GENOMA HUMANO

Prof. Leandro Martins de Freitas, Phd

IMS/UFBA

HISTÓRICO

Complexidade dos organismos é muito grande

Complexidade da formação do fenótipo é muito grande

Muitos modelos hereditários e genéticos

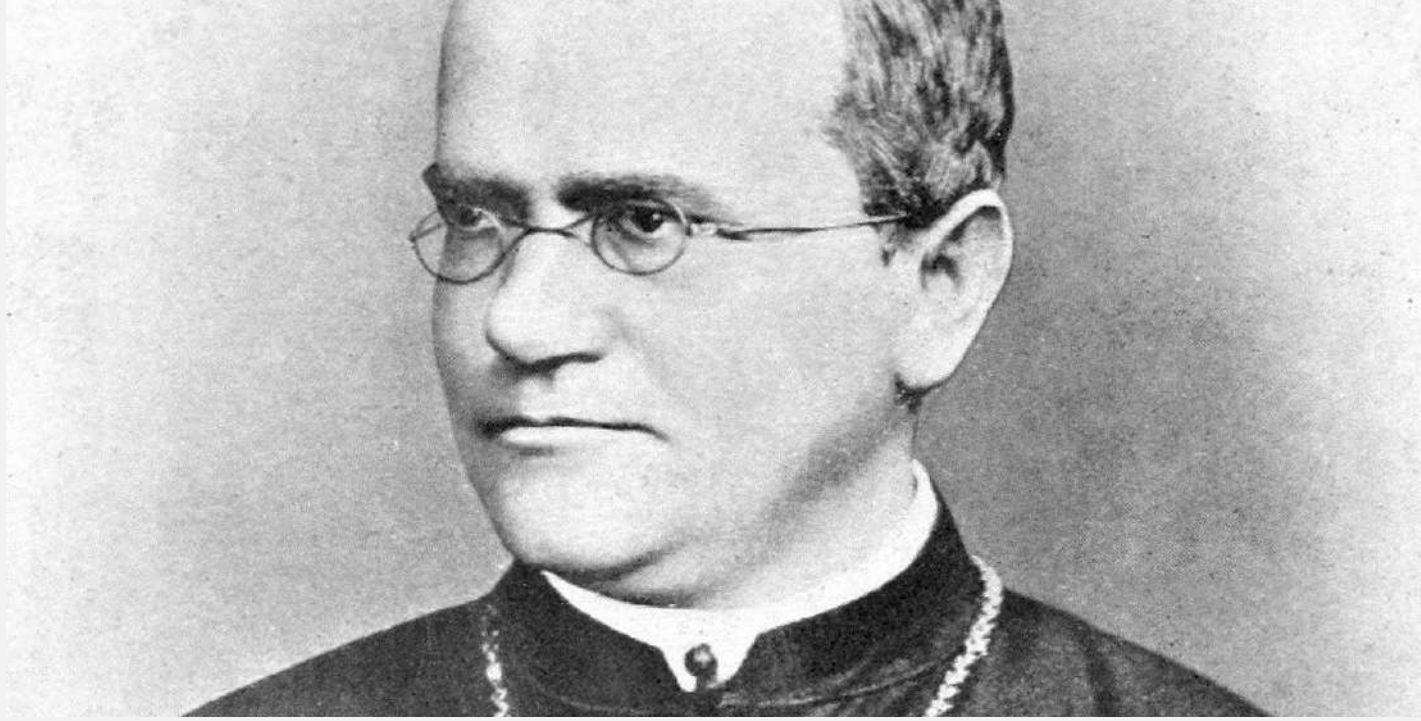
HISTÓRICO

Genes não atuam sozinhos - Interação dos genes

Conhecer somente a sequência de um gene não é suficiente

Necessidade de conhecimento da informação genética

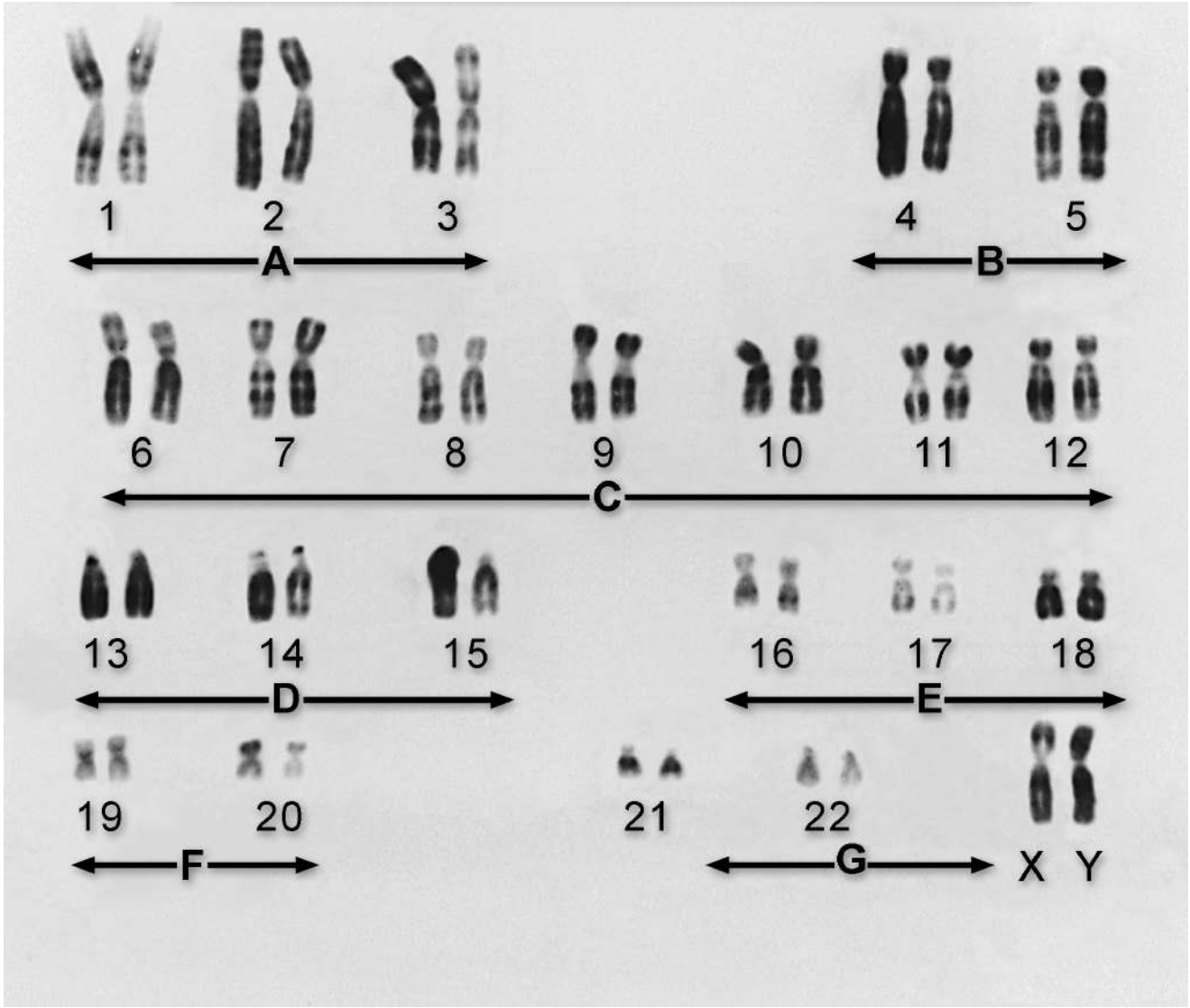
GENÉTICA

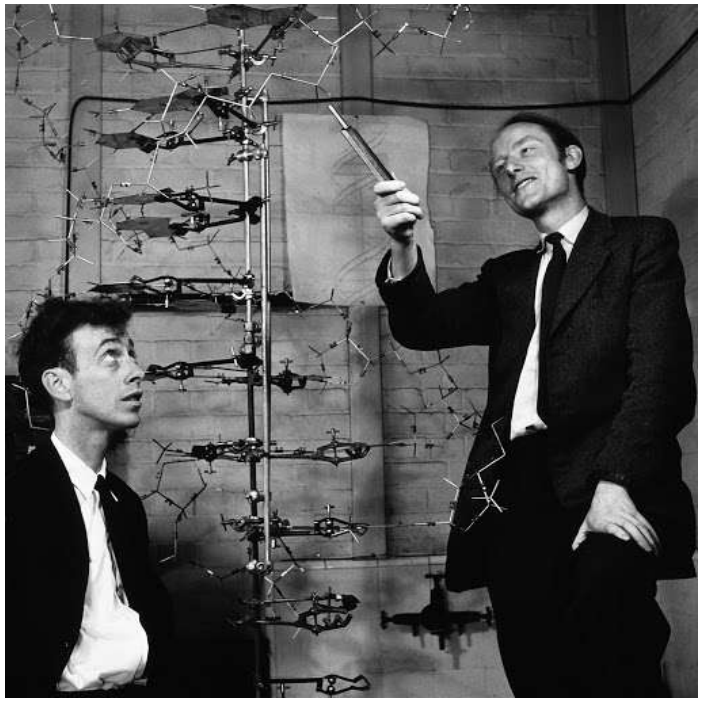


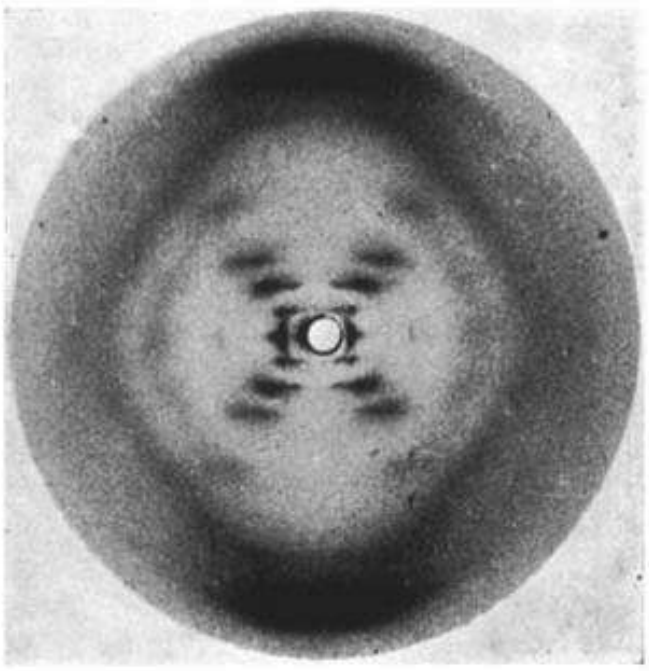
Characteristics of pea plants Gregor Mendel used in his inheritance experiments

Seeds		Flower colour	Pod		Stem	
form	cotyledons		form	colour	position of inflorences	size
round roundish	yellow	white	full	yellow	axial	long
wrinkled	green	violett-red	constricted between the seeds	green	terminal	short

CROMOSSOMOS

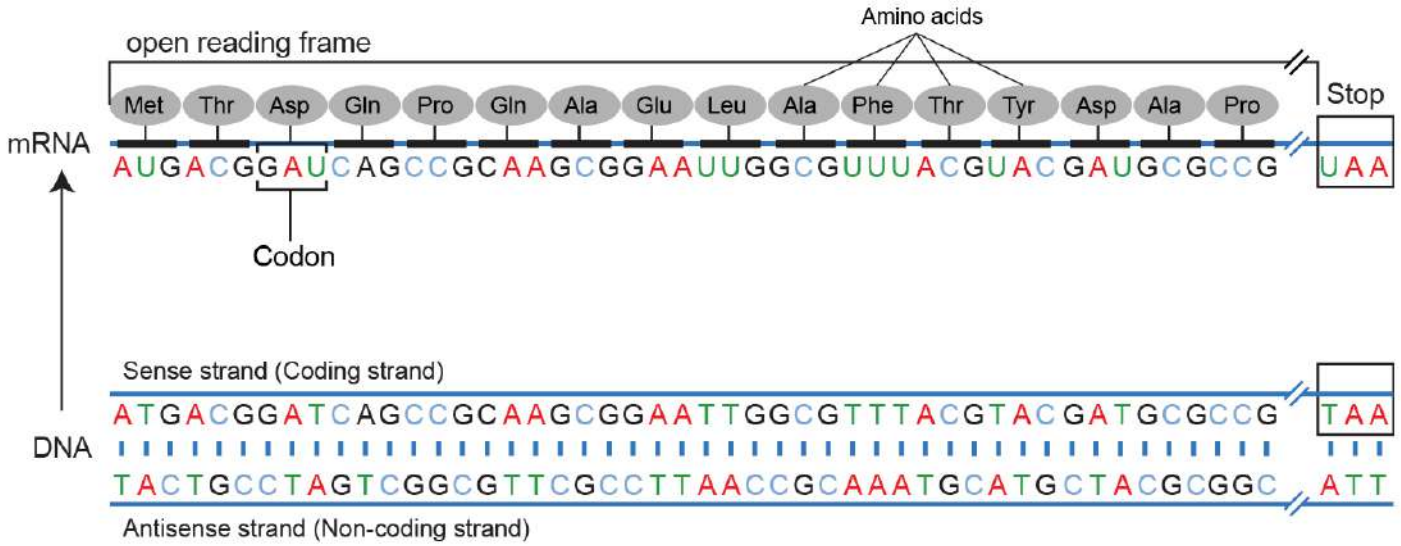






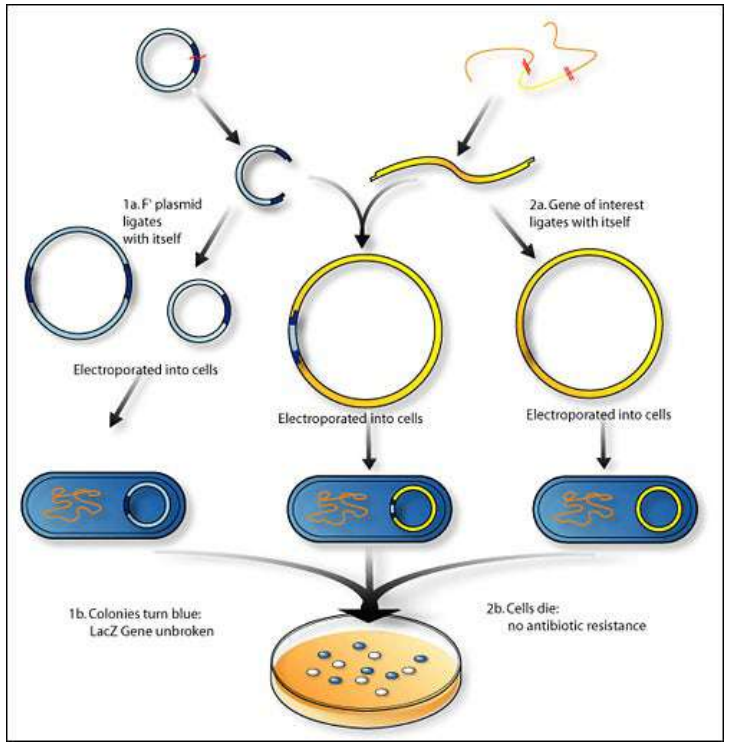
DNA – DUPLA HÉLICE

LEITURA DO DNA

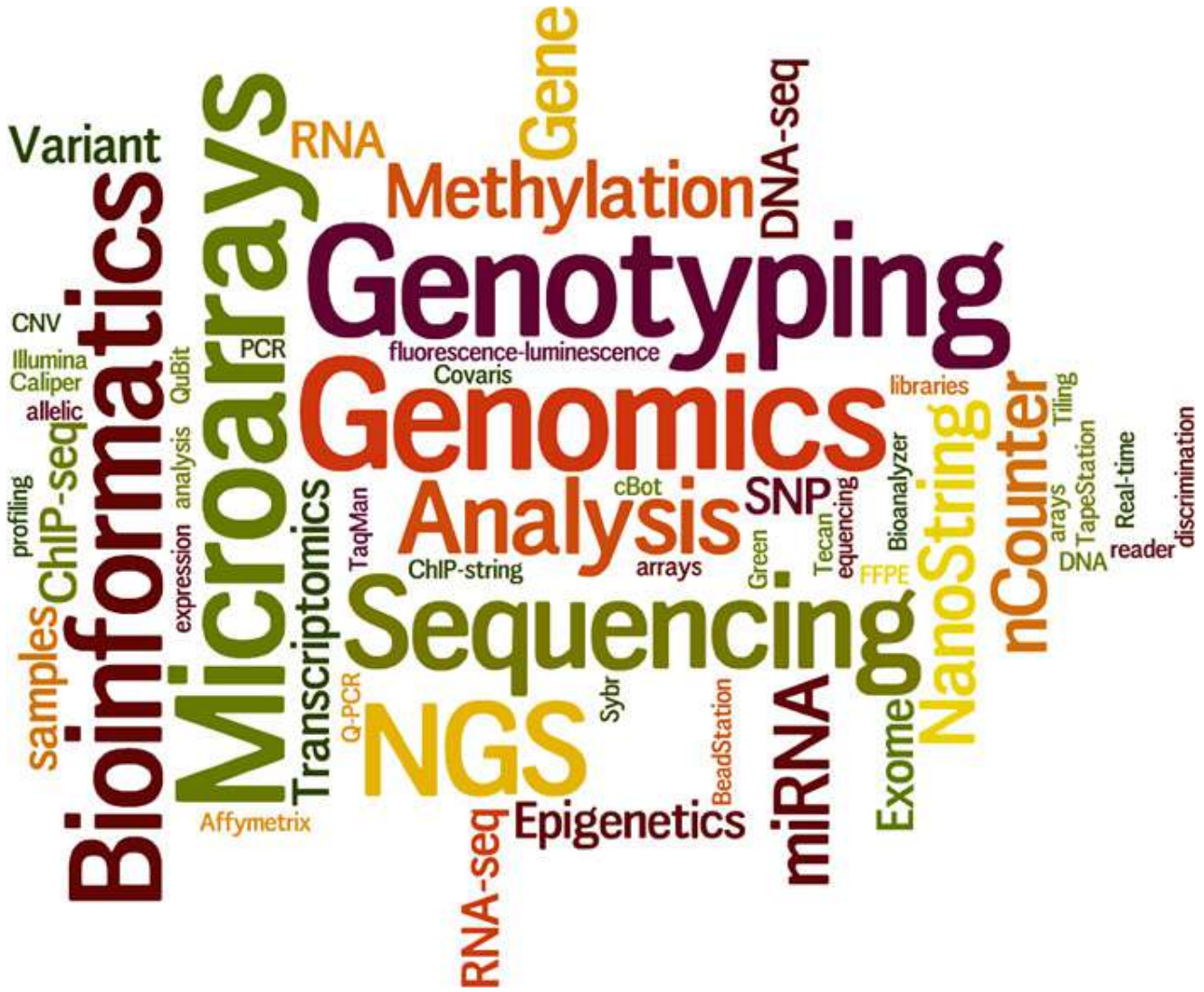


Second letter

First letter	U	C	A	G	
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr **UAA Stop** **UAG Stop**	UGU Cys UGC Cys **UGA Stop** UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile **AUG Met**	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G



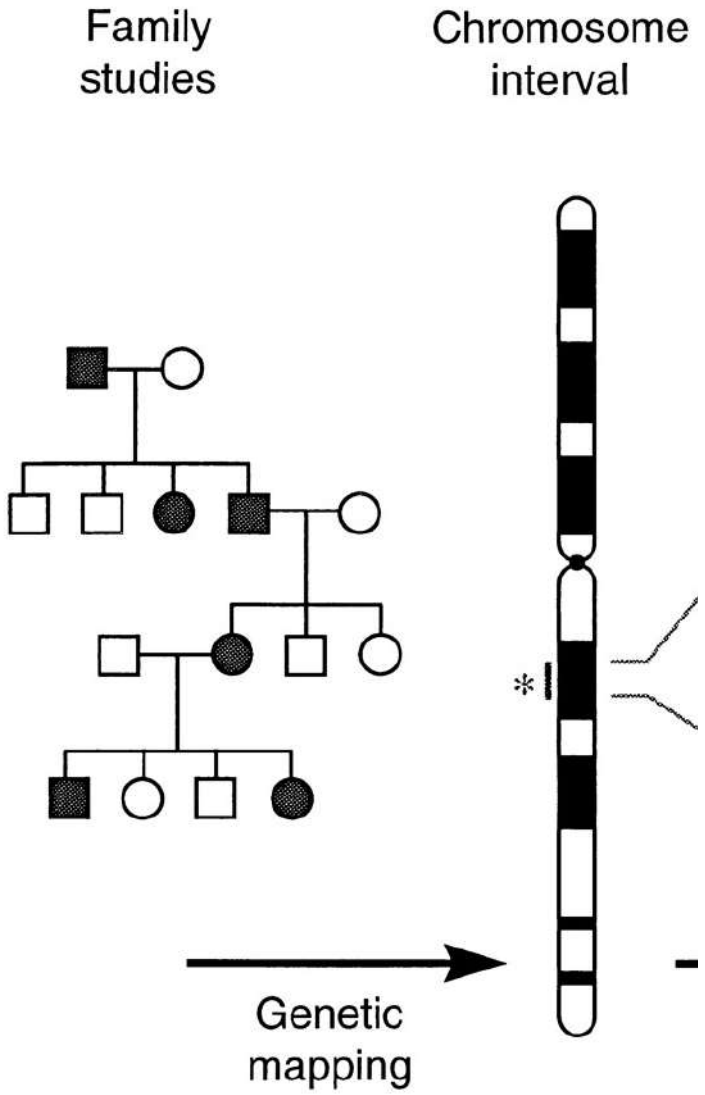
GENÔMICA


Variant
RNA
Gene
DNA-seq
Methylation
Genotyping
Microarrays
Bioinformatics
CNV
Illumina
Caliper
allelic
PCR
fluorescence-luminescence
Covaris
QuBit
Genomics
libraries
Tiling
nCounter
NanoString
ChIP-seq
profiling
analysis
expression
Transcriptomics
TaqMan
cBot
SNP
Analysis
Bioanalyzer
arays
TapeStation
Real-time
DNA
reader
discrimination
ChIP-string
arrays
Green
Tecan
equencing
FFPE
samples
Sequencing
Q-PCR
Sybr
NGS
BeadStation
Exome
miRNA
Affymetrix
RNA-seq
Epigenetics

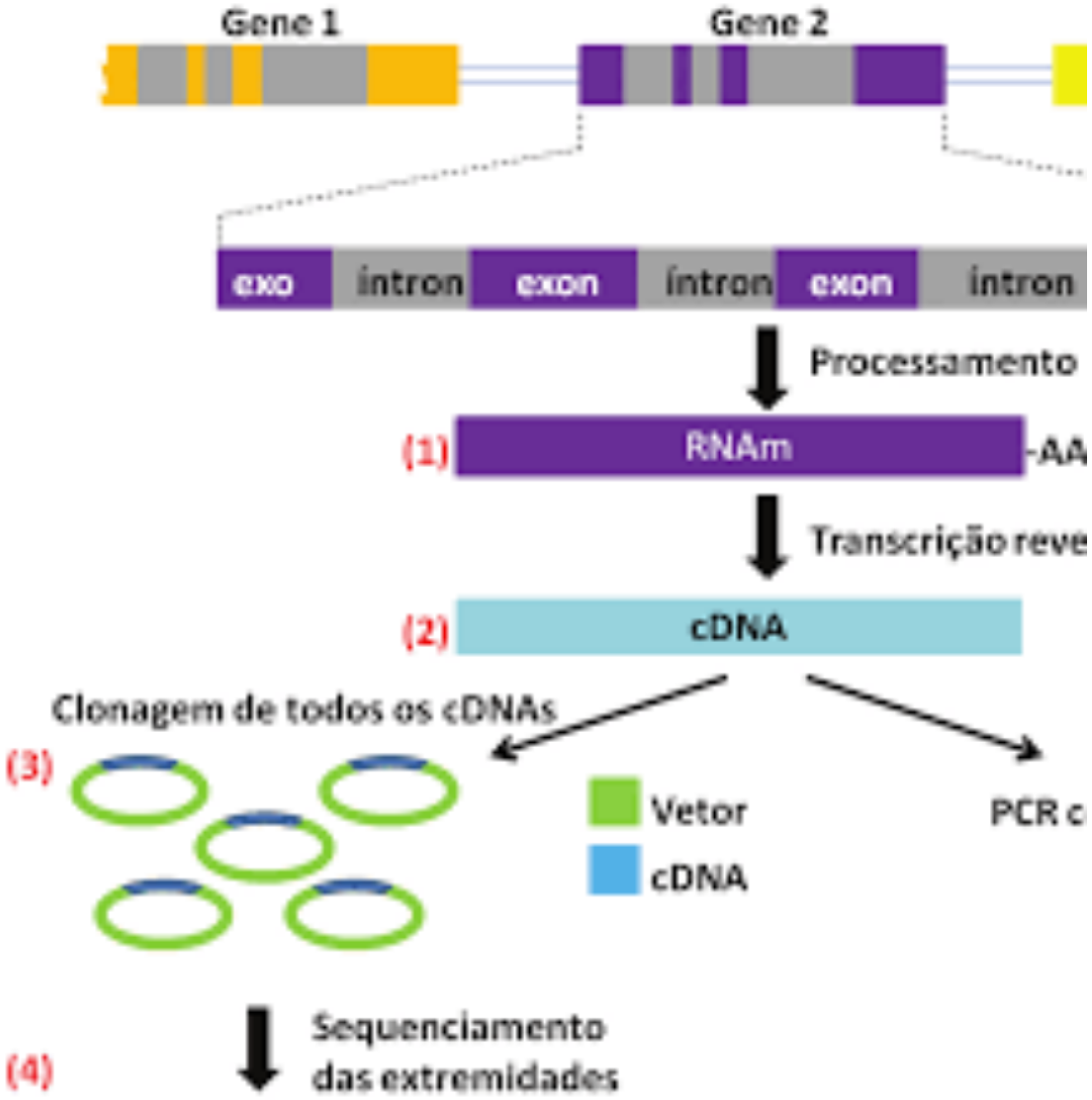
PRIMEIROS GENOMAS

- Virus bacteriano ΦX174 – 1977, 1978
- Virus bacteriano Lambda – 1982
- Vírus SV40 -1978
- Mitocôndria humana 1981

