institutCurie

Ensemble, prenons le cancer de vitesse.

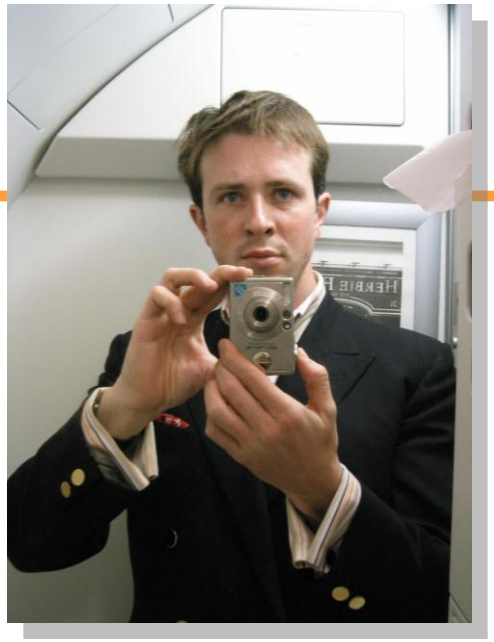
How much non-coding DNA do eukaryotes require?

Andrei Zinovyev

UMR U900

"Computational Systems Biology of Cancer"

Institute Curie/INSERM/Ecole de Mine Paritech

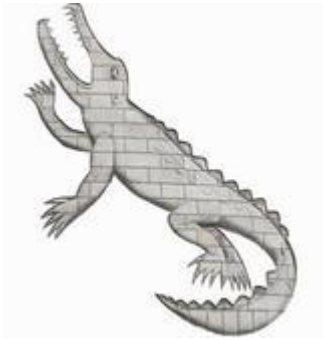


Dr. Thomas Fink

institutCurie
Ensemble, prenons le cancer de vitesse.

Bioinformatics service

Dr. Sebastian Ahnert



Cavendish laboratory,
University of Cambridge

Ahnert, S.E., Fink T. Zinovyev A.
How much non-coding DNA do eukaryotes require? J. Theor. Biol. (2008)

