Proto-genes and *de novo* gene birth

Nature 2012 jul 19

Anne-Ruxandra Carvunis, Thomas Rolland, Ilan Wapinski, Michael A. Calderwood, Muhammed A. Yildirim,Nicolas Simonis, Benoit Charloteaux, Ce´sar A. Hidalgo, Justin Barbette, Balaji Santhanam, Gloria A. Brar,Jonathan S. Weissman, Aviv Regev, Nicolas Thierry-Mieg, Michael E. Cusick & Marc Vidal

Objective

- Formalize an evolutionary model according to witch functional genes evolve *de novo through transitory proto-genes generated by widespread translational activity in non-genic sequences.*

Novel protein-coding genes

- Arise through re-organization of pre-existing genes
  - After gene duplication

- *de novo*
  - *Poorly understood*
  - *Insignificant polypeptides*

De novo gene birth

- non-genic sequences acquire ORFs and become transcribed
- transcripts access the translation machinery

Hard to reconcile this proposed mechanism because

- non-genic sequences should lack translational activity
- if translated, should encode insignificant polypeptides

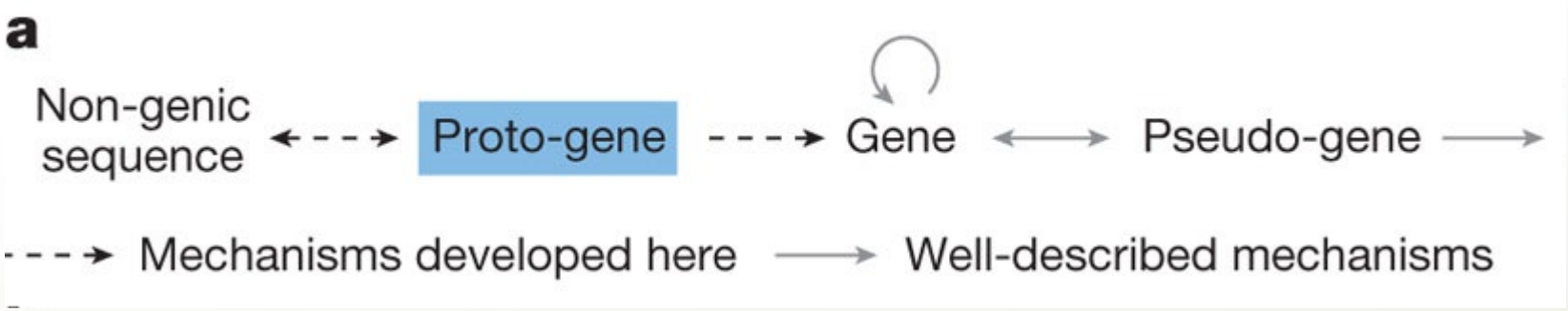
Evolutionary model

- Evidence of associations between non-genic transcripts and ribosomes

- Genes that originate de novo could initially be simple and gradually become more complex over evolutionary time

Evolutionary model

- *de novo* gene birth proceeds through intermediate and reversible proto-gene stages, mirroring the well-described pseudo-gene stages of gene death



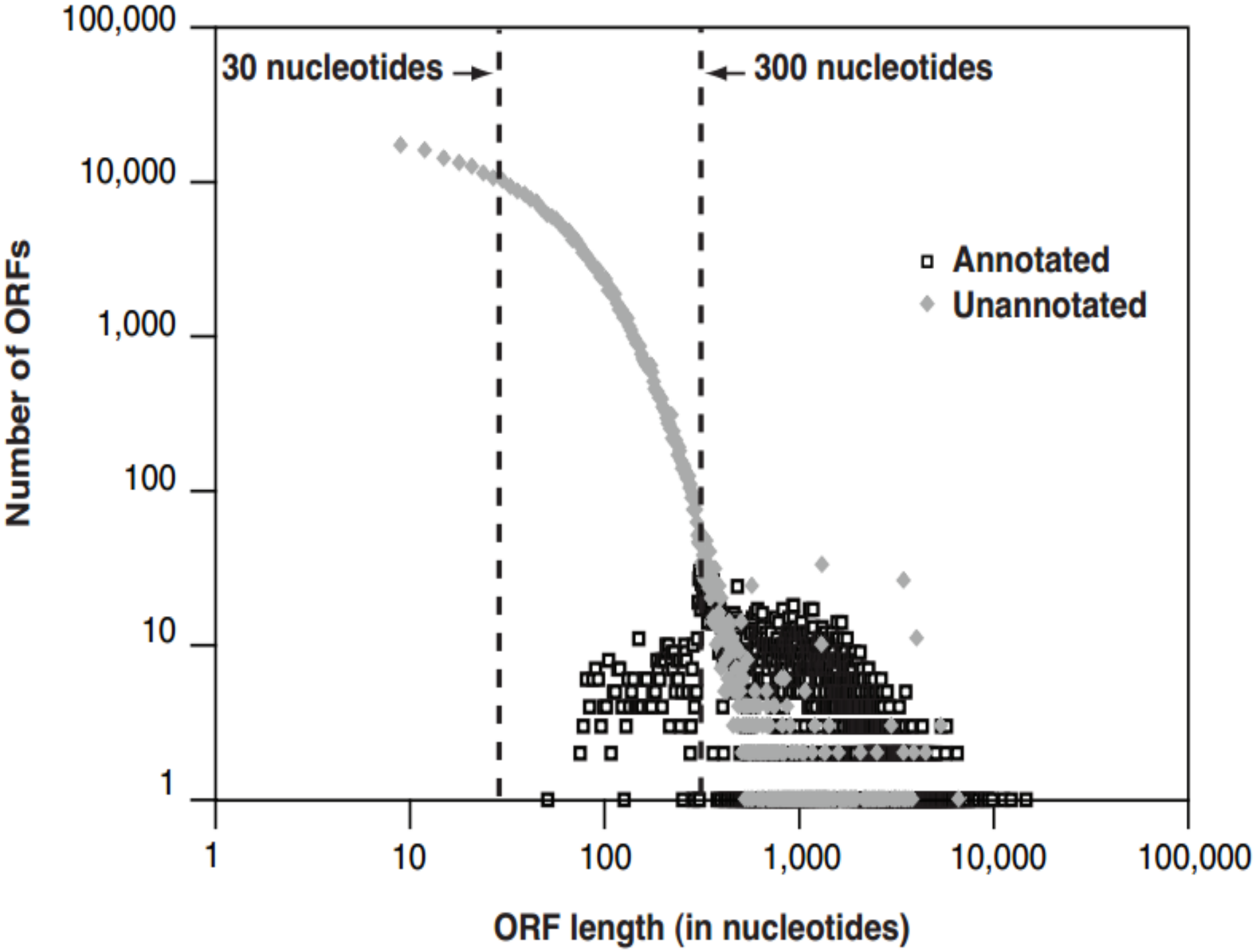
Data

- S288C reference strain of *S. cerevisiae* annotations downloaded in October 2007 by the Saccharomyces Genome Resequencing Project group (SGRP) from the Saccharomyces Genome Database (SGD)
- Paralogy relationships among annotated open reading frames (ORFs) of *S. cerevisiae* were downloaded from the Ensembl Compara website

