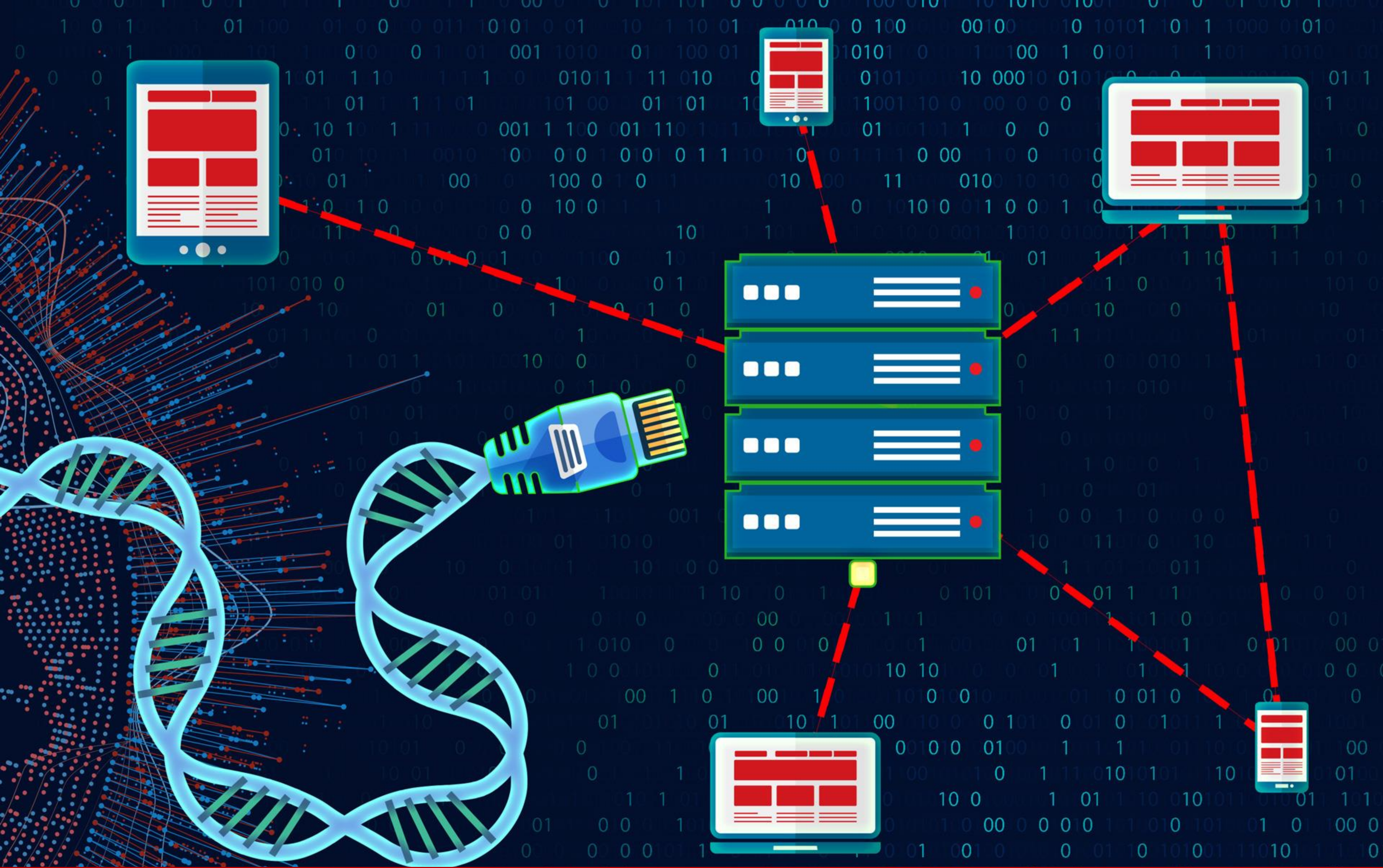


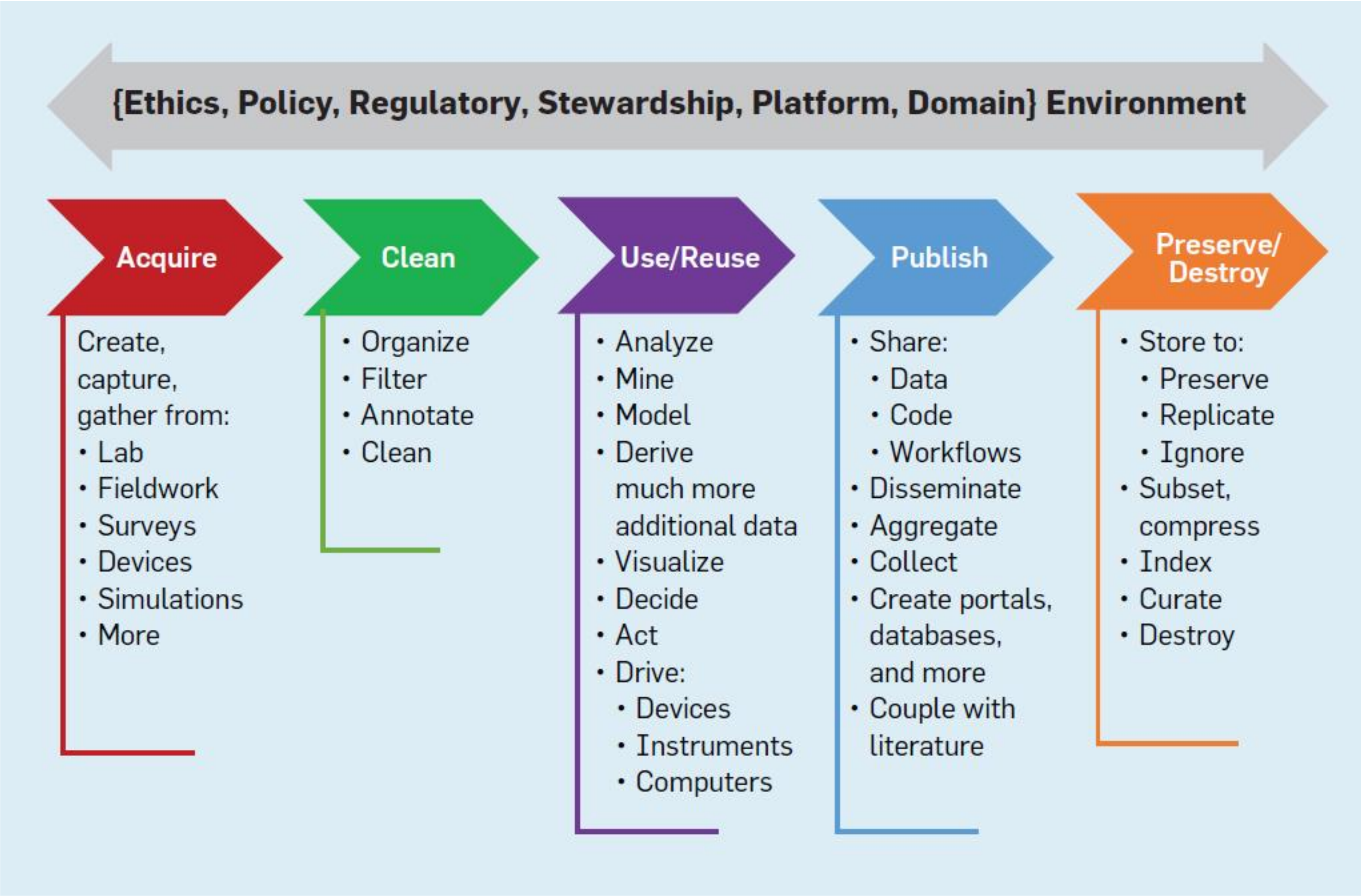
Lecture-L1. 生物信息学导论

本章内容

- 1.1 引言--数据的价值
- 1.2 什么是生物信息学？
- 1.3 生物信息学发展简史
- 1.4 生物信息学的应用领域
- 1.5 大数据时代的生物信息学
- 1.6 总结与展望

第1节：引言——数据的价值

Varied **Big Data** in Our Life



{Ethics, Policy, Regulatory, Stewardship, Platform, Domain} Environment

Acquire

Create, capture, gather from:

- Lab
- Fieldwork
- Surveys
- Devices
- Simulations
- More

Clean

- Organize
- Filter
- Annotate
- Clean

Use/Reuse

- Analyze
- Mine
- Model
- Derive much more additional data
- Visualize
- Decide
- Act
- Drive:
  - Devices
  - Instruments
  - Computers

