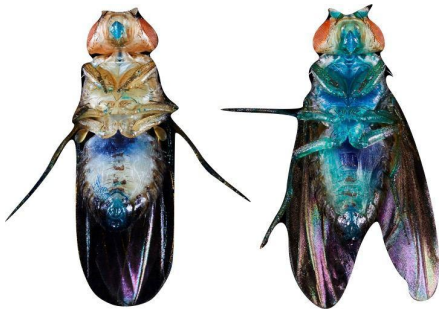
Large-scale assessment of the impact of protein sequence variations on ageing using *Drosophila melanogaster*

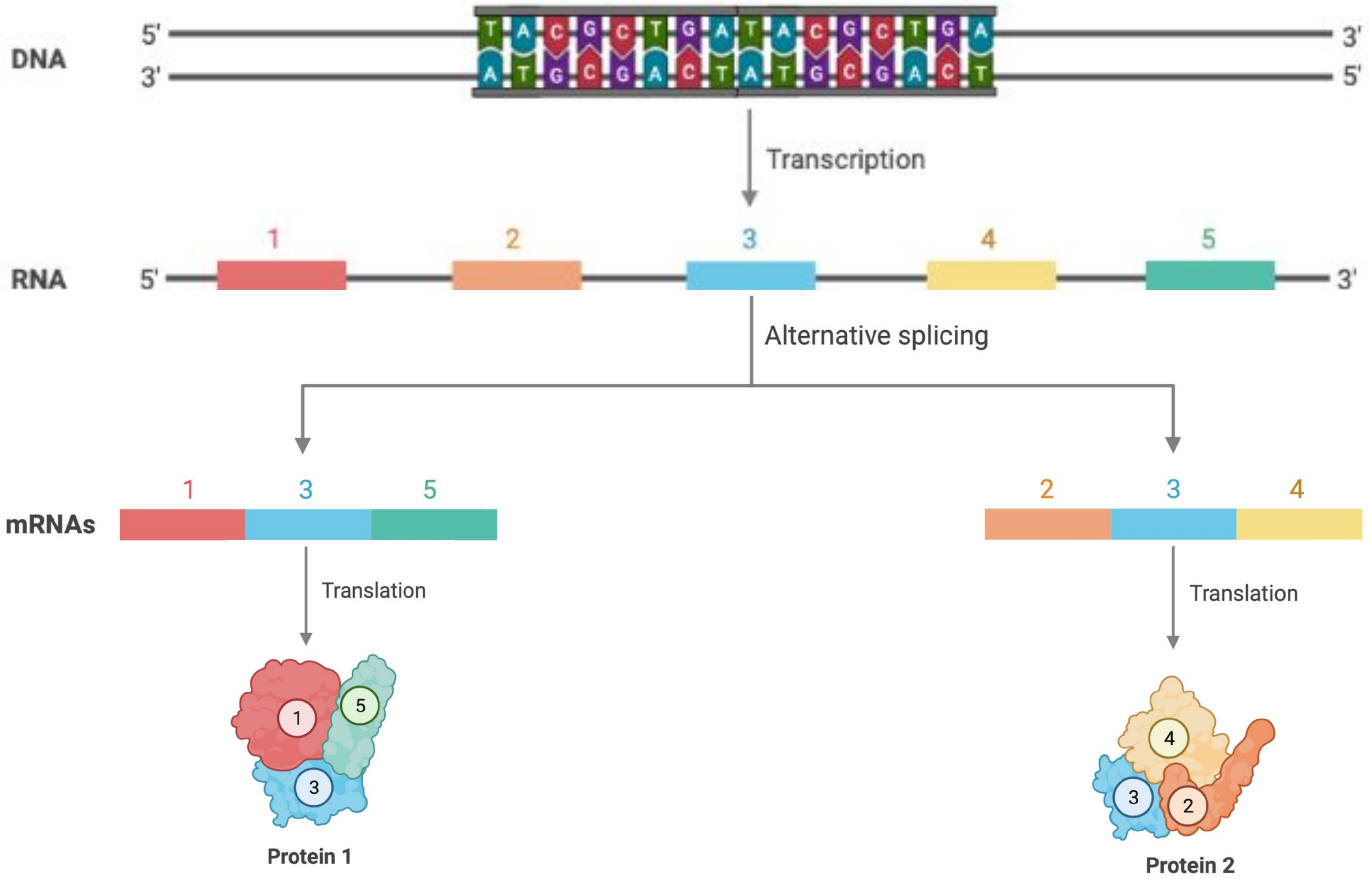
Marina Abakarova

16.06.22 Aussois

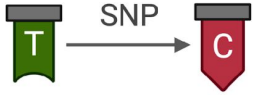
Project supervisors
Elodie Laine, MCF SU
Michael Rera, CR CNRS