institutCurie

Inserm
Institut national
de la santé et de la recherche médicale

MINES PARIS

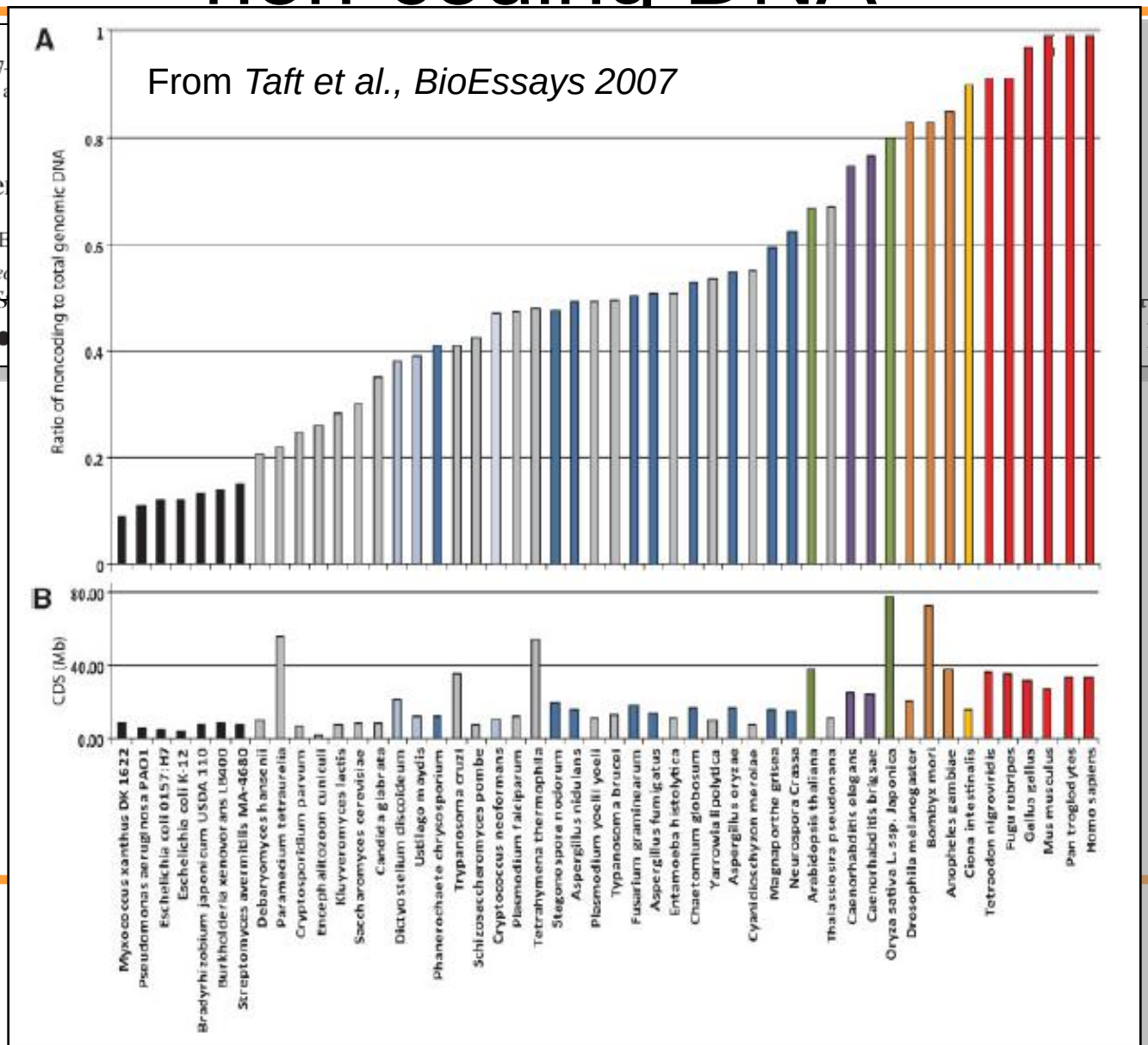
C-value and G-value paradox

- Neither genome length nor gene number account for complexity of an organism
- *Drosophila melanogaster* (fruit fly) C=120Mb
- *Podisma pedestris* (mountain grasshopper) C=1650 Mb

Genome Size (C-value)

- correlates with cell division rates, nucleus size, cell size, rates of basal metabolism, seed size
- deletions of several Mb of the mouse genome in gene deserts seem to not affect the phenotype
- about half of human and mouse genomes are repetitive
- mutational equilibrium models (long insertions/small deletions) partially explain variety of genome sizes

