LGN5809 - Genética Molecular

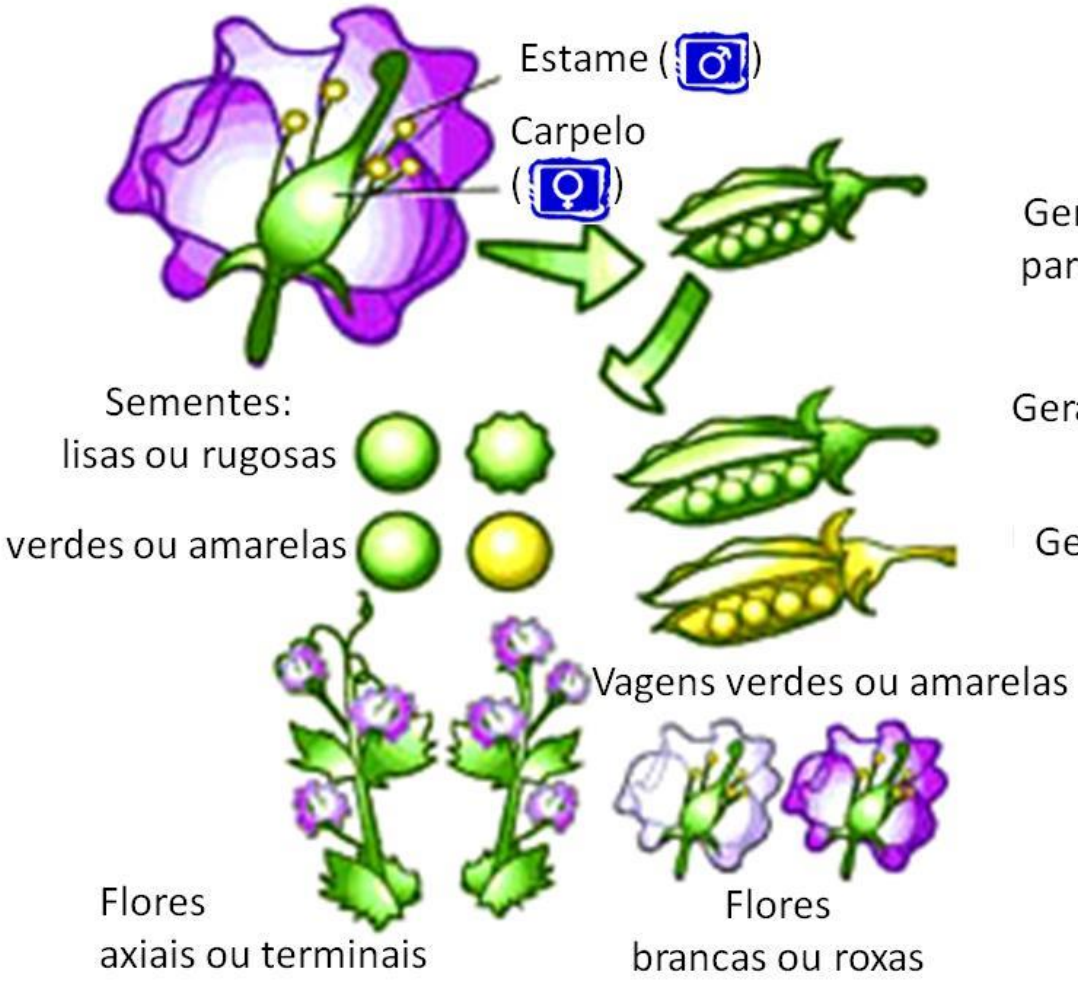
ARQUITETURA GENÔMICA: DOS GENES AO GENOMA

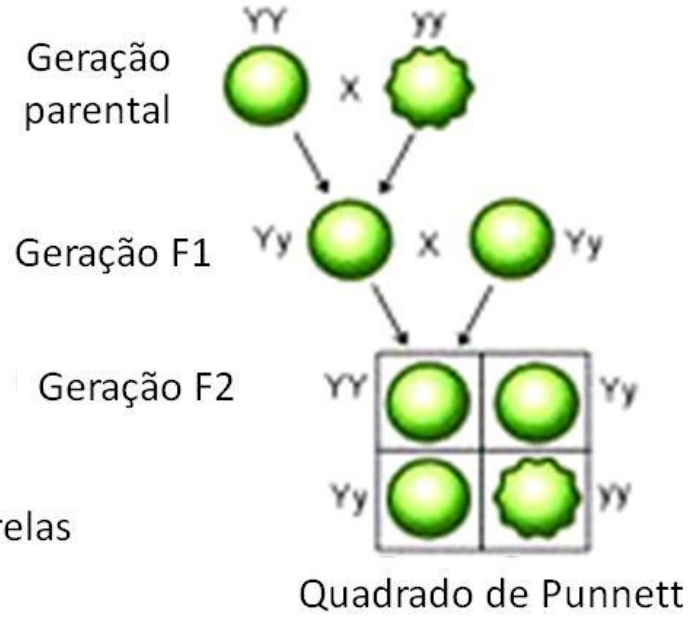
Maria Carolina Quecine
Departamento de Genética
mquecine@usp.br

SUMÁRIO

- Fluxo da informação genética;
- Conceito de genoma;
- Anotação de genomas;
- Conceito de genes;
- DNA lixo;
- Genoma de plantas;
- Desafios da genomica.
- Próxima aula

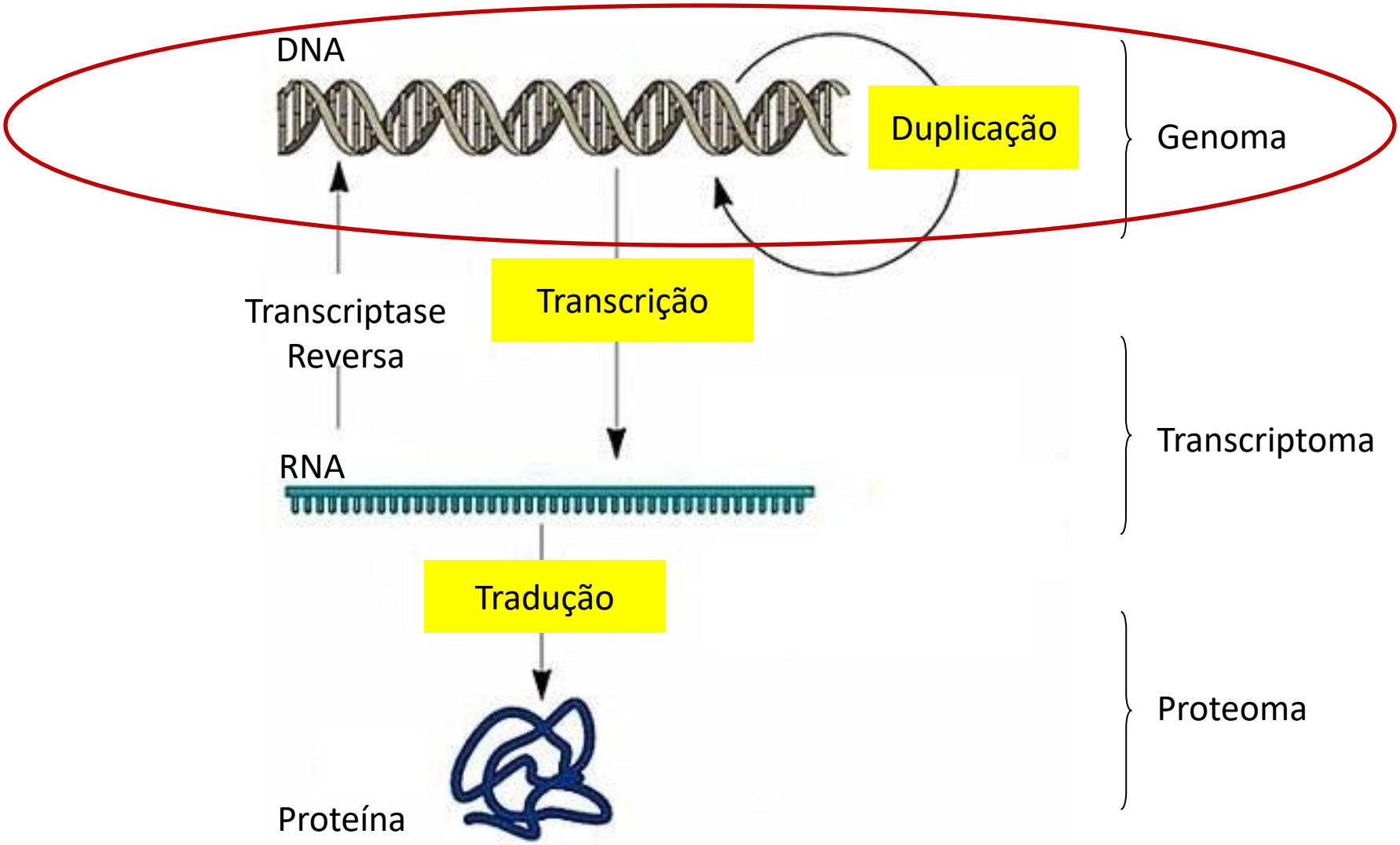
MENDEL: *FATORES CONSTANTES* QUE CONTROLAM CARACTERÍSTICAS FENOTÍPICAS





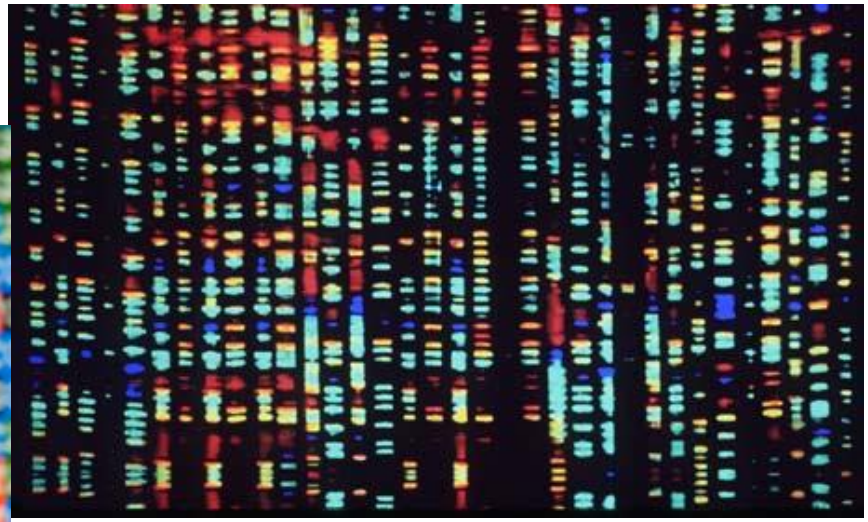
Leis de Mendel (1866)

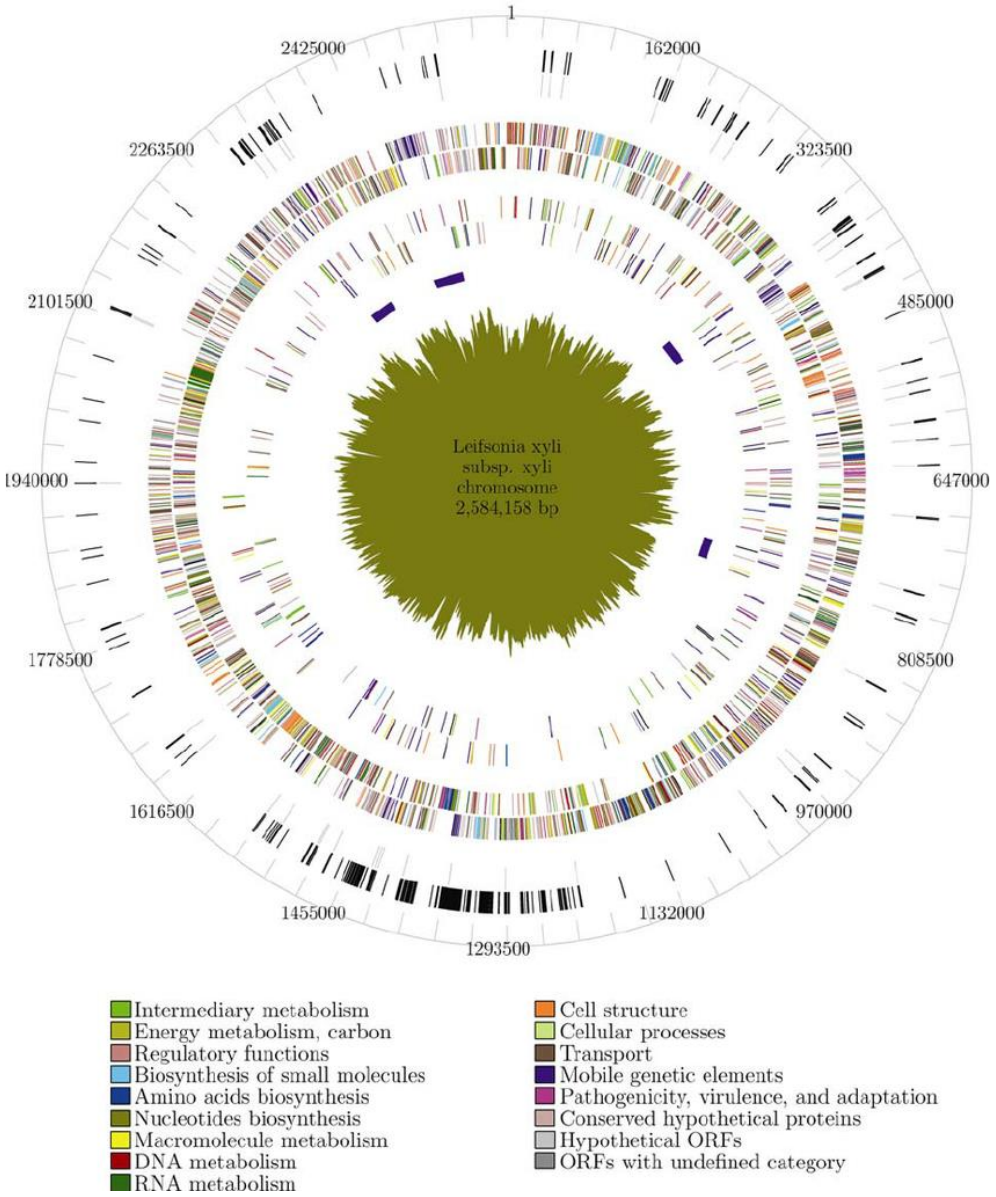
DOGMA CENTRAL DA BIOLOGIA



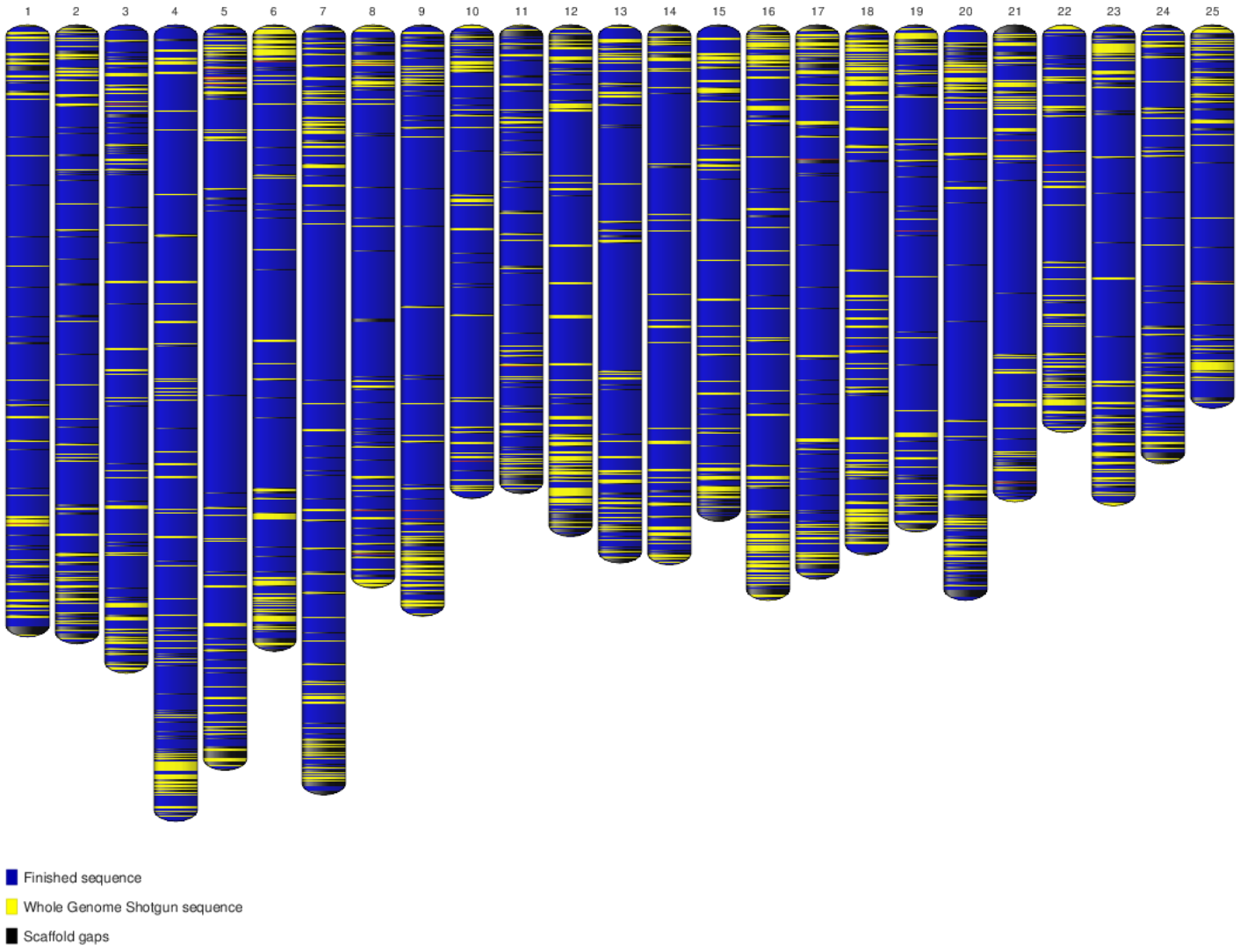
o **genoma** é toda a informação hereditária de um organismo que está codificada em seu **DNA**. Isto inclui tanto os **genes como as sequências não-codificadoras**.

????

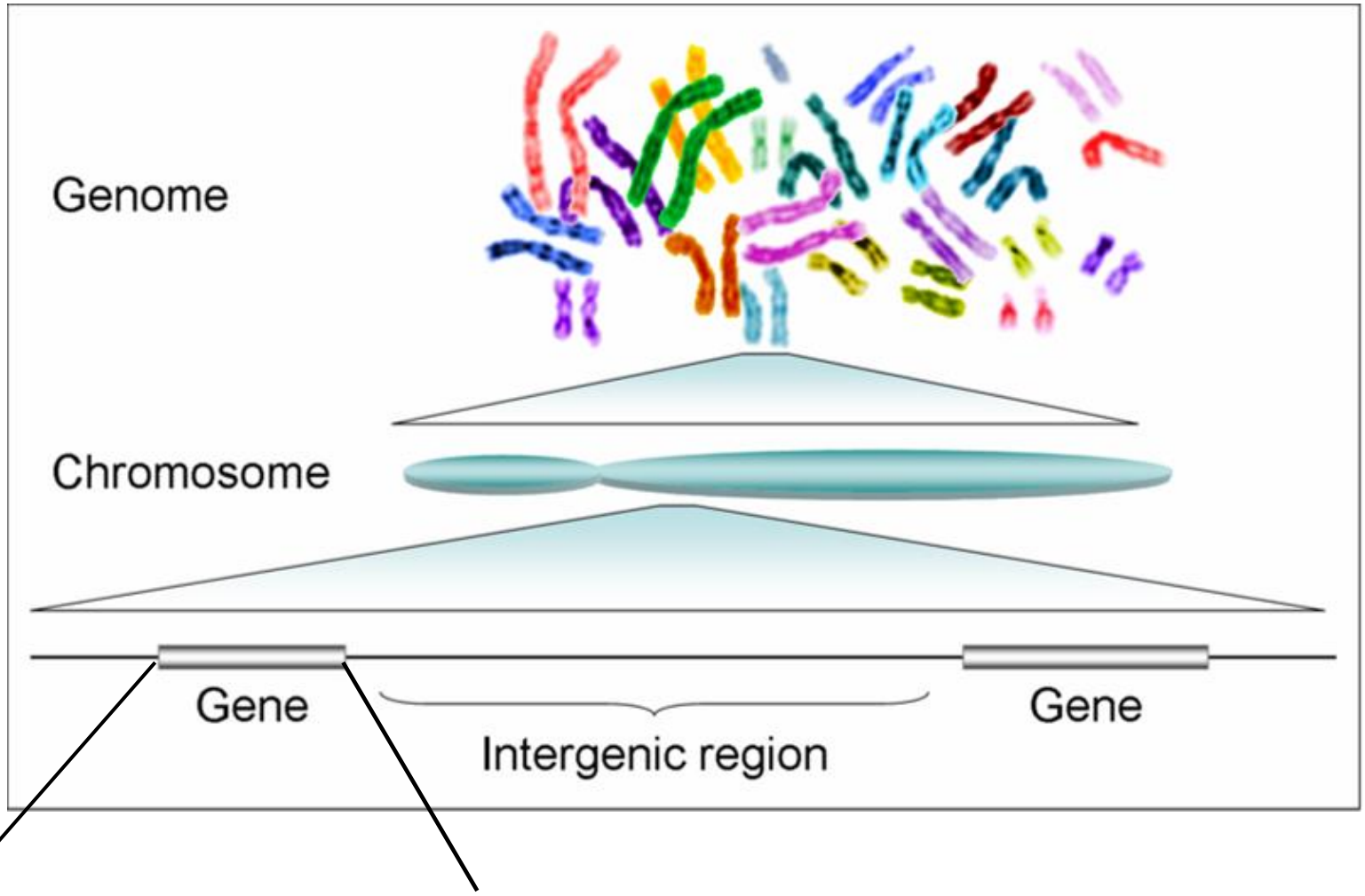




Monteiro-Vitorello et al. 2004



http://www.sanger.ac.uk/science/data/zebrafish-genome-project



attgcgcgataccccggctaa ….