Setup

- *Saccharomyces cerevisiae*
  - Genome ~12Mbp
  - ~6200 genes, about 5800 are functional
  - 31% similarity to human genome

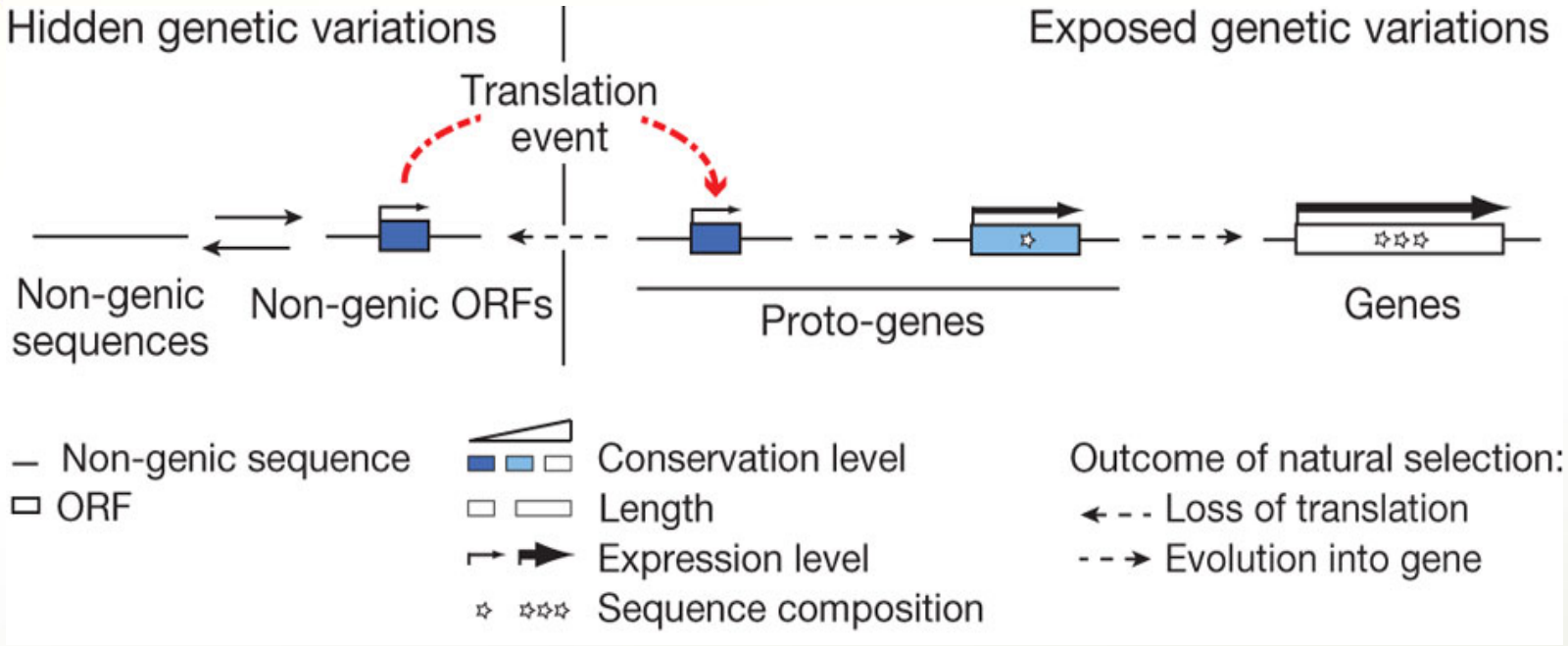
  - Threshold of ORFs were 300 nucleotides
  - 6000 ORF annotated as genes
  - 261 000 unannotated
  - both categories contain some proto-genes


100,000
10,000
1,000
100
10
1
Number of ORFs
30 nucleotides
300 nucleotides
Annotated
Unannotated
1
10
100
1,000
10,000
100,000
ORF length (in nucleotides)

Idea

- Non-genic sequences are broadly transcribed in *S. cerevisiae*, their overexpression is mostly non-toxic
- Translation of non-genic ORFs could be more common than expected.
- Such translation events would not systematically lead to *de novo* gene birth, as the corresponding polypeptides would not necessarily have specific biological functions.