MAPA GENÉTICO



SEQUENCIAMENTO SHOTGUN DE CDNA

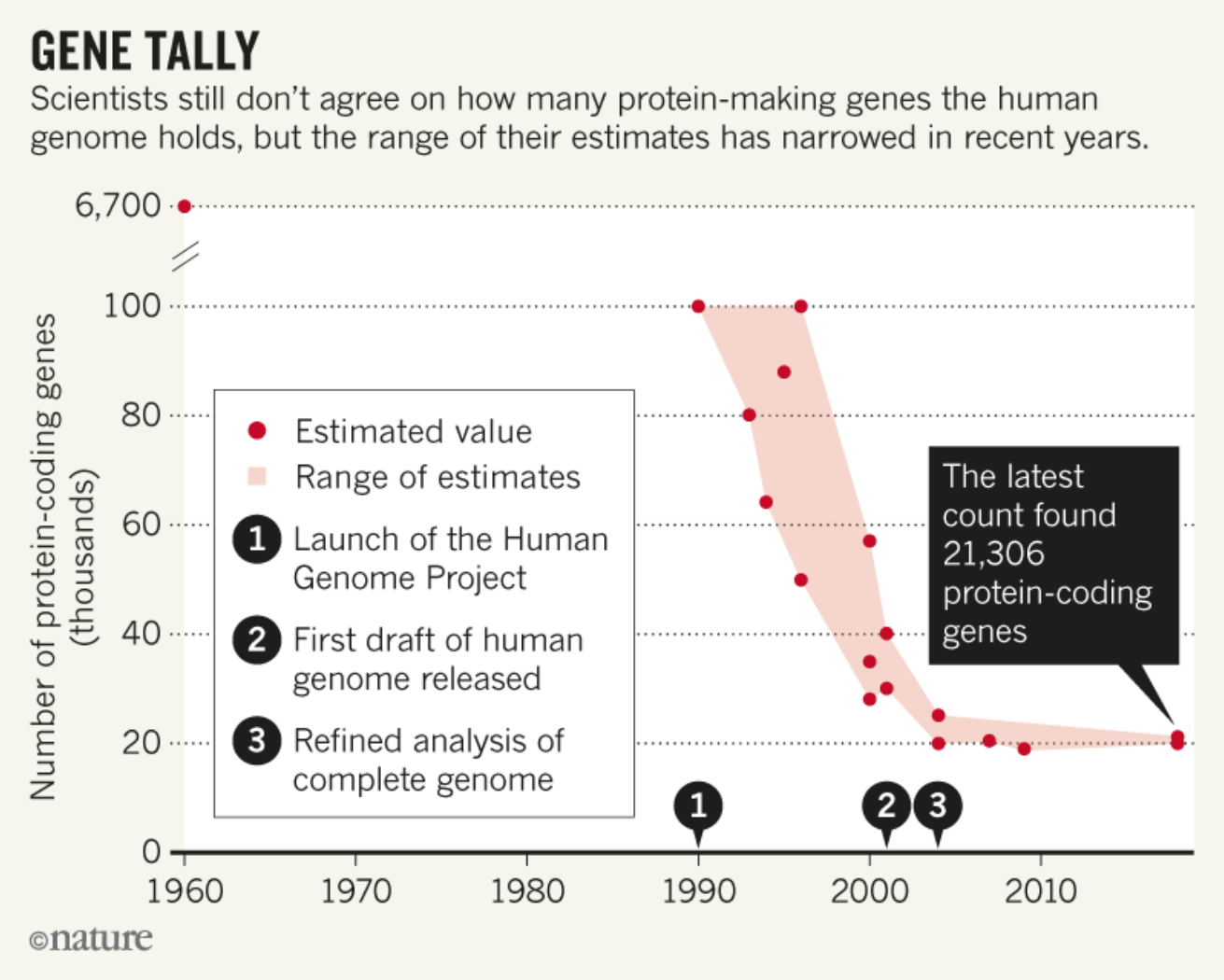


GENOMA HUMANO

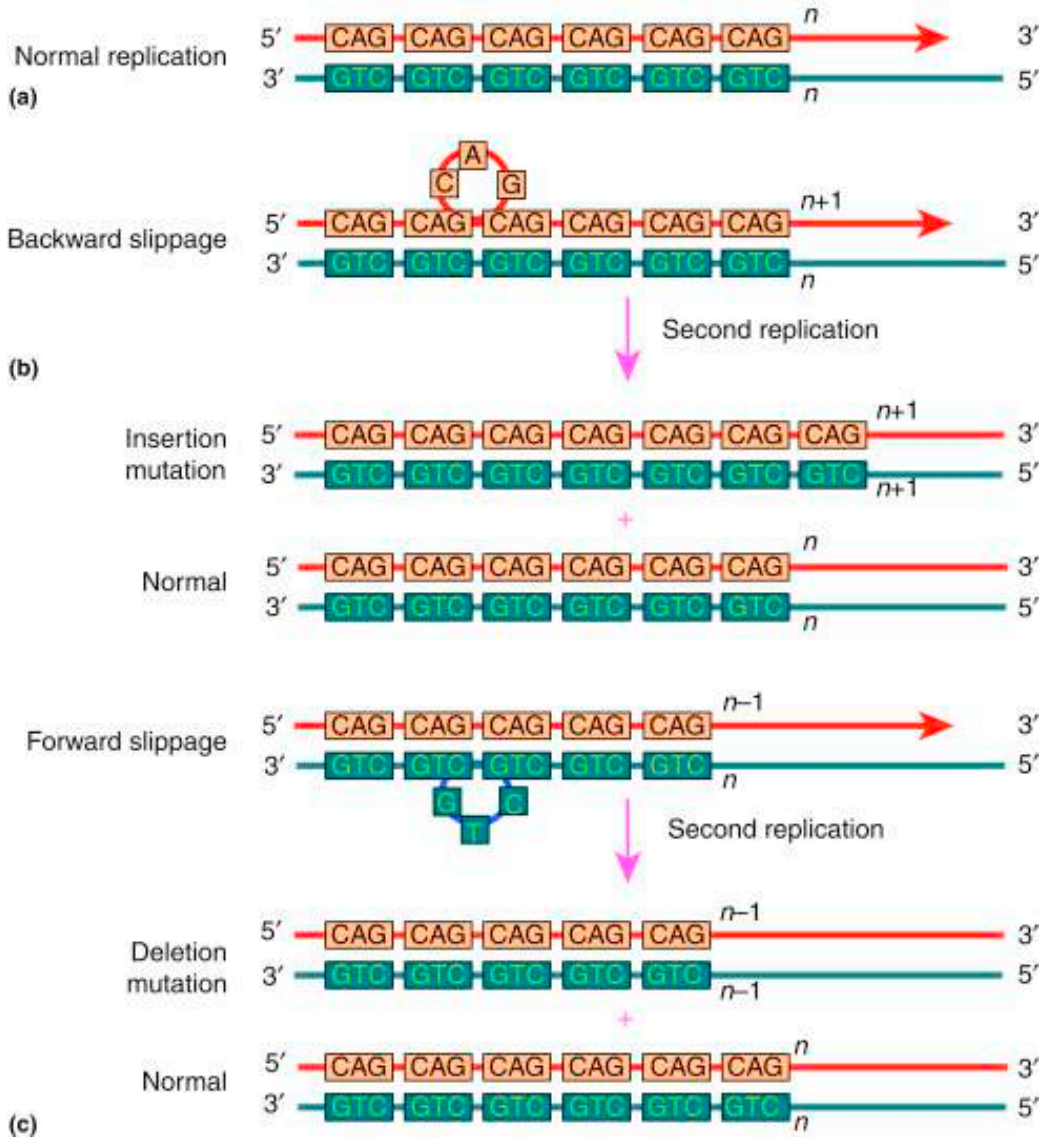
30-40 mil genes, estimativas no início

96% das região de eucromatina foi sequenciado (94% do genoma)

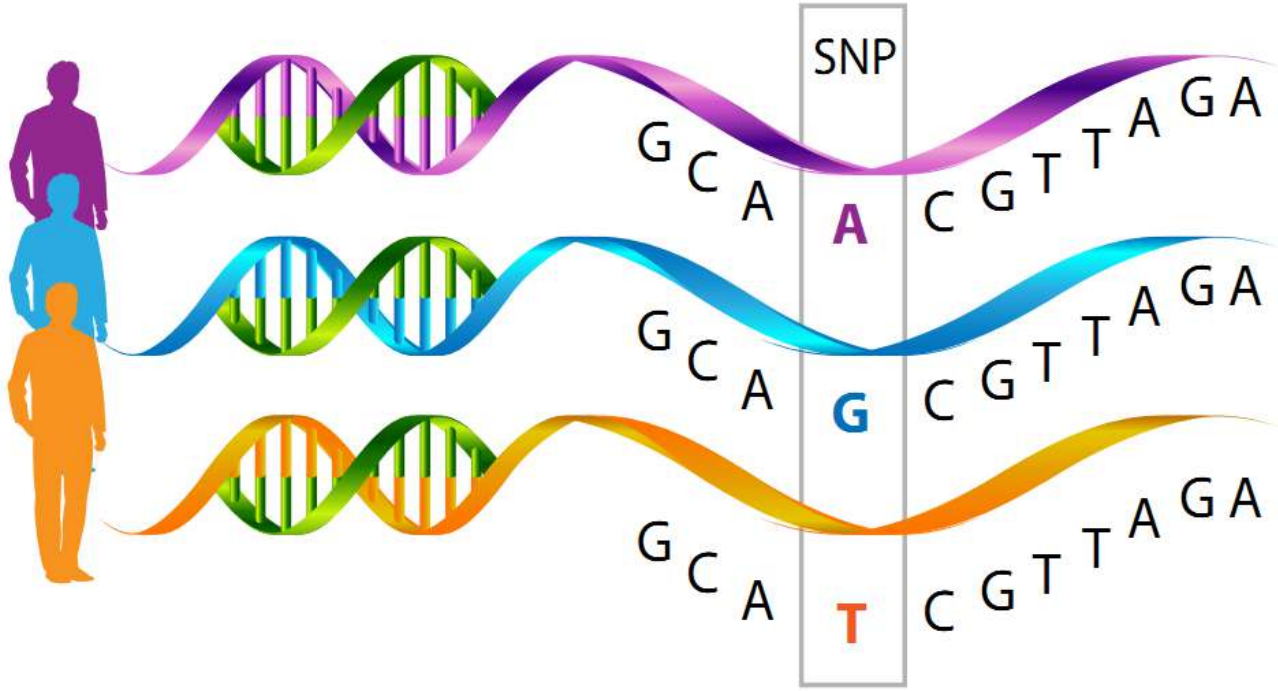
GENES HUMANOS - PROTEINAS



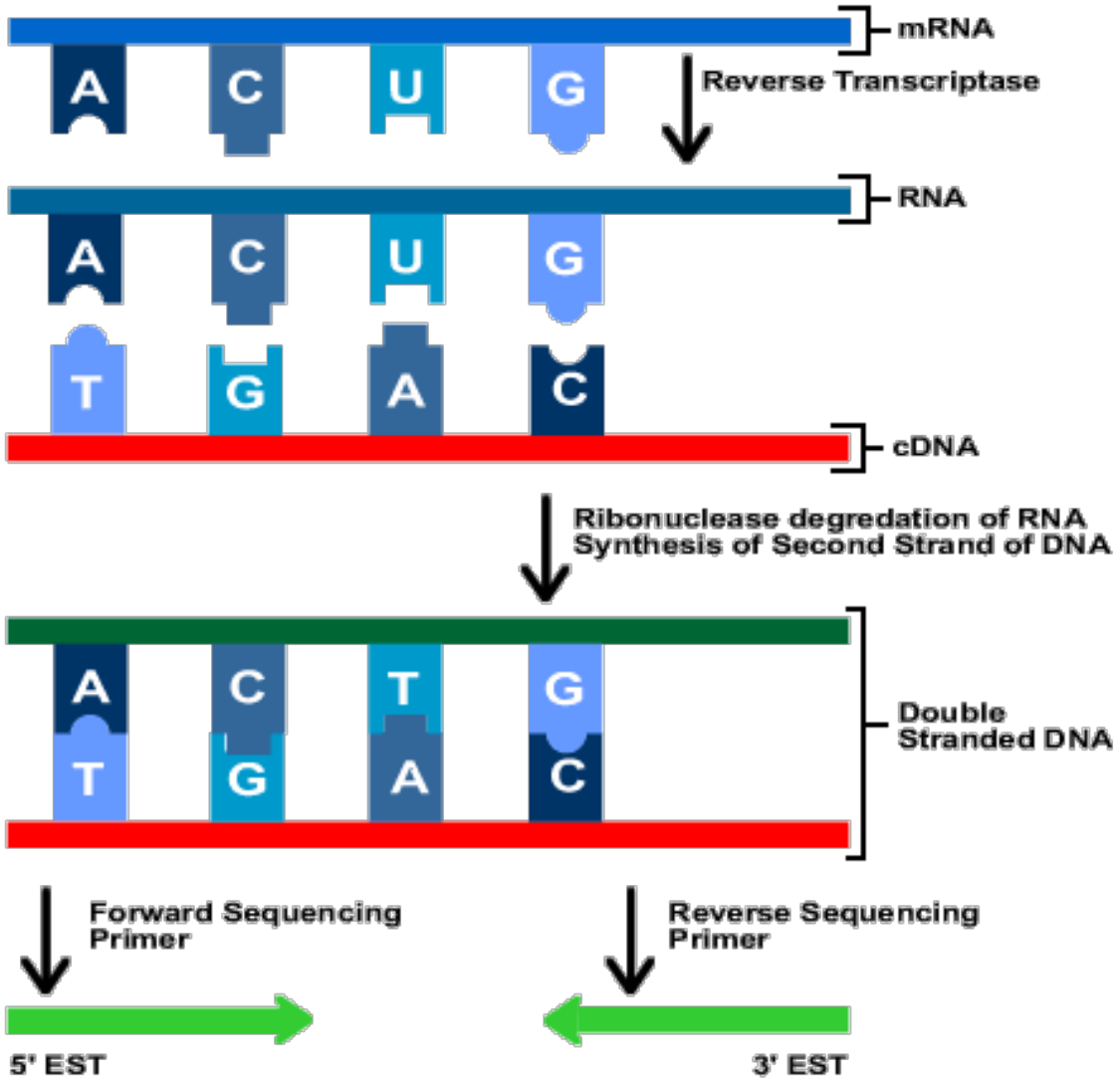
MICROSSATÉLITES

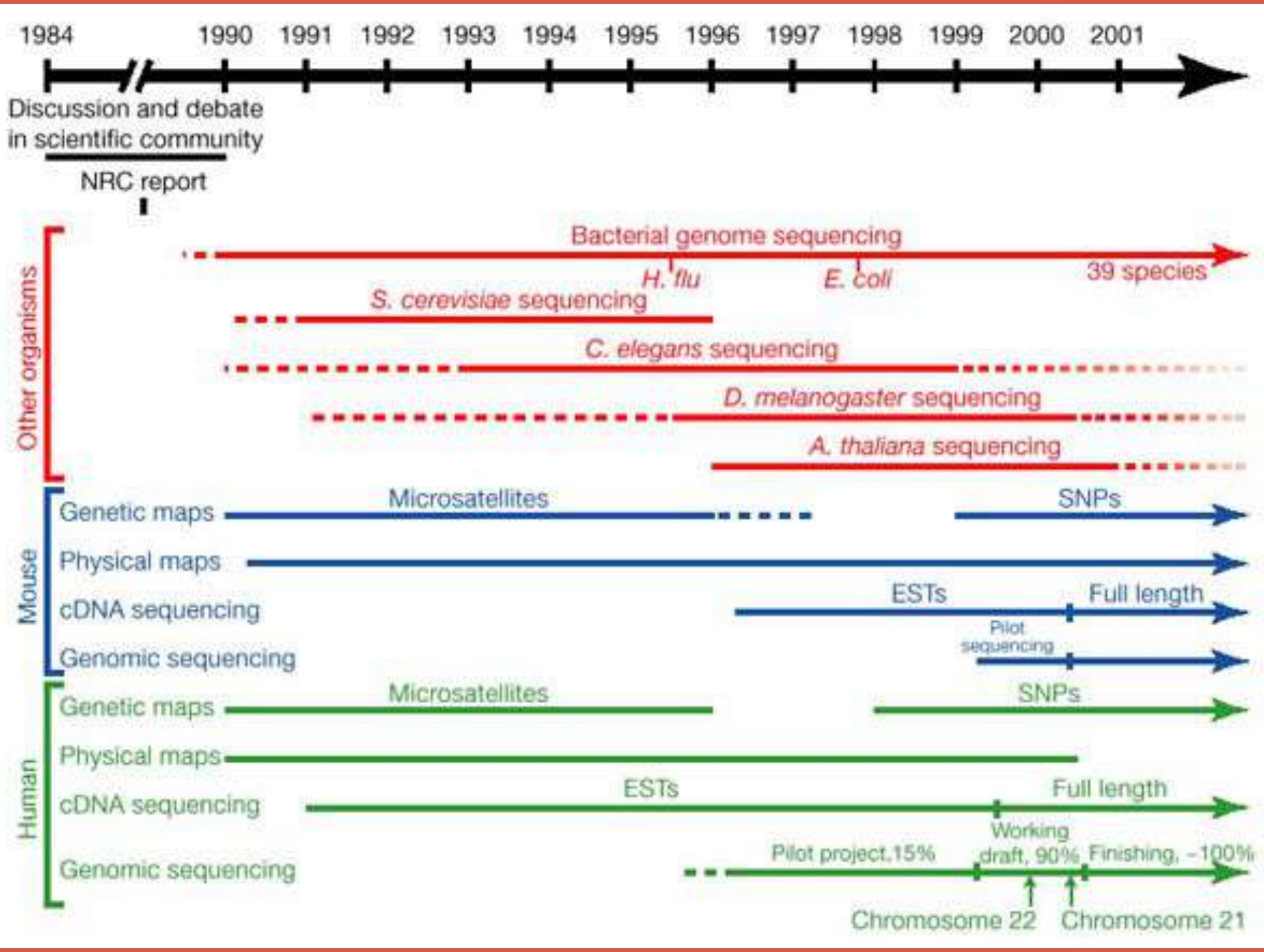


SNP

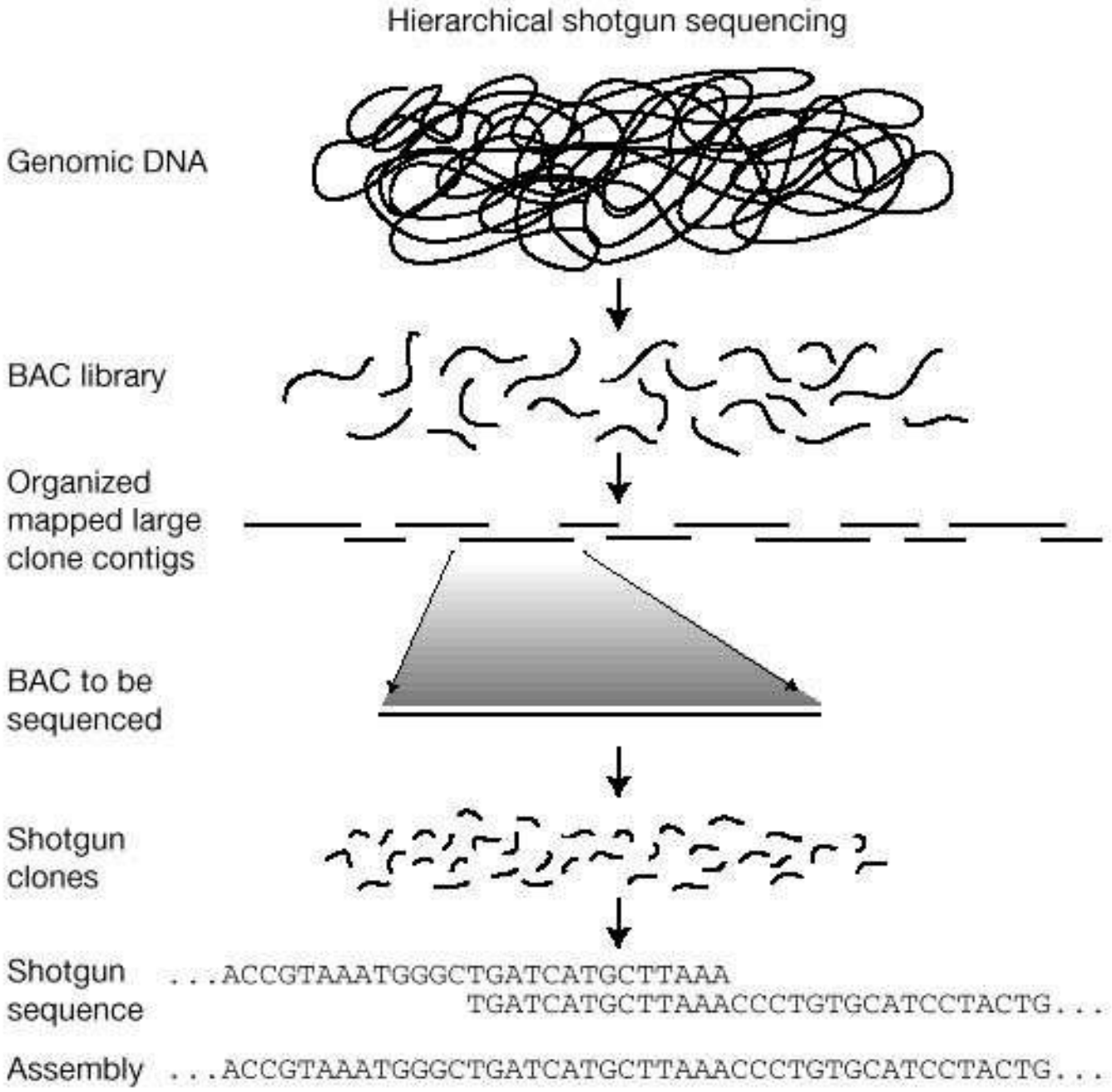


EST - ESPRESSION SEQUENCE TAG

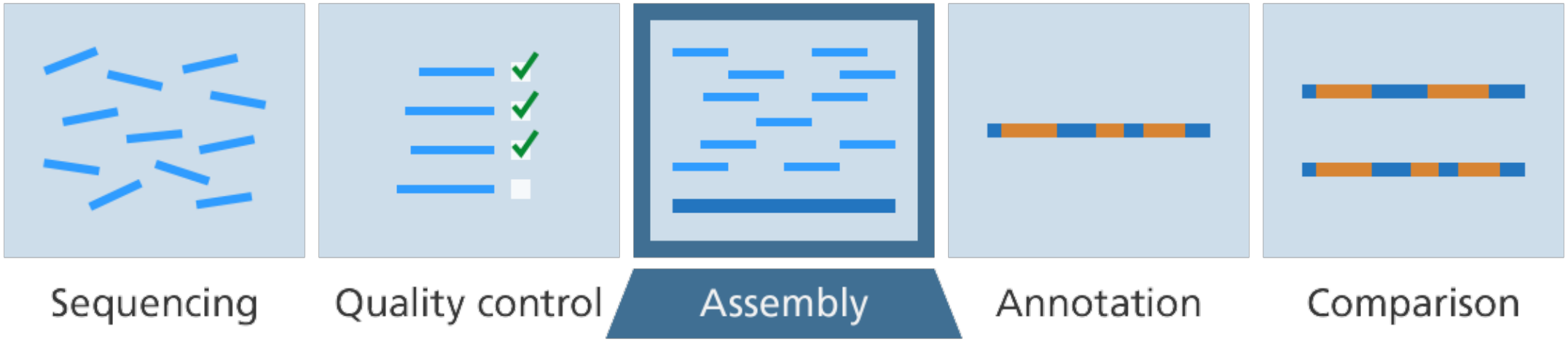



1984
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
Discussion and debate in scientific community
NRC report
Other organisms
Bacterial genome sequencing
H. flu
E. coli
39 species
S. cerevisiae sequencing
C. elegans sequencing
D. melanogaster sequencing
A. thaliana sequencing
Mouse
Genetic maps
Microsatellites
SNPs
Physical maps
cDNA sequencing
ESTs
Full length
Pilot sequencing
Genomic sequencing
Human
Genetic maps
Microsatellites
SNPs
Physical maps
cDNA sequencing
ESTs
Full length
Genomic sequencing
Pilot project,15%
Working draft, 90%
Finishing, ~100%
Chromosome 22
Chromosome 21

GENOMA - SHOTGUN



PRODUÇÃO AUTOMATIZADA - CENTER FOR GENOME RESEARCH



GENOMA E BIOINFORMÁTICA

ESTRATÉGIAS DE MONTAGEM

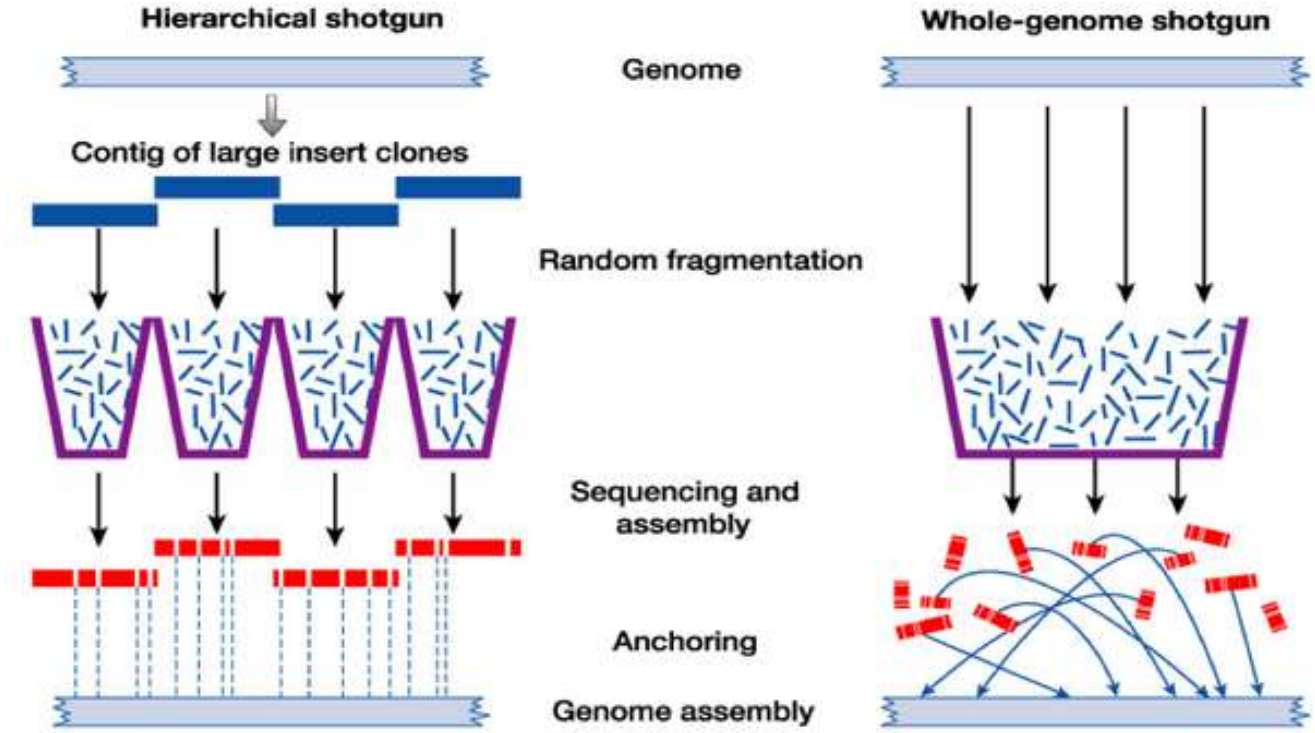


Figure 8-3 Human Molecular Genetics, 3/e. (© Garland Science 2004)

ANALOGIA DO GENOMA

A single recipe is like…

A Gene:
One set of instructions for how to make one protein.

A recipe book is like…

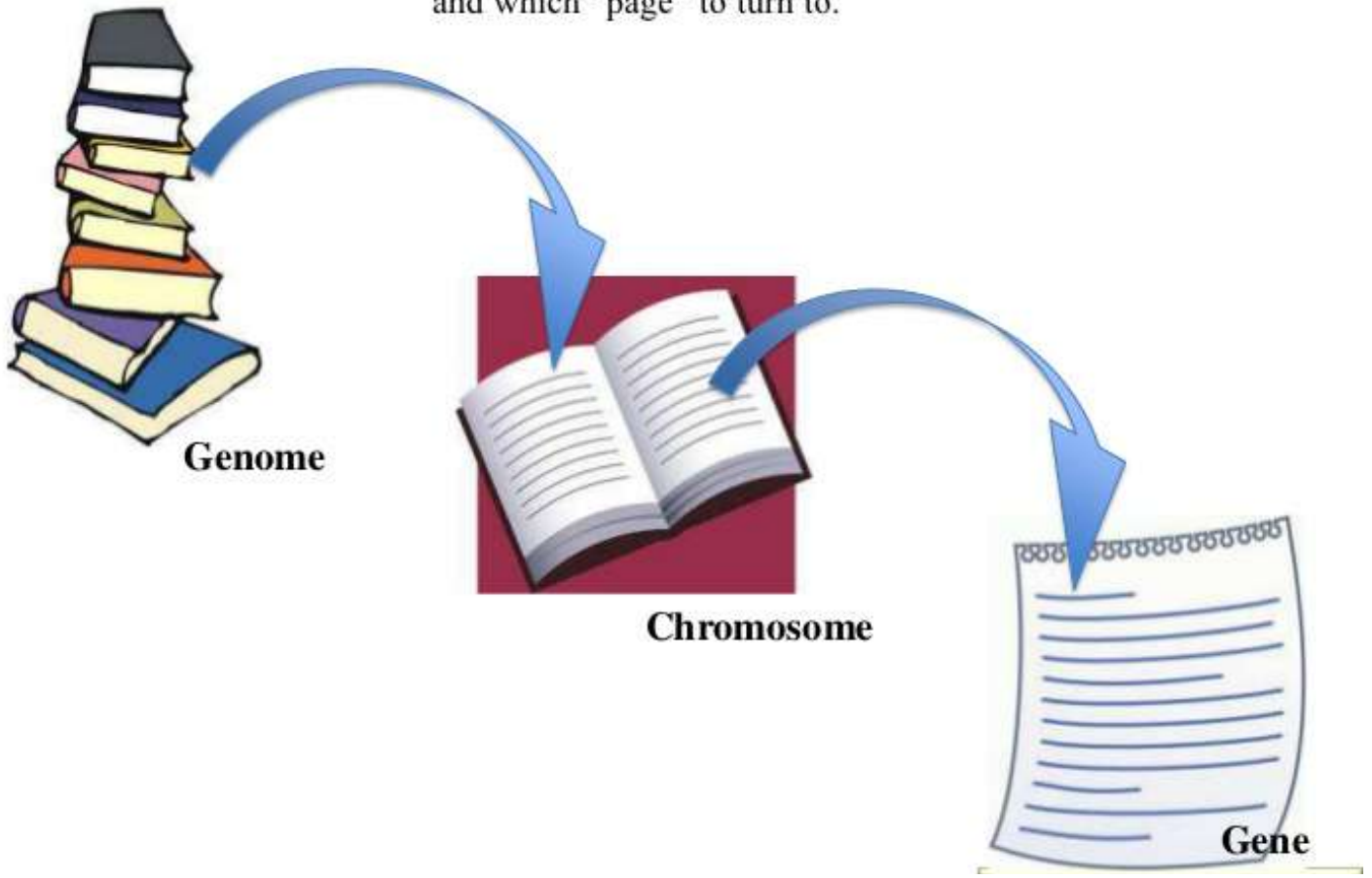
A Chromosome:
Thousands of sets of instructions for how to make thousands of proteins

Two copies of 23 recipe books is like…

Our Genome:
ALL of the sets of instructions for how to make ALL the proteins we need

ALL (gene, chromosome, genome) are written in same the DNA alphabet!!!