Burden and scaling of non-coding DNA



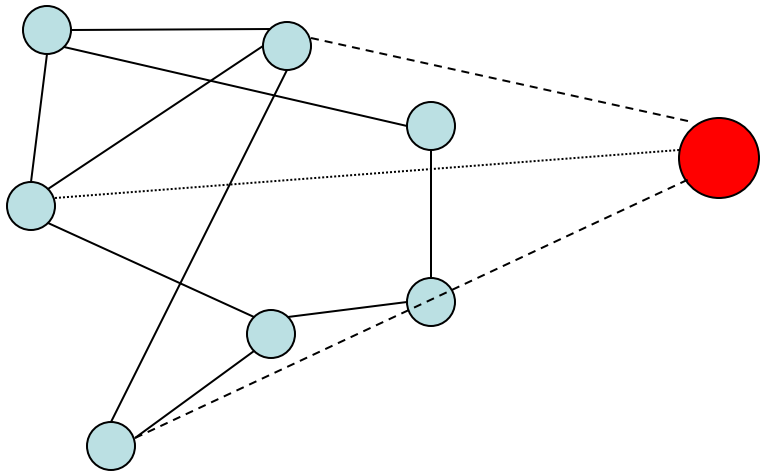
institutCurie

MINES PARIS

Non-linear growth of regulation

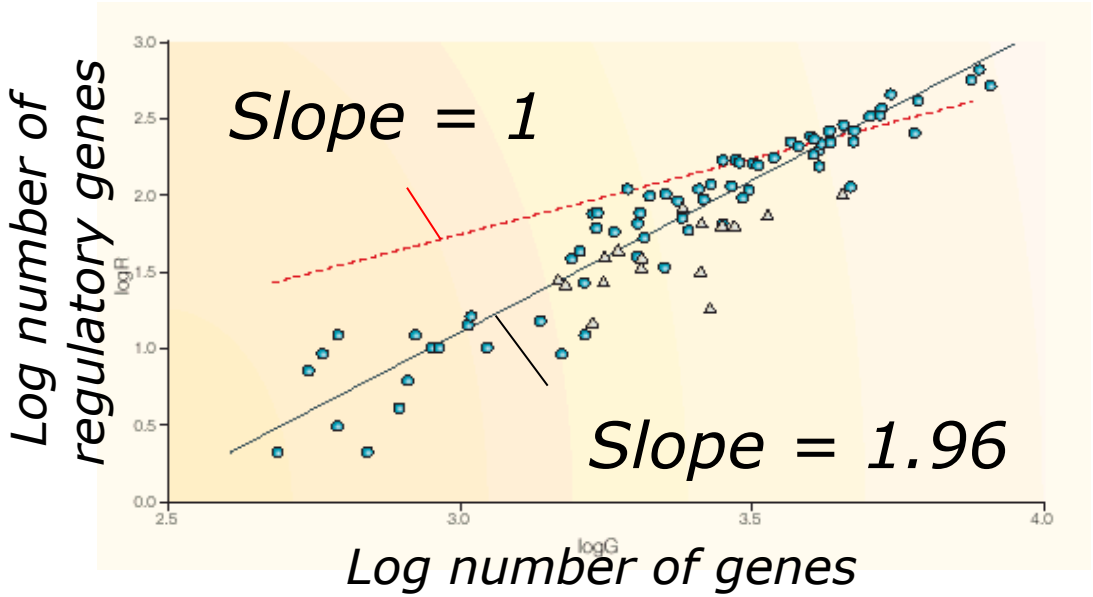
"Amount of regulation" scales non-linearly with the number of genes

N – number of genes
R – number of regulations



$R \sim N^2$

Non-linear growth of regulation



○ *bacteria*

△ *archae*

Number of regulatory genes is predicted by searching all genes for matches to Pfam profiles of protein domains [16] with known regulatory or signalling functions and/or known to be involved in DNA or RNA binding

From Mattick J.S. RNA regulation: a new genetics? Nature Rev genetics, 2004

Complexity ceiling for prokaryotes

- Adding a new function DS requires adding a regulatory overhead DR, the total increase is

  $$\Delta N = \Delta R + \Delta S$$

- Since $R \sim N^2$, at some point DR >> DS,
- *i.e. gain from a new function is too expensive for an organism, it requires too much regulation to be integrated*

There is a maximum possible genome length for prokaryotes (~10Mb)

How eukaryotes bypassed this limitation?