Evolutionary model

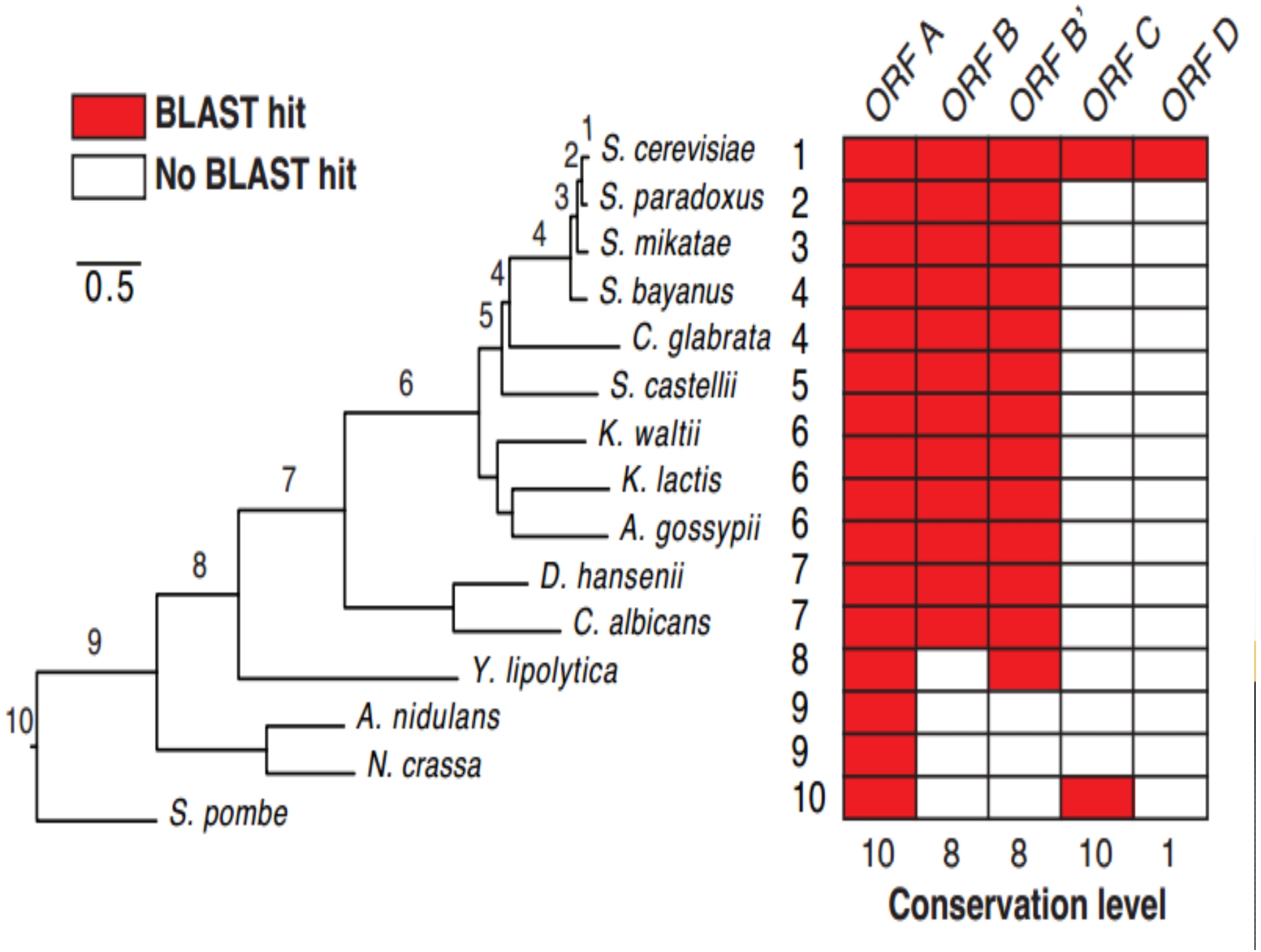


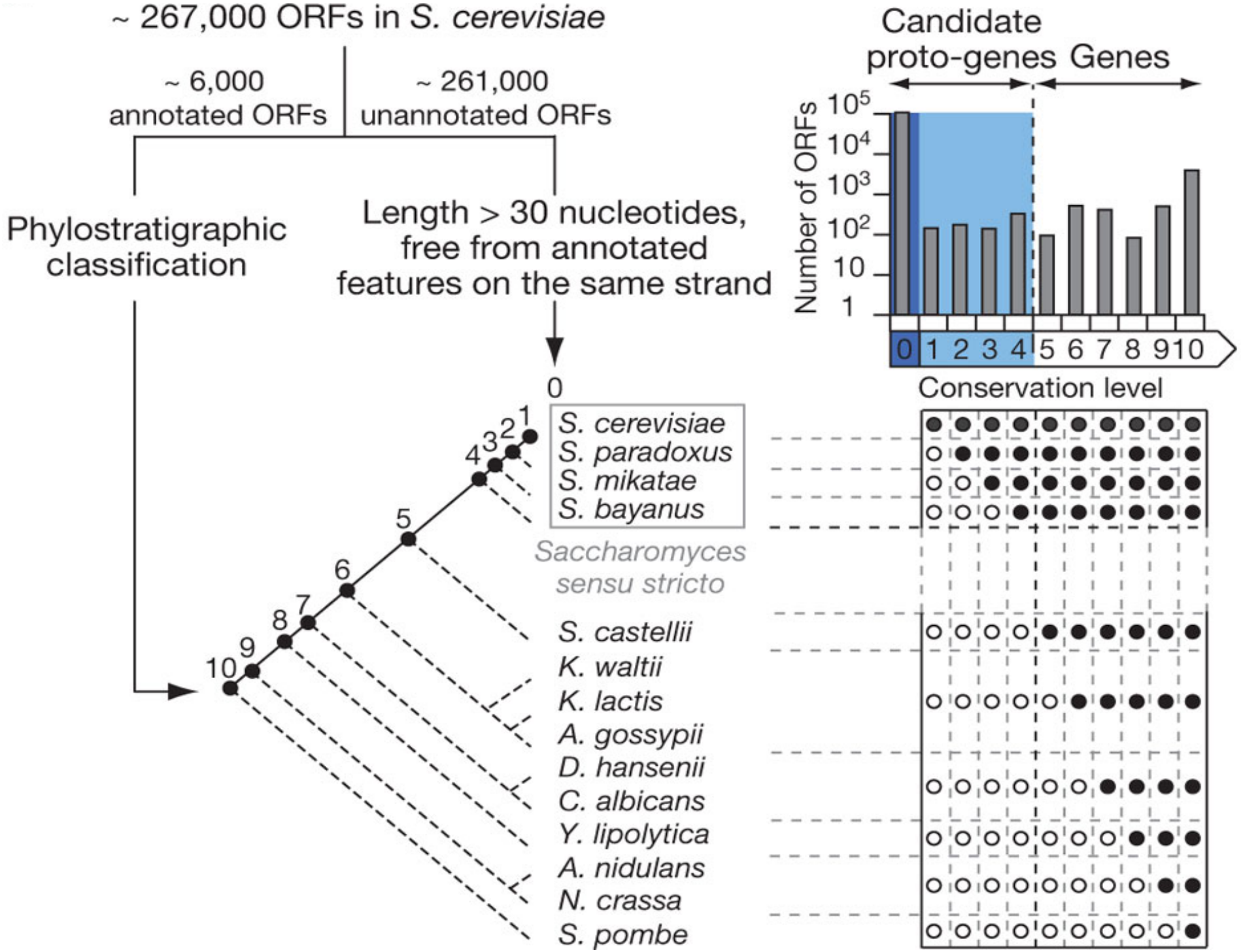
Predictions of this model

1. The structural and functional characteristics of S. cerevisiae ORFs should reflect an evolutionary continuum ranging from non-genic ORFs to genes
2. many non-genic ORFs should be translated
3. ORFs that emerged recently should occasionally have adaptive functions

Setup

- Annotated ORFs were classified into 10 groups based on their conservation throughout the *Ascomycota* phylogeny.
  - **ORFs1 -** 2% of 6000 annotated ORFs are found only in S. cerevisiae.
  - **ORFs1–4 -** 12% are found only in the four closely related Saccharomyces sensu stricto species . These are poorly characterized and their annotation as genes is debatable
  - **ORFs5–10 -** ,88% of annotated ORFs found outside of thisgroup . These are well characterized and can confidently be considered genes.


BLAST hit
No BLAST hit
0.5
S. cerevisiae
S. paradoxus
S. mikatae
S. bayanus
C. glabrata
S. castellii
K. waltii
K. lactis
A. gossypii
D. hansenii
C. albicans
Y. lipolytica
A. nidulans
N. crassa
S. pombe
1
2
3
4
4
5
6
7
8
9
10
ORF A
ORF B
ORF B'
ORF C
ORF D
1
2
3
4
4
5
6
6
6
7
7
8
9
9
10
10 8 8 10 1
Conservation level


~ 267,000 ORFs in S. cerevisiae
~ 6,000 annotated ORFs
~ 261,000 unannotated ORFs
Phylostratigraphic classification
Length > 30 nucleotides, free from annotated features on the same strand
0
1
2
3
4
5
6
7
8
9
10
S. cerevisiae
S. paradoxus
S. mikatae
S. bayanus
Saccharomyces sensu stricto
S. castellii
K. waltii
K. lactis
A. gossypii
D. hansenii
C. albicans
Y. lipolytica
A. nidulans
N. crassa
S. pombe
Candidate proto-genes
Genes
Number of ORFs
10^5
10^4
10^3
10^2
10
1
0 1 2 3 4 5 6 7 8 9 10
Conservation level

Proof

- They estimated that over 97% of ORFs1-4 originated de novo rather by cross-specis transfer

- ORFs1-4 partially overlap ORFs5-10 and this is incompatible with cross-species transfer

