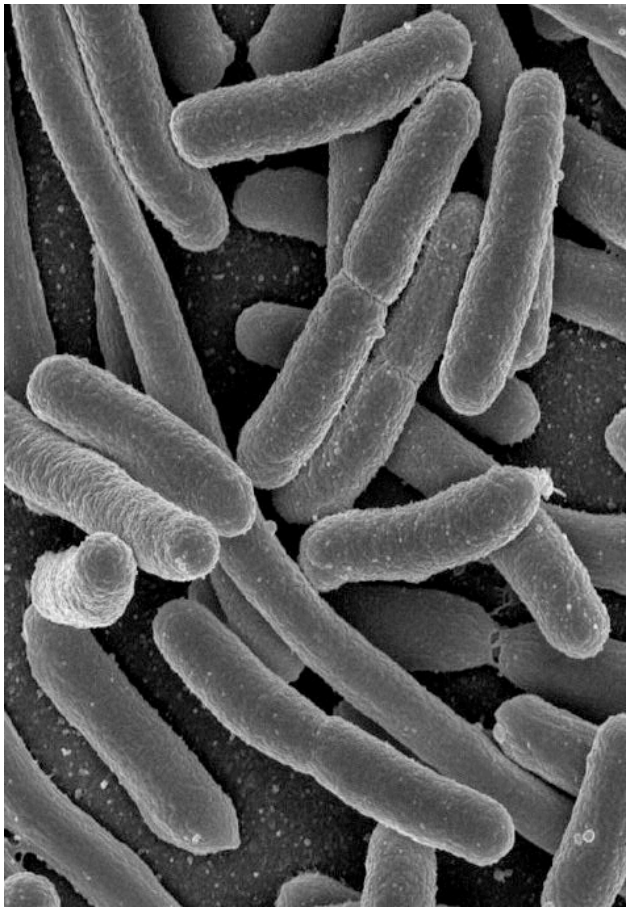
how much DNA do we need?

ESDG
20th June 2007

Sebastian E. Ahnert (TCM)

Thomas M. A. Fink (Institut Curie)
Andrei Zinovyev (Institut Curie)

prokaryotes



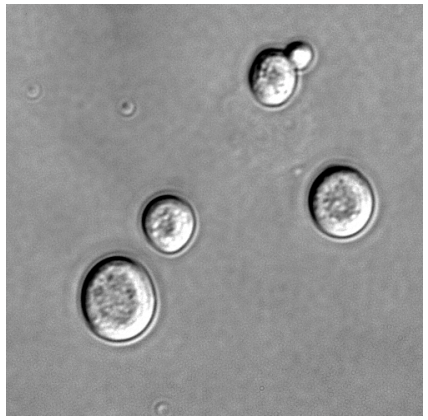
- organisms without a cell nucleus.

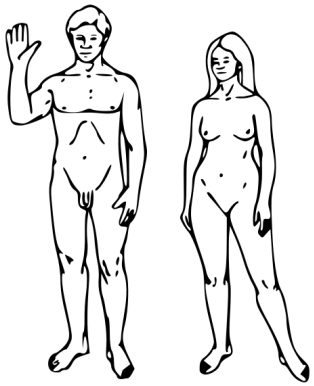
- unicellular (with a few exceptions).

- fall into two categories:

bacteria and *archaea*.

eukaryotes





- organisms with a cell nucleus.

- unicellular or multicellular

- many different kinds, including:

  fungi
  plants
  animals

coding vs. non-coding DNA

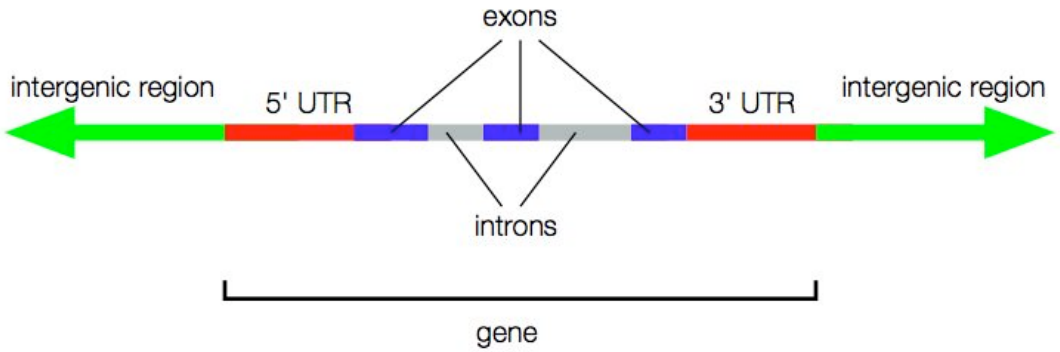
coding DNA:	The part of the DNA that is transcribed and translated into proteins. (Exons)
non-coding DNA:	The other part.

non-coding DNA

Formerly most of it was known as 'junk DNA'.

