Synonymous Codon Usage

Laurent Duret
Laboratoire de Biométrie et Biologie
Evolutive, CNRS, Université Lyon 1

Synonymous Codon Usage

- Synonymous substitutions: neutral?
- E.g. Nematode (~19,000 genes, 7,000,000 codons)

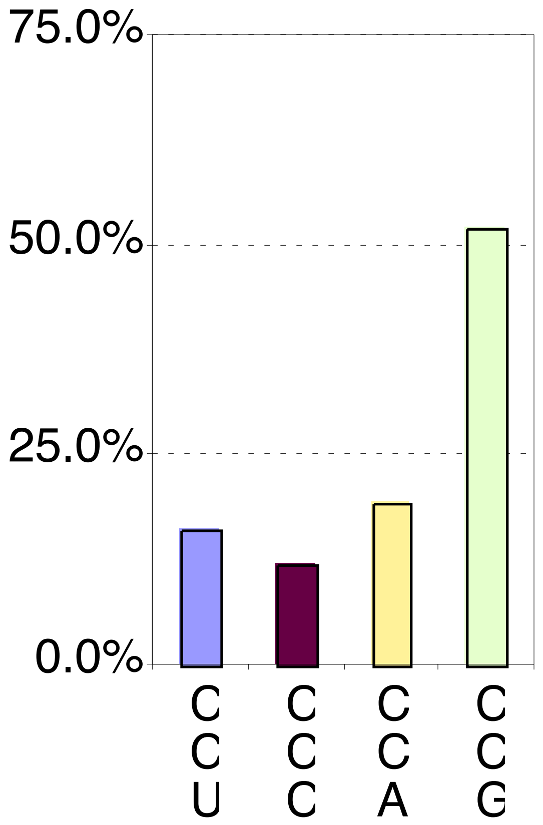
CCT -> CCC
Pro Pro

Silent mutation? Fitness impact ?

Synonymous codon usage

Ikemura, Gautier, Gouy, Grantham, 1980...

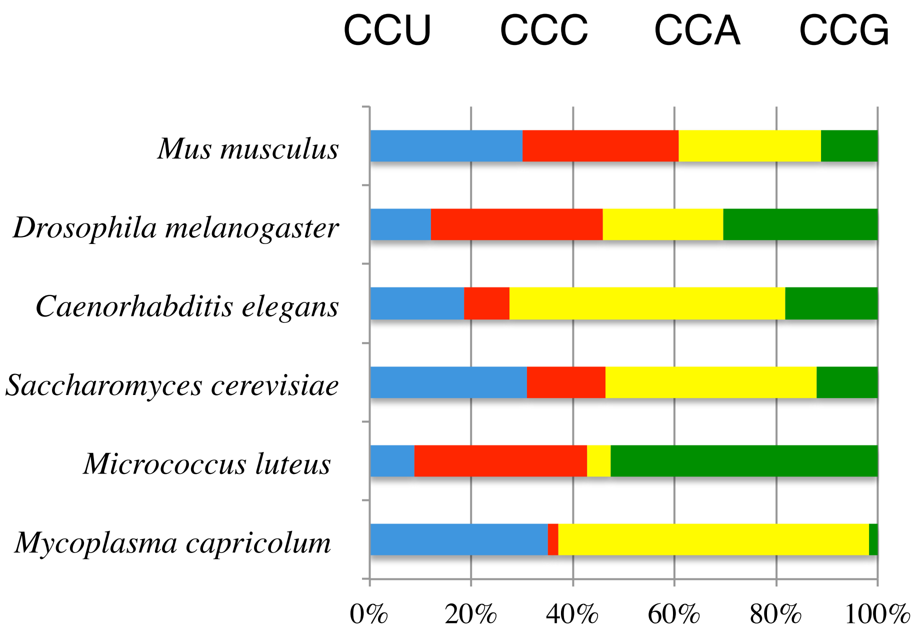
- 61 codons, 20 amino-acids: degenerascy of the genetic code
- Non-random synonymous codon usage: some synonymous codons are preferentially used.
- Synonymous codon usage bias
- Example: frequency of proline codon in *Escherichia coli* genome (4300 genes)



Synonymous codon usage varies ...

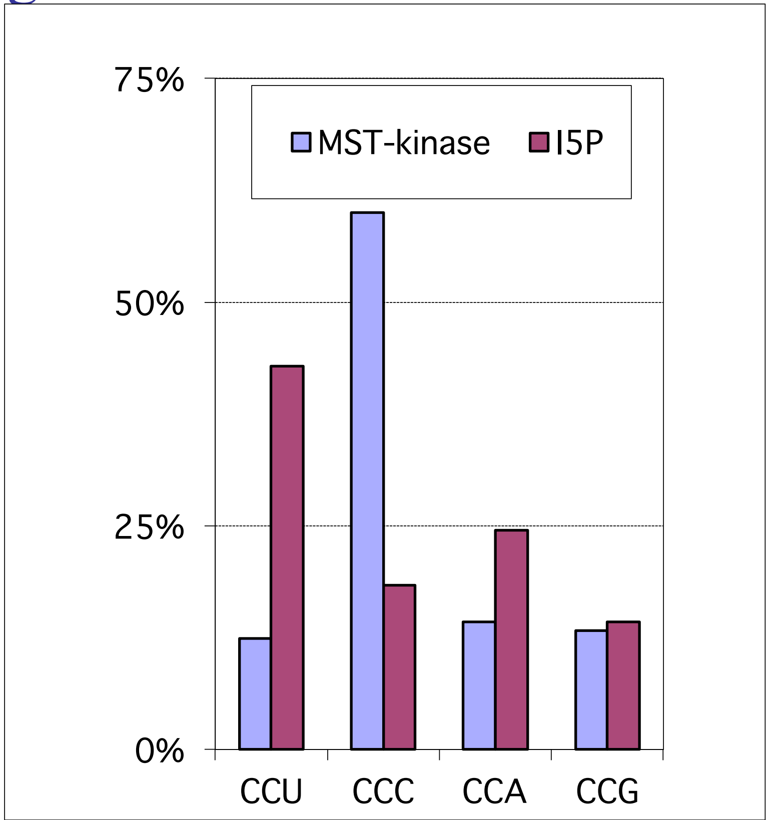
- … among species

- Example: proline codon usage in different species



- … among genes within a genome.

- Example: proline codon usage in different human genes



How to explain synonymous codon usage biases ? Neutralist and selectionist models

- Selection for translation efficiency

- Neutral substitution bias :
  - Mutational bias
  - gBGC

Equilibrium codon frequency (1)

- Lys: 2 synonymous codons: AAG, AAA
- Codon frequency depends on relative substitution rates:

$$G \underset{u}{\overset{v}{\rightleftarrows}} A$$

At equilibrium: Frequency codon AAG = $u / (u + v)$

Equilibrium codon frequency (2)

- Lys: 2 synonymous codons: AAG, AAA
- Codon frequency depends on relative substitution rates:

$$\begin{tikzcd} G \arrow[r, shift left, "v = 2N \times \mu_{GA} \times P(A)"] & A \arrow[l, shift left, "u = 2N \times \mu_{AG} \times P(G)"] \end{tikzcd}$$

μ_{GA} : mutation rate G->A (per bp per generation)
P(A): fixation probability of allele A
N : population size

Neutral substitution bias (1)

- If no selection, no gBGC: $P(A)=P(G)=1/2N$

$$\begin{tikzcd}
G \arrow[r, shift left=1.5ex, "v = \mu_{GA}"] & A \arrow[l, shift left=1.5ex, "u = \mu_{AG}"]
\end{tikzcd}$$

At equilibrium: Frequency codon AAG $= \mu_{AG} / (\mu_{AG} + \mu_{GA})$

=> Mutational pressure (Sueoka, 1962)

Mutational pressure varies among species

- Direct measurement of mutation rates (sequencing of pedigrees, mutation accumulation lines)
- ~20 species (bacteria, eukaryotes)
- *Paramecium tetraurelia*: $\mu_{AG} / (\mu_{AG} + \mu_{GA}) = 0.07$
- Human: $\mu_{AG} / (\mu_{AG} + \mu_{GA}) = 0.32$
- *E. coli*: $\mu_{AG} / (\mu_{AG} + \mu_{GA}) = 0.43$

- => differences in mutational pressure can contribute to differences in codon usage among species

Lynch *PNAS* 2010, Sun et al. *PNAS* 2012

Neutral substitution bias (2)

- If no selection, but gBGC: $P(G) > P(A)$

$$\begin{tikzcd} G \arrow[r, shift left, "v = 2N \times \mu_{GA} \times P(A)"] & A \arrow[l, shift left, "u = 2N \times \mu_{AG} \times P(G)"] \end{tikzcd}$$

Variation in gBGC intensity can contribute to difference in codon usage among species and within genomes (variation in recombination rate along chromosomes)

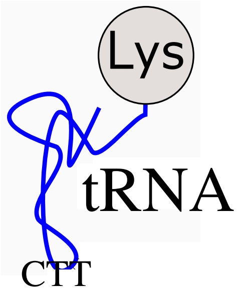
Selection on codon usage

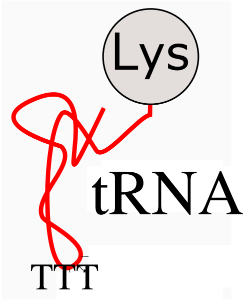
- If selection: $P(G) \neq P(A)$

$$\begin{tikzcd}[column sep=huge]
G \arrow[r, shift left, "v = 2N \times \mu_{GA} \times P(A)"] & A \arrow[l, shift left, "u = 2N \times \mu_{AG} \times P(G)"]
\end{tikzcd}$$

Selection for translation efficiency (translational selection) (1)

