“Baby talk” of genomic DNA. Fundamental role of repetitions.

Edward N. Trifonov
University of Haifa, Israel

Prague, Brno
2013

Baby talk words, perfect repeats

(Russian, if not specified)

Mama
Papa
Baba (grandma)
Pipi
Caca
Sisi (breast)
Bobo (pain)
Baibai (good night)
Tiatia (father)
Niania (nanny)
Ham-ham (eat, Vietnamese)
Ai-ai-ai (mishap)
Ne-ne-ne (no, Czech)
Wong-wong (drink, Vietnamese)

Baby talk words, perfect repeats

Lala (doll, baby)
Kuku (from hiding)
Diadia (man)
Oi-oi-oi (mishap)
Ni-ni-ni (strictly no)
Niam-niam (eat)
Dai-dai-dai (give me)

Sound imitations, mostly babies

Av-av (dog)
Bi-bi (car)
Cococo (chicken)
Kva-kva (frog)
Tik-tak (clock)
Din’din’ (ringbell)
Ga-ga-ga (geese)
Kria-kria (duck)
Tuk-tuk-tuk (knocking)
Kap-kap-kap (rain)
Chmok-chmok (kisses)
Top-top-top (walk)
Skirly-skirly (wooden leg)

Rooster (adults):

Ku ka re ku
Ki ri ko ko (Czech, French)
Cock-a-doodle-doo (English)

Mooring steamer to a pier

Sound imitations from “Adventures of Tom Sawyer” by Mark Twain:

He was boat and captain and engine-bells combined, so he had to imagine himself standing on his own hurricane-deck giving the orders and executing them:
"Stop her, sir! Ting-a-ling-ling!" The headway ran almost out, and he drew up slowly toward the sidewalk.
"Ship up to back! Ting-a-ling-ling!" His arms straightened and stiffened down his sides.
"Set her back on the stabboard! Ting-a-ling-ling! Chow! ch-chow-wow! Chow!"
His right hand, mean-time, describing stately circles—for it was representing a forty-foot wheel.
"Let her go back on the labboard! Ting-a-ling-ling! Chow-ch-chow-chow!"
The left hand began to describe circles.
"Stop the stabboard! Ting-a-ling-ling! Stop the labboard! Come ahead on the stabboard! Stop her! Let your outside turn over slow! Ting-a-ling-ling! Chow-ow-ow!
Get out that head-line! *lively* now! Come—out with your spring-line—what're you about there! Take a turn round that stump with the bight of it! Stand by that stage, now—let her go! Done with the engines, sir! Ting-a-ling-ling! SH'T! S'H'T! SH'T!"
(trying the gauge-cocks).

Adult forms, perfect repeats:

O-o (warning)
Bebe
Da-da (come in)
Ja-ja (yes, German)
Ku-ku (crazy)
Ga-ga (crazy, English)
Hahaha
Nununu (warning to babies)
Tuktuk (Cambodia, moto-rickshaw)
Tamtam (drum)
Tak-tak (all right)
Ks-ks-ks (calling cat)
Nuka-nuka (go ahead)
Chachacha
Leat-leat (slowly, Hebrew)
Tipa-tipa (little bit, Hebrew)
Tilki-tilki (barely fit, Ukrainian)
Trochi-trochi (little bit, Ukrainian)
Rock-rock-rock (Kenya, lullaby)
Langsam-langsam (slowly, Yiddish)

Adult forms, perfect repeats:

E-e (warning)
Ohoho (that much)
Mimimi (sweaty, cuty)
Bumbum (ignorant)
Lalala (empty talk)
Tsatsa (girl showing up)
Vot-vot (in a moment)
Idu-idu (coming)
Kto-kto? (who)
Gde-gde? (where)
Vas'-vas' (friends)
Tiny-tiny
Jele-jele (barely)
Kuda-kuda? (where)
Tolko-tolko (barely fit)
Chut'-chut' (little bit)
Hei-hei-hei (warning)
Chevo-chevo? (what)
Tsip-tsip-tsip (calling chicken)
Skolko-skolko? (how much)
Kak eto, kak eto? (why all of a sudden)

Mutated, imperfect repeats, babies and adults:

Mamy (mother, English)
Baby
Bibika (car)
Mamaya (fruit, Brazil)
Papaya (similar fruit, Brazil)
O-la-la (surprize, French)
Coocook
To-to-je (Aliska, co to je, Czech)
Ta-ra-ram (mess)
Balalaika
Tarataika (type of a cart)
Yin'-yan' (Chinese)
Siusiukat' (imitate baby-talk)
Tsap-tsarap (catch, about cats)
Villi-nilli (against will, Latin)
Meli, Emelia (talking nonsense)
Olgoi-horhoi (Mongolian, ferrytale creature)
Volens-nolens (against will, Latin)
Naziuziukalsa (drunk)
Futy-nuty, lapti gnuty (mishap)

Mutated, imperfect repeats, babies and adults:

Nu-i-nu (surprized)
Kukushka (coocook)
Coca-cola
Tra-ta-ta (thunder)
Futy-nuty (mishap)
Tiap-liap (lousy work)
Trali-vali (menstruation)
Dura duroi (stupid, her)
Figli-migli (flirt)
Shito-kryto (everything is fine)
Tram-tararam (mess)
Durak durakom (stupid, he)
Boogie-woogie
Trach-tararach (thunder)
Postolku-poskolku (as soon as)
Baiu-baiushki-baiu (lullaby)
Tiutelka v tiutelku (just exactly fit)

Counting rhymes for seek and hide game

Ene bene rech
Kenter menter zhech
Ene bene raba
Kenter menter zhaba

Eniki beniki
Eli vareniki
Eniki beniki klotz

Ine mine
Minke tinke
Fade rude
Rolke tolke
Wigel wagel weg (German)

Martin Luther King, 1968:

"Yes, if you want to
say that **I was a drum major**,
say that **I was a drum major** for justice.
Say that **I was a drum major** for peace.
I was a drum major for righteousness."

Criticized misquote:

"I was a drum major for justice,
for piece,
for righteousness."

Human languages, quite likely, originated from simple repetitive words,

continued with their mutated forms,

and even today the languages operate with simple repeats, mutated forms, and longer tandem or dispersed repeats (refrains).