In order for scientists to find a specific gene, they need to know which "book" to look in, and which "page" to turn to.

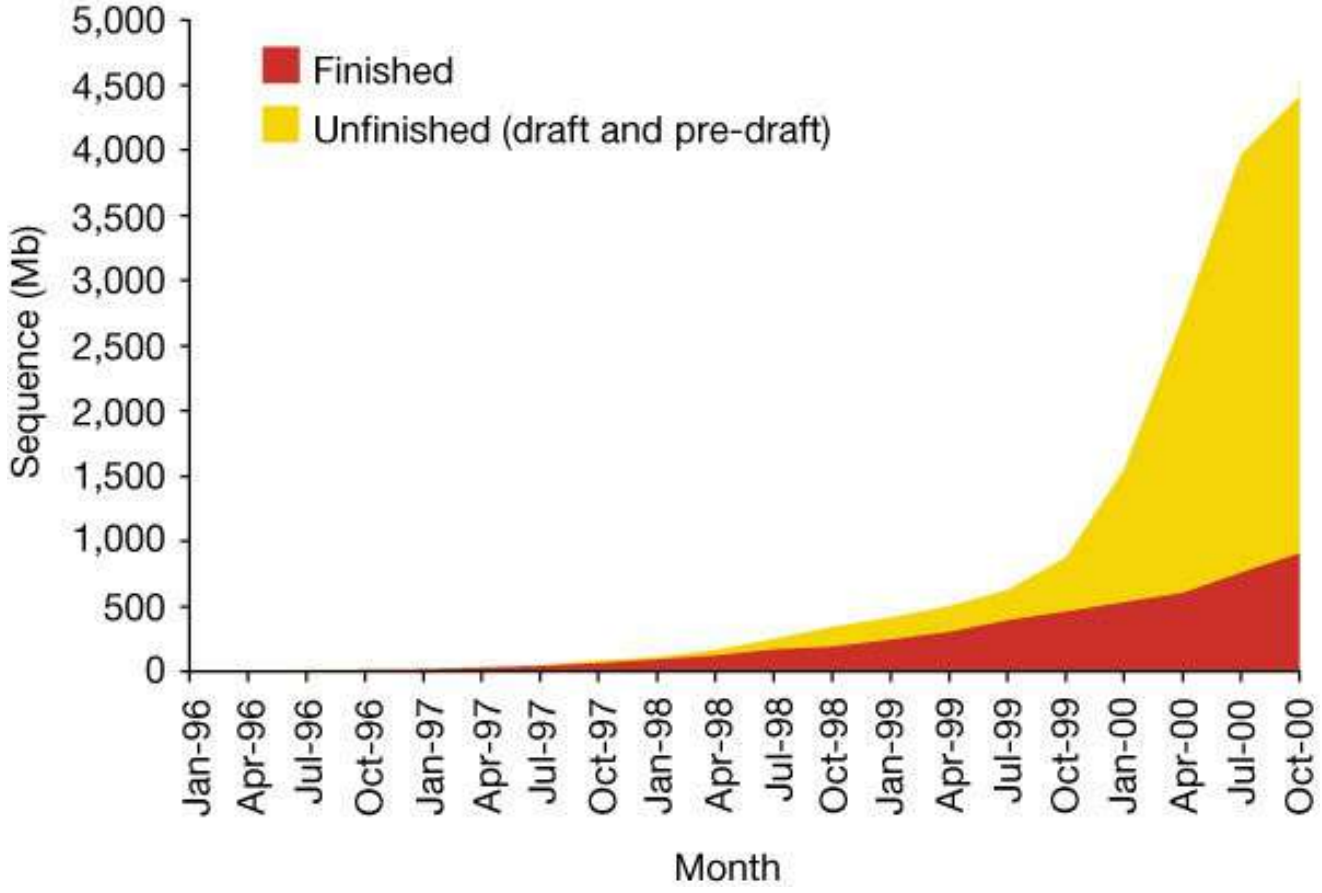


MONTAGEM DE UM GIGANTESO QUEBRA-CABEÇA

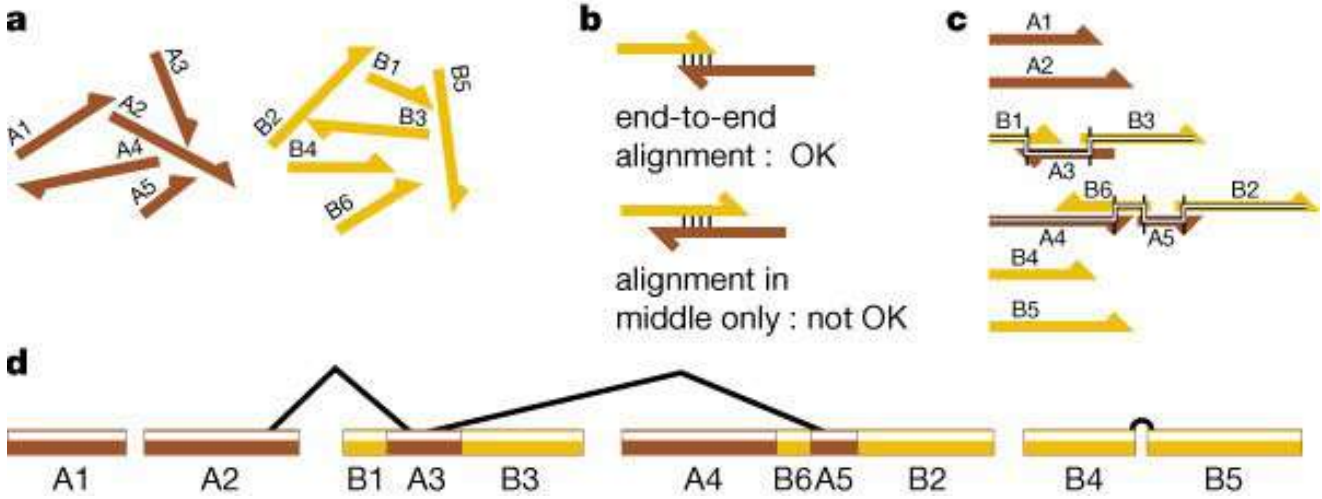
GIGANTESCO QUEBRA-CABEÇA SEM A CAIXA

SÓ FALTA MONTAR

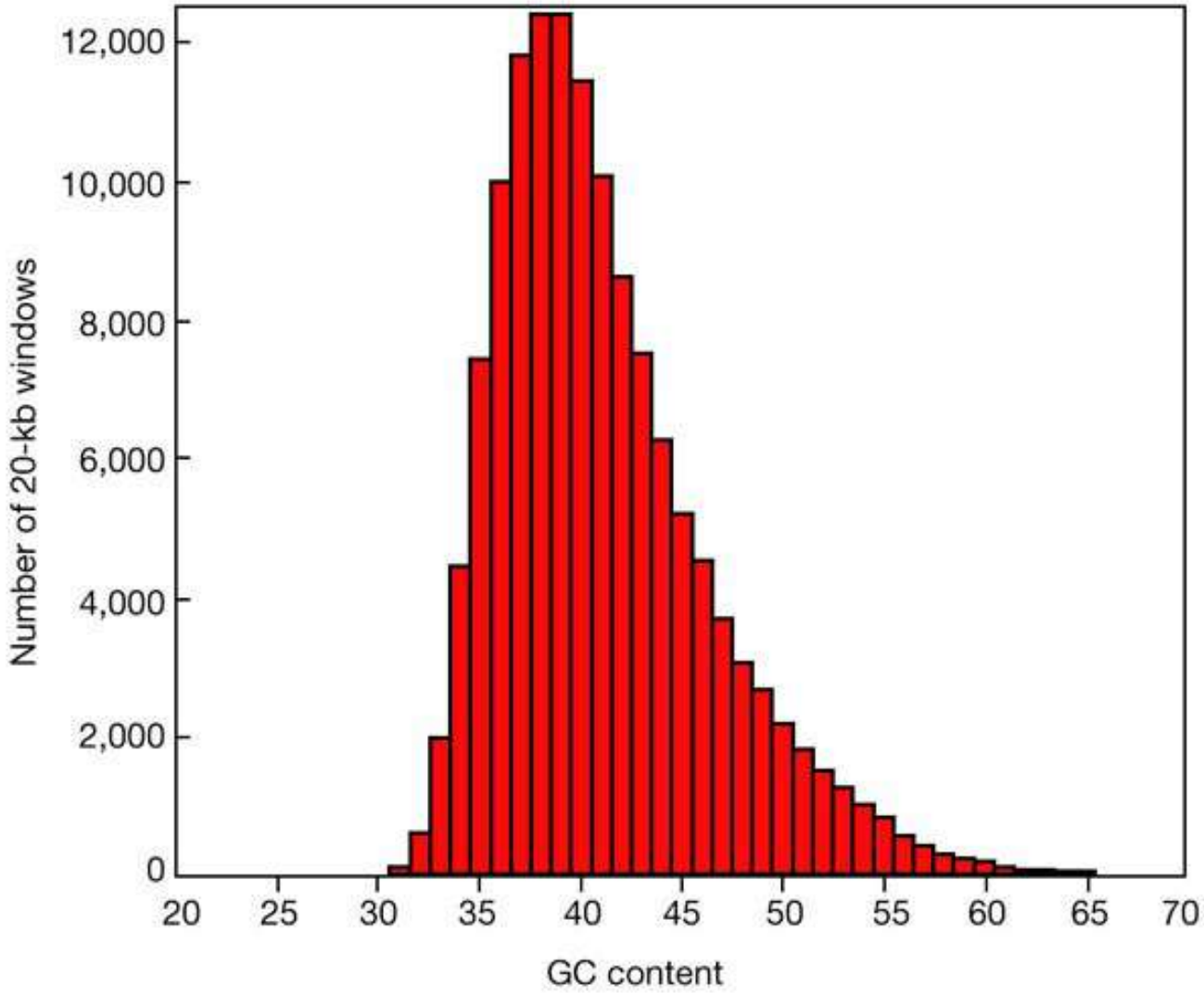
HIGH THROUGHPUT GENOME SEQUENCE



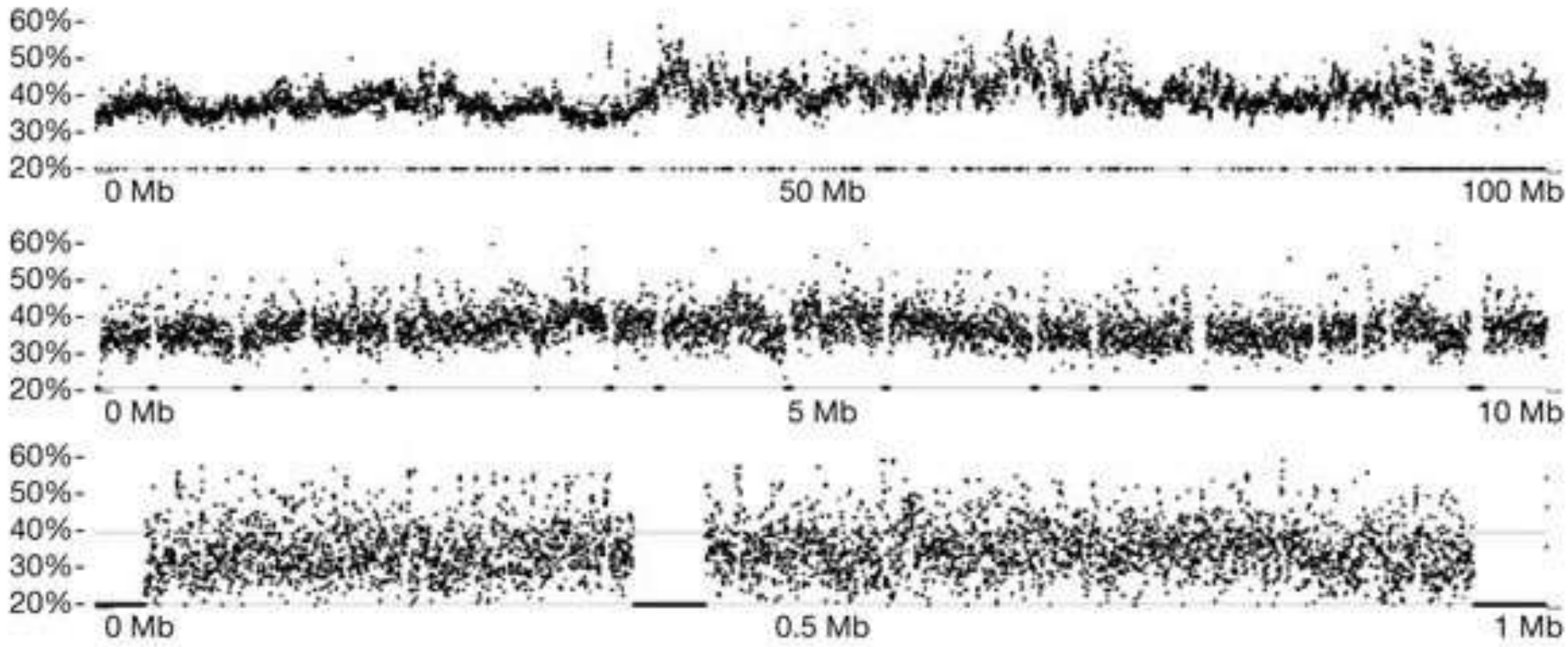
ASSEMBLING CLONES INTO THE DRAFT GENOME



CONTEÚDO GC

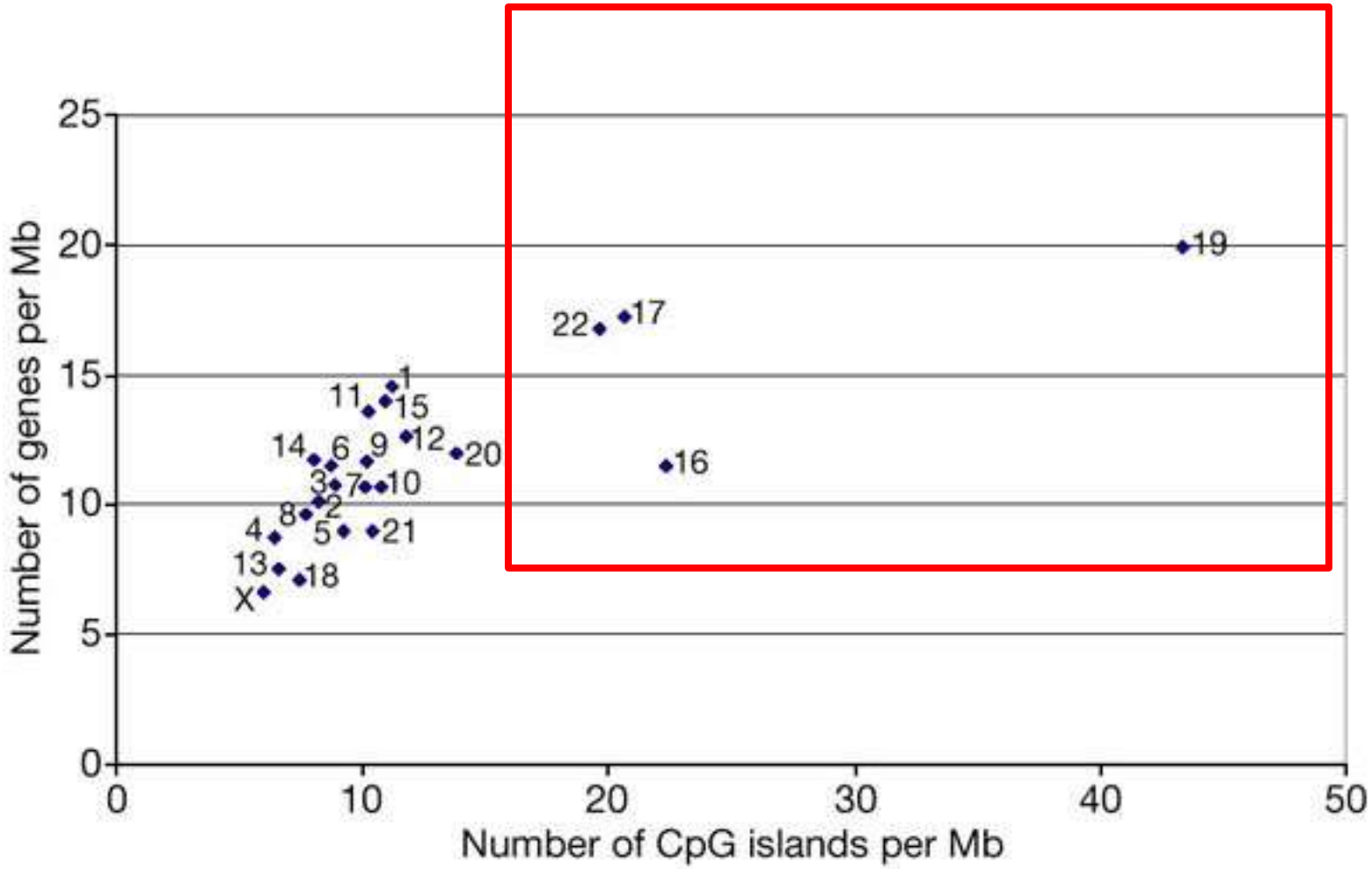

12,000
10,000
8,000
6,000
4,000
2,000
0
Number of 20-kb windows
20
25
30
35
40
45
50
55
60
65
70
GC content

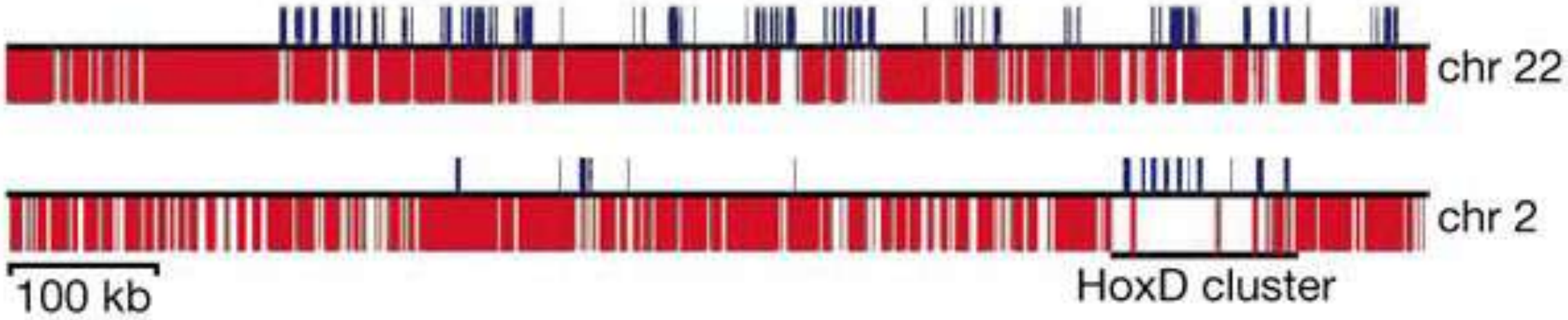
Variation in GC content at various scales.



100-Mb (20-kb windows).
10 Mb (2-kb windows).
1 Mb (200-bp windows)

NUMBER OF CPG ISLANDS PER MB FOR EACH CHROMOSOME

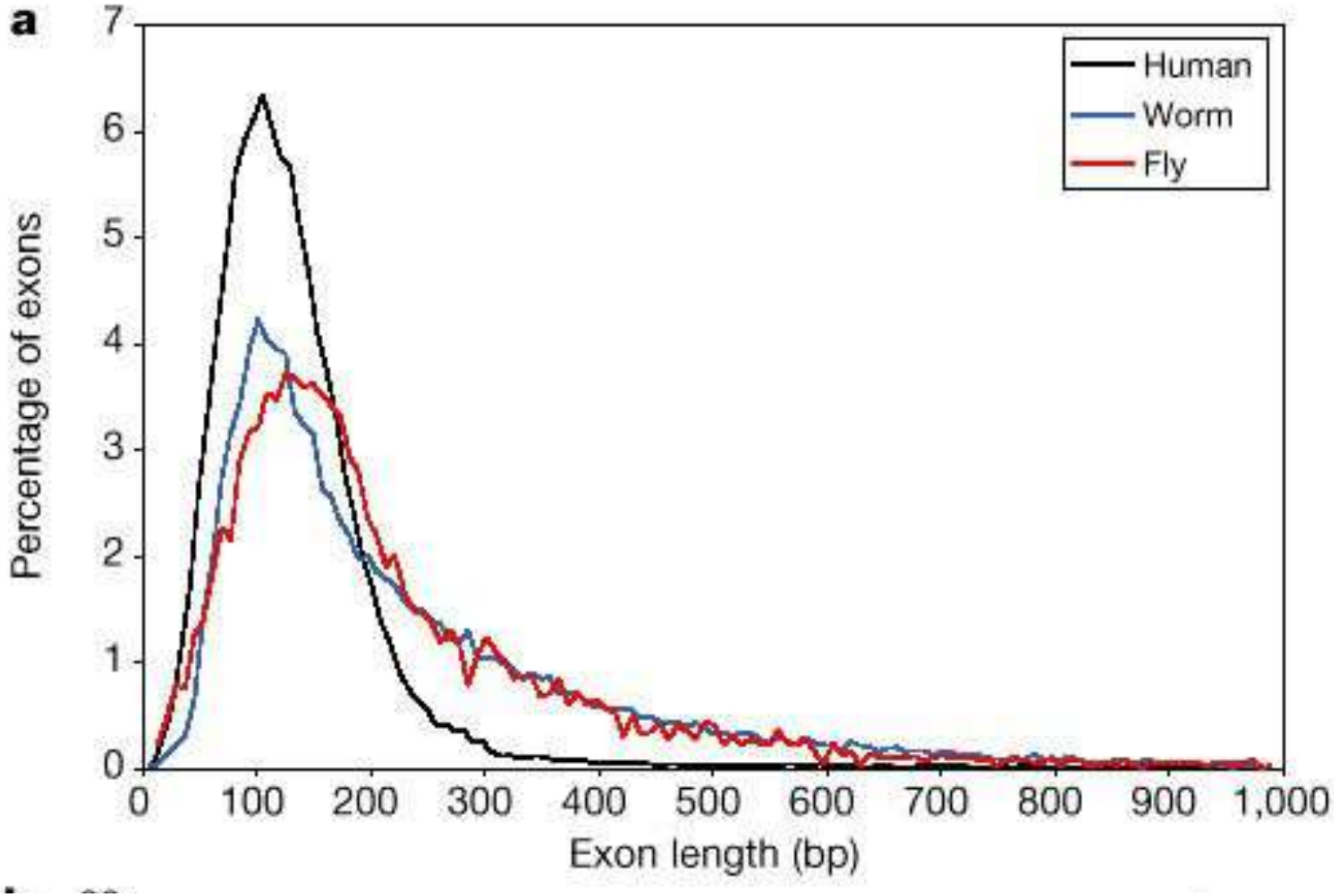




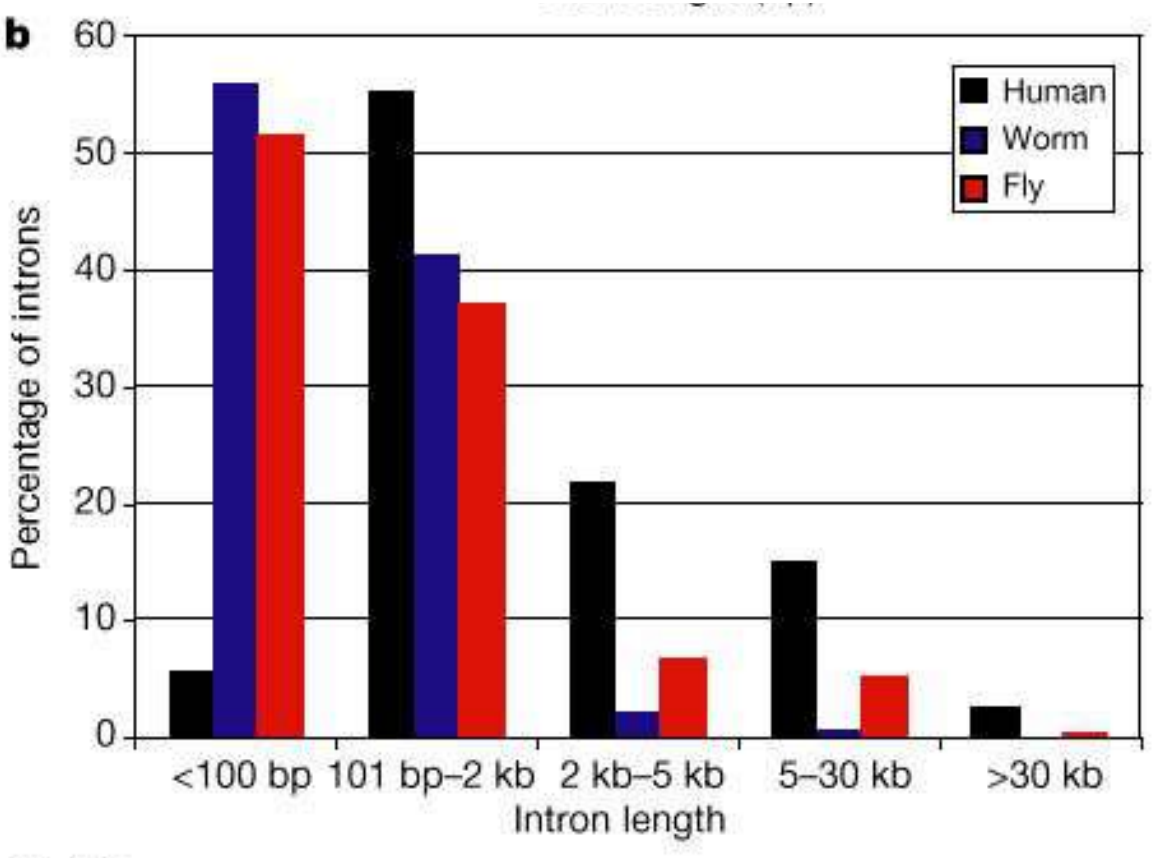
INTERSPERSED REPEATS

interspersed repeats -→ Vermelho
Exons → Azul

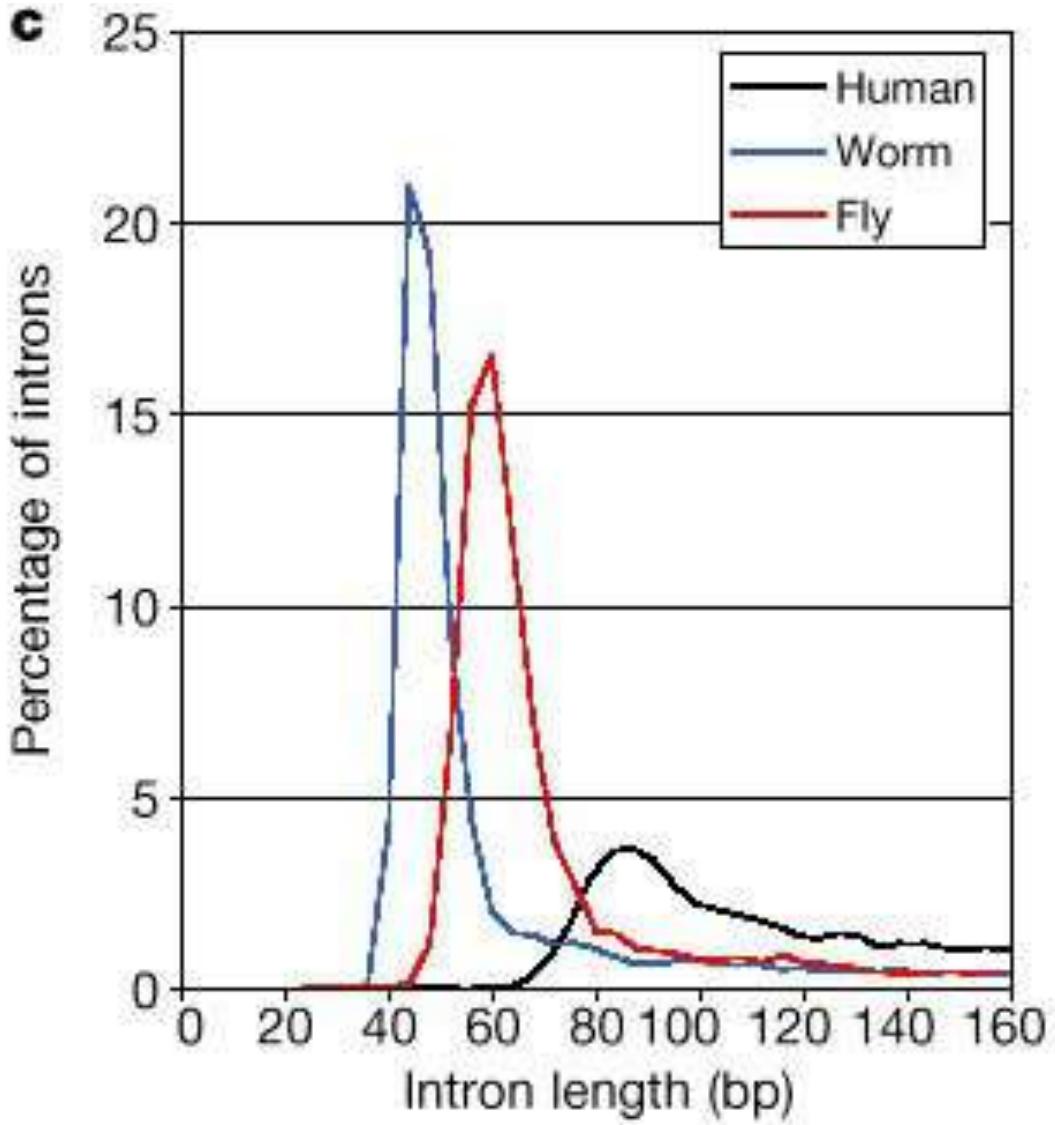
EXONS


a
Percentage of exons
0
1
2
3
4
5
6
7
0
100
200
300
400
500
600
700
800
900
1,000
Exon length (bp)
Human
Worm
Fly