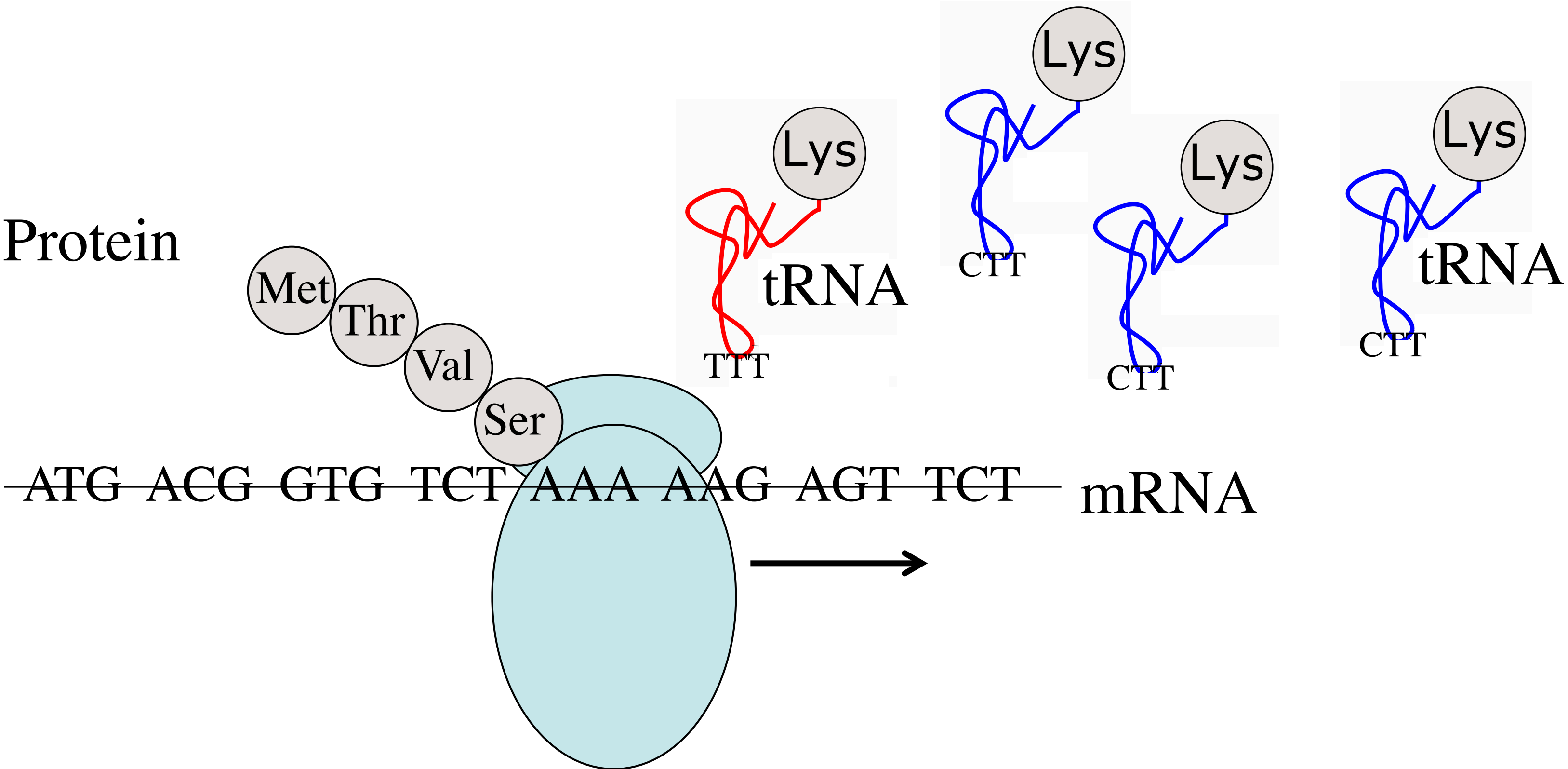
- Some amino-acids are encoded by several synonymous codons, recognized by different tRNAs
- E.g. : Lys
  - 2 synonymous codons: AAG, AAA
  - 2 tRNAs:
    - Anticodon CTT
    - Anticodon TTT





Selection for translation efficiency (translational selection) (2)

- The speed and accuracy of translation of codons depends on the abundance of their corresponding tRNA



Selection for translation efficiency (translational selection) (3)

- Optimal codons = codons corresponding to the most abundant tRNAs
- *Fop*: frequency of optimal codons
- Genes with high *Fop* are translated more accurately and more rapidly

Selection for translation efficiency (translational selection) (4)

- The number of ribosomes present within a cell is a limiting resource
- Highly expressed genes mobilize a large number of ribosomes
- => selective pressure to optimize translation speed in highly expressed genes

How to explain synonymous codon usage biases ?
Neutralist and selectionist models

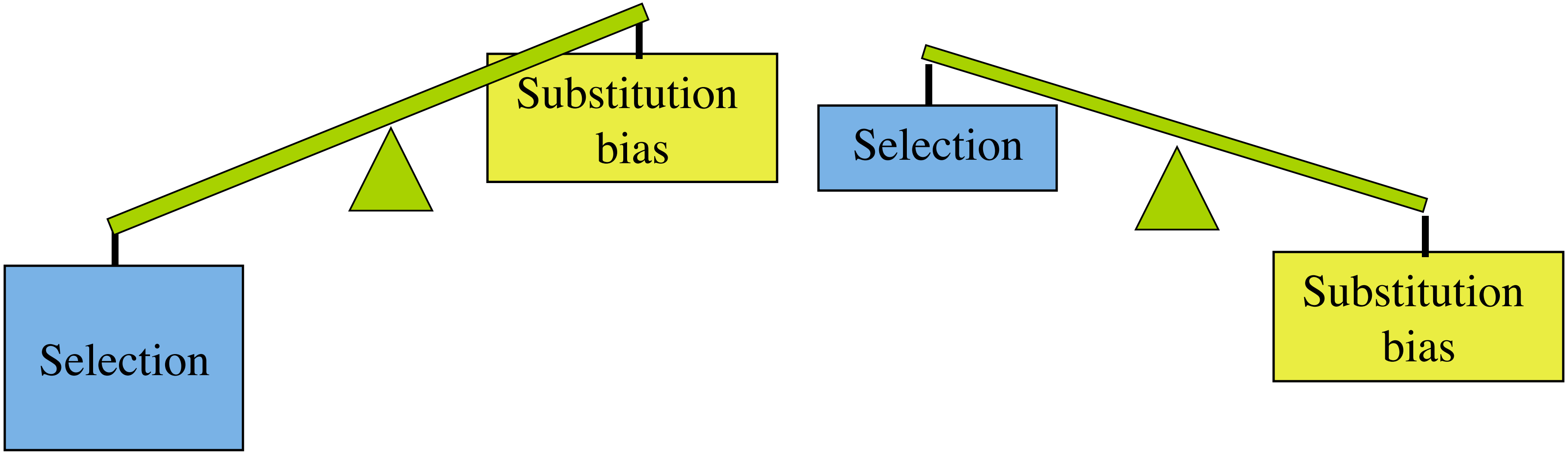
- Selection of codons that are optimal for translation efficiency
- Codon usage should correlate with gene expression level
- Preferred codons in highly expressed genes should correspond to the most abundant tRNAs

- Neutral substitution bias :
  - Mutational bias
  - gBGC
- No relationship with gene expression level
- Substitution biases affect all positions within a genome (not only synonymous codon positions) ➪ correlation between codon usage and genome base composition

Balance mutation-drift-selection-gBGC

Codon usage biases in unicellular organisms

- *Escherichia coli*, *Bacillus subtilis*, yeast
- *Borrelia burgdorferi*, *Mycobacterium tuberculosis*

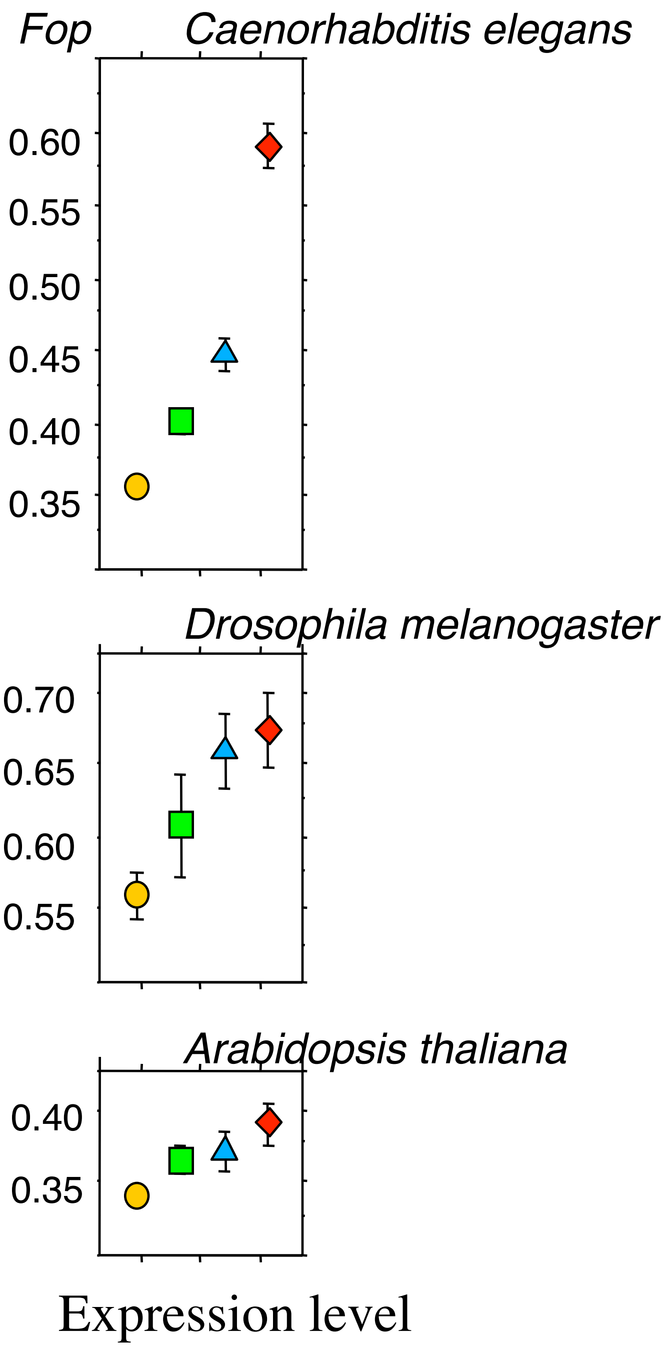


Grantham, Gouy, Gautier (1980…), Ikemura (1980…), Kurland, Bulmer, Sharp, ...

Codon usage biases in pluricellular organisms : selection or neutral substitution bias ?

- Analysis of the relationship between codon usage and gene expression
  - Nematode
  - Drosophila
  - Arabidopsis thaliana
- Transcriptome data:
  - Sanger sequencing of cDNA clones (ESTs)
  - Low-coverage (1999!)