EXACTLY THE SAME CAN BE SAID ABOUT BIOLOGICAL SEQUENCES (nucleic acids and proteins)

All 15-mers of human genome (sorted)

1	1198780	TTTTTTTTTTTTTTT	T_n
2	1190667	AAAAAAAAAAAAAAA	A_n
3	366285	TGTGTGTGTGTGTGT	TG_n
4	362623	ACACACACACACACA	AC_n
5	348215	GTGTGTGTGTGTGTG	GT_n
6	344421	CACACACACACACAC	CA_n
7	223424	GCTGGGATTACAGGC	Alu
8	223011	GCCTGTAATCCCAGC	Alu
9	222894	TATATATATATATAT	TA_n
10	222730	ATATATATATATATA	AT_n
11-67			Alu
68	169033	TTTTTTTTTTTTTTG	T_n
69-72			Alu
73	167889	CAAAAAAAAAAAAAA	A_n
74	167361	CTAAAAATACAAAAA	Alu
75	150349	CTTTTTTTTTTTTTT	T_n
76	149748	AAAAAAAAAAAAAAG	A_n
77-82			Alu

Three known pathologically expanding ("aggressive") classes of triplets

GCU (GCU, CUG, UGC, AGC, GCA, CAG) ,

GCC (GCC, CCG, CGC, GGC, GCG, CGG) and

GAA (AAG, AGA, GAA, CTT, TTC, TCT).

They cause neurodegenerative diseases and chromosome fragility

EVOLUTION OF THE TRIPLET CODE

E. N. Trifonov, December 2007, Chart 101

Consensus temporal order of amino acids:

```
                        UCX           CUX      CGX AGY UGX AGR          UUY UAX
     Gly Ala Asp Val Ser Pro Glu Leu Thr Arg Ser TRM Arg Ile Gln Leu TRM Asn Lys His Phe Cys Met Tyr Trp Sec Pyl

 1   GGC-GCC  .    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .   .|.    .    .    .    .    .    .    .
 2    |    |  GAC-GUC   .    .    .    .    .    .    .    .    .    .    .    .    .    .   .|.    .    .    .    .    .    .    .
 3   GGA--|---|---|--UCC  .    .    .    .    .    .    .    .    .    .    .    .    .    .|.    .    .    .    .    .    .    .
 4   GGG--|---|---|---|--CCC  .    .    .    .    .    .    .    .    .    .    .    .    .|.    .    .    .    .    .    .    .
 5    |    | (gag)-|---|---|--GAG-CUC  .    .    .    .    .    .    .    .    .    .    .   .|.    .    .    .    .    .    .    .
 6   GGU--|---|---|---|---|---|---|--ACC  .    .    .    .    .    .    .    .    .    .    .|.    .    .    .    .    .    .    .
 7    .  GCG--|---|---|---|---|---|---|--CGC  .    .    .    .    .    .    .    .    .    .|.    .    .    .    .    .    .    .
 8    .  GCU--|---|---|---|---|---|---|---|--AGC  .    .    .    .    .    .    .    .    .|.    .    .    .    .    .    .    .
 9    .  GCA--|---|---|---|---|---|---|---|---|--ugc  .    .    .    .    .    .    .    .|.    .   UGC   .    .    .    .    .
10    .    .    |    |    |  CCG--|---|---|--CGG  |    |    .    .    .    .    .    .    .|.    .    |    .    .    .    .    .
11    .    .    |    |    |  CCU--|---|---|---|---|---|--AGG  .    .    .    .    .    .  .|.    .    |    .    .    .    .    .
12    .    .    |    |    |  CCA--|---|---|---|---|--ugg  |    .    .    .    .    .    . .|.    .    |    .    .   UGG   .    .
13    .    .    |    |  UCG------|---|---|--CGA  |    |    |    .    .    .    .    .    .|.    .    |    .    .    .    .    .
14    .    .    |    |  UCU------|---|---|---|---|---|--AGA  .    .    .    .    .    .   .|.    .    |    .    .    .    .    .
15    .    .    |    |  UCA------|---|---|---|---|--UGA  .    .    .    .    .    .    .  .|.    .    |    .    .    .   UGA   .
16    .    .    |    |    .    .    |    |  ACG-CGU  |    |    .    .    .    .    .    .   .|.    .    |    .    .    .    .    .
17    .    .    |    |    .    .    |    |  ACU-----AGU  |    .    .    .    .    .    .    .|.    .    |    .    .    .    .    .
18    .    .    |    |    .    .    |    |  ACA---------ugu  .    .    .    .    .    .    .|.    .   UGU   .    .    .    .    .
19    .    .  GAU--|-----------|---|----------------------AUC  .    .    .    .    .    . .|.    .    .    .    .    .    .    .
20    .    .    .  GUG----------|---|----------------------|--cac  .    .    .    .    . .|CAC   .    .    .    .    .    .    .
21    .    .    .    |    .    .    |  CUG--------------------|--CAG  .    .    .    .    .|  |   .    .    .    .    .    .    .
22    .    .    .    |    .    .    |    |    .    .    .    .    .  aug-cau  .    .    .    .|CAU   .    .   AUG   .    .    .    .
23    .    .    .    |    .    .  GAA--|-----------------------|---|--uuc  .    .    .    .|.   UUC   .    .    .    .    .    .
24    .    .    .  GUA-------------|-----------------------|---|---|--uac  .    .    .|.    |    .    .   UAC   .    .    .
25    .    .    .    |    .    .    .  CUA---------------------|---|---|--UAG  .    .   .|.    |    .    .    |    .    .   UAG
26    .    .    .  GUU-------------|-----------------------|---|---|---|--AAC  .      .|.    |    .    .    |    .    .    .
27    .    .    .    .    .    .    .  CUU---------------------|---|---|---|---|--AAG|.    |    .    .    |    .    .    .
28    .    .    .    .    .    .    .    .    .    .    .    .    .    |  CAA-UUG  |    |    |  |.    |    .    .    |    .    .    .
29    .    .    .    .    .    .    .    .    .    .    .    .    .  AUA------|--uau  |    |  |.    |    .    .   UAU   .    .    .
30    .    .    .    .    .    .    .    .    .    .    .    .    .  AUU------|---|--AAU  |  |.    |    .    .    .    .    .    .
31    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .  UUA-UAA      |  |.    |    .    .    .    .    .    .
32    .    .    .    .    .    .    .    .    .    .    .    .    .    .    .  uuu---------AAA|.   UUU   .    .    .    .    .    .
```

