKEVSEPQDDDYLY CEMCQNFFIDSCAAHGPPTFVKDSAVDKGHPNRSALSLPPGLRIGPSGIPQ AGLGVWNEASDLPLGLHFGPYEGRITEDEEAANNGYSWLITKGRNCYEYVD GKDKSWANWMRYVNCARDDEEQNLVAFQYHRQIFYRTCRVIRPGCELLVWY GDEYGQELGIKWGSKWKKELMAGREPKPEIHPCPSCCLAFSSQKFLSQHVE RNHSSQNFPGPSARKLLQPENPCPGDQNQEQQYPDPHSRNDKTKGQEIKER SKLLNKRTWQREISRAFSSPPKGQMGSCRVGKRIMEEESRTGQKVNPGNTG KLFVGVGISRIAKVKYGECGQGFSVKSDVITHQRTHTGEKLYVCRECGRGF SWKSHLLIHQRIHTGEKPYVCRECGRGFSWQSVLLTHQRTHTGEKPYVCRE CGRGFSRQSVLLTHQRRHTGEKPYVCRECGRGFSRQSVLLTHQRRHTGEKP YVCRECGRGFSWQSVLLTHQRTHTGEKPYVCRECGRGFSWQSVLLTHQRTH TGEKPYVCRECGRGFSNKSHLLRHQRTHTGEKPYVCRECGRGFRDKSHLLR HQRTHTGEKPYVCRECGRGFRDKSNLLSHQRTHTGEKPYVCRECGRGFSNK SHLLRHQRTHTGEKPYVCRECGRGFRNKSHLLRHQRTHTGEKPYVCRECGR GFSDRSSLCYHQRTHTGEKPYVCREDE

Analysis of the Human Proteome

- ~ 27% are CO.
- Periodic structures
  - Abundance of runs and zinc-fingers



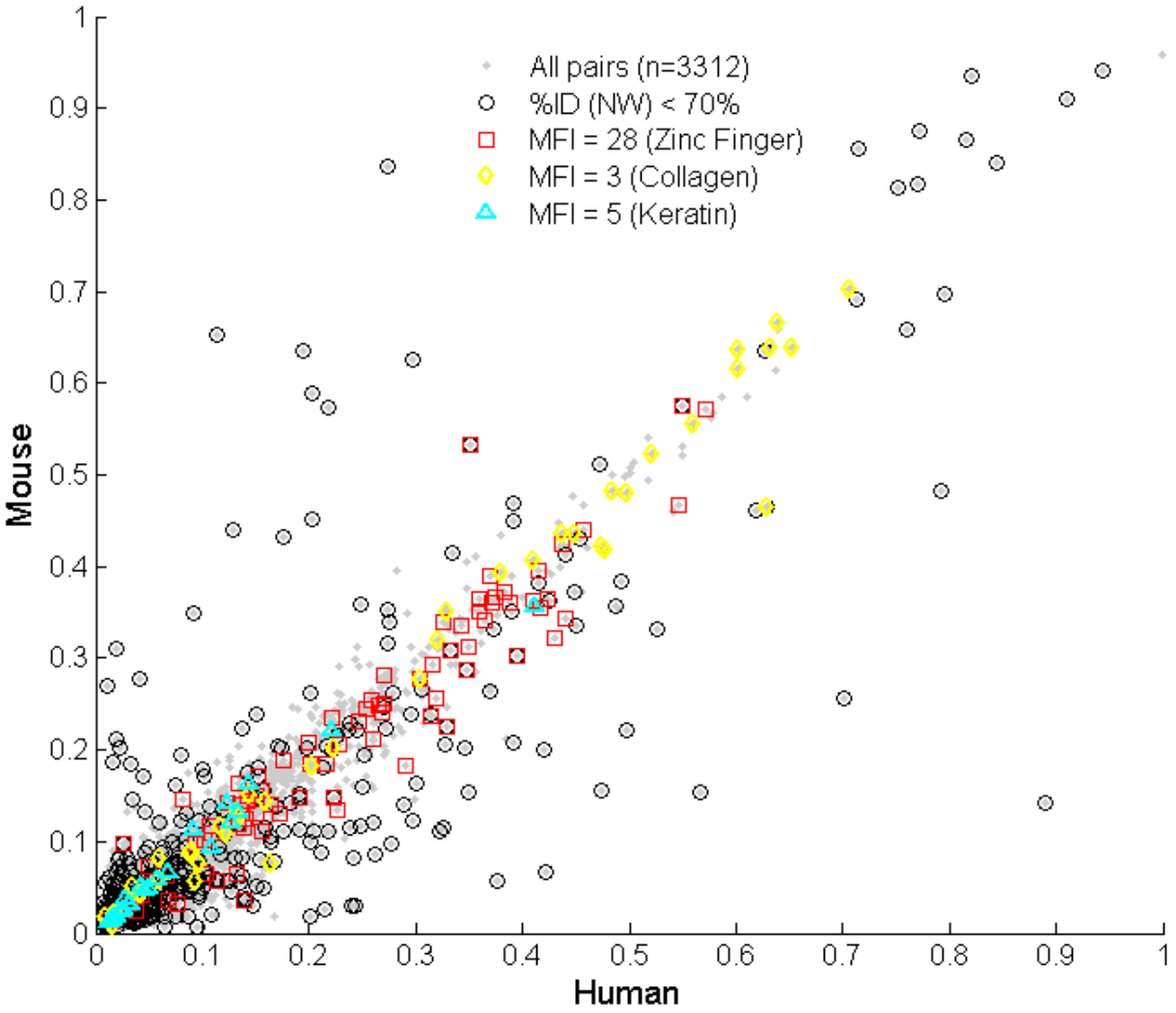
- Some outstanding enrichments

Function	within the proteome	within proteins containing FTs	Mean RP	Mean RC
Disease	2755 (13.6%)	903 (16.4%)	0.3	0.1
Zinc Fingers	1799 (8.9%)	977 (17.7%)	0.43	0.17
Collagen	166 (0.8%)	87 (1.6%)	0.21	0.25
Keratin	162 (0.8%)	100 (1.8%)	0.27	0.39

averages in the CO set: <RP>=0.35 and <RC>=0.1.