Frequency of optimal codons (*Fop*) and gene expression level



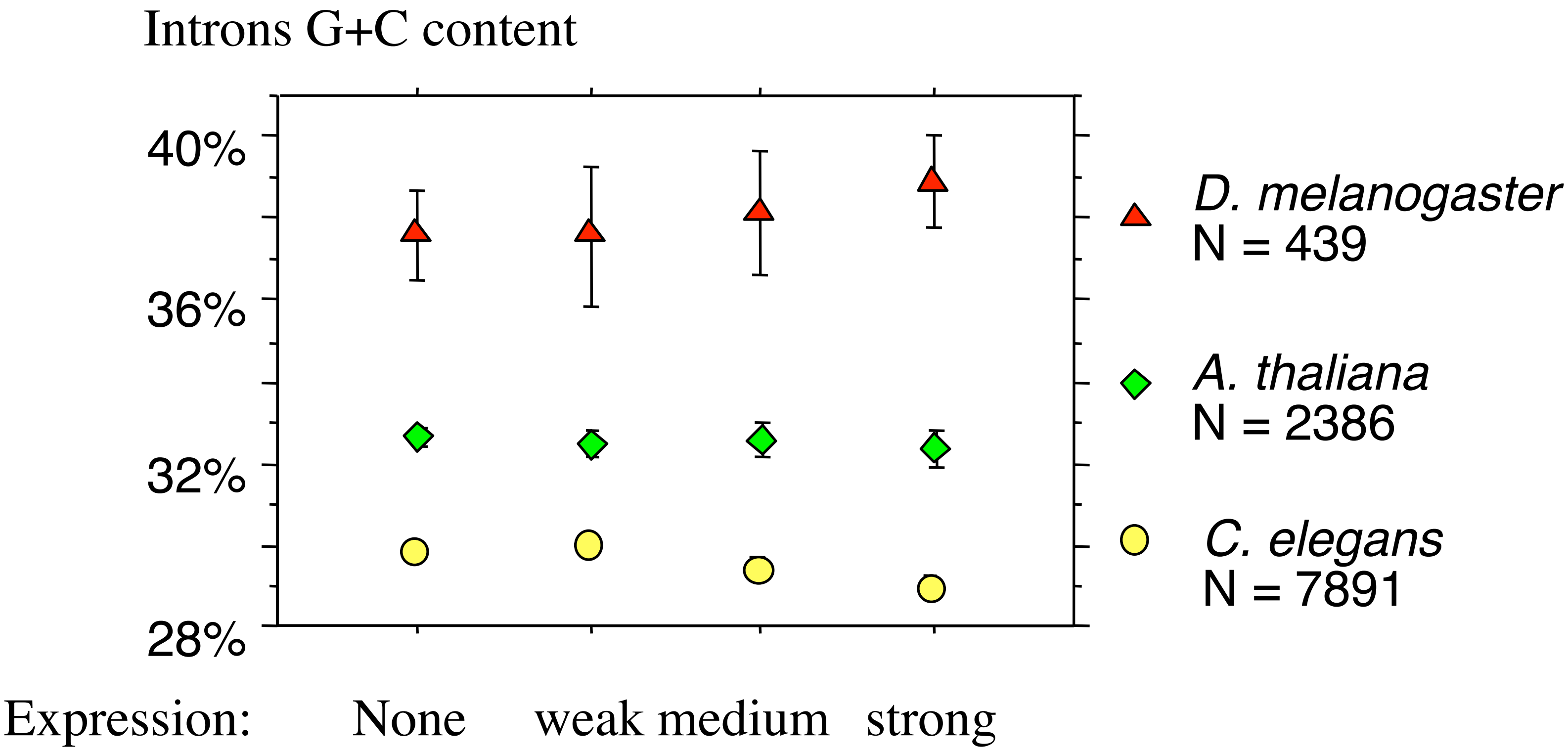
Expression level:

- 🟡 No mRNA detected
- 🟩 Weak
- 🔺 Medium
- 🔶 Strong

Duret & Mouchiroud *PNAS* 1999

Mutational bias or selection ?

In drosophila, *C. elegans* and *A. thaliana,* most optimal codons end in C or G.
Mutational bias toward C and G in highly expressed genes ?



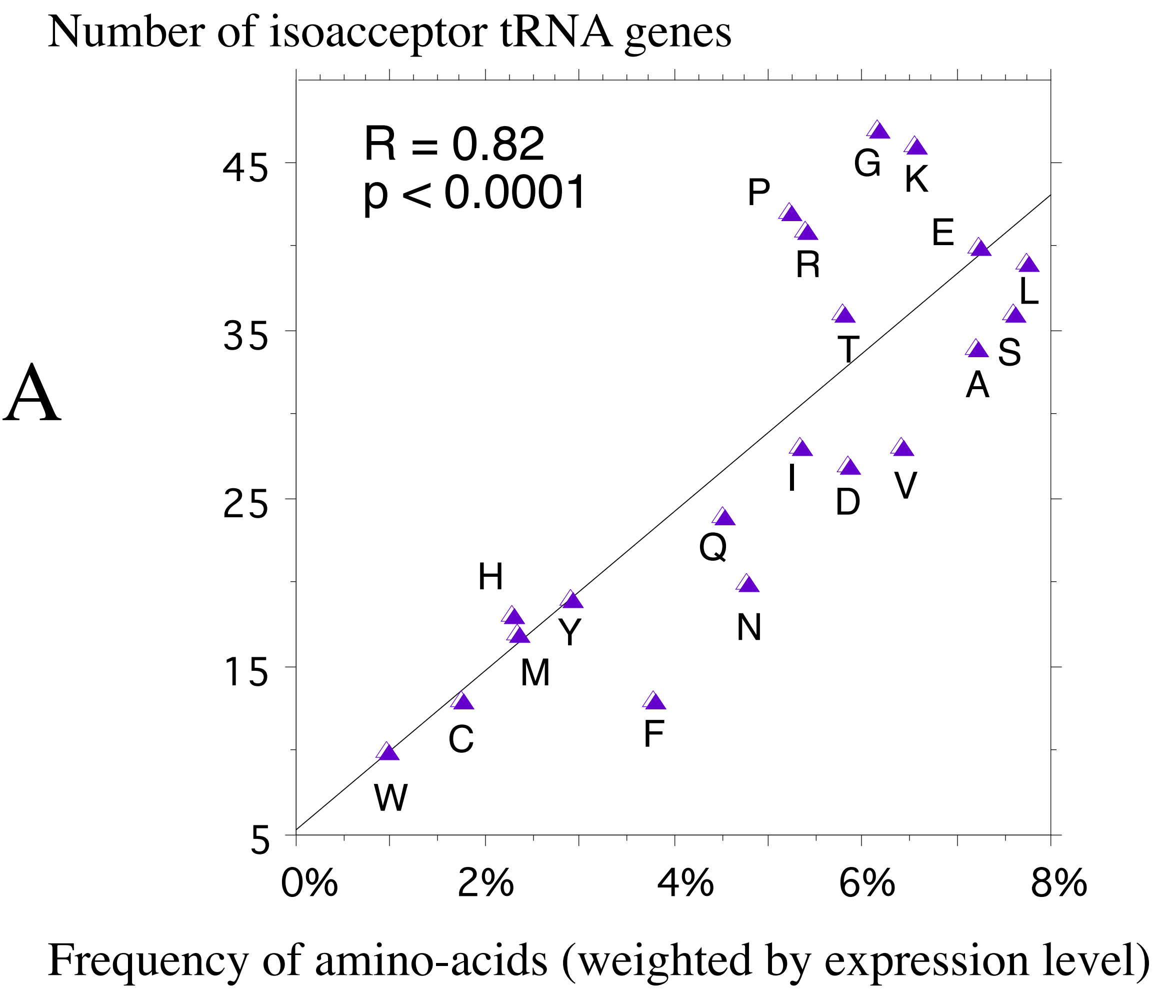
Correlation between synonymous codon usage and tRNA abundance ?

- Nematode:
  - Complete genome: 580 tRNA genes (10 to 46 copies per family of isacceptor tRNA)

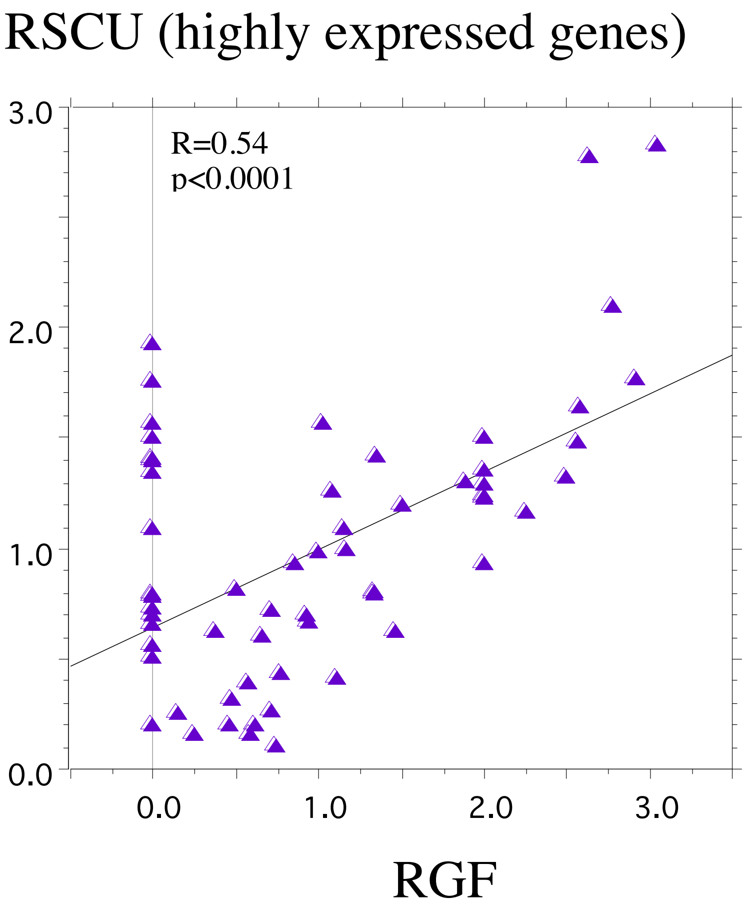
Number of tRNAgenes: indicator of tRNA abundance within the cell ? (bacteria: Ikemura, 99 yeast: Percudani, 97)

Relationship between the number of tRNA genes and the frequency of amino-acids in *C. elegans* proteins

- 580 tRNA genes



Correlation between the relative synonymous codon usage (RSCU) and the relative frequency of tRNA genes (RGF)



Example: 41 tRNA-Pro genes

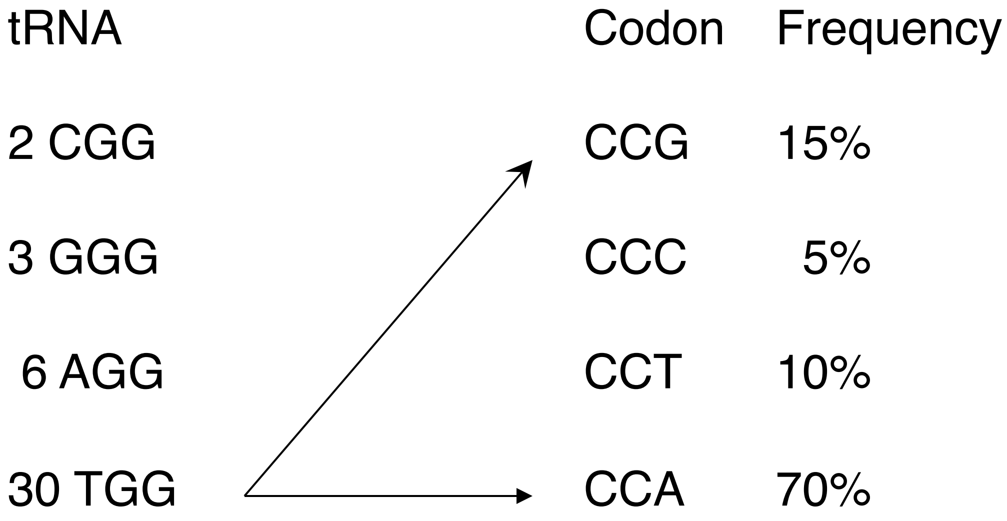
tRNA	N.	Frq.	RGF
CGG	2	5%	0.2
GGG	3	7%	0.3
AGG	6	15%	0.6
TGG	30	73%	2.9
	41	100%	4