- Presumably, they invented a cheaper (digital) regulatory system, based on RNA

- This regulatory information is stored in the “non-coding” DNA

‘Analogue’ vs ‘digital’ or ‘Hardware’ vs ‘software’ regulation

vs

Protein-based regulation

RNA-based regulation

institutCurie

Inserm
Institut national
de la santé et de la recherche médicale

MINES PARIS

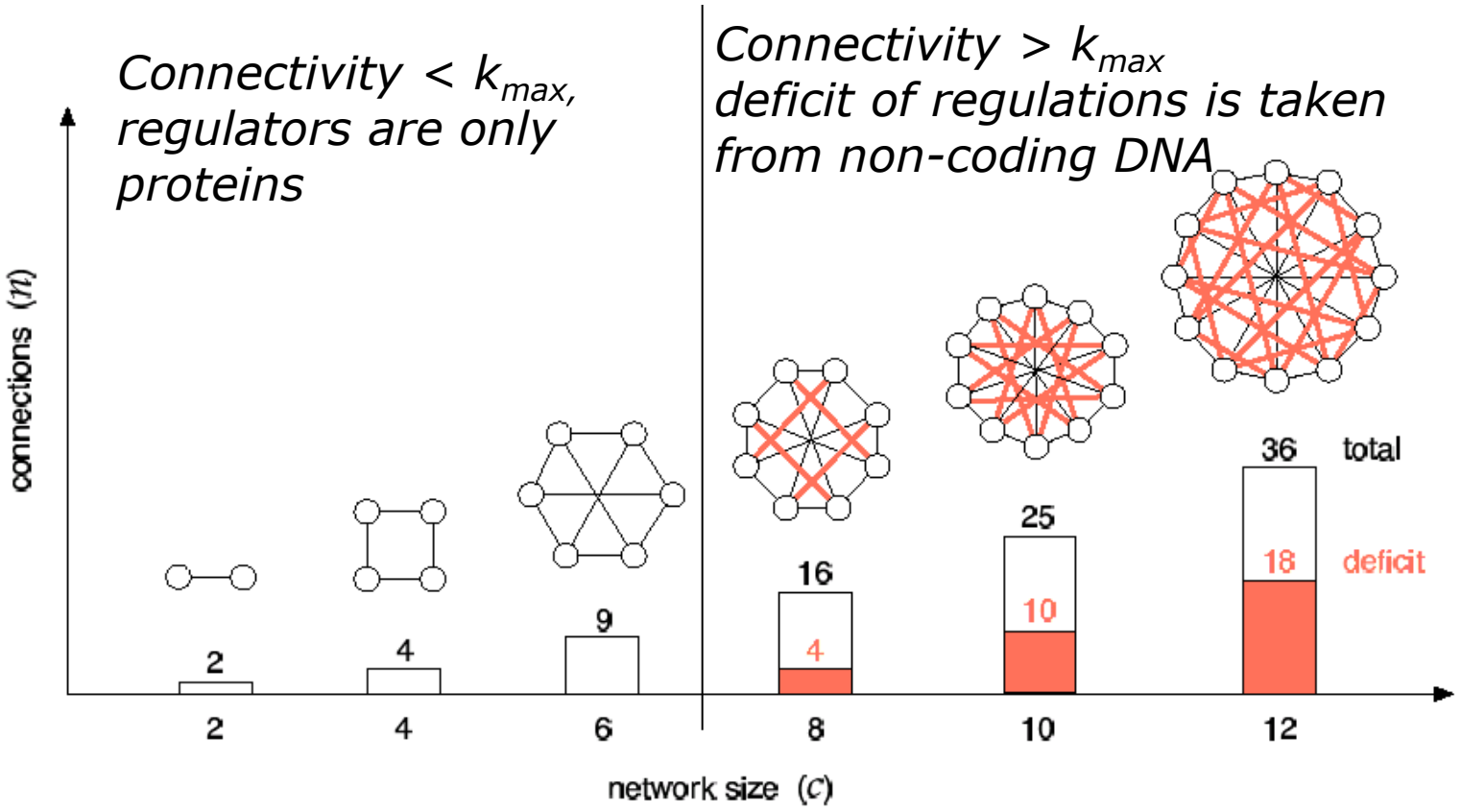
Non-coding RNAs

- The major output of metazoan genomes is non-coding RNA (introns, intergenic, antisense transcripts)
- A key advantage of RNA is its sequence specificity, in that it can direct a precise interaction with its target by base pairing, over short stretches of nucleotides, far more efficiently than can be achieved by proteins
- Simplest way of functioning: by antisense binding inhibiting some other interaction
- Many RNAs function as ‘digital-analogue’ adaptors, with a target sequence-specific address code and separate structural motifs that specify the type of consequent function and bind the appropriate protein

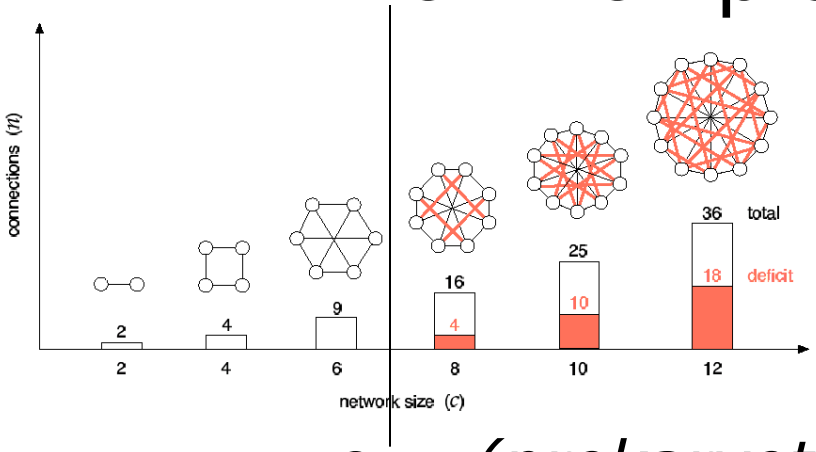
Simple model: Accelerated networks

Node is a gene (*c* genes)
Edge is a "regulation" (*n* edges)

$$n = \alpha c^2$$



How much regulation genome is needed from non-protein-coding DNA?