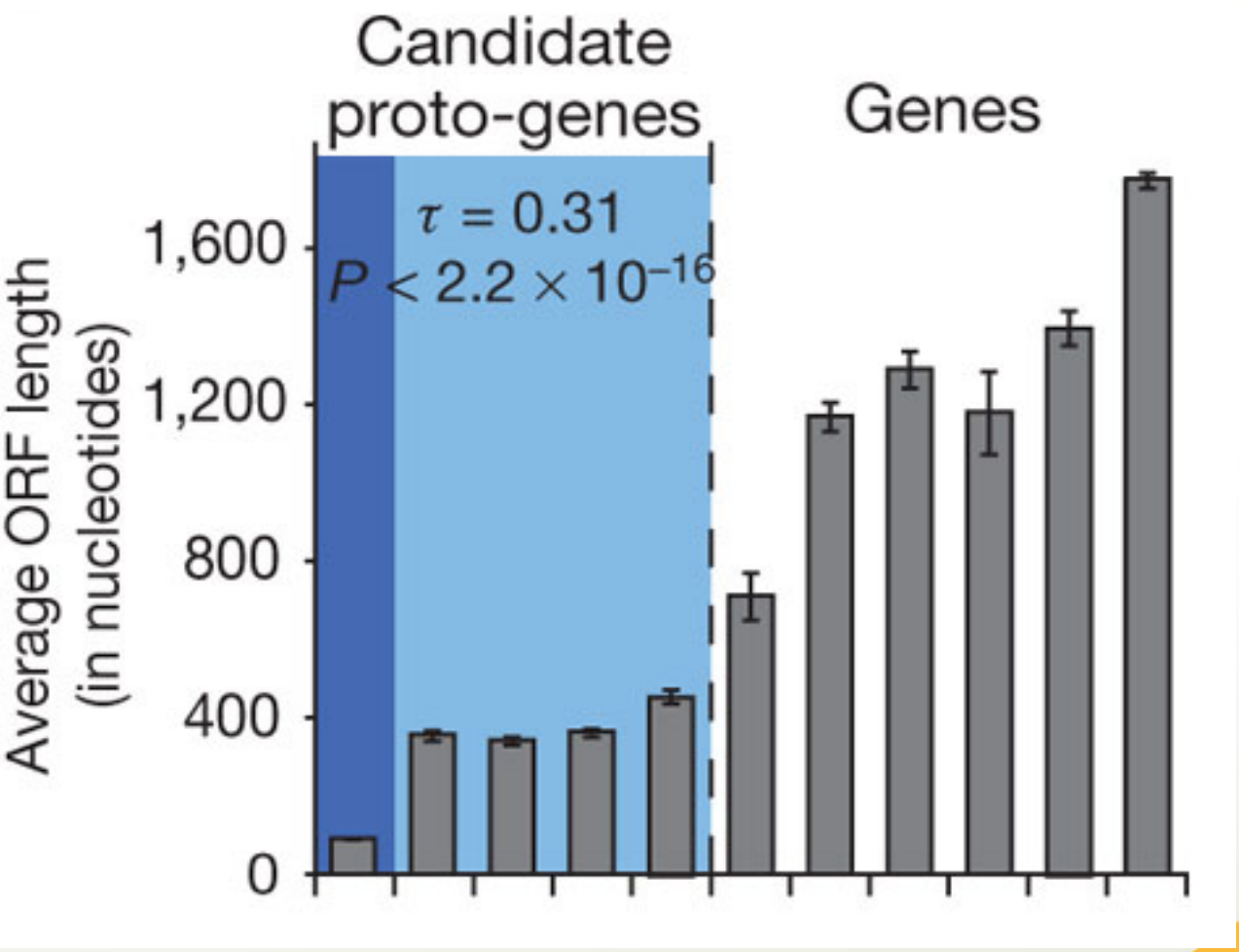
Proof

- Level 0 contains 107,425 unannotated ORFs (ORFs0), defined as any sequence between canonical start and stop codons that is
  - longer than 30 nucleotides
  - a multiple of three
  - not overlapping an annotated ORF, an rRNA, a tRNA, a ncRNA, a snoRNA or an upstream ORF on the same strand.

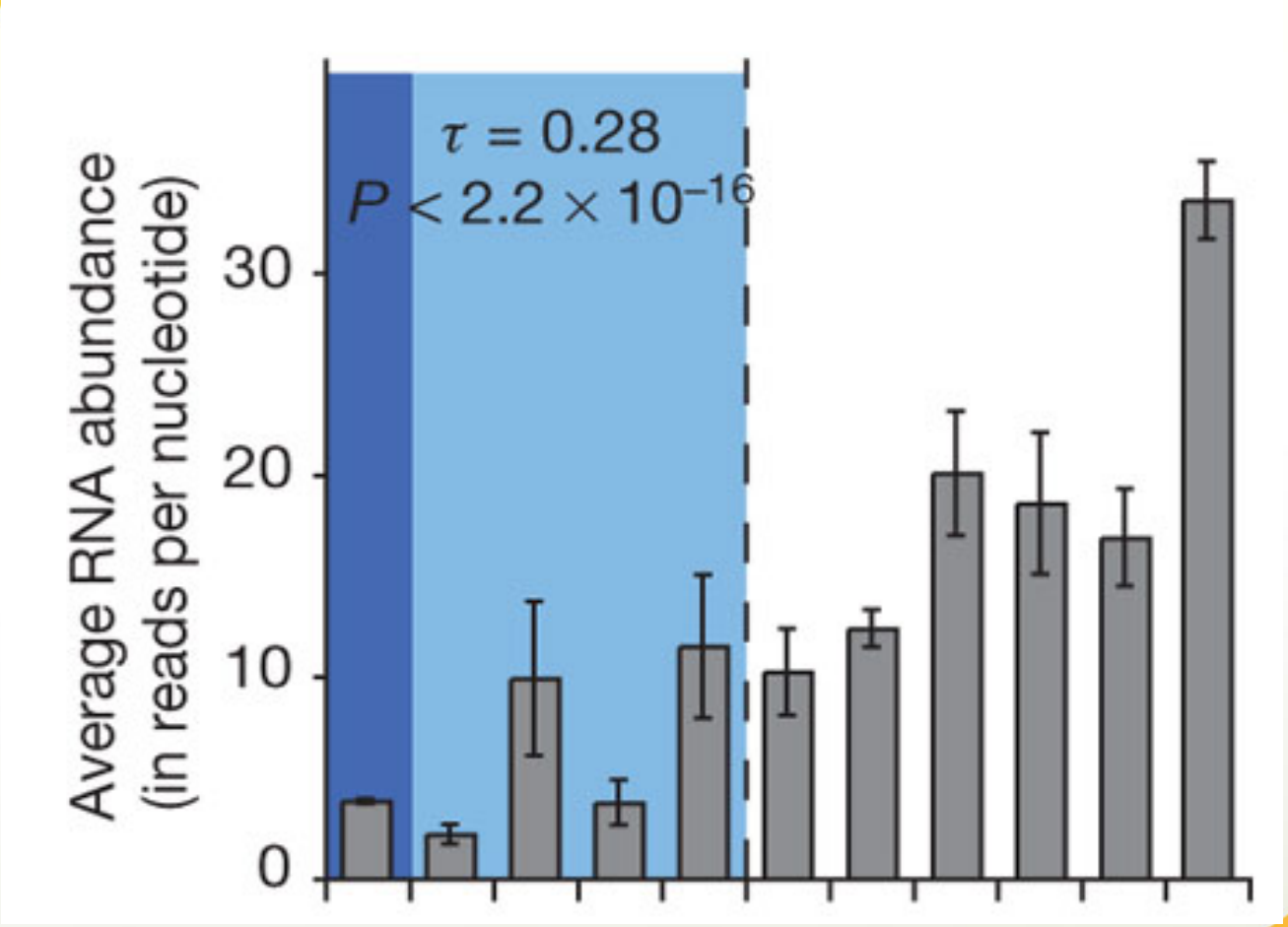
Proof

- To test the evolutionary continuum prediction, they first verified that ORF conservation level correlates positively with length and expression level

