The target identification bottleneck

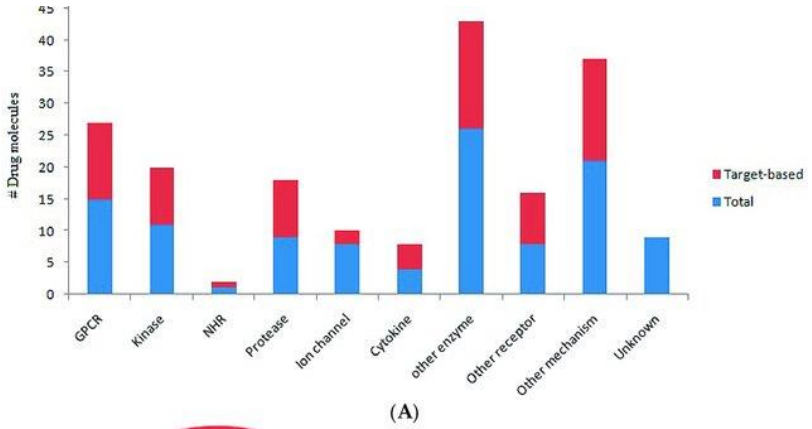
Decline of molecular mechanism discovery after 2000 and the story of the function discovery of c7orf10/SUGCT

Frank Eisenhaber

franke@bii.a-star.edu.sg

24th May 2022

Known drugs address very few human targets



Vasaikar *et al*. Biomedicines (2016) v.4, 27

- **As of May 2022, 813 human protein targets for FDA drugs (source: ProteinAtlas)**
- Mostly receptors, ion channels, enzymes, carrier molecules
- >95% of all drug targets are proteins
- Known biomolecular mechanism of action and genetic proof drastically enhance success rate of drug development and FDA approval
- **Industry complains about increasing lack of new target supply from academia**
- >5000 human genes are considered druggable

~10 years of biomolecular mechanism research on c7orf10

Phenotypic function discovery of SUGCT

Cell Mol Life Sci. 2020 Sep;77(17):3423-3439
PMID: 31722069

In collaboration with

- Joanna Niska
- Philipp Kaldis
- Sebastian Maurer-Stroh

How did we run into C7orf10?

- Early 2008: Search for human protein hits with domain models of enzymes that are below the significance threshold but have conservation of certain critical residues

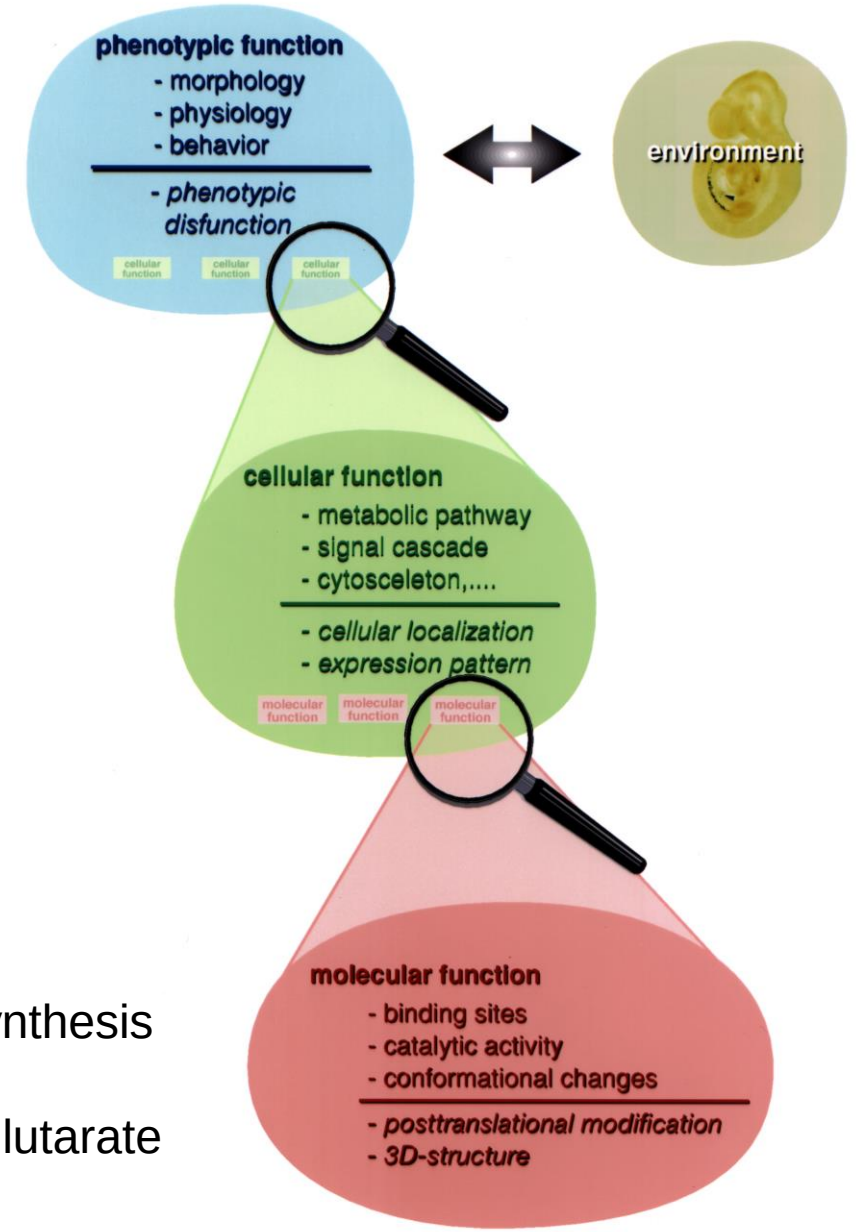
- Hits to:
  - PF02515 (CoA-transferase family)
  - Below threshold hits to CoA transferases for dicarboxylic acids
- Hypothesis: c7orf10 is a CoA transferases for some type of dicarboxylic acid
- 2008~2010: Unsuccessful biochemical experimentation ...
- Since 2011: c7orf10 mouse knock-out studies ...