CONSECUTIVE ASSIGNMENT OF 64 TRIPLETS **CODON CAPTURE**

```
aa "age":
 17   17   16   16   15   14   13   13   12   11               10    9             8    7    6    5    4    3    2    1
```

"... if **variations** useful to any organic being ever do occur, assuredly individuals thus characterized will have the best chance of being preserved in the struggle for life;
and from the strong principle of inheritance, these will tend to **produce offspring similarly characterized**"

Charles Darwin, Origin of Species (1859)

Rephrasing (ET):

Individuals with useful **variations** will **self-reproduce**

self-reproduction and variation

Any system capable of
replication and mutation
is alive (Oparin 1961).

self-reproduction and variation

not Life yet

(self-reproduction only)

Life

(self-reproduction and variations)

```
    Gly   Ala| Val   Asp   Ser   Pro ...
             |
1   GGC--GCC|
2    |     |    GUC--GAC
3   GGA---|----|----|---UCC
4   GGG---|----|----|----|---CCC
.
.
```

Life is self-reproduction with variations

From vocabulary of 123 known definitions of life the following groups of meanings are revealed

LIFE	123
living	47
alive	10
being	6
biological	5
other related words	8
Sum	**199**

SYSTEM	43
systems	22
organization	14
organism	14
order	6
organisms	6
network	5
organized	5
other related words	40
Sum	**155**

MATTER	25
organic	11
materials	10
molecules	6
other related words	36
Sum	**88**

CHEMICAL	17
process	15
metabolism	14
processes	8
reactions	5
other related words	26
Sum	**85**

COMPLEXITY	13
information	8
complex	7
other related words	46
Sum	**74**

REPRODUCTION	10
reproduce	8
replication	7
self-reproduction	5
other related words	33
Sum	**63**

EVOLUTION	10
evolve	7
change	6
mutation	5
other related words	20
Sum	**48**

ENVIRONMENT	20
external	6
other related words	15
Sum	**41**

ENERGY	18
force	5
other related words	17
Sum	**40**

ABILITY	12
able	11
capable	11
capacity	5
other related words	1
Sum	**40**

Life *(definiendum)*

Definientia:

System

Matter

Chemical

Complexity

Reproduction

Evolution

Environment

Energy

Ability

These appear to be both necessary and sufficient for the definition of life

We, thus, come again to the same definition:

Life is self-reproduction with variations

Human Genome Composition

Protein-coding and RNA-coding	3%
Non-coding DNA	97%
of which	
Simple sequence repeats	3% (underestimate)
Transposable elements	45%

"repeat sequences account for at least 50%

and, probably, much more"

From E. S. Lander *et al.* Initial sequencing and analysis of the human genome, Nature 409, 860-921, 2001

Aggressive amino acids encoded by expanding triplets

Amino acid	Triplets
L (leucine)	**CTG CTT**
A (alanine)	**GCT GCA GCC GCG**
G (glycine)	**GGC**
P (proline)	**CCG**
S (serine)	**AGC TCT**
E (glutamate)	**GAA**
R (arginine)	**CGG CGC AGA**
Q (glutamine)	**CAG**
K (lysine)	**AAG**
F (phenylalanine)	**UUC**
C (cysteine)	**UGC**

Majority of homopeptides are built from aggressive amino acids

human tripeptides 1st exons	Score (tripept.)	eukar. (Faux et al.)	prokar. (Faux et al.)
1. L3	**4552**	**1446**	**70(5)**
2. A3	**4046**	**5465(3)**	**251(3)**
3. G3	**2972**	**5002(5)**	**310(2)**
4. P3	**2258**	**4157(7)**	**217(4)**
5. S3	**1981**	**5424(4)**	**378(1)**
6. E3	**1630**	**4334(6)**	**67(6)**
7. R3	**1145**	**462**	**60(8)**
8. Q3	**802**	**8022(1)**	**52(9)**
9. K3	**535**	**1920(9)**	**25**
10. V3	414	94	9
11. H3	273	1049	32
12. D3	269	1554	34
13. T3	267	2492(8)	63(7)
14. I3	109	34	3
15. F3	**103**	**175**	**1**
16. C3	**92**	**38**	**0**
17. N3	79	6962(2)	31
18. M3	34	19	0
19. Y3	32	39	4
20. W3	14	3	0
	92%	**75%**	**89%**

(Z. Koren, 2011)

Could it be that protein sequences, actually, are ALL originally made from the aggressive repetitions?

And we don't see all the original repeats just because they have extensively mutated.

If this view is correct, then we should see in mRNA sequences

1. Ideal repeats of some codons

2. The codons “sandwiched” between two identical codons should be their point mutation derivatives

3. Those codons which are more often in tandem repeats should be also of higher usage in non-repeats

We, thus, undertook analysis

of the largest non-reduntant database of mRNAs available,

of total ~5 000 000 000 codons,

from eukaryotes, prokaryotes, viruses, organelles together

Z. Frenkel, E. Trifonov, JBSD, 30, 201-210 (2012)

22.5 min

Sorted occurrence of the triplet repeats for different groups ("aggressive" triplets)

	group of codons	Occurrence
1	**GCC, CCG, CGC, GGC, GCG, CGC**	**1 784302**
2	**GCA, CAG, AGC, UGC, GCU, CUG**	**1 436660**
3	**GAA, AAG, AGA, UUC, UCU, CUU**	**1 131214**
4	AAU, AUA, uaa, AUU, UUA, UAU	932105 (1 118526)
5	AUC, UCA, CAU, GAU, AUG, uga	735397 (882476)
6	ACC, CCA, CAC, GGU, GUG, UGG	726443
7	AGG, GGA, GAG, CCU, CUC, UCC	706484
8	AAC, ACA, CAA, GUU, UUG, UGU	694387
9	ACG, CGA, GAC, CGU, GUC, UCG	533888
10	ACU, CUA, UAC, AGU, GUA, uag	152747 (183296)