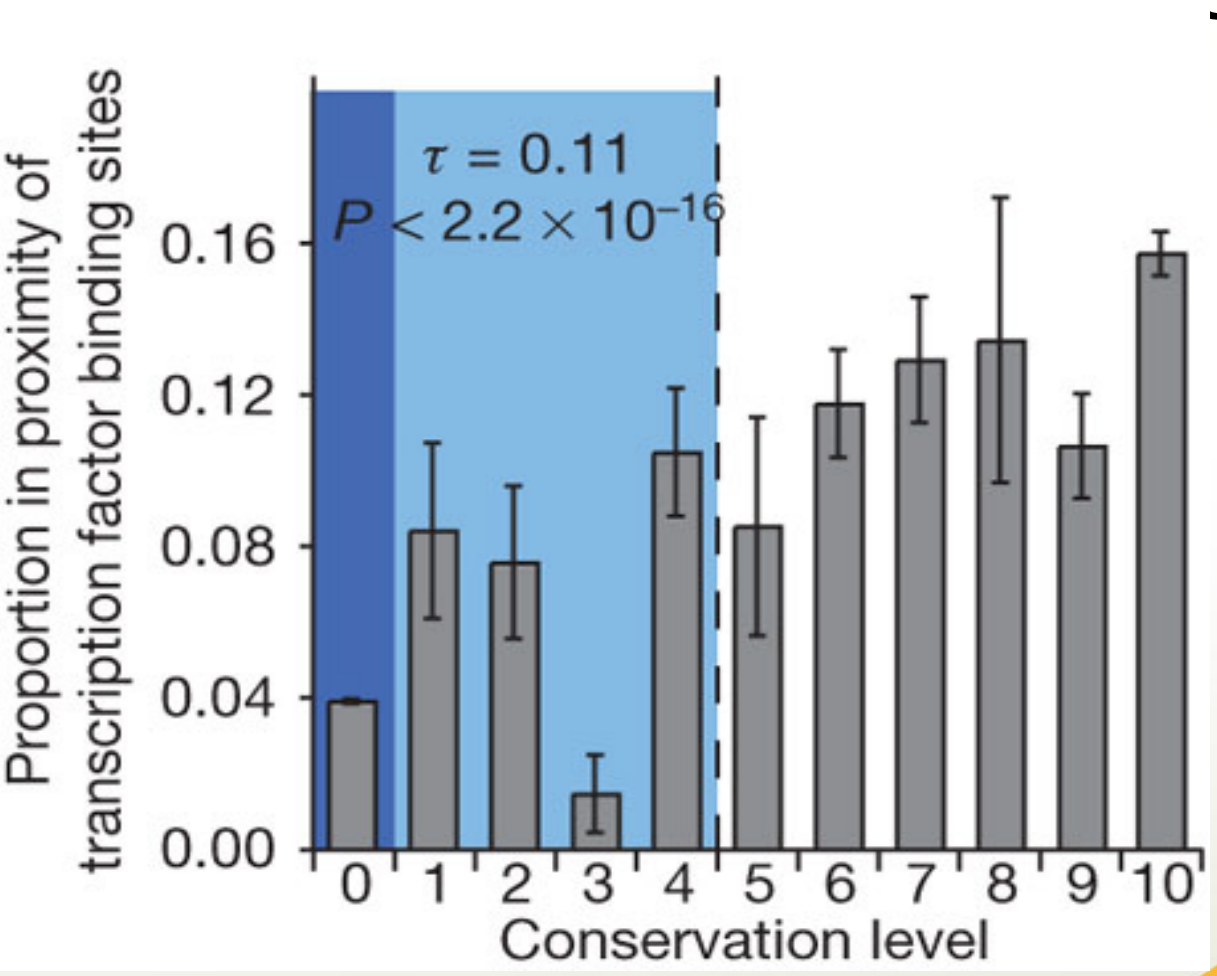
Average ORF lenght



RNA expression level



Proximity to transcription factor binding sites

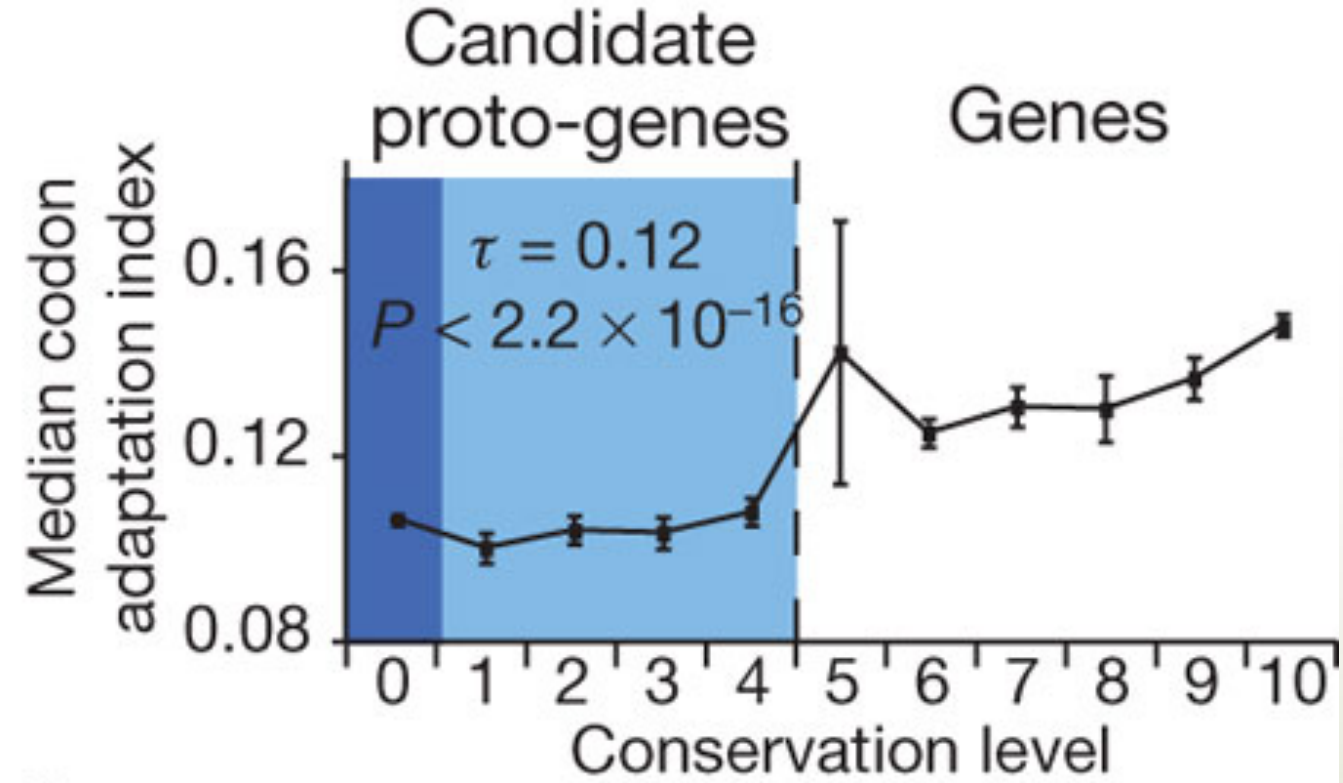


Proof

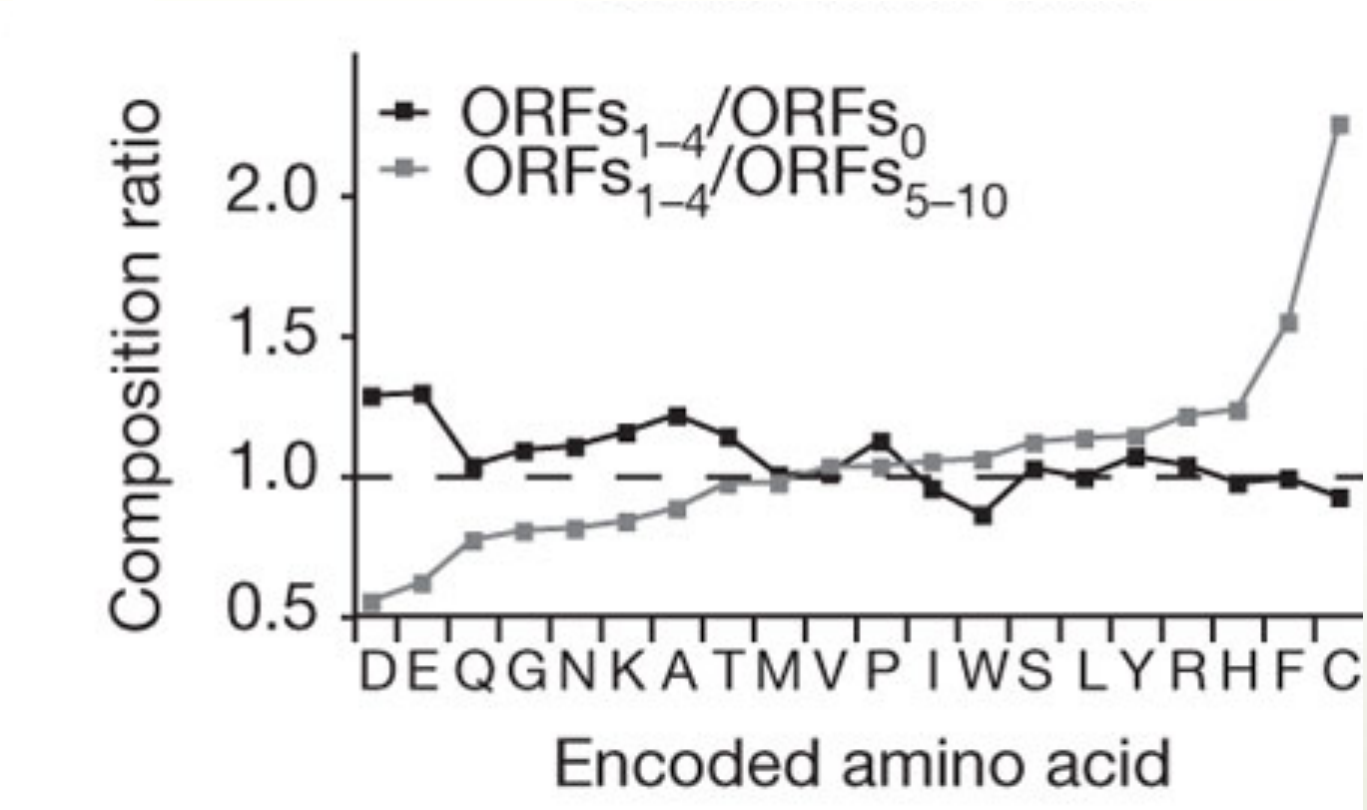
- These correlations suggest that genes evolve from non-genic ORFs that lengthen and increase in expression level over evolutionary time

- Thus, some ORFs may increase in expression level at different rates than they increase in length over evolutionary time