Function of SUGCT/c7orf10

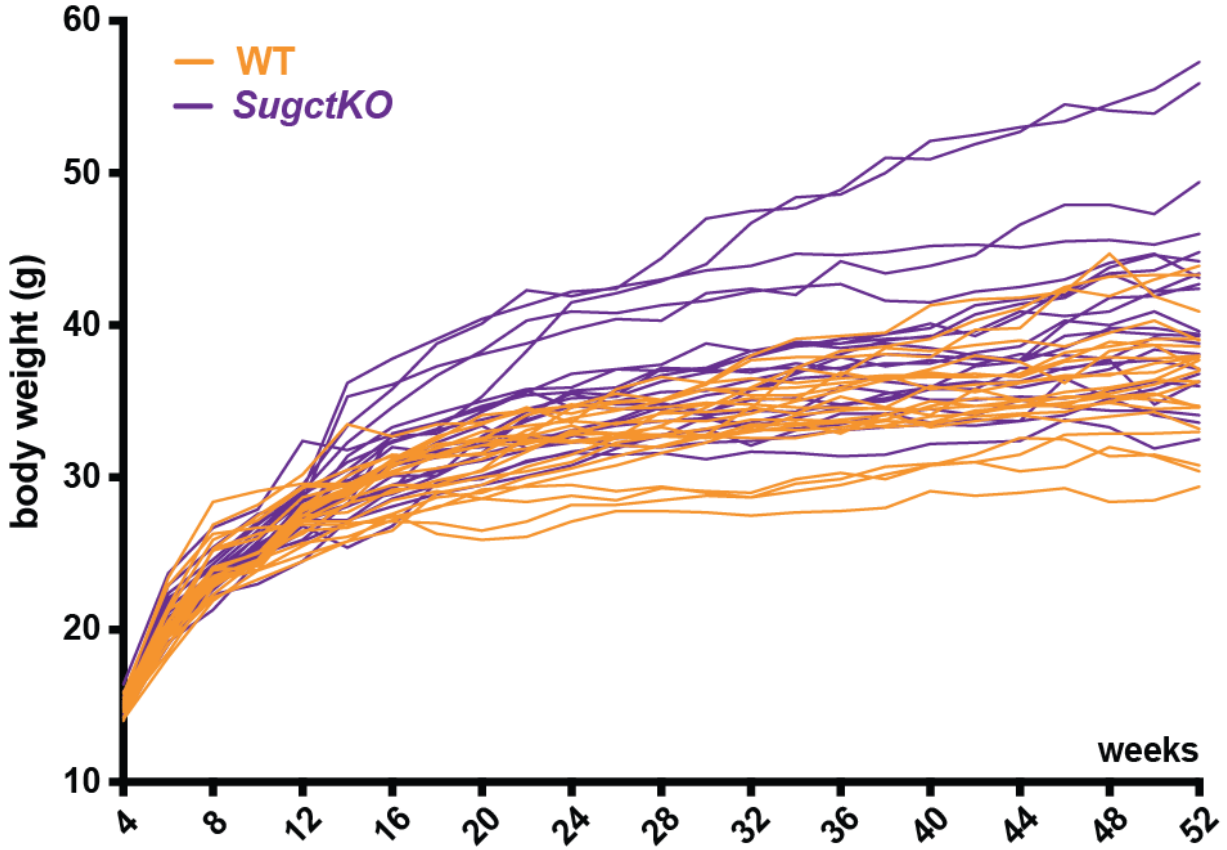
Glutaric aciduria, **no phenotype ????**
from late 2008 (PMID: 18926513)

Mitochondrial localization
(2014 - PMID: 23893049)

Catalyzes some dicarboxyl acid-CoA synthesis
(our prediction from early 2008)
Catalyzes glutaryl-CoA synthesis from glutarate
(2014 - PMID: 23893049)