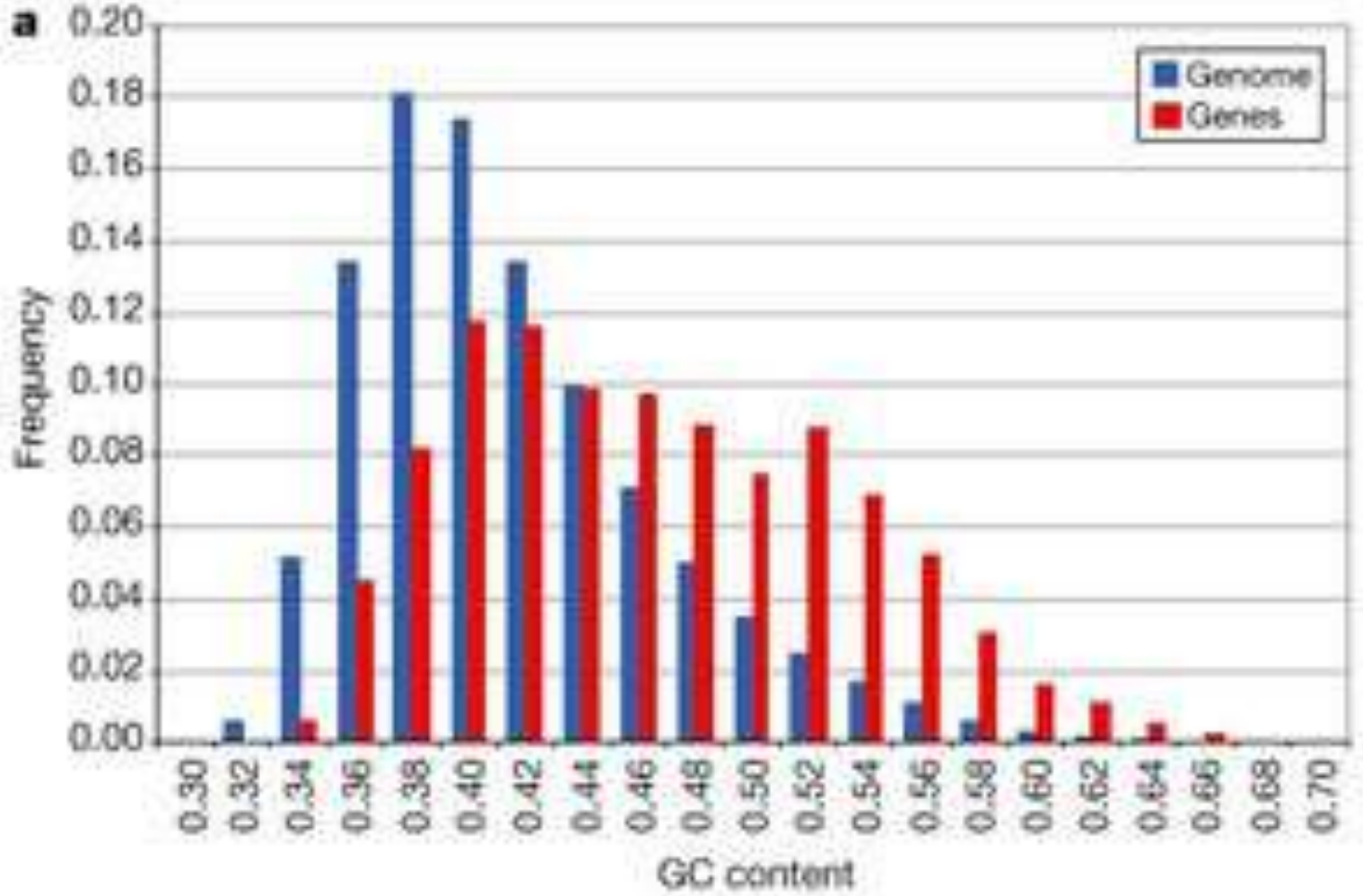
INTRONS


b
60
50
40
30
20
10
0
Percentage of introns
Human
Worm
Fly
<100 bp
101 bp–2 kb
2 kb–5 kb
5–30 kb
>30 kb
Intron length

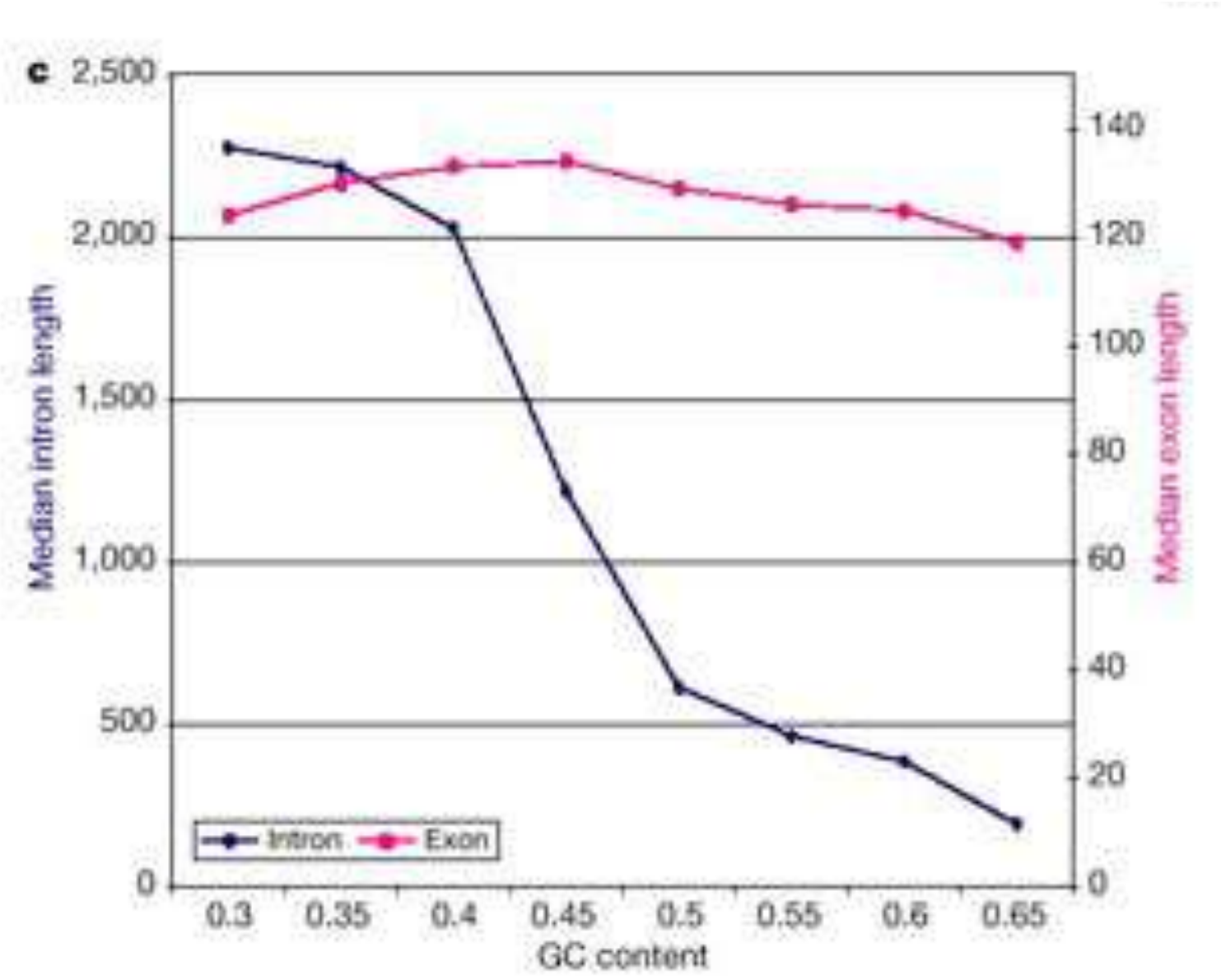
INTRONS


c
25
20
15
10
5
0
Percentage of introns
Human
Worm
Fly
0
20
40
60
80
100
120
140
160
Intron length (bp)

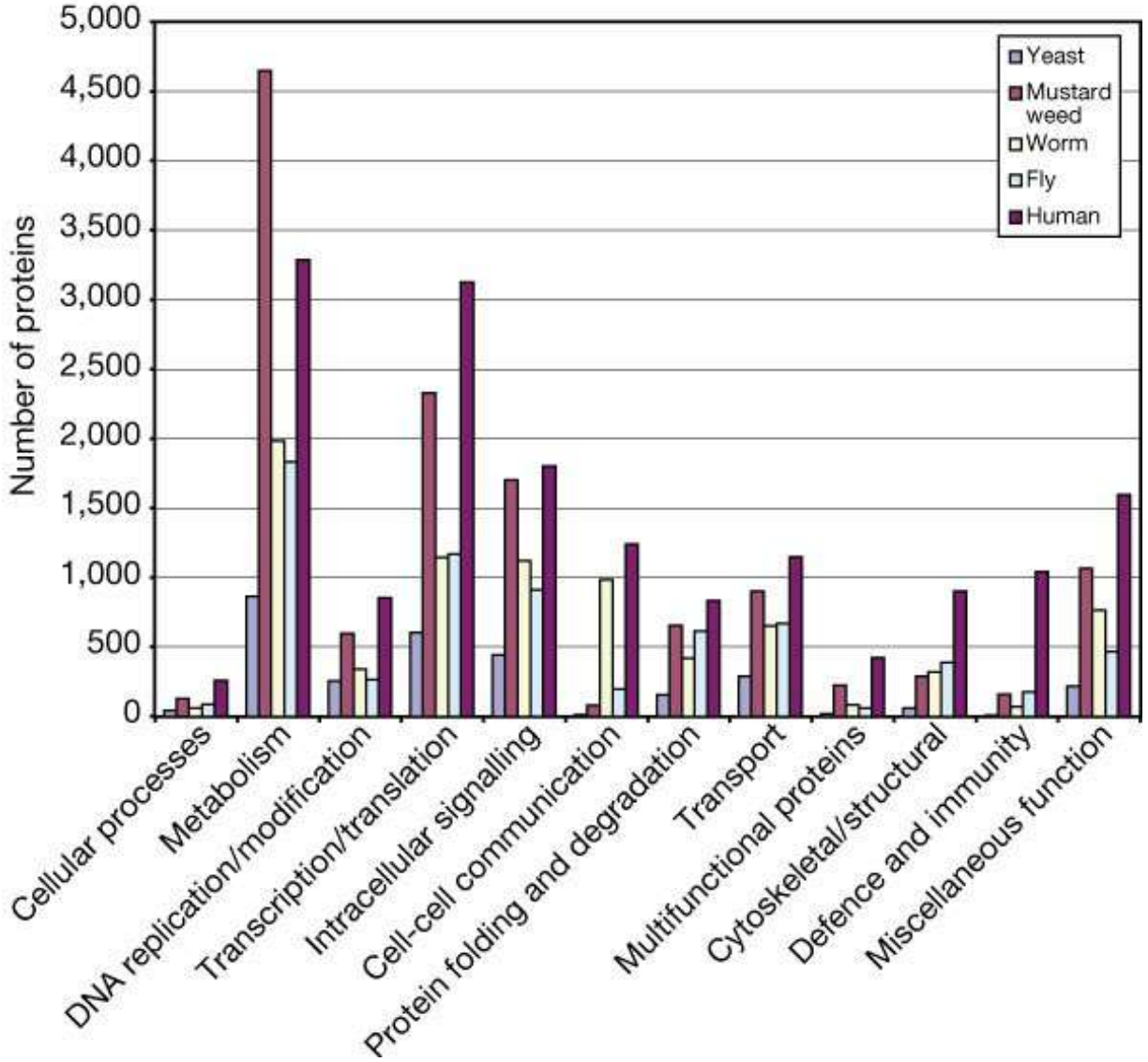
GC% GENOMA GENE



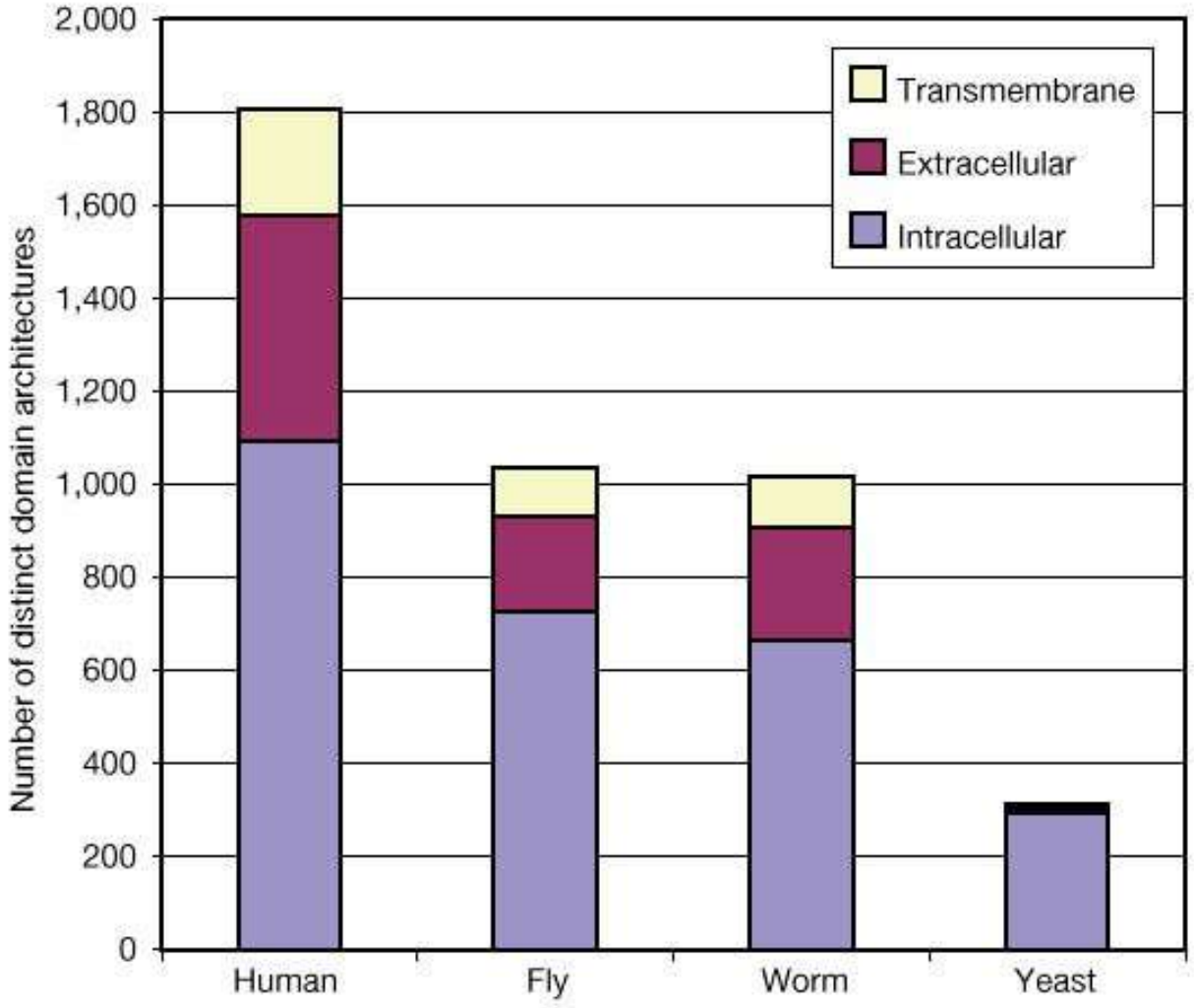
GC% EXON INTRON



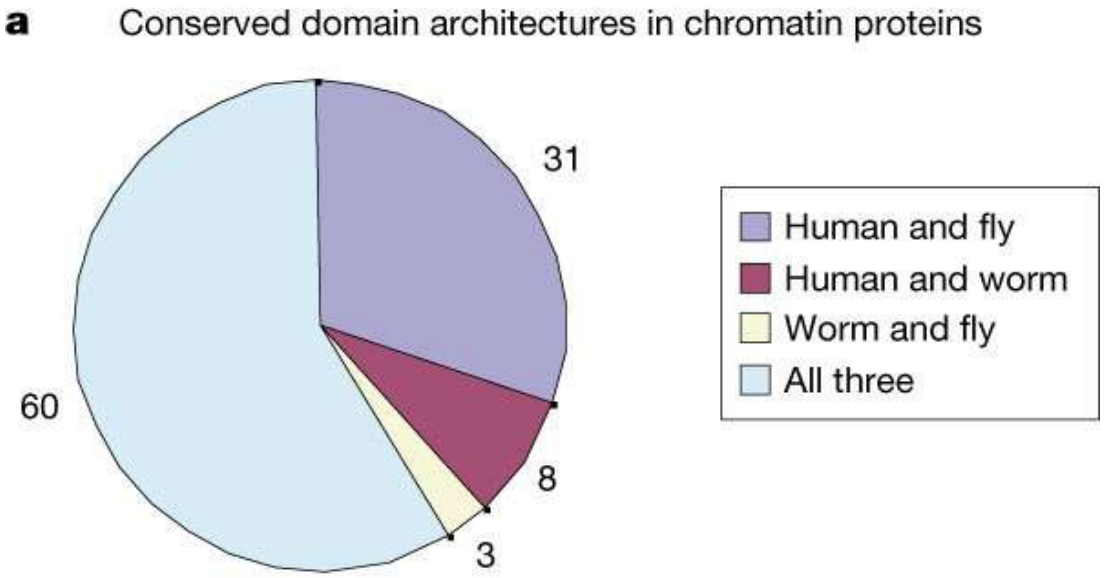
FUNÇÕES DAS PROTEÍNAS

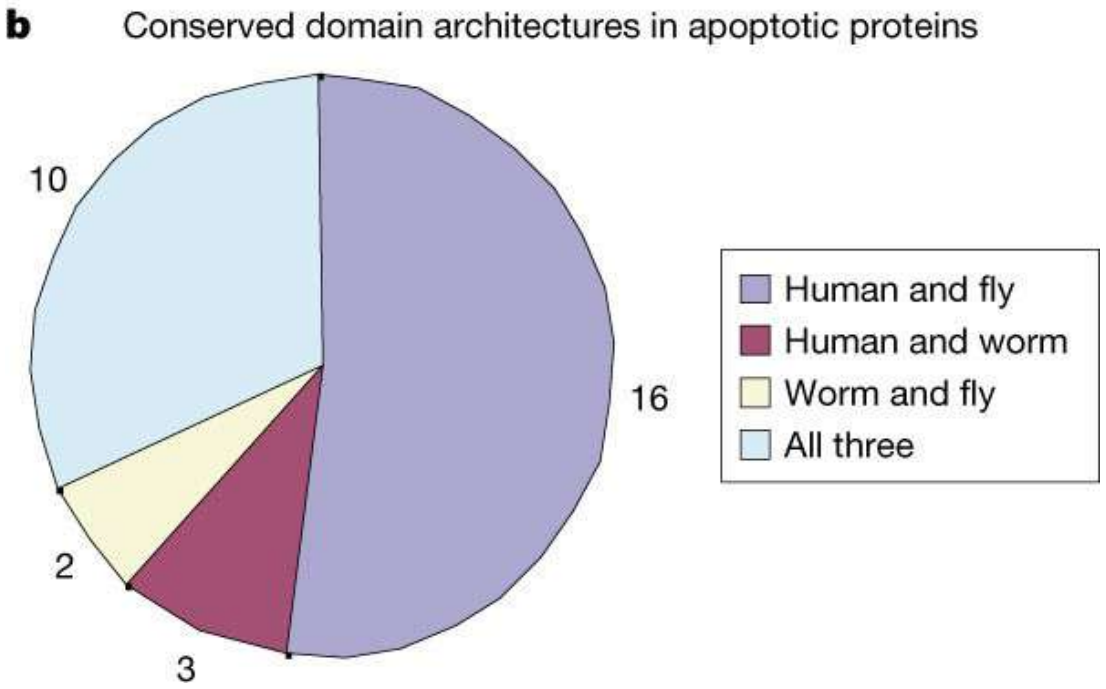


ARQUITETURA DE PROTEÍNAS



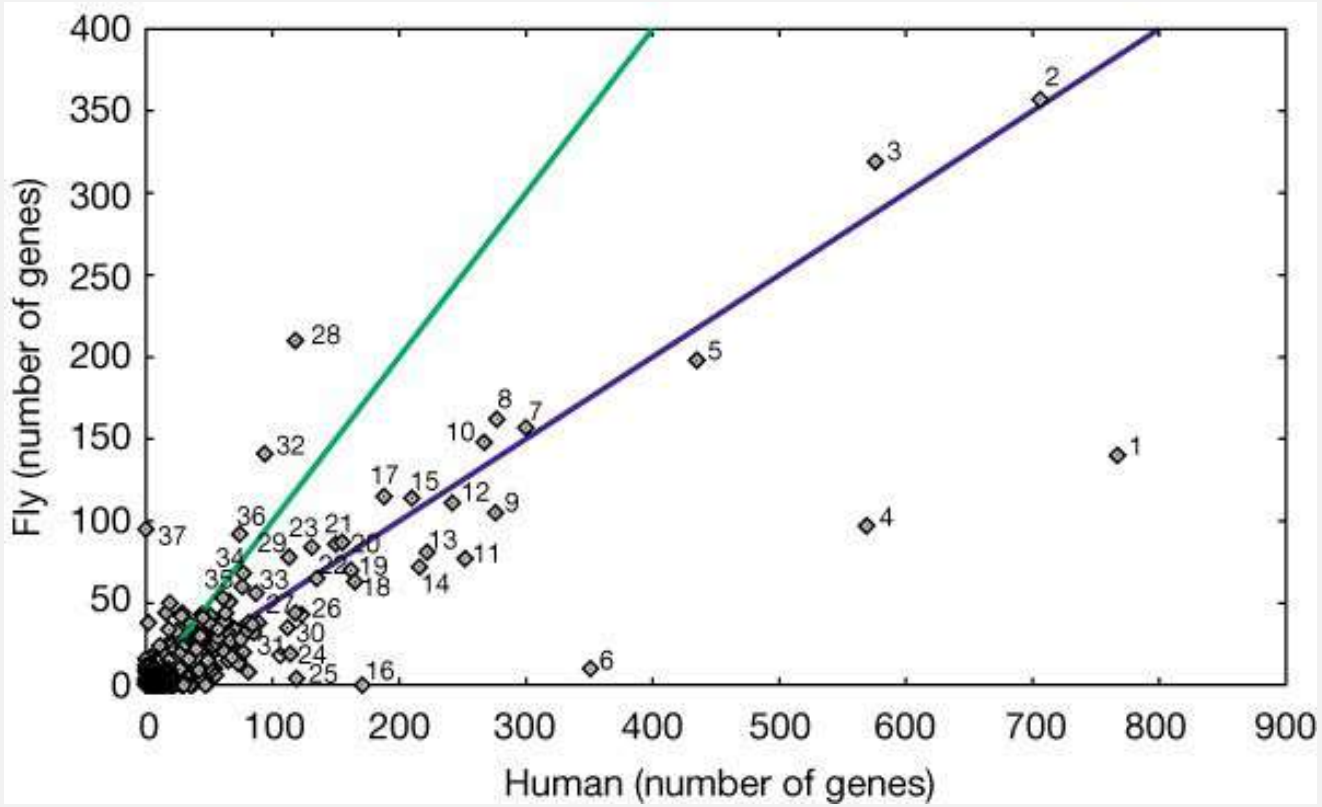
ARQUITETURA DE PROTEÍNAS



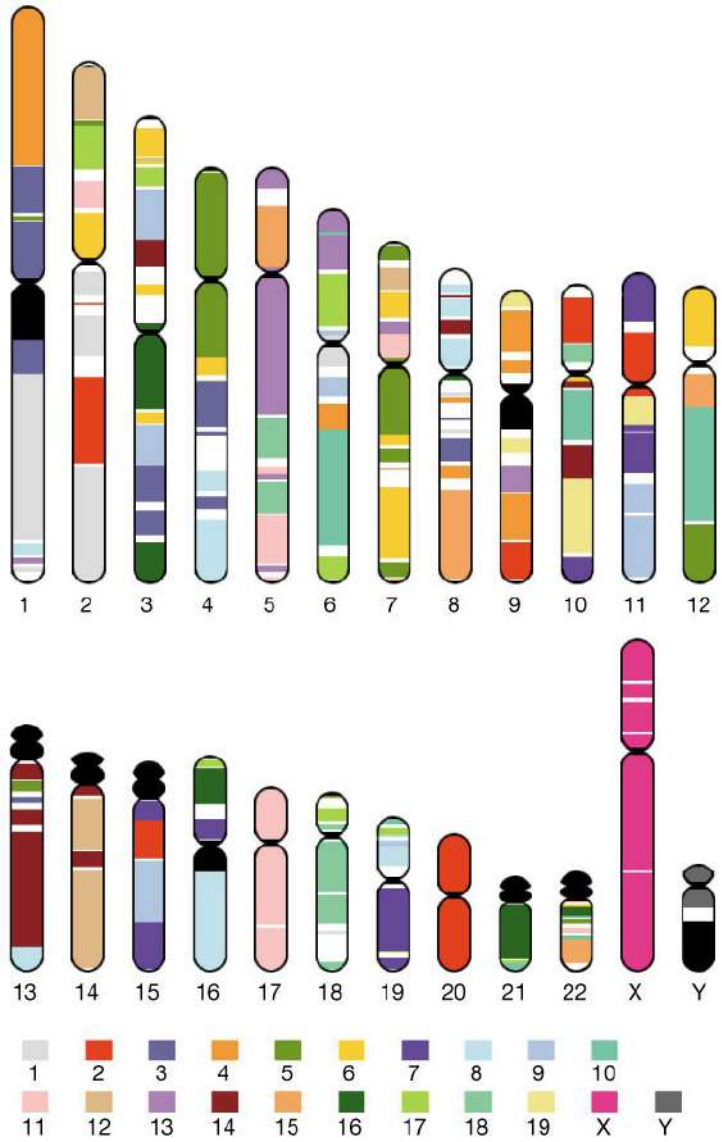


EXPANSÃO DAS FAMÍLIAS DE PROTEÍNAS

- (1) immunoglobulin domain
- (4) rhodopsin-like GPCR superfamily [IPR000276]
- (6) reverse transcriptase (RNA-dependent DNA polymerase)
- (28) serine proteases, trypsin family [IPR001254]



CONSERVAÇÃO DE GENOMAS HUMANO X MICE

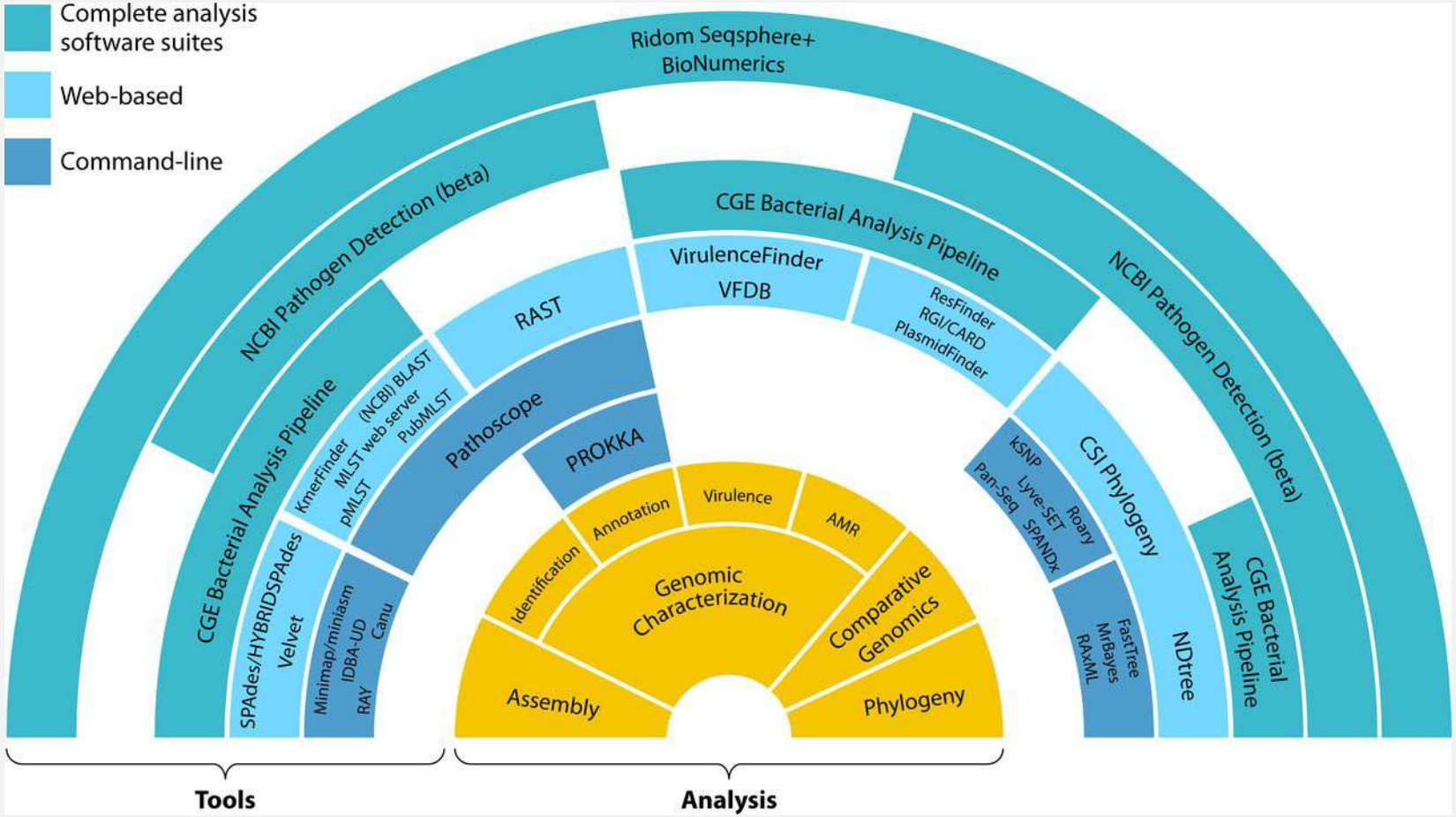


MONTAGEM DE GENOMA

Prof. Leandro Martins de Freitas, Phd

IMS/UFBA

WGS outbreak analysis tools.