Correlation between codon usage and conservation level

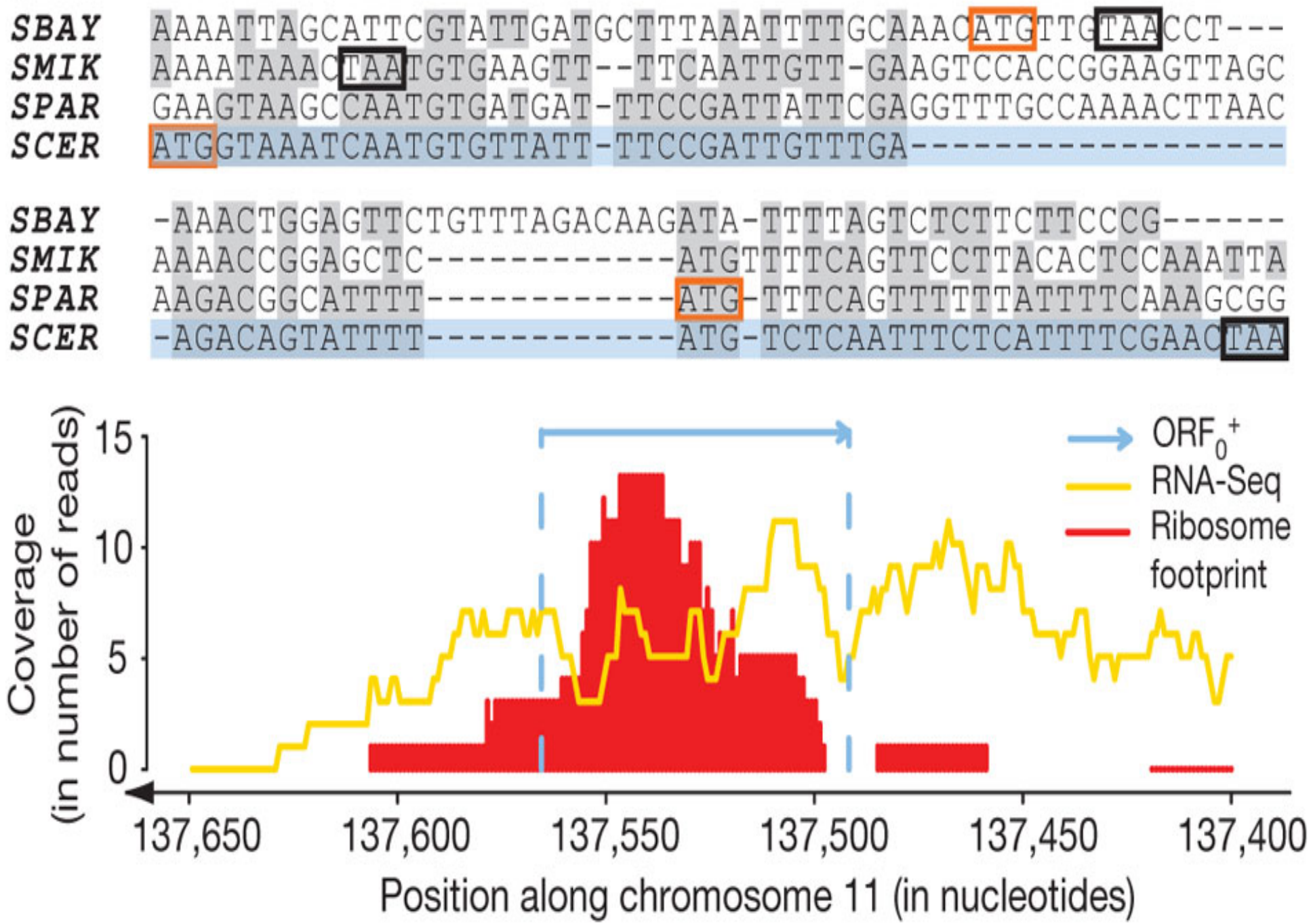


Relative amino acid abundances shift with increasing conservation level



Assessing the extent

- Taken together, their observations support the existence of an evolutionary continuum ranging from non-genic ORFs to genes
- They searched for signatures of translation of ORFs0 in a ribosome footprinting data
  - Rich and starvation condtions
- They developed a stringent pipeline to detect unequivocal translation signatures for ORFs0 located on transcripts associated with ribosomes
  - Bowtie short read mapping program

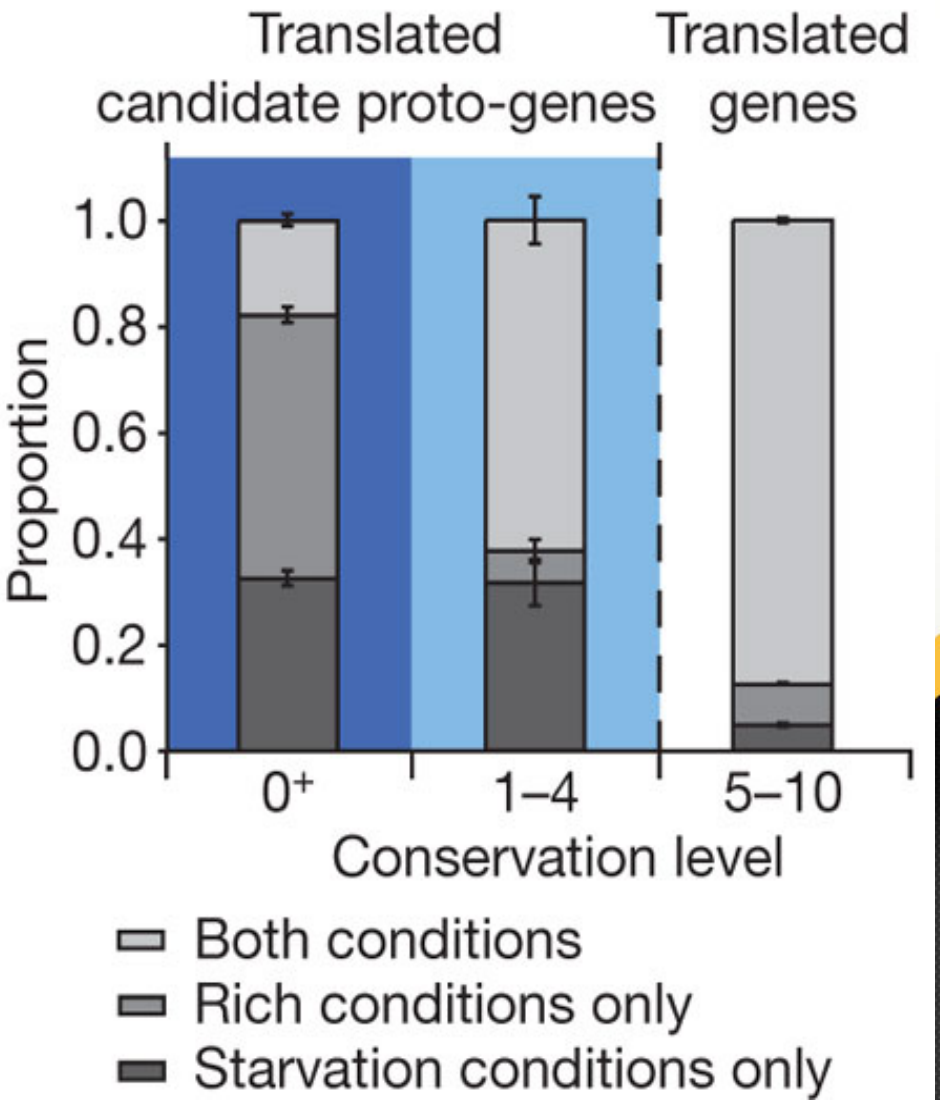

SBAY AAAATTAGCATTCGTATTGATGCTTTAAATTTTGCAAACATGTTGTAACCT---
SMIK AAAATAAACTAATGTGAAGTT--TTCAATTGTT-GAAGTCCACCGGAAGTTAGC
SPAR GAAGTAAGCCAATGTGATGAT-TTCCGATTATTCGAGGTTTGCCAAAACTTAAC
SCER ATGGTAAATCAATGTGTTATT-TTCCGATTGTTTGA----------
SBAY -AAACTGGAGTTCTGTTTAGACAAGATA-TTTTAGTCTCTTCTTCCCG------
SMIK AAAACCGGAGCTC----------ATGTTTTCAGTTCCTTACACTCCAAATTA
SPAR AAGACGGCATTTT----------ATG-TTTCAGTTTTTTATTTTCAAAGCGG
SCER -AGACAGTATTTT----------ATG-TCTCAATTTCTCATTTTCGAACTAA
Coverage (in number of reads)
15
10
5
0
ORF0+
RNA-Seq
Ribosome footprint
137,650
137,600
137,550
137,500
137,450
137,400
Position along chromosome 11 (in nucleotides)

Results of mapping ribosome footprinting data

- They found that 1,139 of 108 000 ORFs0 show such evidence of translation (ORFs0+)
- We verified that ORFs0+ did not originate from gene duplication or cross-species transfer and are not genes that have failed to be annotated due to their short length
- The 1,139 ORFs0+ therefore appear to be translated non-genic ORFs

Different translation in both contitions

- We detected strong differential translation of ORFs0+ and ORFs1–4 in starvation or rich conditions, whereas most ORFs5–10 are translated in both conditions