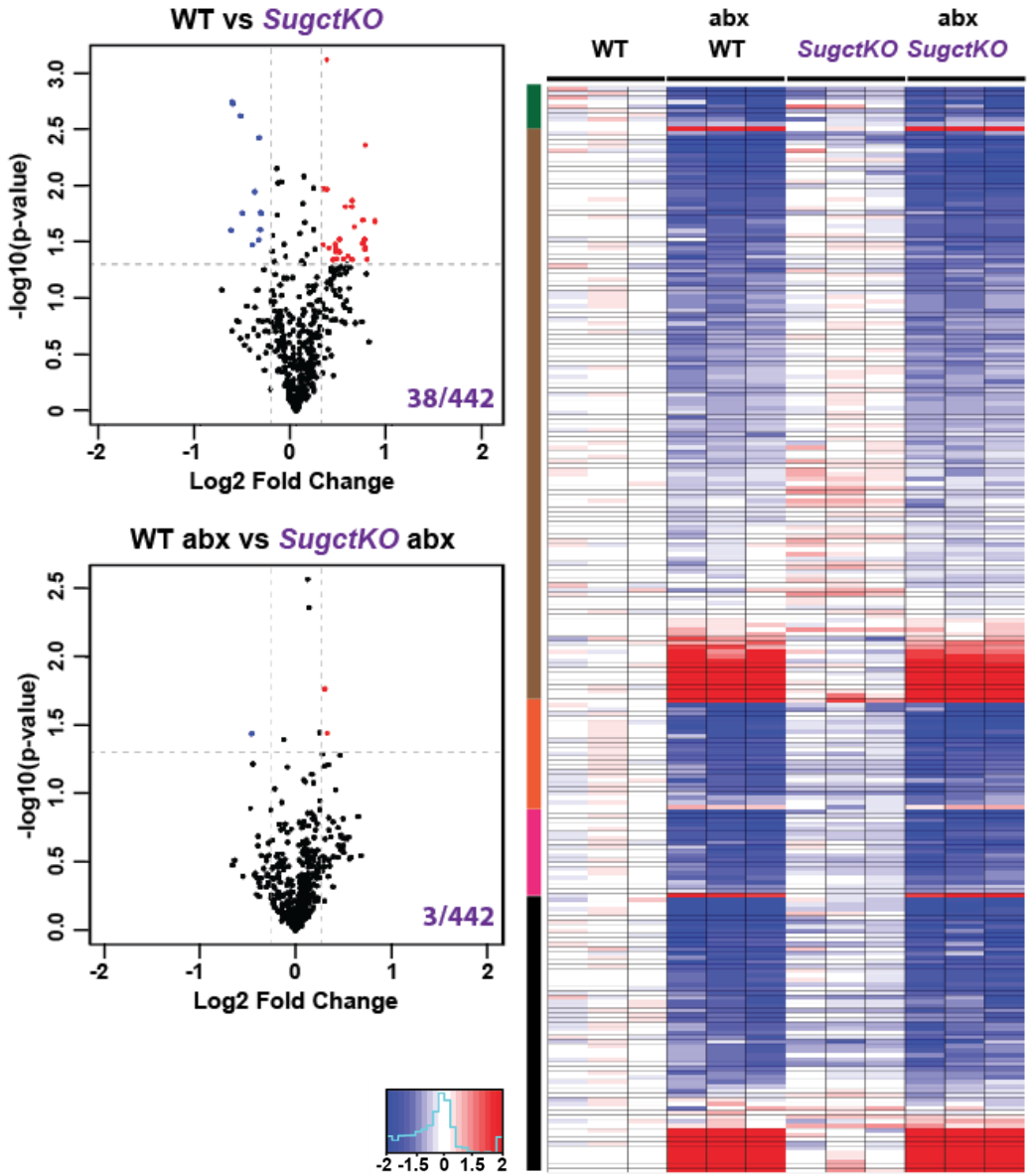
Q: Is physical and functional well-being decreased in aged *SugctKO* mice?



trend towards obesity

plus chronic inflammation in liver and kidney

Results: Metabolic differences between WT and *SugctKO* mouse plasma vanished after abx treatment



5% Possible bacterial origin
39.5% Lipids
10.5% Acylcarnitines
7.5 % Glycine conjugates
45% Others

In collaboration with Yoganathan Kanagasundaram, BII

Function of SUGCT/c7orf10 – update 2020

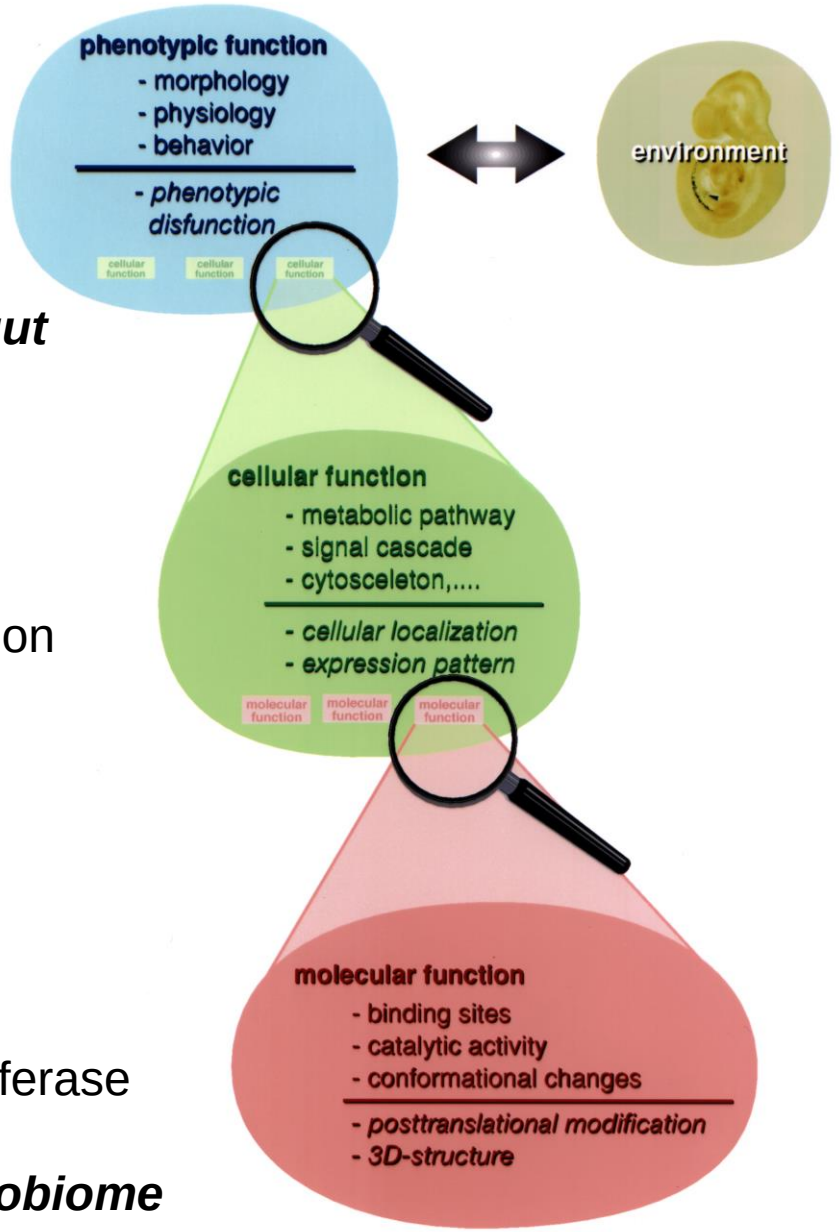
Glutaric aciduria, no acute effects
but chronic inflammation of kidney and liver
+ trend towards metabolic syndrome and obesity