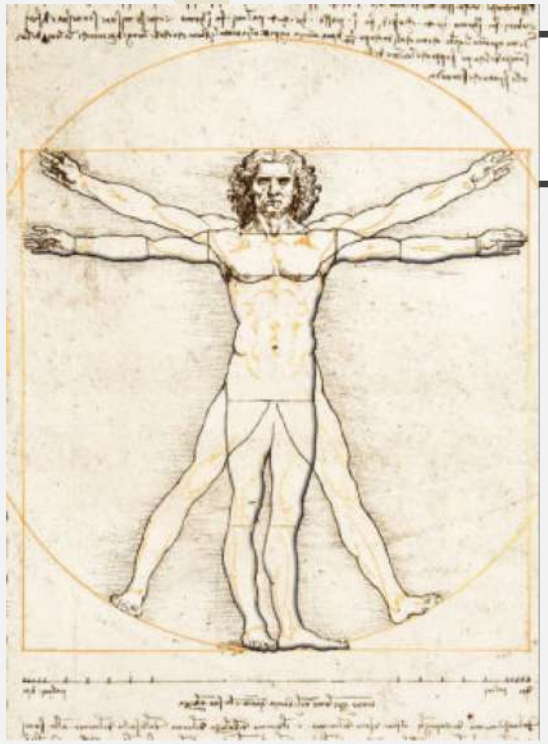
Scott Quainoo et al. Clin. Microbiol. Rev. 2017;
doi:10.1128/CMR.00016-17

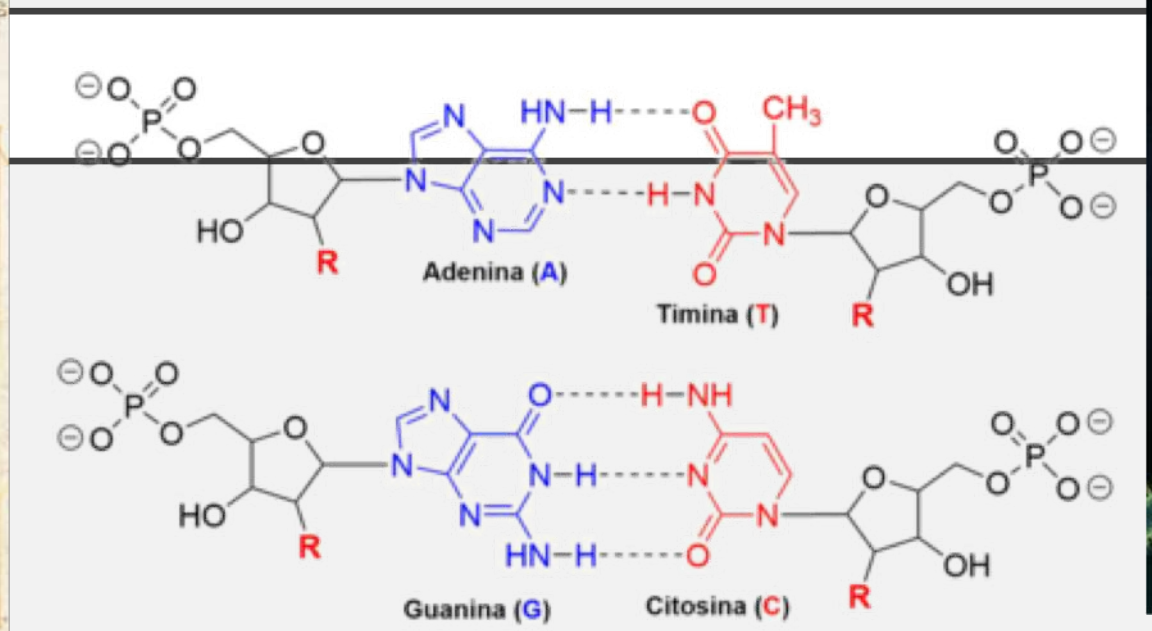
GENOMA

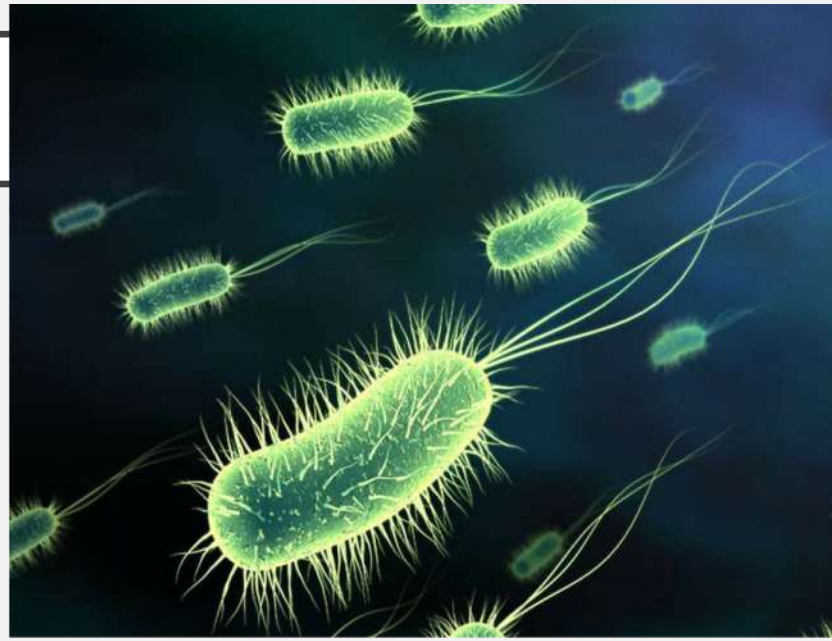
Genoma é todo o conteúdo de DNA presente em um organimo

- Todos os cromossomos
- Todos os genes
- Elementos regulatórios e estruturais

Genômica: estudo em larga escala de genes, elementos regulatórios e estruturais contidos na sequência de DNA de um organismo







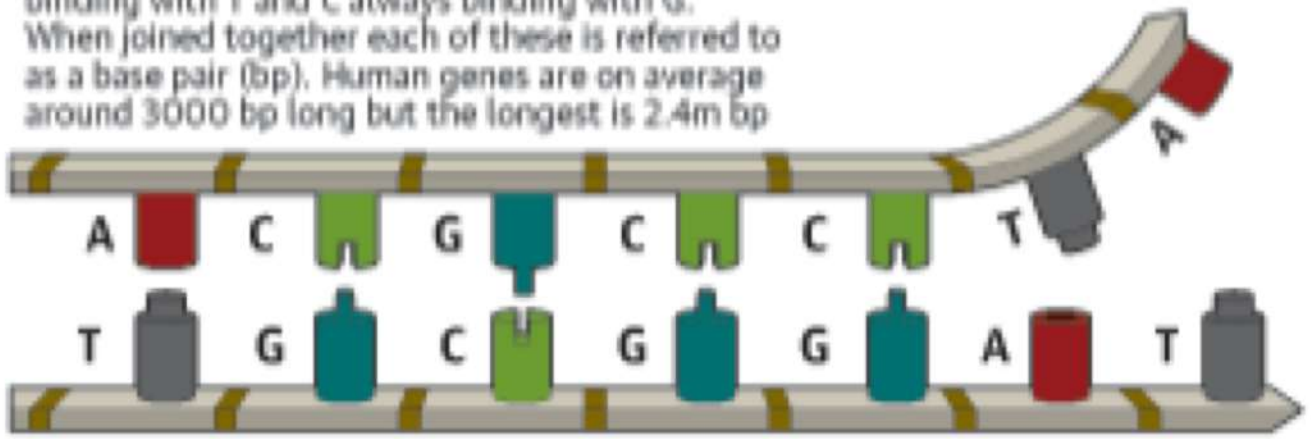
3 bilhões de bases

22 mil genes

3 milhões de bases

3 mil genes

The two DNA strands that make make up the double helix are joined by bonds between each of the bases. These complementary with A always binding with T and C always binding with G. When joined together each of these is referred to as a base pair (bp). Human genes are on average around 3000 bp long but the longest is 2.4m bp



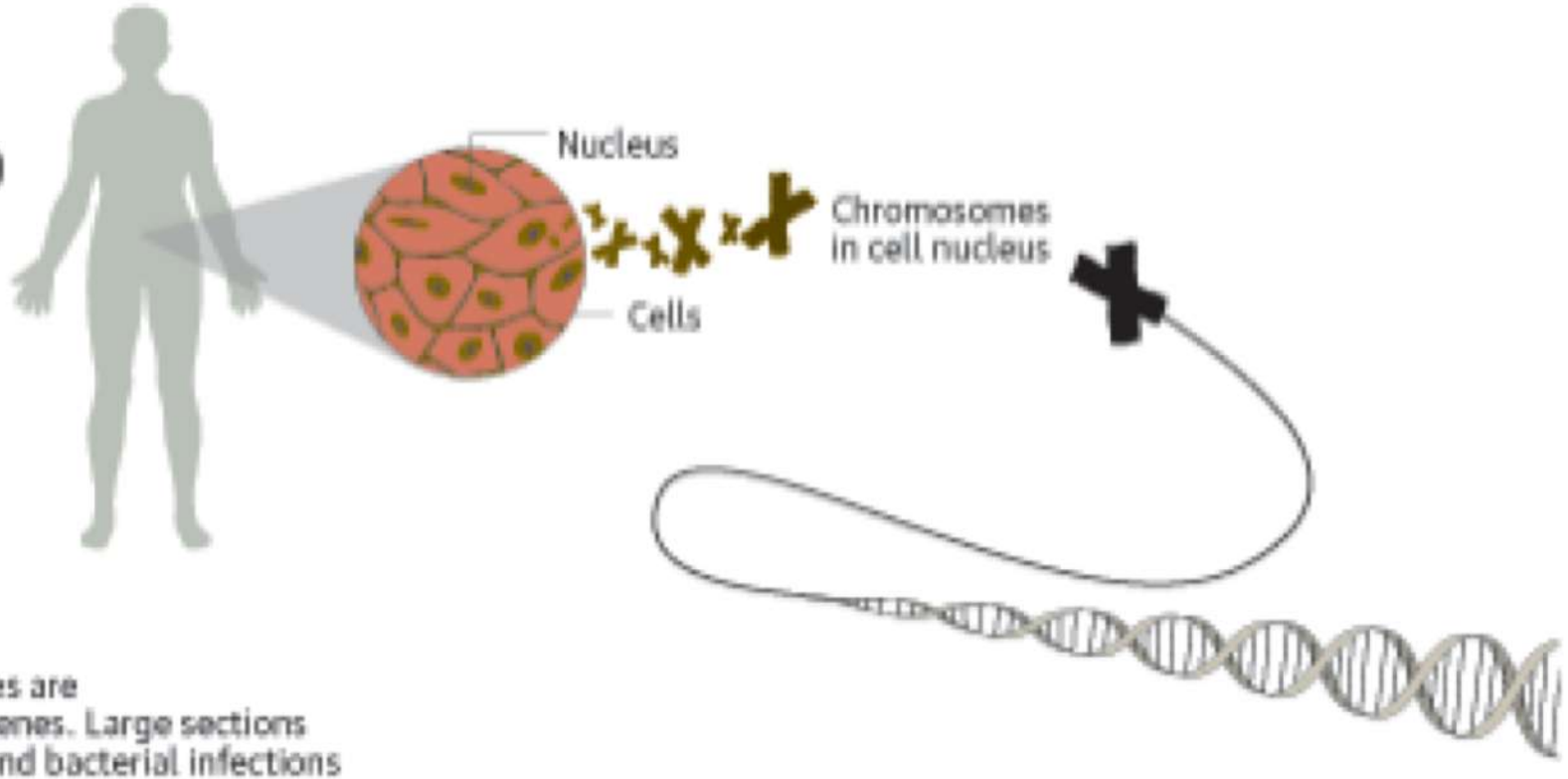
Different species vary greatly in the size of their genomes and the number of genes they have. There is a trend for increasing gene number with increasing

CHARTING THE HUMAN GENOME MAP

All humans share

99.8%

of their genetic material



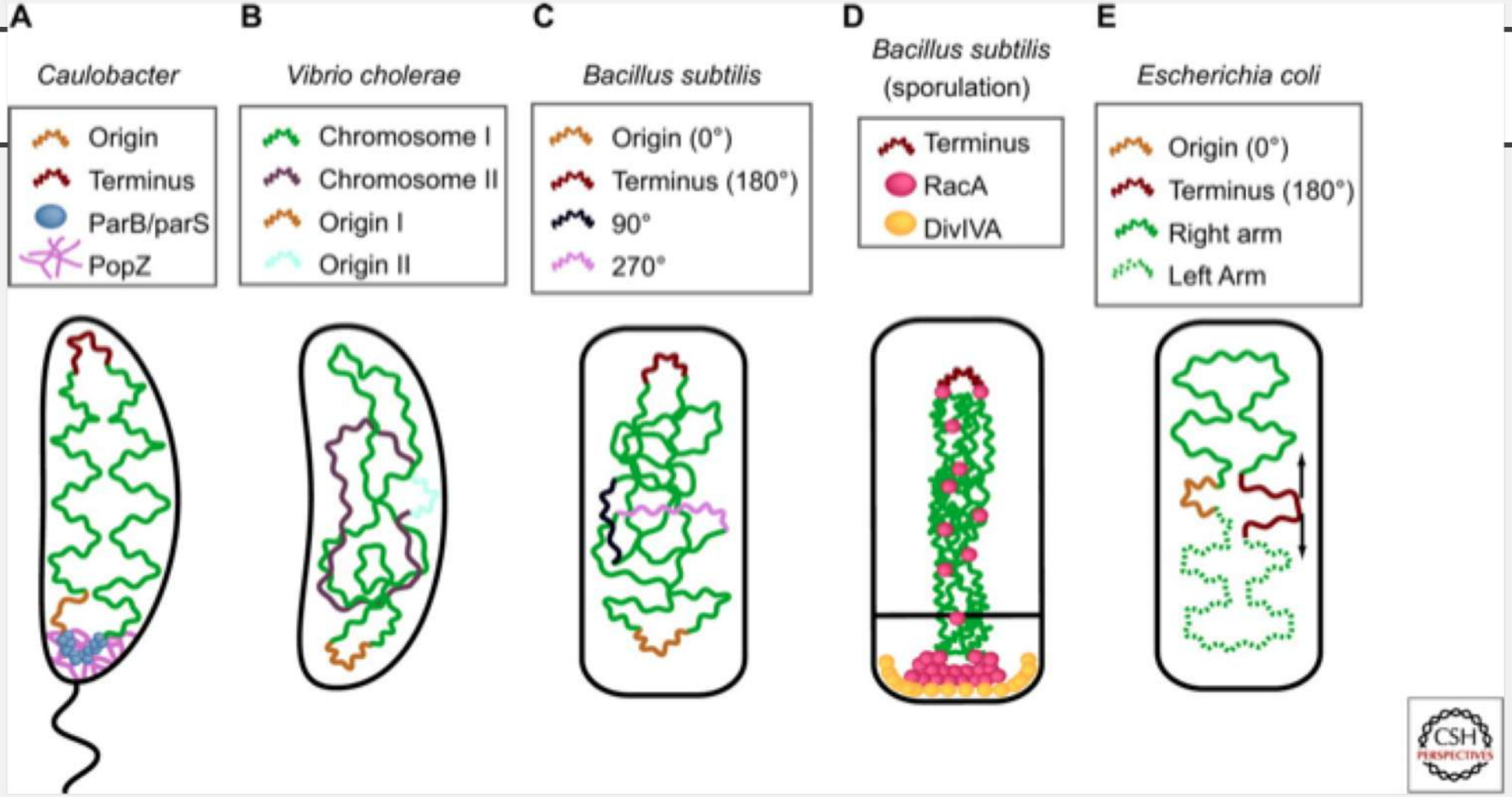
98%

of human chromosomes are "junk DNA" between genes. Large sections are remnants of viral and bacterial infections

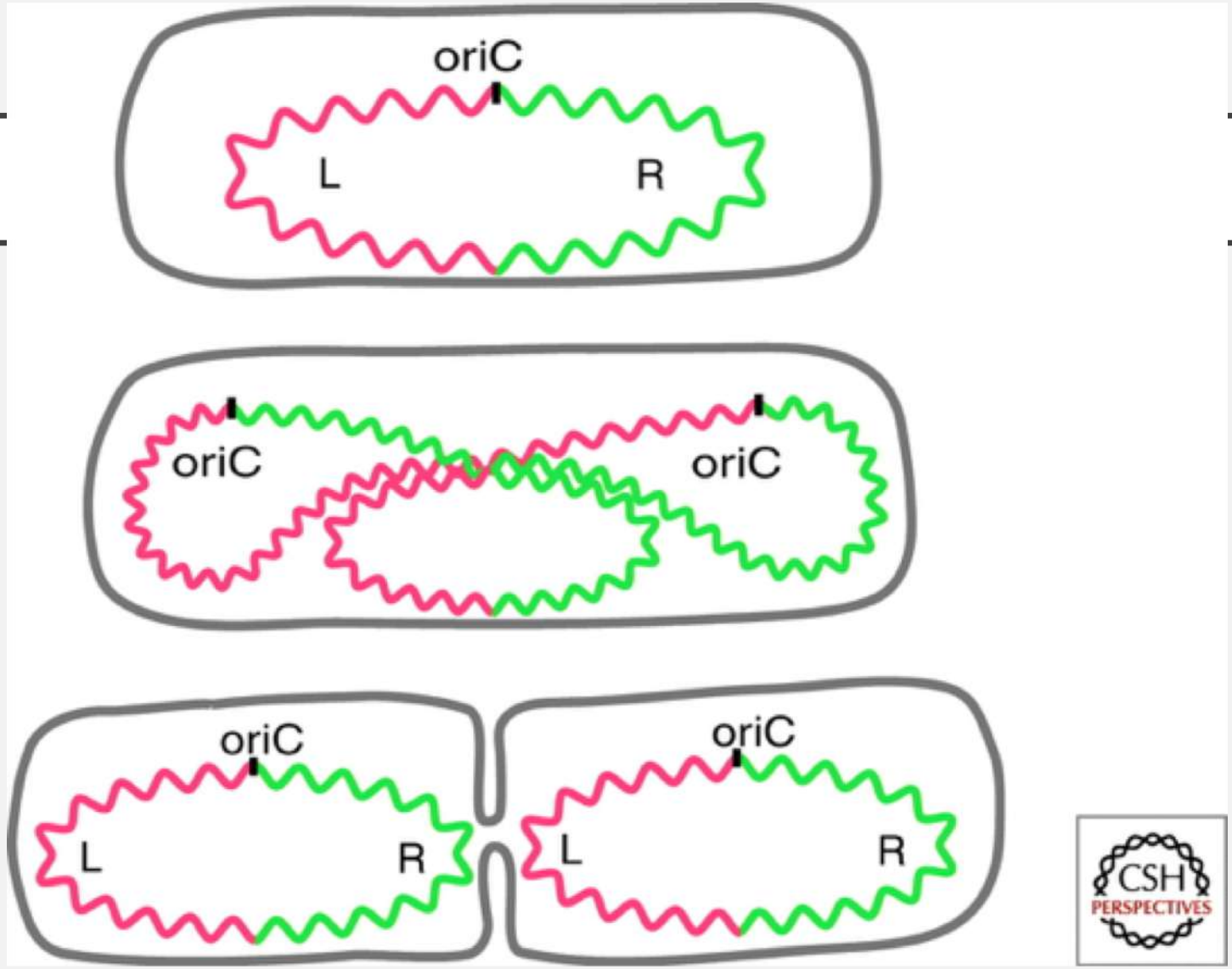
Different species vary greatly in the size of their genomes and the number of genes they have. There is a trend for increasing gene number with increasing body complexity, but there are numerous exceptions. Some species such as the mountain grasshopper, with 16.5 bn base pairs, have much larger genomes than us.

	Human	Thale cress	Earthworm	Fruit fly	Yeast
Genes	**20-25,000**	**26,000**	**18,000**	**13,600**	**6,000**
Base pairs	c.3 billion	117m	91m	116m	12m
Complexity	12 genes per million base pairs	221	197	117	500

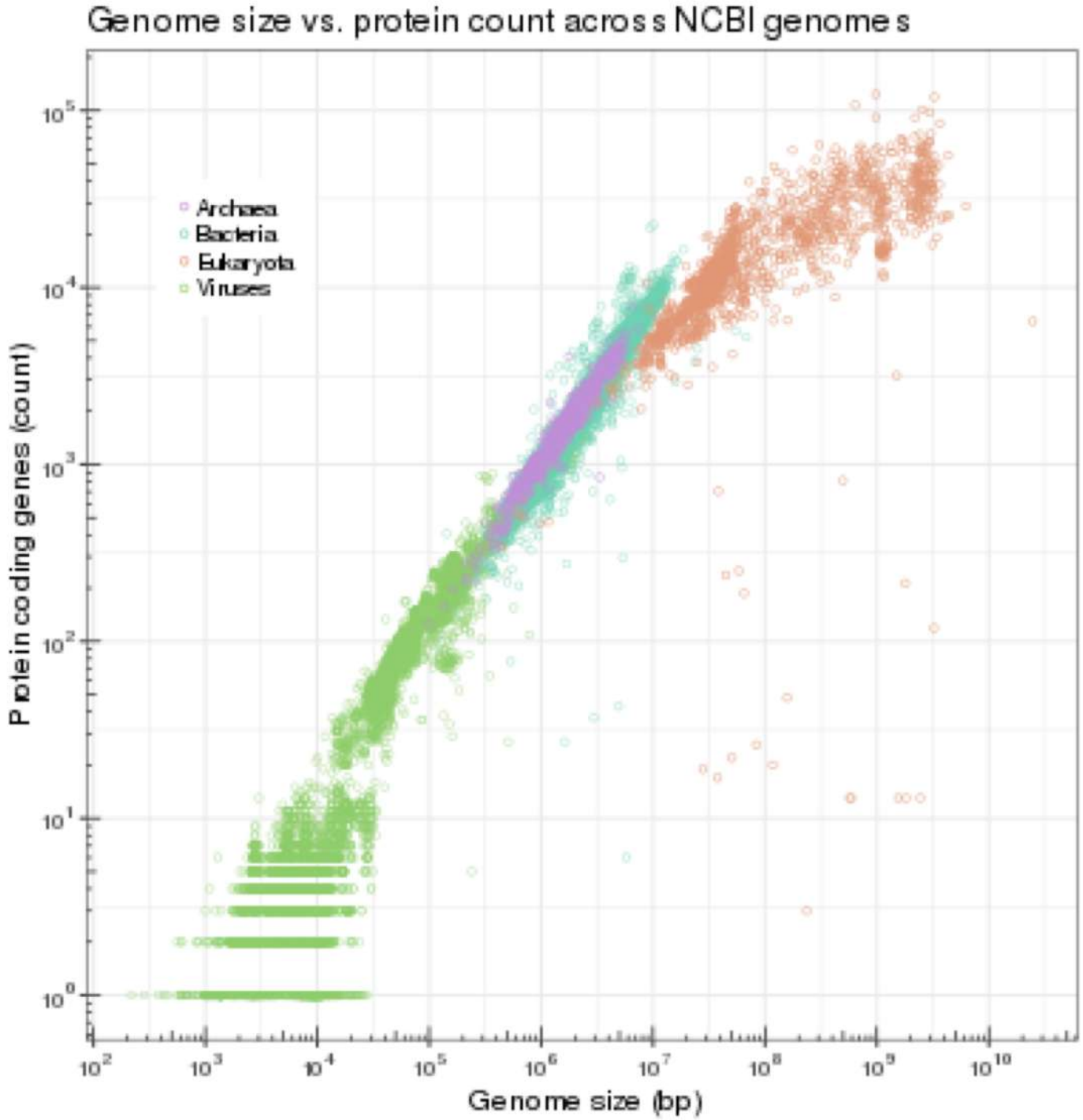
Chromosome organization in model bacteria.