c_{max} *(prokaryotic ceiling)*

Average degree of connectivity $k = 2n/c$

Number of 'expensive' edges per node is limited (k_{max})

Maximum number of 'expensive' edges possible $n_p = c_{max}\, k_{max}/2$

Number of deficit 'cheap' edges needed $n_{DEF} = \frac{k_{\max}}{2}\frac{c}{c_{\max}}(c - c_{\max})$

Amount of non-coding DNA needed for cheap regulators

Encoding the regulation in genome

$$n_{DEF} = \frac{k_{\max}}{2} \frac{c}{c_{\max}} (c - c_{\max})$$

Cost of encoding one protein-coding gene with its (protein-based) regulation (node) l_c

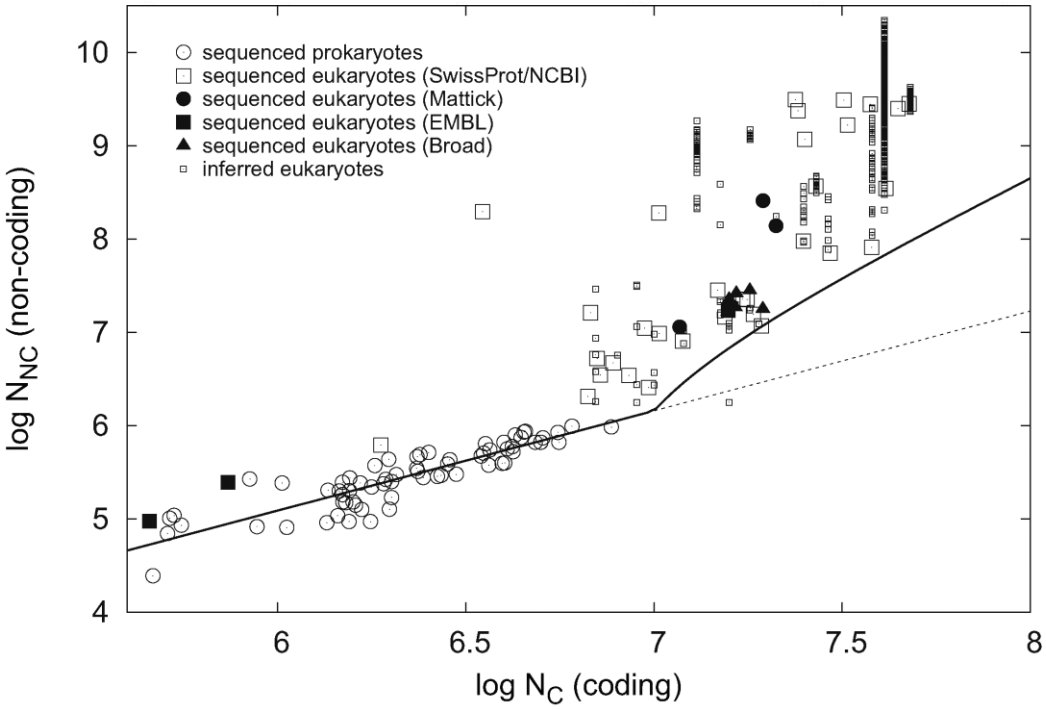
Cost of encoding one non-protein-based regulation (cheap connector) l_n

$$N_C = l_c c \ , \ N_{CC} = l_c \, c_{max} \ , \ N_{DEF} = l_n n_{DEF} \ , \ \beta = k_{max} \, l_n / l_c$$

$$N_{DEF} = \left(\frac{\beta}{2}\right)\left(\frac{N_C}{N_{CC}}\right)(N_C - N_{CC})$$