1. Tandem repeats of all 61 different codons are observed, strongest for aggressive groups, as expected

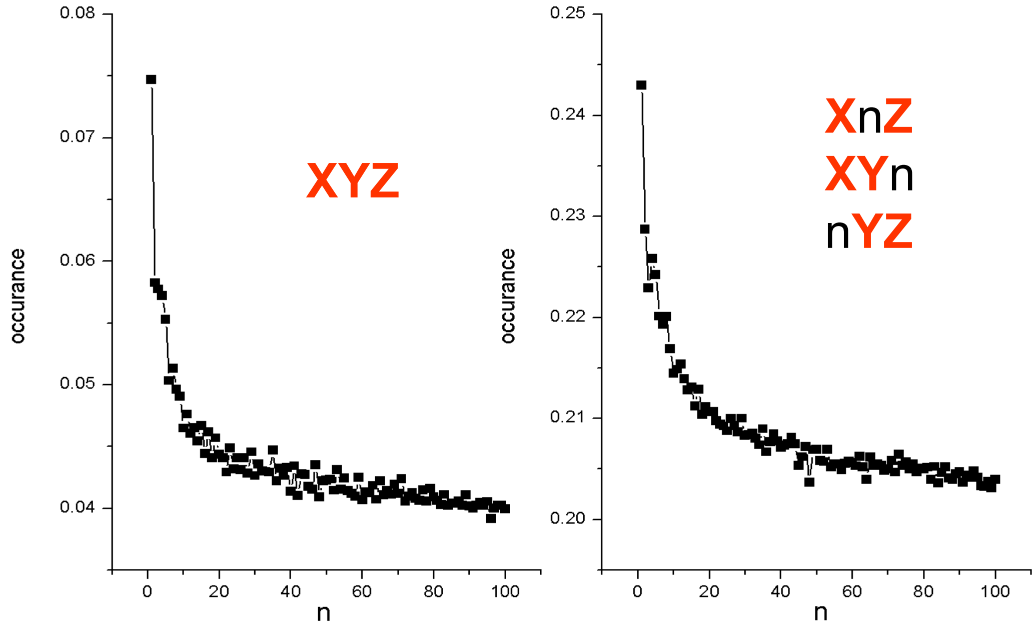
2. Middle codons abc in "sandwiches" GCUabcGCU (total 3 168 933) are most often first derivatives of GCU

GCU	243706	
GG**U**	125946	
GA**U**	115500	
GAA	114278	the topmost in codon usage
GU**U**	102550	
GCA	95493	
GCC	92153	
AUU	89648	
UUU	87861	
AAA	84194	next topmost in codon usage
UUA	80660	
GGA	74934	
GGC	71770	
	…	

This also holds for most of other codons

„Thick“ sandwiches

XYZ$abc_1abc_2...abc_n$XYZ



A B

Occurrence of the triplet **XYZ** (A) and its first derivatives (B) in the middle sequence $abc_1abc_2...abc_n$