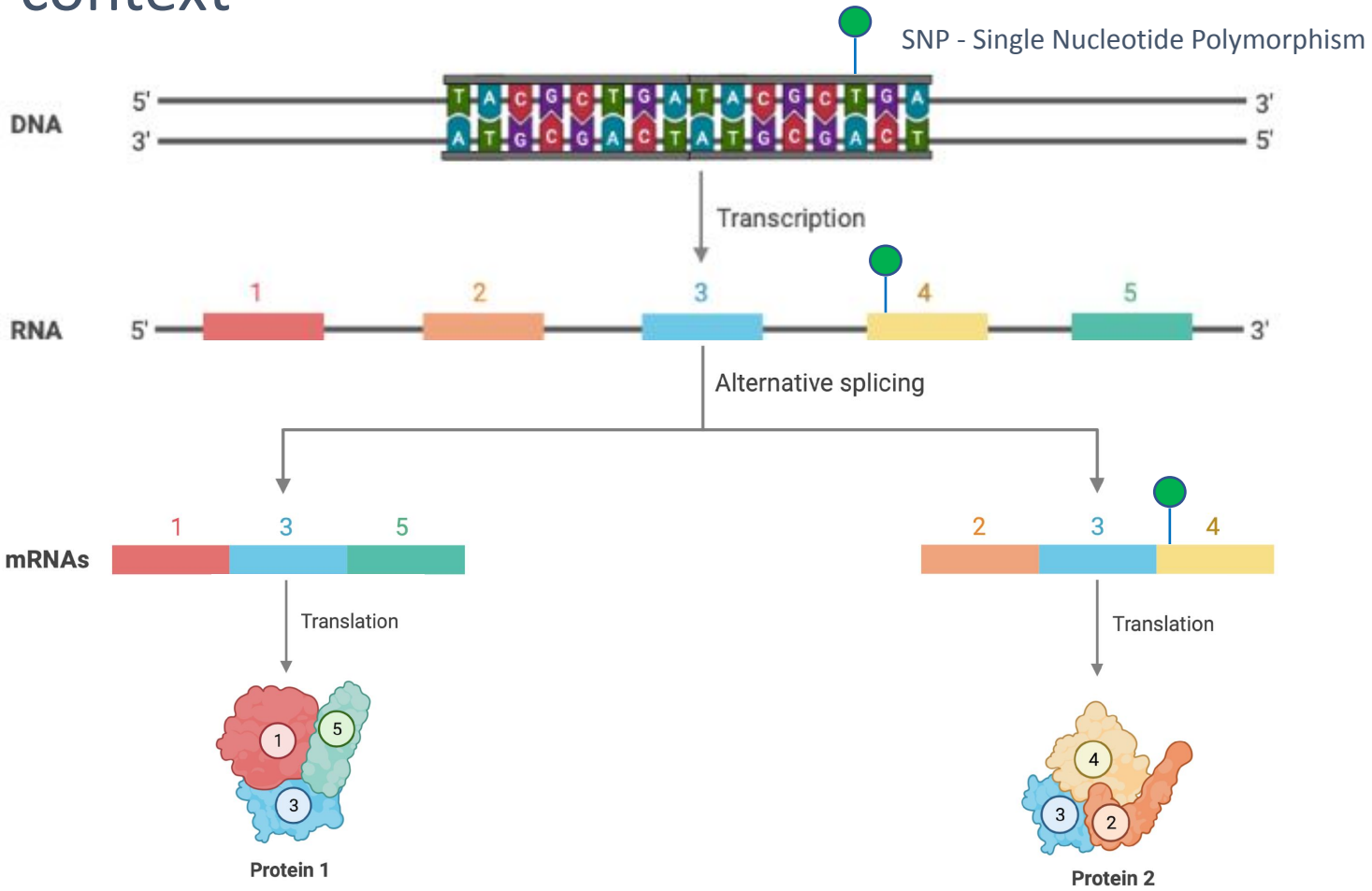


Biological context

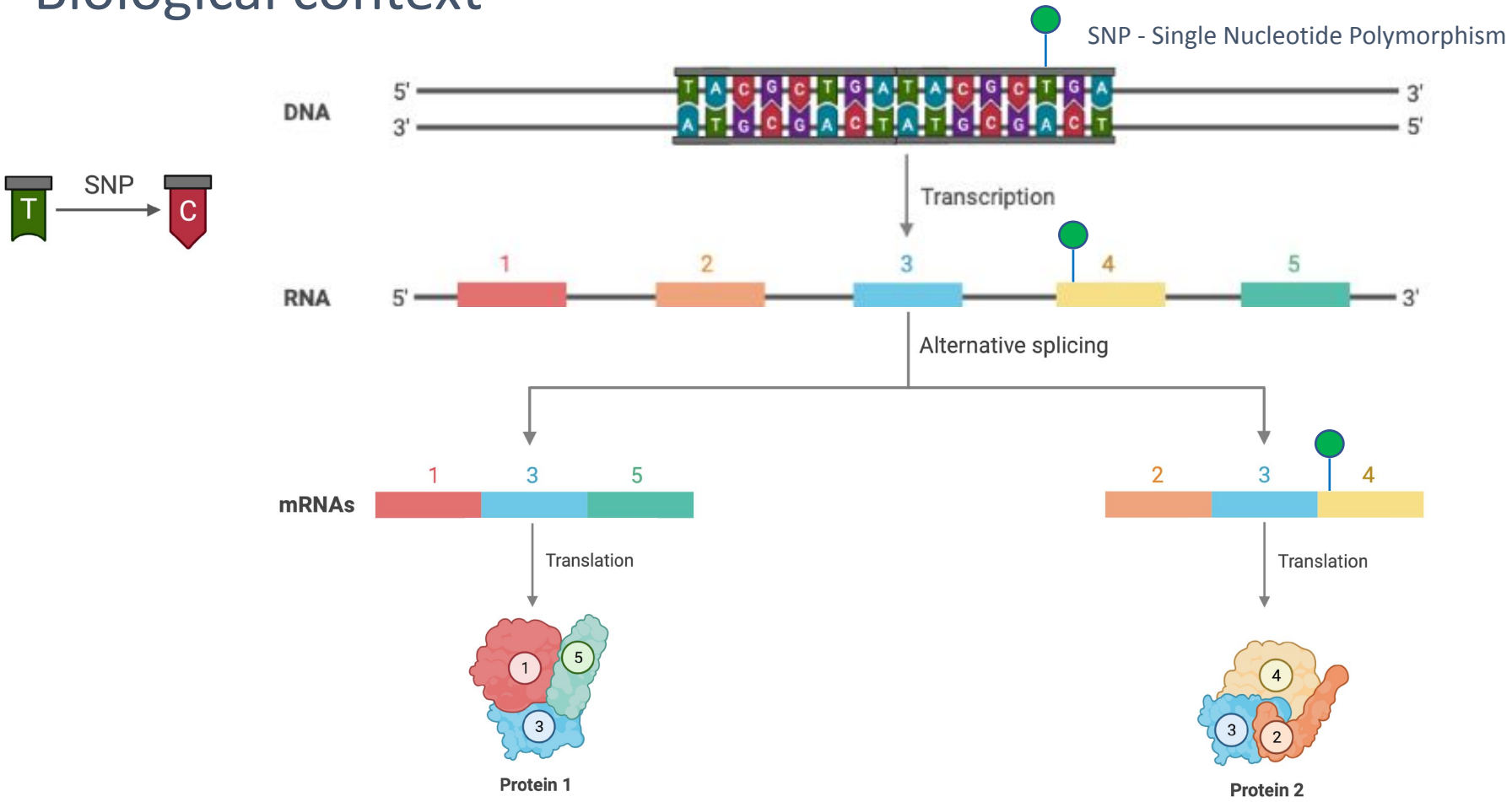


Biological context

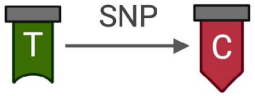


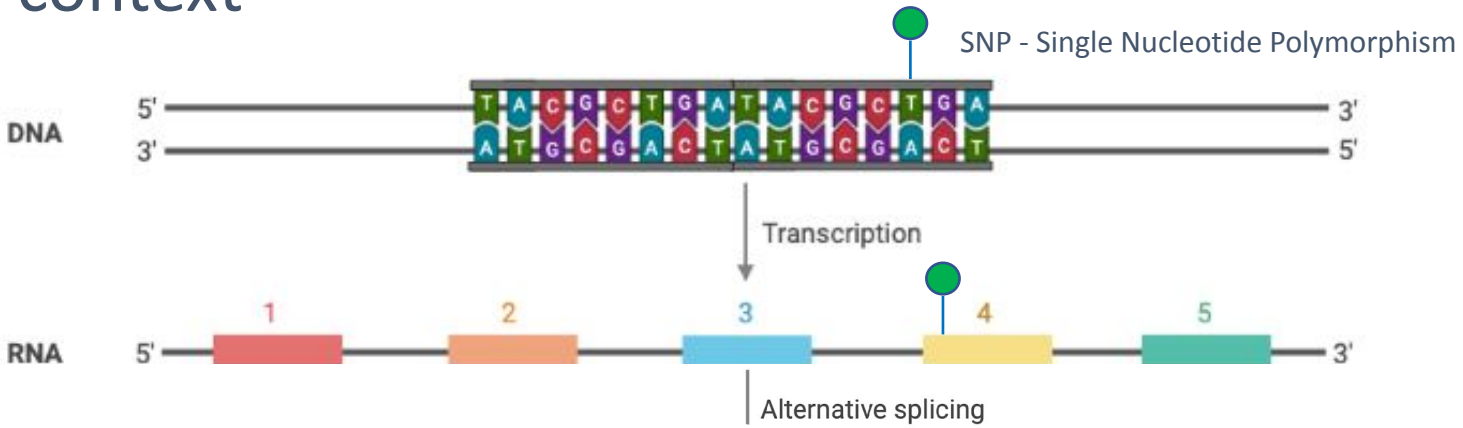


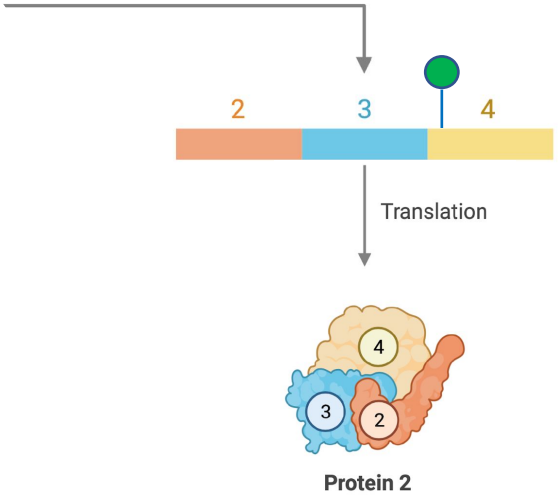
Biological context



Biological context



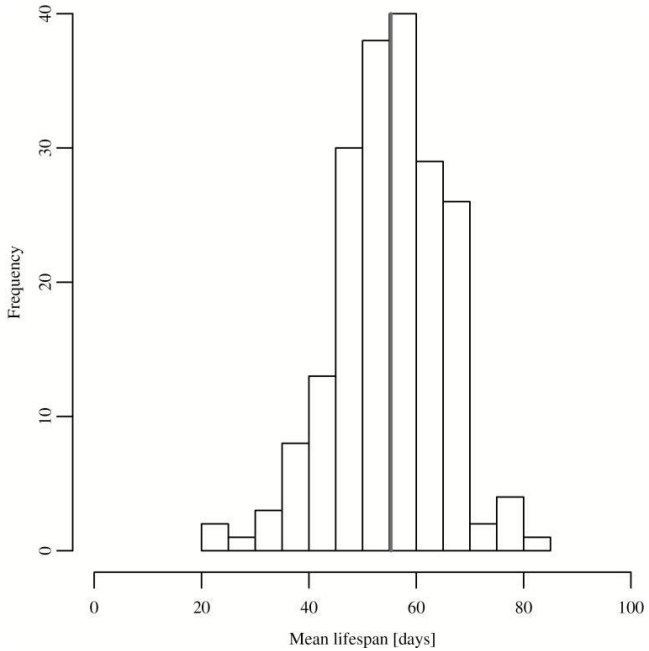




Problem to tackle: What is the impact of such sequence variations on lifespan?

Ageing phenotype: a continuous quantitative trait

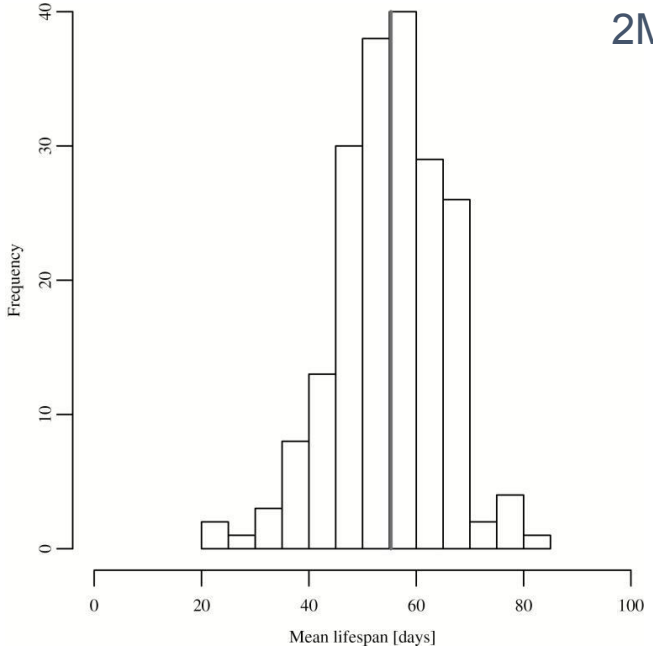
broad continuous range of life expectancies