Experimental data suggests $\beta \sim 1$

Observation:
coding length *vs* non-coding



1) Prokaryotes scale linearly

$$N_{NC} = 0.181 \; N_C^{0.975}$$

2) Transition from prokaryotes to eukaryotes is approximately continuous

3) Maximum prokaryote genome length

4) Existence of empty region in the eukaryote part of the graph (A)

330 Prokaryotes: GenBank annotations
37 Eukaryotes: RefSeq database estimates

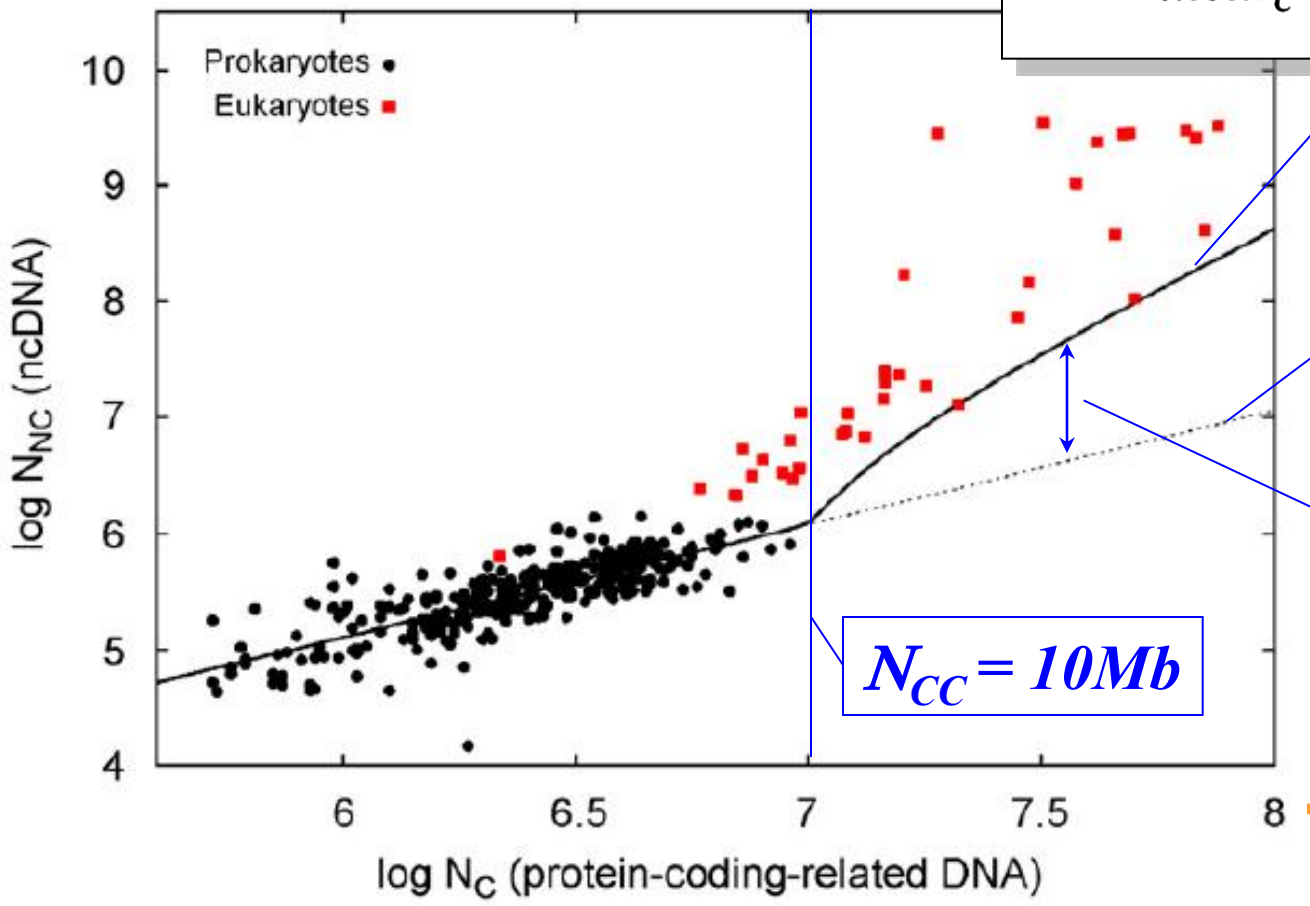
institutCurie

Inserm
Institut national
de la santé et de la recherche médicale

MINES PARIS

Observation:
coding length *vs* non-coding

$$N_{MIN} = N_P + N_{DEF}$$
$$= 0.181 N_C^{0.975} + \left(\frac{\beta}{2}\right)\left(\frac{N_C}{N_{CC}}\right)(N_C - N_{CC})$$