Codon	Frq.	RSCU
CCG	15%	0.6
CCC	5%	0.2
CCT	10%	0.4
CCA	70%	2.8
	100%	4

Duret (2000) Trends Genet

tRNA / codon pairing (wooble)

Example: proline



tRNA	Codon	Frequency
2 CGG	CCG	15%
3 GGG	CCC	5%
6 AGG	CCT	10%
30 TGG	CCA	70%

In all cases (but Gln), optimal codons are decoded by the tRNA having the highest copy number in the genome (Duret, 2000)

Synonymous codon usage in pluricellular eukaryotes

- Nematode, drosophila, arabidopsis:

CCT → CCC
Pro Pro

Phenotypic impact ~~?~~

Translation efficiency
- speed
- accuracy (Akashi 1994, Marais & Duret, 2001)

Synonymous codon usage in mammals

- Selectionist/neutralist controversy
- Neutralists:
  - Kanaya et al. (2001) *JME*, Duret (2002), Sémon et al (2004) *Hum Mol Genet*, dos Reis et al (2004) *NAR*, Sémon et al (2006) *Mol Biol Evol*
- Selectionists:
  - Plotkin et al. (2004) *PNAS*, Kudla et al. (2006) *Plos Biol*, Gingold et al. (2014) *Cell*

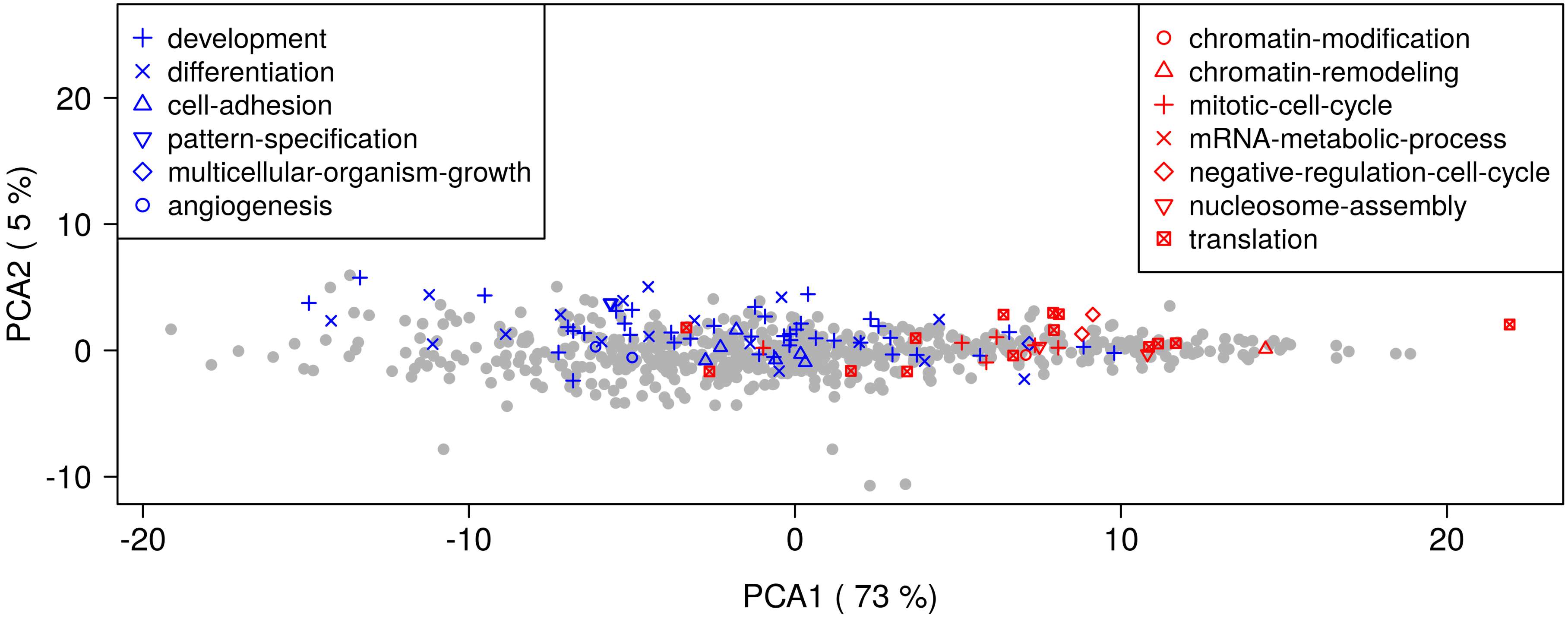
Synonymous codon usage in humans (Gingold et al. 2014 *Cell*)

- N=19,766 protein-coding genes
- Analysis of genes involved in different functions
  - 687 GO categories with > 40 genes
  - Codon usage of each GO gene set
- Principal Component Analysis
- Comparison of gene categories involved in differentiation or proliferation

Synonymous codon usage in humans (Gingold et al. 2014 *Cell*)

Differentiation

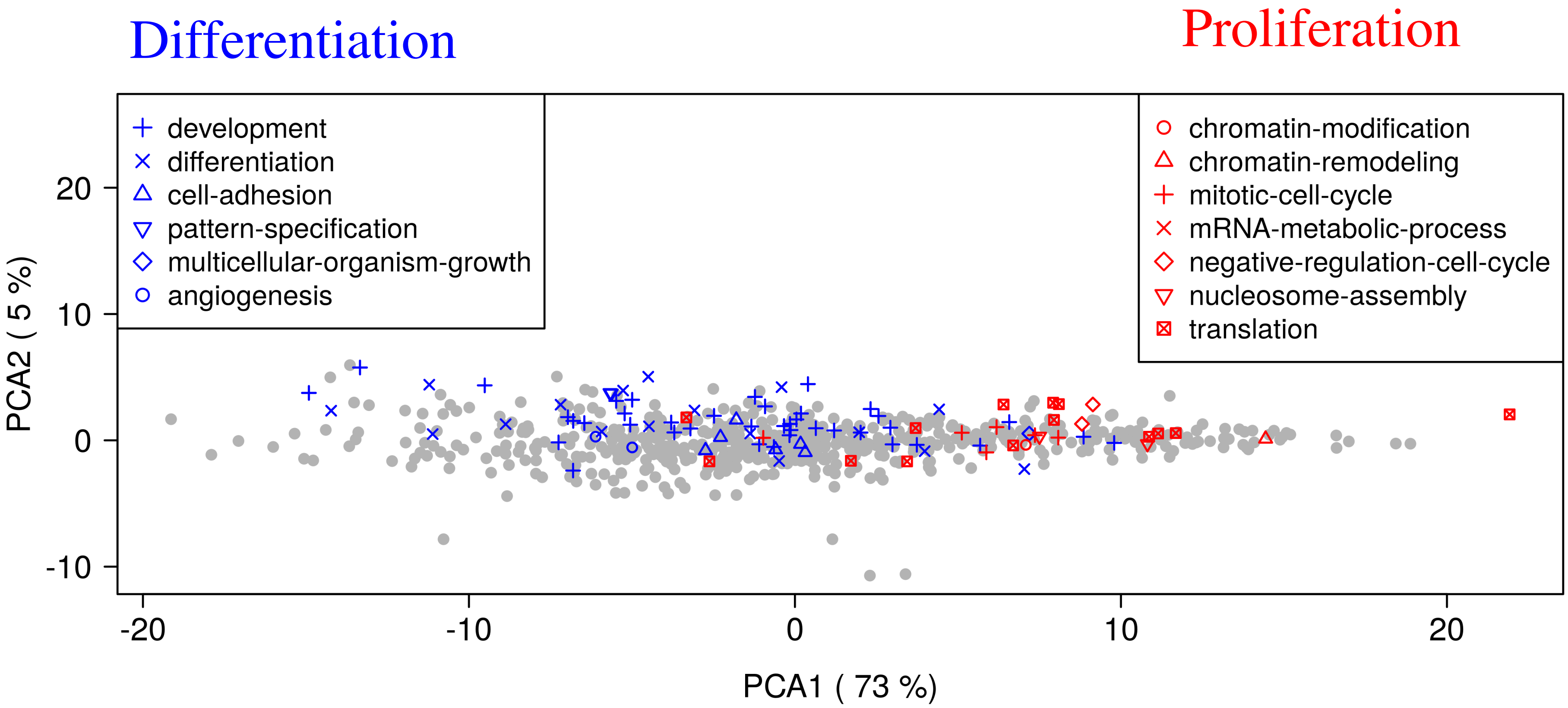
Proliferation



Synonymous codon usage in humans (Gingold et al. 2014 *Cell*)

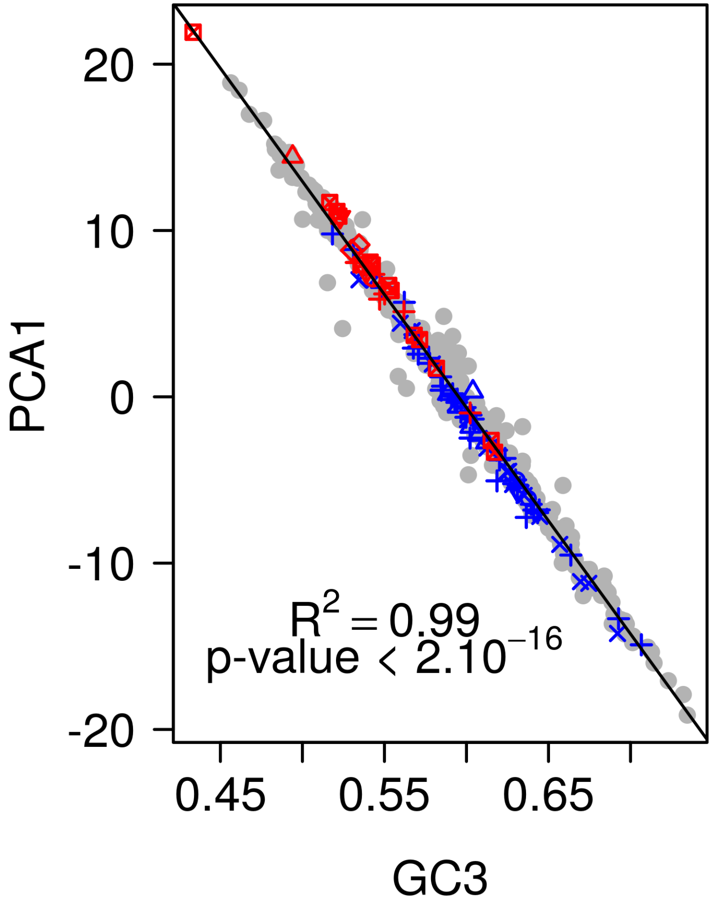
- Differences in synonymous codon usage between genes involved in cell differentiation vs. cell proliferation
- Variation in tRNA abundance during differentiation
- Conclusion: co-adaptation of tRNA abundance and synonymous codon usage to fine-tune the expression of genes involved in cellular differentiation

Why does synonymous codon usage vary among functional categories?



What are the variables that correspond to PCA1?