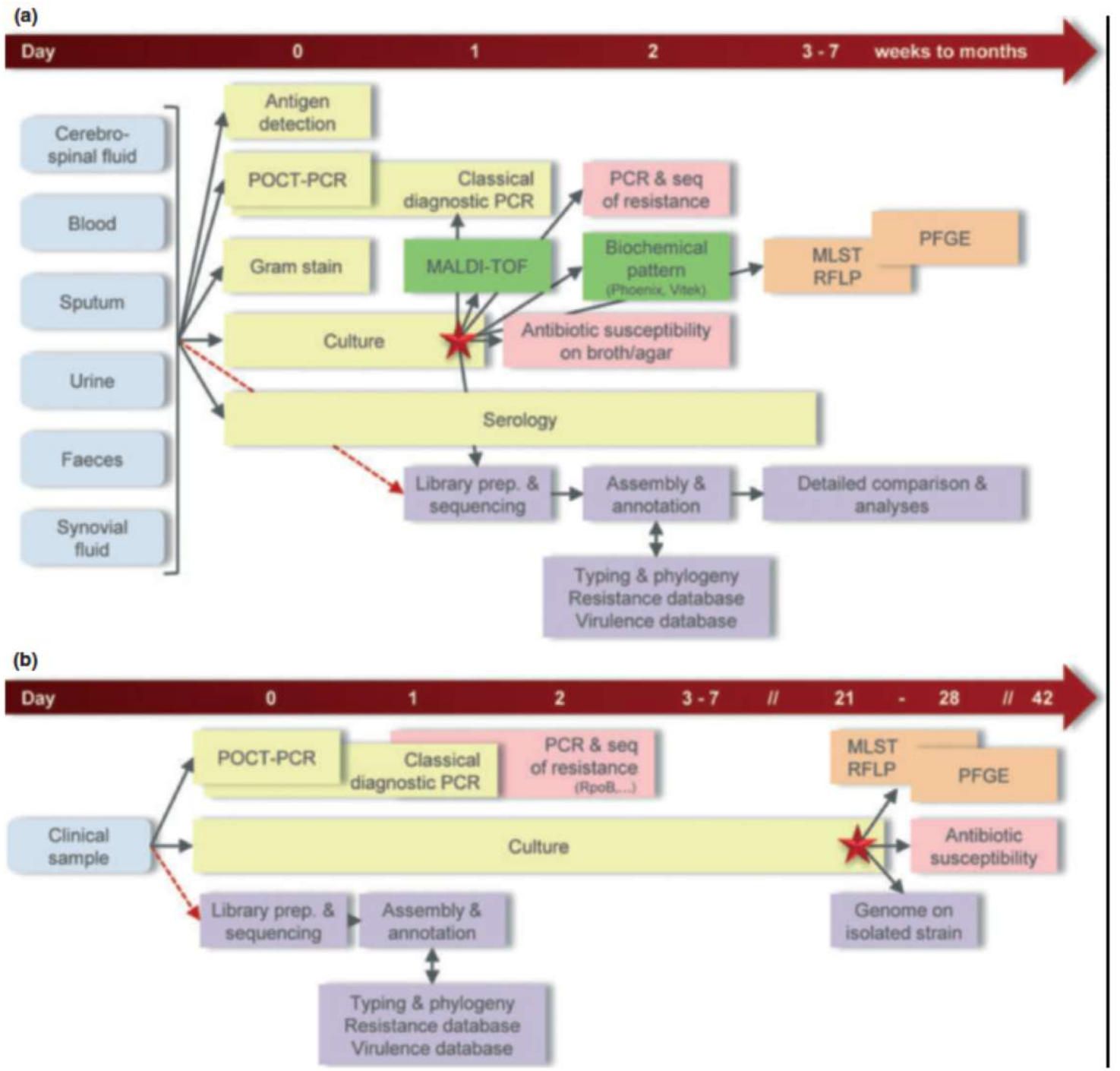


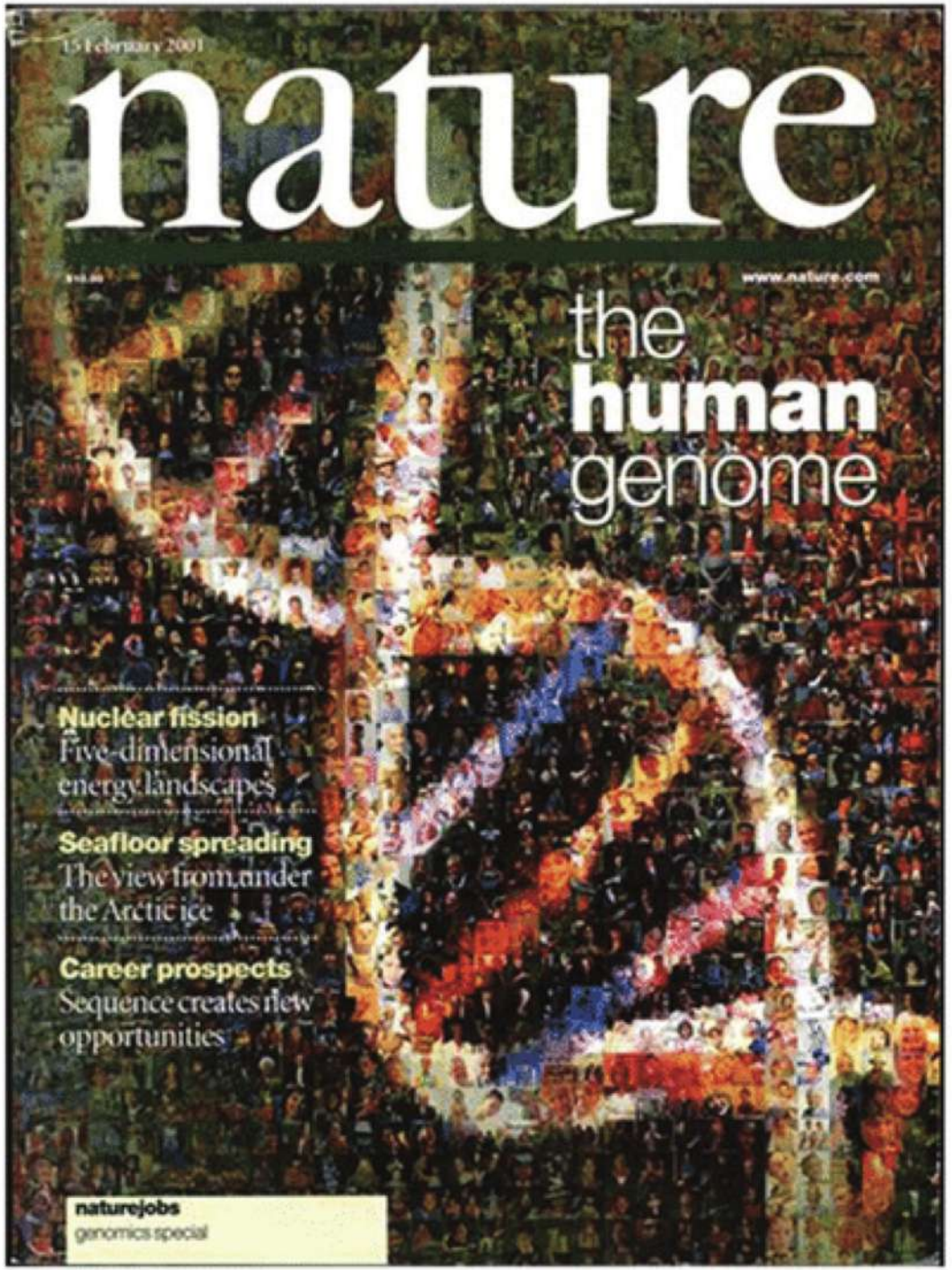
Chromosome organization in *E. coli*



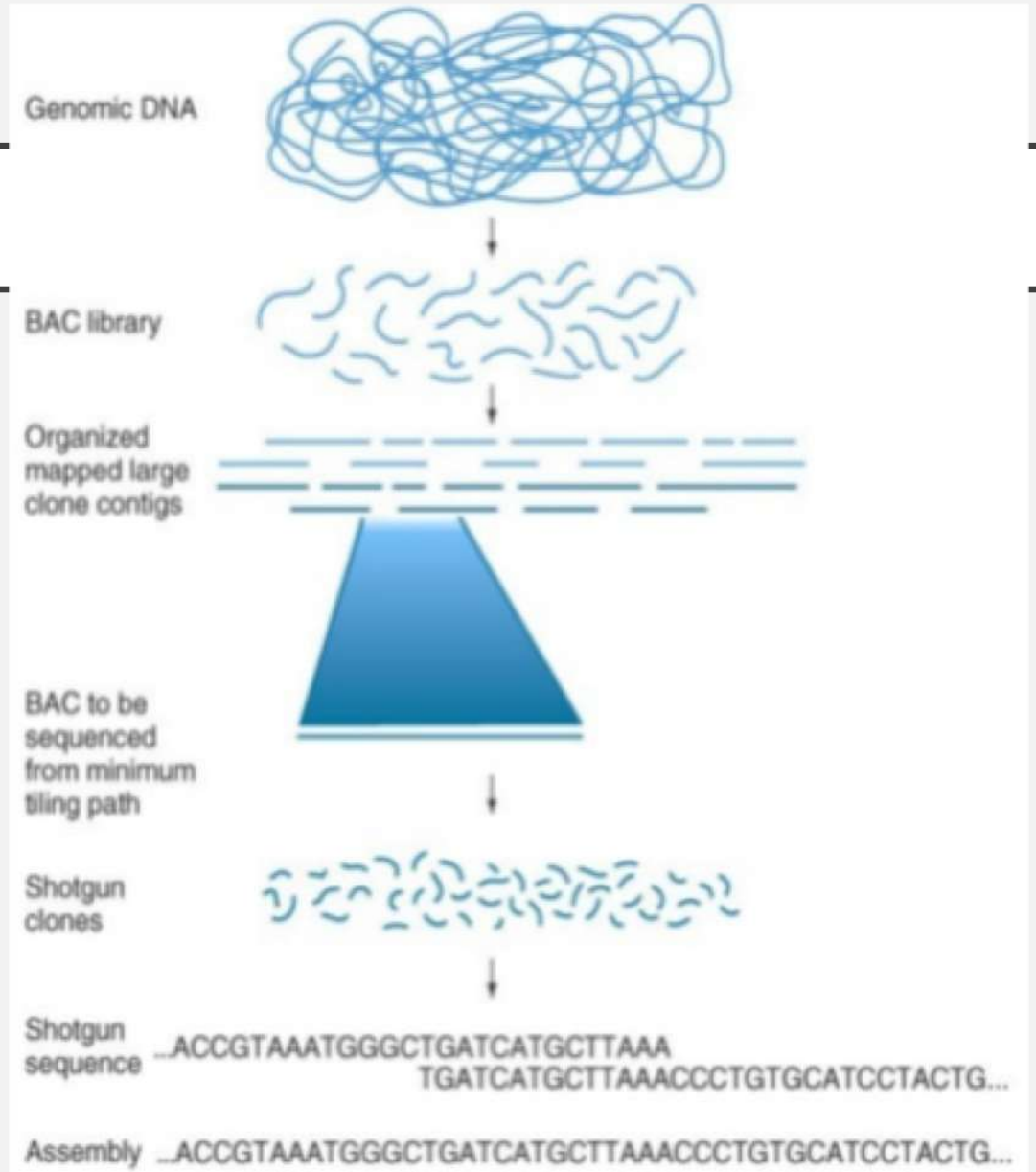
Tamanho do genoma vs número de proteínas

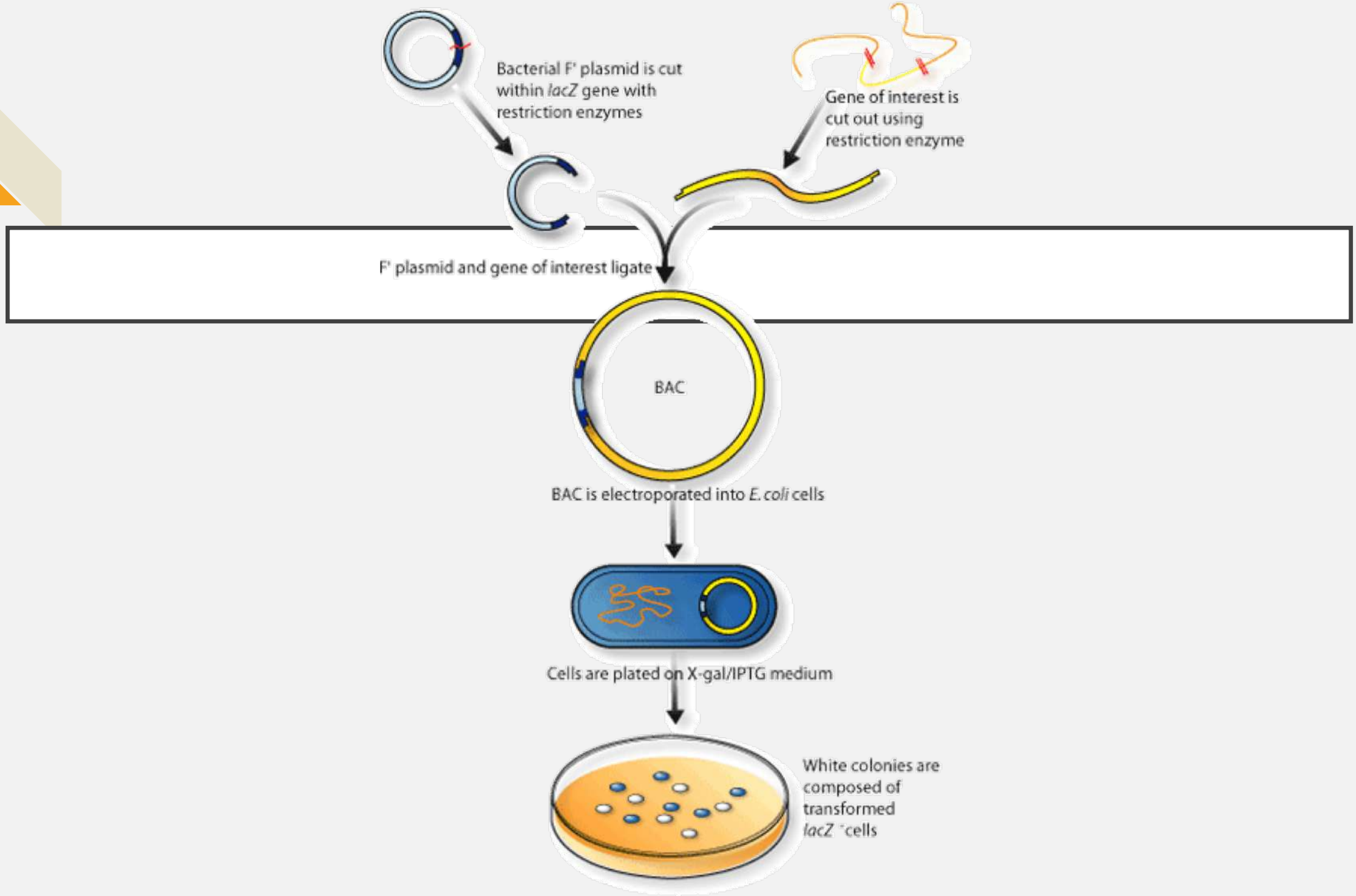


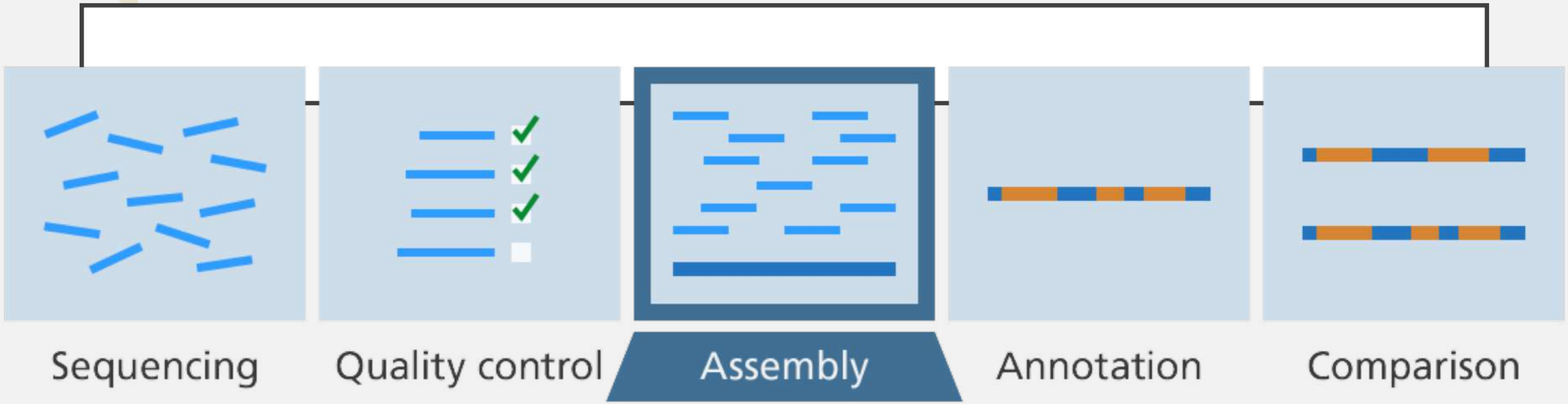

(a)
Day
0
1
2
3 - 7
weeks to months
Cerebro-spinal fluid
Blood
Sputum
Urine
Faeces
Synovial fluid
Antigen detection
POCT-PCR
Classical diagnostic PCR
PCR & seq of resistance
Gram stain
MALDI-TOF
Biochemical pattern
(Phoenix, Vitek)
MLST
RFLP
PFGE
Culture
Antibiotic susceptibility on broth/agar
Serology
Library prep. & sequencing
Assembly & annotation
Detailed comparison & analyses
Typing & phylogeny
Resistance database
Virulence database
(b)
Day
0
1
2
3 - 7
//
21
-
28
//
42
POCT-PCR
Classical diagnostic PCR
PCR & seq of resistance
(RpoB,...)
MLST
RFLP
PFGE
Clinical sample
Culture
Antibiotic susceptibility
Library prep. & sequencing
Assembly & annotation
Genome on isolated strain
Typing & phylogeny
Resistance database
Virulence database


15 February 2001
nature
www.nature.com
the human genome
Nuclear fission
Five-dimensional energy landscapes
Seafloor spreading
The view from under the Arctic ice
Career prospects
Sequence creates new opportunities
naturejobs
genomics special


Science
THE HUMAN GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE


Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced from minimum tiling path
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...


Bacterial F' plasmid is cut within *lacZ* gene with restriction enzymes
Gene of interest is cut out using restriction enzyme
F' plasmid and gene of interest ligate
BAC
BAC is electroporated into *E. coli* cells
Cells are plated on X-gal/IPTG medium
White colonies are composed of transformed *lacZ* cells


Sequencing
Quality control
Assembly
Annotation
Comparison

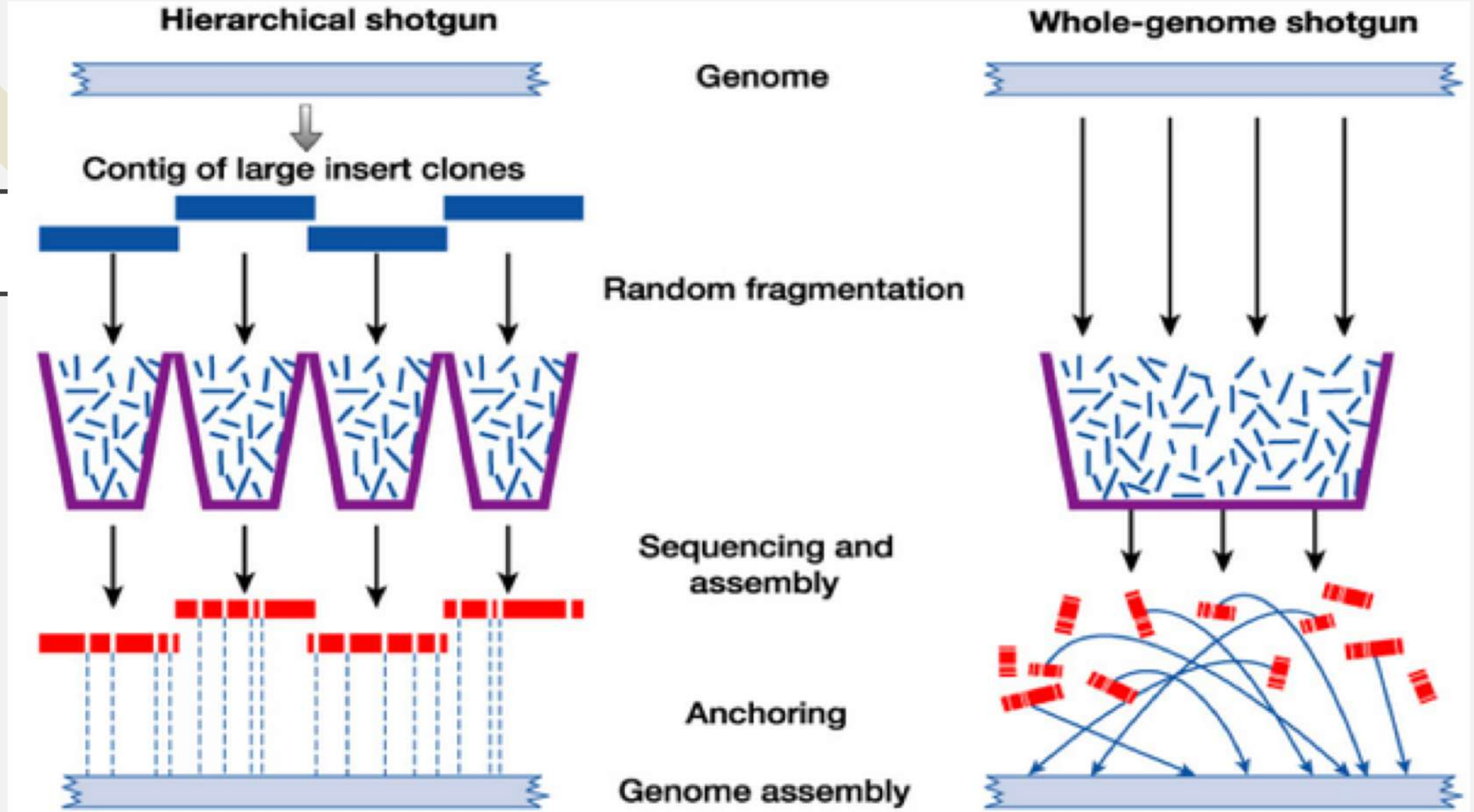


Figure 8-3 Human Molecular Genetics, 3/e. (© Garland Science 2004)

A single recipe is like…

A Gene:
One set of instructions for how to make one protein.

A recipe book is like…

A Chromosome:
Thousands of sets of instructions for how to make thousands of proteins

Two copies of 23 recipe books is like…

Our Genome:
ALL of the sets of instructions for how to make ALL the proteins we need

ALL (gene, chromosome, genome) are written in same the DNA alphabet!!!

In order for scientists to find a specific gene, they need to know which "book" to look in, and which "page" to turn to.