Highly enriched in RC: Keratin, collagen, filament and cell adhesion proteins **fast evolving**
Highly enriched in RP: Neuro and immune system proteins **new functions**
Non-monotonic behavior: DNA binding, regulation transcription **enrichment with run length**

Human Mouse Orthologs: RC (mouse protein) *vs* RC (human protein)



- On diagonal find ZF, collagen and keratin, where function conserved CO structure.
- Off diagonal correlate with low homology and display large deviations between the CO structures, pointing to larger losses in the mouse lineage, presumably because of higher substitution rate.

modified orthologs

ASPX – 61% ID

HUMAN

MFI = 5, LFT = SGE, RP = 0.35, RC = 0.24

83AA...

SGEHAESEHA(10)SGEPAATEHAEGEHTVGEQP(20)
SGEQP(5)SGEHL(5)SGEQPLSELE(10)SGEQPSDE
QP(10)SGEHG(5)SGEQP(5)SGEQA(5)SGEQP(5)
SGEHA(5)

...96AA

MOUSE

MFI = 14, LFT = SGE, RP = 0.2, RC = 0.11

89AA...

SGEQSSEHMSGDHM(14)SGEHLSEHTSEEHS(14)SGE
HTSTEHT(10)SGEQPATEQSSSDQPSEAS(19)SGE

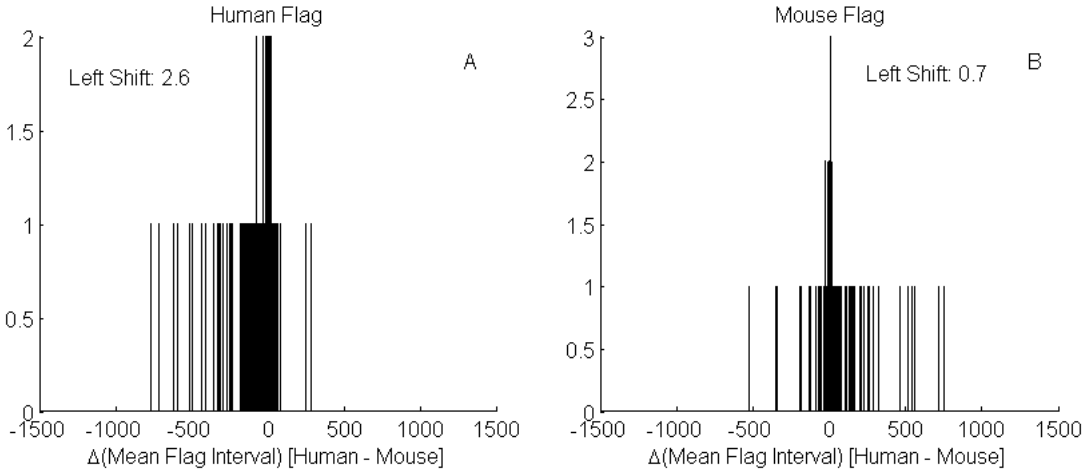
...112AA

The human protein exhibits high RP with MFI=5, but also harmonics of 10 and 20, suggesting rapid evolvement. The mouse protein has intervals of 10, 14 ,19 and lower RC, suggesting deterioration of the periodic structure due to high substitution rate.

In general mouse exhibits higher harmonics, seen in a study of over 200 orthologs with low sequence similarity but with periodic structures.

Flag=leading FT

Left shift= Δ <0 / Δ >0



Non-Orthologs: more 'innovation' in the human lineage

species	Orthology	# of CO proteins	**RP**(p-value)	**RC**(p-value)
Human (n=5511)	V (CO in mouse)	3312	0.33	0.09
	V (NO in mouse)	831	**0.4**(2.1×10^{-35})	**0.03**(6.02×10^{-68})
	X	1368	**0.36**(2.25×10^{-11})	**0.19**(7.56×10^{-62})
Mouse (n=4063)	V (CO in human)	3312	0.33	0.08
	V (NO in human)	626	**0.44**(1.16×10^{-34})	**0.04** (1.18×10^{-51})
	X	125	0.34	**0.16**(9.8×10^{-5})

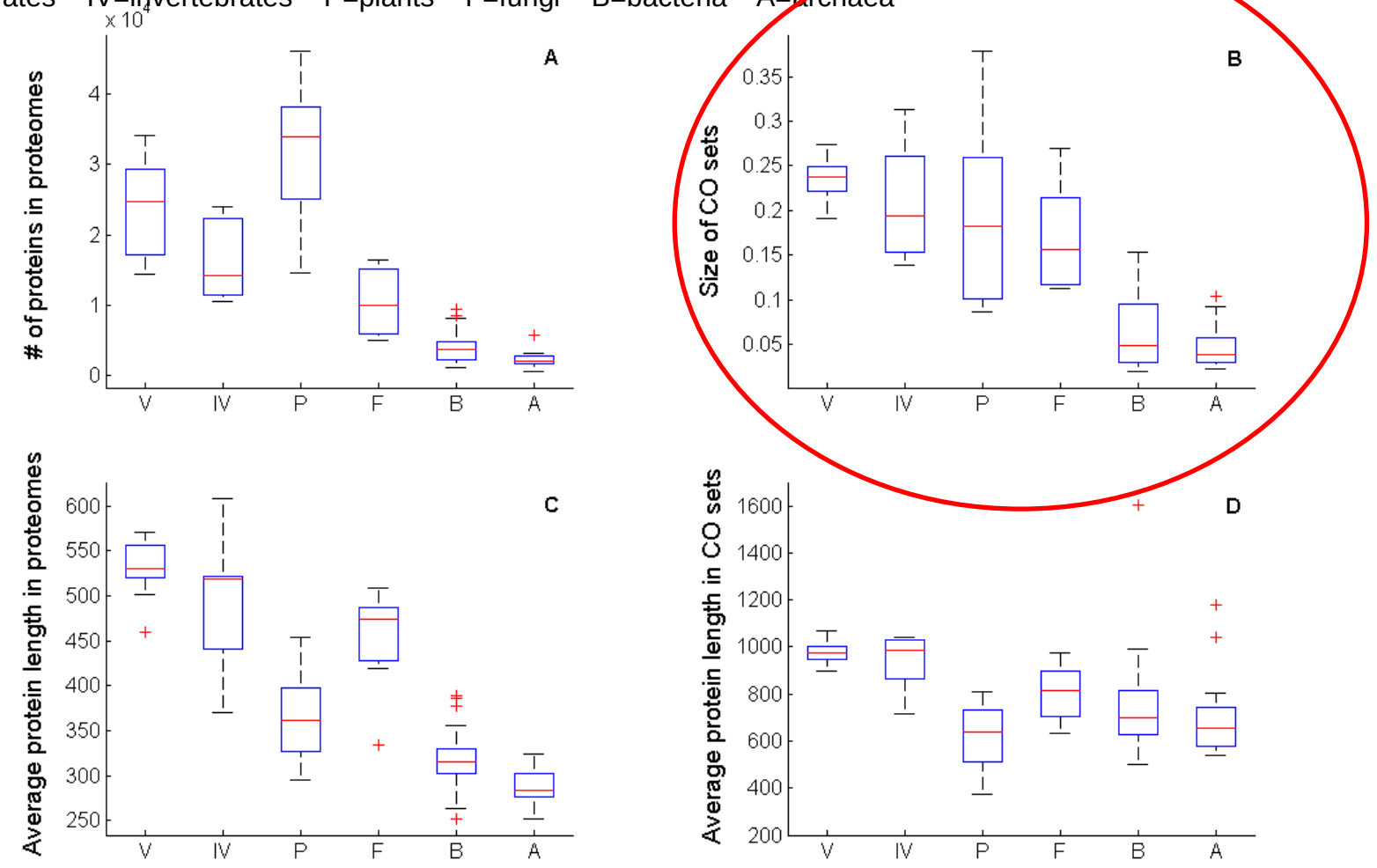
- CO proteins whose orthologs lost/gain CO (in the respective species) have high RP but low RC values. Similar GO contents - many are nervous system related.