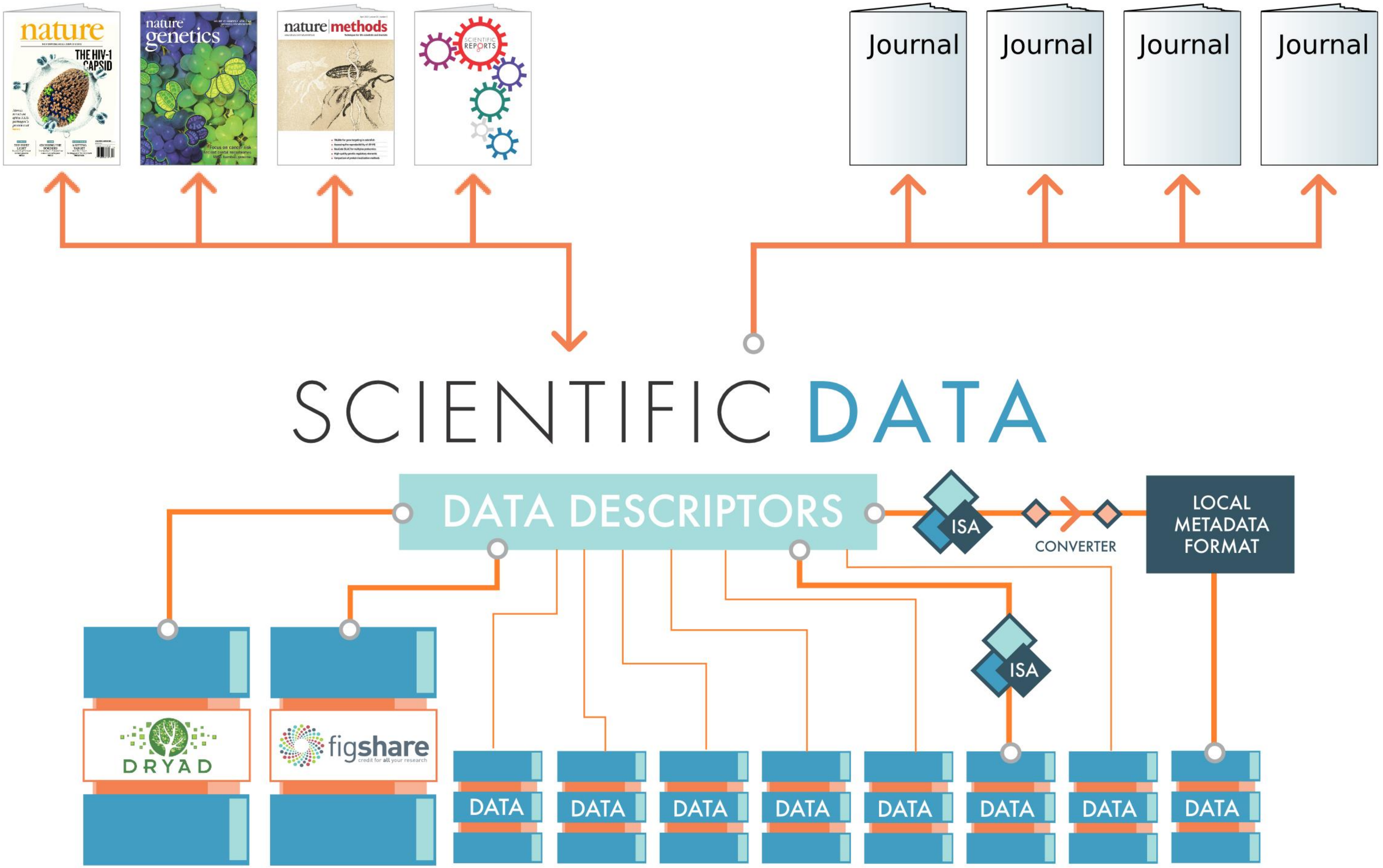
Publish

- Share:
  - Data
  - Code
  - Workflows
- Disseminate
- Aggregate
- Collect
- Create portals, databases, and more
- Couple with literature

Preserve/ Destroy

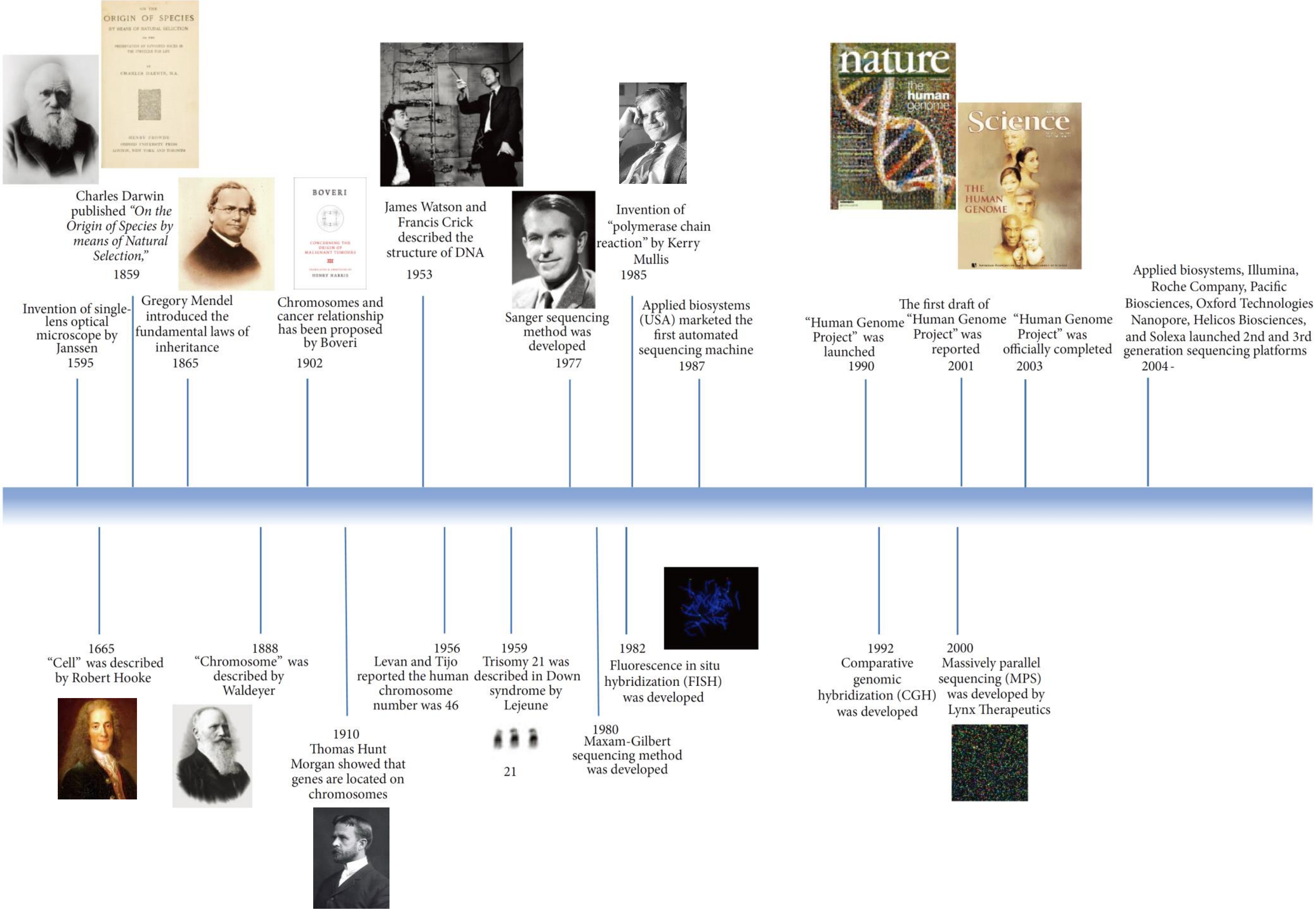
- Store to:
  - Preserve
  - Replicate
  - Ignore
- Subset, compress
- Index
- Curate
- Destroy