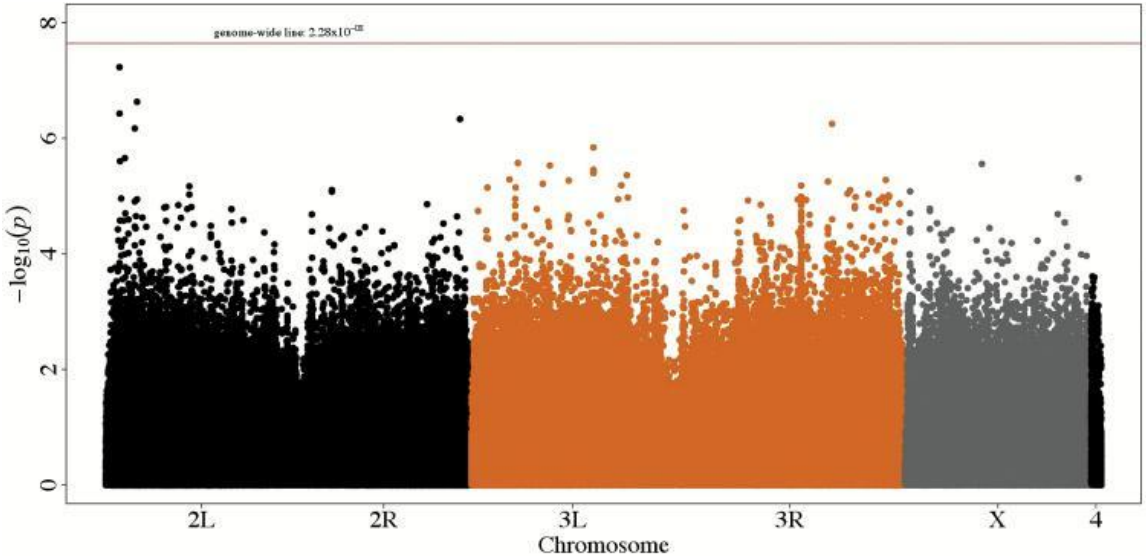


Ivanov et al. 2015
Mackay et al. 2012

Ageing phenotype: a continuous quantitative trait

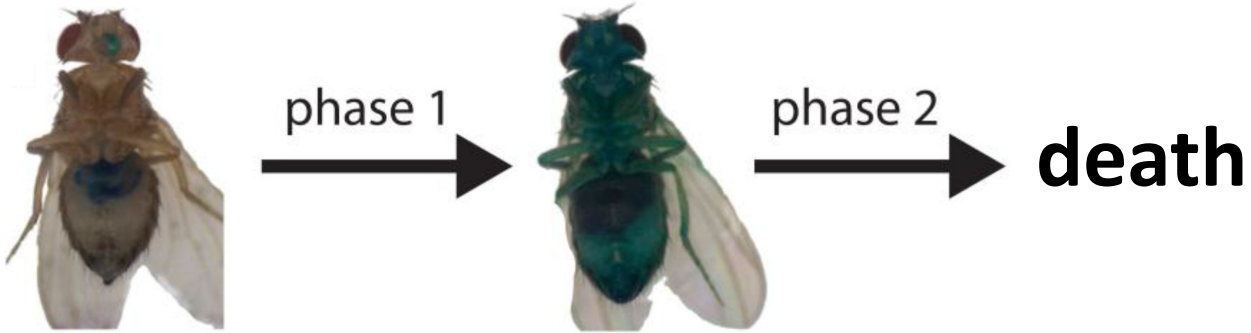
GWAS - Genome Wide Association Study
Drosophila melanogaster Genetic Reference Panel (DGRP)
2M of tested SNPs, only ~4.7% describes life expectancy variation