Gene knockout is compensated by suppression of gut microflora with antibiotics

Mitochondrial localization

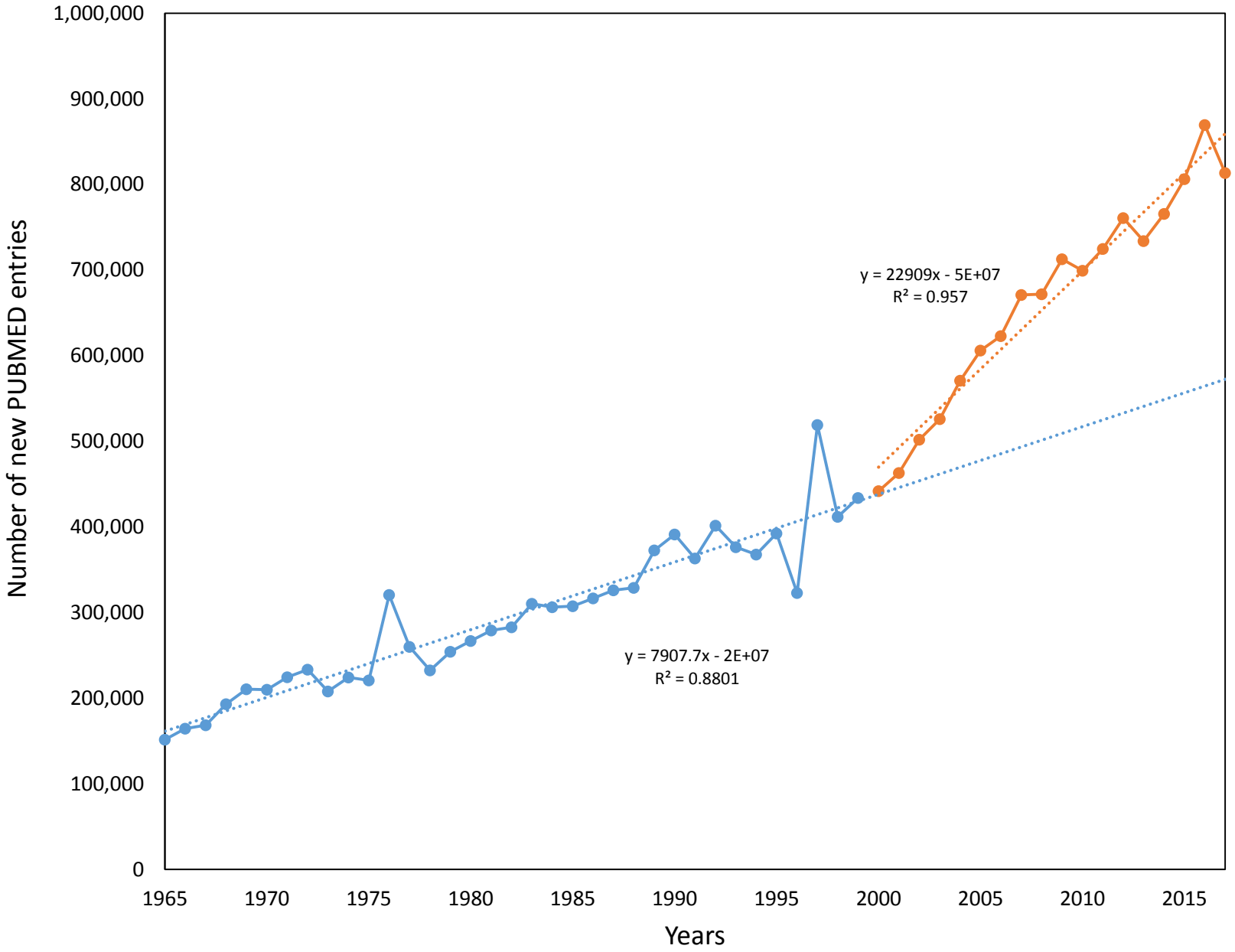
Catalyzes glutaryl-CoA synthesis from glutarate
SUGCT=succinate hydroxymethylglutarate CoA-transferase

Reserve pathway for the metabolization of gut microbiome metabolites



Many yet uncharacterized human genes have role in aging and chronic disease

as they are not essential for short-term survival



1970: 1.1 million entries
2000: 10.7 million entries
2017: 24.3 million entries

1965 – 2000
Growth of annual growth by ~8,000 entries per year

2000 – 2017
Growth of annual growth by ~23,000 entries per year

Mentioning of gene names as macro-indicator for life science research innovative quality

Usage of new or rarely occurring gene name indicates innovative life science research and new biomolecular mechanism discovery

About the darkness in the human gene and protein function space: Widely modest or absenct illumination by the life science literature and the decline in protein function discoveries since 2000

Swati Sinha, Birgit Eisenhaber, Lars Juhl Jensen, Bharata Kalbuaji, Frank Eisenhaber

Proteomics. 2018 Nov;18(21-22):e1800093. doi: 10.1002/pmic.201800093. Epub 2018 Oct 30.