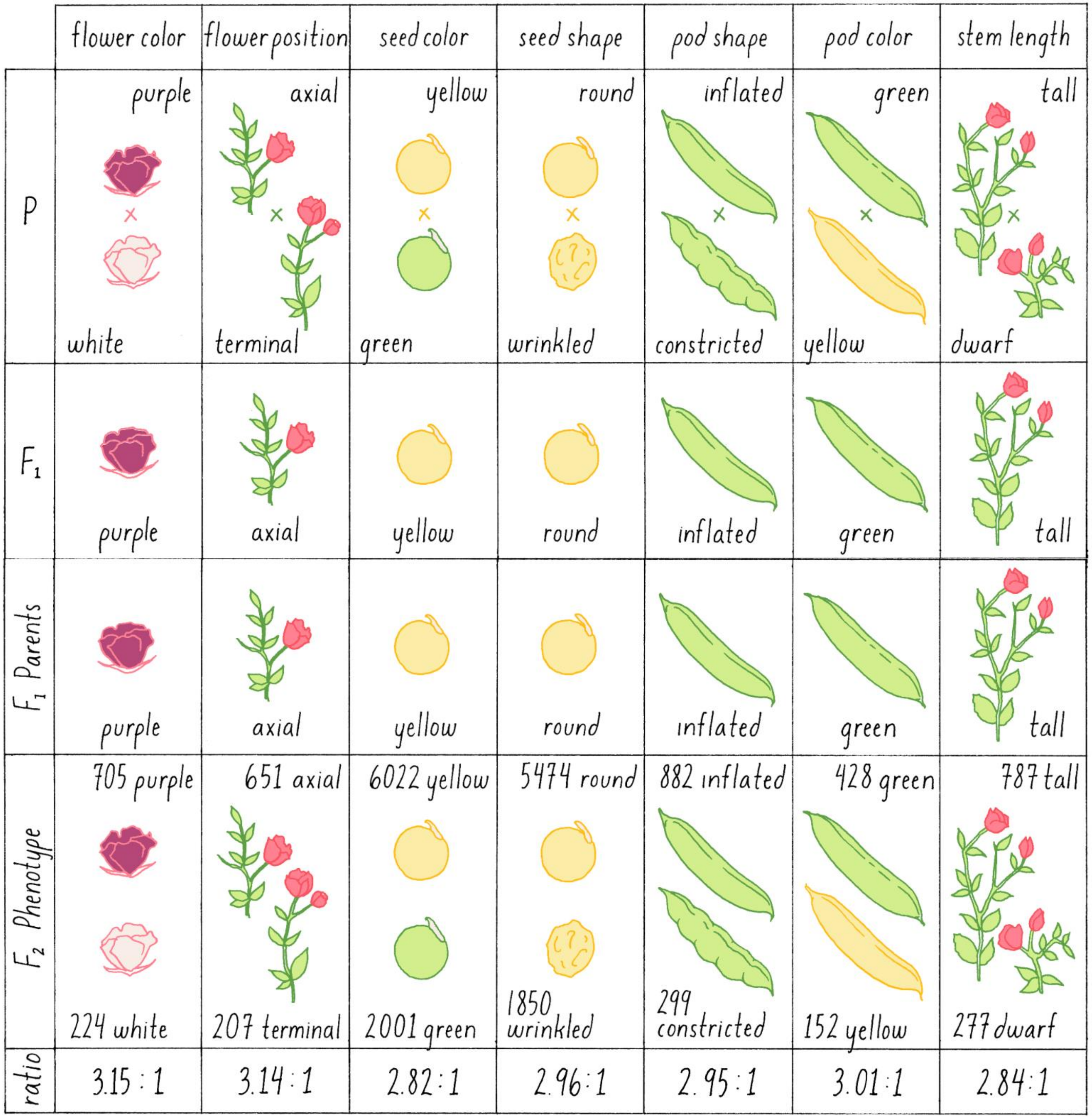
Framework used in processing Big Data



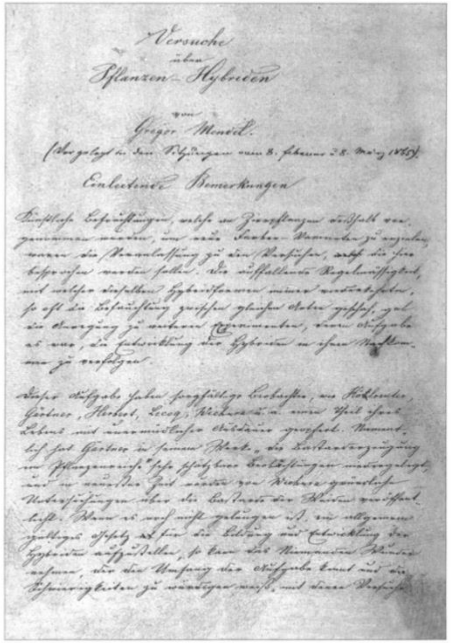
Ref: http://blogs.nature.com/

Scientific Data to complement and promote public data repositories


Invention of single-lens optical microscope by Janssen 1595
1665 "Cell" was described by Robert Hooke
Charles Darwin published *"On the Origin of Species by means of Natural Selection,"* 1859
ORIGIN OF SPECIES
Gregory Mendel introduced the fundamental laws of inheritance 1865
1888 "Chromosome" was described by Waldeyer
Chromosomes and cancer relationship has been proposed by Boveri 1902
BOVERI
1910 Thomas Hunt Morgan showed that genes are located on chromosomes
James Watson and Francis Crick described the structure of DNA 1953
1956 Levan and Tijo reported the human chromosome number was 46
1959 Trisomy 21 was described in Down syndrome by Lejeune
21
Sanger sequencing method was developed 1977
1980 Maxam-Gilbert sequencing method was developed
1982 Fluorescence in situ hybridization (FISH) was developed
Invention of "polymerase chain reaction" by Kerry Mullis 1985
Applied biosystems (USA) marketed the first automated sequencing machine 1987
"Human Genome Project" was launched 1990
1992 Comparative genomic hybridization (CGH) was developed
2000 Massively parallel sequencing (MPS) was developed by Lynx Therapeutics
The first draft of "Human Genome Project" was reported 2001
nature the human genome
Science THE HUMAN GENOME
"Human Genome Project" was officially completed 2003
Applied biosystems, Illumina, Roche Company, Pacific Biosciences, Oxford Technologies Nanopore, Helicos Biosciences, and Solexa launched 2nd and 3rd generation sequencing platforms 2004 -



	flower color	flower position	seed color	seed shape	pod shape	pod color	stem length
P	purple × white	axial × terminal	yellow × green	round × wrinkled	inflated × constricted	green × yellow	tall × dwarf