2. The first derivatives between the identical codons in mRNA keep memory of initial tandem repetition of the codons

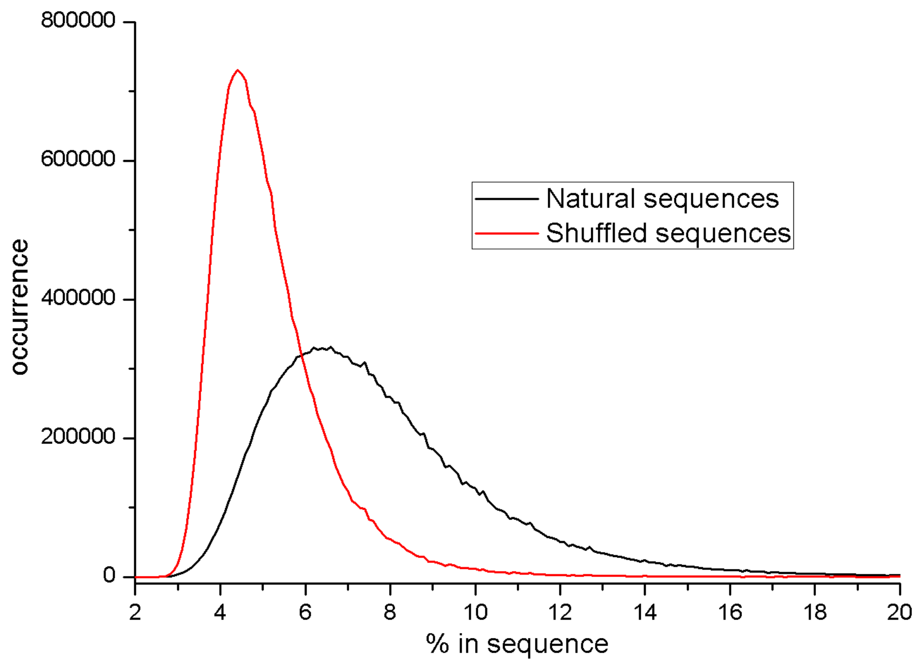
The sequences like

XYZ nnn nnn nnn nnn XYZ nnn nnn nnn nnn nnn nnn XYZ

are likely descendants of

XYZ XYZ XYZ XYZ XYZ XYZ XYZ XYZ…

Enrichment of mRNA sequences by one or another dominant codon



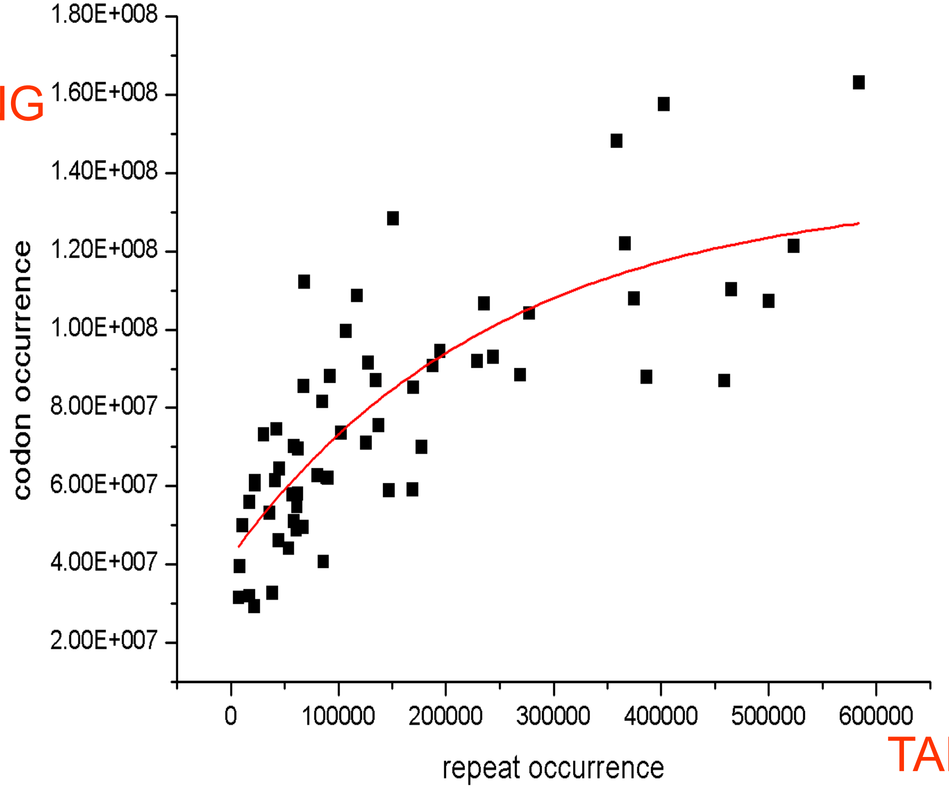
ATG GCT CTA ACC AAA GAA GAT ATT TTA AAC GCA ATT GCT **GAA** ATG CCA GTA ATG
GAC CTT GTT GAG CTT ATC **GAA** GCT GCA **GAA** **GAA** AAA TTC GGT GTA ACA **GCT** ACT
GCT **GCT** GTT **GCT** GCC **GCT** **GCT** CCT **GCT** **GCT** GGC GGT GAA **GCT** **GCT** GCA GAA CAA
ACT GAA TTT GAT GTT GTT TTG ACA TCT TTC GGT GGT AAC AAA GTT **GCT** GTA ATC
AAA GCG GTA CGT GGC GCA ACT GGT CTT GGC TTG AAA **GAA** GCT AAA **GAA** GTA GTT
GAA GCT GCA CCG AAA GCG ATT AAA **GAA** GGC GTT GCT AAA **GAA** **GAA** GCT **GAA** **GAA**
CTT AAG AAG ACG CTT GAA GAA GCT GGC GCT GAA GTT GAG CTT AAG

GAA and **GCT** “bricks” in mRNA of
ribosomal protein L12 of *Ps. Atlantica*

Frequent triplets make clusters, remnants of original ideal repeats

3. The more frequently the codon appears in tandem the more frequent it is also in non-repeating regions of mRNA

NON-REPEATING REGIONS



TANDEMS

Ala	**GCC**	**110**	**465**	Arg	**CGC**	**70**	**177**	Arg	**AGA**	**55**	**62**
	GCA	94	195		CGU	46	45		AGG	29	22
	GCU	93	245		CGG	41	86				
	GCG	88	386		CGA	33	39				
Asn	**AAU**	**121**	**523**	Asp	**GAU**	**148**	**359**	Cys	**UGC**	**31.9**	**18**
	AAC	85	170		GAC	107	236		UGU	31.5	7
Gln	CAA	88	269	Glu	**GAA**	**163**	**584**	**Gly**	**GGC**	**107**	**500**
	CAG	87	459		GAG	122	367		GGU	92	229
									GGA	87	135
									GGG	56	17
His	**CAU**	**58**	**62**	Ile	**AUU**	**128**	**151**	Leu	**UUA**	**91**	**127**
	CAC	49	61		AUC	100	107		UUG	73	30
					AUA	70	63				
Leu	**CUG**	**108**	**375**	Lys	**AAA**	**158**	**403**	Met	AUG	109	117
	CUU	75	43		AAG	104	277				
	CUC	70	59								
	CUA	40	8								
Phe	UUU	112	68	Pro	CCA	62	89	Ser	UCU	63	81
	UUC	82	85		CCG	59	169		UCA	62	90
					CCU	58	59		UCC	50	67
					CCC	50	11		UCG	44	54
Ser	**AGC**	**59**	**147**	Thr	**ACC**	**76**	**138**	Trp	UGG	60	22
	AGU	53	36		ACA	71	126				
					ACU	65	45				
					ACG	51	59				
Tyr	**UAU**	**86**	**68**	Val	**GUG**	**91**	**187**				
	UAC	61	41		GUU	88	92				
					GUC	74	103				
					GUA	61	23				

1st columns - codons
(millions)

2nd columns - repeats
(thousands)

In 17 of 21 codon repertoires
the most frequent codon
is also the most repetitive

This result came as a surprize,
considering zelions of factors
known to influence the codon usage

More frequent codons keep memory of
tandem repetition of these codons
in the past

The triplet expansion of codons
is the major single factor
shaping the codon usage

According to the Theory of Early Molecular Evolution based on the Evolutionary Chart of Codons

the very first genes have been repeats

…GGC GGC GGC GGC GGC GGC…

and complementary

…GCC GCC GCC GCC GCC GCC…

encoding Gly_n and Ala_n, respectively

Thus, life started with the replication (and expansion)
and subsequent mutations
of tandemly repeating triplets GGC and GCC.
(self-reproduction with variation)

Life continued then to spontaneously emerge
within the primitive early genomes and further on,
in form of replication and expansion
and subsequent mutations
of other tandem repeats as well
(self-reproduction with variation)

Life never stopped emerging

"… if (and oh what a big if) we could conceive in some warm little pond with all sort of ammonia and phosphoric salts, - light, heat, electricity etc., present,
that a protein compound was chemically formed,
ready to undergo still more complex changes,
at the present day such matter would be
instantly devoured, or absorbed,
which would not have been the case
before living creatures were formed." (Darwin 1871)

With the new view on genome origin and evolution the emerging life is not consumed by the earlier life, but rather protected by the environment within the cell.

The tandem repeats have been considered as a class of
"selfish DNA" (Orgel and Crick, 1980; Doolittle and Sapienza, 1980).

They are, actually, more than just parasites tolerated by genome.
They are even more than
building material for the genome (Ohno, Junk DNA, 1972).

The tandem repeats represent constantly emerging life,
and genomes are products of their everlasting domestication.

Genomes are built by the expansion and
mutational domestication of the tandem repeats

Genomes ARE the repeats
(some already unrecognizable)

Painful symbiosis of repeats with genomes

For genomes

accepted repeats are useful.

new repeats are dangerous.