N_{MIN}, $\beta=1$

N_P

$N_{CC} = 10Mb$

N_{DEF} = Minimum non-coding length needed for the «deficit» regulation

Hypothesis

- Prokaryotes:

<Non-coding length> = α <**C**oding length>

$\alpha = 5–15\%$ (little constant add-on, promoters, UTRs…)

- Eukaryotes

$\mathbf{N}_{reg} = \beta/2\ \mathbf{C}/\mathbf{C}_{maxprok}(\mathbf{C}-\mathbf{C}_{maxprok}) \sim \mathbf{C}^2$,

$\mathbf{C}_{maxprok} \approx 10Mb,\ \beta \approx 1$

This is the amount necessary for regulation, but repeats, genome parasites, etc., might make a genome much bigger

Prediction on amount of functional non-coding DNA

Species	L	N_{NC}	N_C (% of L)	N_{DEF} (% of L)	N_{MIN} (% of L)
Homo sapiens	3107	3043	64 (2.1)	159 (5.1)	167 (5.4)
Rattus norvegicus	2834	2787	47 (1.7)	80 (2.8)	86 (3.0)
Mus musculus	2664	2597	68 (2.5)	179 (6.7)	186 (7.0)
Gallus gallus	1100	1063	37 (3.4)	47 (4.3)	51 (4.7)
Oryza sativa	430	384	45 (10.6)	73 (17.1)	79 (18.3)
Drosophila melanogaster	176	147	30 (16.8)	27 (15.1)	30 (17.1)
Caenorhabiditis elegans	100	72	28 (28.1)	23 (23.2)	27 (26.5)
Dictyostellum discoideum	34	13	21 (61.8)	10 (30.8)	13 (38.2)
Plasmodium falciparum	23	11	12 (53.2)	1 (5.2)	3 (11.6)

Functionality and conservation in human genome

- Prediction on the N_{MIN} for human genome:
- N_{DEF} = 167 Mb = 5.4% of genome length

N_C ~ 48 Mb = 1.7%

$N_{DEF} + N_C$ = 7.1%

● By comparing the extent of genome-wide sequence conservation to the neutral rate, the proportion of small (50–100 bp) segments in the mammalian estimated to be be explained by genome contai regions, regulat mosomal stru function.
● The mamma

ublishing Group

generated constraint annotations that integrate these data. Three annotation sets emerged from this integration: a "loose" set, defined by the union of all bases predicted as constrained for any method on any alignm union of all bases predi ods on at least two alig intersection of all three the loose, moderate, a and 2.4% of the ENCO

***Genome Res.* 2007 17: 760-774**

It seems clear that 5% is a minimum estimate of the fraction of the human genome that is functional, and that the true extent is likely to be significantly greater. If the upper figure of 11.8% under common purifying selection in mammals from ENCODE (Margulies et al. 2007) is realistic across the genome as whole, and if turnover and positive selection approximately doubles this figure (Smith et al. 2004), then the functional portion of the genome may exceed 20%. It is also now clear that the majority of