F_1	purple	axial	yellow	round	inflated	green	tall
F_1 Parents	purple	axial	yellow	round	inflated	green	tall
F_2 Phenotype	705 purple 224 white	651 axial 207 terminal	6022 yellow 2001 green	5474 round 1850 wrinkled	882 inflated 299 constricted	428 green 152 yellow	787 tall 277 dwarf
ratio	3.15 : 1	3.14 : 1	2.82 : 1	2.96 : 1	2.95 : 1	3.01 : 1	2.84 : 1



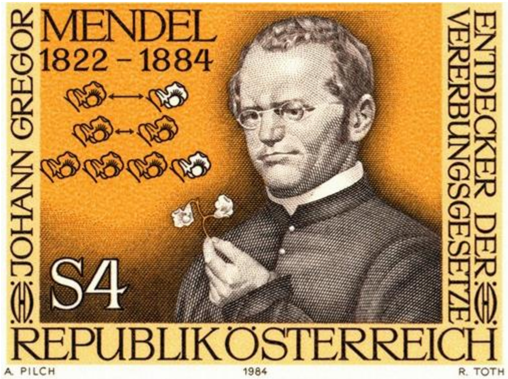
Versuche
über
Pflanzen-Hybriden
von
Gregor Mendel.
(Vorgelegt in den Sitzungen vom 8. Februar u. 8. März 1865.)

Einleitende Bemerkungen

Künstliche Befruchtungen, welche an Zierpflanzen deshalb vorgenommen wurden, um neue Farben-Varianten zu erzielen, waren die Veranlassung zu den Versuchen, welche hier besprochen werden sollen. Die auffallende Regelmässigkeit, mit welcher dieselben Hybridformen immer wiederkehrten, so oft die Befruchtung zwischen gleichen Arten geschah, gab die Anregung zu weiteren Experimenten, deren Aufgabe es war, die Entwicklung der Hybriden in ihren Nachkommen zu verfolgen.

Dieser Aufgabe haben sorgfältige Beobachter, wie Kölreuter, Gärtner, Herbert, Lecoq, Wichura u. a. einen Theil ihres Lebens mit unermüdlicher Ausdauer geopfert. Namentlich hat Gärtner in seinem Werke „die Bastarderzeugung im Pflanzenreiche" sehr schätzbare Beobachtungen niedergelegt, und in neuester Zeit wurden von Wichura gründliche Untersuchungen über die Bastarde der Weiden veröffentlicht. Wenn es noch nicht gelungen ist, ein allgemein giltiges Gesetz für die Bildung und Entwicklung der Hybriden aufzustellen, so kann das Niemanden Wunder nehmen, der den Umfang der Aufgabe kennt und die Schwierigkeiten zu würdigen weiss, mit denen Versuche

The first page of the manuscript of Mendel's Experiments in Plant Hybridization, which was published in 1865.

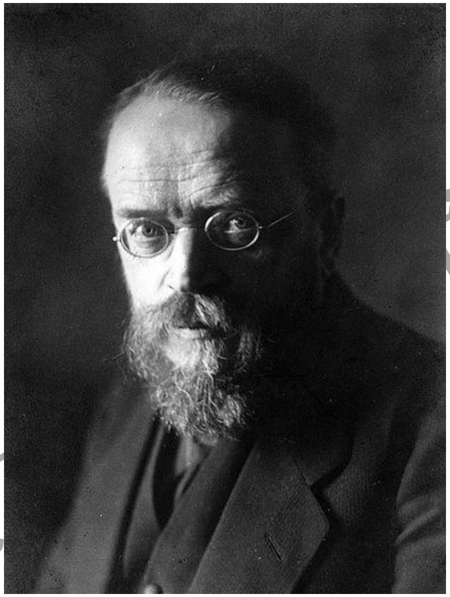


一个超越时代的天才！

Rediscovery of Mendel’s work

Hugo de Vries

(*Netherlands*)