Mapping of life science literature onto the human genome

Status 31st December 2017

FPE = full publication equivalent

- Analysis of publicly available articles (>2 million), Medline abstracts and full-text patents for mentioning genes/genomic regions
- In total, ~4.6 million FPEs found for ~20000 named genomic entities
- Fractional count of mentioning a gene/protein/non-coding RNA in a given paper:

$$f_i = \sum_{j \in D} \frac{n_{ij}}{n_{.j}}$$

n_{ij}: Mentioning of gene i in paper j

$n_{.j}$: Mentioning of any gene in paper j

- 1 FPE = one paper dedicated solely to the function of one gene

10 most scored proteins	ENSP00000250971	145832	Insulin (INS)
	ENSP00000295897	97977	Serum Albumin (ALB)
	ENSP00000269305	49819	Cellular tumor antigen p53 (TP53)
	ENSP00000398698	37409	Tumor necrosis factor (TNF)
	ENSP00000359663	53382	CD40 ligand (CD40LG)
	ENSP00000264708	36442	Pro-opiomelanocortin (POMC)
	ENSP00000255030	35943	C-reactive protein (CRP)
	ENSP00000308541	33341	Prothrombin (F2)
	ENSP00000272190	33025	Renin (REN)
	ENSP00000447378	32291	Maltase-glucoamylase (MGAM)

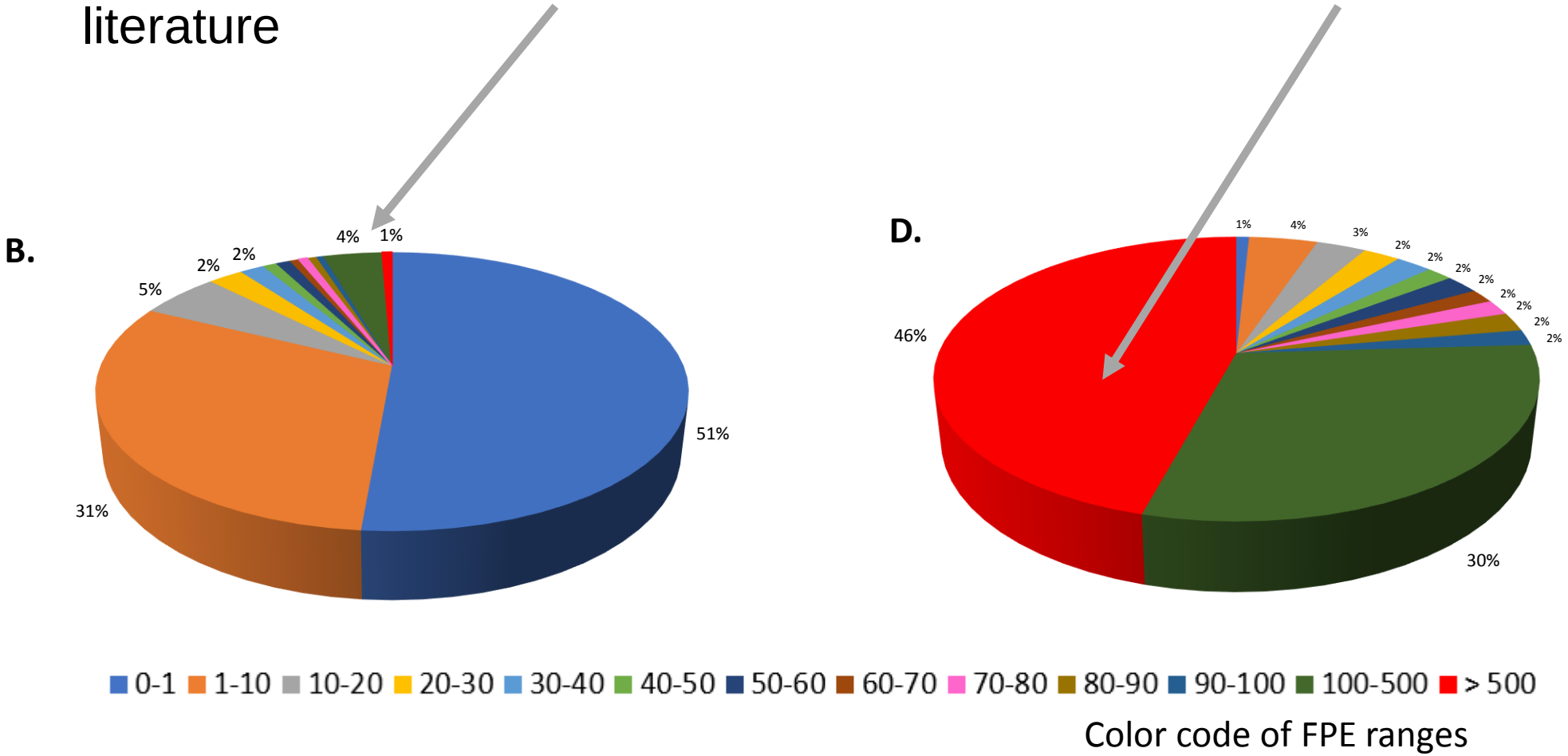
Proteins

Range for S	Number of proteins	% of **17824** proteins	Total literature score for all targets	% of total score
0-1	**1997**	11.2	836.2	0.02
1-10	**4571**	25.6	20379.0	0.44
10-20	1935	10.8	27957.7	0.60
20-30	1175	6.6	28863.9	0.62
30-40	857	4.8	29730.9	0.64
40-50	638	3.6	28597.8	0.61
50-60	531	3.0	29090.5	0.62
60-70	352	2.0	22808.3	0.49
70-80	357	2.0	26759.0	0.57
80-90	312	1.8	26456.9	0.57
90-100	282	1.6	26779.5	0.57
100-500	**3207**	18.0	736109.8	**15.75**
>500	**1610**	9.0	3666853.2	**78.50**