- More Novel CO proteins in human 1368/125.
  Many are Zinc fingers (433/977), keratin-associated proteins (61/94) and protocadherins (44/55).

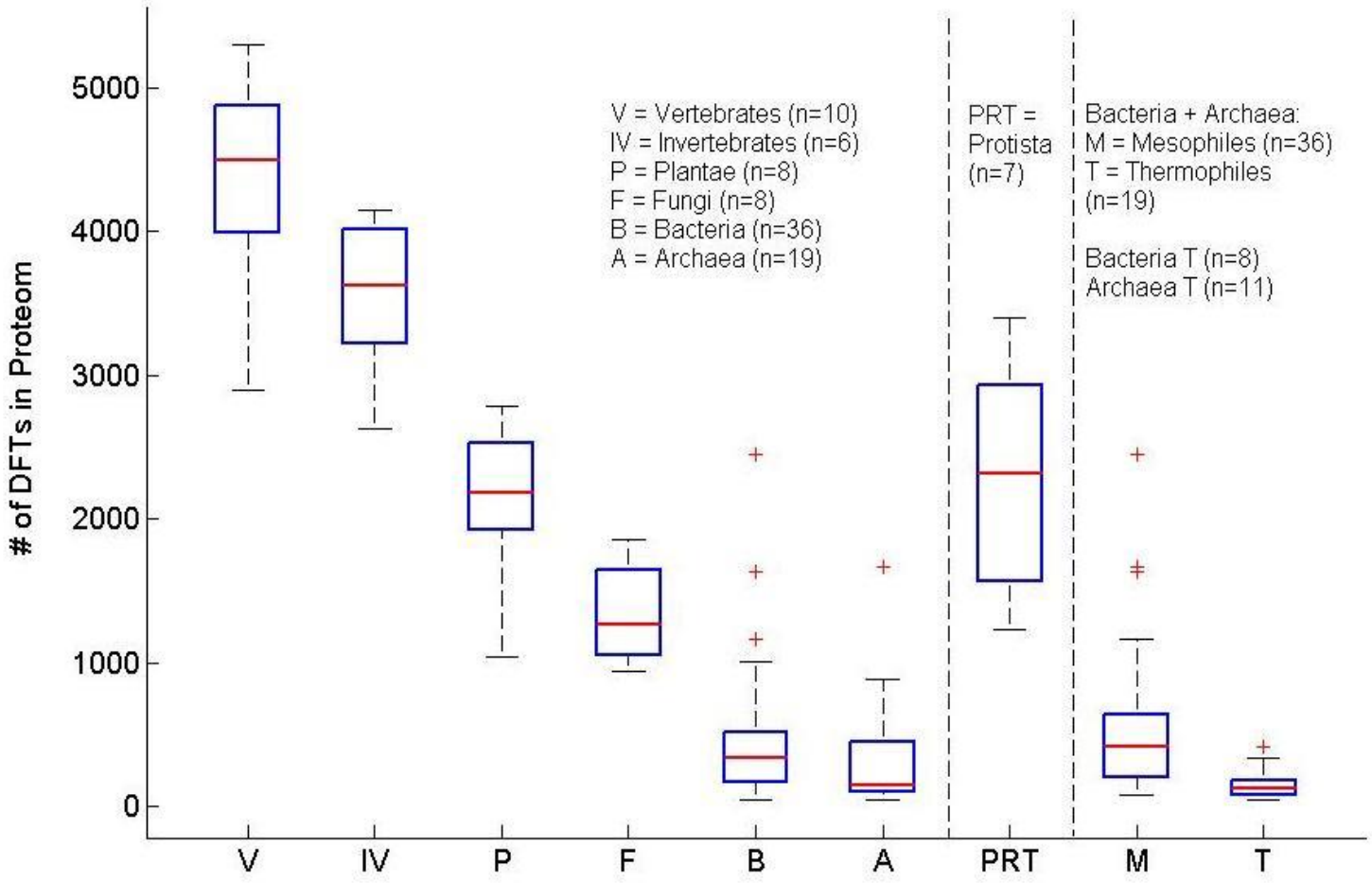
Sequence-information Markers of Evolution ?

- (A,C) - No obvious distinguishing observable (length, # of genes).
- (B) - CO proteins are more abundant in Eukaryotes (Marcotte et al 1998).
- (C,D) - The length of proteins in the CO sets are larger and more homogenous than all proteins in the proteome.