TTCATACTTGGTTAAGACCTTTACAAGCCGACCAACGTGGTGACAGTGTCGTCCTTTA
CGCACCGAATCCCTTTATCATTGAATTAGTAGAAGAGCGATACTTAGGACGTCTTCGG
ATGGAATCTTGGTCCCGTTGCCTGGAACGTCTTGAAACTGAATTCCCGCCAGAAGATG
TTCATACTTGGTTAAGACCTTTACAAGCCGACCAACGTGGTGACAGTGTCGTCCTTTA
CGCACCGAATCCCTTTATCATATTGAATTAGTAGAAGAGCGATACTTAGGACGTCTTC
GGGAATTGTTATCCTATTTCTCAGGAATACGTGAAGTAGTCCTTGCAATTGGCTCACG
ACCTAAAACAACAGAACTACCCGTACCAGTAGACACTACAGGACGTTTGTCTTCAACA
GTCCCATTTAACGGAAATCTCGACACACACTATAACTTTGATAATTTTGTTGAGGGAC
GAAGCAATCAACTCGCTCGTGCTGCAGCTTGGCAAGCGGCACAGAAACCGGGAGACCG
TACTCACAACCCTCTATTGCTCTATGGTGGGACTGGTTTGGGTAAAACCCATTTAATG
TTTGCTGCAGGTAACGTAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATCTTC
GTTCGGAACAGTTTTTCAGCGCCATGATAAGAGCGTACAAGATAAAAGTATGGATCAT
AAGGGTAAAACCCATTTAATGTTTGCTGCAGGTAACGTAATGCGGCAAGTAAACCCAA
CTTATAAAGTAATGTATCTTCGTTCGGAACAGTTTTTCAGCGCCATGATAAGAGCGTA
CAAGATAAAAGTATGGATCATAAGGGTAAAACCCATTTAATGTTTGCTGCAGGTAACG
TAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATCTTCGTTCGGAACAGTTTTT
CAGCGCCATGATAAGAGCGTACAAGATAAAAGTATGGATCATAAAACGTAATGCGGCA
AGTAAACCCAACTTATAAAGTAATGTATCTTCGTTCGGAACAGGGTAAAACCCATTTA
ATGTTTGCTGCAGGTAACGTAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATC
TTCGTTCGGAACAGTTTTTCAGCGCCATGATAAGAGCGTACAAGATAAAAGTATGGAT
CATAAAACGTAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATCTTCGTTCGGA
ACAAAAACGTAATGCGGCAAGTAAACCCAACTTATAAAGTAATGTATCTTCGTTCGGA
AG

GENOME ANNOTATION: FROM SEQUENCE TO BIOLOGY

Lincoln Stein

O que buscamos são sequências com alguma função biológica!!!

No mínimo - entre 5 - 15% dos genes são negligênciados!!

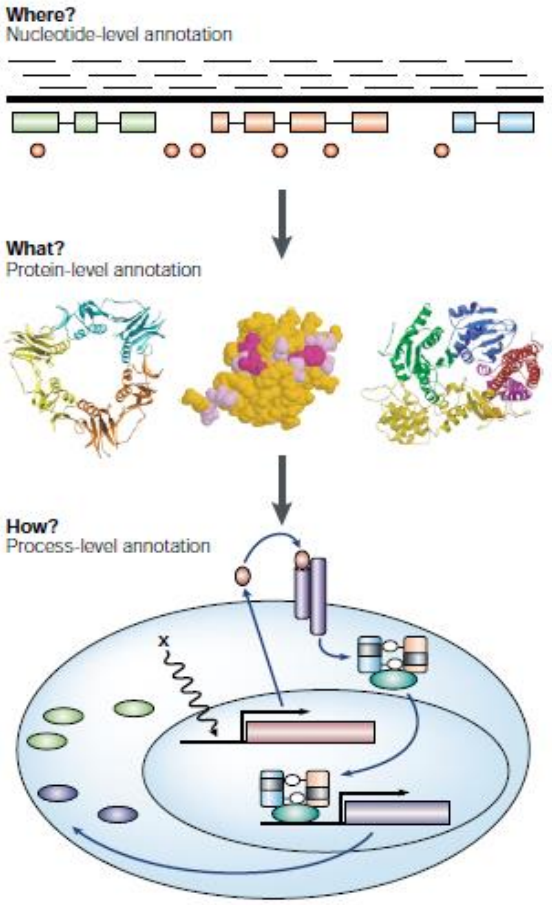


Figure 1 | **The three layers of genome annotation: where, what and how?**

Doi: 10.1038/350805296

Haemophilus influenza - 85% dos 1.8-Mb são regiães codantes.
Leveduras – menos que 70%.

Aves e vermes – menos que 25%.
Humanos – menos que 1%

Vários programas de predição

Box 1 | **Resources and tools I**

The following list provides brief descriptions of some of the software tools and resources mentioned in the article. Online, links are provided to these resources from this box, and from the text.

BLASTN, BLASTX, BLASTP, PSI-BLAST ..... http://www.ncbi.nlm.nih.gov/BLAST/
This family of sequence-similarity search tools allows you to rapidly search a query protein or nucleotide sequence against a large database of sequences, to identify sequences that are similar to the search sequence.

Ensembl ..... http://www.ensembl.org
This web site, a joint project of the European Bioinformatics Institute and the Sanger Centre, seeks to make available a high-quality, consistent set of annotations on the human and mouse genomes.

e-PCR ..... http://www.ncbi.nlm.nih.gov/genome/sts/epcr.cgi
BLAST does not work well with short sequences, such as PCR primers. The ePCR program was developed to search rapidly for a primer pair (using their sequences and known physical separation) in a large sequence, such as a genome.

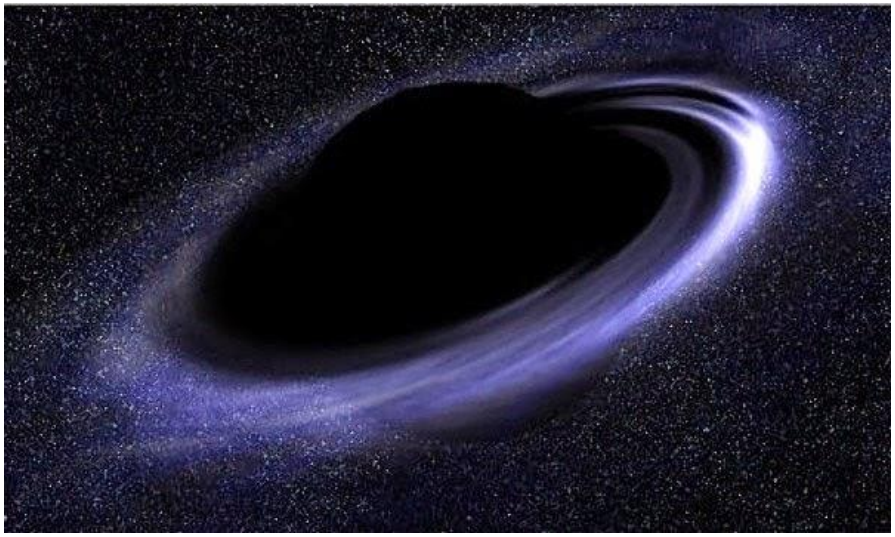
FlyBase ..... http://www.flybase.org/
A fully realized model organism system database, presenting curated annotations on the *Drosophila melanogaster* genome, as well as rich information on the genetics of the organism, mutant strains and molecular resources.

GeneMark.hmm ..... http://genemark.biology.gatech.edu/GeneMark/
Another gene-prediction program that uses hidden Markov models (HMMs). The online version supports numerous eukaryotic and prokaryotic genomes.

Genie ..... http://www.fruitfly.org/seq_tools/genie.html
The gene-prediction program used to annotate genes in *Drosophila melanogaster*. The online version has been trained for human and *Drosophila* sequence.

GENSCAN ..... http://genes.mit.edu/GENSCAN.html
This is probably the most widely used gene-prediction program. It uses HMMs to predict the presence of a gene given the raw DNA sequence. The online version provides prediction services for vertebrates, *Arabidopsis thaliana* and maize.

Grail http://compbio.ornl.gov/Grail-1.3/
One of the oldest gene-prediction programs still in use, this software uses a neural network to predict genes. The online version provides gene-prediction services for human, mouse, *Arabidopsis*, *Drosophila* and *Escherichia coli* sequences.



Auxilio com cDNA
Mas tem problemas:
Perdas de transcritos (raros)
Contaminação

Doi: 10.1038/350805296

biology

MDPI

Review

Review on the Computational Genome Annotation of Sequences Obtained by Next-Generation Sequencing

Girum Fitihamlak Ejigu and Jaehee Jung *

Department of Information and Communication Engineering, Myongji University, Yongin-si 17058, Gyeonggi-do, Korea; girumfitex@gmail.com

* Correspondence: jhjung@mju.ac.kr

Received: 21 August 2020; Accepted: 16 September 2020; Published: 18 September 2020

check for updates

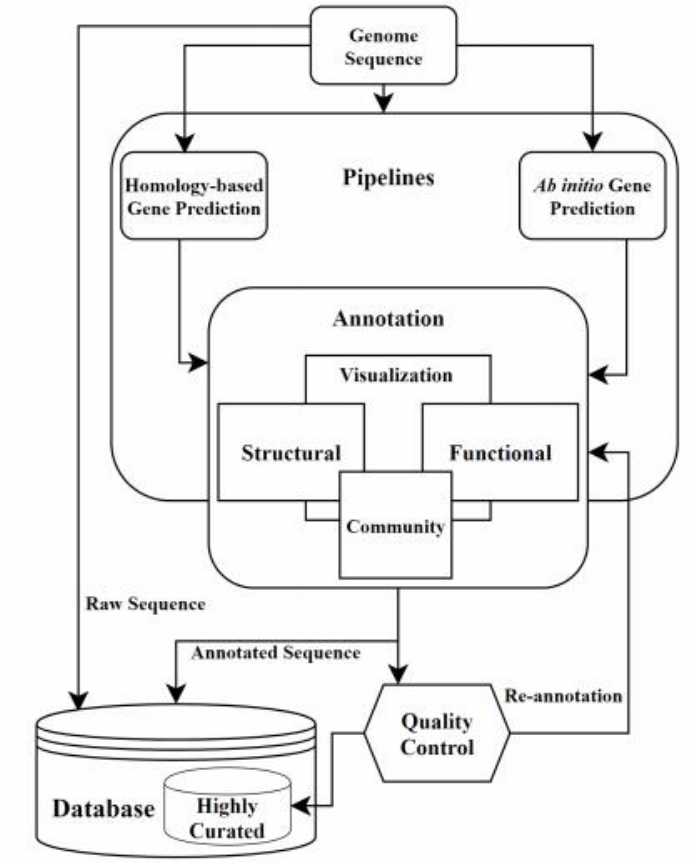
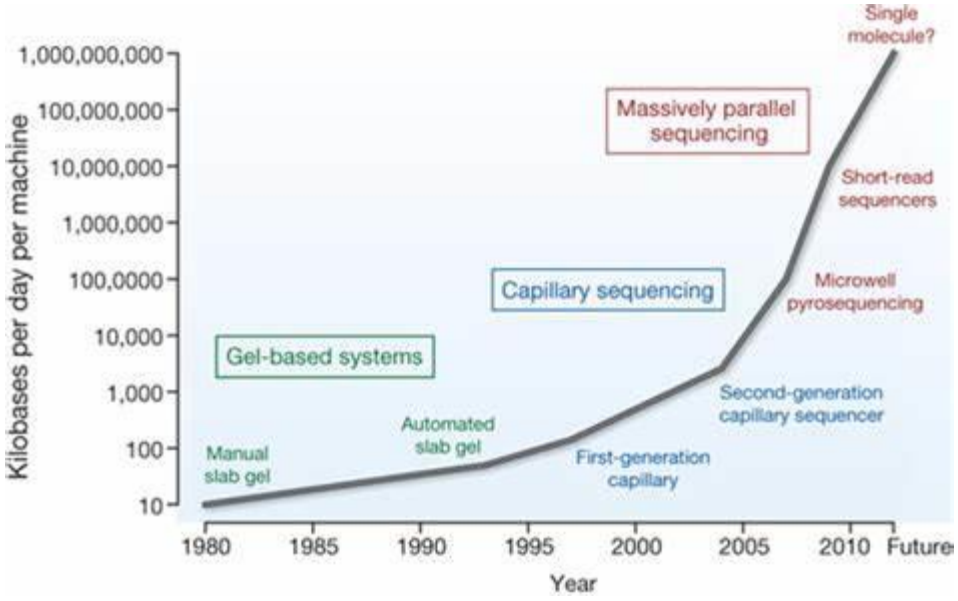
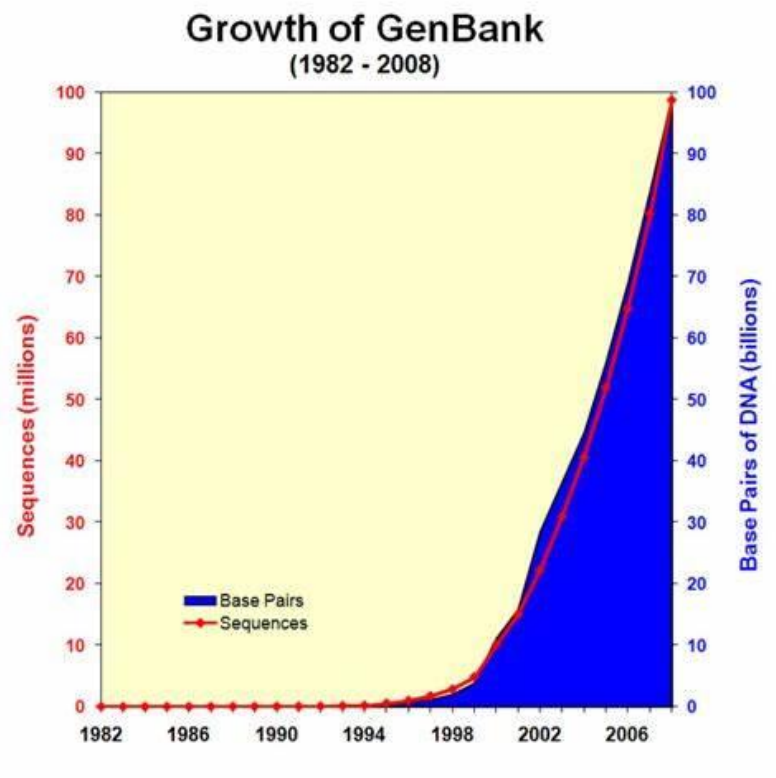


Figure 1. Genome annotation workflow.

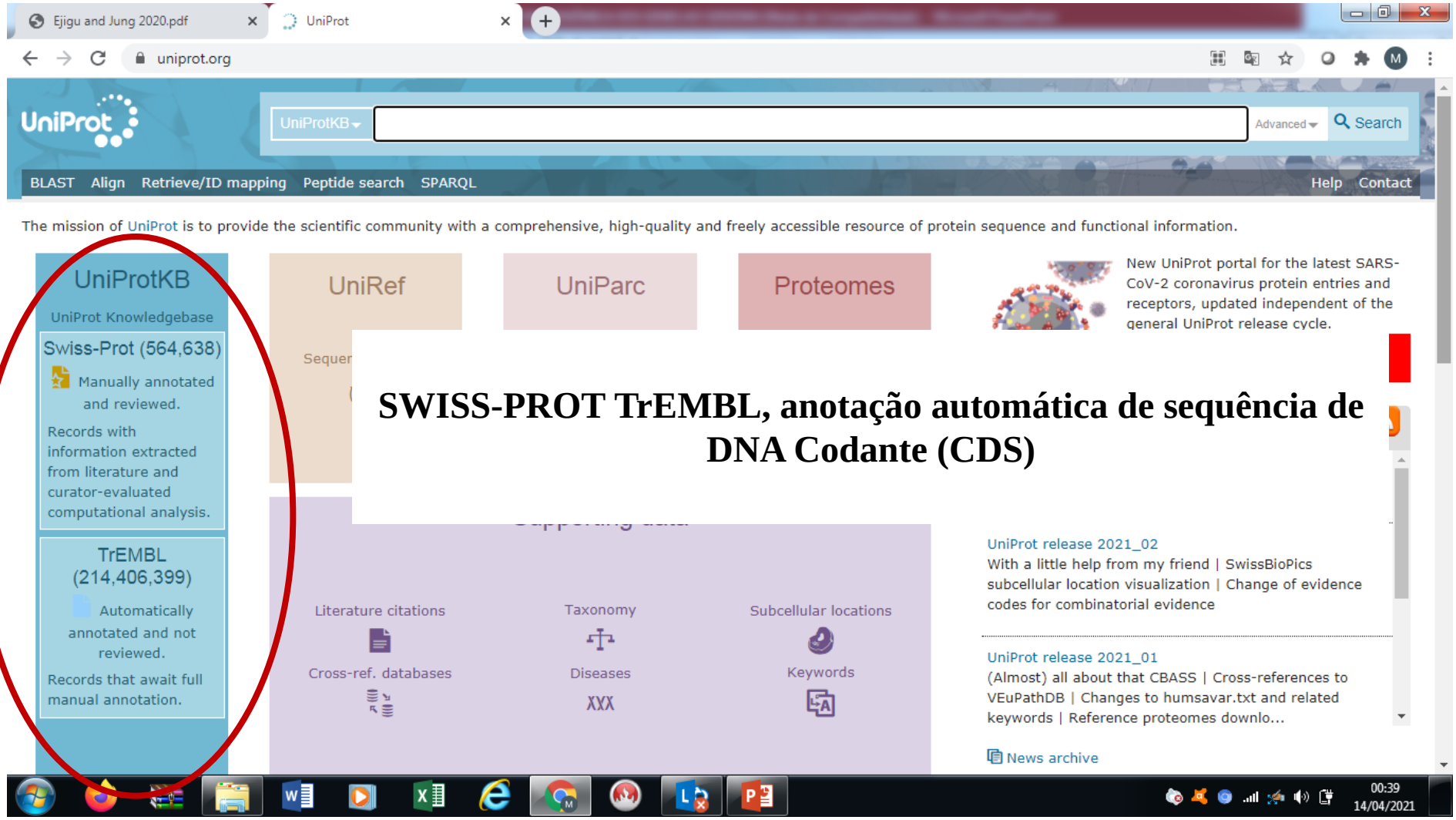


Desatualizados...mas, é muita sequencia!!

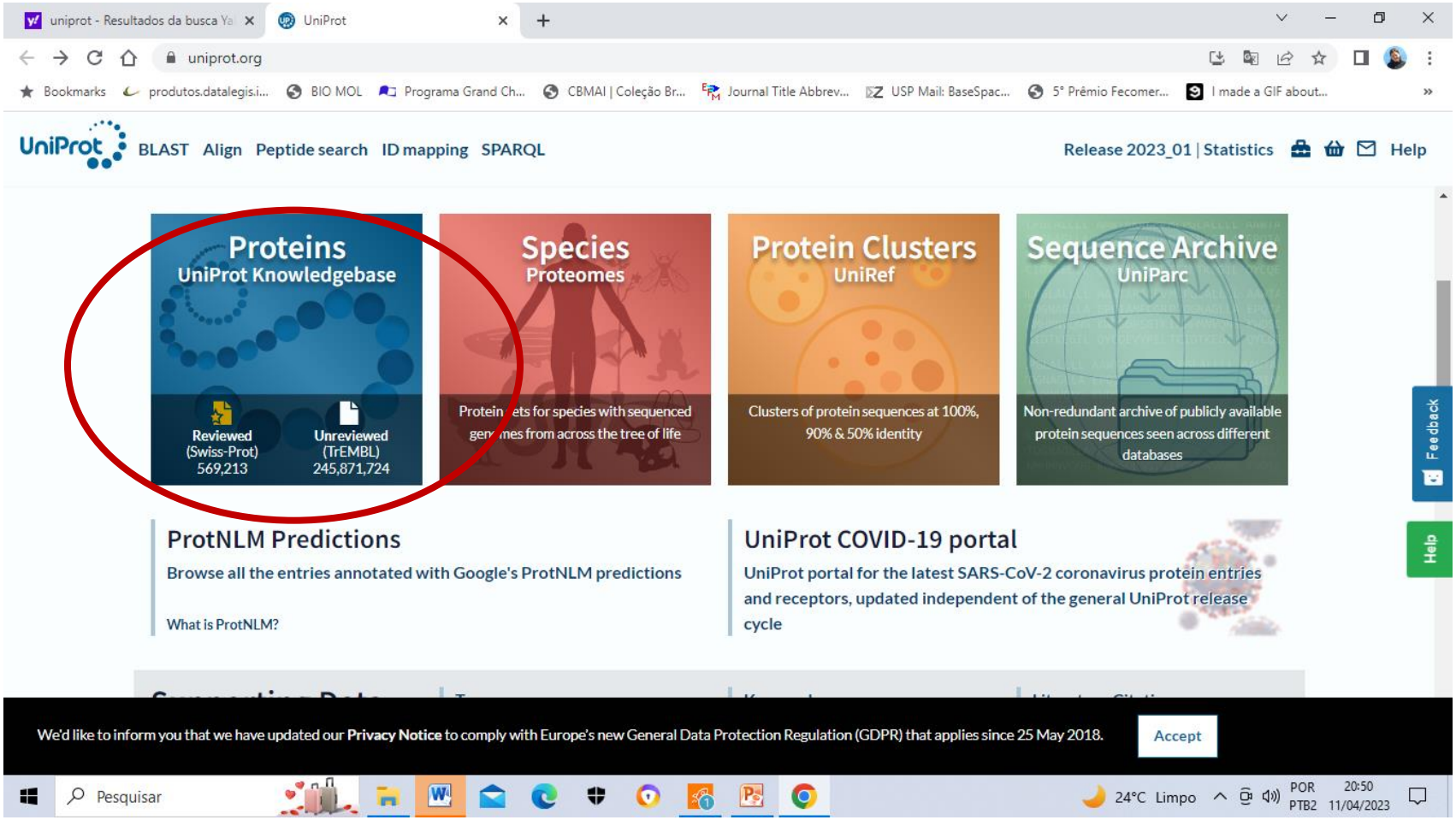


Muito dado – 2017!