Conclusions: Status of life science literature in 2017

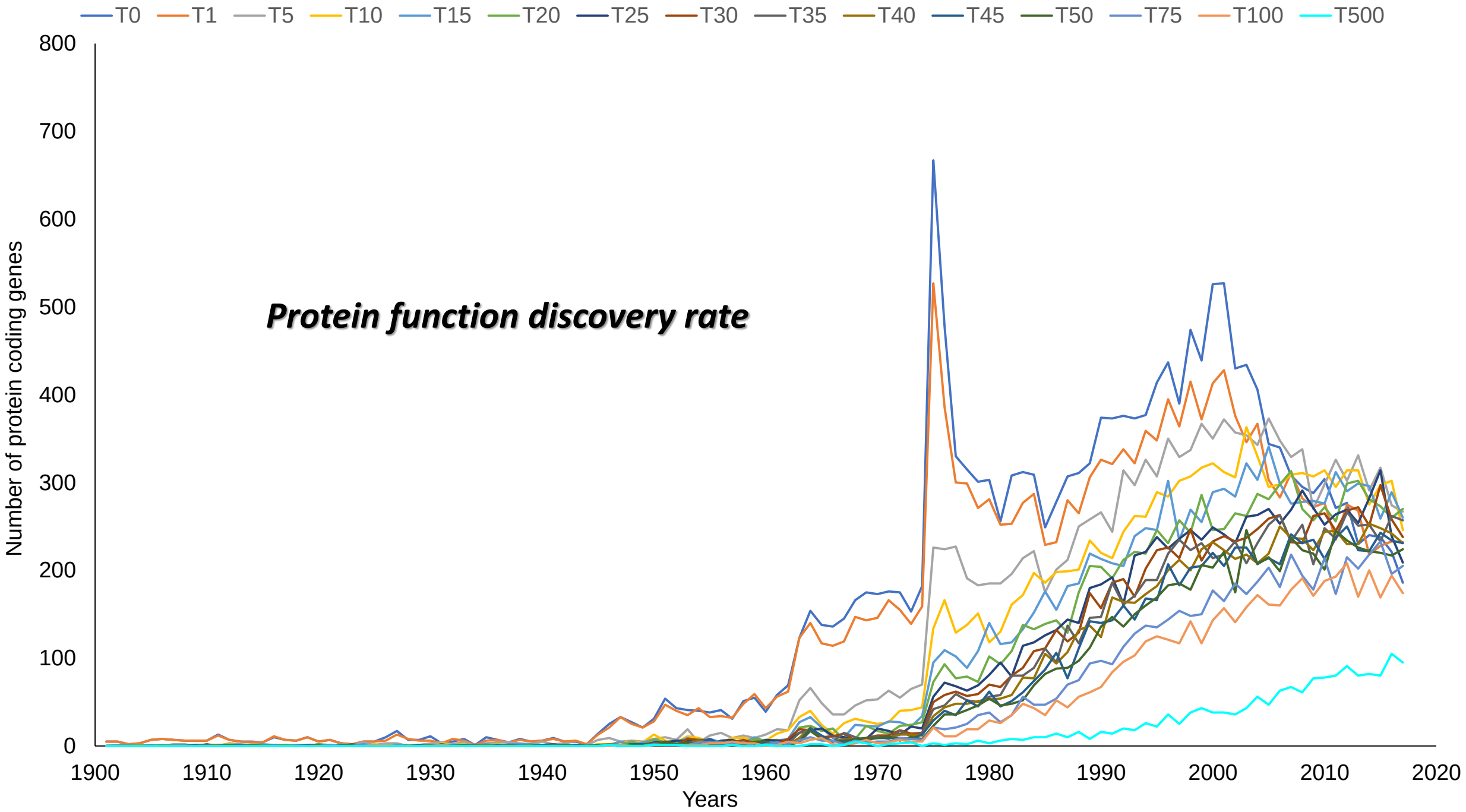
- Elite group of ~4800 protein-coding genes attracted almost 95% of all literature on protein function
- ~4000 proteins have never been studied (no literature at all)
- ~7000 proteins are almost not studied (<0.5% of all literature)
  - ~2000 proteins have less than 1 FPE
  - ~5000 proteins have less than 10 FPEs

- Similar picture for ncRNAs
- In total: 2641 ncRNA genes mentioned in the literature until 2017
- ~2200 ncRNAs (83% of all ncRNAs) attracted 5% of all ncRNA articles
- 119 elite ncRNAs (~5% of all ncRNAs) are covered by 76% of the relevant literature