V=vertebrates IV=invertebrates P=plants F=fungi B=bacteria A=archaea

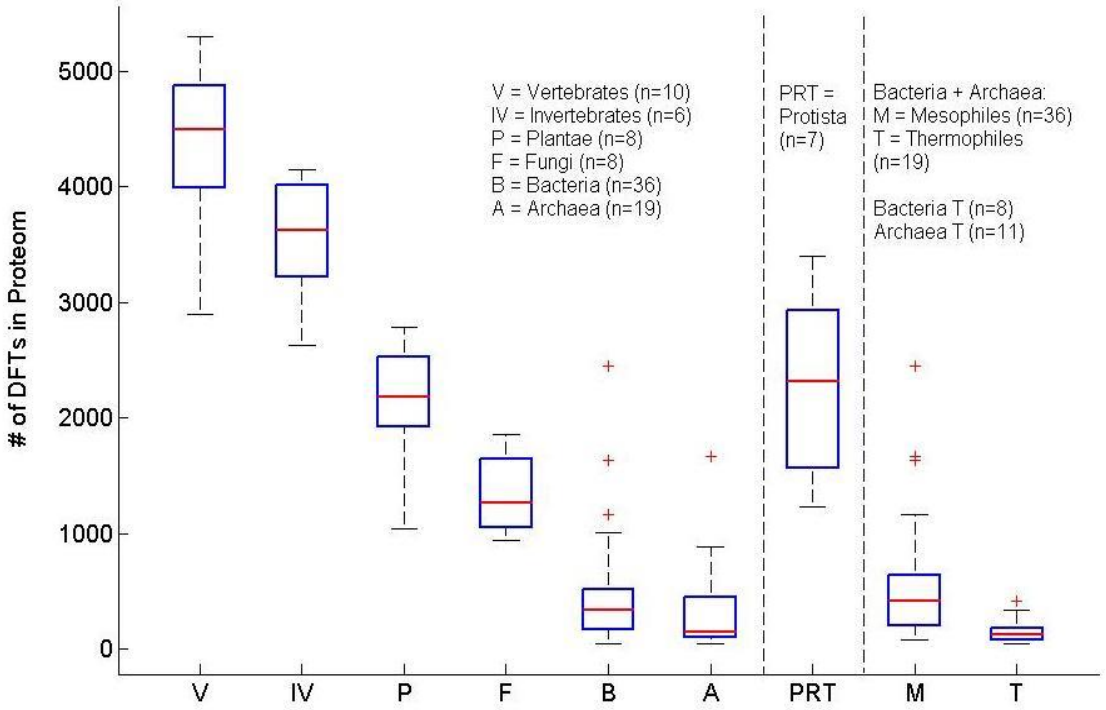


CO Vocabulary - A Rare Marker of Evolution



P-values according to two-sample Kolmogorov-Smirnov test are:
2.5×10^{-2} (**V-IV**), 6.5×10^{-3}(**IV-P**), 9×10^{-3} (**P-F**), 1.7×10^{-5} (**F-B**), and 1.4×10^{-4} (**M-T**).

Conclusion: major CO generation may occur during the creation of completely new species, i.e. during macroevolutionary events.



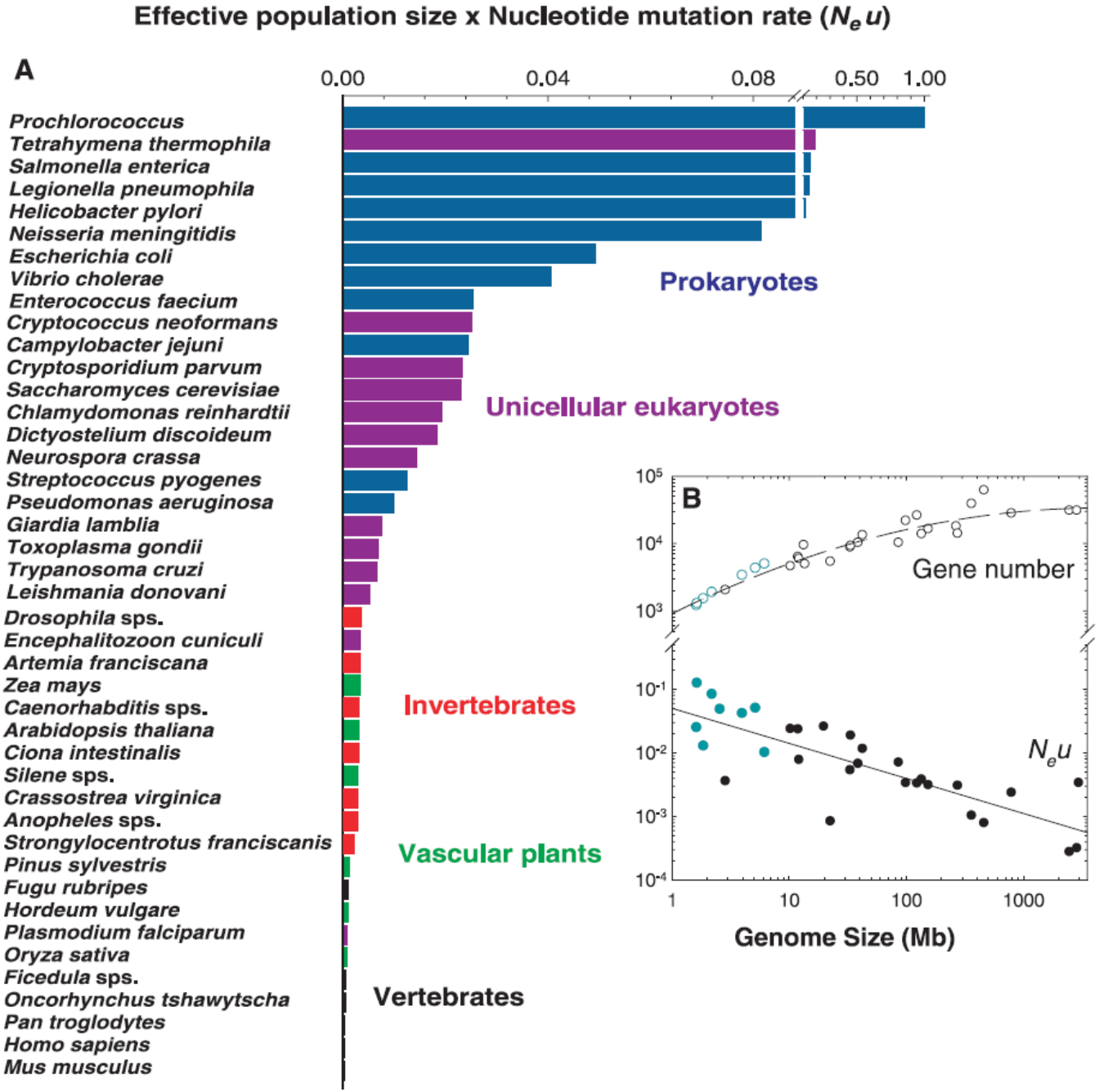
Common knowledge: Macroevolutionary changes are invariably connected to major genomic changes. Novel taxa and novel functions are marked by gene and chromosome rearrangement, and segmental duplications.