Genome Res. 17:1245-1253, 200

institutCurie

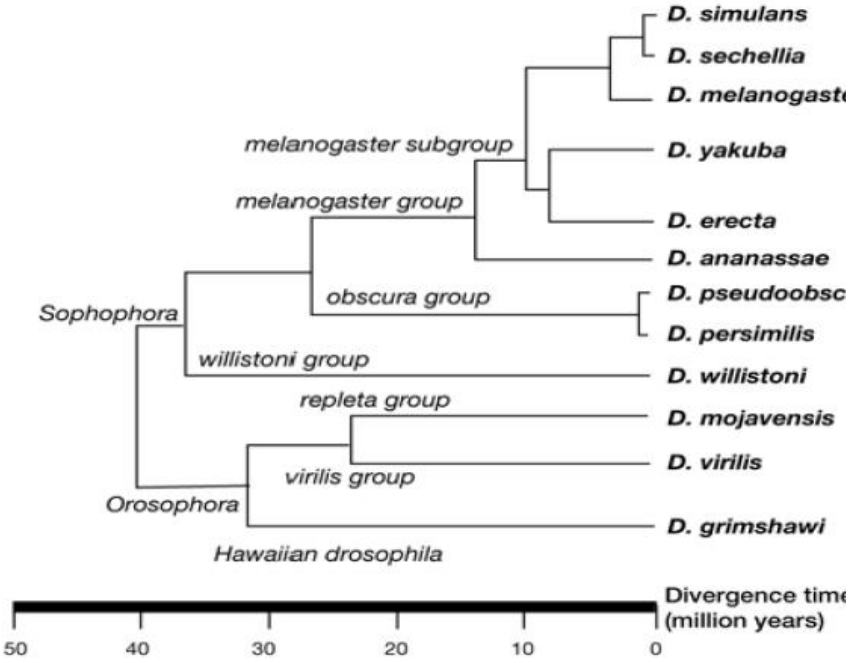
Institut national de la santé et de la recherche médicale

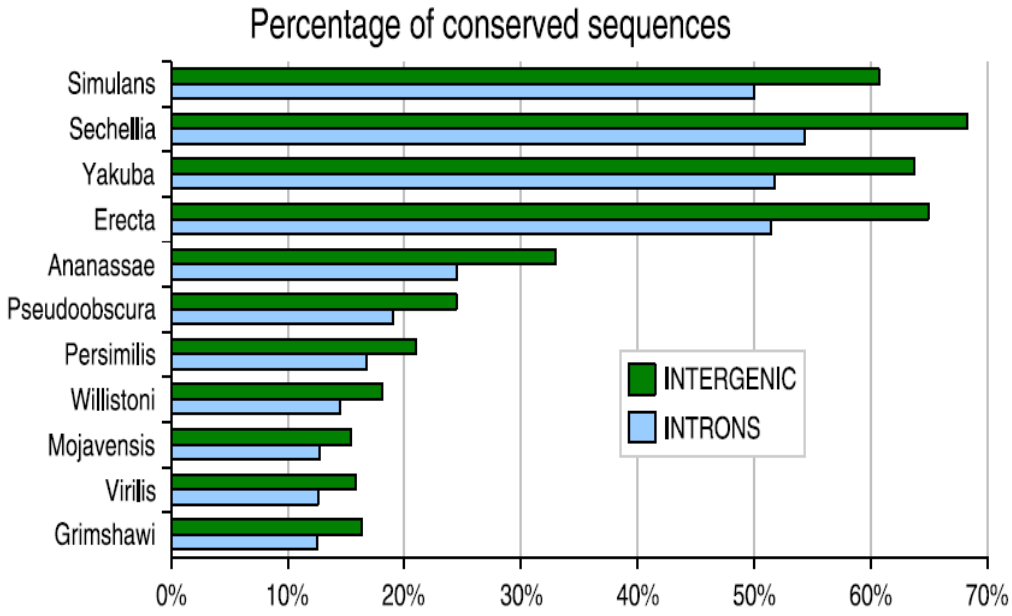
MINES PARIS

Prediction for drosophila genome

- N_{MIN} = 30Mb = 17% of genome length
- Distant drosophila species have ~13% of genome conserved in non-coding regions

Martignetti, L., Caselle, M., Jacq, B., Herrmann, C., 2007. DrosOCB: a high-resolution map of conserved non-coding sequences in Drosophila. ⟨http://arxiv.org/abs/0710.1570⟩.





Summary

- Simply looking at the lengths of coding and non-coding genome parts over all available genomes, we can formulate a hypothesis about that certain amount of non-coding DNA is functional
- The low-boundary estimate on the amount of functional non-coding DNA grows approximately quadratically with the length of the coding part, similar to the amount of regulation in simple organisms
- This allows to attribute the function of this part of DNA to the presence of the regulatory layer alternative to proteins: non-coding RNAs
- Low-boundary estimate on the functional genome part roughly corresponds to the amount of conserved DNA in several well-studied organisms