For repeats

genomes are natural habitats.

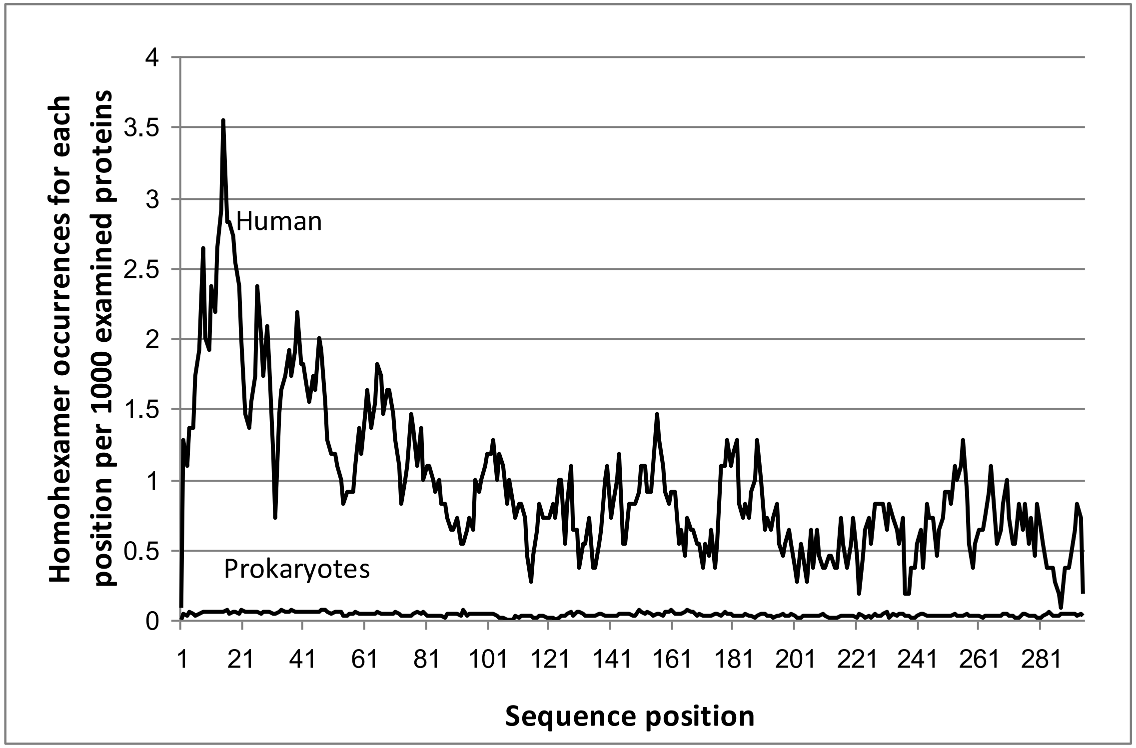
initiation is at high risk

PREDICTION:

GENOMES SHOULD BE EQUIPPED BY

DEFENSE SYSTEMS

AGAINST CONSTANTLY EMERGING REPEATS



The amino acid repeats in prokaryotes
are far less frequent compared to eukaryotes.

Defense in prokaryotes:

Brutal negative selection,
death of individuals contracting the repeats

Defense in eukaryotes:

Expulsion of the repeats into introns and intergenic sequences?
(Alternative splicing as an intermediate stage)

Possible defense devices:

Prevention of slippage. Nucleosomes.
Excision of slippage loops.
Methylation of repeats.
Sequence-specific nucleases
…..

The simplest life forms – simple tandem repeats – represent a whole class of pathological agents, not considered as such up to now.

Genomes evolve under constant attacks by various repeats.

Apparently, most of the attacks are normally stopped by the defense system.

Some of the new expansions or insertions are accomodated by the genomes.

Some are neither stopped, nor accomodated, causing disaster.

A DIFFERENT VIEW ON CANCER, EXPANSION DISEASES
AND DISEASES WITH UNKNOWN CAUSATIVE AGENT:

The repeats in the diseases are not symptoms.
They are cause of the diseases.