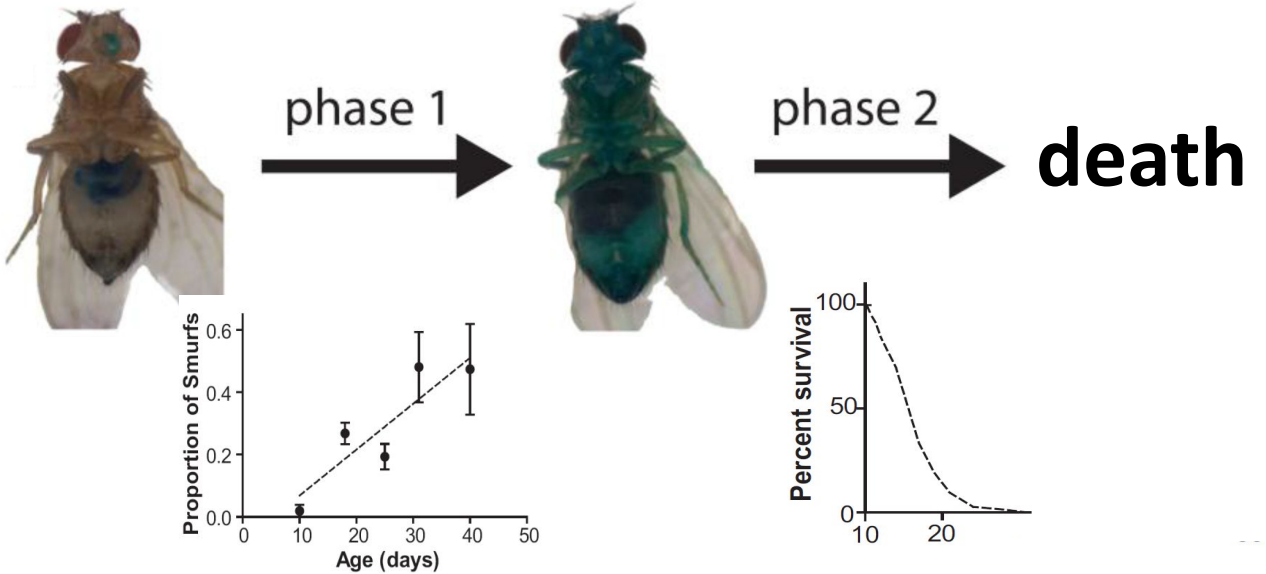
Ivanov et al. 2015
Mackay et al. 2012

Ageing phenotype: the two-phase model



Tricoire&Rera *PloS One* 2015

Ageing phenotype: the two-phase model

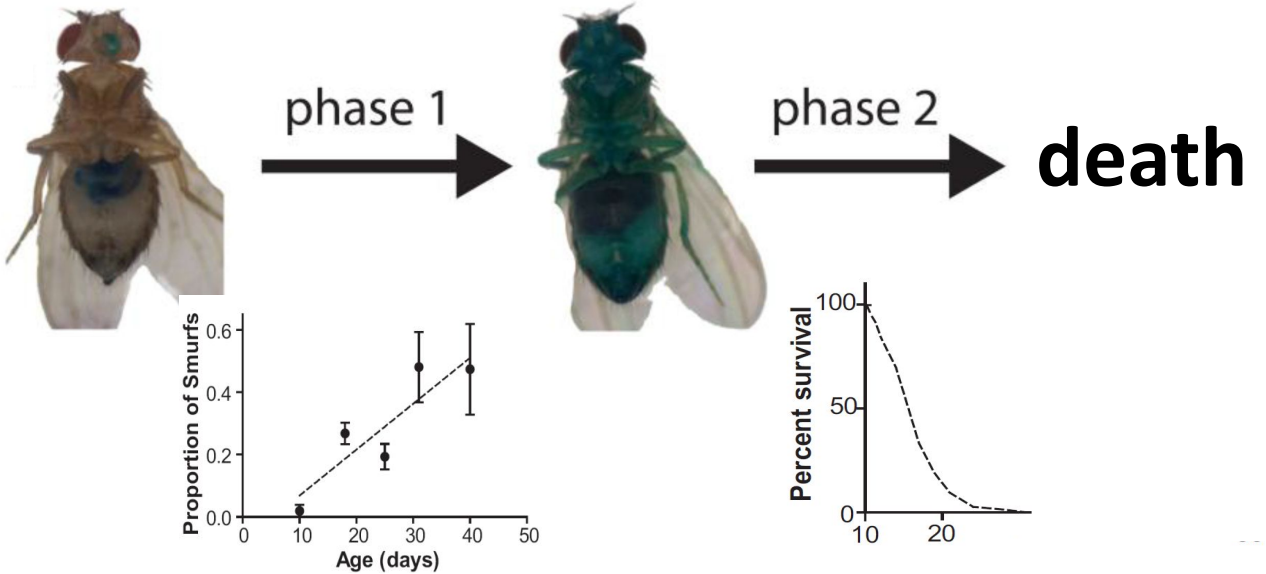


Smurfs = a*t + b

Smurf survival = e^{-k*t}

Tricoire&Rera *PloS One* 2015

Ageing phenotype: the two-phase model

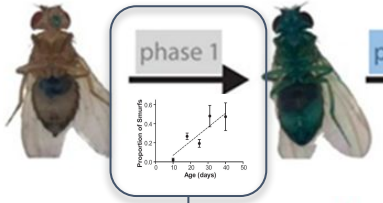


Smurfs = a*t + b

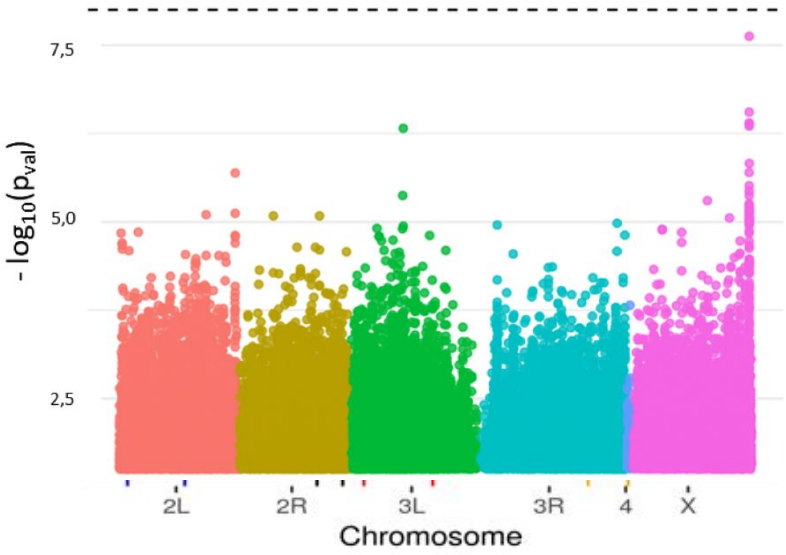
Smurf survival = e^{-k*t}

Tricoire&Rera *PloS One* 2015

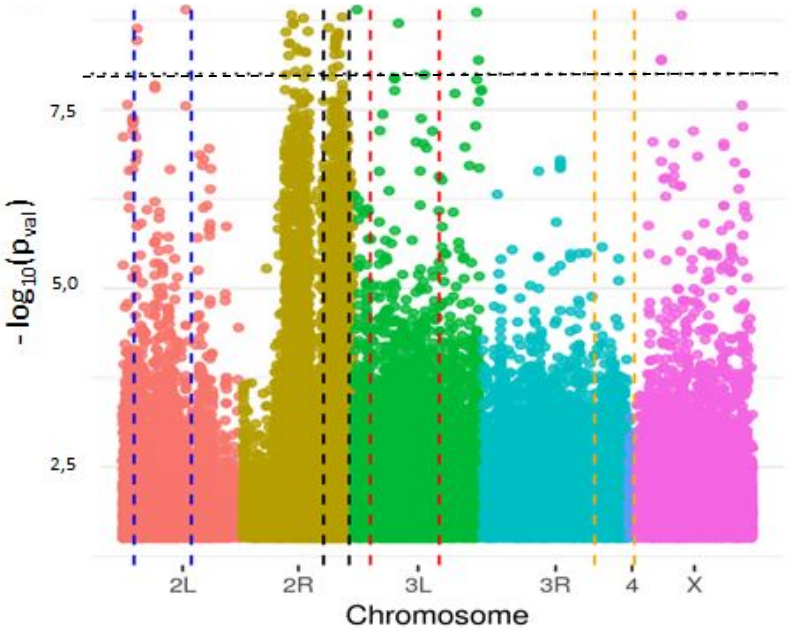
Ageing phenotype: the two-phase model



Lifespan-based GWAS



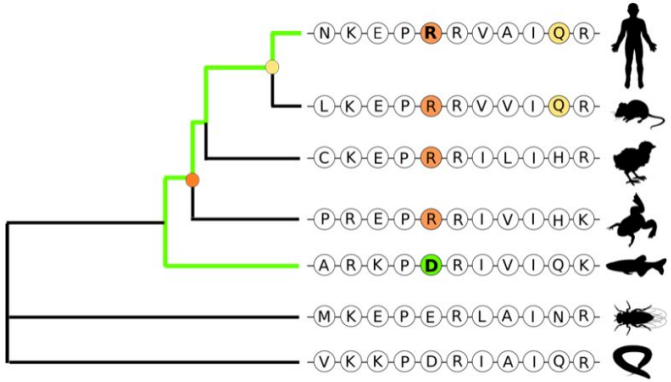
***a** parameter - based GWAS*

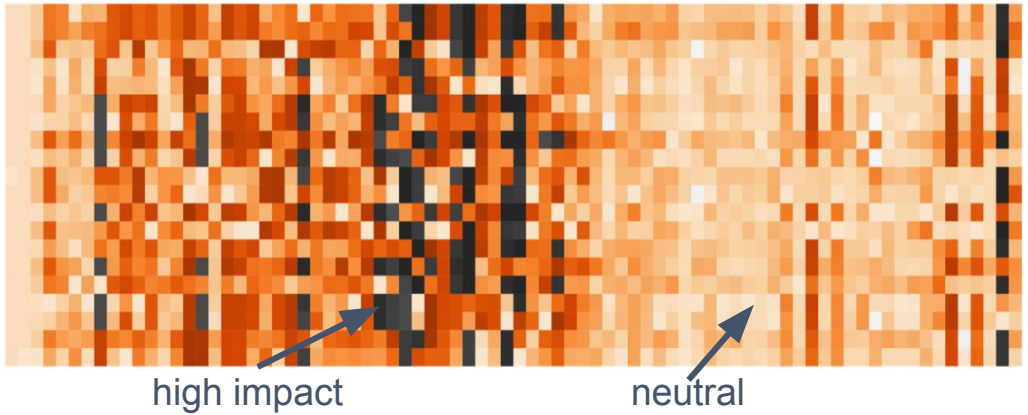


Tricoire&Rera *PloS One* 2015

GEMME - an evolutionary-informed model

a protein full mutational landscape





Takes as input a Multiple Sequence Alignment (MSA) and explicitly accounts for the evolutionary relationships between natural protein sequences.

E.Laine *et al.* **MBE 2019**

GEMME - an evolutionary-informed model

Main hypotheses

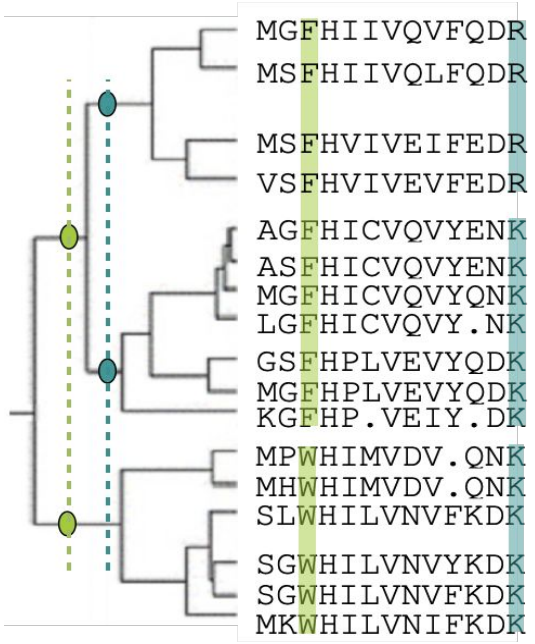
- **conservation** is an indicator of mutational sensitivity
- **epistasis**: positions interact with each other

```
MGFHIIVQVFQDR
MSFHIIVQLFQDR
MSFHVIVEIFEDR
VSFHVIVEVFEDR
AGFHICVQVYENK
ASFHICVQVYENK
MGFHICVQVYQNK
LGFHICVQVY.NK
GSFHPLVEVYQDK
KGFHP.VEIY.DK
MPWHIMVDV.QNK
MHWHIMVDV.QNK
SLWHILVNVFKDK
SGWHILVNVYKDK
SGWHILVNVFKDK
MKWHILVNIFKDK
KGFHP.VEIY.DK
MPWHIMVDV.QNK
MHWHIMVDV.QNK
SLWHILVNVFKDK
SGWHILVNVYKDK
SGWHILVNVFKDK
MKWHILVNIFKDK
```

$\sqrt{N}$ sequences

Gibbs sampling

$$T_{\mathrm{JET}}(i) = \frac{1}{M_i} \sum_{t=1}^{M_i} \frac{L_t - l_i^t}{L_t}$$



MSA with N sequences

Joint Evolutionary Trees

Engelen et *al.* 2009

GEMME - an evolutionary-informed model

Main hypotheses

- **conservation** is an indicator of mutational sensitivity
- **epistasis**: positions interact with each other

$$D_{evol}(q,s) = \sum_{i=1}^{n} T_{JET}(i)^2 * \mathbf{1}_{X_i^q \neq X_i^s}(i)$$

E.Laine *et al.* **MBE 2019**

GEMME - an evolutionary-informed model

Main hypotheses

- **conservation** is an indicator of mutational sensitivity
- **epistasis**: positions interact with each other

$$D_{evol}(q,s) = \sum_{i=1}^{n} T_{JET}(i)^2 * \mathbf{1}_{X_i^q \neq X_i^s}(i)$$

Epistatic contribution: $PE^{Epi}(Y_i) = \min[D_{evol}(q,s)]$

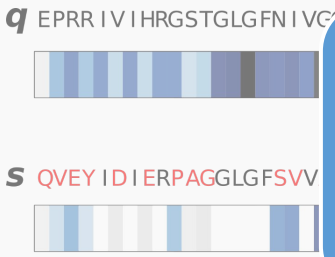
Independent contribution: $PE^{Ind}(Y_i) = -\log\left[\frac{\max(1,|S_{Y_i}|)}{|S_{X_i}|}\right]$

E.Laine *et al.* **MBE 2019**

GEMME - an evolutionary-informed model

Main hypotheses

- **conservation** is an indicator of mutational sensitivity
- **epistasis**: positions interact with each other



How to scale to entire proteomes?

$$\sum_{:1} T_{JET}(i)^2 * 1_{X_i^q \neq X_i^s}(i)$$

Epistatic contribution: $PE^{Epi}(Y_i) = \min[D_{evol}(q,s)]$

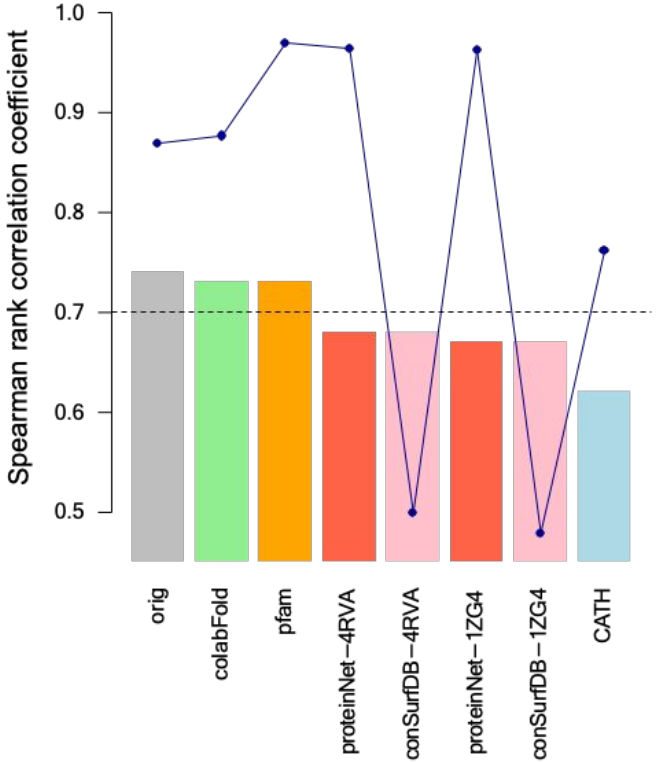
Independent contribution: $PE^{Ind}(Y_i) = -\log\left[\frac{\max(1,|S_{Y_i}|)}{|S_{X_i}|}\right]$

E.Laine *et al.* **MBE 2019**

Development of a scalable protocol

SACADO
SERVICE D'AIDE AU CALCUL ET À L'ANALYSE DE DONNÉES
SORBONNE UNIVERSITÉ

GEMME predictions for BLAT (~5000 mut)



MSA	Method	#(seqs)
orig	JackHMMER (5 it), Uniref100	~15 000
colabFold	MMseqs2, Uniref30+Env.	~9 000
Pfam	from UniProt	~43 000
ProteinNet	JackHMMER (5 it), Uniparc + JGI	~220-250 000
ConSurfDB	HMMER (1 it), Uniref90	300
CATH	Muscle on CATH superfamily	300