Genome

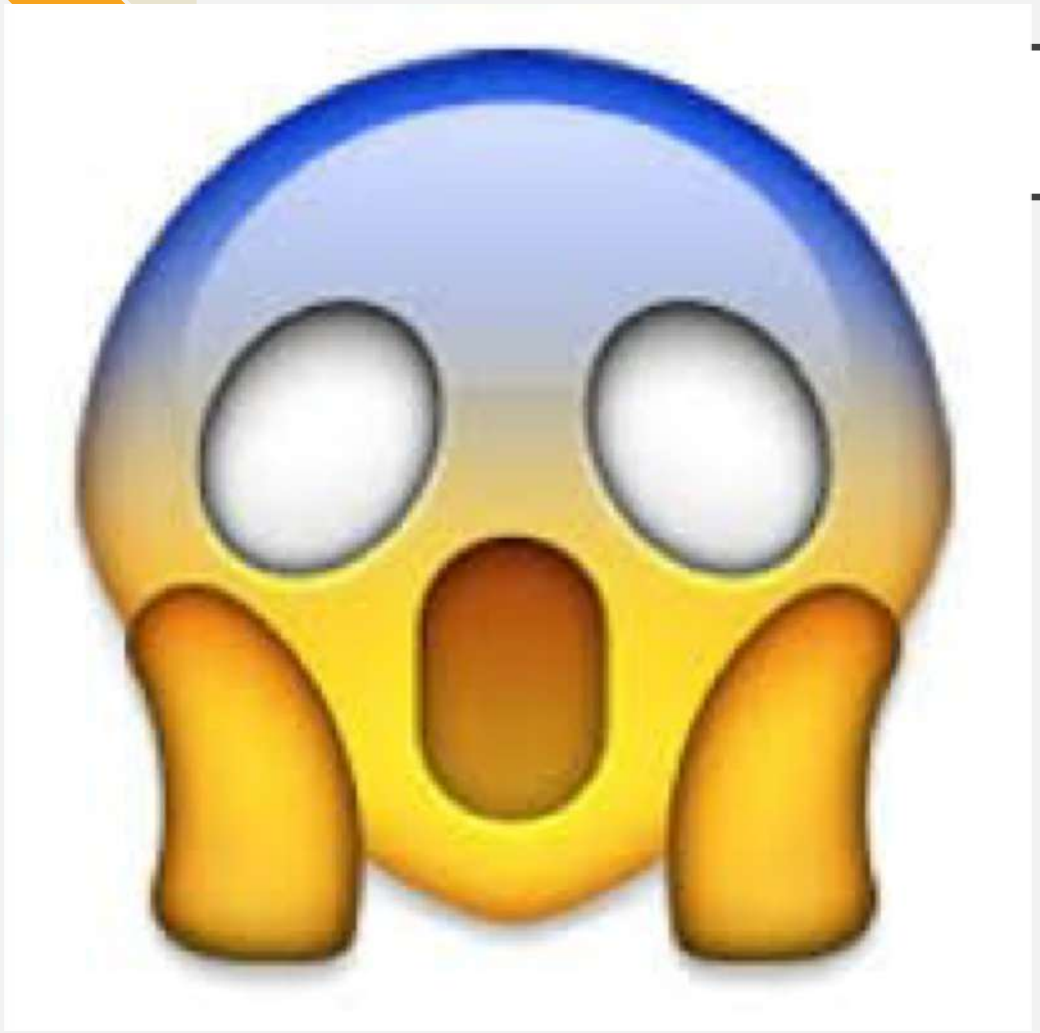
Chromosome

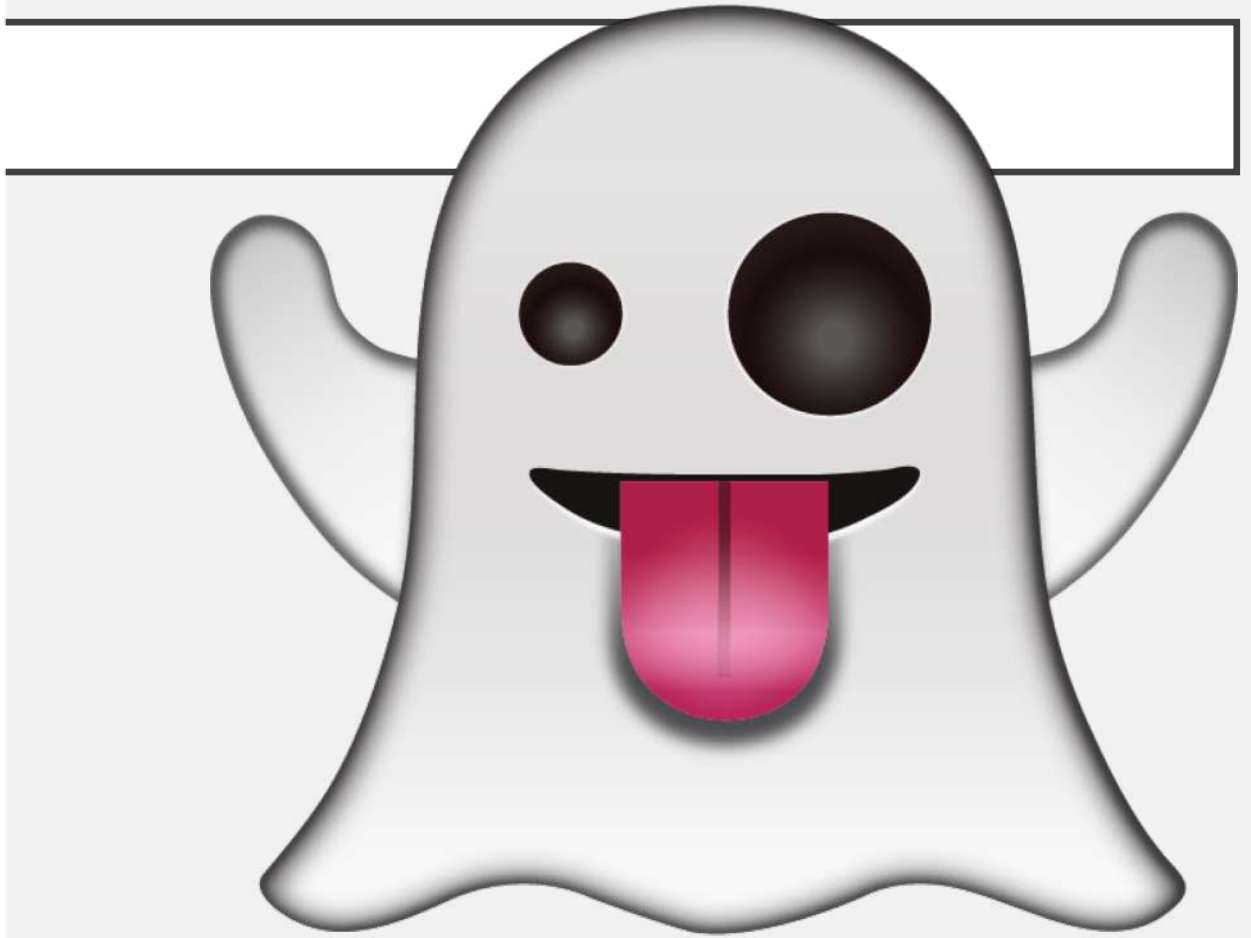
Gene

MONTAGEM DE UM GIGANTESO QUEBRA-CABEÇA

GIGANTESCO QUEBRA-CABEÇA SEM A CAIXA


Weihnachts-Geld



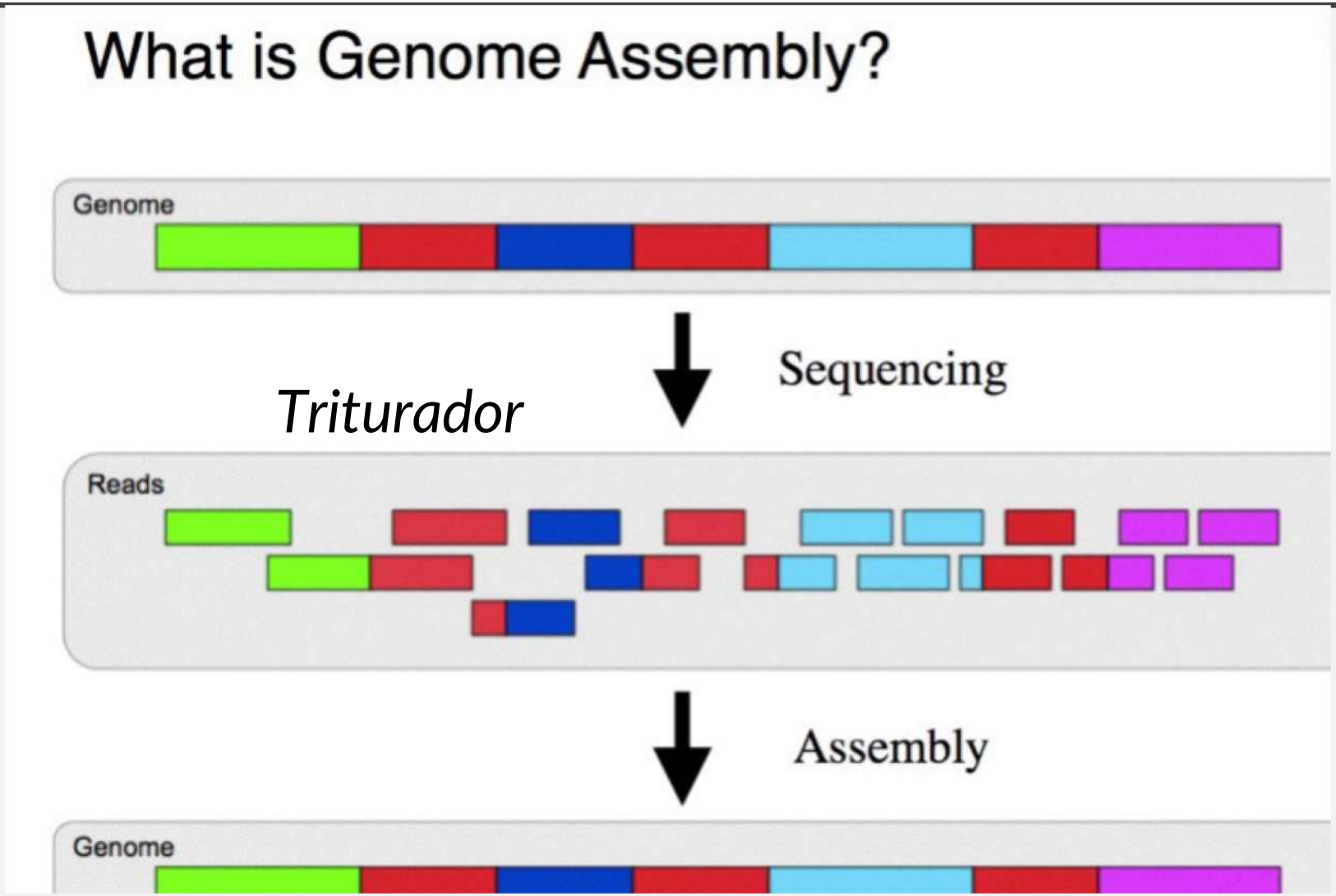


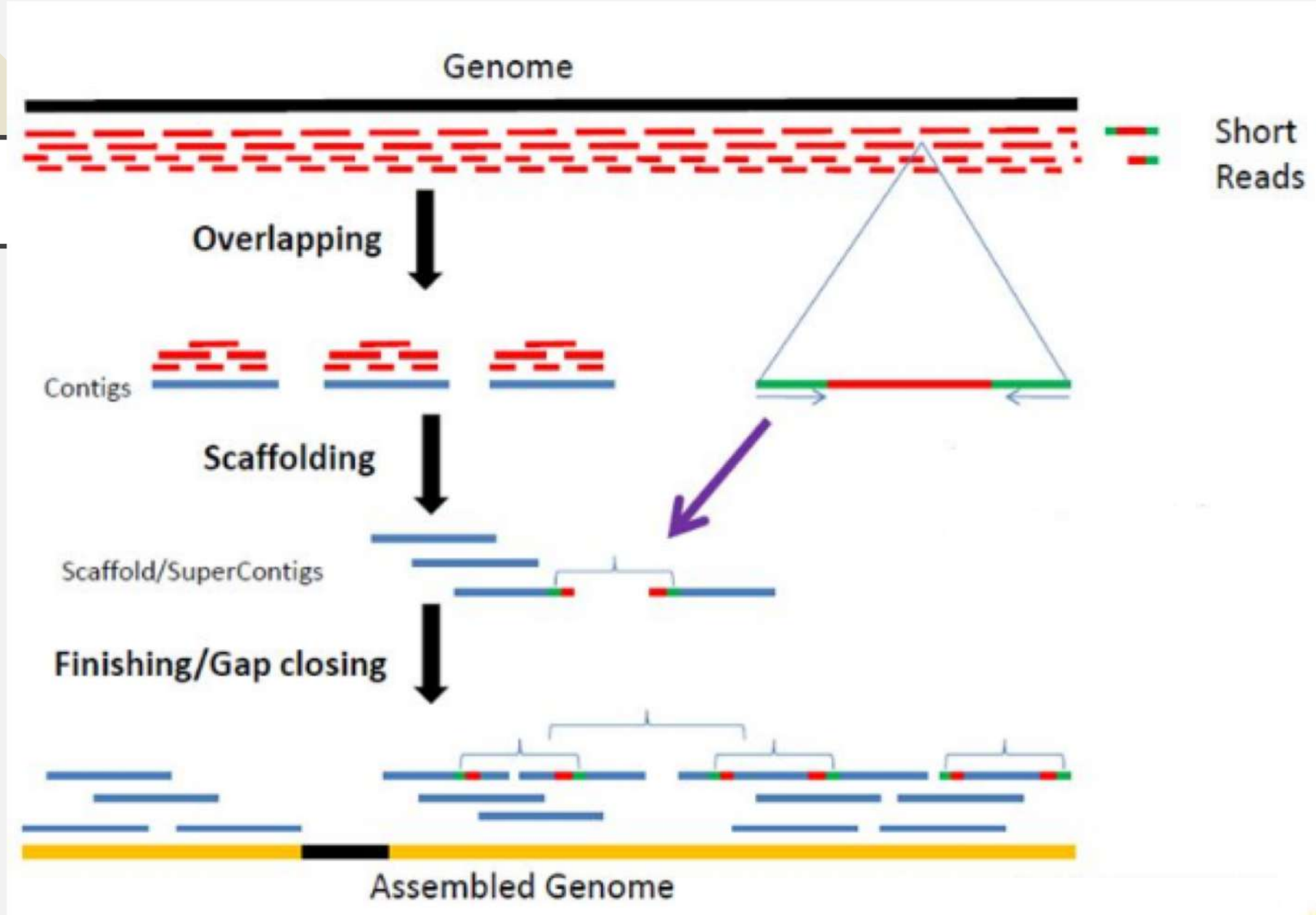
MONTAGEM *DE NOVO*

Método computacional para montar fragmentos de sequências de DNA - usando sobreposição das extremidades

Leituras obtidas de métodos de sequenciamento para formar sequências de DNA mais longas

MONTAGEM DE GENOMA - BÁSICO



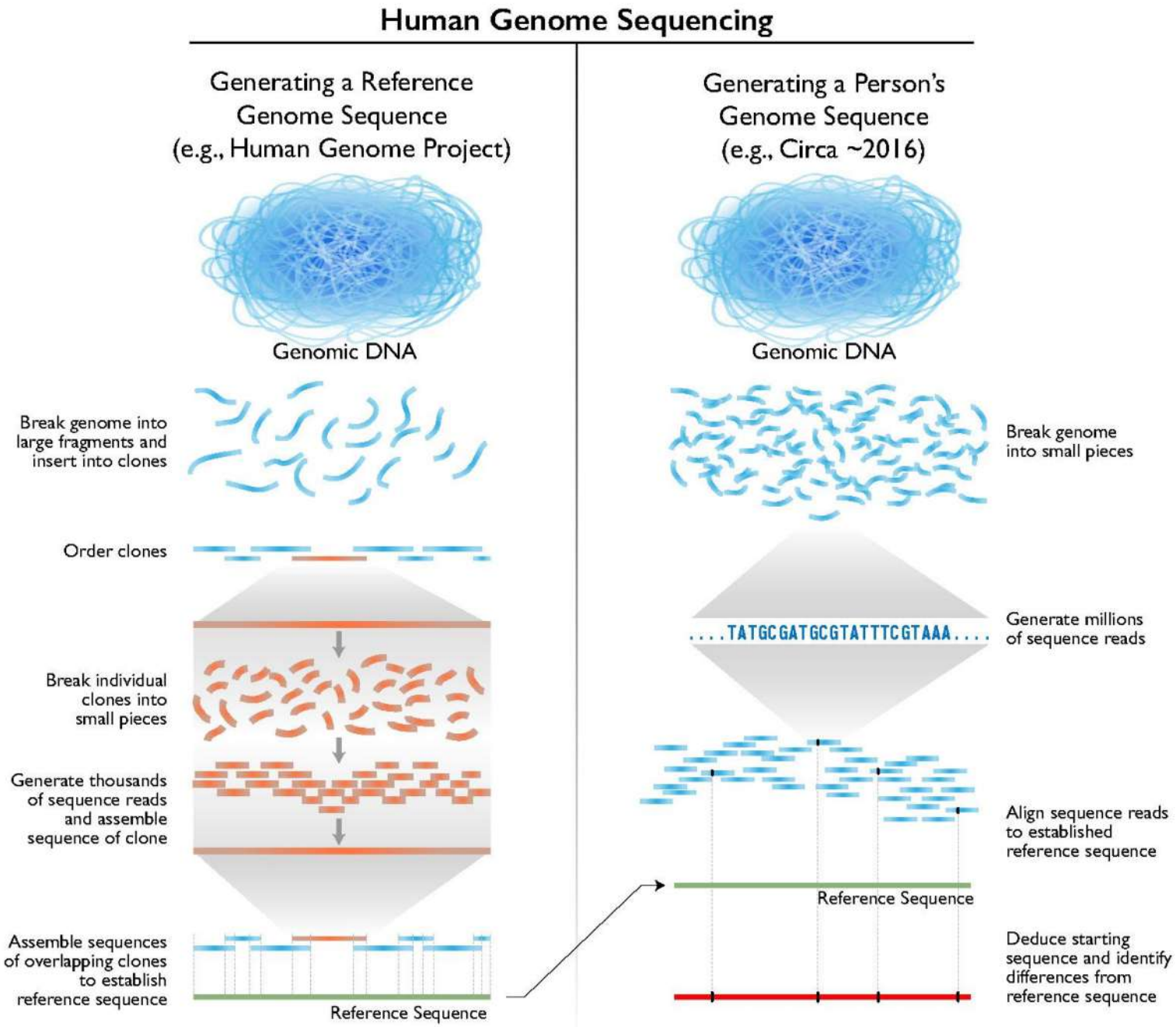

Genome
Short
Reads
Overlapping
Contigs
Scaffolding
Scaffold/SuperContigs
Finishing/Gap closing
Assembled Genome

MONTAGEM GENOMA REFERÊNCIA

Montagem usando sequência referência existente

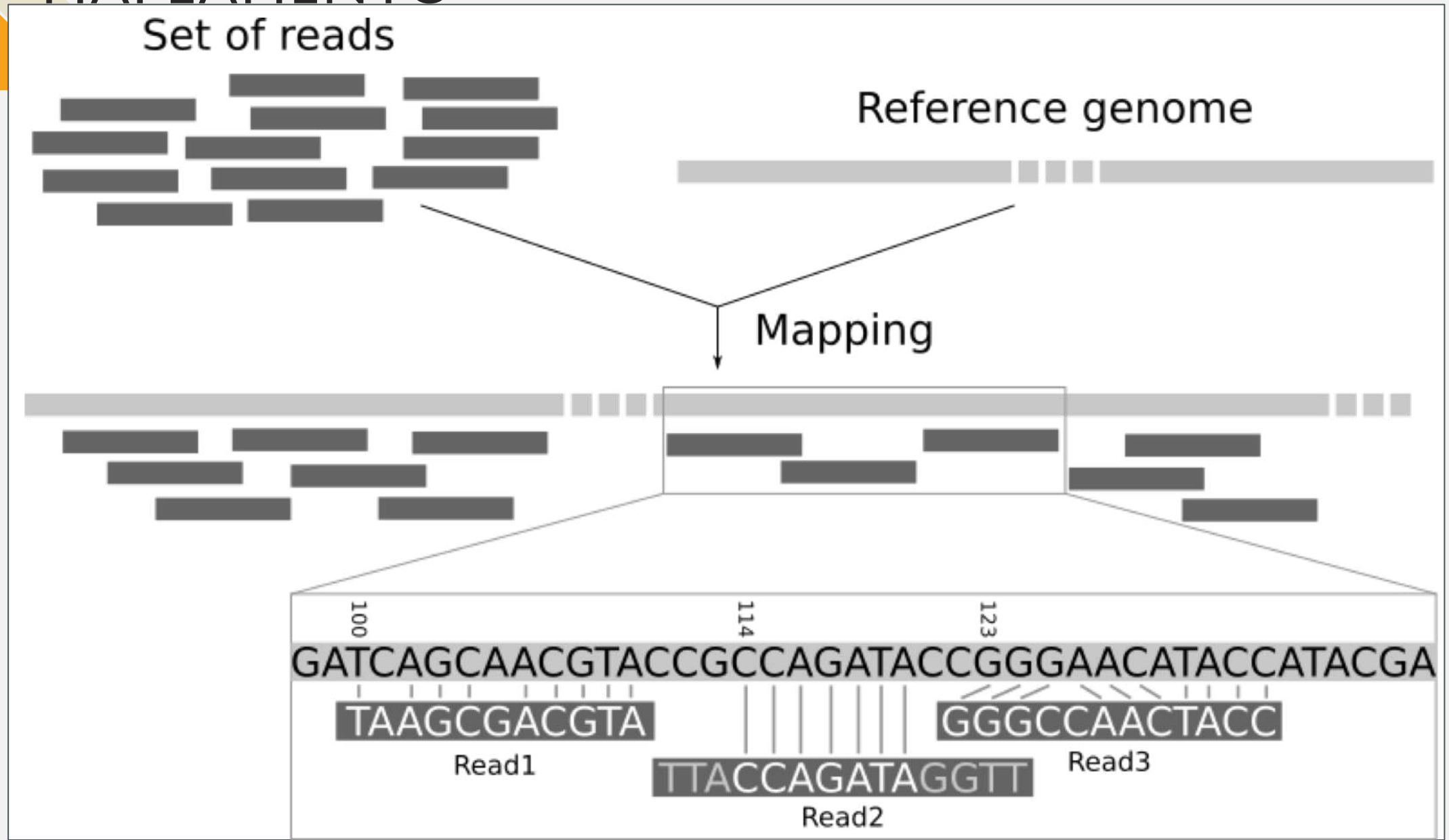
Montagem apenas constrói uma sequência semelhante a referência

Algoritmo de mapeamento

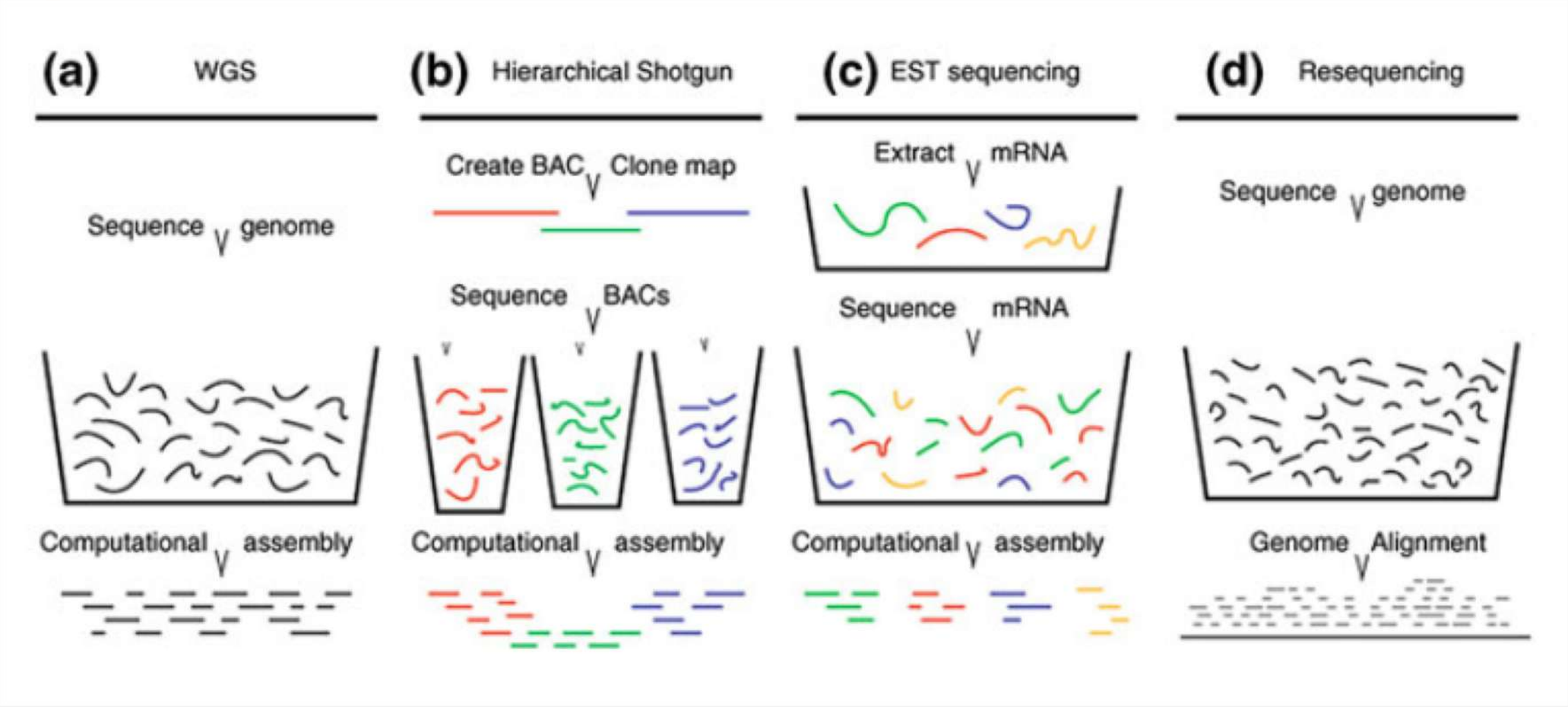

Human Genome Sequencing
Generating a Reference Genome Sequence (e.g., Human Genome Project)
Generating a Person's Genome Sequence (e.g., Circa ~2016)
Genomic DNA
Genomic DNA
Break genome into large fragments and insert into clones
Break genome into small pieces
Order clones
....TATGCGATGCGTATTTCGTAAA....
Generate millions of sequence reads
Break individual clones into small pieces
Generate thousands of sequence reads and assemble sequence of clone
Align sequence reads to established reference sequence
Reference Sequence
Assemble sequences of overlapping clones to establish reference sequence
Deduce starting sequence and identify differences from reference sequence
Reference Sequence

MONTAGEM GENOMA REFERÊNCIA

MONTAGEM DE GENOMA REFERÊNCIA - MAPEAMENTO



MÉTODOS DE SEQUENCIAMENTO E MONTAGEM



TAMANHO DE *READS*

TABLE 1. Main sequencing technologies and their characteristics

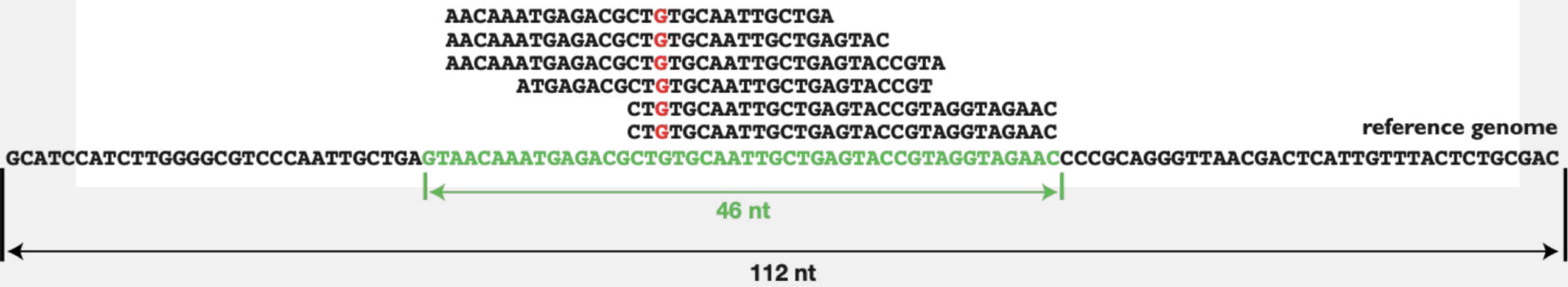
Technology (manufacturer)	Sequencing chemistry	Platform	Read length (bp)	Throughput (Gb/h run)	Best suited for:
Sanger	Dye terminator	ABI 3730xl	700–900		*De novo* and metagenomics
454 (Roche)	Pyrosequencing	GS FLX	400–700	0.04	*De novo* and metagenomics
		GS Junior	400	0.004	*De novo* and metagenomics
Solexa (Illumina)	Sequencing by synthesis with reversible terminators	GaIIx	36–150	0.3	Resequencing
		HiSeq2000	36–100	2.9	Resequencing
		MiSeq	36–250	0.2	Resequencing
SOLiD (ABI)	Sequencing by ligation	5500xl	35–75	1	Resequencing
Heliscope (Helicos)	Sequencing by synthesis with virtual terminators	tSMS	25–55	1	Resequencing
Ion Torrent (Life Technologies)	Semiconductor sequencing	Ion torrent PGM	100–200	0.2	Resequencing
		Ion proton sequencer	100–200	2.5	Resequencing
PacBio (Pacific Bioscience)	SMRT technology	PacBioRS	250–10 000	0.1	Genome structure and metagenomics
Nanopore (Oxford Nanopore Technologies)	Ionic current sensing	GridION and MinION	10 000–50 000*	*	*De novo* and genome structure

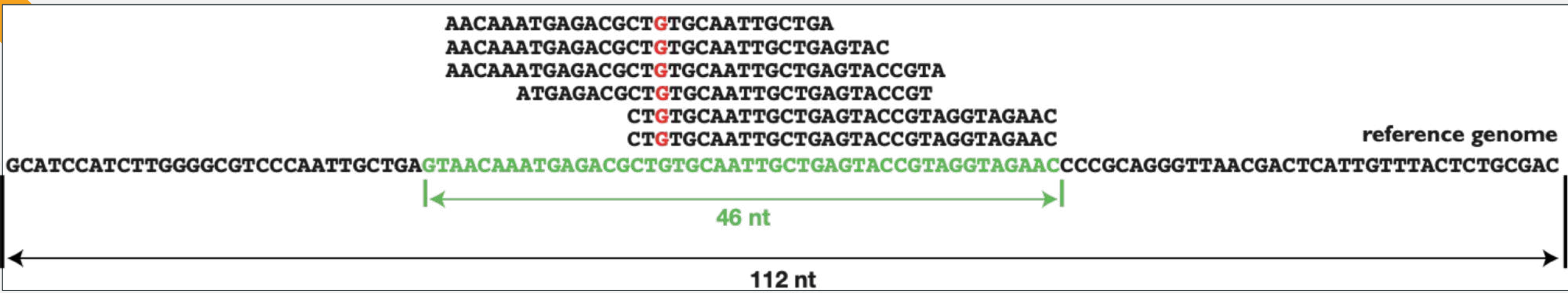
*Data not available on the corporate website.

Para minimizar o problema de tamanhos muito pequenos dos reads deve aumentar a cobertura

COBERTURA

Número de *reads* que representam um dado nucleotídeo na sequência do contig





Reference point	Calculation	Example (see Fig.2)
Whole genome	(# of sequenced bases)* / (genome size)	188 / 112 = 1,68 fold
One locus	(# of bases mapping to the locus) / (size of locus)	188 / 46 = 4,09 fold
One position	(# of reads overlapping with one position)	6 fold

K-MER (LEITURAS DE COMPRIMENTO K)

Estimativa de tendências do sequenciamento

Conteúdo de repetições

K-MER (LEITURAS DE COMPRIMENTO K)

Cobertura do sequenciamento

Detecta heterozigotos

Erros na montagem de genomas

Table 3.1 The mean number of false placements of k-mers on the genome [5]

K	*Escherichia coli*	*Saccharomyces cerevisiae*	*Arabidopsis thaliana*	*Homo sapiens*
200	0.063	0.26	0.053	0.18
160	0.068	0.31	0.064	0.49
120	0.074	0.39	0.086	1.7
80	0.082	0.49	0.15	7.2
60	0.088	0.58	0.27	18
50	0.091	0.63	0.39	32
40	0.095	0.69	0.65	78
30	0.11	0.77	1.5	330
20	0.15	1.0	5.7	2,100
10	18	63.8	880	40,000

Montagem incorreta do k-mers, depois de remover as montagens corretas

SOBREPOSIÇÃO DO *READ*

Reads que se sobrepõem a mais de um read formarão um ramo na montagem (incerteza)

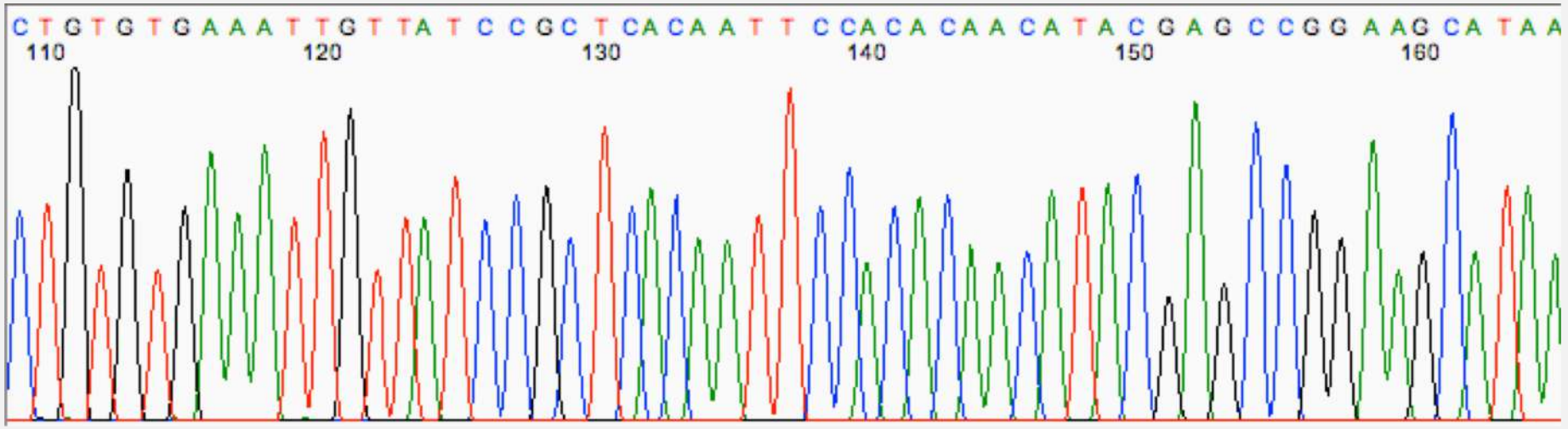
Um dos métodos para abordar essas ramificações é usando leituras em pares (*paired-end reads*)

QUALIDADE DO *READ*

Os erros de sequenciamento podem atrapalhar o processo de montagem

O uso da informação da qualidade da base pode ajudar na montagem e descobrir erros de sequenciamento

QUALIDADE DA BASE NO *READ*



Nem todos os programas usam a qualidade das bases porque isso demanda mais esforço computacional

PROBLEMAS COM SEQUENCIAMENTO

Leituras (*reads*) curtas

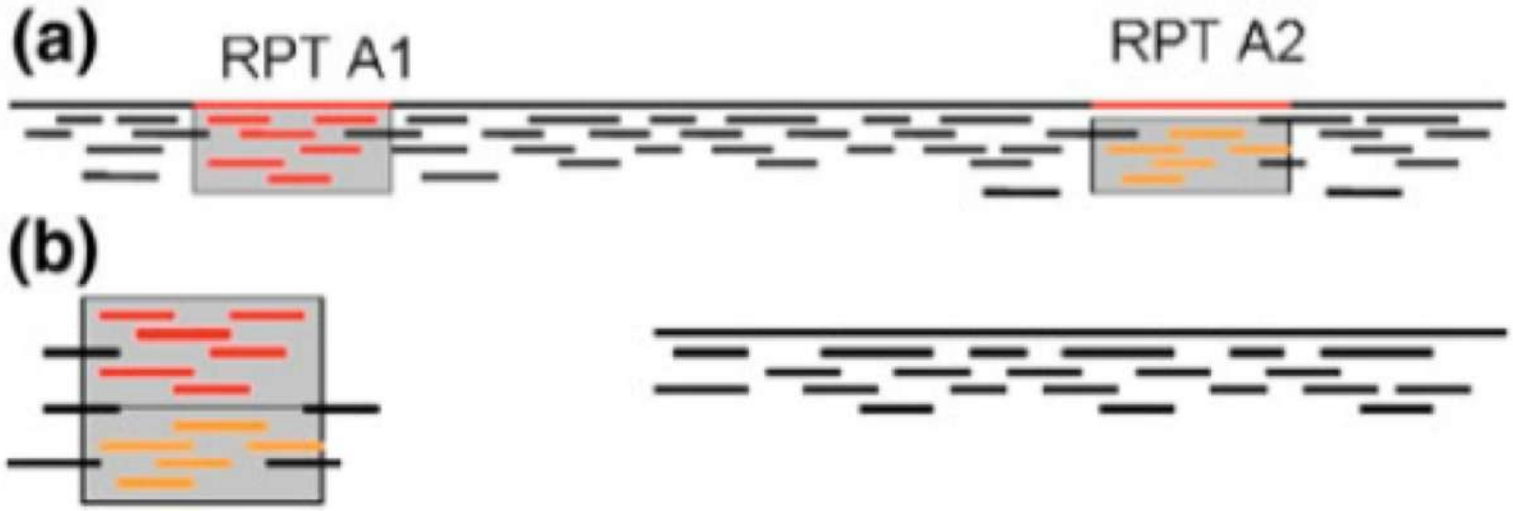
Regiões de baixa cobertura

PROBLEMAS COM SEQUENCIAMENTO

Regiões repetitivas

Erros de sequenciamento

PROBLEMAS COM SEQUENCIAMENTO



Além disso, algumas partes do genoma podem permanecer não sequenciadas

REPETIÇÕES PODEM SER MINIMIZADAS COM *PAIR-END-READS*

Pair-end-reads são fragmentos de DNA com ambas as extremidades sequenciadas.