It must also include duplication of sections of genes, large and small motifs, and formation of novel CO material.

Lynch and Cannery: **The Origins of Genome Complexity**. Science 2003

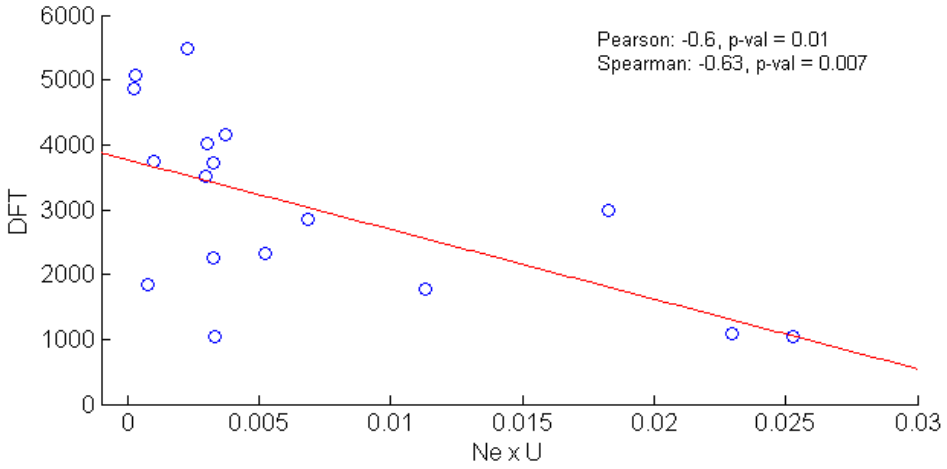
Transitions from prokaryotes to unicellular eukaryotes to multicelular eukaryotes are associated with orders-of-magnitude reductions in population size. By magnifying the power of *random genetic drift*, this provides a permissive environment for proliferation of genomic features that would otherwise be eliminated by purifying selection.

Silent-site variations among alleles provide an estimate of 'effective population size' X 'mutation rate', Ne*u, of a species.