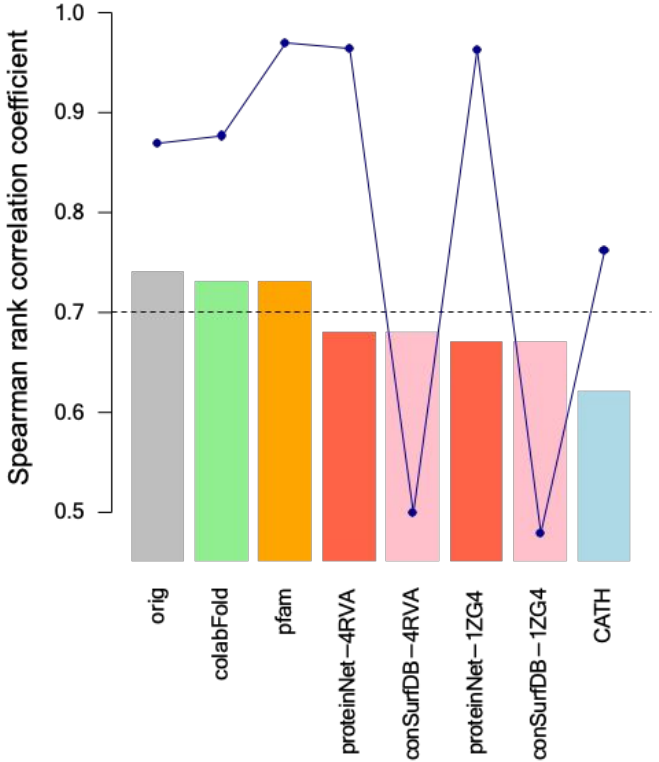
Landscape covered by the MSA

$$1 - \frac{\#(\text{missing substitutions})}{\#(\text{substitutions})}$$

Development of a scalable protocol

SACADO
SERVICE D'AIDE AU CALCUL ET À L'ANALYSE DE DONNÉES
SORBONNE UNIVERSITÉ

GEMME predictions for BLAT (~5000 mut)



Spearman correlation orig: 0.74

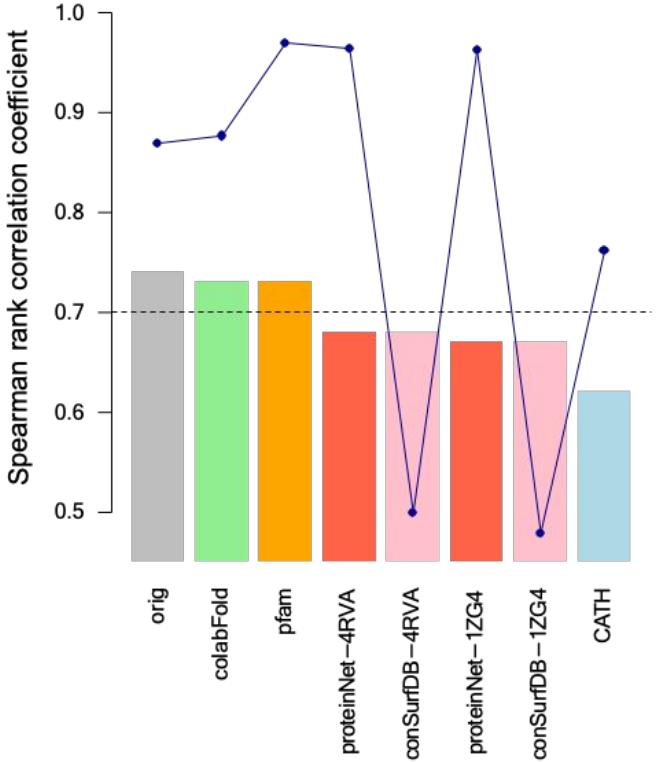
Landscape covered by the MSA

$$1 - \frac{\#(\text{missing substitutions})}{\#(\text{substitutions})}$$

Development of a scalable protocol

SACADO
SERVICE D'AIDE AU CALCUL ET À L'ANALYSE DE DONNÉES
SORBONNE UNIVERSITÉ

GEMME predictions for BLAT (~5000 mut)



Spearman correlation orig: 0.74

Confirmed on the benchmark set (24 proteins)

ColabFold (MMseqs) > Pfam

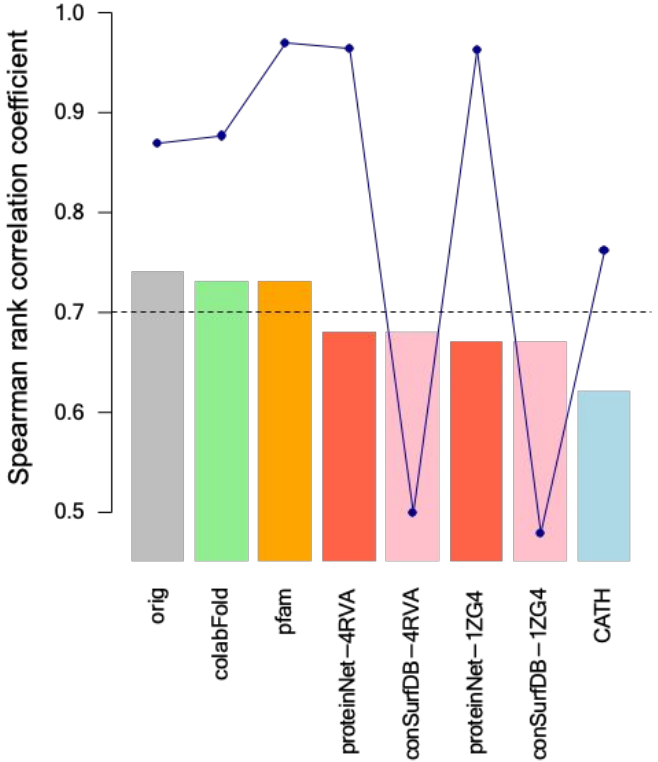
Landscape covered by the MSA

$$1 - \frac{\#(\text{missing substitutions})}{\#(\text{substitutions})}$$

Development of a scalable protocol

SACADO
SERVICE D'AIDE AU CALCUL ET À L'ANALYSE DE DONNÉES
SORBONNE UNIVERSITÉ

GEMME predictions for BLAT (~5000 mut)



Spearman correlation orig: 0.74

Confirmed on the benchmark set (24 proteins)

ColabFold (MMseqs) > Pfam

Scanning of the entire Fly proteome
(95%: 29 339 sequences)

Execution time :
3 ans ⟶ 4 jours

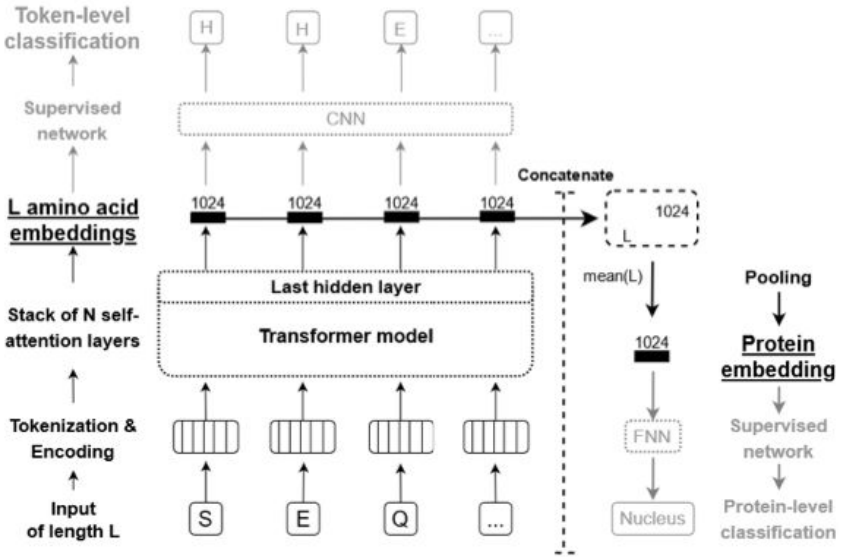
Landscape covered by the MSA

$$1 - \frac{\#(\text{missing substitutions})}{\#(\text{substitutions})}$$

Protein language models

ROSTLAB.
TUM, Germany

VESPA - Variant Effect Score Prediction without Alignments



pLM-embeddings - protT5

➔ no alignment needed
➔ very fast

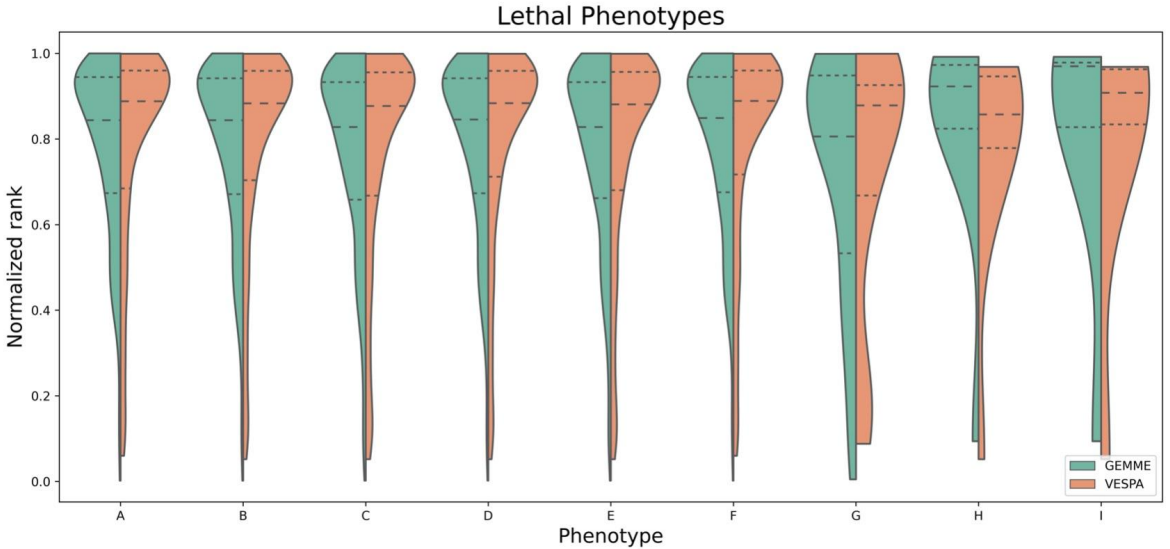
Embeddings and VESPA predictions for
for the entire proteome (100%: 30 784 sequences)