…2021…


Ejigu and Jung 2020.pdf
UniProt
uniprot.org
UniProt
UniProtKB
Advanced
Search
BLAST
Align
Retrieve/ID mapping
Peptide search
SPARQL
Help
Contact
The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.
UniProtKB
UniProt Knowledgebase
Swiss-Prot (564,638)
Manually annotated and reviewed.
Records with information extracted from literature and curator-evaluated computational analysis.
TrEMBL (214,406,399)
Automatically annotated and not reviewed.
Records that await full manual annotation.
UniRef
UniParc
Proteomes
New UniProt portal for the latest SARS-CoV-2 coronavirus protein entries and receptors, updated independent of the general UniProt release cycle.
SWISS-PROT TrEMBL, anotação automática de sequência de DNA Codante (CDS)
Literature citations
Taxonomy
Subcellular locations
Cross-ref. databases
Diseases
Keywords
XXX
UniProt release 2021_02
With a little help from my friend | SwissBioPics subcellular location visualization | Change of evidence codes for combinatorial evidence
UniProt release 2021_01
(Almost) all about that CBASS | Cross-references to VEuPathDB | Changes to humsavar.txt and related keywords | Reference proteomes downlo...
News archive
00:39
14/04/2021

Muito dado - 2023!



Grande desafio…

RNAs não codantes??

Regiões regulatórias??

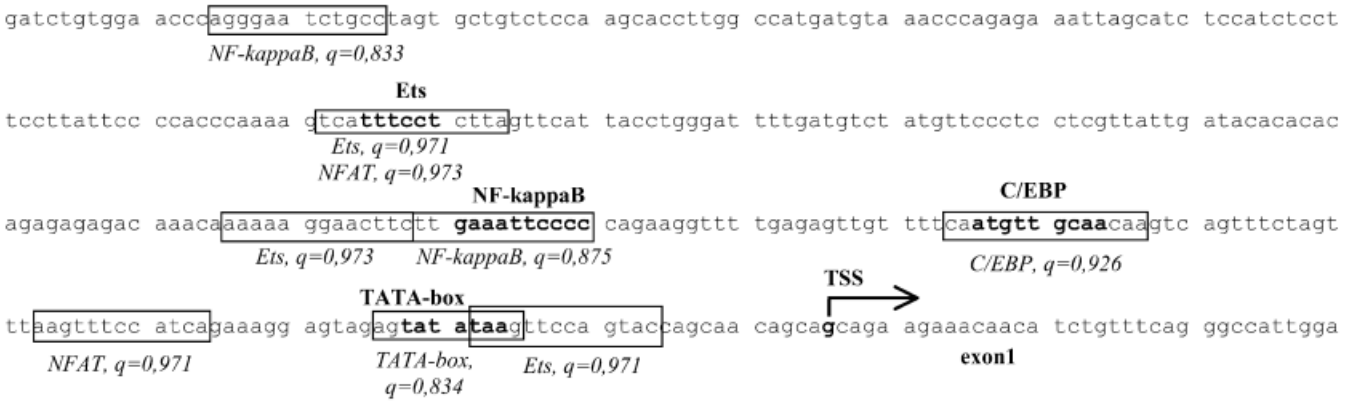
374–378 Nucleic Acids Research, 2003, Vol. 31, No. 1
DOI: 10.1093/nar/gkg108

TRANSFAC®: transcriptional regulation, from patterns to profiles

V. Matys[1,*], E. Fricke[1], R. Geffers[1], E. Gößling[1], M. Haubrock[1], R. Hehl[2], K. Hornischer[1], D. Karas[1], A. E. Kel[1], O. V. Kel-Margoulis[1], D.-U. Kloos[1], S. Land[1], B. Lewicki-Potapov[1], H. Michael[2], R. Münch[1], I. Reuter[1], S. Rotert[1], H. Saxel[1], M. Scheer[1], S. Thiele[1] and E. Wingender[1,3]

[1]BIOBASE GmbH, Halchtersche Strasse 33, D-38304 Wolfenbüttel, Germany, [2]Institut für Genetik-Biozentrum, Technische Universität Braunschweig, Spielmannstrasse. 7, D-38106 Braunschweig, Germany and [3]Gesellschaft für Biotechnologische Forschung mbH, Mascheroder Weg 1, D-38124 Braunschweig, Germany

Received September 16, 2002; Revised October 11, 2002; Accepted October 27, 2002



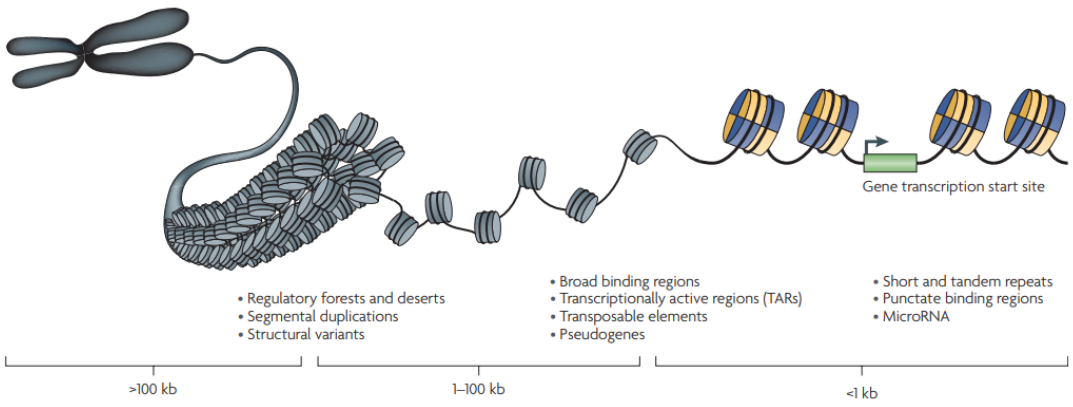
Annotating non-coding regions of the genome

Roger P. Alexander[*‡], *Gang Fang*[*‡], *Joel Rozowsky*[‡], *Michael Snyder*[§] *and Mark B. Gerstein*[*‡||]

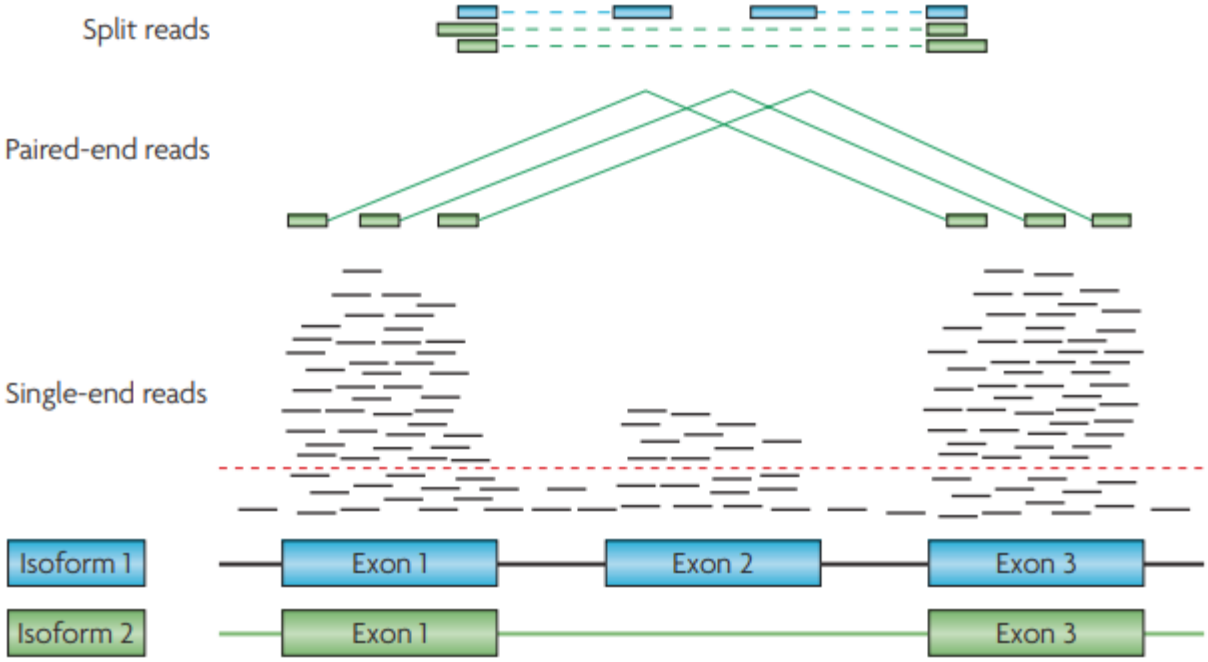
Table 1 | **Length, number and genome coverage of a representative collection of non-coding features**

Classification	Property	Length (nucleotides)		Number of items	Genome coverage (Mb)	Genome coverage (%)
		Average	Longest			
From comparative analysis						
Short and tandem repeats	Simple repeat	63	2,961	415,917	26.1	0.84
	Satellite	1,444	160,602	8,997	13.0	0.42
	Low complexity	46	2,023	370,102	17.0	0.55
DNA transposons		215	3,625	459,524	98.6	3.17
Retrotransposons	LINEs	426	8,505	1,490,241	634.6	20.4
	Alu SINE element	261	614	1,186,885	309.7	9.97
Pseudogenes	Duplicated	6,607	181,882	2413	15.9	0.51
	Processed	723	15,732	8303	6.0	0.19
Segmental duplications		5,740	630 kb	26,469	151.9	4.89
Structural variants		8,761	3.3 Mb	96,874	848.8	27.3
From functional analysis						
Punctate binding sites	STAT1	446	9,079	~2,300	1.0	0.03
	CTCF	1,181	79,200	~35,000	41.4	1.33
	H3K4me3	1,759	71,025	~62,000	110.2	
Broad binding sites	H3K36me3	4,518	380,076	~130,000	589	
MicroRNA		89	150	718	0.063	
TARs		72	1,854	644,200	46.7	
Regulatory forests		3,890	35,165	68,900	268	
Regulatory deserts		27,107	203,691	72,500	1,970	

Box 1 | **Catalogue of non-coding elements**



Doi: 10.1038/nrg2814


Split reads
Paired-end reads
Single-end reads
Isoform 1
Exon 1
Exon 2
Exon 3
Isoform 2
Exon 1
Exon 3

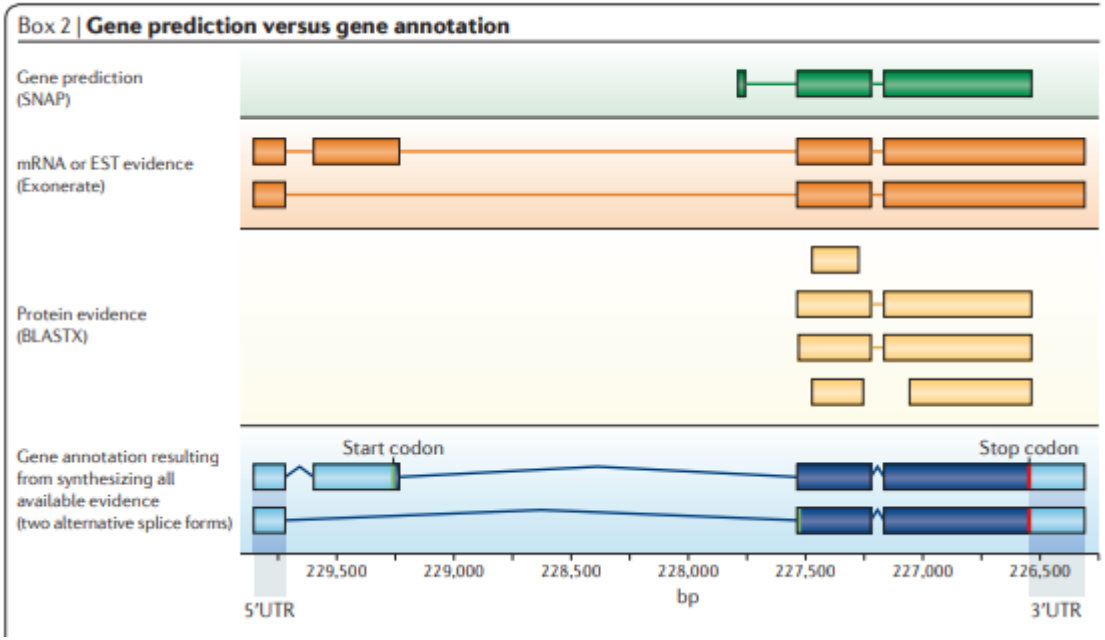

XIAP
STAG2
Tiling array signal
Called TARs
Relaxed threshold
Stringent threshold

STUDY DESIGNS

A beginner's guide to eukaryotic genome annotation

Mark Yandell and Daniel Ence

Abstract | The falling cost of genome sequencing is having a marked impact on the research community with respect to which genomes are sequenced and how and where they are annotated. Genome annotation projects have generally become small-scale affairs that are often carried out by an individual laboratory. Although annotating a eukaryotic genome assembly is now within the reach of non-experts, it remains a challenging task. Here we provide an overview of the genome annotation process and the available tools and describe some best-practice approaches.



Doi: 10.1038/nrg3174

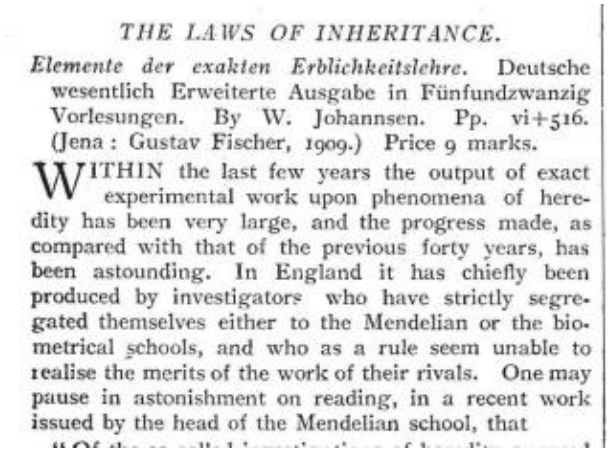
MAS O QUE É UM GENE?

DEFINIÇÃO DE GENE

Pioneiro - Wilhelm Johannsen

1909 → gene

“The word gene is completely free of any hypothesis; it expresses only the evident fact that, in any case, many characteristics of the organism are specified in the germ cells by means of special conditions, foundations, and determiners which are present in unique, separate, and thereby independent way”



THE LAWS OF INHERITANCE.

Elemente der exakten Erblichkeitslehre. Deutsche wesentlich Erweiterte Ausgabe in Fünfundzwanzig Vorlesungen. By W. Johannsen. Pp. vi+516. (Jena : Gustav Fischer, 1909.) Price 9 marks.

WITHIN the last few years the output of exact experimental work upon phenomena of heredity has been very large, and the progress made, as compared with that of the previous forty years, has been astounding. In England it has chiefly been produced by investigators who have strictly segregated themselves either to the Mendelian or the biometrical schools, and who as a rule seem unable to realise the merits of the work of their rivals. One may pause in astonishment on reading, in a recent work issued by the head of the Mendelian school, that

GENE TÍPICO DE PROCARIOTOS

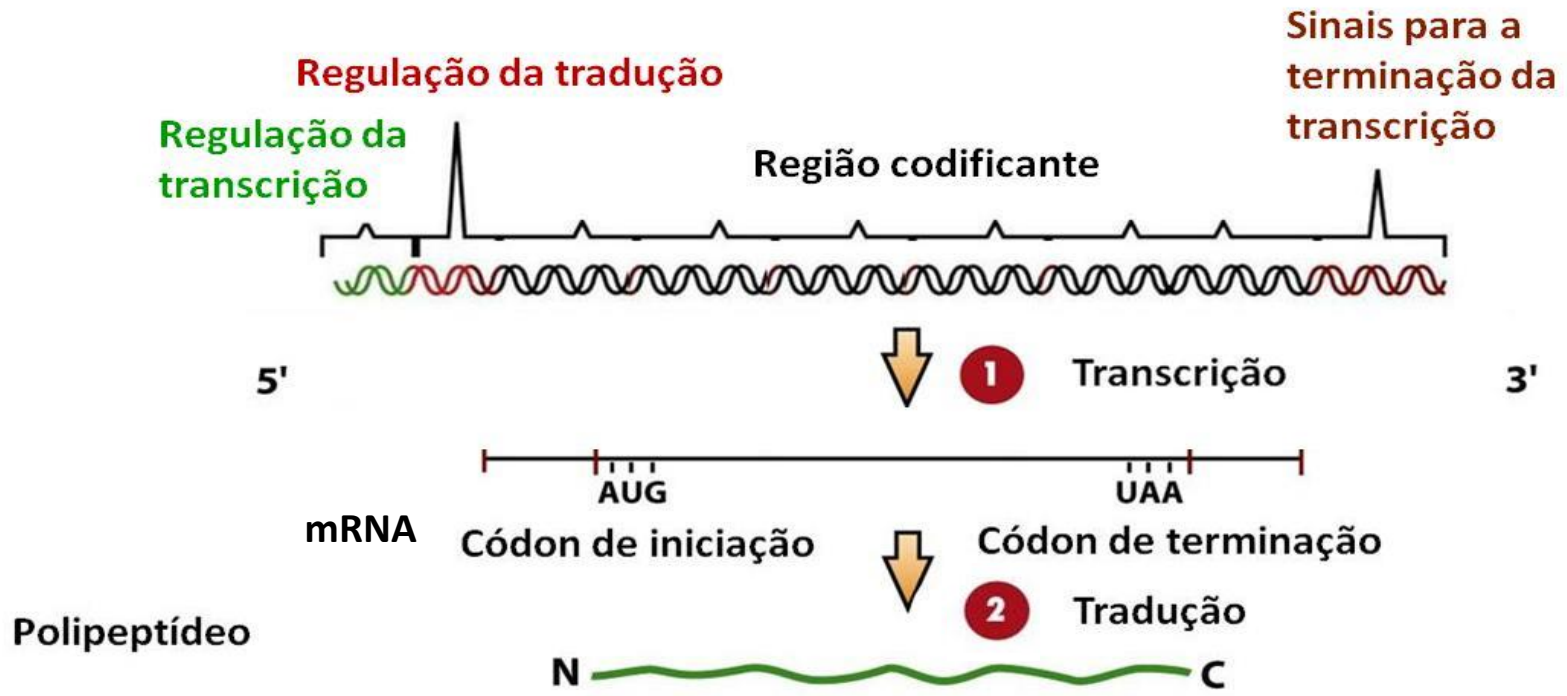


Figure 14-1b Principles of Genetics, 4/e

Unidade da informação genética que controla a síntese de polipeptídios!