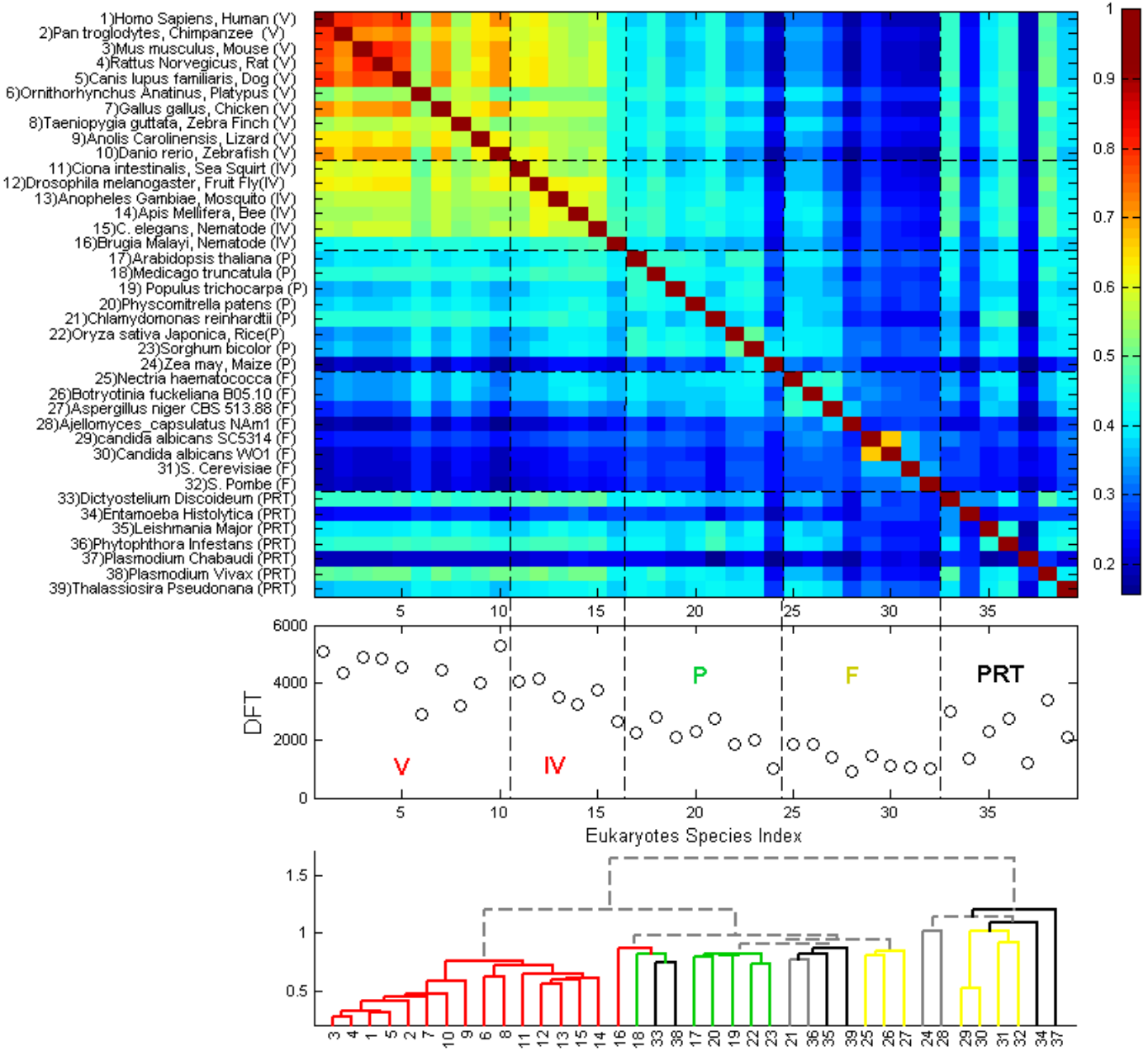


DFT vs. Ne x U

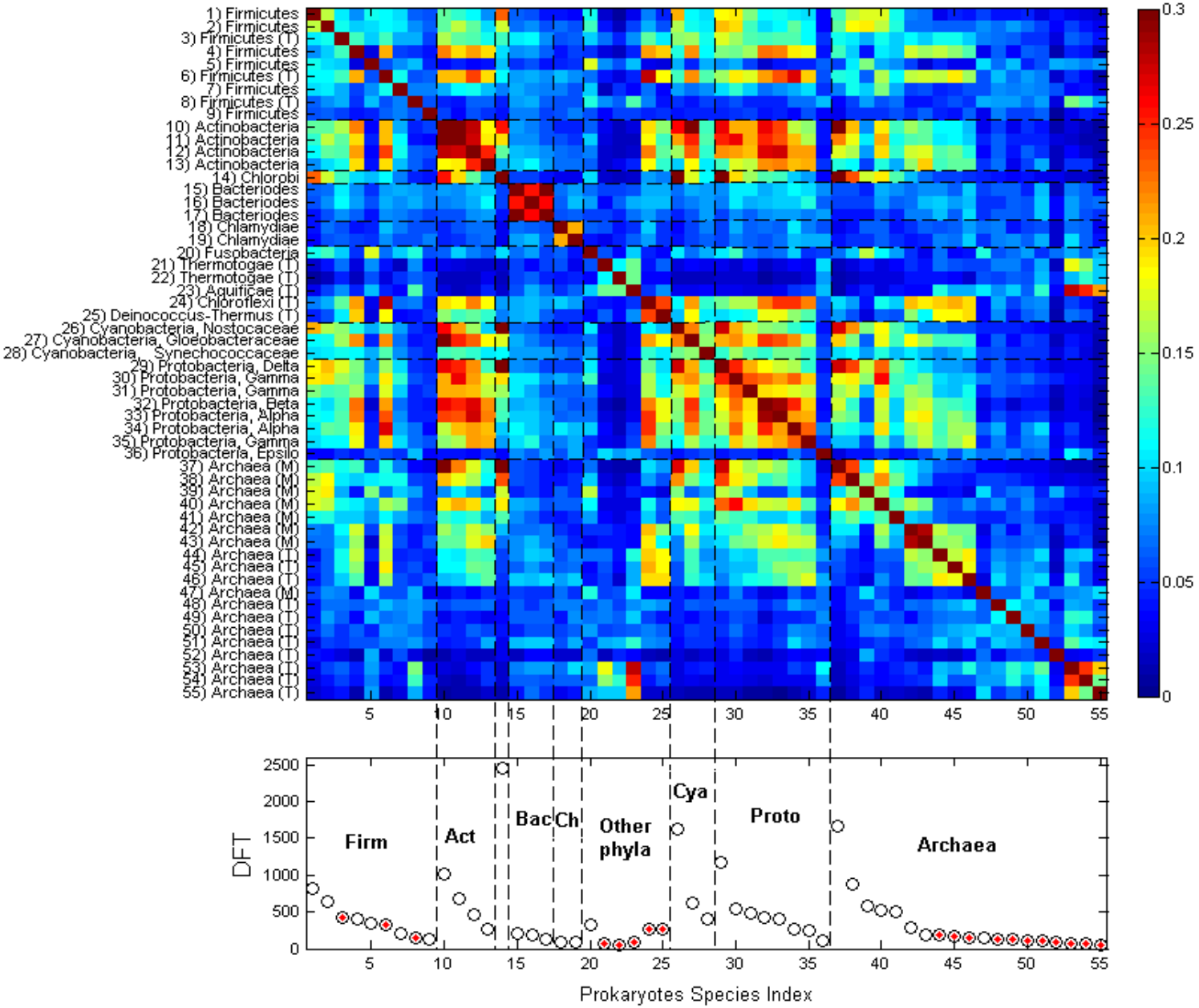
Species	DFT	Clade	Species	Ne x u	Clade
Zea mays	1037	P	Cryptococcus neoformans	0.02526	F
Cryptococcus neoformans	1050	F	Saccharomyces cerevisiae	0.02294	F
Saccharomyces cerevisiae	1077	F	Dictyostelium discoideum	0.01825	PRT
Neurospora crassa	1780	F	Neurospora crassa	0.0113	F
Oryza sativa	1846	P	Toxoplasma gondii	0.00688	PRT
Arabidopsis thaliana	2262	P	Leishmania major	0.00521	PRT
Leishmania major	2319	PRT	Drosophila melanogaster	0.00374	IV
Toxoplasma gondii	2840	PRT	Zea mays	0.0033	P
Dictyostelium discoideum	2990	PRT	Caenorhabditis elegans	0.00328	IV
Anopheles gambiae	3518	IV	Arabidopsis thaliana	0.00323	P
Caenorhabditis elegans	3722	IV	Ciona intestinalis	0.00305	IV
Fugu rubripes	3746	IV	Anopheles gambiae	0.00298	IV
Ciona intestinalis	4019	IV	Strongylocentrotus purpuratus	0.0023	IV
Drosophila melanogaster	4146	IV	Fugu rubripes	0.00101	IV
Mus musculus	4873	V	Oryza sativa	0.00077	P
Homo sapiens	5076	V	Homo sapiens	0.00031	V
Strongylocentrotus purpuratus	5477	IV	Mus musculus	0.00027	V