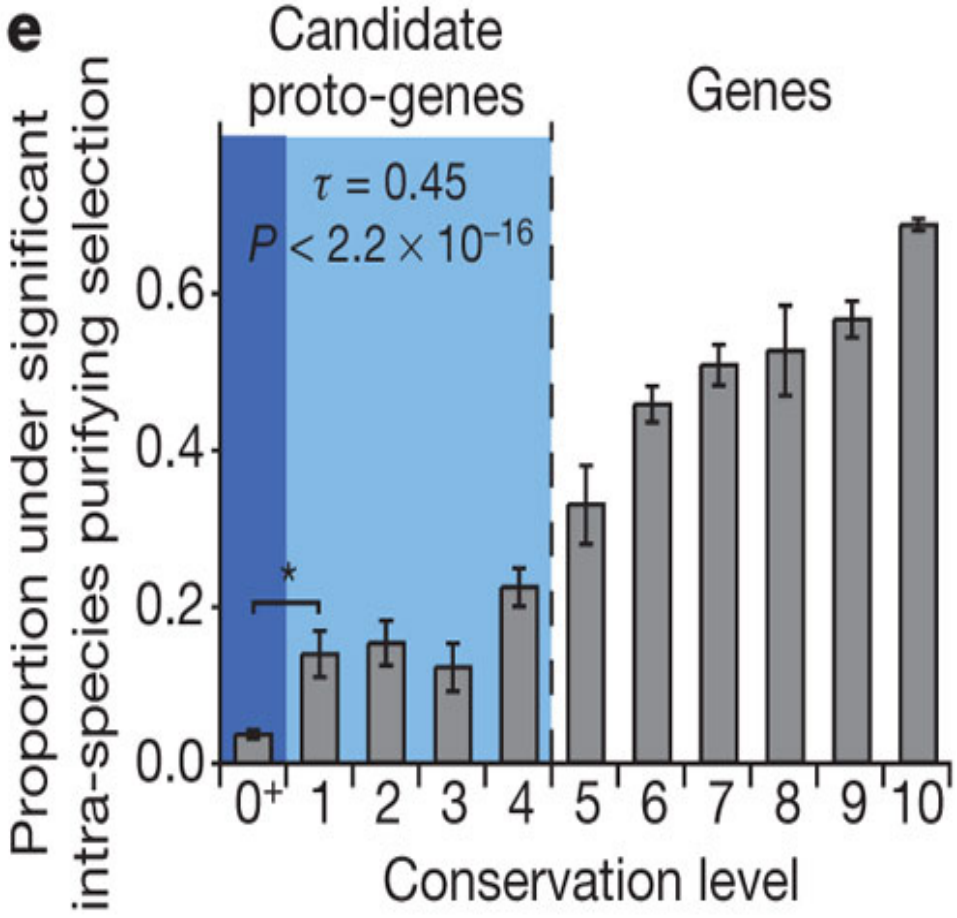


Purifying selection

- Most of ORFs0+ and ORFs1-4 do not exhibit a significant deviation from neutral evolution.
- But ~3% of ORFs0+ and 9-25% of ORFs1-4 appear under purifying selection.
- This fraction increases with conservation level,in line with proposed evolutionary continuum

Purifying selection

Our observations suggest that recently emerged ORFs occasionally acquire adaptive functions that are retained by natural selection, in agreement with findings in primates and with evolutionary models derived from inter-species comparisons



Conclusion

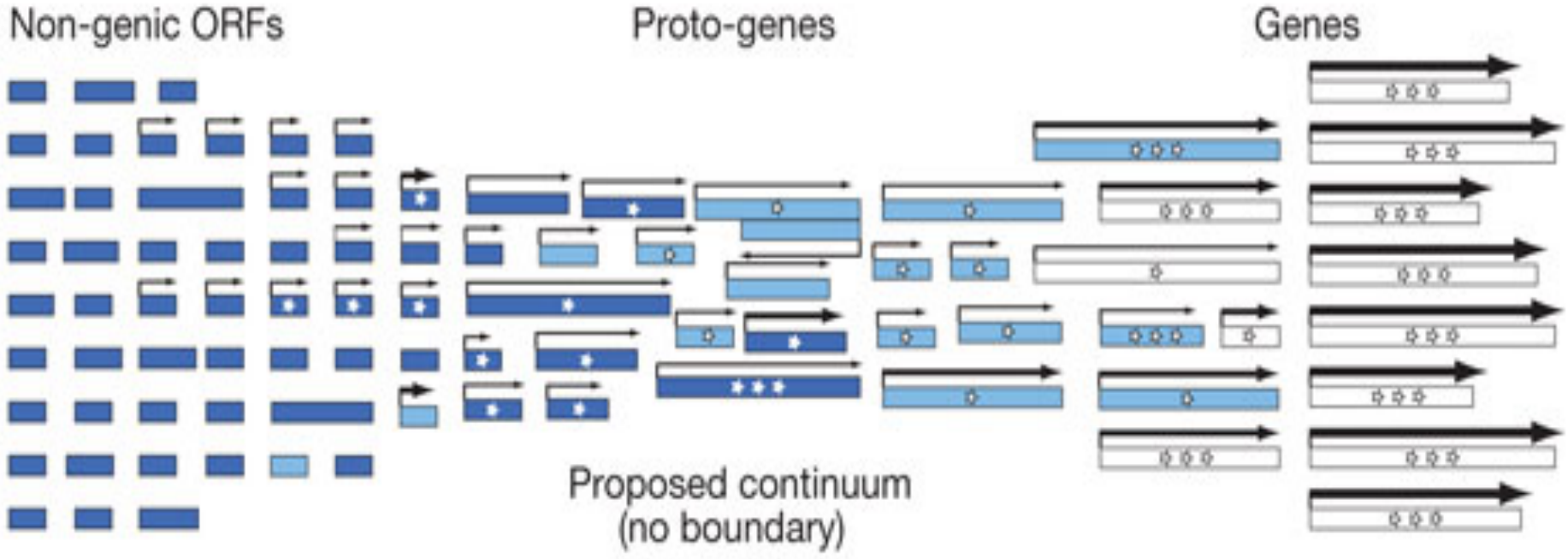
- Overall, our results show that de novo gene birth could proceed through proto-genes.

- The remaining 1891 (1139 ORFs0+ + 752 ORFs1-4 - 25 ORFs4 that are thought to be genes) presents characteristics intermediate between non-genic ORFs genes, meeting our proto-gene designation.

Conclusion

They proposed to place these ORFs
in a continuum where strict annotation
boundaries no longer have to
be set

Conclusion


Non-genic ORFs
Proto-genes
Genes
Proposed continuum
(no boundary)

Conclusion

- Gene birth by re-organization of pre-existing genes, duplication – have long been regarded as the predominant source of evolutionary innovation, but..

- *S. cerevisiae* vs *S. paradoxus* (since split)
  - Duplication 5 novel genes
  - 19 of 143 ORFs1 arose *de novo*

Conclusion

Therefore, *de novo* gene birth seems to be more prevalent than previously supposed

The End

Thank You For Listening