GENE TÍPICO DE EUCARIOTOS

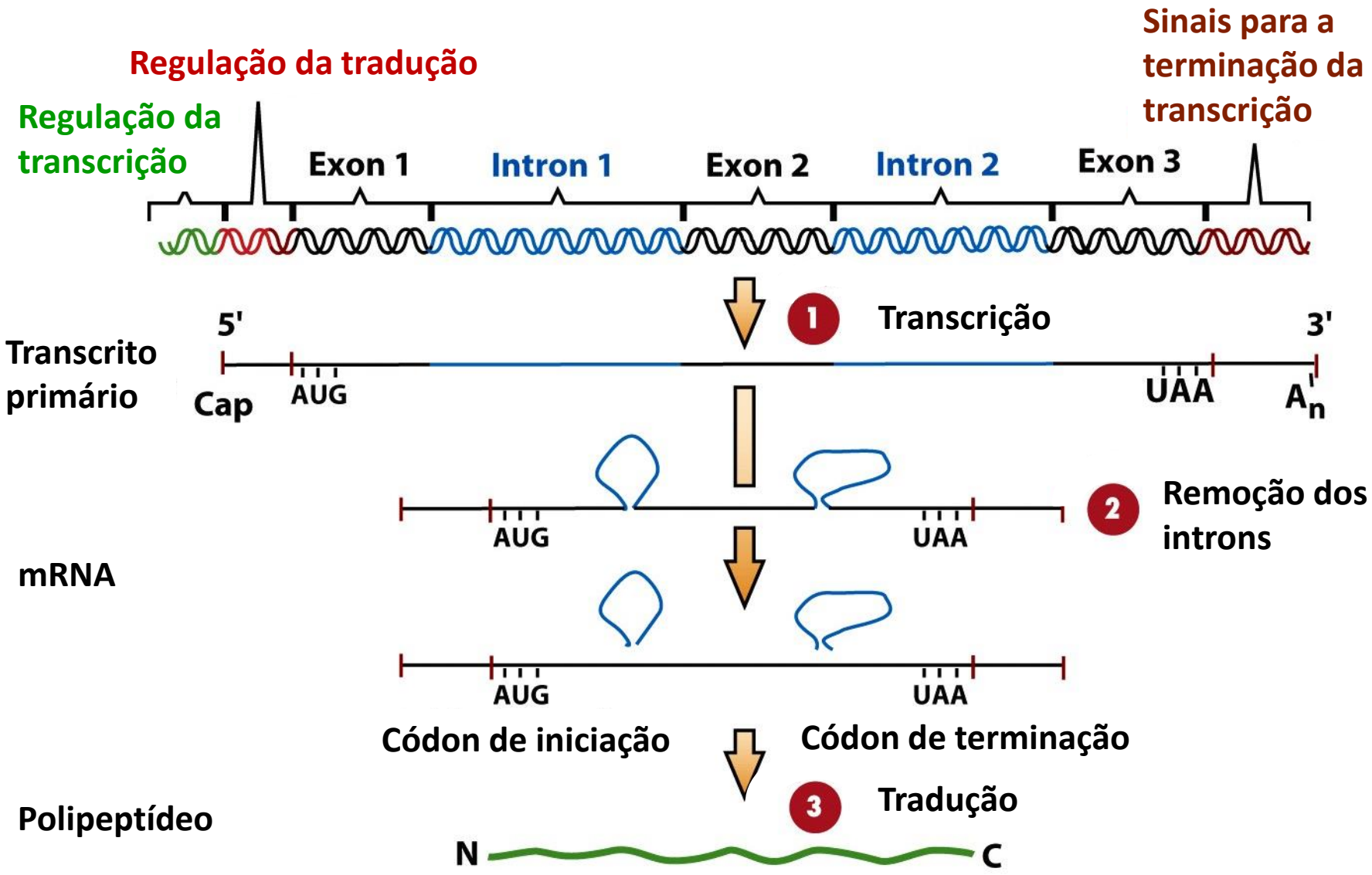
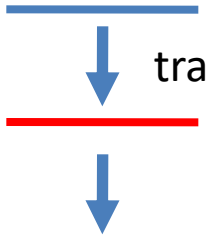


Figure 14-1b Principles of Genetics, 4/e

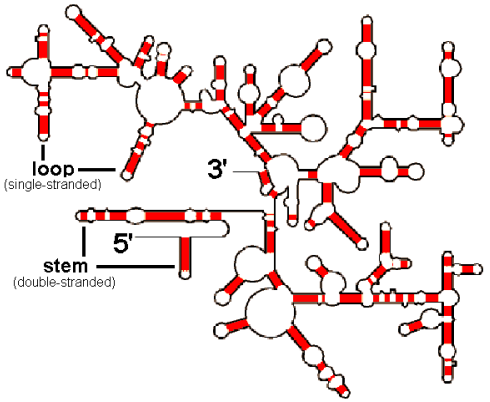
CONCEITO TAMBÉM SERVE PARA RNA ESTRUTURAL

DNA

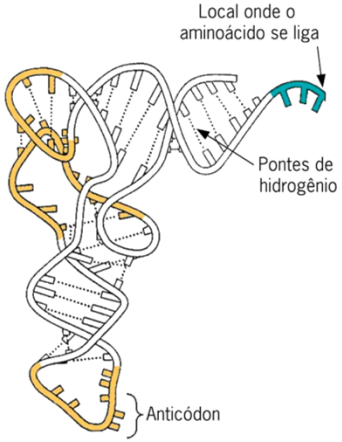


transcrição

RNA ribossomal (rRNA)



RNA transportador (tRNA)



RNAt em três dimensões.

RNA transportador (tRNA)
RNA ribossomal (rRNA)

E os outros RNAs????

Perspective

What is a gene, post-ENCODE? History and updated definition

Mark B. Gerstein,[1,2,3,9] Can Bruce,[2,4] Joel S. Rozowsky,[2] Deyou Zheng,[2] Jiang Du,[3] Jan O. Korbel,[2,5] Olof Emanuelsson,[6] Zhengdong D. Zhang,[2] Sherman Weissman,[7] and Michael Snyder[2,8]

"A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products"

Doi: 10.1101/gr.6339607

HISTÓRICO DA DEFINIÇÃO DE GENE

Gene é uma unidade discrete de hereditariedade;

Gene é um locus distinto;

Gene codifica uma proteína;

Gene é uma molécula fisica;

Gene é uma unidade de transcrito;

Gene é um quadro aberto de leitura (ORF);

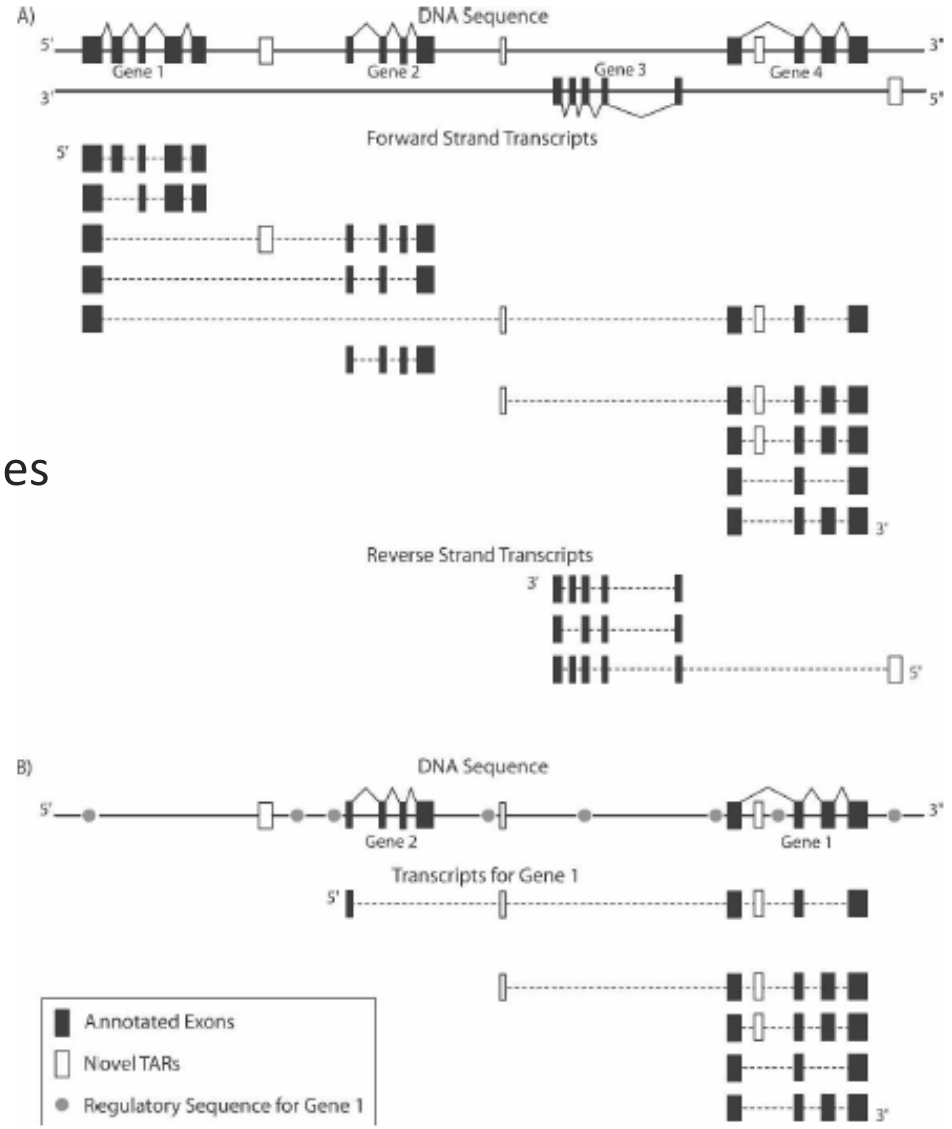
SERÁ?

Table 1. Phenomena complicating the concept of the gene

Phenomenon	Description	Issue
Gene location and structure		
Intronic genes	A gene exists within an intron of another (Henikoff et al. 1986)	Two genes in the same locus
Genes with overlapping reading frames	A DNA region may code for two different protein products in different reading frames (Contreras et al. 1977)	No one-to-one correspondence between DNA and protein sequence
Enhancers, silencers	Distant regulatory elements (Spilianakis et al. 2005)	DNA sequences determining expression can be widely separated from one another in genome. Many-to-many relationship between genes and their enhancers.
Structural variation		
Mobile elements	Genetic element appears in new locations over generations (McClintock 1948)	A genetic element may be not constant in its location
Gene rearrangements/structural variants	DNA rearrangement or splicing in somatic cells results in many alternative gene products (Early et al. 1980)	Gene structure is not hereditary, or structure may differ across individuals or cells/tissues
Copy-number variants	Copy number of genes/regulatory elements may differ between individuals (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005)	Genetic elements may differ in their number
Epigenetics and chromosome structure		
Epigenetic modifications, imprinting	Inherited information may not be DNA-sequence based (e.g., Dobrovic et al. 1988); a gene's expression depends on whether it is of paternal or maternal origin (Sager and Kitchin 1975)	Phenotype is not determined strictly by genotype
Effect of chromatin structure	Chromatin structure, which does influence gene expression, only loosely associated with particular DNA sequences (Paul 1972)	Gene expression depends on packing of DNA. DNA sequence is not enough to predict gene product.
Post-transcriptional events		
Alternative splicing of RNA	One transcript can generate multiple mRNAs, resulting in different protein products (Berget et al. 1977; Gelinas and Roberts 1977)	Multiple products from one genetic locus; information in DNA not linearly related to that on protein
Alternatively spliced products with alternate reading frames	Alternative reading frames of the INK4a tumor suppressor gene encodes two unrelated proteins (Quelle et al. 1995)	Two alternative splicing products of a pre-mRNA produce protein products with no sequence in common
RNA *trans*-splicing, homotypic *trans*-splicing	Distant DNA sequences can code for transcripts ligated in various combinations (Borst 1986). Two identical transcripts of a gene can *trans*-splice to generate an mRNA where the same exon sequence is repeated (Takahara et al. 2000).	A protein can result from the combined information encoded in multiple transcripts
RNA editing	RNA is enzymatically modified (Eisen 1988)	The information on the DNA is not encoded directly into RNA sequence
Post-translational events		
Protein splicing, viral polyproteins	Protein product self-cleaves and can generate multiple functional products (Villa-Komaroff et al. 1975)	Start and end sites of protein not determined by genetic code
Protein *trans*-splicing	Distinct proteins can be spliced together in the absence of a *trans*-spliced transcript (Handa et al. 1996)	Start and end sites of protein not determined by genetic code
Protein modification	Protein is modified to alter structure and function of the final product (Wold 1981)	The information on the DNA is not encoded directly into protein sequence
Pseudogenes and retrogenes		
Retrogenes	A retrogene is formed from reverse transcription of its parent gene's mRNA (Vanin et al. 1980) and by insertion of the DNA product into a genome	RNA-to-DNA flow of information
Transcribed pseudogenes	A pseudogene is transcribed (Zheng et al. 2005, 2007)	Biochemical activity of supposedly dead elements

Doi: 10.1101/gr.6339607

NOVA VISÃO DOS GENES



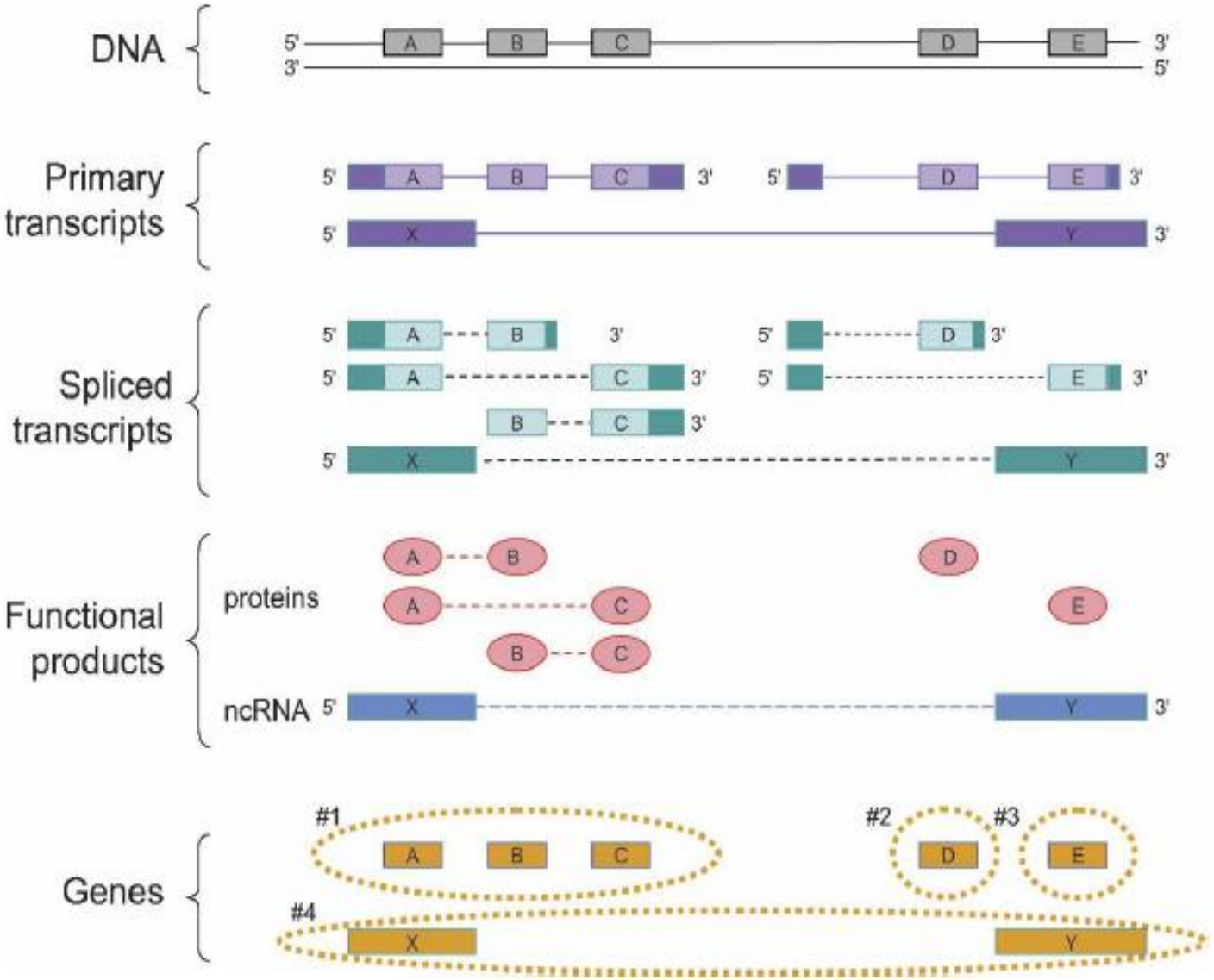
Sequências regulatória dentro dos genes

Muitas possibilidades...

TAR – regiões ativas de transcrição

TSS – sitio de inicio de transcrição

Doi: 10.1101/gr.6339607



Doi: 10.1101/gr.6339607

The Evolving Definition of the Term "Gene"

Petter Portin[*,1] **and Adam Wilkins**[†]
*Laboratory of Genetics, Department of Biology, University of Turku, 20014, Finland and †Institute of Theoretical Biology, Humboldt Universität zu Berlin, 10115, Germany

Table 1 Abridged list of different propositions for a definition of the gene in the current era given by different authors

Essential Content or Character of the Proposition	Classification	Author(s)
These three first operational definitions give criteria, formal, experimental and computational, for identifying genes in the DNA sequences of genomes, annotation of genomes, and for specifying the function of genes	Operational Operational Operational	Snyder and Gerstein (2003) Pesole (2008) Stadler *et al.* (2009)
In these three following definitions, classified as molecular, the structural and the functional gene are conceptually distinguished and separated	Molecular Molecular Molecular	Scherrer and Jost (2007) Keller and Harel (2007) Burian (2004)
In this definition two gene concepts, "gene-P (preformationist)" and "gene-D (developmental)", are distinguished	Complex	Moss (2003)
This definition presents three different concepts of the gene: instrumental, nominal and postgenomic	Complex	Griffiths and Stotz (2006)
This definition aims at to define the gene on the basis of its products and separates it from DNA	A new kind of redefinition	Waters (1994)

Received: 14 April 2021 | Revised: 5 October 2021 | Accepted: 8 December 2021

DOI: 10.1111/bioe.13006

ORIGINAL ARTICLE

bioethics WILEY

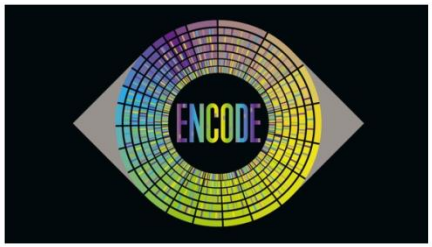
The ethical gene

Reuven Brandt

COLLECTION | 05 SEPTEMBER 2012

Encode

Access the collected papers by exploring the thematic threads that run through them, with topics such as DNA methylation, RNA or machine learning.



https://www.nature.com/collections/aghcdefffg/

ENCODE

3 Characterization of intergenic regions and gene definition

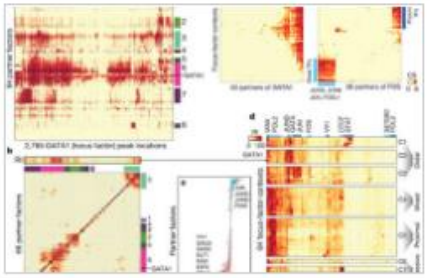
The prevalence and analysis of ENCODE data are changing the definition and characterization of intergenic and genic regions

The cumulative coverage of transcribed regions in the 15 cell lines across the human genome is 62.1% and 74.7% for processed and primary transcripts, respectively (Supplementary Table 10 and Supplementary Fig. 22). On average, for each cell line, 39% of the genome is covered by primary transcripts and 22% by processed RNAs. No cell line showed transcription of more than 56.7% of the union of the expressed transcriptomes across all cell lines. When mapping the current RNA-seq data to the ENCODE pilot regions (Supplementary Table 10), we observed a similar, albeit higher, extent of transcriptional coverage of 73.3% for processed RNAs and 84.5% for primary transcripts. Previously reported estimates in these regions for processed and primary transcripts were 24% and 93%, respectively (Supplementary Table 2.4.3 and ref. 3). The increased genome coverage by processed RNAs stems largely from the inclusion of non-polyadenylated RNAs in the current study. Other than that, given the differences in the samples studied, the selection of pilot regions with high genic content, the increase of annotated genomic regions over time, and the different technologies used to interrogate transcription, both estimates are in reasonable agreement.