DFT Correlation among Eukaryotes



DFT Correlation among Prokaryotes

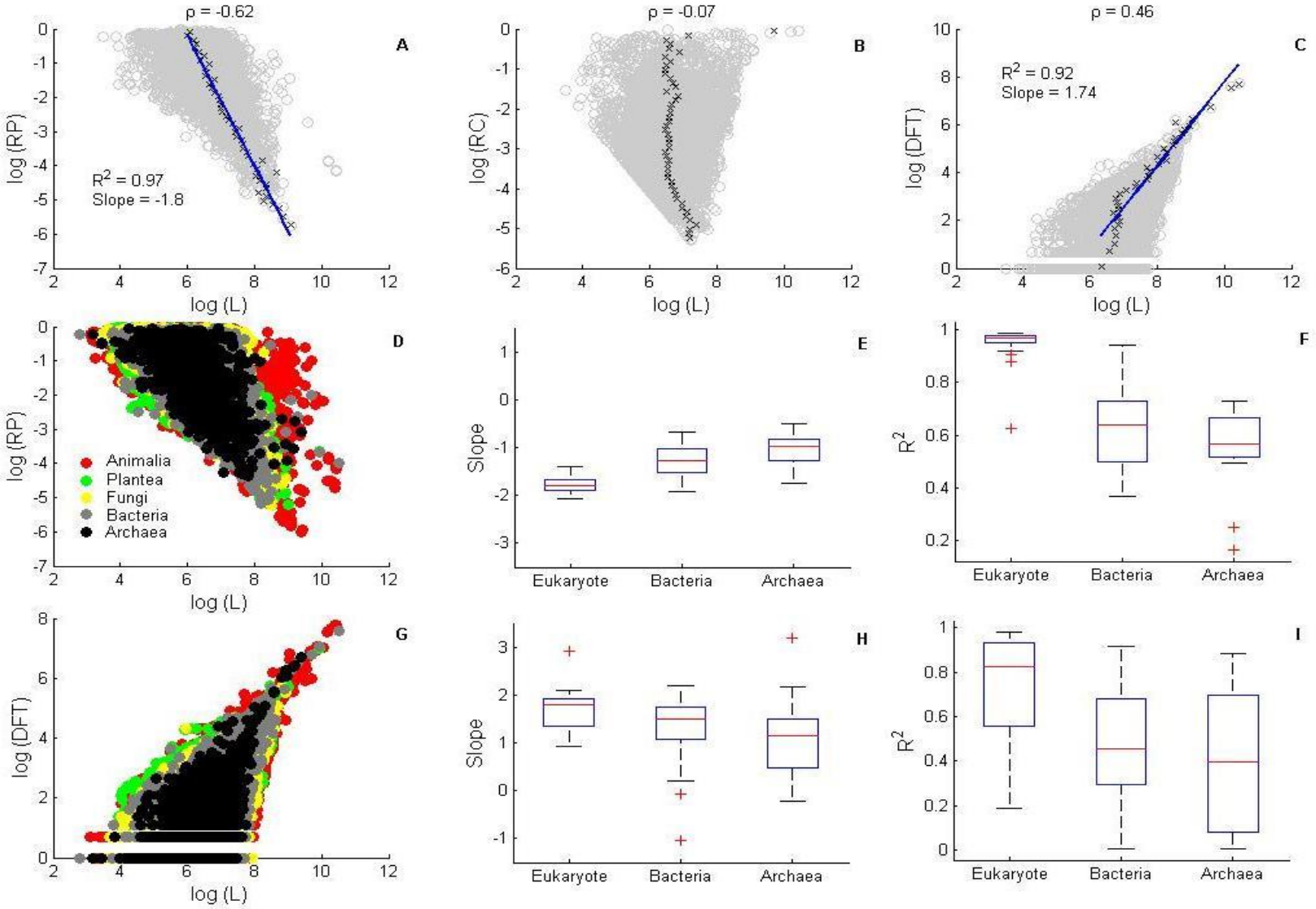


Universal dependence of RP and DFT on protein length

Human RP, RC, DFT *vs* L

RP *vs* L for all species

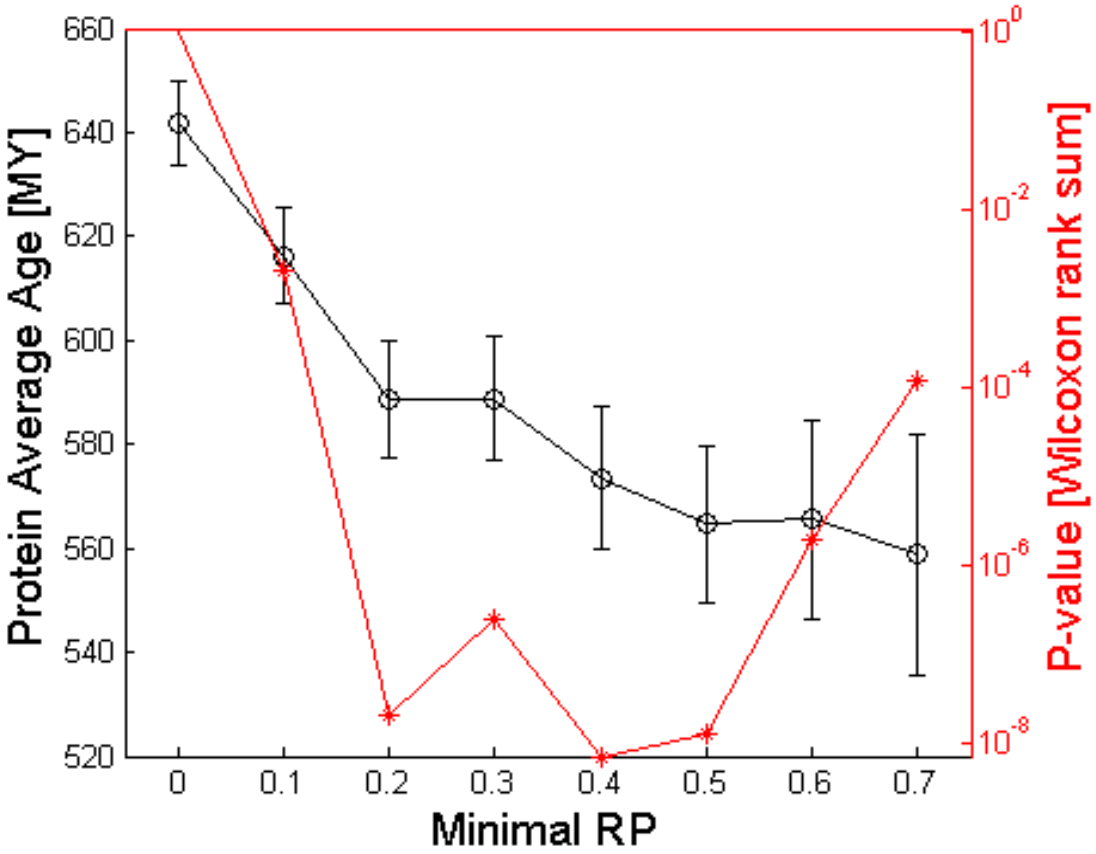
DFT *vs* L for all species



possiblle explanation: large L=older proteins, exhibit large DFT indicating growth through CO material, and decrease of RP due to mutations accumulated during evolutionary history

High RP is associated with young proteins

Capra JA et al (2012). **ProteinHistorian: Tools for the Comparative Analysis of Eukaryote Protein Origin**. *PLoS Comp Biol*



- The average age of proteins (black) is shown *vs* RP.
- The statistical significance of the difference between the age distribution for a given RP threshold and the age distribution of the entire CO set was estimated according to Wilcoxon rank-sum test (red).

Conclusions

- Large scale study of Compositional Order is facilitated by employing FTs and the measures RC, RP and DFT.
- DFT serves as an effective measure of macroevolution (a “stamp” on the proteome).
- Macroevolution may be associated with increase of CO leading to genomic innovation: new raw material which is fast evolving and facilitates adaptation and acquisition of functions.
- Compositional order, as accounted for by measures of regularity, periodicity and richness, has universal characteristics, yet displays species-specific contents.