Carl Correns

(*Germany*)

Erich Tschermak

(*Austria*)

英勇的人民子弟兵在抢救地震中的受难群众（*2008.05.12*）

第 21卷第 3期
2006年 9月

灾　害　学
JOURNAL OF CATASTROPHOLOGY

Vol. 21 No. 3
Sep. 2006

基于可公度方法的川滇地区地震趋势研究*

龙小霞，延军平，孙虎，王祖正

(陕西师范大学 旅游与环境学院，陕西 西安　710062)

摘要: 川滇地区为我国大陆最显著的强震活动区域，地震活动频繁。在对川滇地区强震灾害数据分析的基础上，应用三元、四元、五元可公度法分别预测了该地区下（几）次可能发生强震的趋势，以便能更好地配合防震减灾工

数据是极其有用的，值得挖掘！

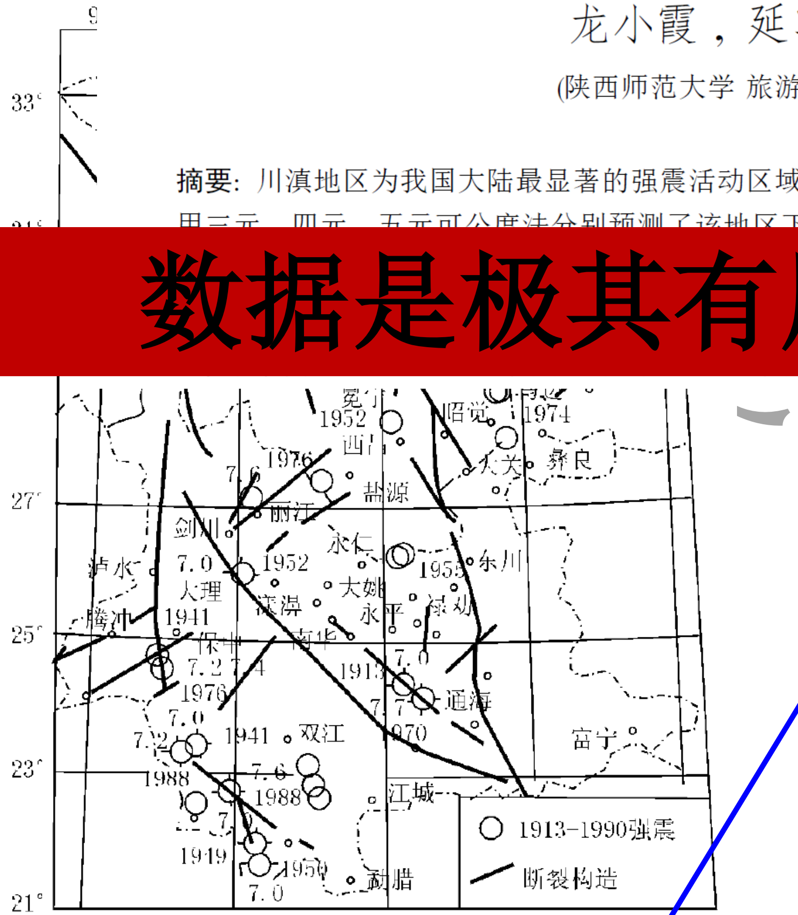


图 1　川滇地区 20世纪以来强震分布格局图

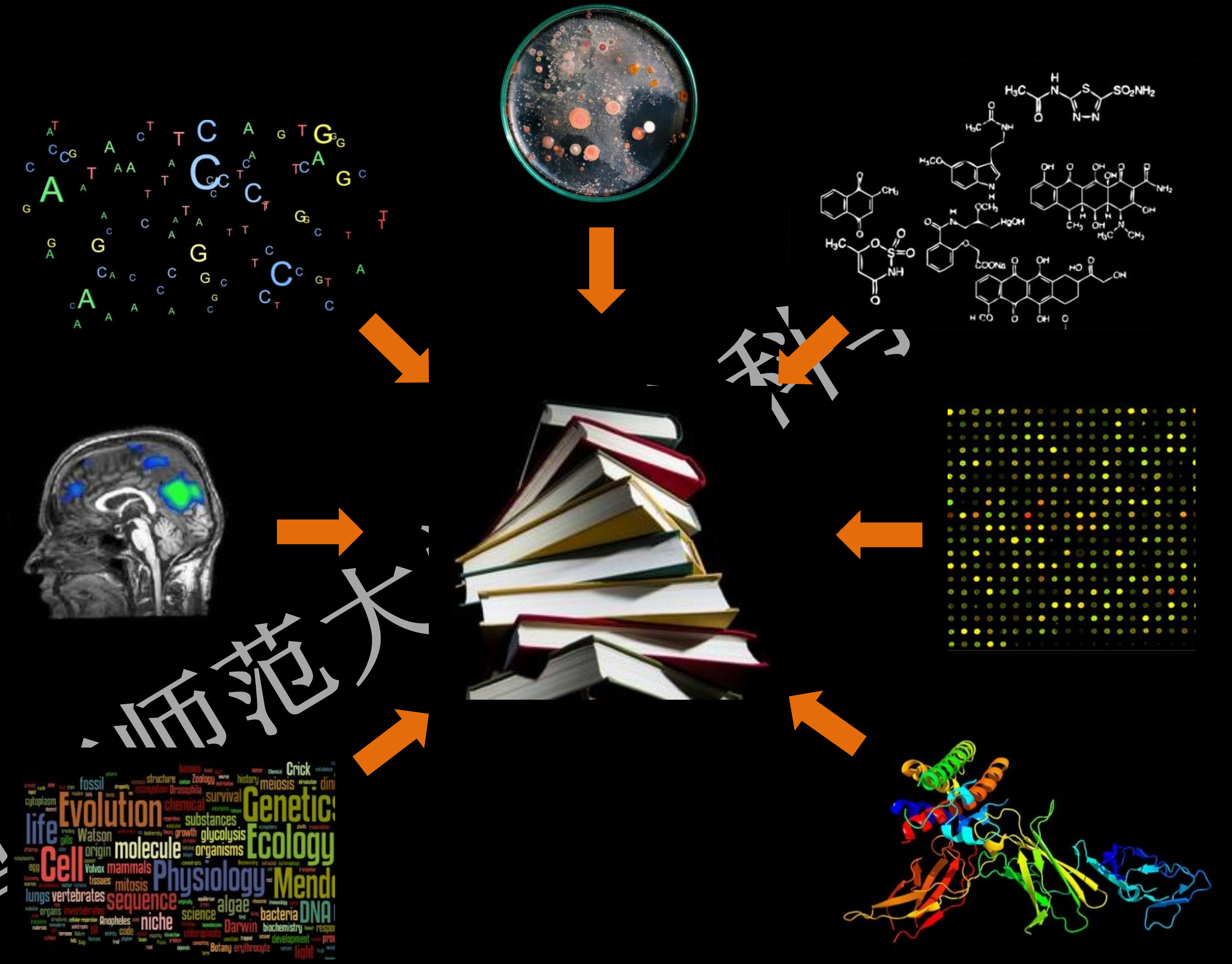
数据

总结以上几种预测结果，可以看出从灾害信息来讲，2007年和 2008年的灾害信号比较强，尤其是2008年更符合已有地震资料的统计规律，因此川滇地区下（几）次可能发生≥ 6.7级地震的年份为 2008年。