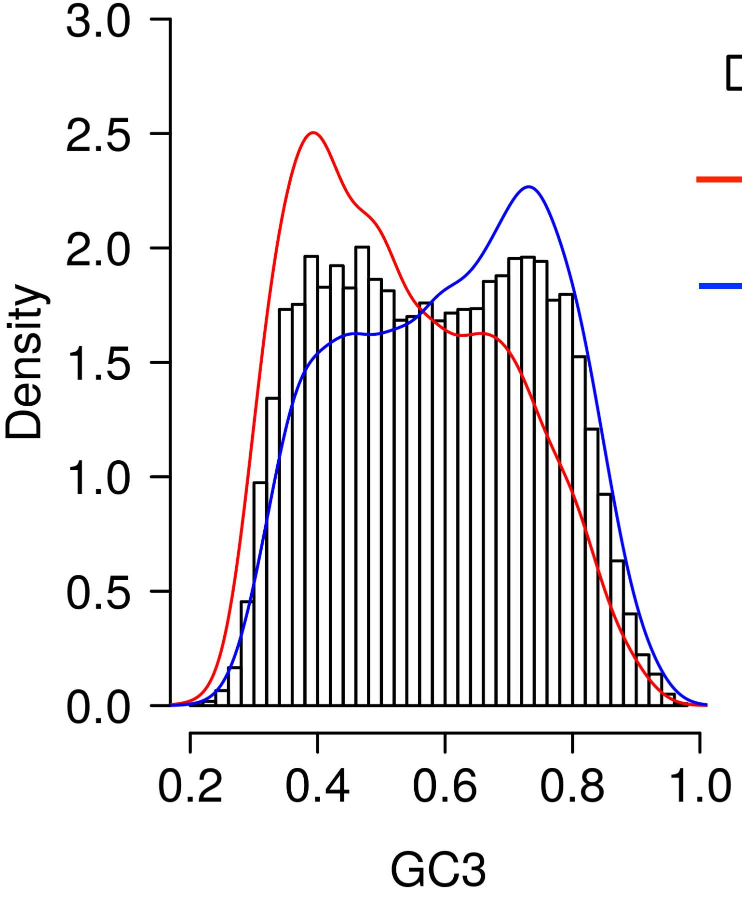
In humans, variation in codon usage corresponds to variation in GC-content at 3rd codon position



- N=687 GO gene sets

! Points are not independent !

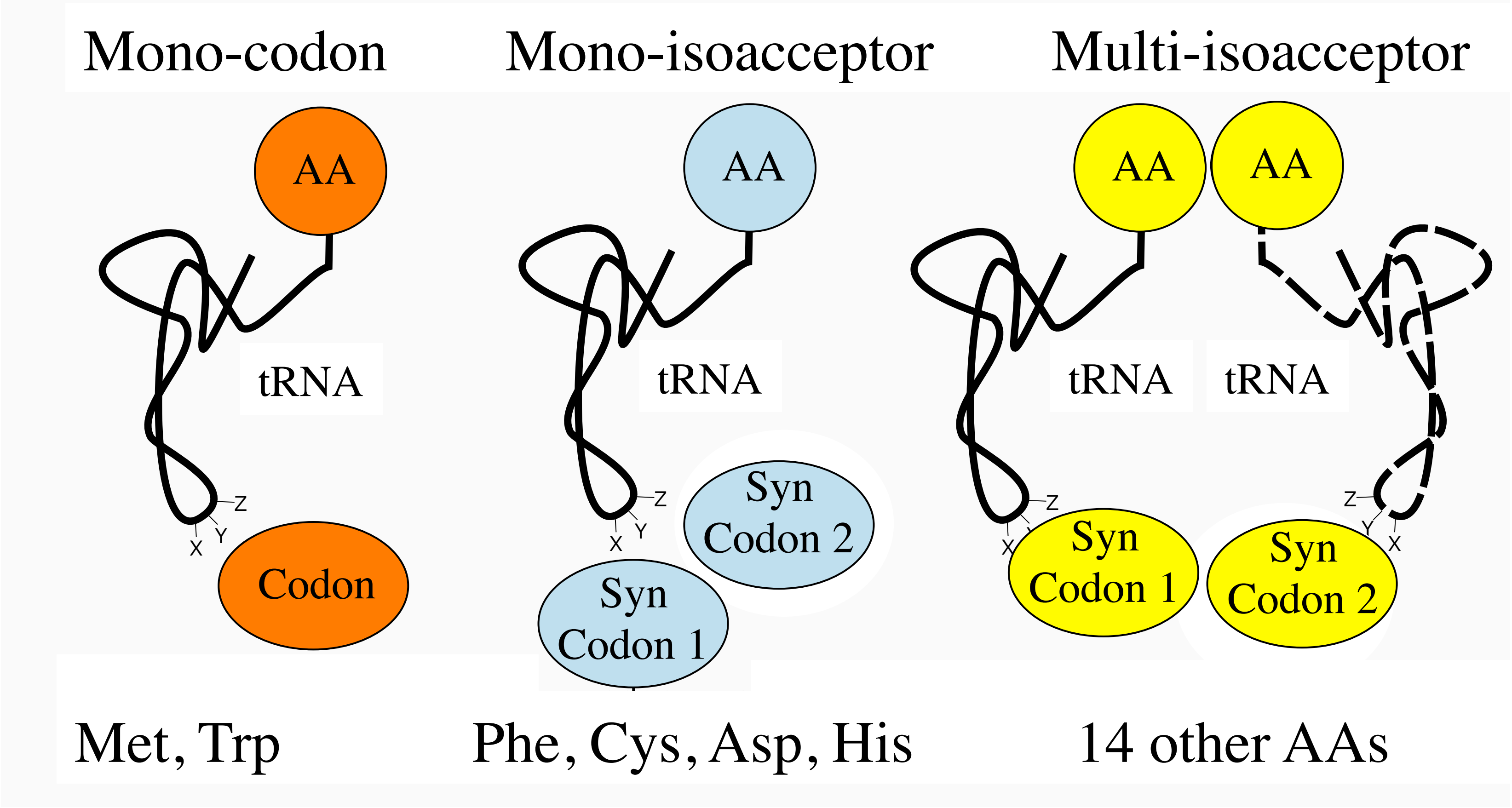
In humans, variation in codon usage corresponds to variation in GC-content at 3rd codon position



=> Differences in GC3 between "Proliferation" and "Differentiation" genes

Selection for translation efficiency?

- Three sets of amino-acids



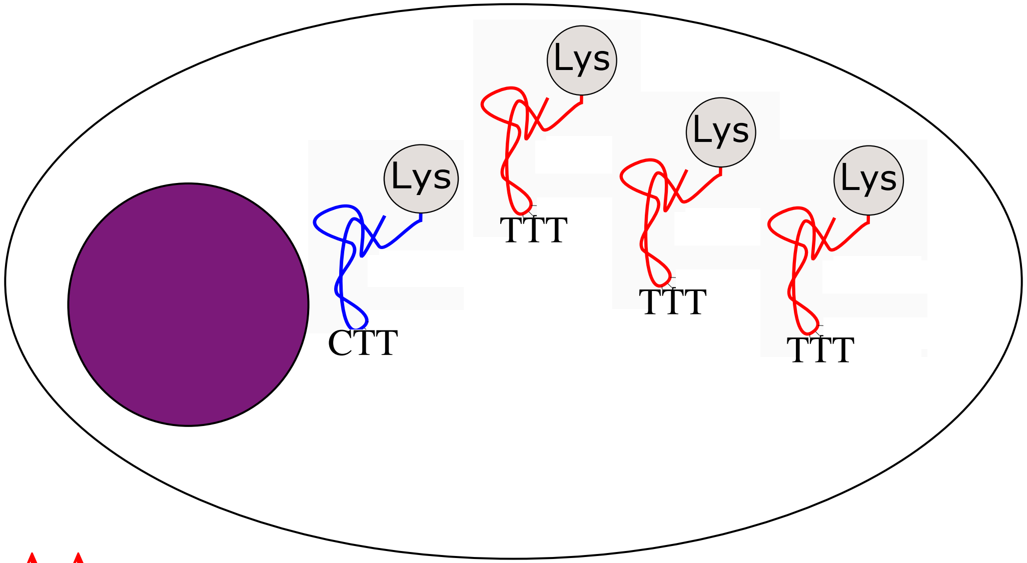
Mono-codon: Met, Trp

Mono-isoacceptor: Phe, Cys, Asp, His

Multi-isoacceptor: 14 other AAs

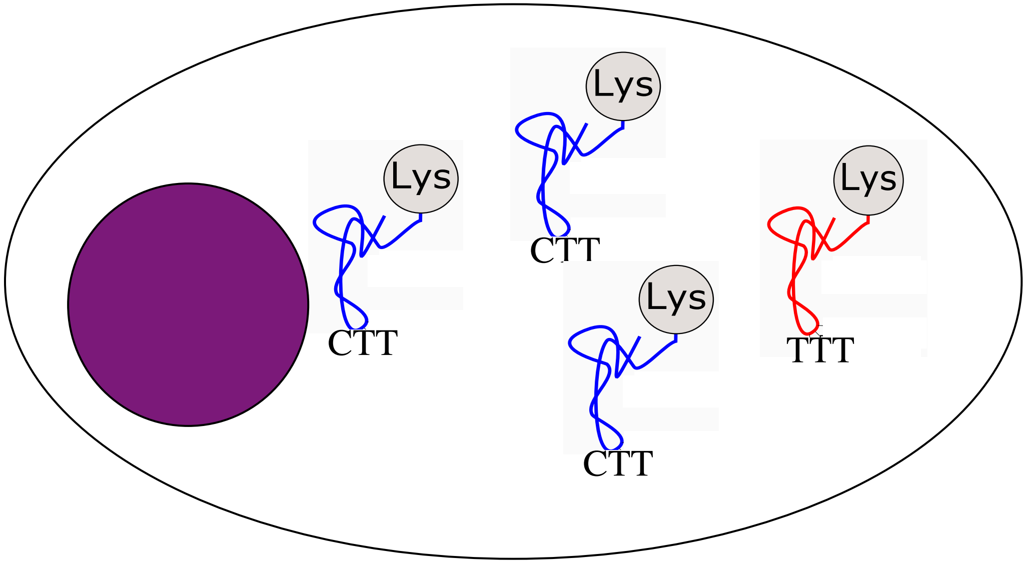
Selection for translation efficiency ?