Podem ser considerados como duas leituras com uma distância específica entre essas leituras.

PRÉ PROCESSAMENTO

Quando os dados não estão limpos e prontos para serem usados pelo algoritmo de montagem

Remover reads de baixa qualidade

Remover extremidades de baixa qualidade

COMO FILTRAR OS READS

Reads confiáveis : livres de erros e não-repetitivos

Se k-mer ocorre uma única vez , provavelmente é um erro de sequenciamento

Se k-mer ocorre muitas vezes , provavelmente é uma repetição

COMO FILTRAR OS READS

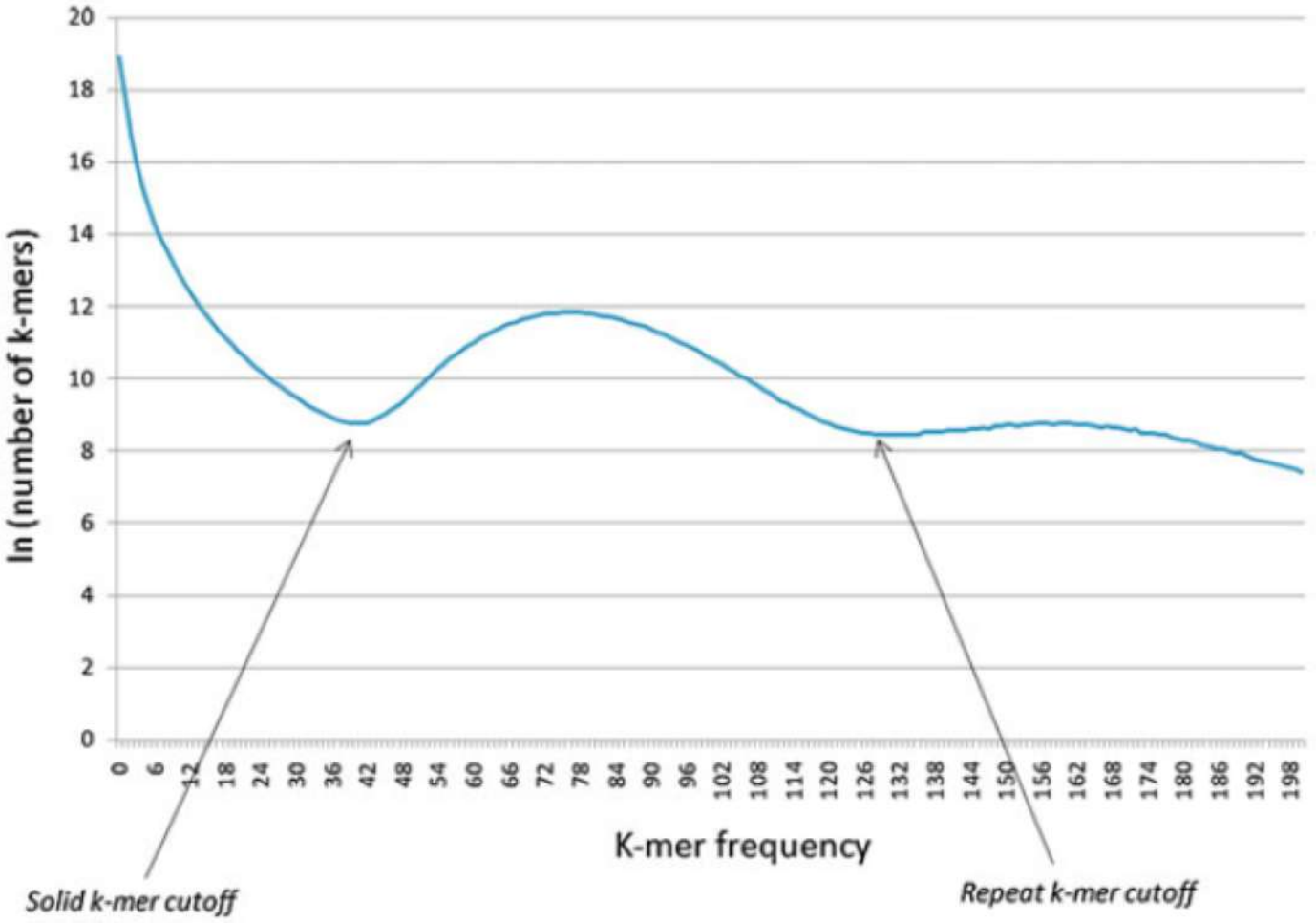


Fig. 3.4 Statistical analysis for filtering erroneous reads [6]

AVALIAÇÃO DA MONTAGEM

Pontuação de acerto com base em um genoma referência

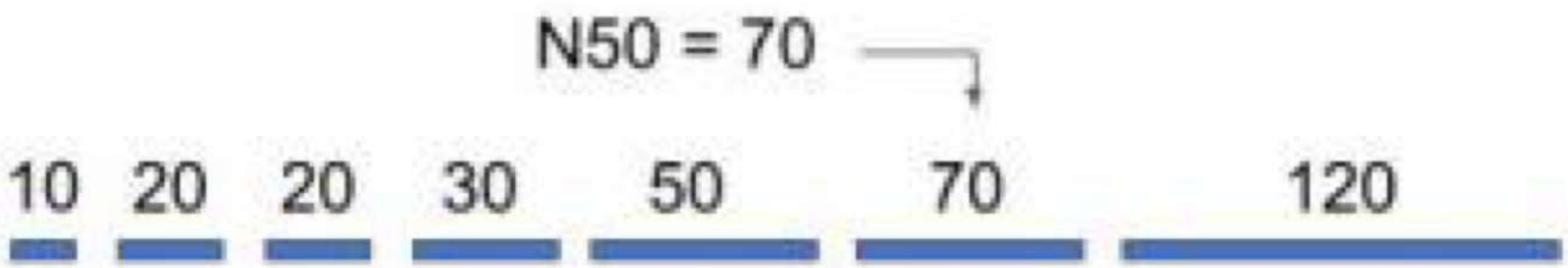
Estatísticas de tamanho, N50

DISTRIBUIÇÃO DO COMPRIMENTO CONTIG E SCAFFOLD

N50: estatística para análise de comprimentos de contig

N50: qual o menor comprimento de contigs que cobrem pelo menos metade do genoma

N50



Total length 10+20+20+30+50+70+120 = 320

50% = 160

10+20+20+30+50+70 = 200

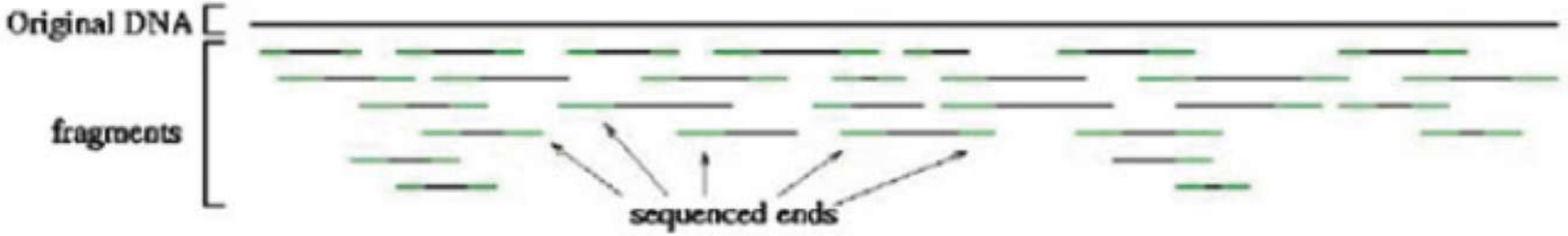
ALGORITMO DE MONTAGEM

Algoritmo de montagem deve sobrepor os reads formando uma sequência longa e contínua (cromossomo)

Devido a erros de sequenciamento e à reads não sequenciados, contigs obtidos não são completos o suficiente e não formam um cromossomo.

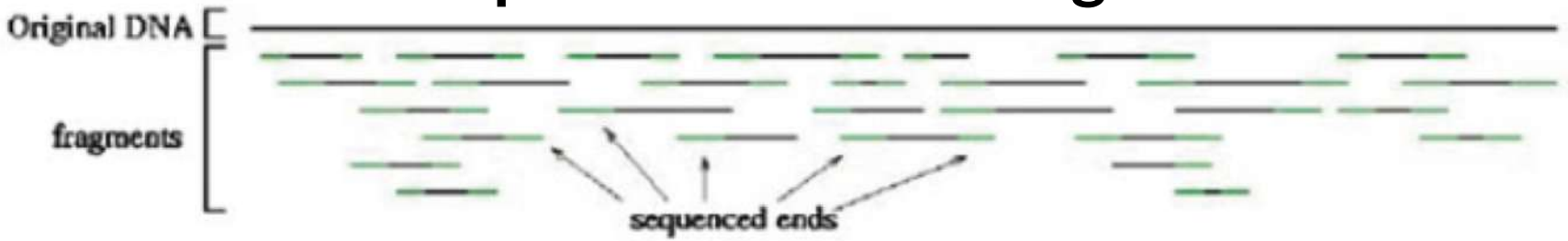
Passos no processo de montagem

(a) Original DNA broken into a collection of fragments

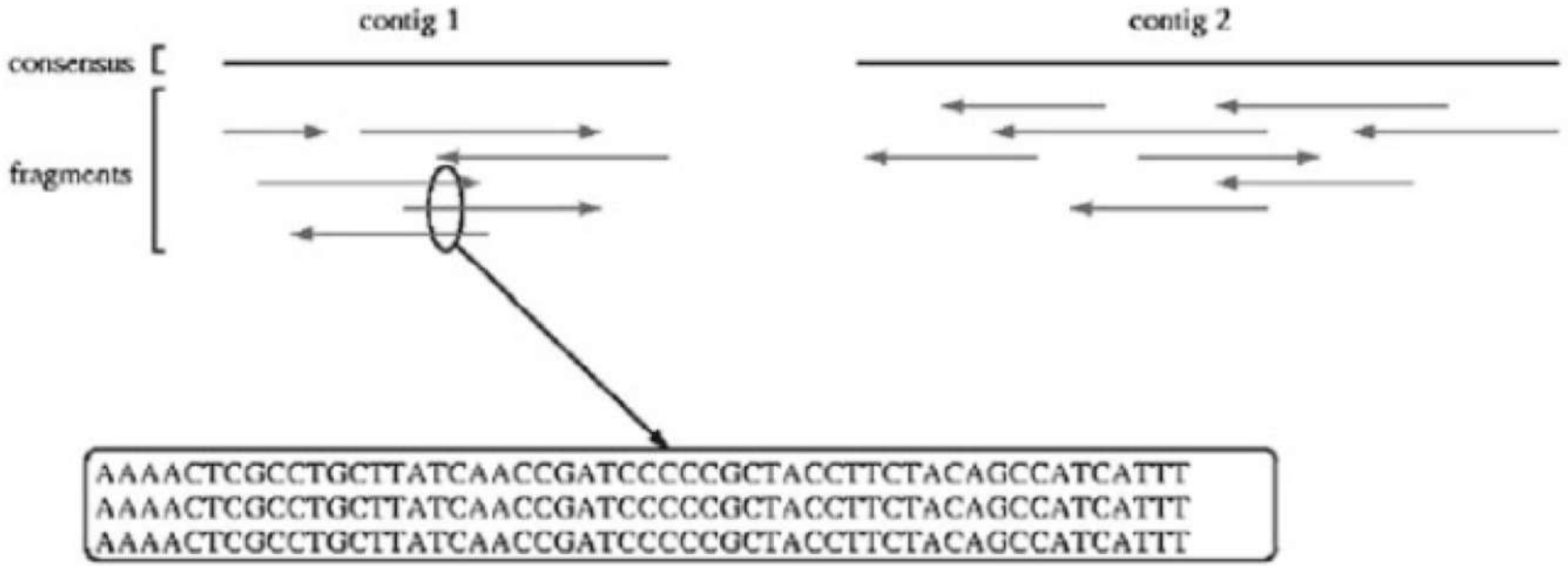


(b) The end of each fragment (drawn in green) are sequenced

Passos no processo de montagem



(b) The end of each fragment (drawn in green) are sequenced



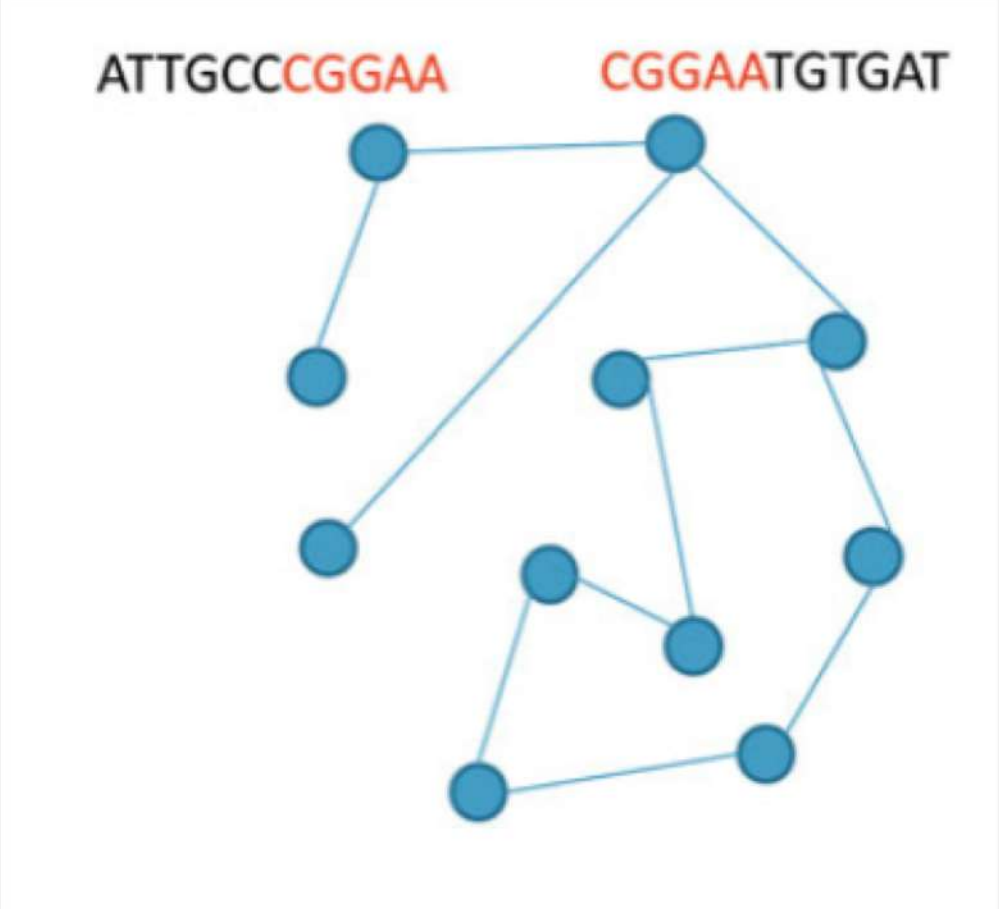
(c) Merging reads to form contigs.

Passos no processo de montagem

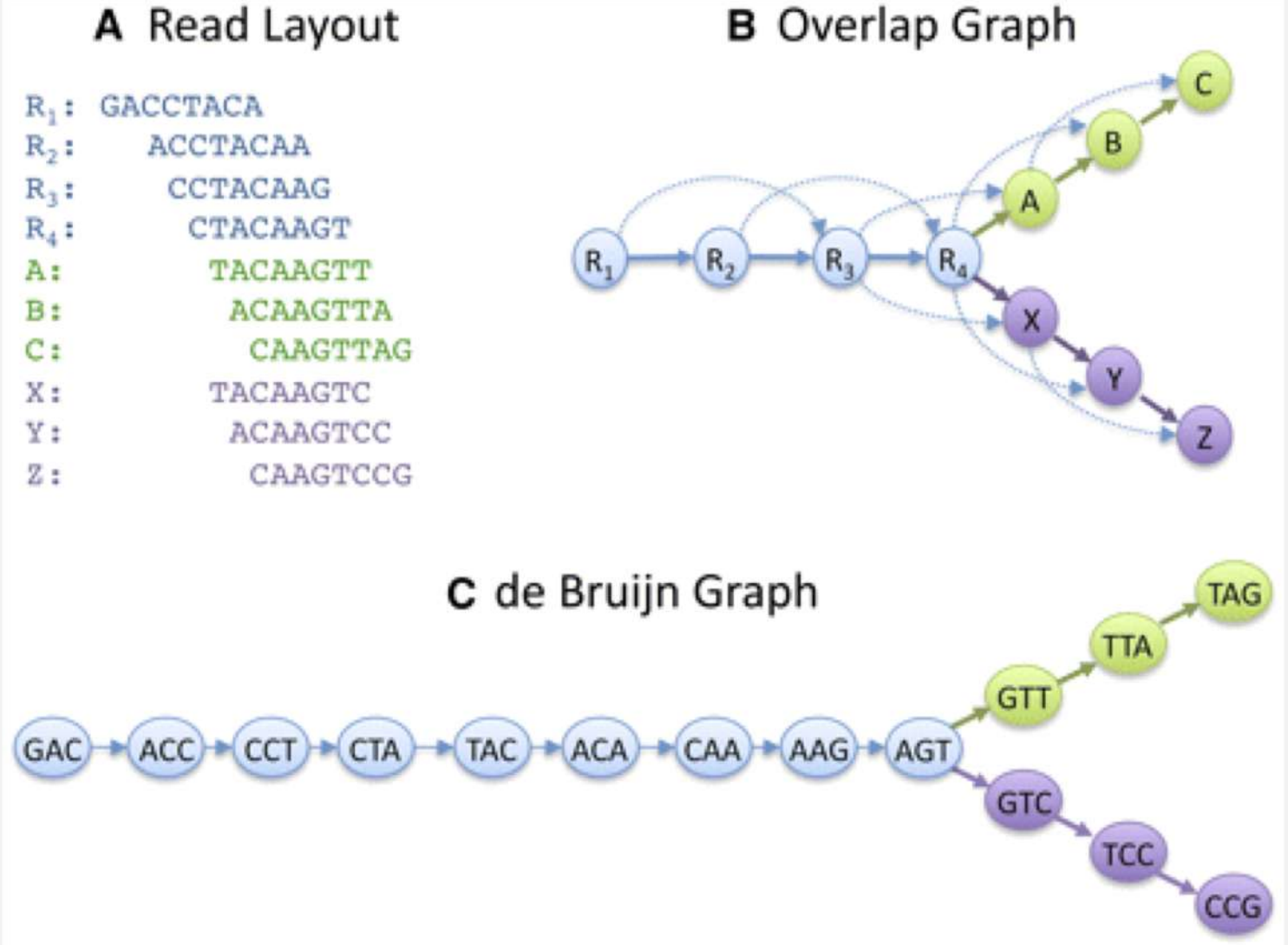
PAIR-END-READS

Necessário existir 3-5 pair-end-read com distância esperada para suportar um *scafflod*

MONTAGEM - PROBLEMA DOS GRAFOS



MONTAGEM SOBREPOSIÇÃO E GRAFO BRUIJN

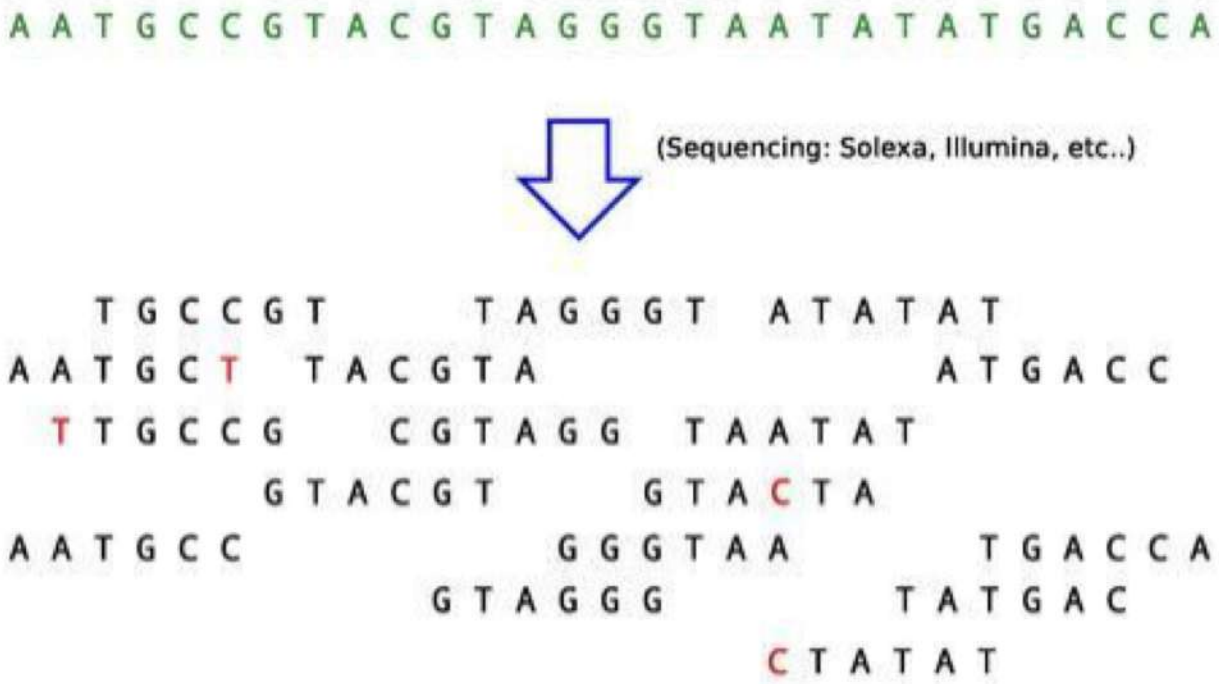


VELVET

Velvet é um algoritmo para montagem *de novo*

Adequado para montagem de *reads* curtos (25–50 bp)

N50 de 50kb para genomas procariotos

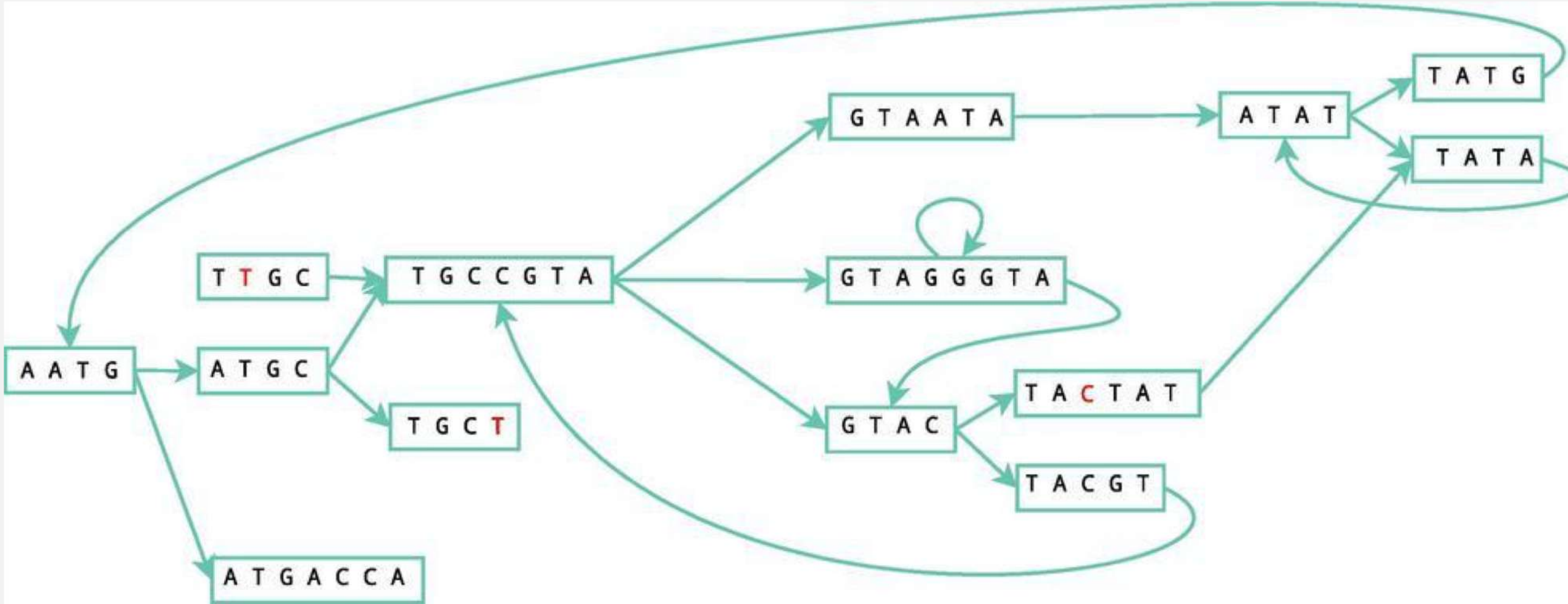


Compute k-mer with k=4

Crearte Graph for the set of k-mers

A A T A
A A T G (x2)
A C C A
A C G T (x2)
A C T A
A G G G (x2)
A T A T (x4)
A T G A (x2)
A T G C (x2)
C C G T
C G T A (x2)
C T A T
G A C C (x2)
G C C G (x2)
G G G T (x2)
G G T A
G T A A
G T A C (x2)
G T A G (x2)
T A A T
T A C G (x2)
T A C T
T A G G (x3)
T A T A (x2)
T A T G
T G A C (x3)
T G C C (x3)
T G C T
T T G C

VELVET


A A T G
A T G C
T T G C
T G C C G T A
T G C T
A T G A C C A
G T A A T A
G T A G G G T A
G T A C
A T A T
T A T G
T A T A
T A C T A T
T A C G T

N50

N50 size

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome

N50 size = 30 kbp

(300k+100k+45k+45k+30k = 520k >= 500kbp)

A greater N50 is indicative of improvement in every dimension:

- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence for all downstream analysis