Still not much is known about the vast majority of it, but more and more indications of its importance are emerging.

There are various types of non-coding DNA:



- 5' and 3' UTRs
- Introns
- Intergenic regions

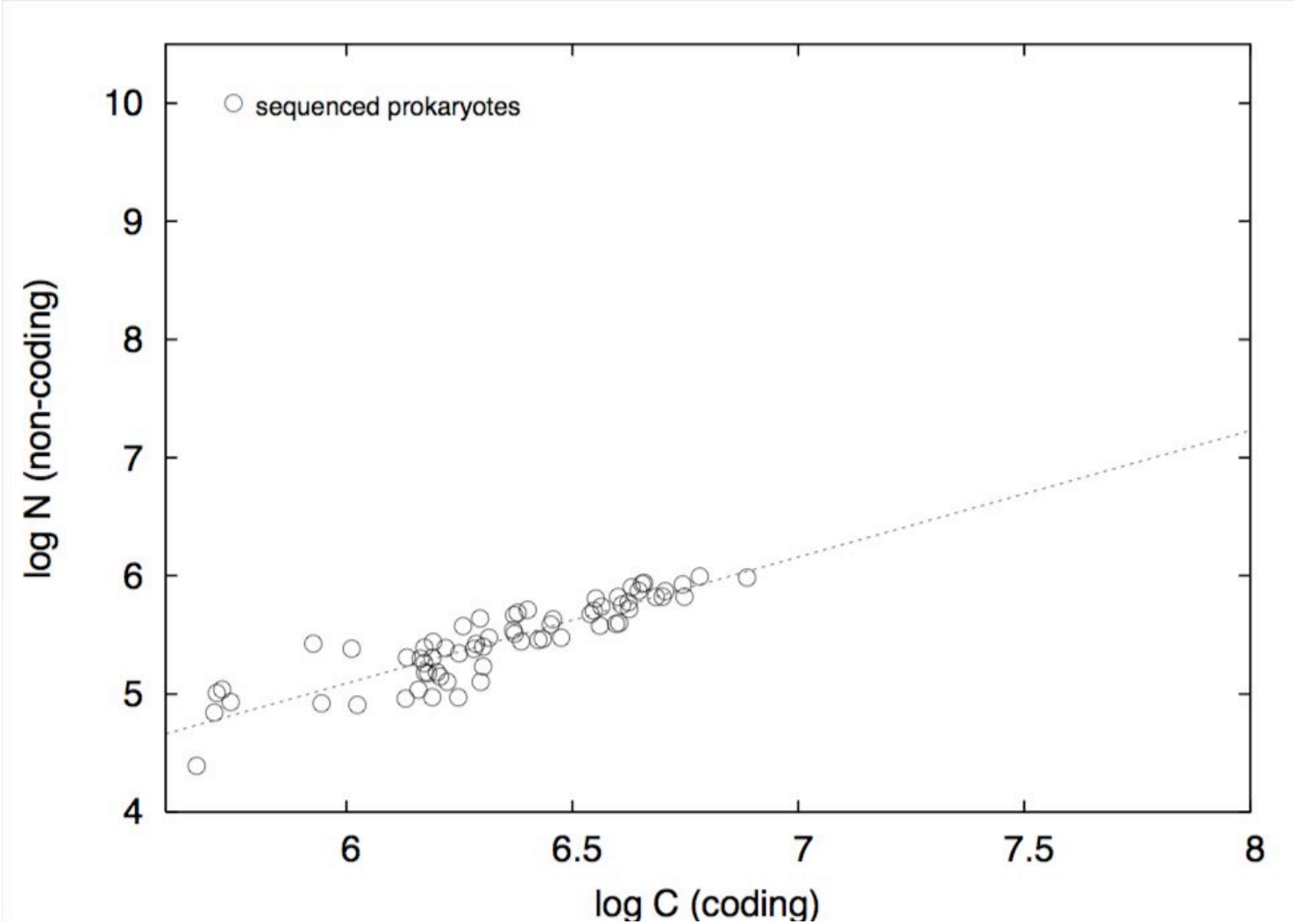
What we did

We collected data for the total amount of coding (C) and non-coding (N) sequence in the genomes of