pLM - protein Language Model

C.Marquet *et al.* 2021
Elnaggar *et al.* 2021

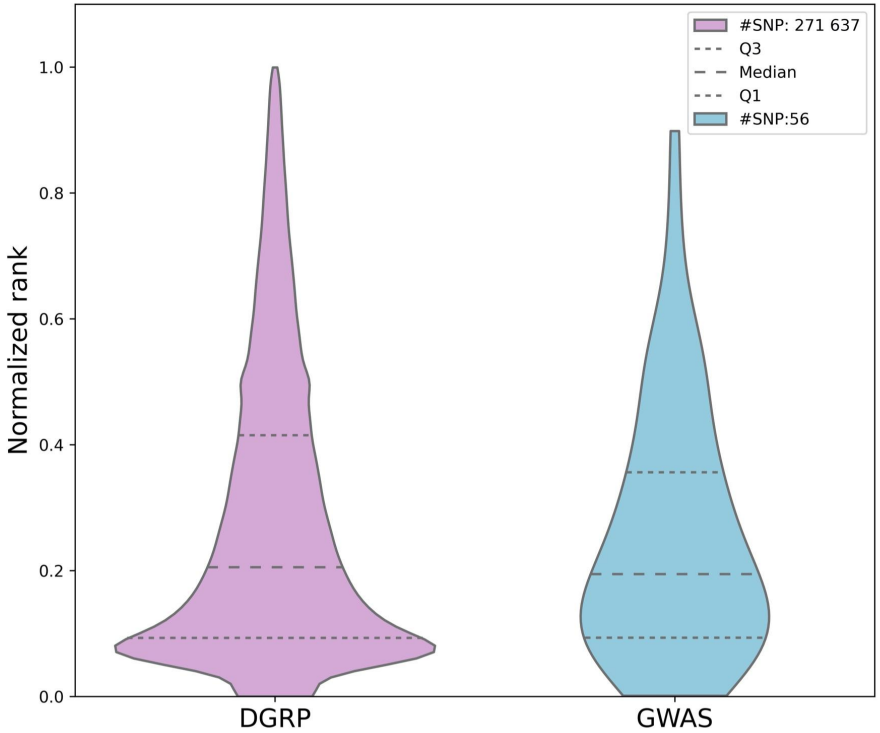
Lethal Phenotype

A - all die before end of first instar larval stage
B - all die before larval stage
C - all die before end of P-stage
D - all die before end of prepupal stage
E - all die before end of pupal stage
F - all die during embryonic stage
G - all die during larval stage
H - all die during P-stage
I - all die during pupal stage



Ageing phenotype

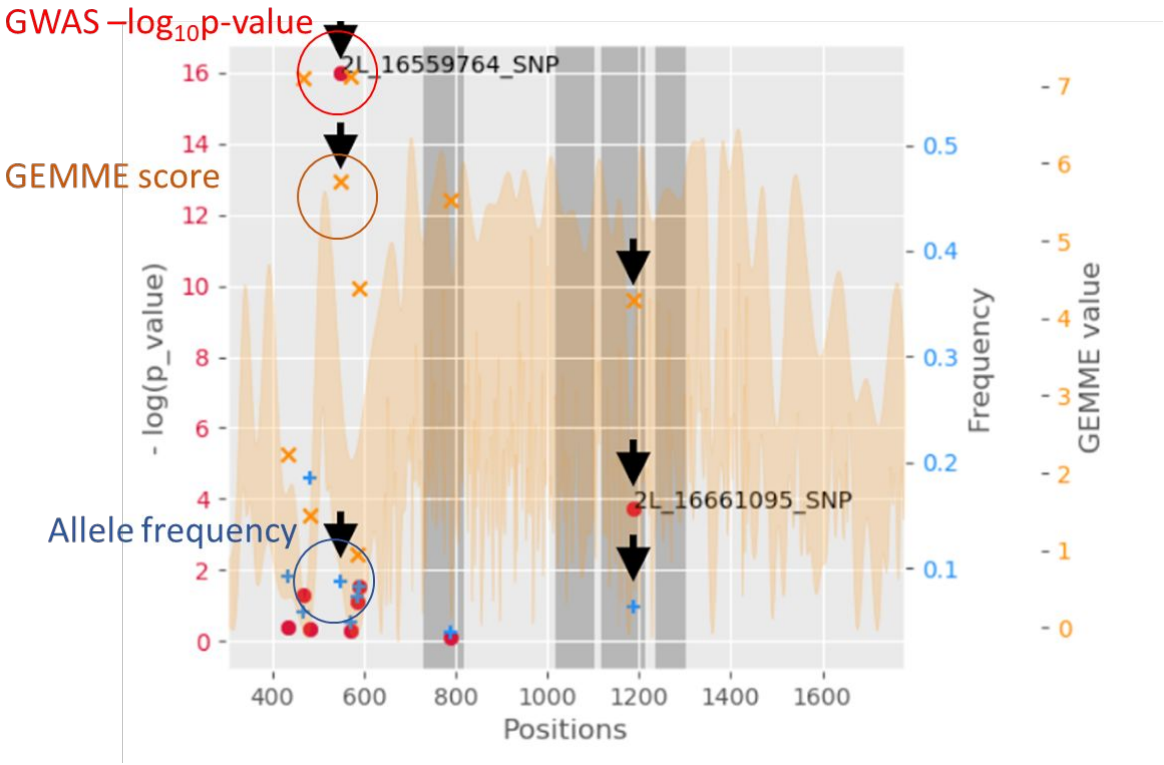
Distribution of mutational outcomes predicted by GEMME



The outcomes predicted for the genetic polymorphism observed in DGRP lines, and more specifically for the SNPs displaying highly significant p-values in the GWAS study span a wide range of values.

A specific example…

“miles to go” gene, crucial for the neuromuscular growth and branching.



PhD projet

Identify a set of non-synonymous single nucleotide polymorphism (SNPs) highly relevant for ageing and validate them by in vivo experiments on flies.

Investigate SNPs annotated in **lethal phenotypes** with relatively low mutational effect.

Build a predictive and interpretable model to characterize the way alternative splicing shapes Drosophila's proteome and interactome during ageing.

Expand our methods to non coding genome.

Explore evolutionary semantics in **language models**.

M2 project: predicting SNPs relevant to longevity in Drosophila

Evolutionary information encoded in proteins

GEMME VESPA

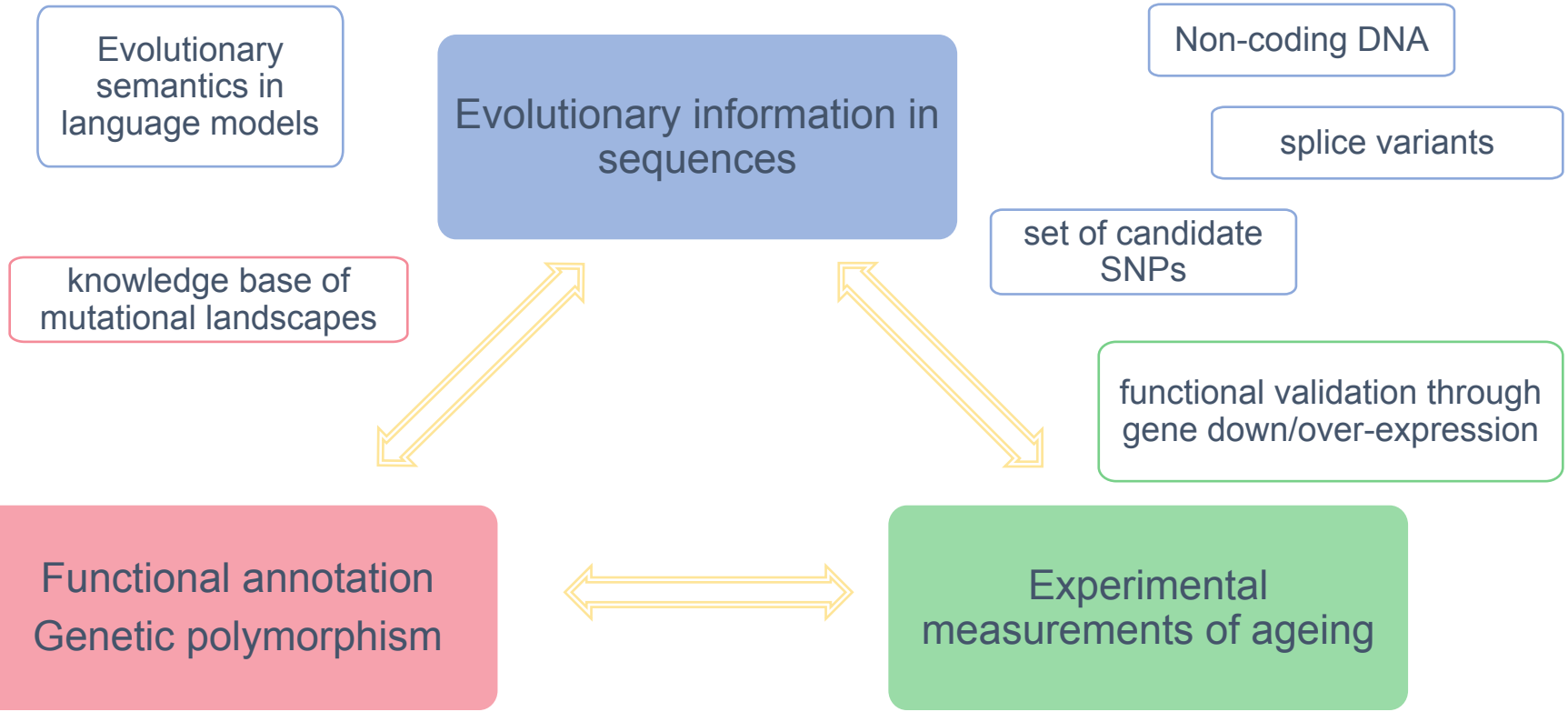
Functional annotation

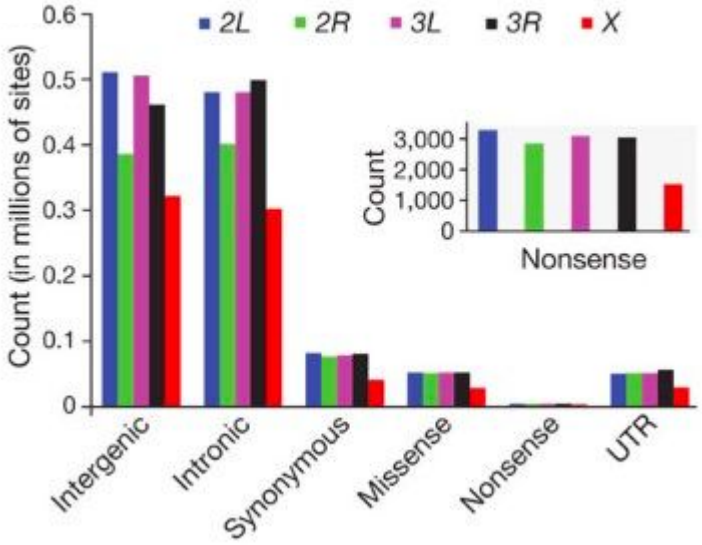
Genetic polymorphism

FlyBase DGRP

Experimental measurements of ageing

Large-scale assessment of the impact of protein sequence variations on ageing using *Drosophila melanogaster*