Doi: 10.1038/nature28172

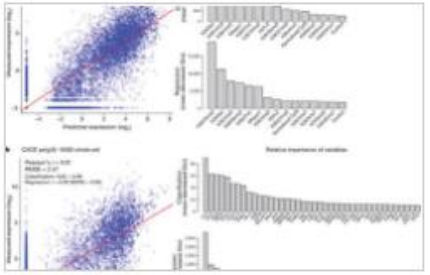
10 Characterization of network topology

ENCODE data analysis helps to describe the various types of regulatory "wiring" implicit in the genome



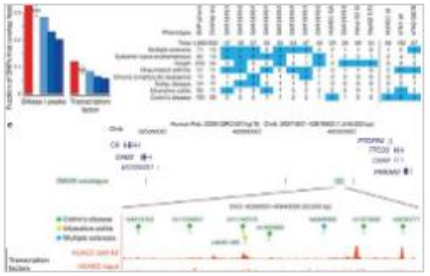
11 Machine learning approaches to genomics

ENCODE has applied machine learning approaches to enable integration and exploration of large and diverse data



12 Impact of functional information on understanding variation

ENCODE provides an initial interpretation of many human variants and plausible leads for the role of many variants identified in genome-wide association studies



CONSTRUÇÃO PRESENTE NA SOJA RR®

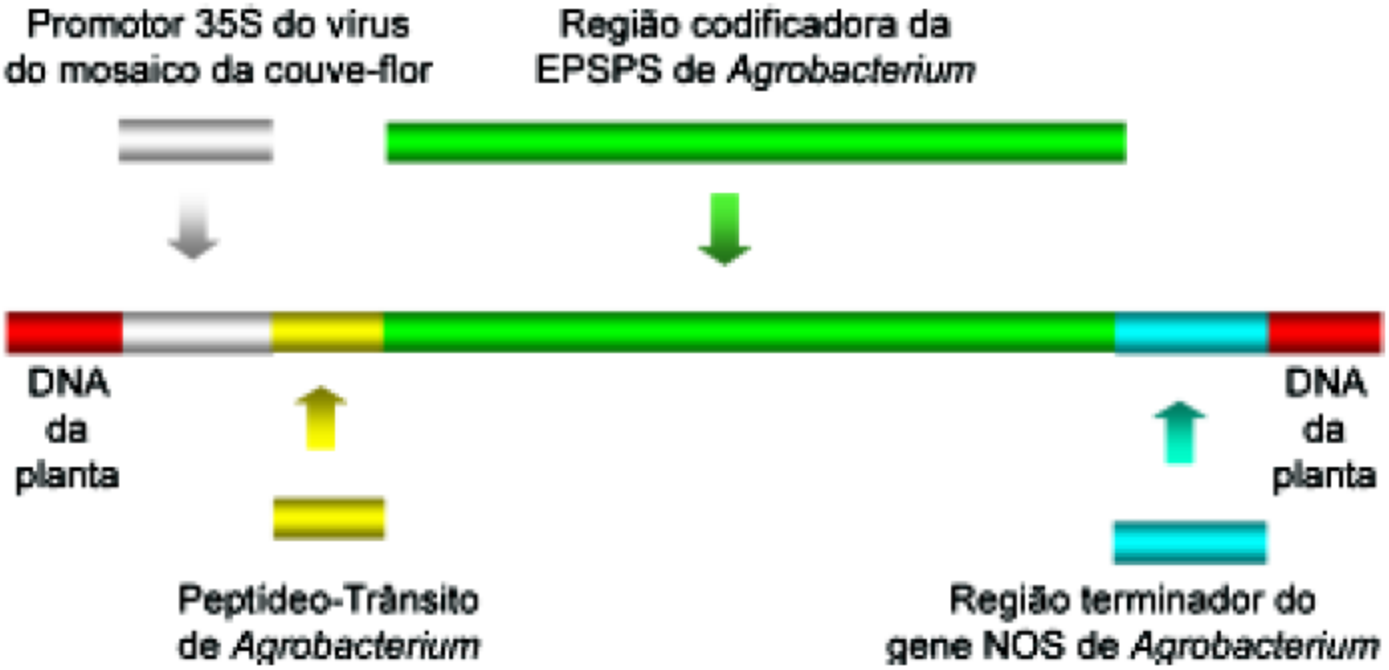
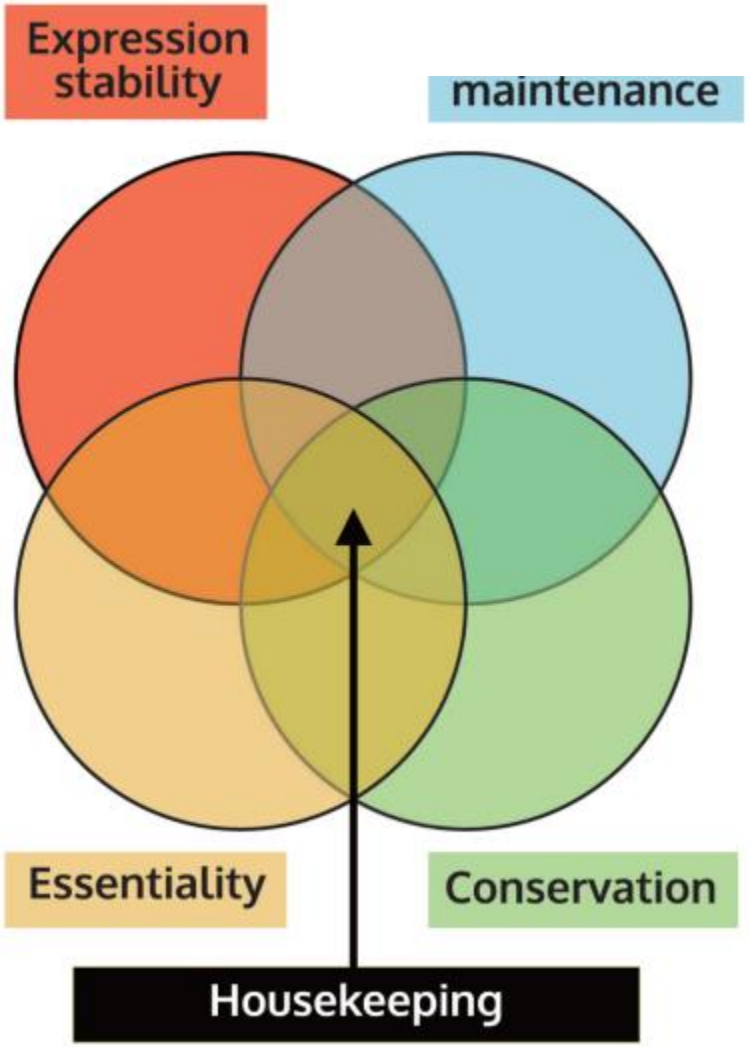


Figura 1 - Representação da construção presente na soja RR® (*Roundup Ready*). Região promotora 35S do vírus do mosaico da couve flor, peptídeo de trânsito de Petúnia, gene que codifica a proteína EPSPS, que confere a resistência ao herbicida, e o terminador do gene da nopalina sintase (NOS).

PLOS COMPUTATIONAL BIOLOGY



RESEARCH ARTICLE

What are housekeeping genes?

Chintan J. Joshi[1], Wenfan Ke[2], Anna Drangowska-Way[2], Eyleen J. O'Rourke[2]*, Nathan E. Lewis[1,3,4,5]*

Bioinformatics, 38(24), 2022, 5390–5397
https://doi.org/10.1093/bioinformatics/btac692
Advance Access Publication Date: 22 October 2022
Original Paper

OXFORD

Systems biology

Defining the extent of gene function using ROC curvature

Stephan Fischer [1,2] **and Jesse Gillis** [1,3,*]

[1]Cold Spring Harbor Laboratory, Stanley Institute for Cognitive Genomics, Cold Spring Harbor, NY 11724, USA, [2]Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris F-75015, France and [3]Department of Physiology, University of Toronto, Toronto, ON, Canada

Systems biology

Transfer learning across ontologies for phenome–genome association prediction

Raphael Petegrosso[1], Sunho Park[2], Tae Hyun Hwang[2] and Rui Kuang[1,*]

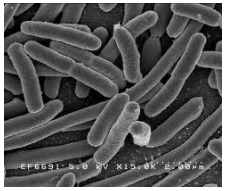
[1]Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA and [2]Department of Clinical Sciences, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

O MELHOR ENTENDIMENTO DOS GENES CONTRIBUI COM O MELHOR ENTENDIMENTO DA ARQUITETURA GENÔMICA!!!

HIV tipo I -19.750 b

Milho
2.5 Gb

Mamute
4.17 Gb



Escherichia coli
5 Mb

Humano
3 Gb

NÚMERO DE GENES EM EUCARIOTOS

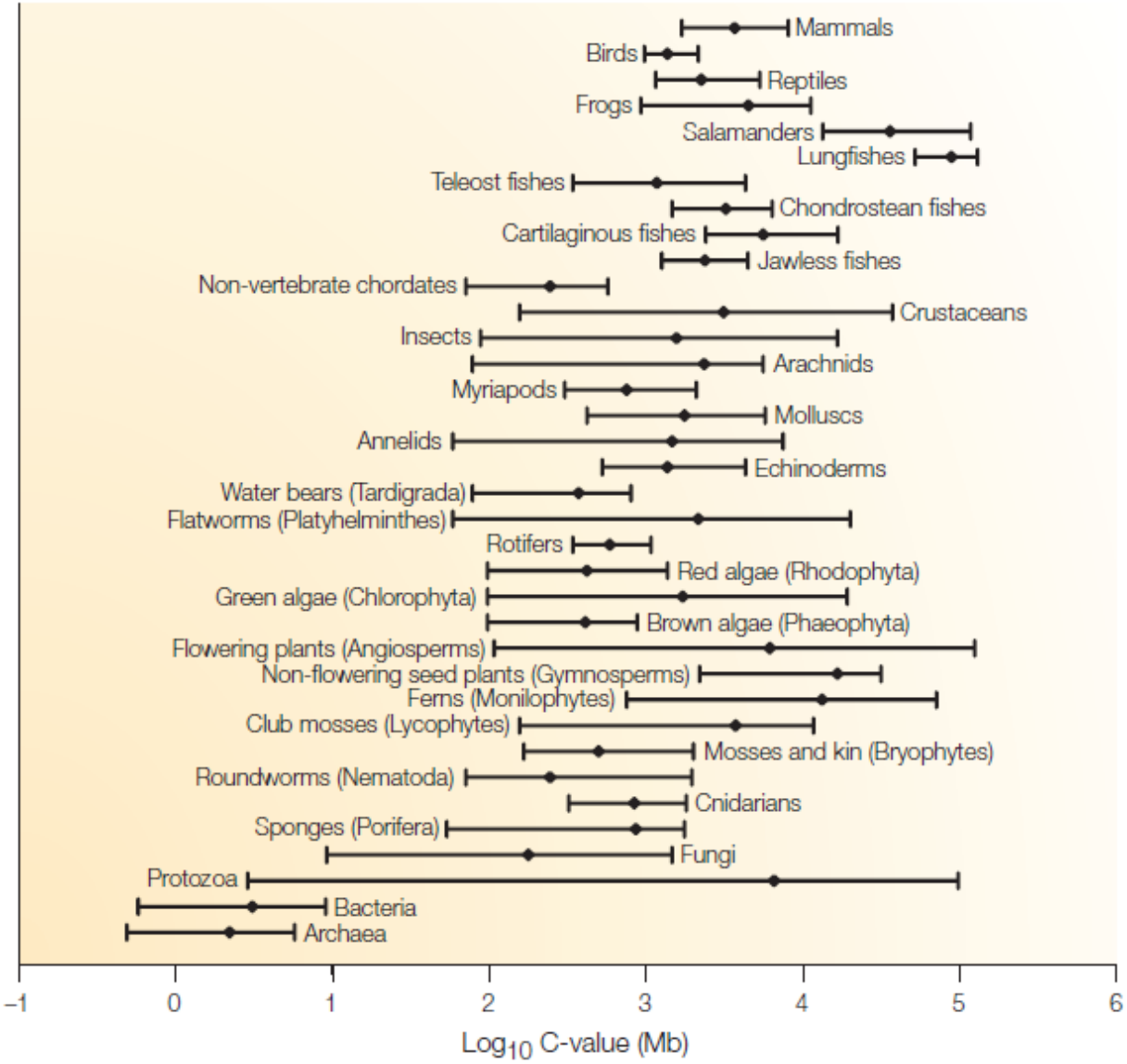
Espécies	Genoma (Mb)	Genes
D. melanogaster	165	~12.000
S. cerevisiae	13	~6.000
C. elegans	97	~20.000
H. sapiens	3.300	~25.000

Table 2 | **Percentage of non-coding DNA in selected sequenced genomes**

Species name		Genome size (Mb)	Fraction of genome (%)				Source of gene annotations
Common	Scientific		Genic	Exonic	Non-coding		
					Intronic	Intergenic	
Yeast	*Saccharomyces cerevisiae*	12.2	73.5	72.9	0.6	26.6	Saccharomyces Genome Database (June 2008 build)
Nematode worm	*Caenorhabditis elegans*	100.3	59.2	28.1	31.2	40.8	WormBase (WS190)
Fruitfly	*Drosophila melanogaster*	168.7	48.2	18.3	30.0	51.8	FlyBase and Berkeley Drosophila Genome Project (BDGP; release no. 5)
Human	*Homo sapiens*	3,107	45.1	2.8	42.3	54.9	UCSC Genome Browser Known Genes table (hg18)

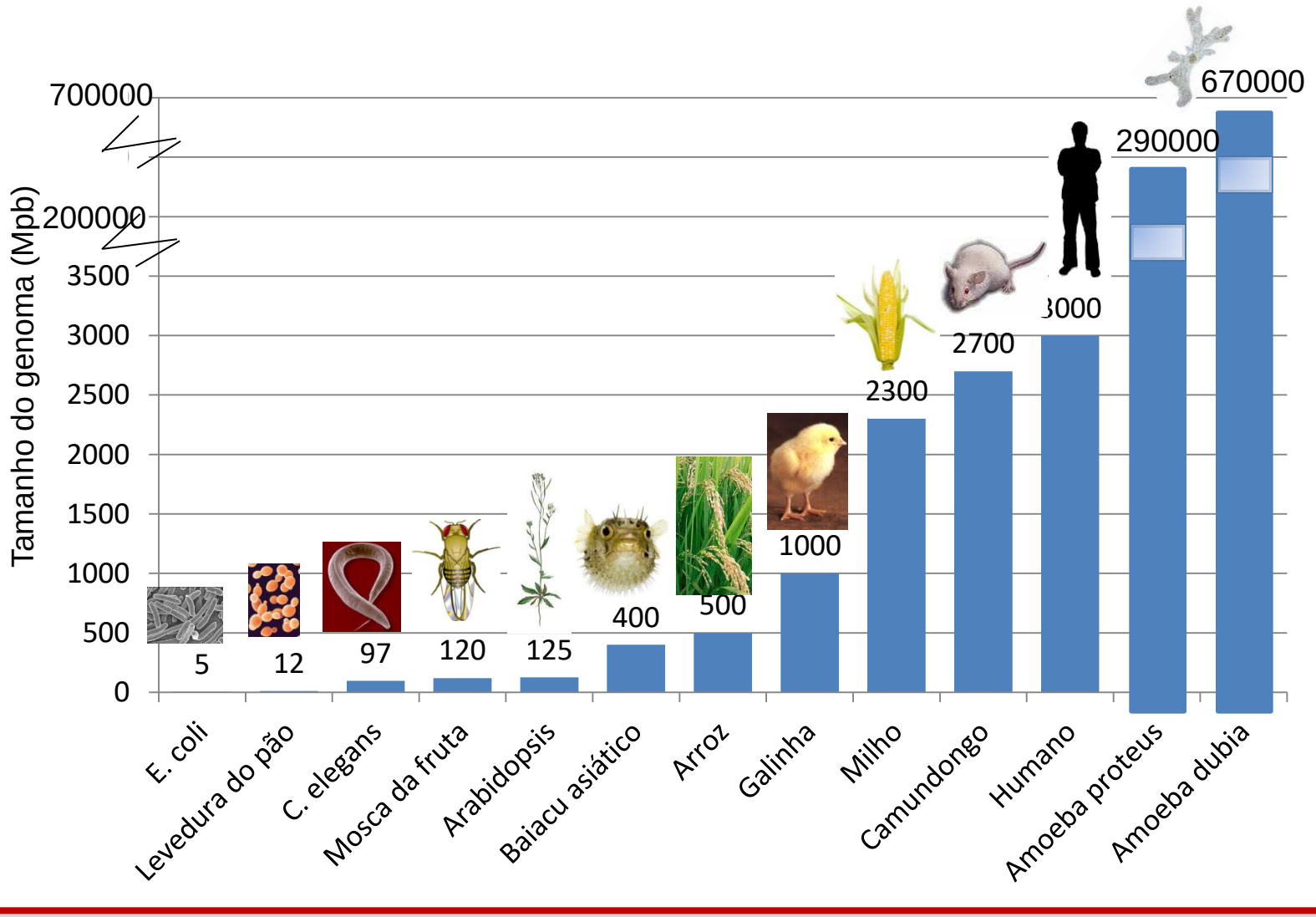
The genic fraction consists of both exonic and intronic sequence. The exonic fraction consists of both coding sequence (CDS) and 5′ and 3′ UTRs. Strictly speaking, UTRs are non-coding, so the exon fraction is a slight overestimate of the fraction of coding sequence in the genome.

Doi: 10.1038/nrg2814

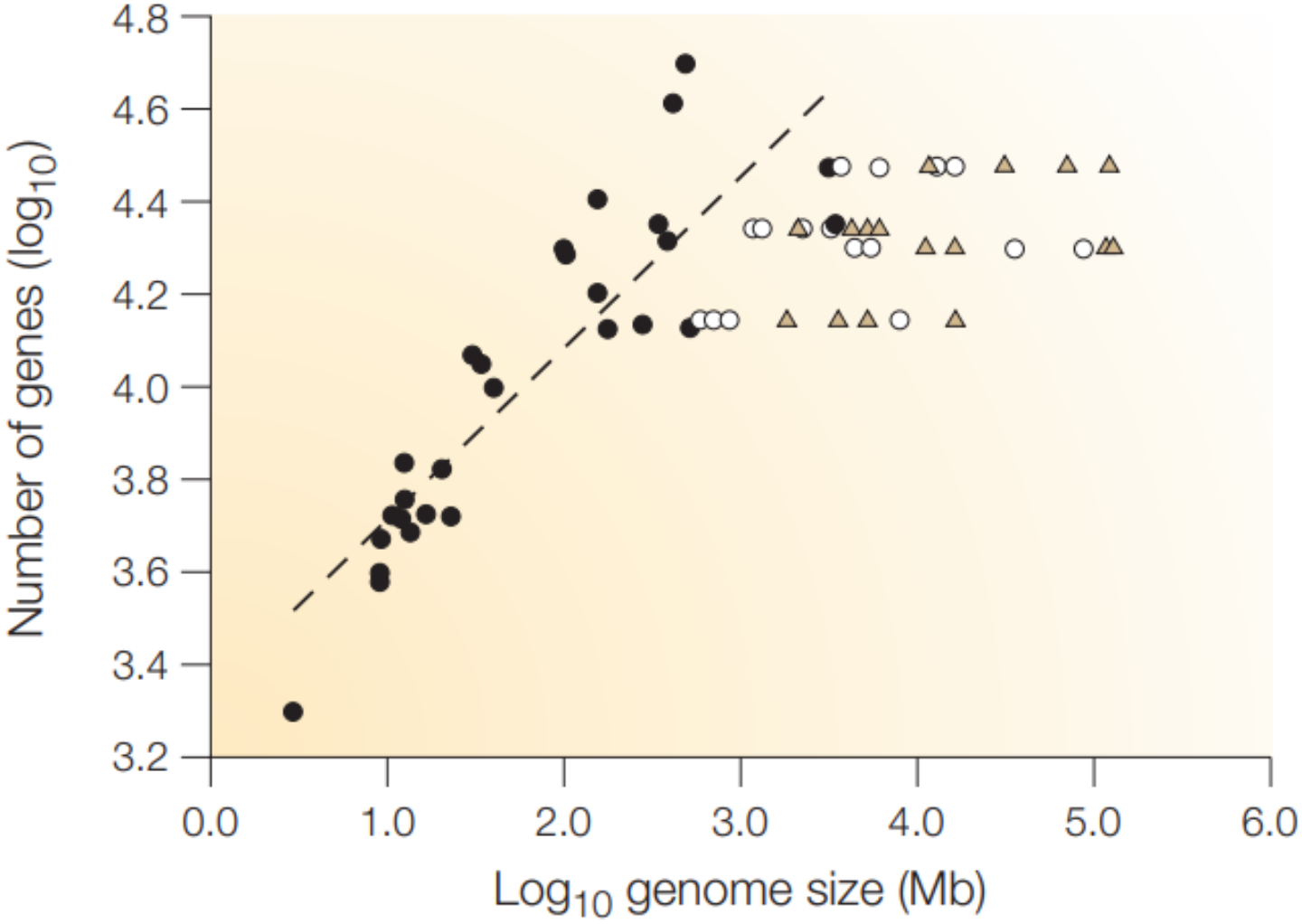


doi: 10.1038/nrg1674

COMPARAÇÃO NO TAMANHO DE GENOMAS

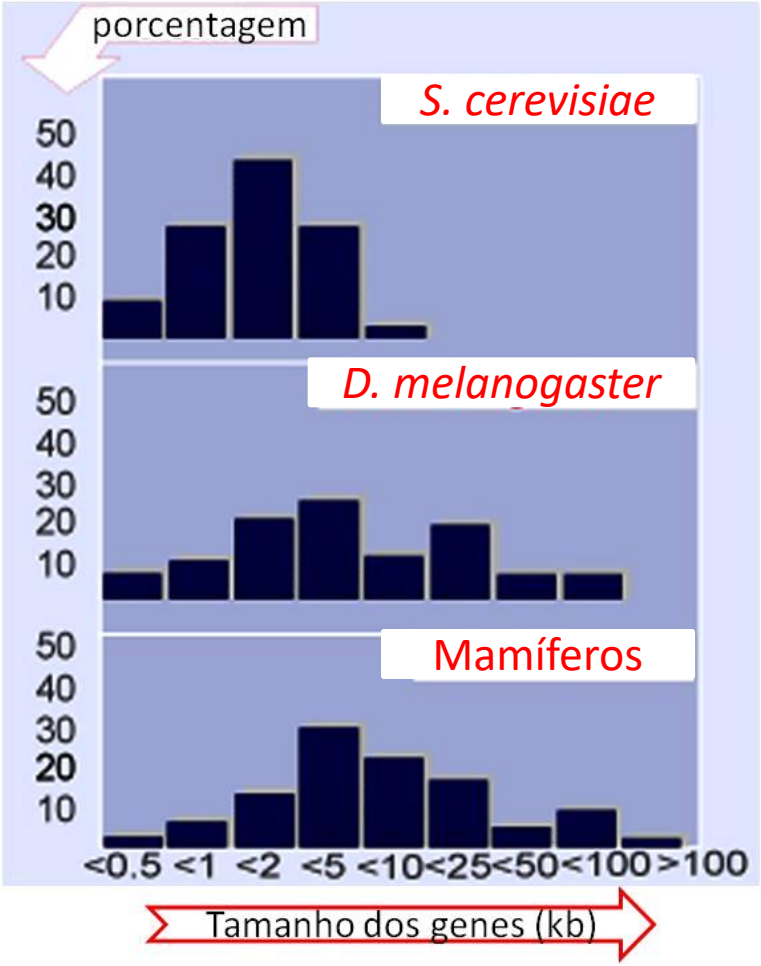


A complexidade de um organismo não é diretamente proporcional ao tamanho do genoma; alguns organismos unicelulares possuem muito mais DNA que os humanos.