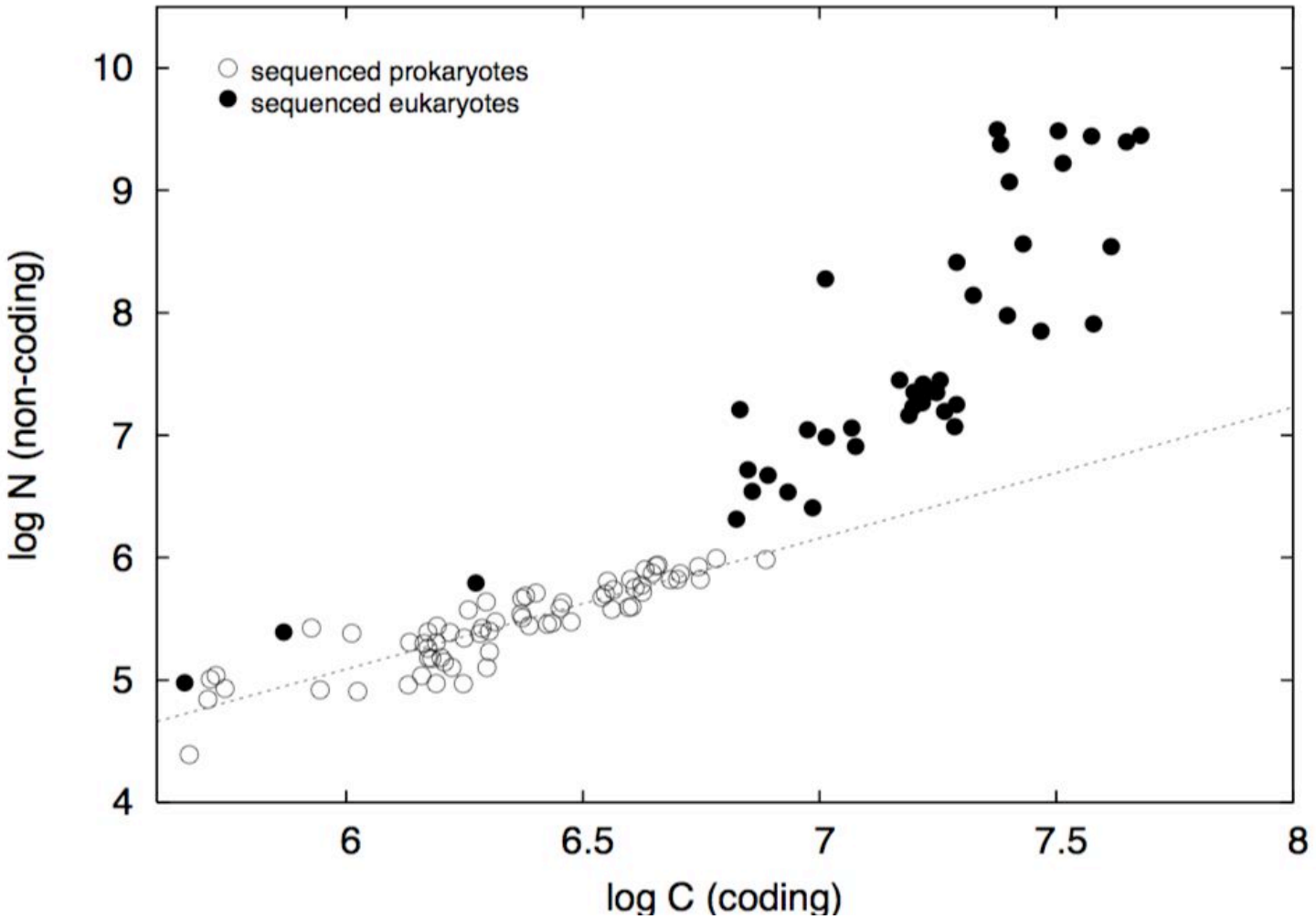
67 prokaryotes and 43 eukaryotes

…and plotted N versus C for these 110 species.

prokaryote data



prokaryote & eukaryote data



observations

- prokaryotes and eukaryotes almost completely disjoint
- dividing line around $C = 10^7$ bp (largest known prokaryote)
- prokaryotes show scaling $\sim C^{1,07}$ (almost linear)
- eukaryotes show approximately *quadratic* scaling
- hence: eukaryotes *require more non-coding DNA*

some observations

In prokaryotes it has been observed that the number of regulatory genes scales *quadratically* with the total number of genes.

This is like a network in which the number of connections grows quadratically with the number of nodes. Such networks are called *accelerated networks*.