Addendum:
Does cancer evolution involve CO growth?

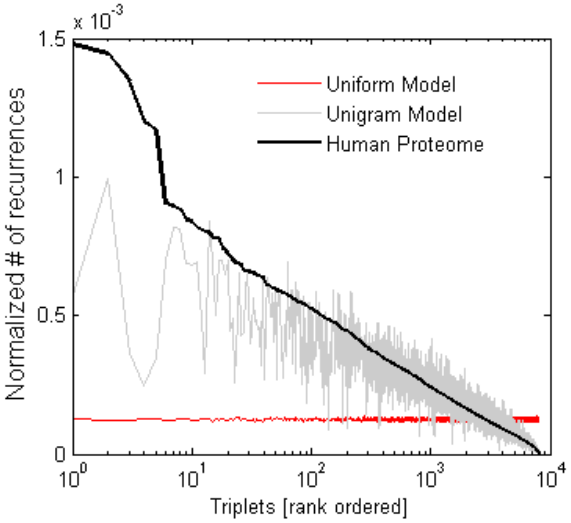
On-going research in collaboration with Tami Geiger (TAU), based on proteomic data of 10 breast cancer patients currently studied in her lab.

Proteomic data contain peptides with 10-30 amino-acids, hence one needs new suitable definition of CO measures.

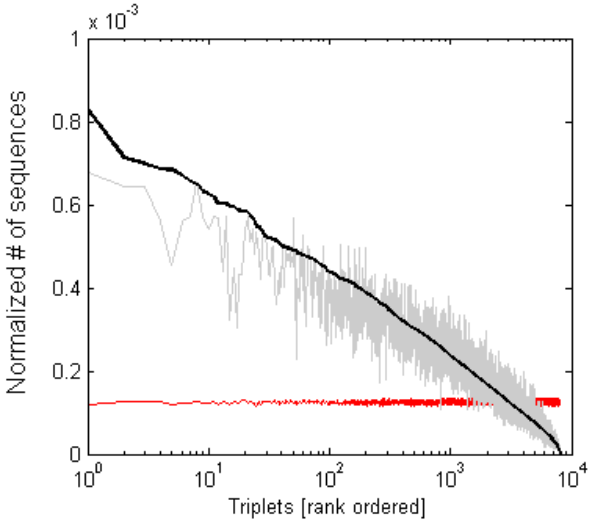
Using 8000 Triplets as the preferred vocabulary, we find that neither # of triplet recurrences, nor # of sequences (peptides) on which a triplet occurs, are suitable but

CR= CO ratio = # of recurrences / # of sequences

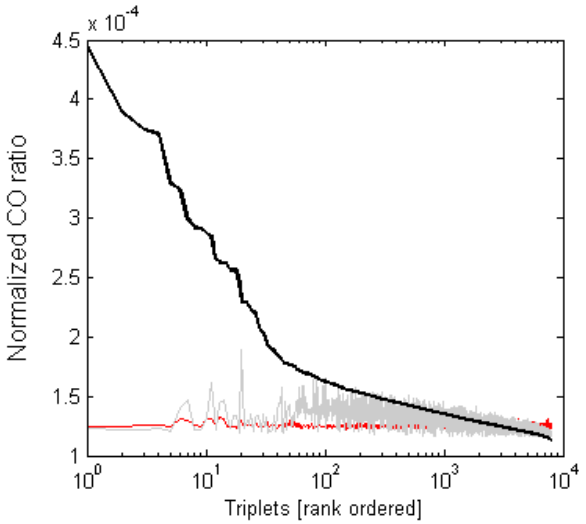
works well: data are significantly different from random models.



# of recurrences



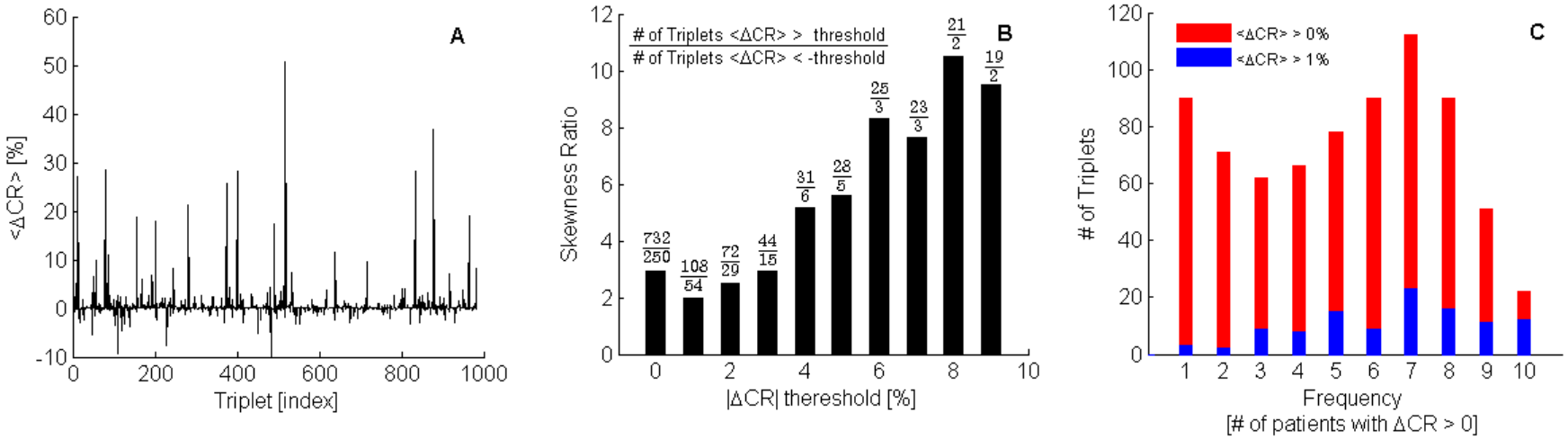
# of sequences



CR

Define, for given triplet and matched tumor to normal data, ΔCR=CR(tumor)-CR(normal).

Preliminary results indicate usefulness of this approach



A: ΔCR in % per triplet. B: Skewness ratio for different thresholds. C:Histogram of frequencies, i.e. # triplets as function of # of patients for which their ΔCR increases (limited to triplets whose <ΔCR>>0).

Large scale testing will allow us to conclude if

- CO triplets can serve as predictive features
- CO triplets increase with evolutionary stage of the tumor
- CO triplets are characteristic of the cancer type

Thank you !