Mackay et al., 2012

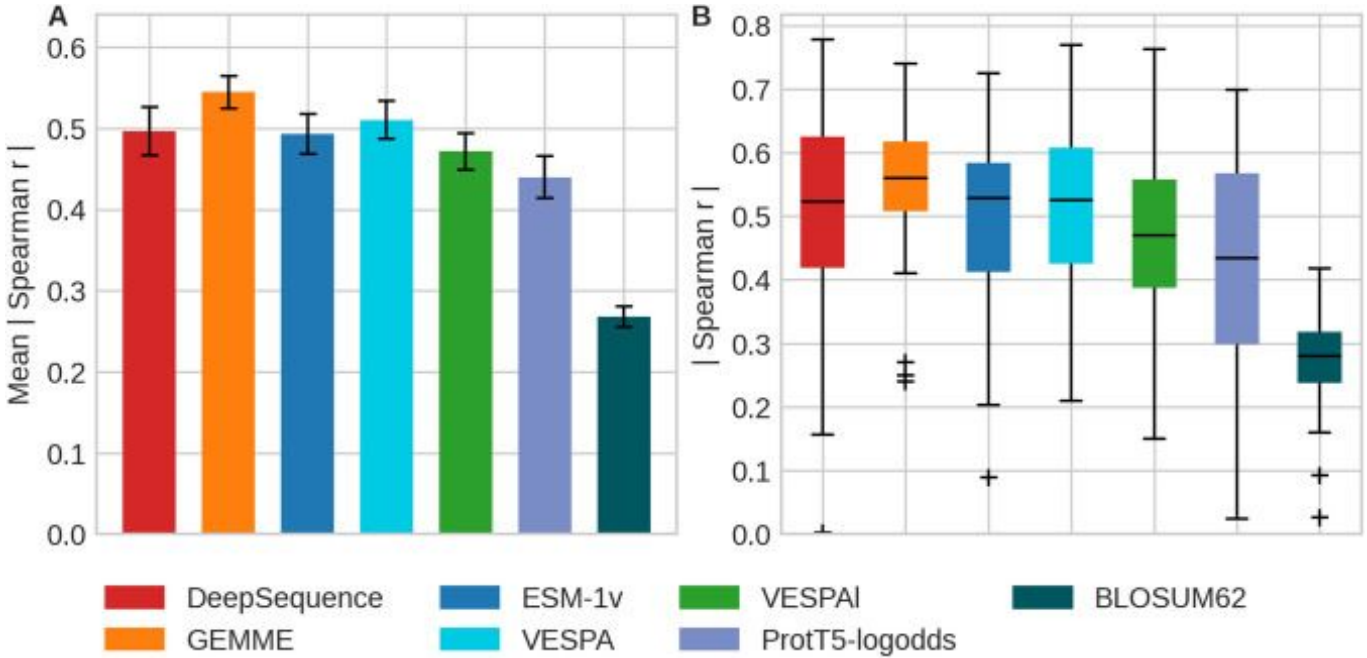
117 lines 27k female flies in total
M plot - inversion in 2R chr associated with 30% lifespan increase for homoz

INFO+

- 117 DGRP, une resource publics des lignées isogénique
- c’est la proportion qui augmentte avec le temps et pas la quantité
- Femelle: elles sont plus grande - donc plus facilement observable, les smurf mal meurent plus tot que les femelle. et pour diminuer la quantités (pourquoi comme ça?)
- Croisement avec les males que les deux premiers jours
- lifespan de 15 à 67 jours
- studying ageing allows us to identify previously hidden features of ageing

Computational methods:
Benchmarking with state-of-the-art methods

Spearman correlation between DMS experiment and predictions by different methods



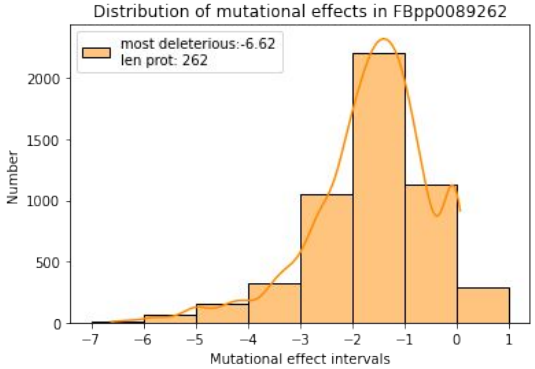
DMS - Deep mutational scanning

C.Marquet *et al.* 2021

Bottleneck

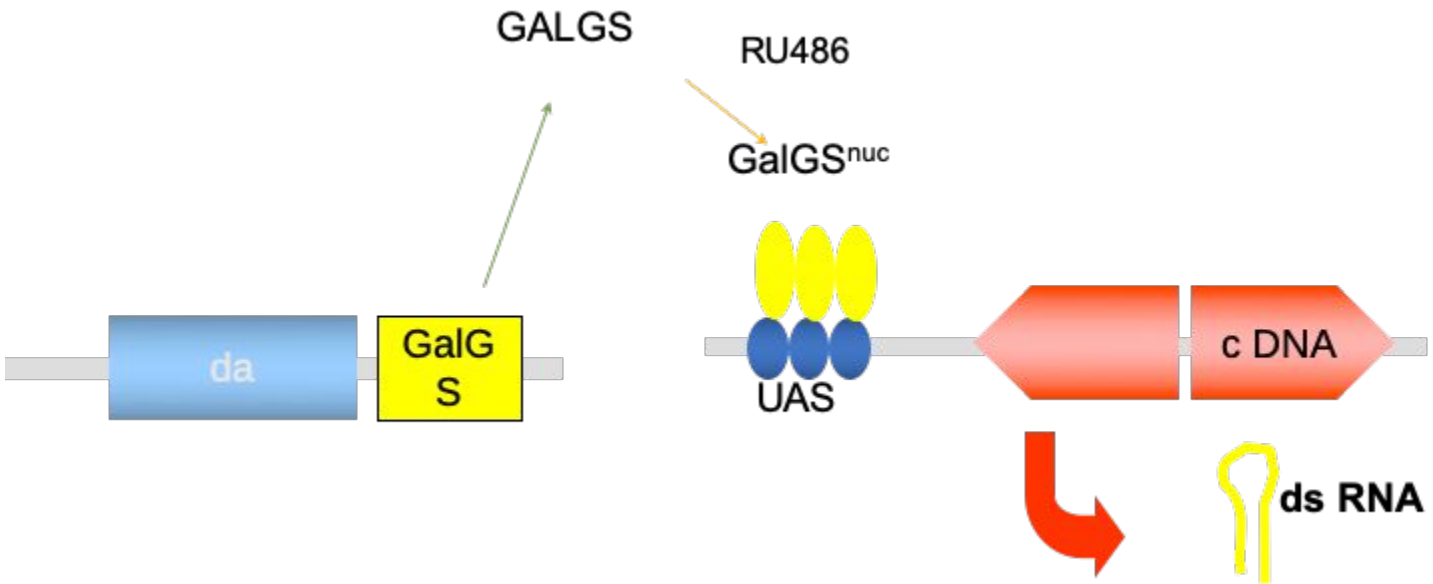
Computational power problem with ColabFold:

- MeSU cluster : a very large shared-memory computer with 16 TB memory
- RAM memory 810-840Go for 3000 sequences
- running time: 4 hours in average

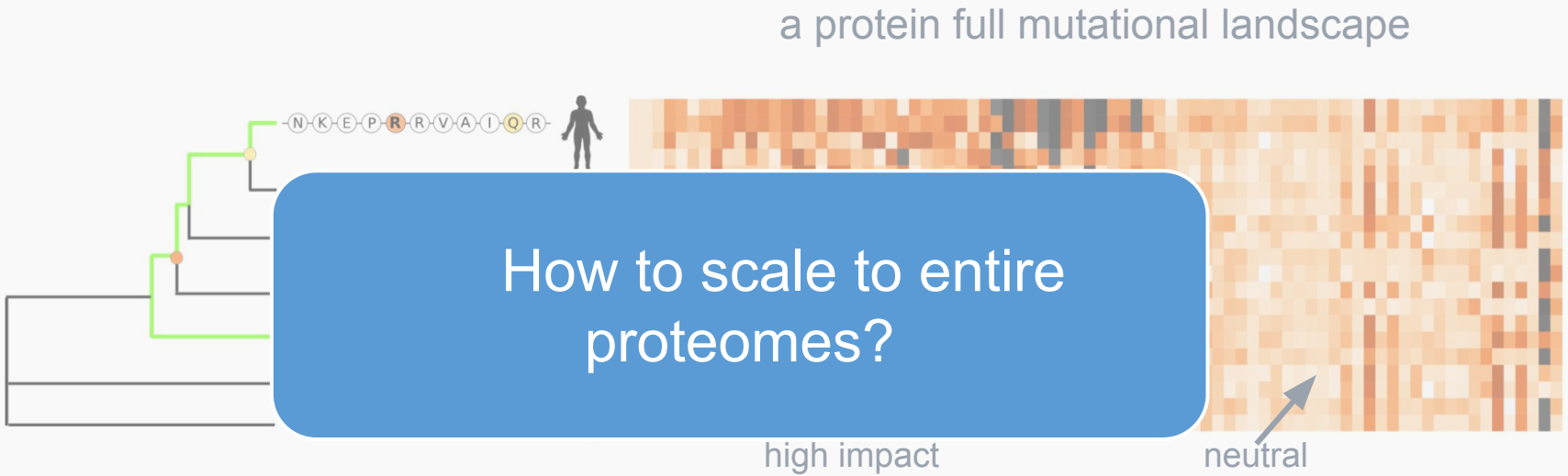


MSA	Method	#(seqs)
orig	JackHMMER (5 it), Uniref100	~15 000
colabFold	MMseqs2, Uniref30+Env.	~9 000
Pfam	from UniProt	~43 000
ProteinNet	JackHMMER (5 it), Uniparc + JGI	~220-250 000
ConSurfDB	HMMER (1 it), Uniref90	300
CATH	Muscle on CATH superfamily	300

Manipulating gene expression



GEMME, an evolutionary-informed model



Takes as input a Multiple Sequence Alignment (MSA) and explicitly accounts for the evolutionary relationships between natural protein sequences.