THANKS TO

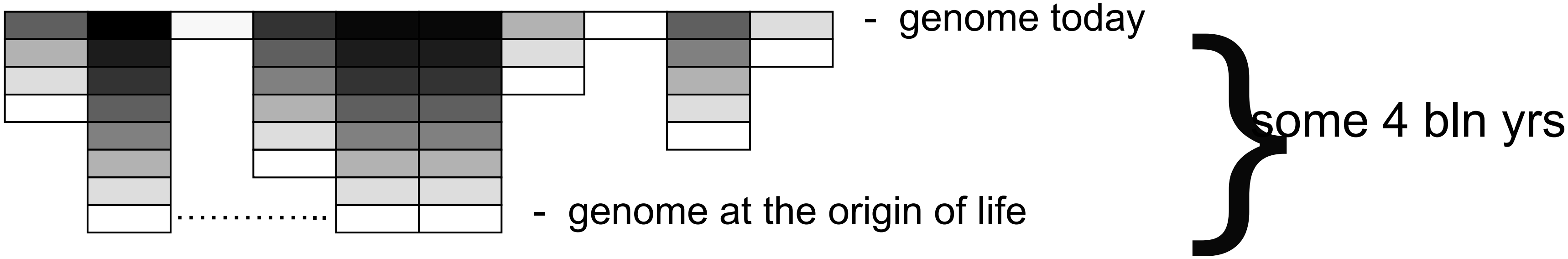
ZOHAR **KOREN**,

ZAKHARIA **FRENKEL**, **Haifa**

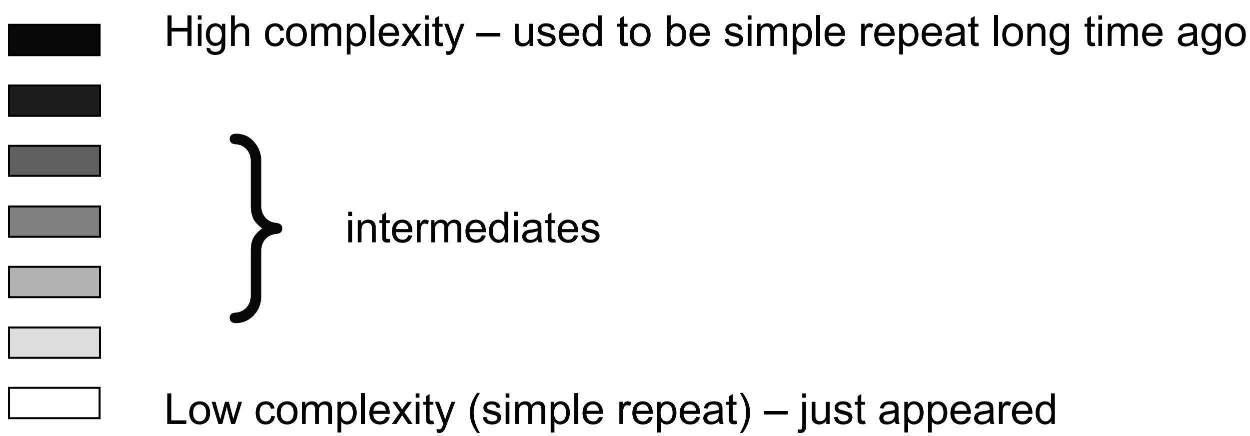
ALEXANDRA **RAPOPORT,**

THOMAS **BETTECKEN** **München**

MISA **ZEMKOVA** **Prague**



Genomes are all built from simple repeats.
Just many of them already unrecognizable



GAA GAA GAA GAA GAA GAA GAA GAA GAA GAA GAA GAA GAA

GAA GAA CAA GAA GGA GAU GAA GAA UAC GAG GAA GAA AAA

CAA GAA CAA GGA GGA AAU GAA GCA UAC GAG GAA GGA AAU

CAG GUA CAG GGU GGA AAU GAA GCC UUC GGG GAA CGG ACU

CAG AUA CCG GGU GGG AAU UAC GCC UUC UGG AAA CGG ACU

CCG AUA CCG UGU GGG ACU UAC UCC UUC UGG AAC CGG ACU

CCG AUC CCG UGU UGG ACU UCC UCC UUC UGG AGC CGG ACU

83	138448	TTTTTTTTTTTTTGA	T_n
84	137643	TCAAAAAAAAAAAAA	A_n
85	135070	TTTTTTTTTTTTGAG	T_n
86	134465	TTTTTTTTTTTGAGA	T_n
87	134262	CTCAAAAAAAAAAAA	A_n
88	133917	TCTCAAAAAAAAAAA	A_n
------------------------- Alu and variants of the above			
185	85432	TTTATTTATTTATTT	TTTA_n
186	85142	AAATAAATAAATAAA	AAAT_n
293	70591	AGAGAGAGAGAGAGA	AG_n
298	70411	TCTCTCTCTCTCTCT	TC_n
945	33435	AATAATAATAATAAT	AAT_n
999	31742	CTTCCTTCCTTCCTT	TTCC_n

The list ends at line ~700 000 000

~300 000 000 15-mers do not appear at all
(of total 1 073 741 824)

GCTGGGATTACAGGC

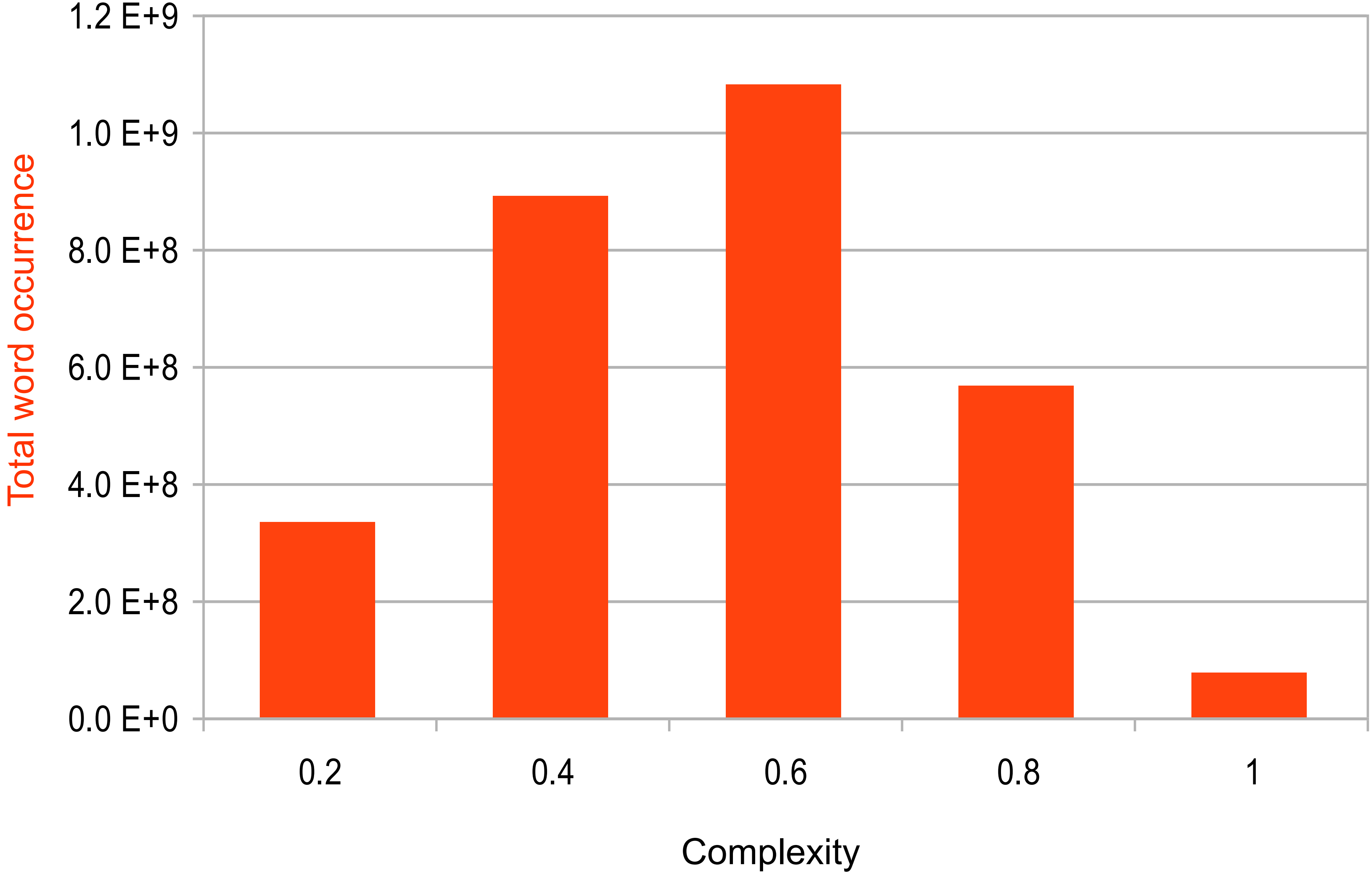
GCT	RYY
GGG	RRR
ATT	RYY
ACA	RYR
GGC	RRY

$(\mathrm{Gct})_n$ $(\mathrm{RYY})_n$

In the vocabulary of human genome 15-mers the simple repeats (low complexity words) dominate.