Proliferation



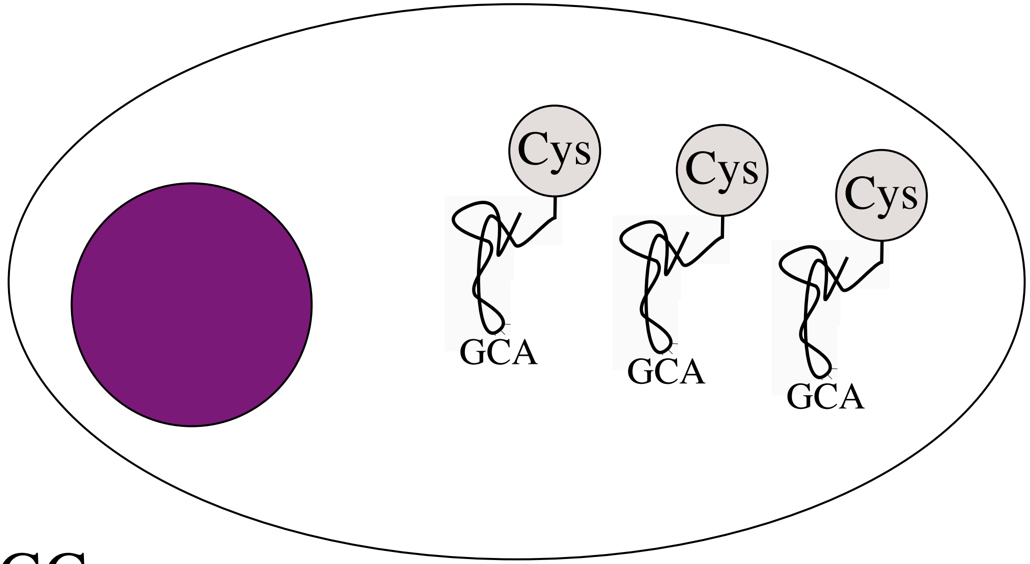
Lys codons: AAG, AAA

Differentiation



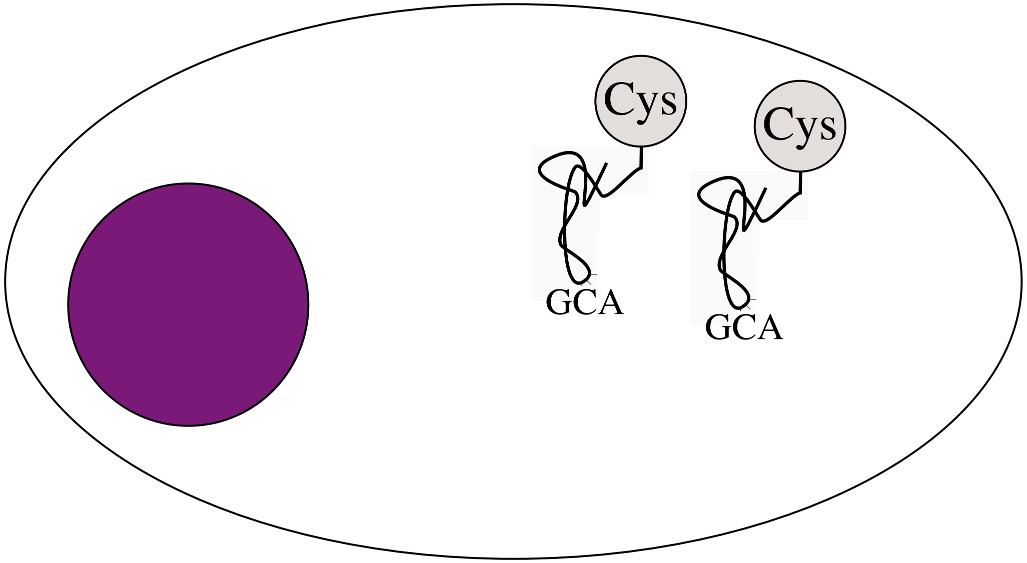
Selection for translation efficiency ?

Proliferation



Cys codons: TGT, TGC

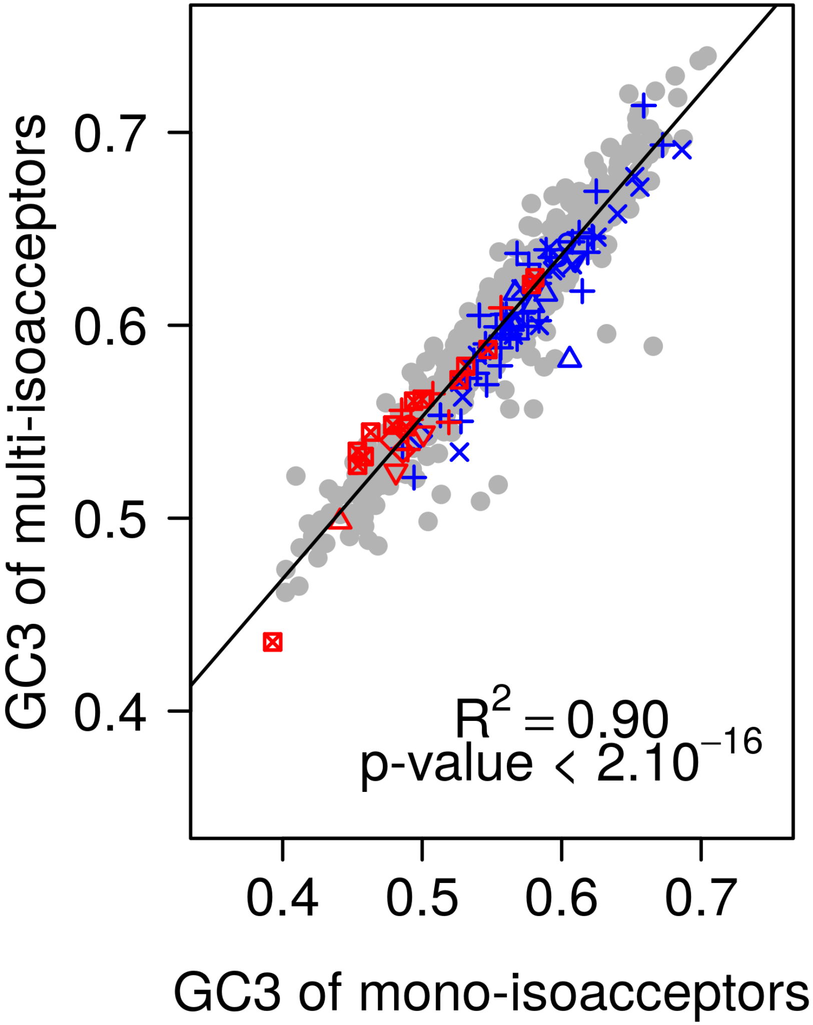
Differentiation



Selection for translation efficiency ?

- If variations in synonymous codon usage between "proliferation" and "differentiation" genes were due to selection for translation efficiency, these variations should affect only codons of multi-isoacceptor amino-acids

Selection for translation efficiency ?



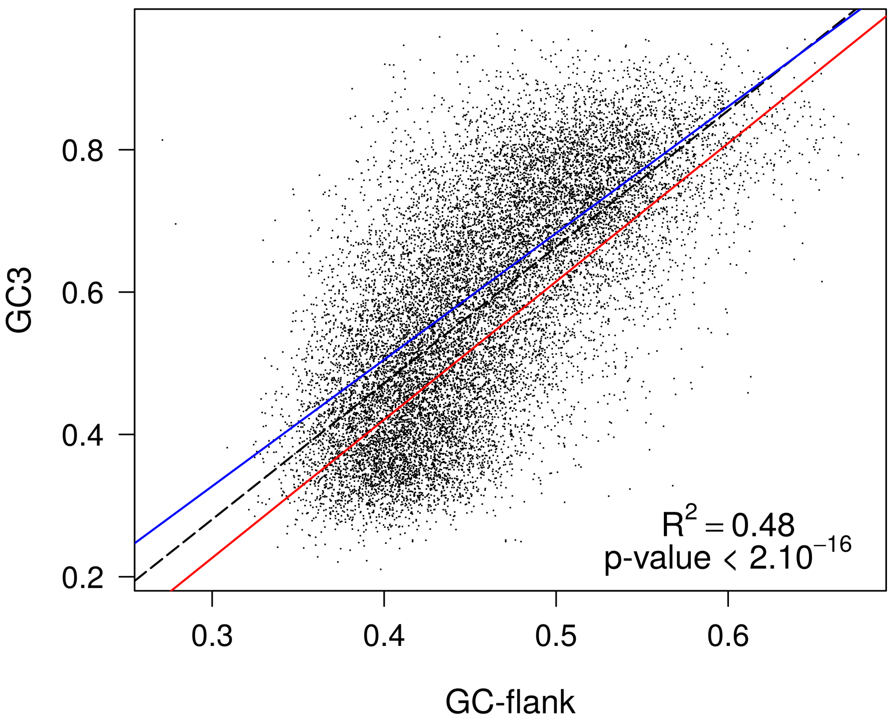
=> The process that drives differences in GC3 between "Proliferation" and "Differentiation" genes affects both mono- and multi-isoacceptor aminoacids

=> Not compatible with the hypothesis of selection for translation efficiency

gBGC ?

- Recombination rate (and hence gBGC intensity) vary along chromosomes
- => large-scale variation in GC-content along chromosomes, affecting all sites (intergenic, introns, exons)
- => correlation between GC3 and GC-content in flanking intergenic regions

gBGC ?

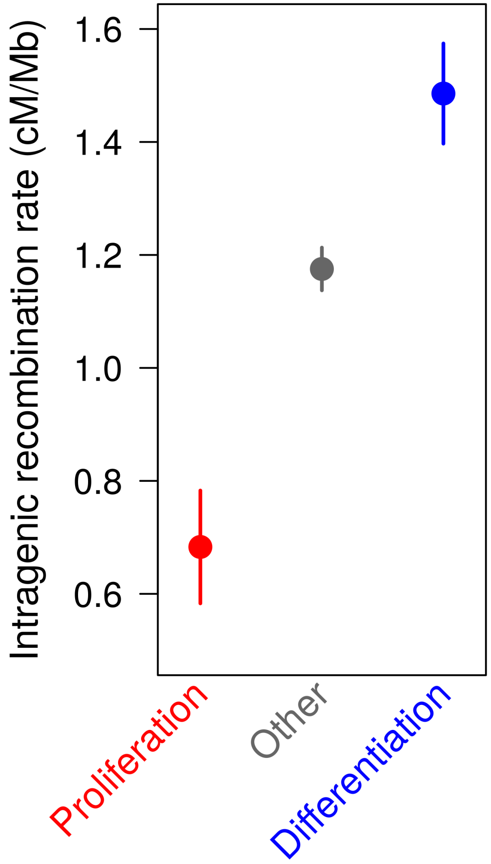


GC-flank = GC-content in flanking intergenic regions (10 kb upstream + 10 kb downstream)

- A large fraction of the variance in GC3 is explained by regional variations in GC-content (i.e. nothing to do with translation efficiency!)
- For a same GC-flank, GC3 "Proliferation" < GC3 "Differentiation"