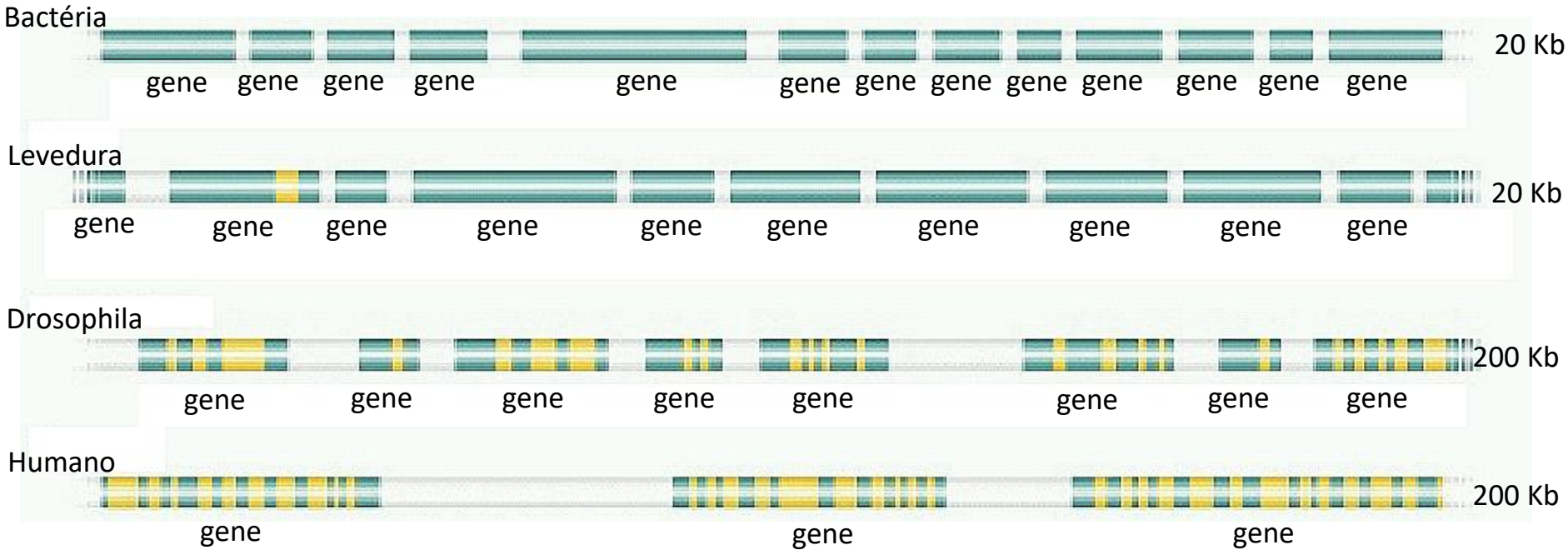
doi: 10.1038/nrg1674

ORGANIZAÇÃO DOS GENES EM EUCARIOTOS



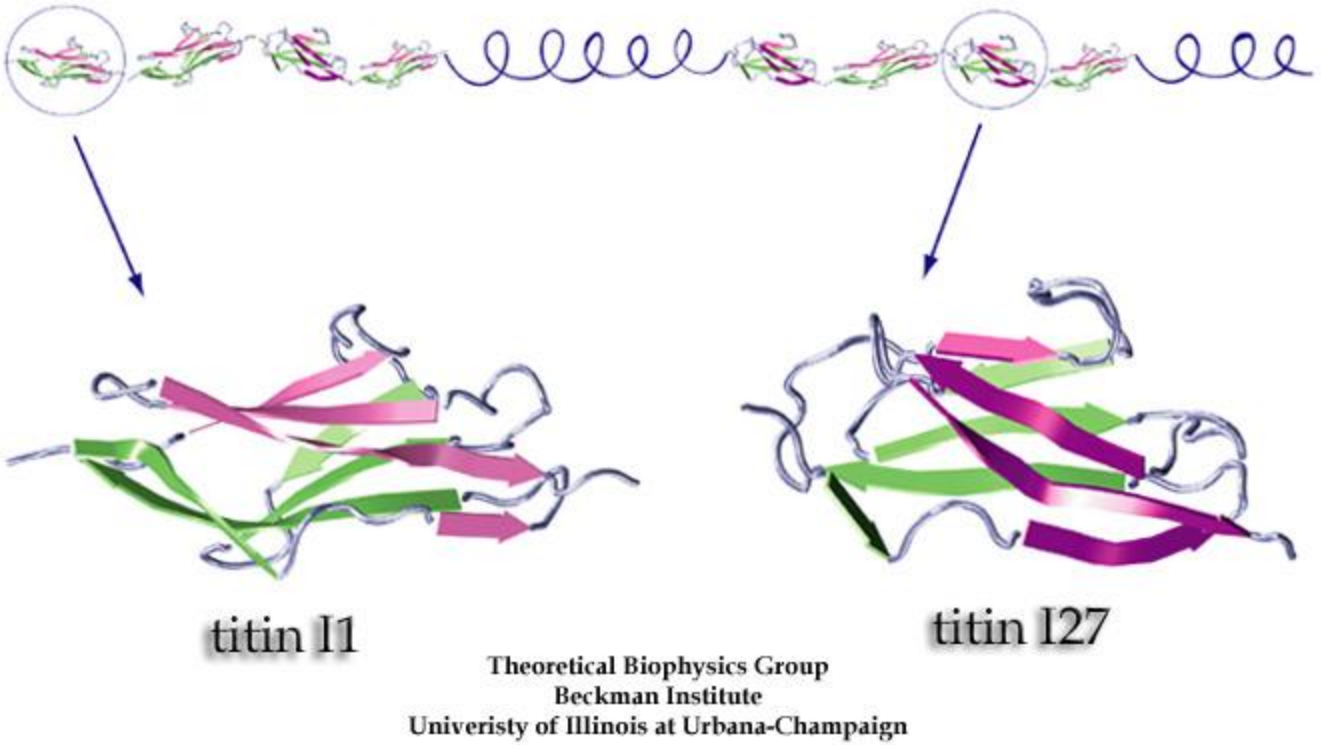
Enorme variação no tamanho dos genes dos mamíferos!
De onde vem isso?

GENES NA MOLÉCULA DE DNA



Grande variação nos tamanho dos genes geradas pela presença dos íntrons!

A proteína gigante do músculo titina contém 38 138 resíduos de aminoácidos (contém 363 exons) que tem um papel importante na contração e elasticidade dos músculos.

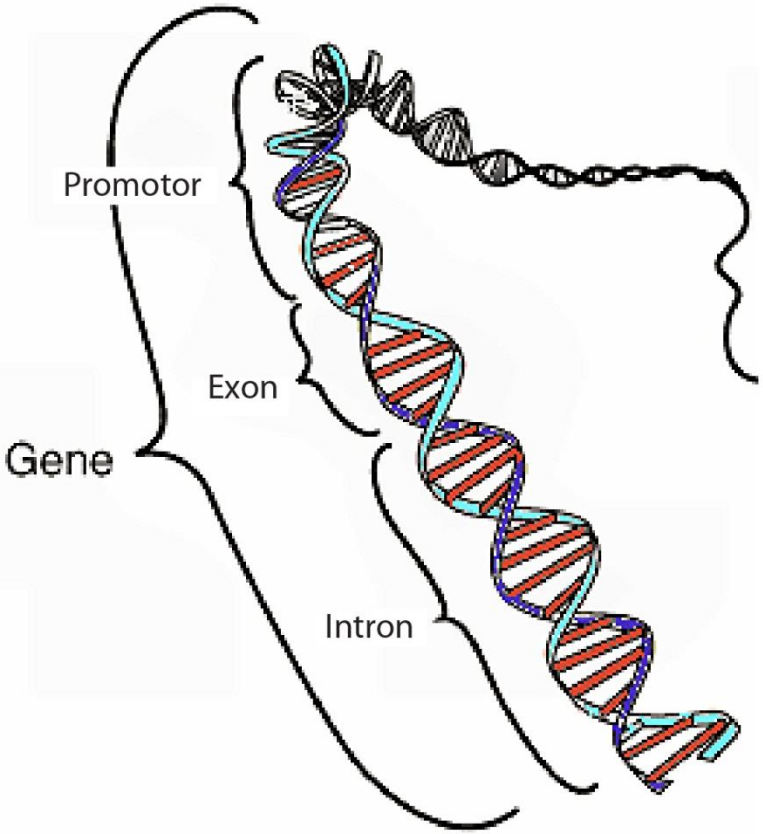


Estrutura de exon-intron do gene titin (293 kb)

O TAMANHO MÉDIO DE UM ÉXON HUMANO É DE 150 pb!!!!

Bang, M.-L. et al. Circ Res 2001;89:1065-1072

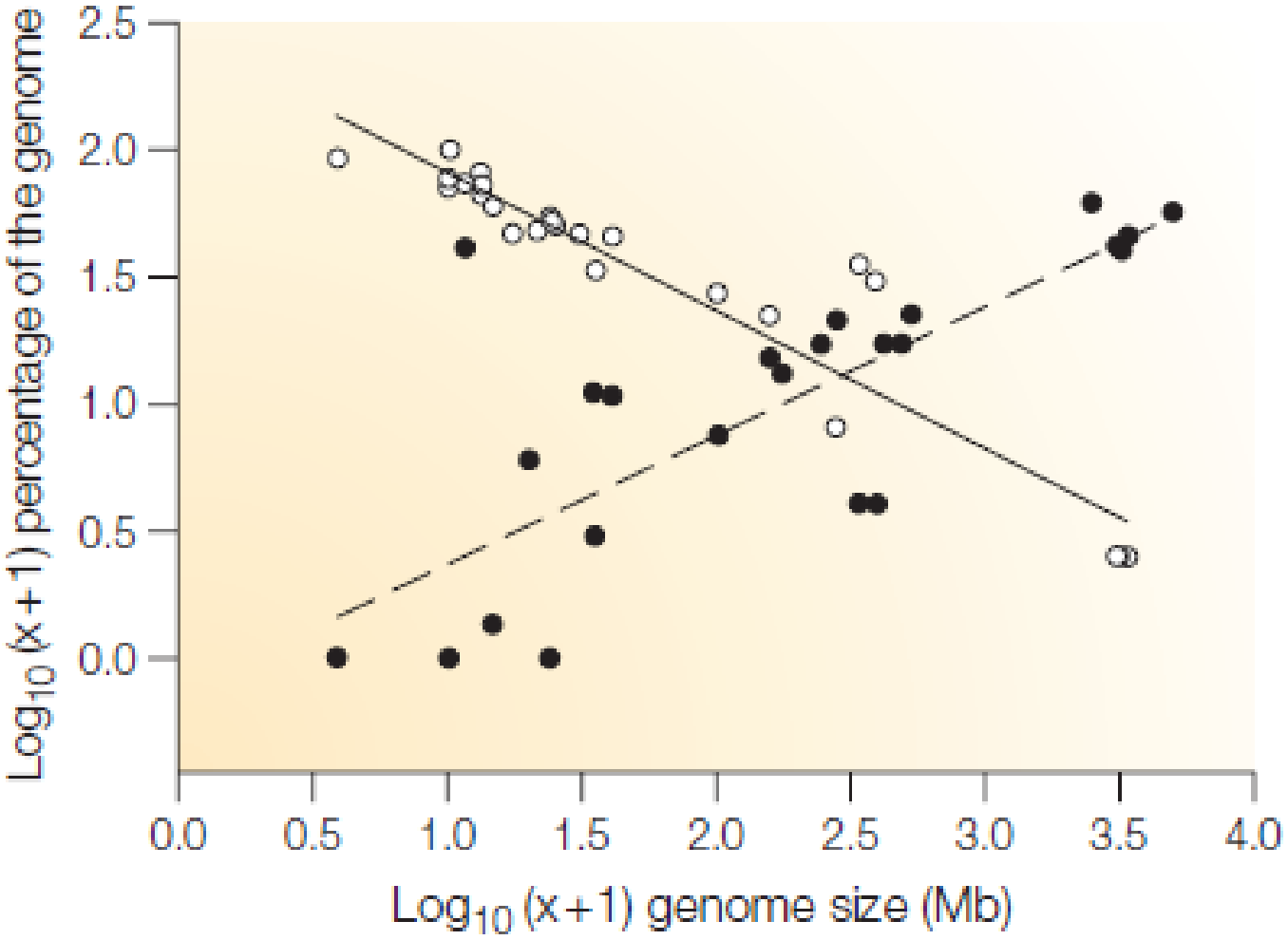
Circulation Research


Promotor
Exon
Gene
Intron

DNA lixo???

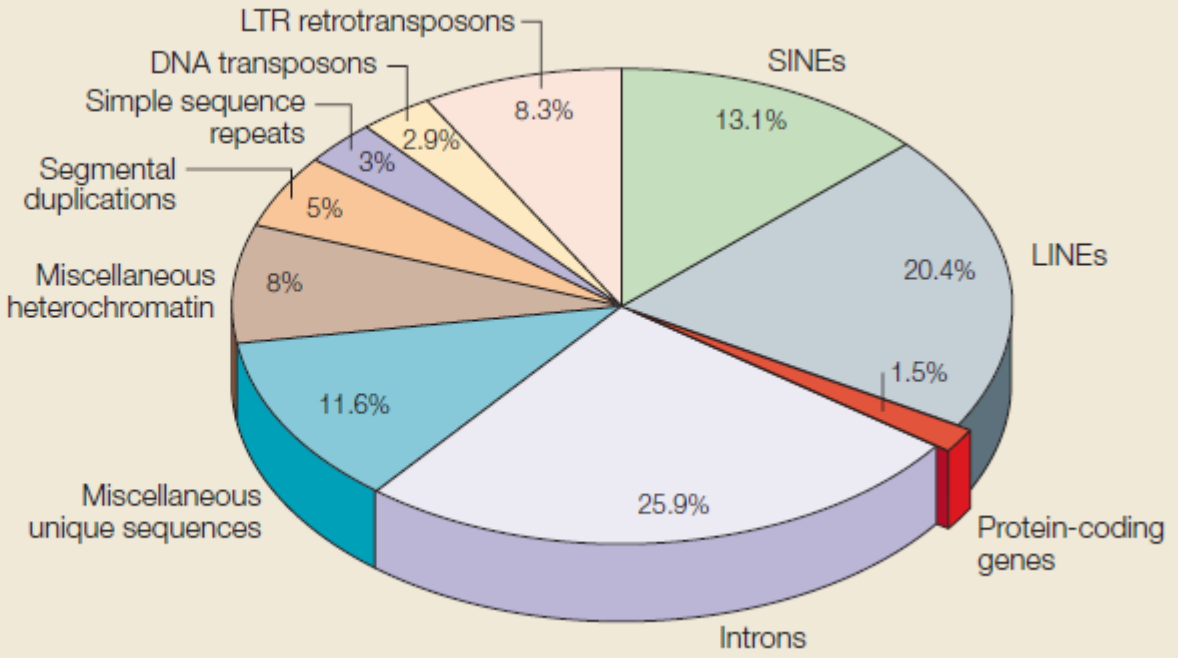

Just because of you don't understand, you can't call us "Junk"!
http://www.biocomicals.com

CORRELAÇÃO ENTRE TAMANHO DE GENOMAS E ELEMENTOS DE TRANSPOSIÇÃO



doi: 10.1038/nrg1674

The figure provides a summary of the different components of the human genome. Less than 1.5% of the genome consists of the suspected 20,000–25,000 protein-coding sequences. By contrast, a large majority is made up of non-coding sequences such as introns (almost 26%) and (mostly defunct) transposable elements (nearly 45%). Data are taken from REF. 16.

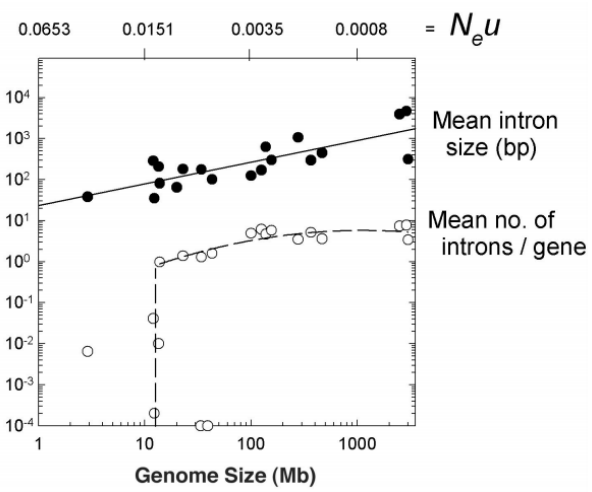


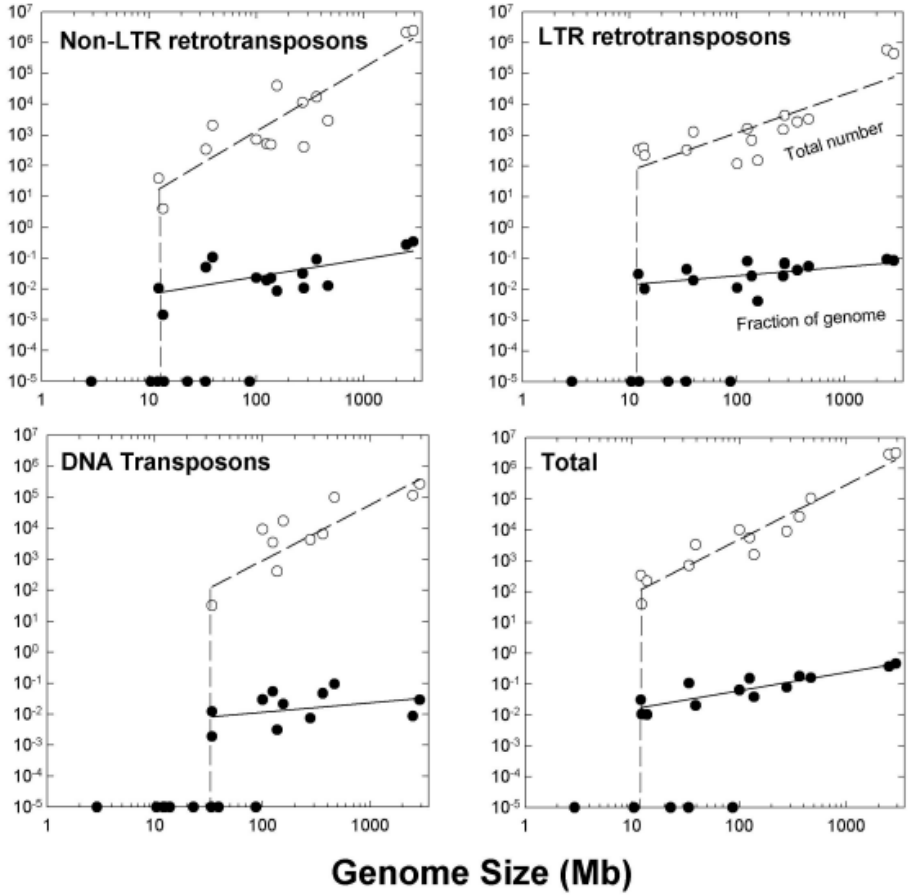
O surgimento de elementos de transposição tem proporcionado o duplicação do tamanho de genoma em humanos

doi: 10.1038/nrg1674

The Origins of Genome Complexity

Michael Lynch[1]* and John S. Conery[2]

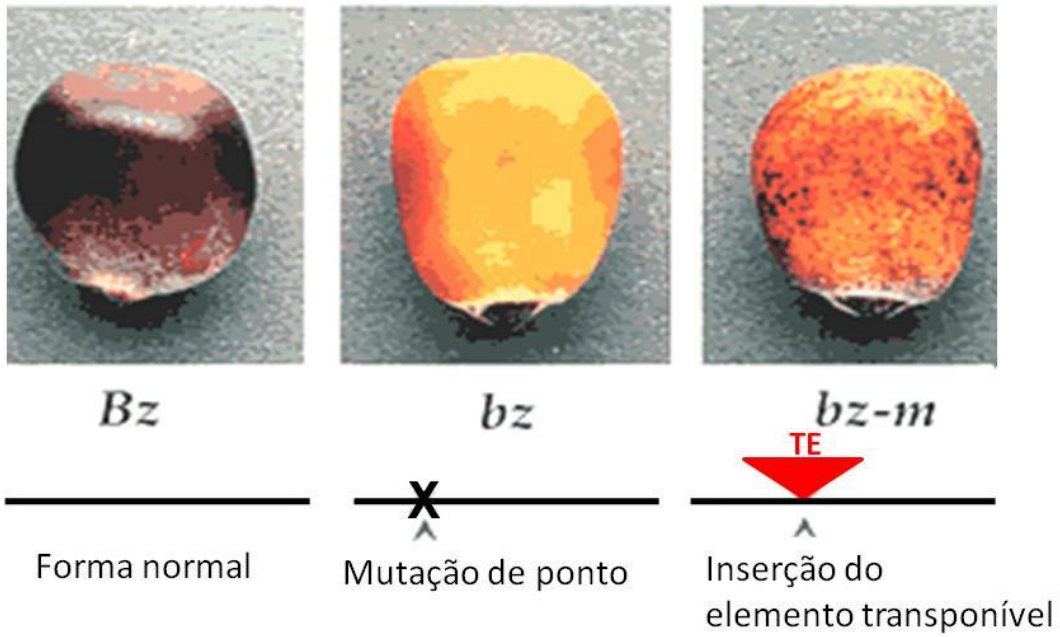




DOI: 10.1126/science.1089370

Barbara McClintock (1902-1992)

Nobel Prize in 1983



Genetica **115:** 49–63, 2002.