The high complexity words (of no repeat structure) are expected to be rather avoided.

Occurrences of simple sequence 15-mers are anomalously high



GCTGGGATTACAGGC (Alu sequence)
(complexity 0.68)

GCT

GGG

AT**T**

A**C**A

GGC

repeating
RYY_5

$\mathbf{GCT_5}$ aggressive triplet

TWO STRANDS OF THE SAME REPEATING DUPLEX

ARE REPRESENTED IN mRNA SEQUENCE BY 6 DIFFERENT TRIPLETS

GCUGCUGCUGCUGCUGCUGCUGCUGCUGCUGCUGCU

GCU GCU GCU GCU GCU GCU GCU GCU GCU GCU GCU GCU $(GCU)_n$

G CUG CUG CUG CUG CUG CUG CUG CUG CUG CUG CUG CU $(CUG)_n$

GC UGC UGC UGC UGC UGC UGC UGC UGC UGC UGC UGC U $(UGC)_n$

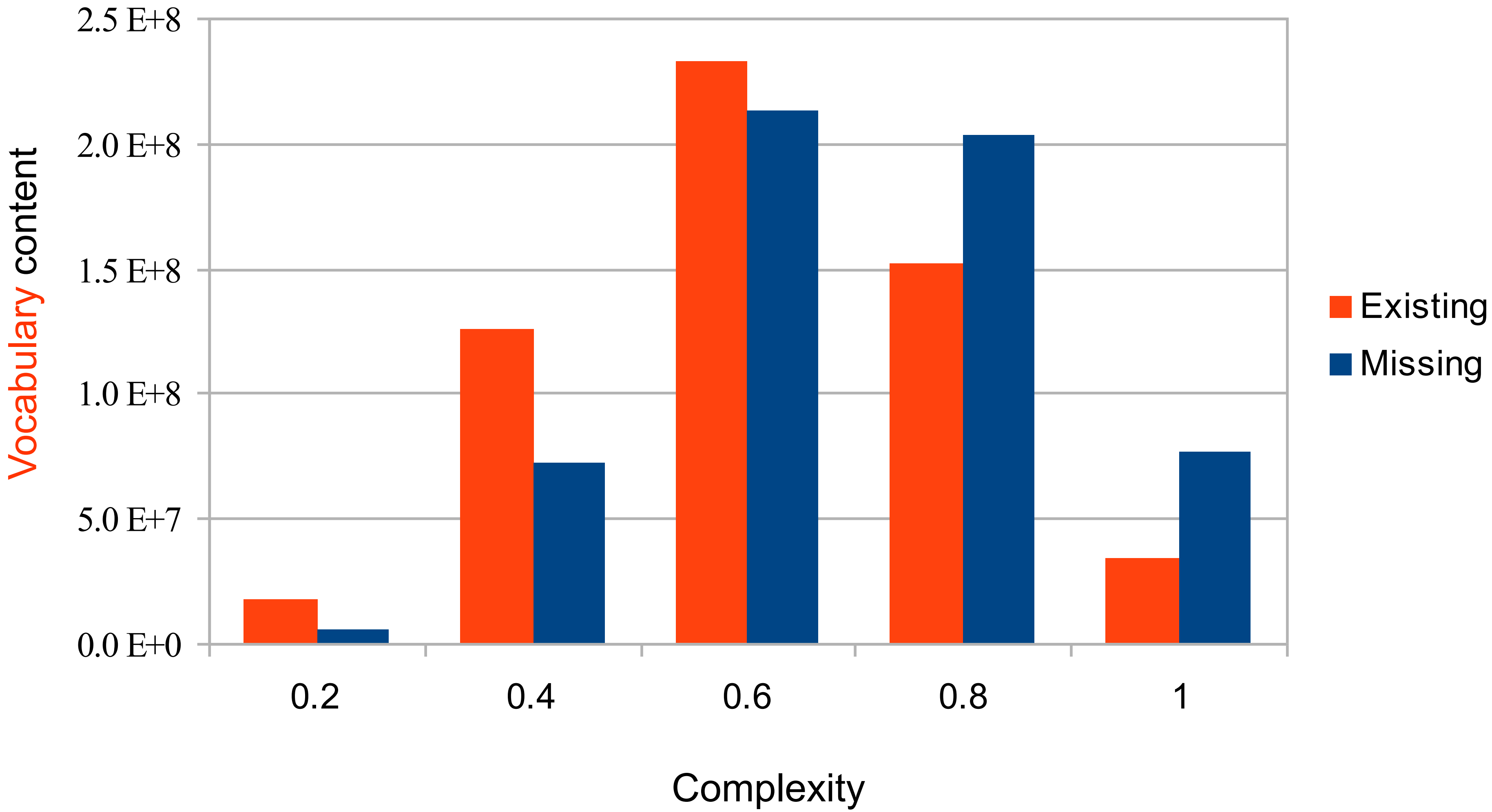
AGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGCAGC

AGC AGC AGC AGC AGC AGC AGC AGC AGC AGC AGC AGC $(AGC)_n$

A GCA GCA GCA GCA GCA GCA GCA GCA GCA GCA GCA GC $(GCA)_n$

AG CAG CAG CAG CAG CAG CAG CAG CAG CAG CAG CAG C $(CAG)_n$

15-mers of human genome are on low sequence complexity side.
High complexity words are rather avoided



Genomes are simpler than we have thought

They are dominated by simple sequences

because they originate from simple sequences,

as non-stop local births of new life