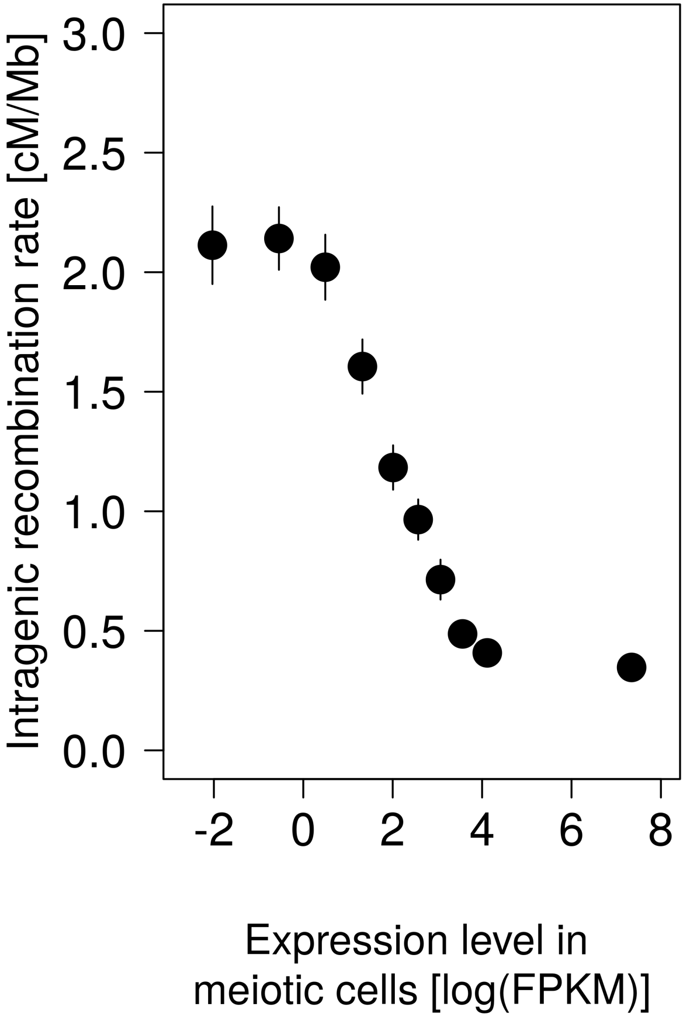
gBGC ?

- If the difference in GC3 is caused by gBGC, it should correlate with variation in recombination rate



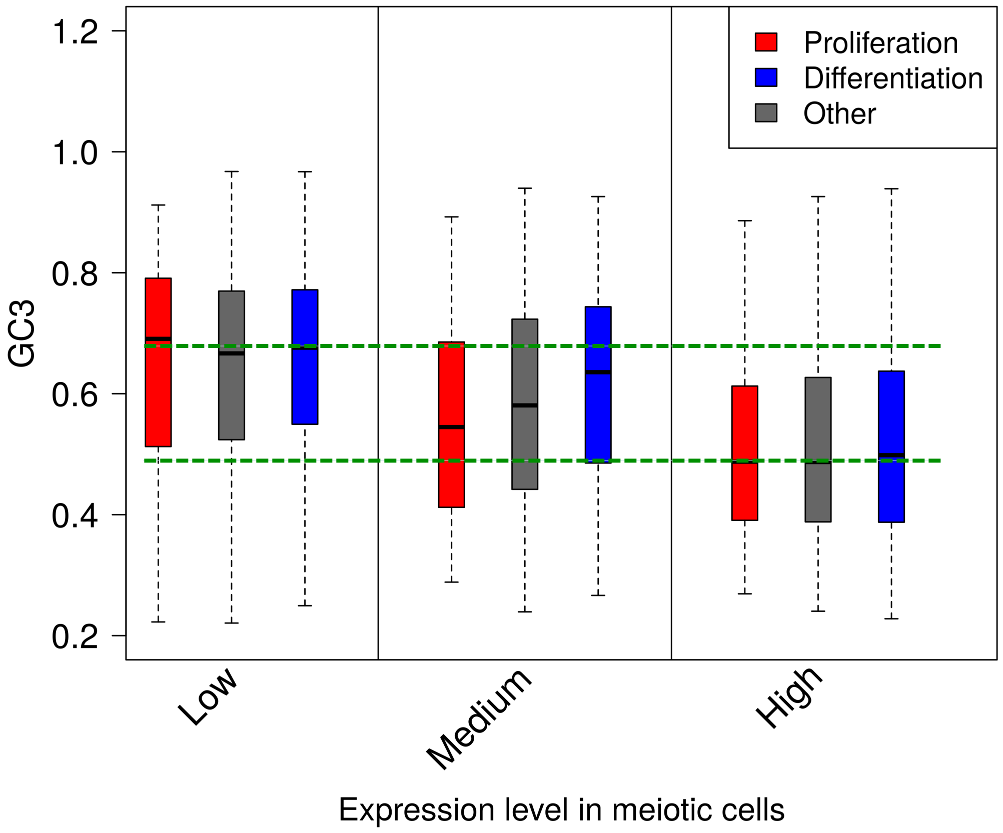
Why does recombination rate vary between "proliferation" and "differentiation" genes?

- McVicker & Green (2010):
  - the intragenic recombination rate correlates negatively with gene expression level in meiotic cells
  - Interference transcription/ recombination



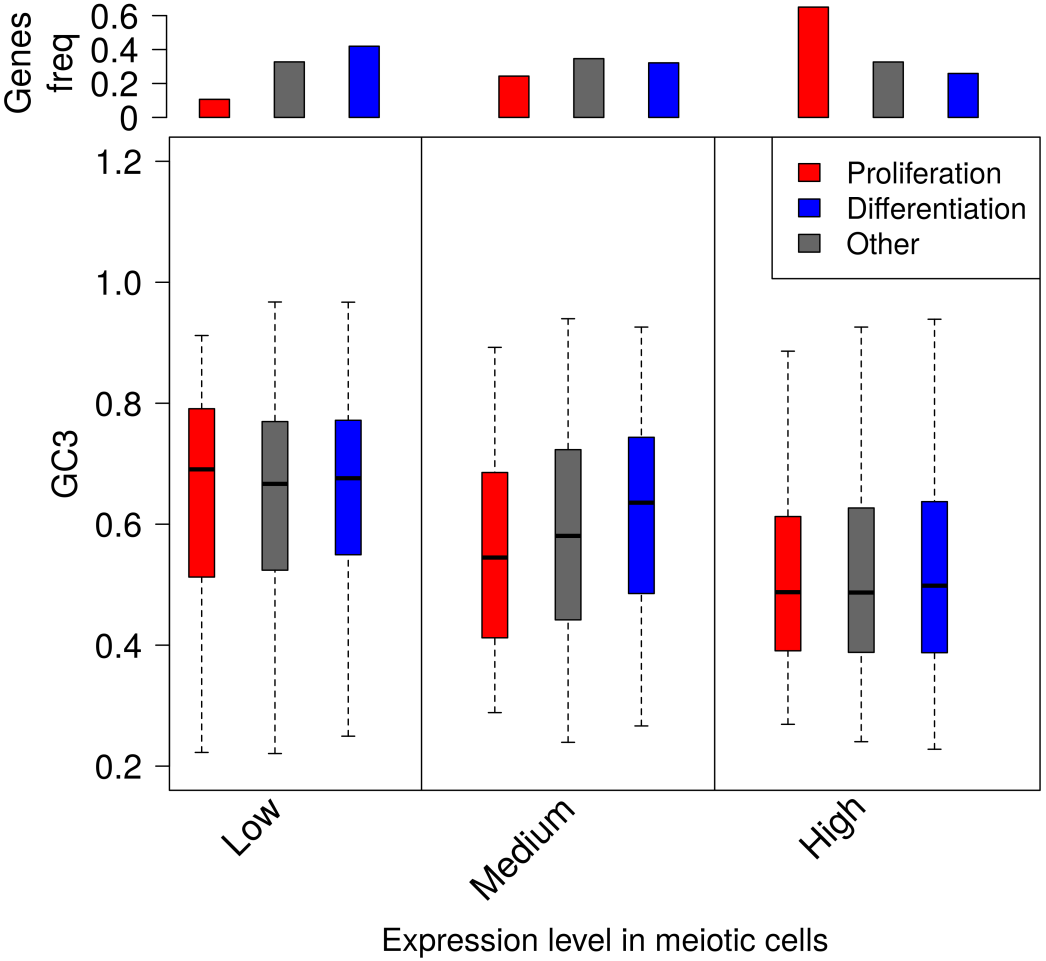
Does GC3 vary with meiotic expression level or with functional category ?

- Distribution of GC3



Does GC3 vary with meiotic expression level or with functional category ?

- Distribution of genes



Why does GC3 vary between "proliferation" and "differentiation" genes?

- "Proliferation" genes:
  - housekeeping genes, expressed in many tissues (including meiotic cells)
  - => low recombination rate
  - => low GC-content
- "Differentiation" genes:
  - generally tissue-specific
  - => low expression in meiotic cells
  - => higher recombination rate
  - => higher GC-content

What fraction of the variance in GC3 is explained by gBGC ?

- gBGC model:

  GC3 = f(long-term intragenic recombination rate)

- Proxies for long-term intragenic recombination rate:
  - Present-day intragenic recombination rate
  - Meiotic expression level
  - Intron GC-content

What fraction of the variance in GC3 is explained by gBGC ?

GC3 predictors	Pairwise R^2	p-value
GCi	62.7%	$<2.10^{-16}$

What fraction of the variance in GC3 is explained by gBGC ?

GC3 predictors	Pairwise R^2	p-value	Model R^2	F statistic	p-value
GCi	62.7%	$<2.10^{-16}$	62.7%	30232.4	$<2.10^{-16}$
+ GC-flank	48.1%	$<2.10^{-16}$	63.0%	126.8	$<2.10^{-16}$

What fraction of the variance in GC3 is explained by gBGC ?

GC3 predictors	Pairwise R^2	p-value	Model R^2	F statistic	p-value
GCi	62.7%	$<2.10^{-16}$	62.7%	30232.4	$<2.10^{-16}$
+ GC-flank	48.1%	$<2.10^{-16}$	63.0%	126.8	$<2.10^{-16}$
+ Intragenic recombination rate	13.0%	$<2.10^{-16}$	66.9%	1453.3	$<2.10^{-16}$

What fraction of the variance in GC3 is explained by gBGC ?

GC3 predictors	Pairwise R^2	p-value	Model R^2	F statistic	p-value
GCi	62.7%	$<2.10^{-16}$	62.7%	30232.4	$<2.10^{-16}$
+ GC-flank	48.1%	$<2.10^{-16}$	63.0%	126.8	$<2.10^{-16}$
+ Intragenic recombination rate	13.0%	$<2.10^{-16}$	66.9%	1453.3	$<2.10^{-16}$
+ Expression level in meiosis	8.8%	$<2.10^{-16}$	68.2%	875.7	$<2.10^{-16}$

What fraction of the variance in GC3 is explained by gBGC ?