But in any physical network, there will be a limited capacity of every node to connect.

some assumptions

Let us assume:

The point at which it becomes inefficient to add a new regulatory gene is when C grows above 10^7 bp.

Let us say that this happens when there is roughly one regulatory gene for every non-regulatory gene.

This defines the transition from prokaryotes to eukaryotes.

Then eukaryotes require a additional source of regulatory connections.

a simple model

How many additional connections do we require?

Imagine we have a eukaryote with $C = n \times 10^7$ bp.

Connections grow quadratically, so we require n^2 times as many regulatory connections as the maximum size prokaryote at $C = 10^7$ bp, where we assumed that of order half the DNA codes for regulatory connections.

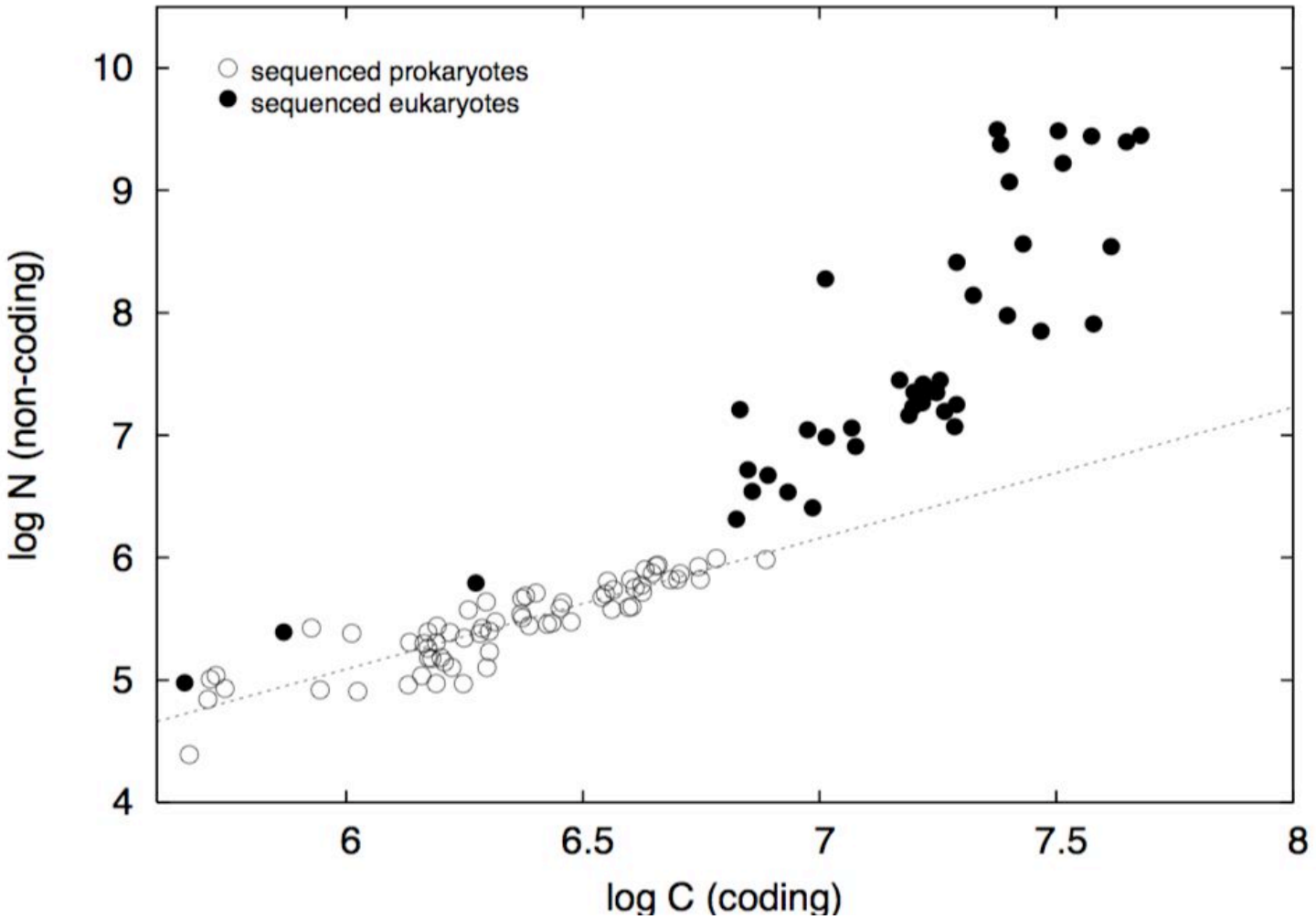
Regulatory genes can account for the regulation connections *inside* each of the *n* blocks, but not for the regulation *between* them.