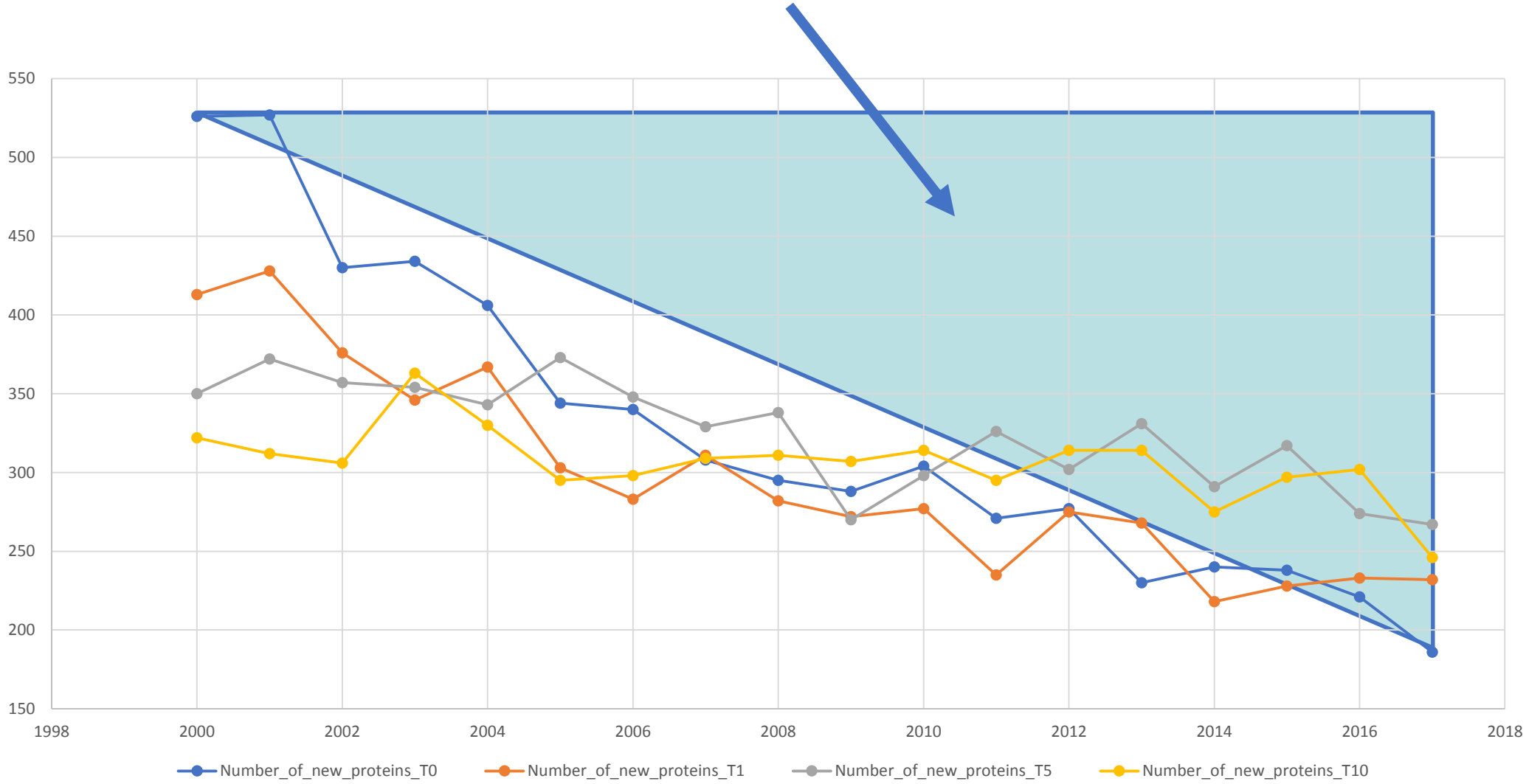
Mapping of life science literature onto the human genome

Trends in function discovery 1900 - 2017


T0
T1
T5
T10
T15
T20
T25
T30
T35
T40
T45
T50
T75
T100
T500
Protein function discovery rate
Number of protein coding genes
Years
0
100
200
300
400
500
600
700
800
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020

	1945-1974				1980-1999				2000-2017			
Tx	Slope	R^2	ρ	P-value	Slope	R^2	ρ	P-value	Slope	R^2	ρ	P-value
0	6.4	0.8495	0.922	**4.91e-13**	9.5	0.8078	0.899	**7.28e-08**	-17.97	0.9026	-0.950	**1.67e-09**
1	5.6	0.8602	0.927	**1.74e-13**	8.0	0.7779	0.882	**2.73e-07**	-11.1	0.8515	-0.923	**4.98e-08**
5	2.3	0.7811	0.8838	**9.72e-11**	9.97	0.9028	0.950	**1.51e-10**	-5.1	0.6523	-0.807	**5.06e-05**
10	1.3	0.7499	0.866	**6.41e-10**	9.6	0.9601	0.980	**4.8e-14**	-2.6	0.3582	-0.598	**0.008687**
15	0.98	0.6756	0.822	**2.56e-08**	8.4	0.8814	0.939	**9.12e-10**	-1.5	0.1637	-0.404	0.0958
20	0.8	0.6705	0.819	**3.19e-08**	9.4	0.9439	0.972	**1.05e-12**	0.9	0.0561	0.236	0.3441
25	0.6	0.6776	0.823	**2.34e-08**	9.3	0.9461	0.973	**7.26e-13**	0.5	0.0137	0.116	0.644
30	0.5	0.6230	0.789	**2.17e-07**	9.3	0.9409	0.970	**1.67e-12**	1.7	0.2719	0.521	0.02647
35	0.5	0.5395	0.735	**3.82e-06**	9.7	0.9644	0.982	**1.7e-14**	2.1	0.3438	0.586	0.01054
40	0.4	0.5399	0.735	**3.77e-06**	9.2	0.9733	0.987	**1.31e-15**	1.4	0.3268	0.571	0.01319
45	0.4	0.4743	0.689	**2.58e-05**	8.9	0.9459	0.973	**7.55e-13**	1.3	0.2801	0.529	0.0239
50	0.3	0.5031	0.709	**1.14e-05**	8.7	0.9741	0.987	**9.8e-16**	1.1	0.1087	0.329	0.1815
75	0.3	0.6105	0.781	**3.46e-07**	7.3	0.9618	0.981	**3.29e-14**	2.3	0.3994	0.631	**0.0049**
100	0.2	0.4628	0.680	**3.53e-05**	6.2	0.9187	0.958	**3.0e-11**	2.5	0.4842	0.695	**0.00134**
500	0.1	0.2538	0.504	**0.004537**	1.7	0.8232	0.9073	**3.41e-08**	3.7	0.9029	0.950	**1.61e-09**

T0 – at least one mentioning of gene name in the scientific literature
~2800 protein functions missing out of ~4000 not mentioned yet