Transposable elements and the evolution of genome size in eukaryotes

Margaret G. Kidwell
Department of Ecology and Evolutionary Biology, The University of Arizona, Tucson, AZ 85721, USA (Phone: (520) 621-1784; Fax: (520) 621-9190; E-mail: kidwell@azstarnet.com)

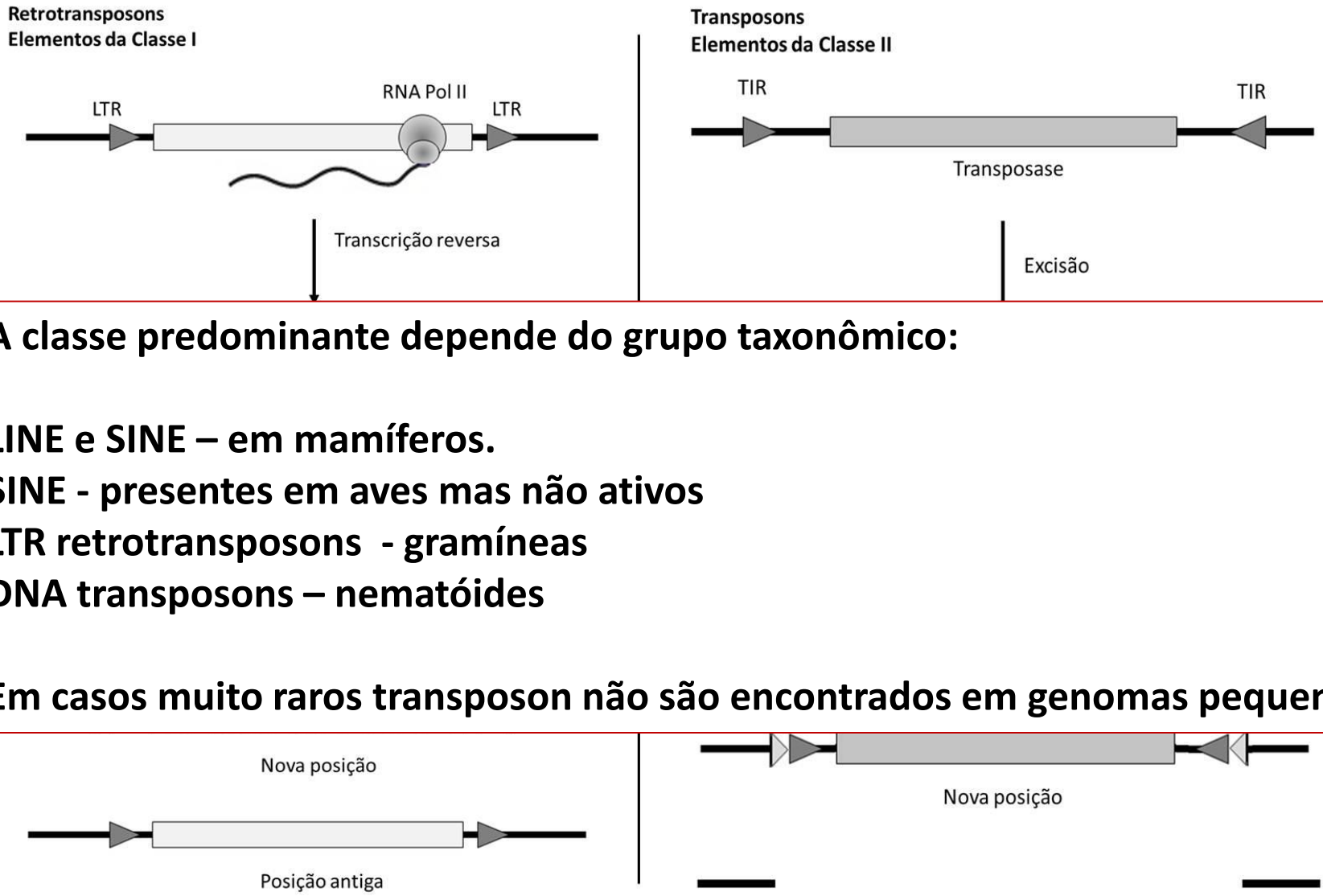
Key words: genome size, molecular evolution, transposable element

Table 1. Characteristics of some widely distributed types of transposable elements

Class	Subclass	Superfamily	Family examples	Approximate size range (bp)
I. Retroelements RNA-mediated elements	LTR retro-transposons	*Ty1-copia*	*Opie-1* in maize	3000–12,000
		BEL	*Cer7* in *C. elegans*	3000–20,000
		DIRS-1	*DIRS-1* in *Dictyostelium*	~5000
		Ty3-gypsy	*Gypsy* in *Drosophila*	5000–14,000
	Non-LTR retroposons	*LINEs*	*LINE-1* in humans; *I* element in *Drosophila*	1000–7000
		SINEs	*Alu* in humans	100–500
II. Transposons DNA-mediated elements	Cut and paste transposition DDE signature present	*mariner-Tc1*	*Tc1* in *C. elegans*; *mariner* in *Drosophila*	1000–2000
		Mu	*Mu* in maize; *MULEs* in *Arabidopsis*	400–20,000
		MITEs	*Tourist* in maize	100–500
	Cut and paste transposition DDE signature absent	*hAT*	*hobo* in *Drosophila*: *Ac in* maize; *Tam-3* in *Anthirinhum*	500–4600
		P	*P* in *Drosophila*	
	Rolling circle (RC) transposition	*Helitrons*	*Helitrons* in *A. thaliana*, *O. sativa*, and *C. elegans*	5500–17,500
Unclassified		*Foldback*	*Galileo* in *D. buzzatii*	Large, but highly variable
		Mini-me	*Mini-me* in many *Drosophila* species	500–1200

Novas classes surgem quase diariamente!

Banco : RepBase - http://www.girinst.org/repbase/



A classe predominante depende do grupo taxonômico:

LINE e SINE – em mamíferos.
SINE - presentes em aves mas não ativos
LTR retrotransposons - gramíneas
DNA transposons – nematóides

Em casos muito raros transposon não são encontrados em genomas pequenos

DESAFIO NO GENOMA DE PLANTAS…

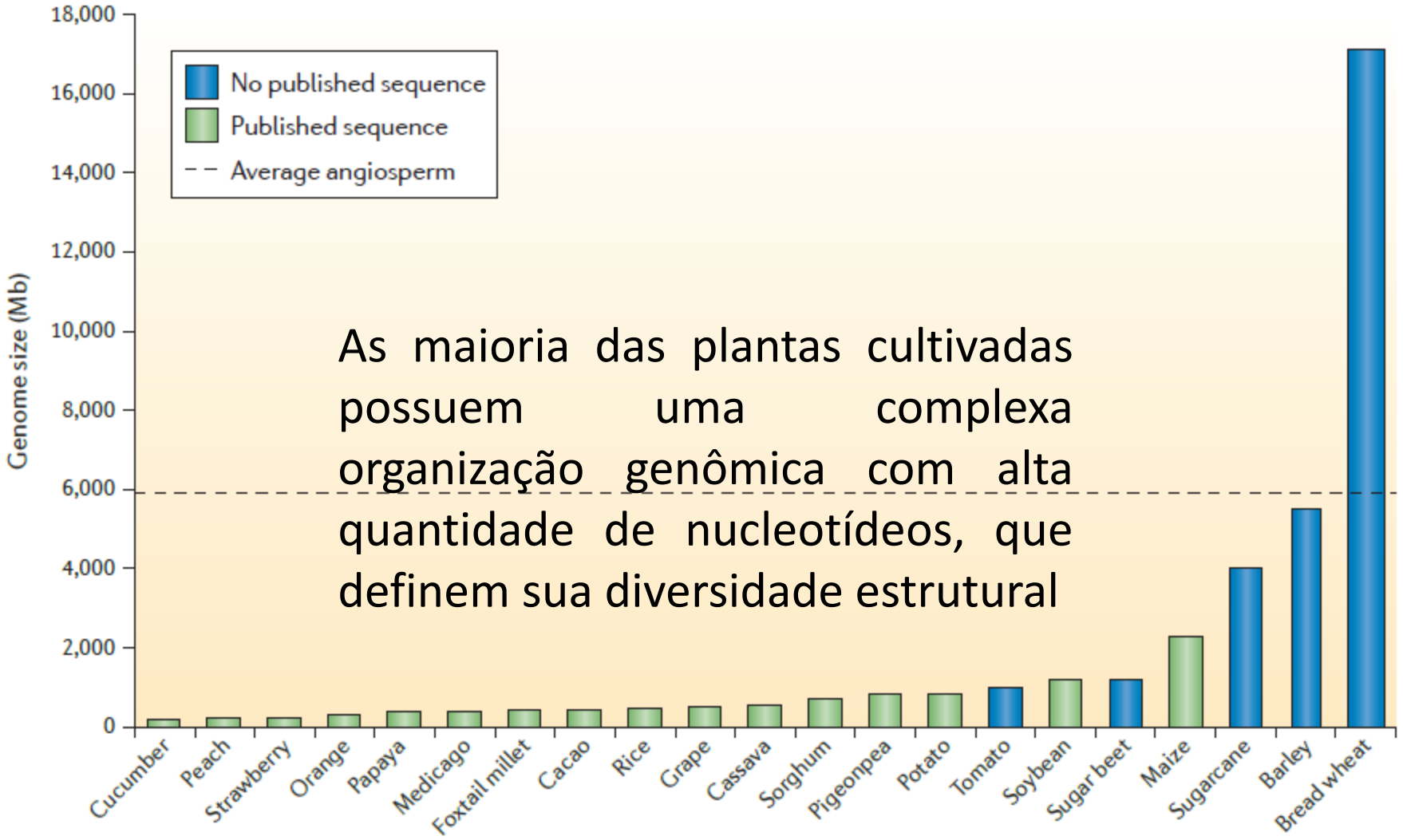
"Uma das primeiras e mais proeminentes observações nas propriedades moleculares do material genético de plantas superiores foi a tremenda variação no tamanho do genoma":

Bennetzen and Kellog, 1997 - The Plant Cell, 9: 1509-1514

Muitos modelos estatisticos não funciona, para plantas..
Muitas familias multigênicas
Poliplóides
Muitos parálogos

Fritillaria assyriaca - 110000 Mpb

Arabidopsis thaliana - 110 Mpb



As maioria das plantas cultivadas possuem uma complexa organização genômica com alta quantidade de nucleotídeos, que definem sua diversidade estrutural

doi: 10.1038/nrg3097

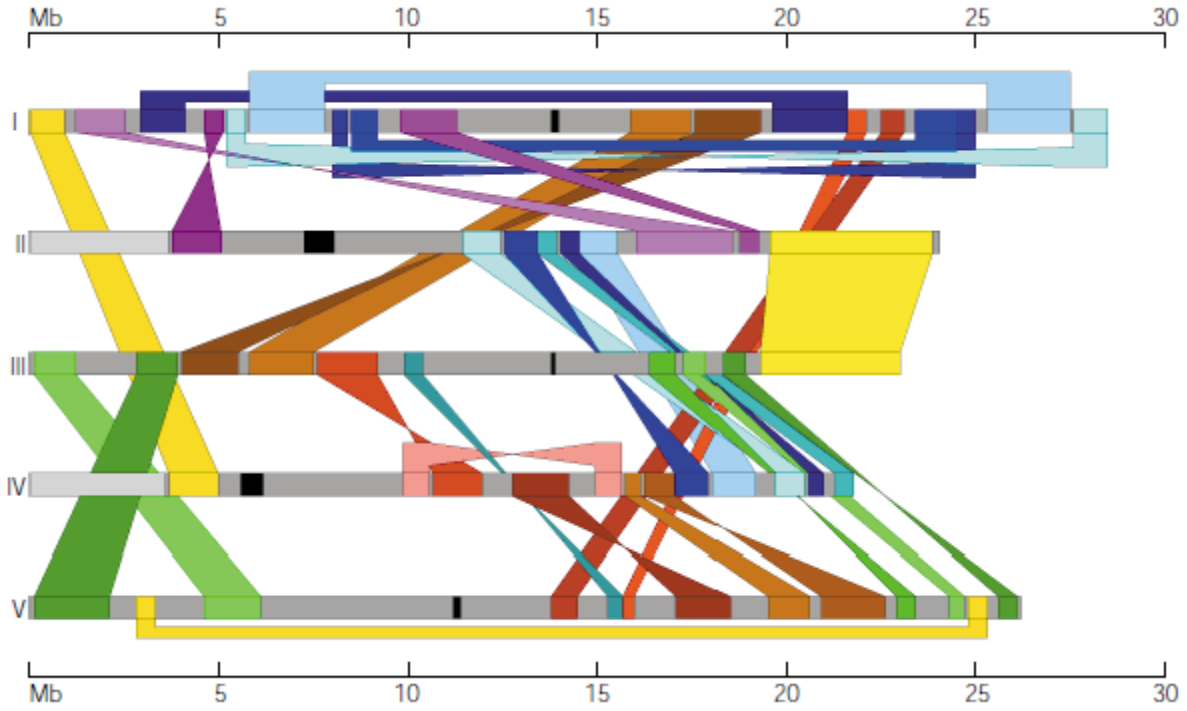
Table 1 | **Crop genome characteristics**

Crop	Genome size (Mb)	Gene number	Transposable element content (%)	Refs
Cucumber	200	21,000		Phytozome
Peach	230	28,000		Phytozome
Strawberry*	240	35,000	22	16
Orange	320	25,000		Phytozome
Papaya	370	29,000	52	149
Medicago*	375	48,000	30	15
Foxtail millet	410	35,000		Phytozome
Cacao	430	29,000	24	150
Rice	450	41,000	25	151
Grape	490	30,000	41	24
Cassava	530	31,000		Phytozome
Sorghum	730	28,000	63	109
Pigeonpea	833	49,000	52	152
Potato	840	39,000	62	17
Tomato	1,000			Kew
Soybean	1,200	46,000	59	153
Sugar beet	1,200			Kew
Maize	2,300	33,000	85	154
Sugarcane	4,000			Kew
Barley	5,500			Kew
Bread wheat	17,100			Kew
Average angiosperm	5,900			Kew

The table shows the genome size, gene and transposable element content for the world's ten top production crops and all other crops with sequenced genomes. In the 'Refs' column, 'Phytozome' refers to http://www.phytozome.net and 'Kew' refers to http://data.kew.org/cvalues. *Sequence of a species related to the main crop.

doi: 10.1038/nrg3097

Genoma de *A. thaliana* – muitos parálogos



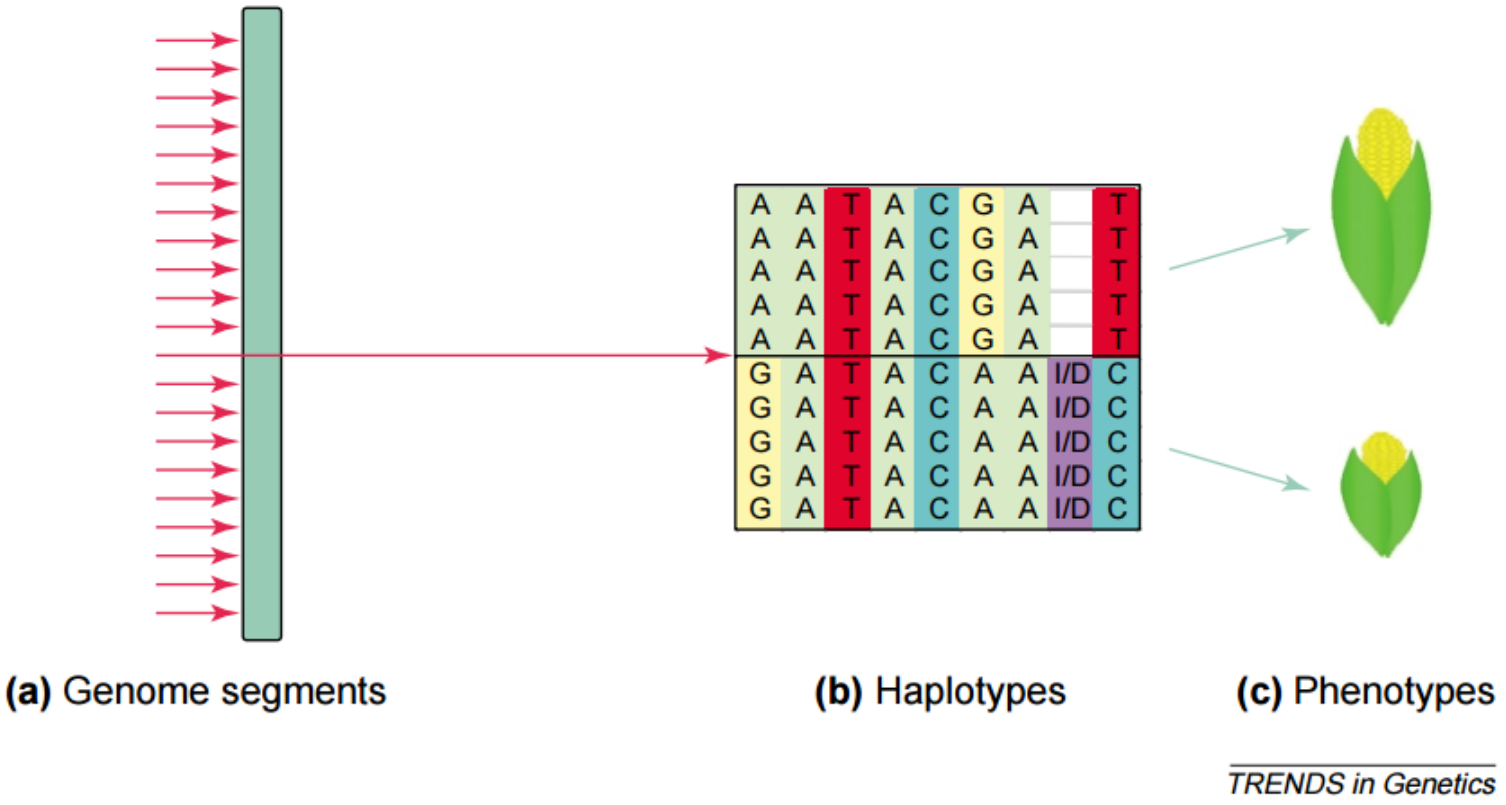
Doi: 10.1038/350805296

Review TRENDS in Genetics Vol.20 No.2 February 2004

Corn and humans: recombination and linkage disequilibrium in two genomes of similar size

Antoni Rafalski[1] and Michele Morgante[2]

[1]DuPont Crop Genetics, Experimental Station E353, PO Box 80353, Wilmington DE, 19880-0353, USA
[2]Dipartimento di Produzione Vegetale e Tecnologie Agrarie, Universita' di Udine, Via delle Scienze 208, Polo Scientifico Rizzi – 33100 Udine, Italy



Influência na eficiência de marcadores moleculares devido ao desequilibrio de ligação

Table 2. Linkage disequilibrium (LD) in different species

Species	LD	Criterion
Human	60 kb	D′ half-length, North Europeans
Human	5 kb	D′ half-length, Yorubans
Cattle	>10 cM	D′ half length
Arabidopsis thaliana	50–100 kb	r^2, half length
Soybean	>50 kb	Little LD decay found
Norway spruce	~100 bp	r^2 half length
	~200 bp	$r^2 = 0.2$
Grape	>500 bp	r^2 half length
Maize	~400 bp	$r^2 = 0.2$
Maize (inbreds from the USA)	~1 kb	$r^2 = 0.2$
Maize	200–1500 bp	$r^2 = 0.2$

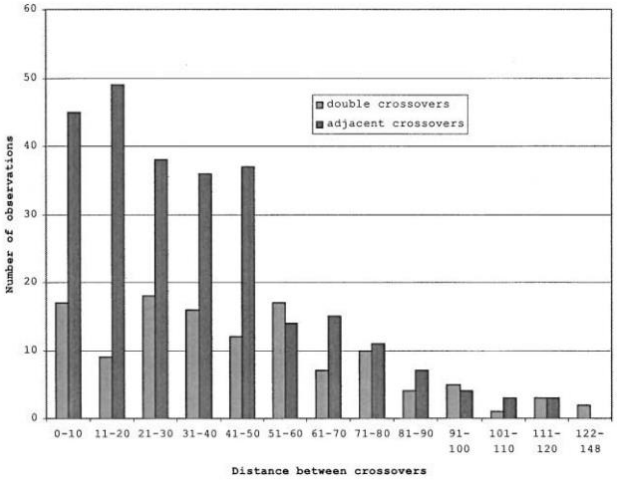
[a]S. Degli Ivanissevich and M. Morgante, unpublished.
[b]F. Cattonaro *et al.*, unpublished.

A história evolutiva dos genomas são especificas!!!