GC3 predictors	Pairwise R^2	p-value	Model R^2	F statistic	p-value
GCi	62.7%	$<2.10^{-16}$	62.7%	30232.4	$<2.10^{-16}$
+ GC-flank	48.1%	$<2.10^{-16}$	63.0%	126.8	$<2.10^{-16}$
+ Intragenic recombination rate	13.0%	$<2.10^{-16}$	66.9%	1453.3	$<2.10^{-16}$
+ Expression level in meiosis	8.8%	$<2.10^{-16}$	68.2%	875.7	$<2.10^{-16}$
+ Functional category	1%	$<2.10^{-16}$	68.3%	30.43	$<2.10^{-16}$

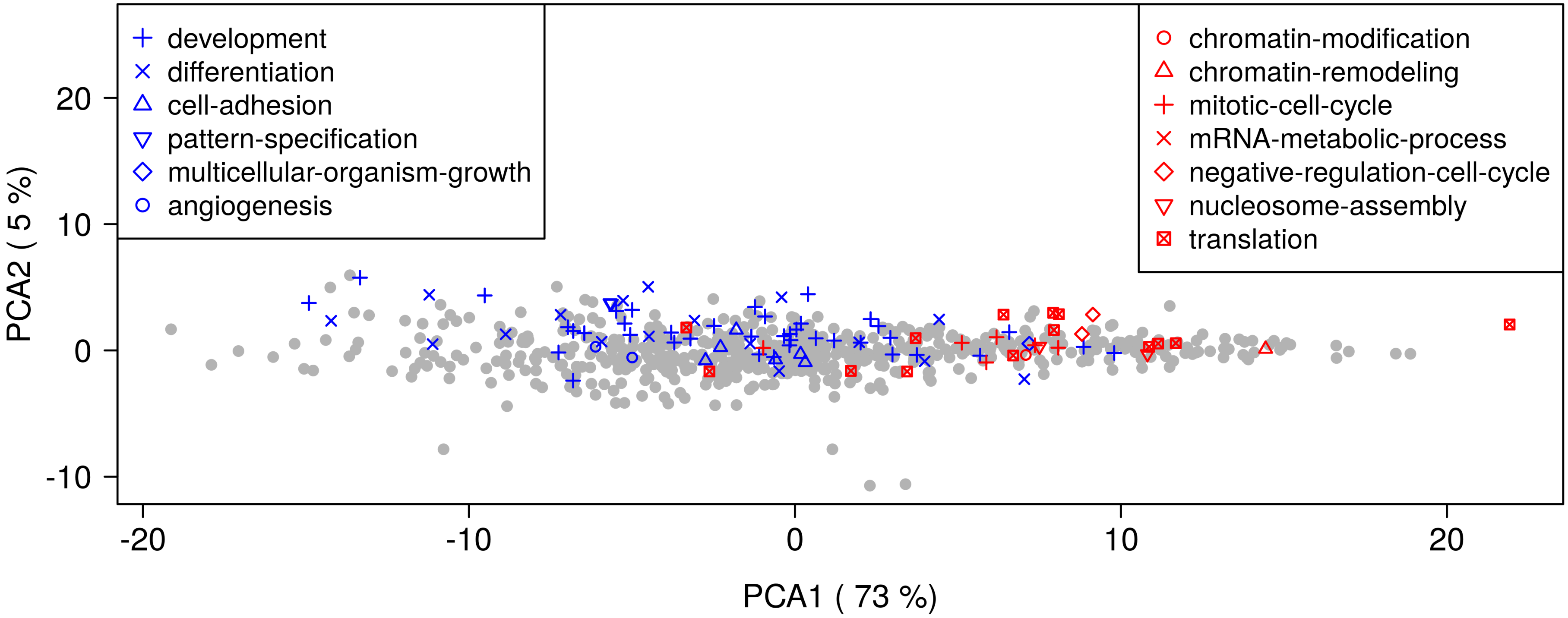
68.2% of the variance in GC3 is explained by gBGC

What fraction of the variance in GC3 is explained by gBGC ?

- 68.2% of the variance in GC3 is explained by gBGC

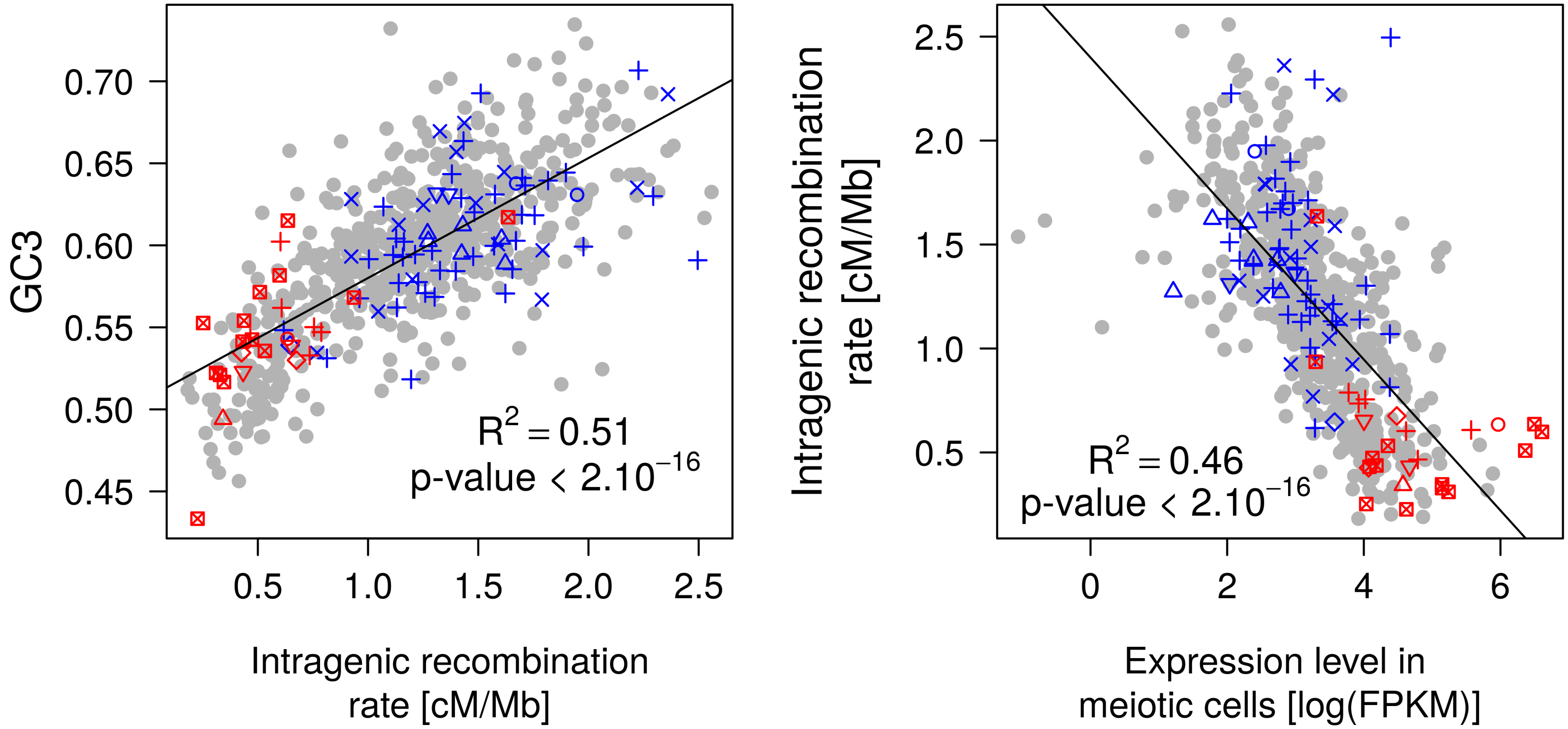
- NB:
  - The number of codons in a gene is limited
  - A part of the variance in GC3 simply results from stochastic sampling
  - In fact, **80% of the explainable variance in GC3 is explained by gBGC**

What about other functional categories?



What about other functional categories?

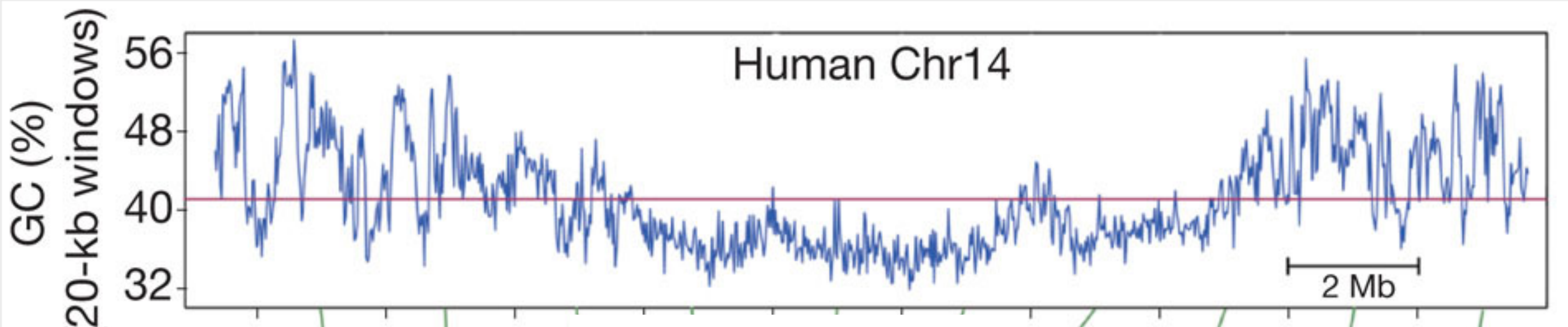
GC3 of GO gene sets
N=687 functional categories with > 40 genes

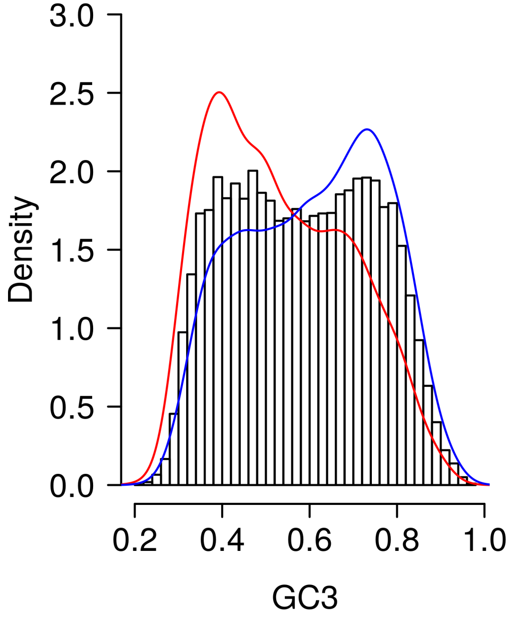


Conclusion (1): why does synonymous codon usage vary among functional categories of human genes?

- Transcription during meiosis interferes with recombination
- Genes that are expressed during meiosis have a lower recombination rate => weaker gBGC => lower GC-content
- Different functional category have different patterns of expression => different GC-content => different codon usage

Conclusion (2): the main determinant of codon usage = large-scale variation in recombination rate





Conclusion (3): no evidence of selection on translation efficiency in humans

- Nematode, drosophila, arabidopsis: selection on translation efficiency
- Why not in mammals?
  - Low Ne => selection is less efficient
  - gBGC + variation in recombination rate along chromosome => strong heterogeneity in GC3
  - => impossible to adapt the pool of tRNA to the demand in codon usage

Reference

- The last part of this lecture (slides 28-56) corresponds to unpublished work by Fanny Pouyet, Dominique Mouchiroud, Laurent Duret & Marie Sémon

Further readings

- M. Lynch: The origins of genome architecture