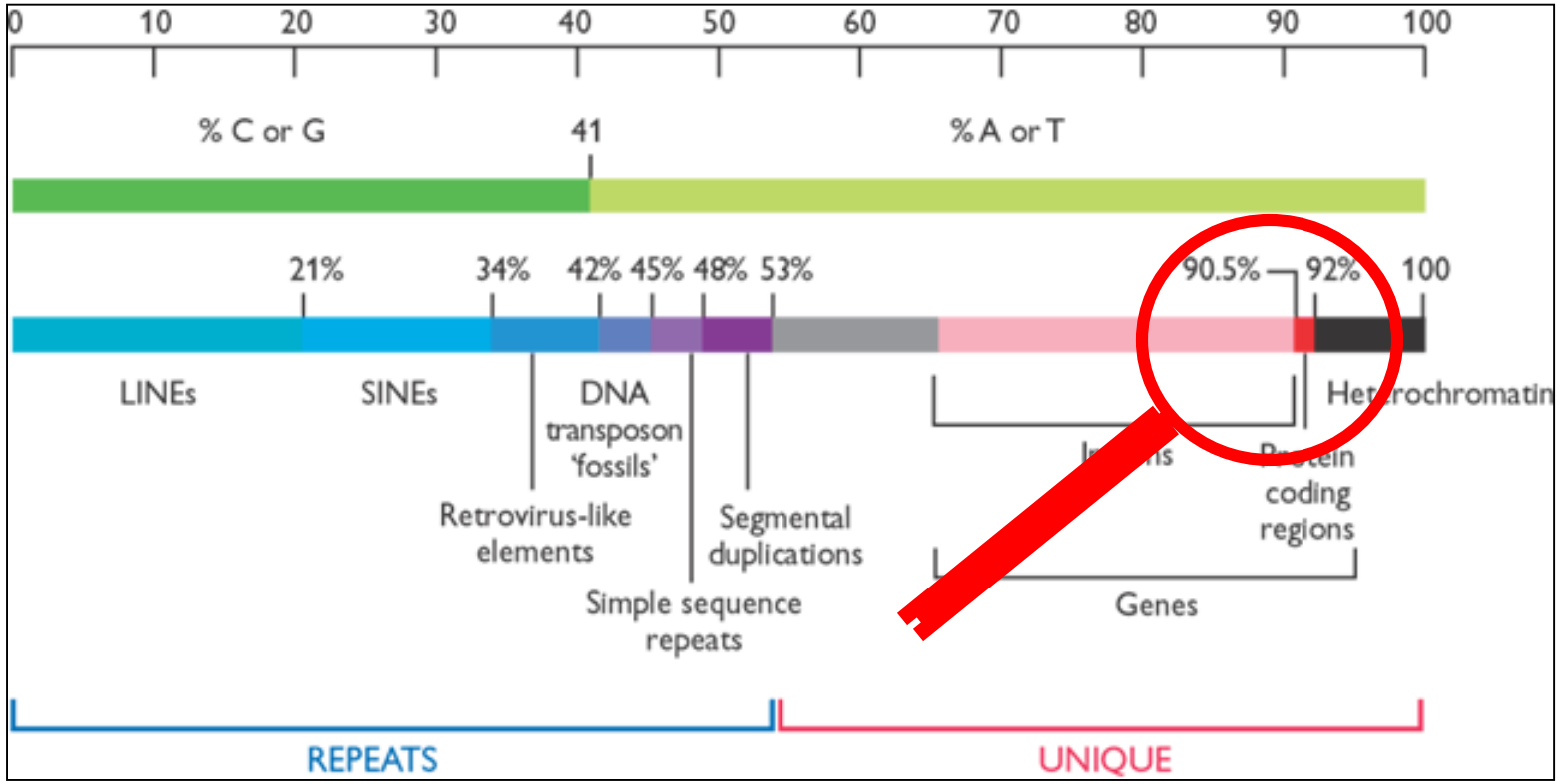
Protein function discovery rate

- Trend for strong, super-linear growth of new function discovery until 2000
- Steady decline of function discovery after 2000
- Accumulated deficit 2000-2018: about 2000 protein functions should have been known, if rate has remained at least constant since 2000
- Presently fastest growing group of proteins: T500 and T100

Darkness in the Human Gene and Protein Function Space: Widely Modest or Absent Illumination by the Life Science Literature and the Trend for Fewer Protein Function Discoveries Since 2000.

Sinha S, Eisenhaber B, Jensen LJ, Kalbuaji B, Eisenhaber F. Proteomics. 2018 Sep 28:e1800093. doi: 10.1002/pmic.201800093

The human genome – not only 20000-25000 genes



~10000 human genes are functionally not or almost not characterized; >1000 biomolecular pathways remain to be discovered , 50%-80% of human biology is still unknown

98.5% of the genome does not code for proteins; functions of its derived RNAs are mostly enigmatic

When will we understand the human genome?

If extrapolated from the decade 2001-2011, it will take at least a century!

A decade after the first full human genome sequencing: when will we understand our own genome?
Eisenhaber F. J Bioinform Comput Biol. 2012 10:1271001

A new gene function is reported not every week; less than 100 per year; a century for >10000 genes yet to characterize (~75% of human biology at the level of pathways is yet to be discovered?)

Yet,

faster and cheaper sequencing

means mainly (much)

more non-understood sequences

Possible Explanations

- ***Decline is not due to universities in Cairo, Dhaka or Nairobi but due to well-known top places***
- One gene function discovery = …
  - sophisticated application of top research technologies
  - more than a decade with a couple of teams engaged
  - ~10 million USD
- ***Structure of financing drives people away from risky projects***
  - Commissions select projects instead of brilliant individuals
  - Short-term contracts even for faculty/Pis
  - Need for preliminary data at grant application stage (with no budget funding for initial stage

Conclusions

- Despite the tremendous societal biomedical and biotechnology needs, many of them will not be satisfied in the time to come (~decades?) due to scientific blocks.
- ***Deficit of good drug targets***

- Dramatic improvement where
  - Only if biomolecular mechanisms are understood
    - Targeted sequencing of genes for mutations with known effects and drugs
  - Comparison of non-understood genome sections suffices

  - Technical developments (miniaturization, computerized support, etc.) are helpful (diagnostics, verification of source, etc.)

- Trend for lottery-type business (pharma, stem cells)
- ***Great market opportunities for charlatans and quacksalvers of all shades to exploit hopes*** (life style genomics, cancer treatment, slimming, mental boosting, bust enhancement, etc.).