A High-Density Genetic Recombination Map of Sequence-Tagged Sites for *Sorghum*, as a Framework for Comparative Structural and Evolutionary Genomics of Tropical Grains and Grasses

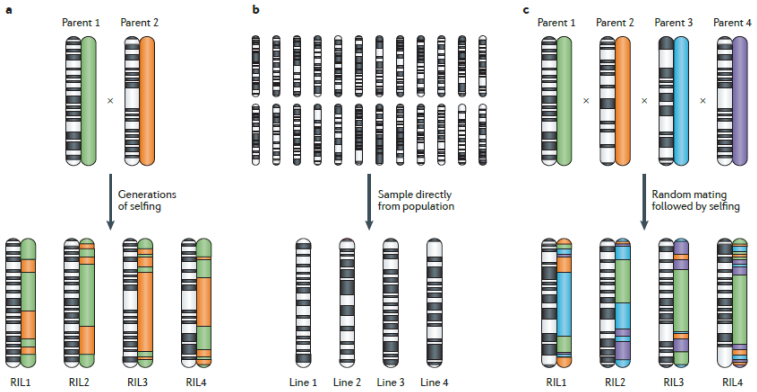
John E. Bowers,* Colette Abbey,† Sharon Anderson,† Charlene Chang,† Xavier Draye,† Alison H. Hoppe,† Russell Jessup,† Cornelia Lemke,* Jennifer Lennington,† Zhikang Li,† Yann-rong Lin,† Sin-chieh Liu,† Lijun Luo,† Barry S. Marler,* Reiguang Ming,† Sharon E. Mitchell,‡ Dou Qiang,† Kim Reischmann,† Stefan R. Schulze,* D. Neil Skinner,* Yue-wen Wang,† Stephen Kresovich,‡ Keith F. Schertz† and Andrew H. Paterson*,†,[1]

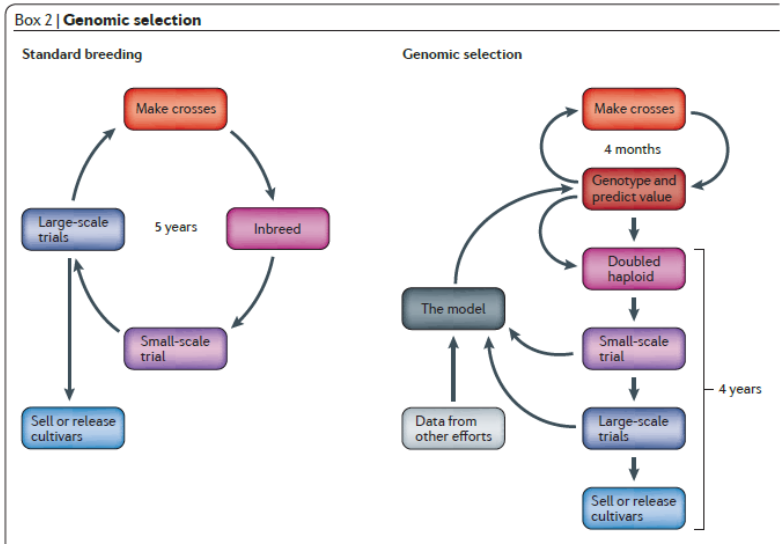
Uso de 2050 sondas de RFLP



Crop genomics: advances and applications

Peter L. Morrell[1], Edward S. Buckler[2] and Jeffrey Ross-Ibarra[3]





doi: 10.1038/nrg3097

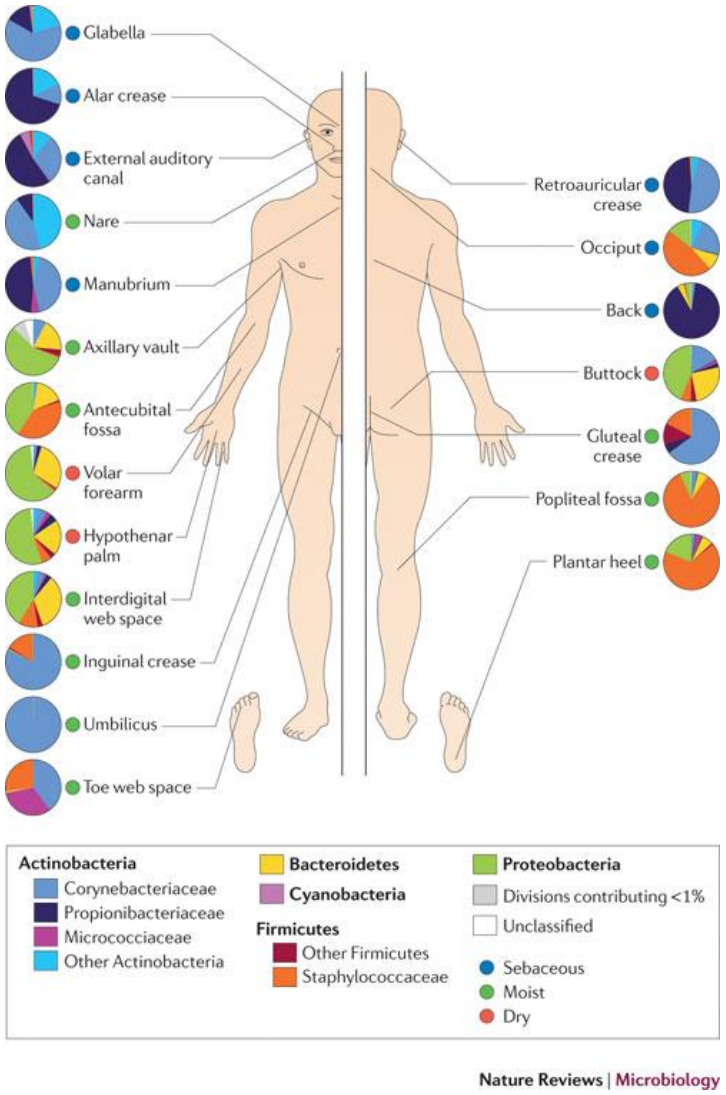
AINDA TEM MUITO O QUE FAZER…

Expectativas de genes em humanos:
antes – 60.000–120.000
draft do genoma – 30.000–35.000
atualmente - 20.000–25.000 na montagem final

QUANTO MAIS SE APRENDE, MAIS PERGUNTAS SURGEM…

- Que tipo de sequências compõem a maioria de regiões não codantes no genoma?
- Como essas sequencias são ganhadas e perdidas nos genomas ao longo da evolução?
- Quais efeitos ou talvez funções esse DNA não codante tem no fenótipo?
- Porque alguns genomas são tão enxutos (aves) e outros tão grandes (salamandras)?

O HOMEM NÃO É UMA ILHA...

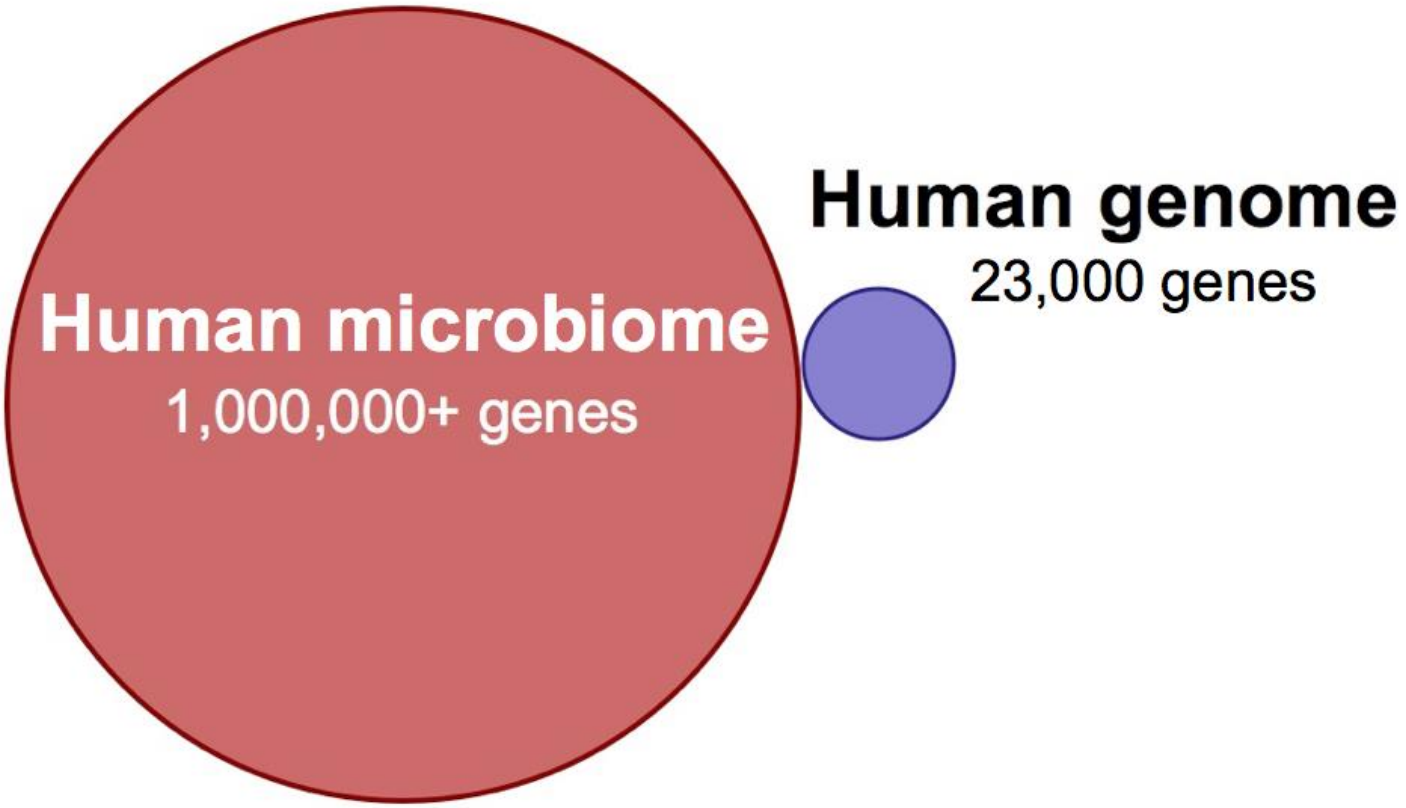


Structure, function and diversity of the healthy human microbiome

The Human Microbiome Project Consortium*

c

Gastrointestinal



FENÓTIPO = GENÓTIPO + AMBIENTE?

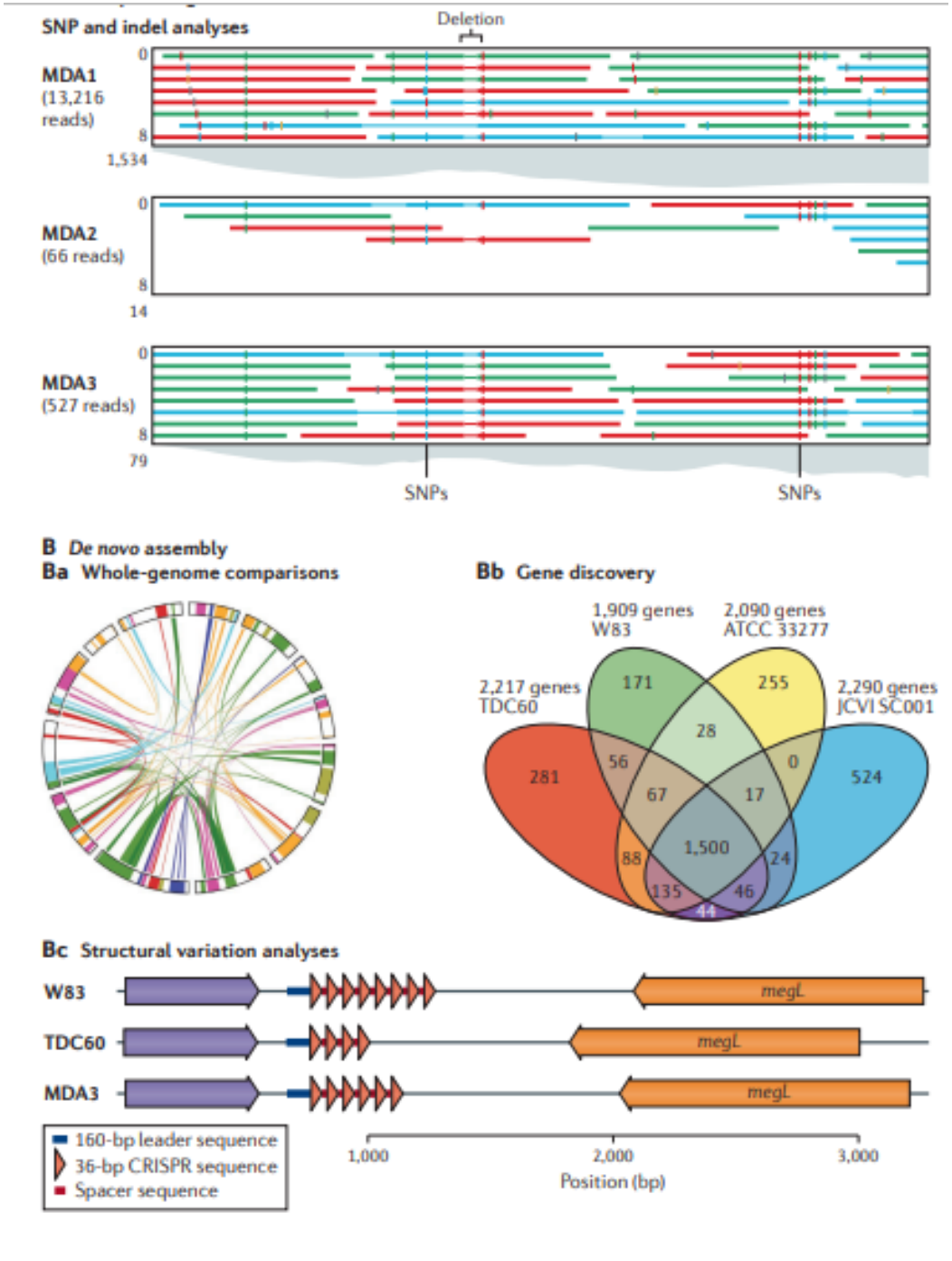
Microbial genome-enabled insights into plant–microorganism interactions

David S. Guttman[1], Alice C. McHardy[2,3] and Paul Schulze-Lefert[3,4]

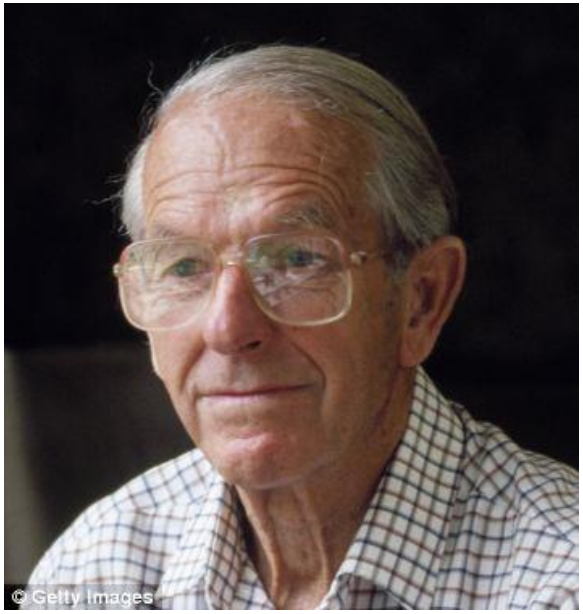
APPLICATIONS OF NEXT-GENERATION SEQUENCING

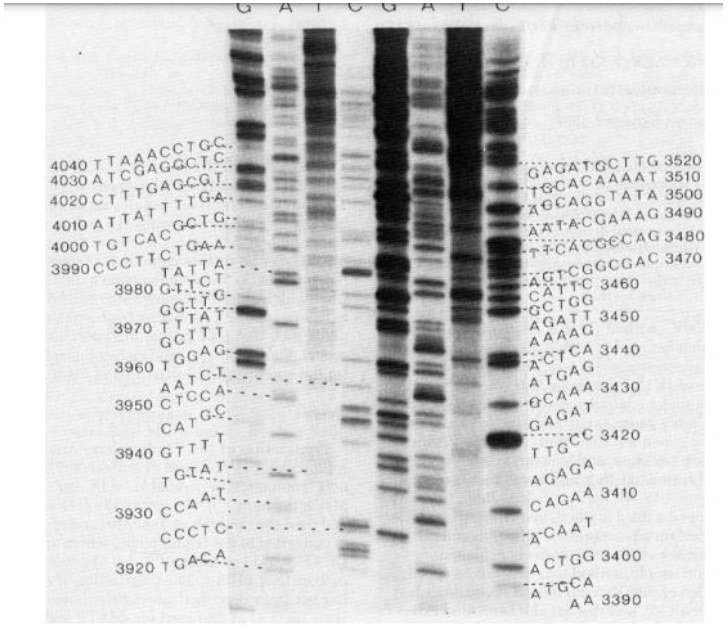
Recent advances in genomic DNA sequencing of microbial species from single cells

Roger S. Lasken and Jeffrey S. McLean


SNP and indel analyses
Deletion
MDA1
(13,216
reads)
0
8
1,534
MDA2
(66 reads)
0
8
14
MDA3
(527 reads)
0
8
79
SNPs
SNPs
B De novo assembly
Ba Whole-genome comparisons
Bb Gene discovery
1,909 genes
W83
2,090 genes
ATCC 33277
2,217 genes
TDC60
2,290 genes
JCVI SC001
171
255
28
281
56
0
524
67
17
1,500
88
24
135
46
44
Bc Structural variation analyses
W83
TDC60
MDA3
megL
megL
megL
160-bp leader sequence
36-bp CRISPR sequence
Spacer sequence
1,000
2,000
3,000
Position (bp)

A CULPA É DO SANGER...





Nobel de Química
1980

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

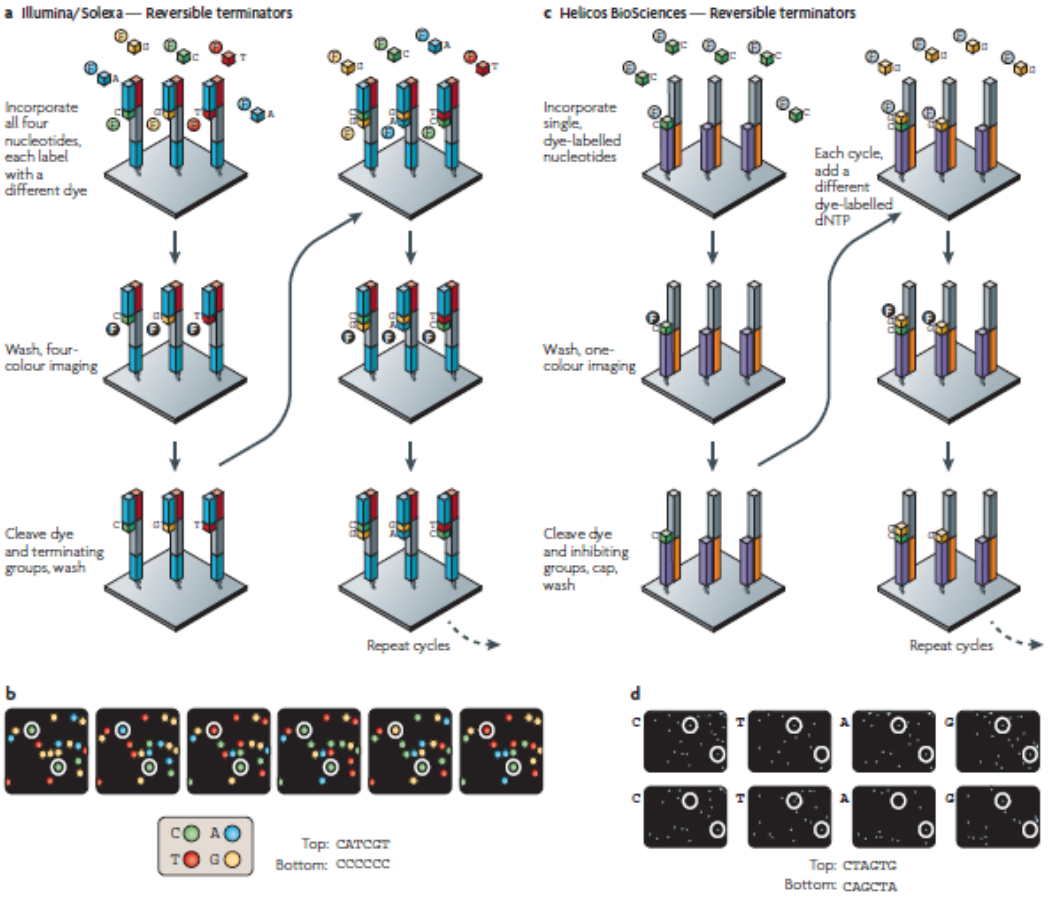
F. SANGER, S. NICKLEN, AND A. R. COULSON

Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977

Sequencing technologies — the next generation

*Michael L. Metzker**‡



APPLICATIONS OF NEXT-GENERATION SEQUENCING

Repetitive DNA and next-generation sequencing: computational challenges and solutions

Todd J. Treangen[1] and Steven L. Salzberg[1,2]

Sequencing depth and coverage: key considerations in genomic analyses

David Sims, Ian Sudbery, Nicholas E. Ilott, Andreas Heger and Chris P. Ponting

SINGLE-CELL OMICS

Single-cell genome sequencing: current state of the science

Charles Gawad[1], Winston Koh[2,3] and Stephen R. Quake[2,3]

High-throughput functional genomics using CRISPR–Cas9

Ophir Shalem, Neville E. Sanjana and Feng Zhang

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Unravelling the genomic targets of small molecules using high-throughput sequencing

Raphaël Rodriguez[1,2] and Kyle M. Miller[3]

Genomes by design

Adrian D. Haimovich[1,2], Paul Muir[1–3] and Farren J. Isaacs[1,2]

19 de Abril - Aula “Sinalização celular”

Plasmídios e novas tecnologias de clonagem: Gibson, Gateway e outros