Hence our shortfall is: $S = (n^2/2 - n/2)10^7 = (1/2)(C / 10^7)(C - 10^7)$

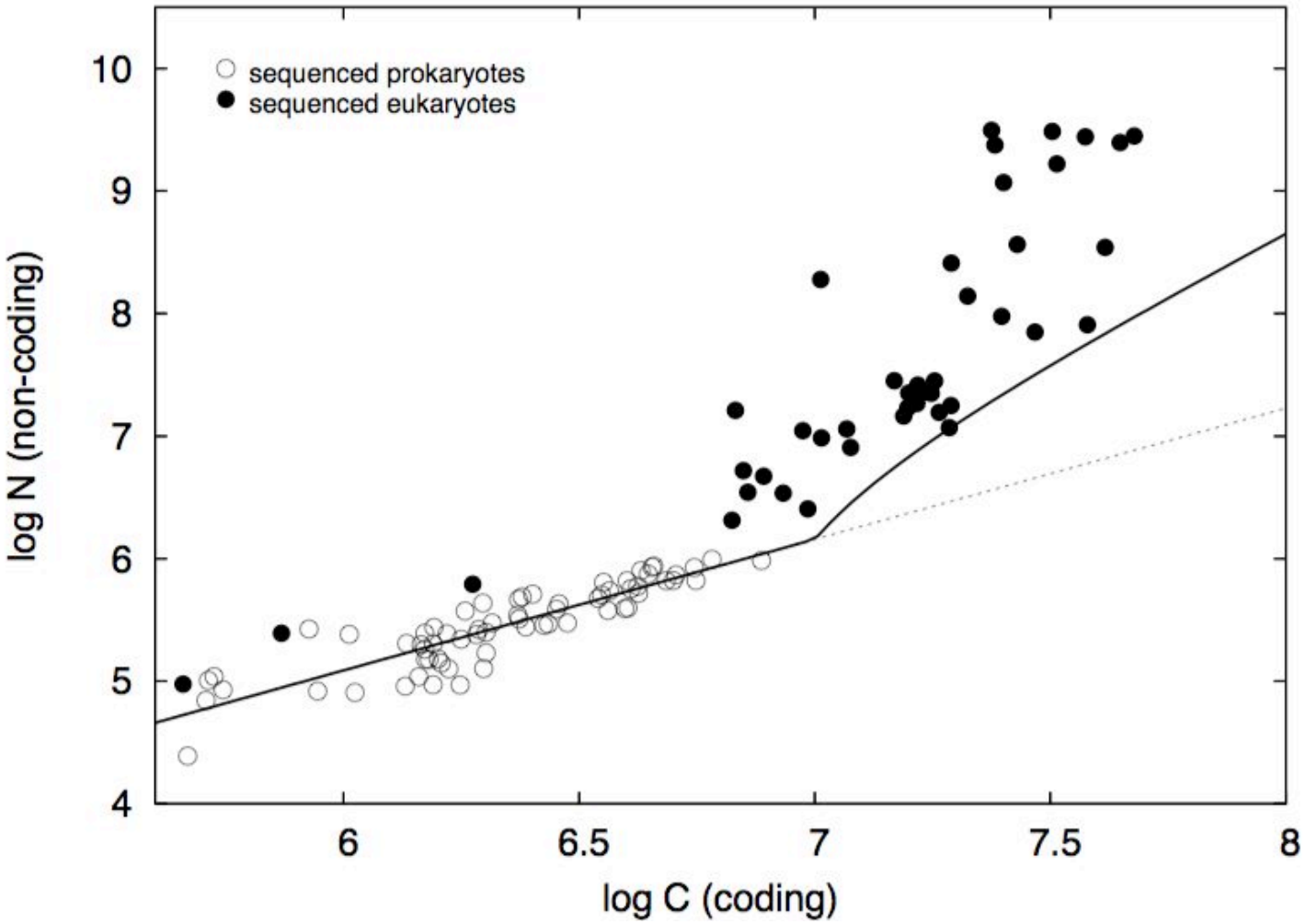
the big question

Do eukaryotes cover this shortfall by recruiting non-coding DNA for regulation?

prokaryote & eukaryote data



prokaryotes, eukaryotes & model



conclusions

- The model matches the data very well.

- It seems that eukaryotes indeed recruit non-coding DNA to cover their regulatory deficit.

- Percentage of minimum necessary amount of non-coding DNA varies widely, from a few percent (in humans) to half of the genome (in simpler eukaryotes).

- Interesting future question: Can we narrow down which parts of the non-coding DNA are used?