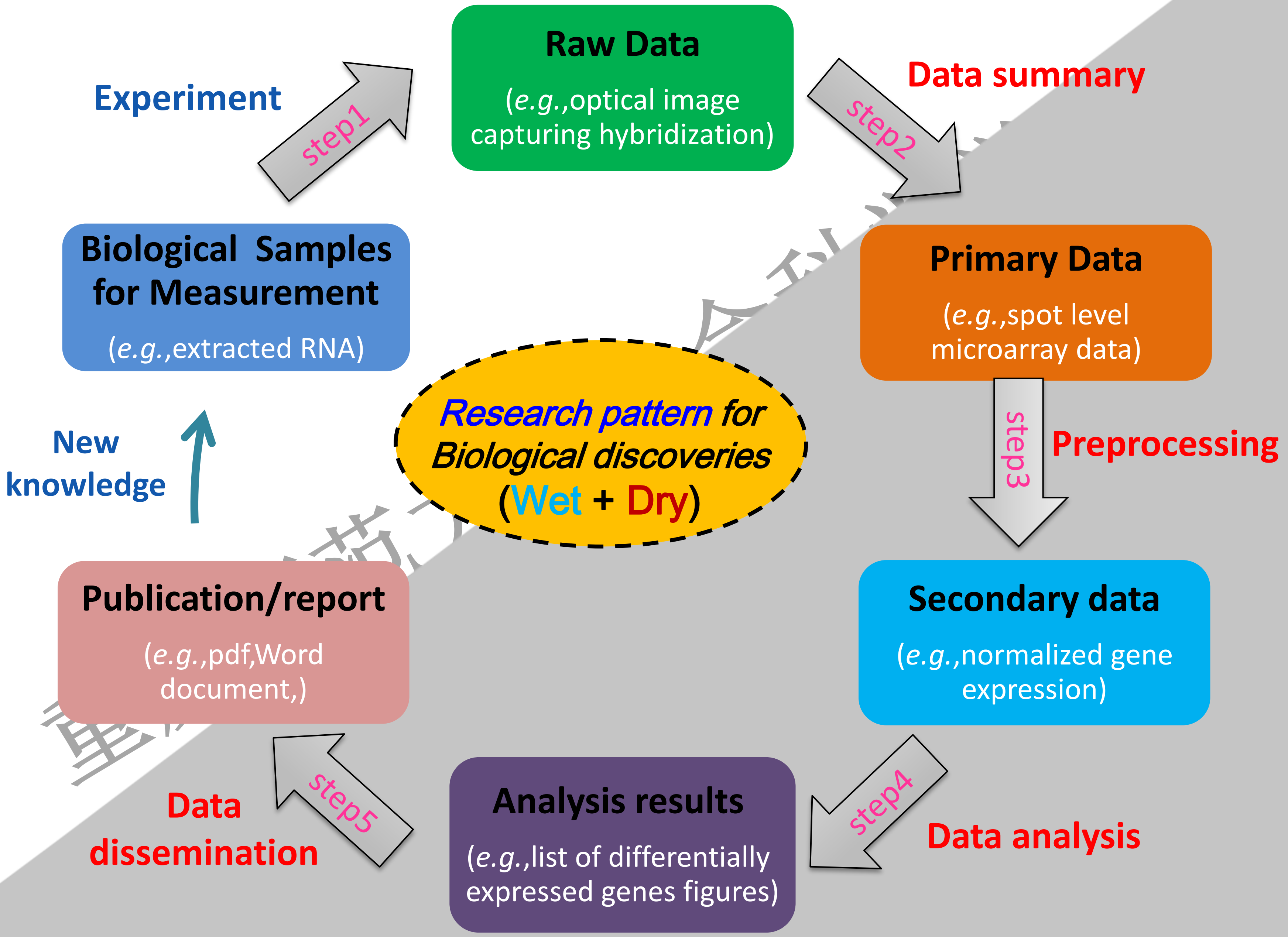
3　结论与建议

从以上所进行的推算与预测结果看，在 2008年左右，川滇地区有可能发生≥ 6.7级强烈地震。为了更好地配合防震减灾活动，笔者提出以下建议。

常见的生物学数据



Bioinformatics: *from data to knowledge*



Research pattern for Biological discoveries (*Wet* + *Dry*)

Biological Samples for Measurement

(*e.g.*,extracted RNA)

Experiment

step1

Raw Data

(*e.g.*,optical image capturing hybridization)

Data summary

step2

Primary Data

(*e.g.*,spot level microarray data)

step3

Preprocessing

Secondary data

(*e.g.*,normalized gene expression)

step4

Data analysis

Analysis results

(*e.g.*,list of differentially expressed genes figures)

step5

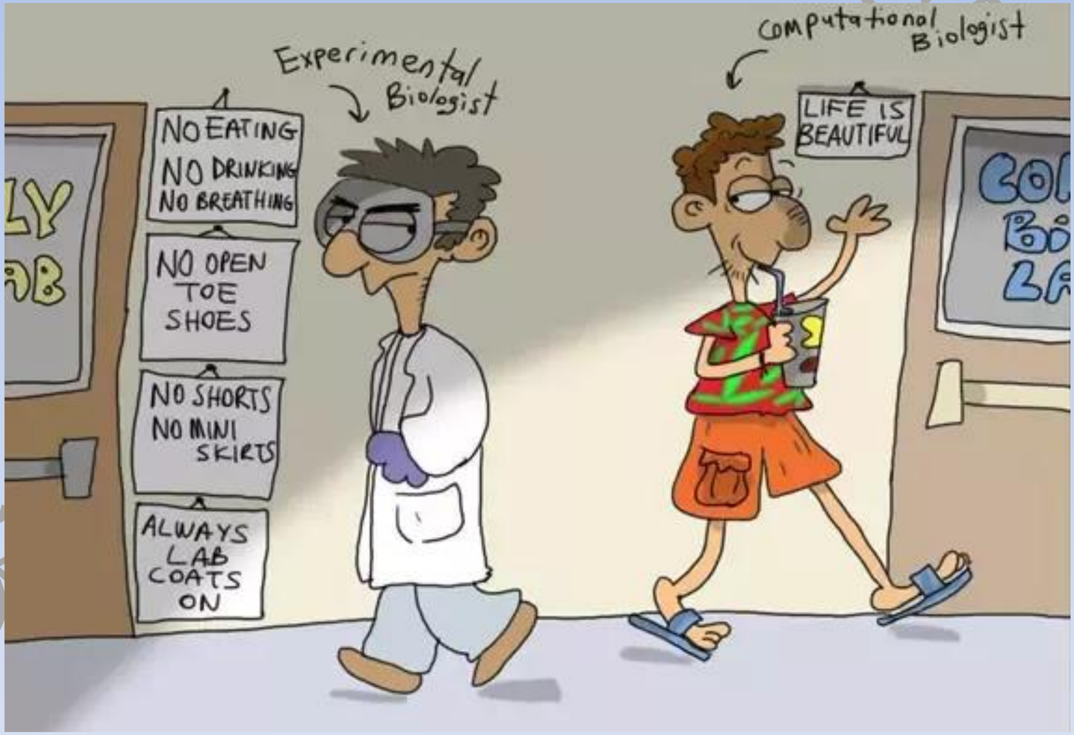
Data dissemination

Publication/report

(*e.g.*,pdf,Word document,)

New knowledge

Biology, experimental or computational?