How to build the input MSA?
Default method - Jackhammer 5 iterations ~ 1h for a protein

HMMER

E.Laine *et al.* **MBE 2019**

Bibliography

1. H.Tricoire, **M.Rera**. *A New, Discontinuous 2 Phases of Aging Model: Lessons from Drosophila melanogaster*. PLoS ONE 2015
2. **Laine, E**.; Karami, Y.; Carbone, A. *GEMME: A Simple and Fast Global Epistatic Model Predicting Mutational Effects.* Mol. Biol. Evol. 2019
3. Rera, M., Clark, R. I. & Walker, D. W. *Intestinal barrier dysfunction links metabolic and inflammatory markers of aging to death in Drosophila*. PNAS 2012
4. Mirdita M., Schütz K., Moriwaki Y., Heo L., Ovchinnikov S., Steinegger M. ColabFold - *Making protein folding accessible to all biorxiv*
5. Marquet C, Heinzinger M, Olenyi T, Dallago C, Erckert K, Bernhofer M, Nechaev D, Rost B. *Embeddings from protein language models predict conservation and variant effects.* Hum Genet. 2021
6. Zea, D. J.; Laskina, S.; Baudin, A.; Richard, H.; **Laine, E.** *Assessing Conservation of Alternative Splicing with Evolutionary Splicing Graphs*. *Genome Res.* 2021
7. Riesselman AJ, Ingraham JB, Marks DS. *Deep generative models of genetic variation capture the effects of mutations*. Nat Methods. 2018

How to build the input MSA?

JackHMMER iterative search based on profile Hidden Markov Model against Uniref100

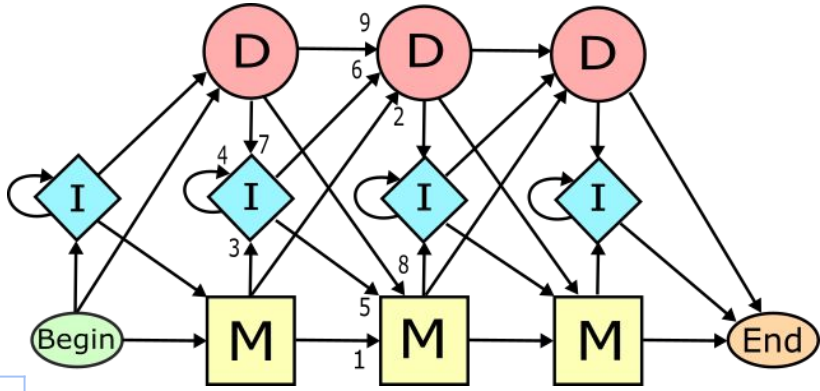
ColabFold Many against Many sequence searching (MMseqs)

Pfam

ProteinNet

ConSurfDB

CATH



Profile Hidden Markov Model

MSA	Method	#(seqs)
orig	JackHMMER (5 it), Uniref100	~15 000
colabFold	MMseqs2, Uniref30+Env.	~9 000
Pfam	from UniProt	~43 000

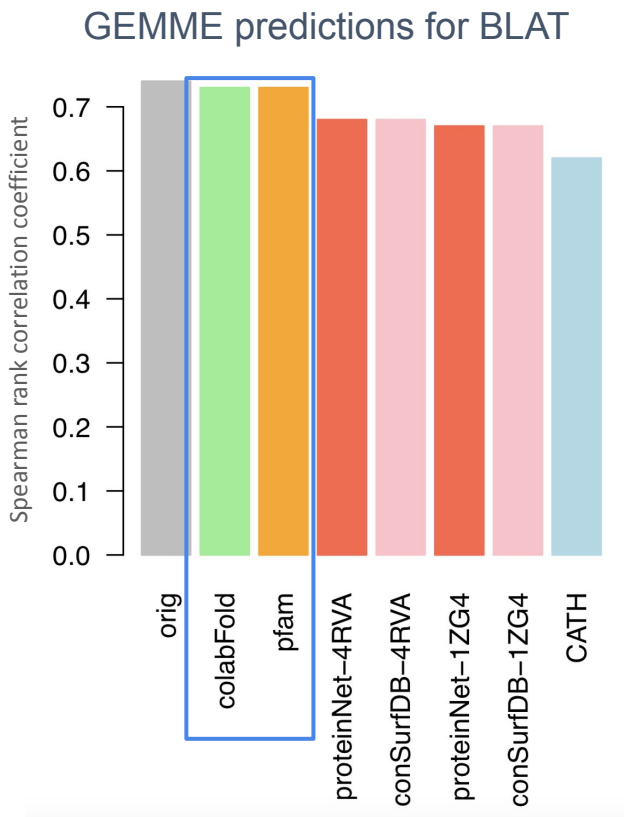
Projet

1) GEMME, basé sur l'évolution - JET(relation hierarchique) VS pLM

comparaison des prédictions ->trouver ce que GEMME a par rapport

1) gènes candidats
2) SNP dans les lignées -> représentatif ou pas, comment ça se compare aux phenotypes?
3) identify genetic variants associated with longevity
4) est-ce qu'on a suff d'information dans les parties transcrites -> challenge : Region non codante: adapter GEMME et Thorax

Development of a scalable protocol



Spearman correlation orig: 0.74

Confirmed on the benchmark set (24 proteins)

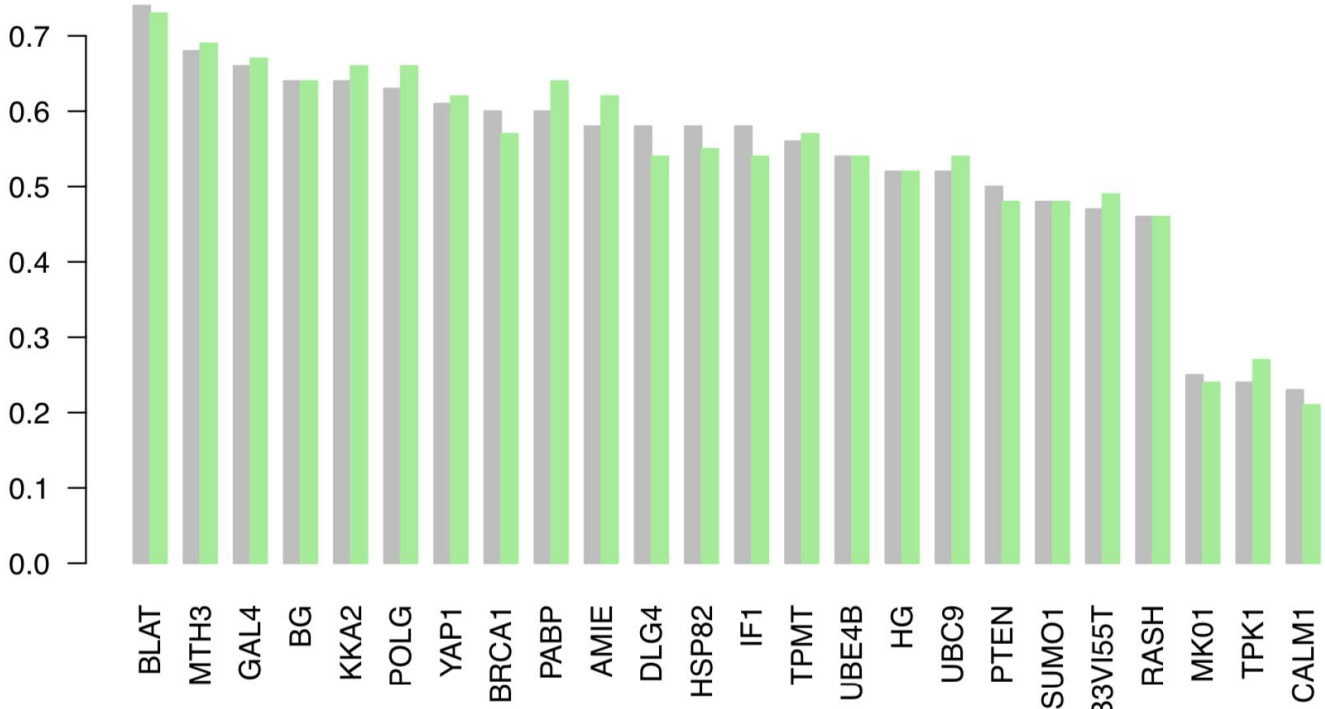
ColabFold (MMseqs) > Pfam

Scanning of the entire Fly proteome
(95%: 29 339 sequences)

Execution time for building the alignments:
4 hours per 3 000 sequences

Alignments and GEMME predictions were generated in 2-3 days for the whole proteome

M.Mirdita *et al.* biorxiv **2021**


0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
BLAT
MTH3
GAL4
BG
KKA2
POLG
YAP1
BRCA1
PABP
AMIE
DLG4
HSP82
IF1
TPMT
UBE4B
HG
UBC9
PTEN
SUMO1
B3VI55T
RASH
MK01
TPK1
CALM1