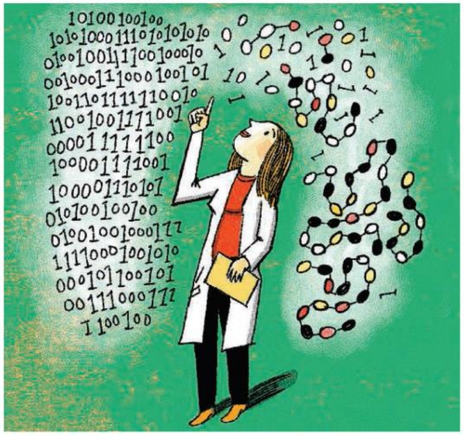
WORKING LIFE

By Elisabeth Pain

Group Leader
Dr Sarah Teichmann

Biology, wet and dry

Sarah Teichmann's work on how cells regulate gene expression and build protein complexes recently won her a European Molecular Biology Organization Gold Medal. At 40, Teichmann holds a joint appointment with the European Bioinformatics Institute and the Wellcome Trust Sanger Institute in Hinxton, U.K. She leads a systems biology group of 17 researchers that uses both computational methods and lab experimentation. *Science* Careers asked Teichmann how she combines the two approaches. This interview has been edited for clarity and brevity.



"What unifies both wet and dry work is the conceptual part of the science."

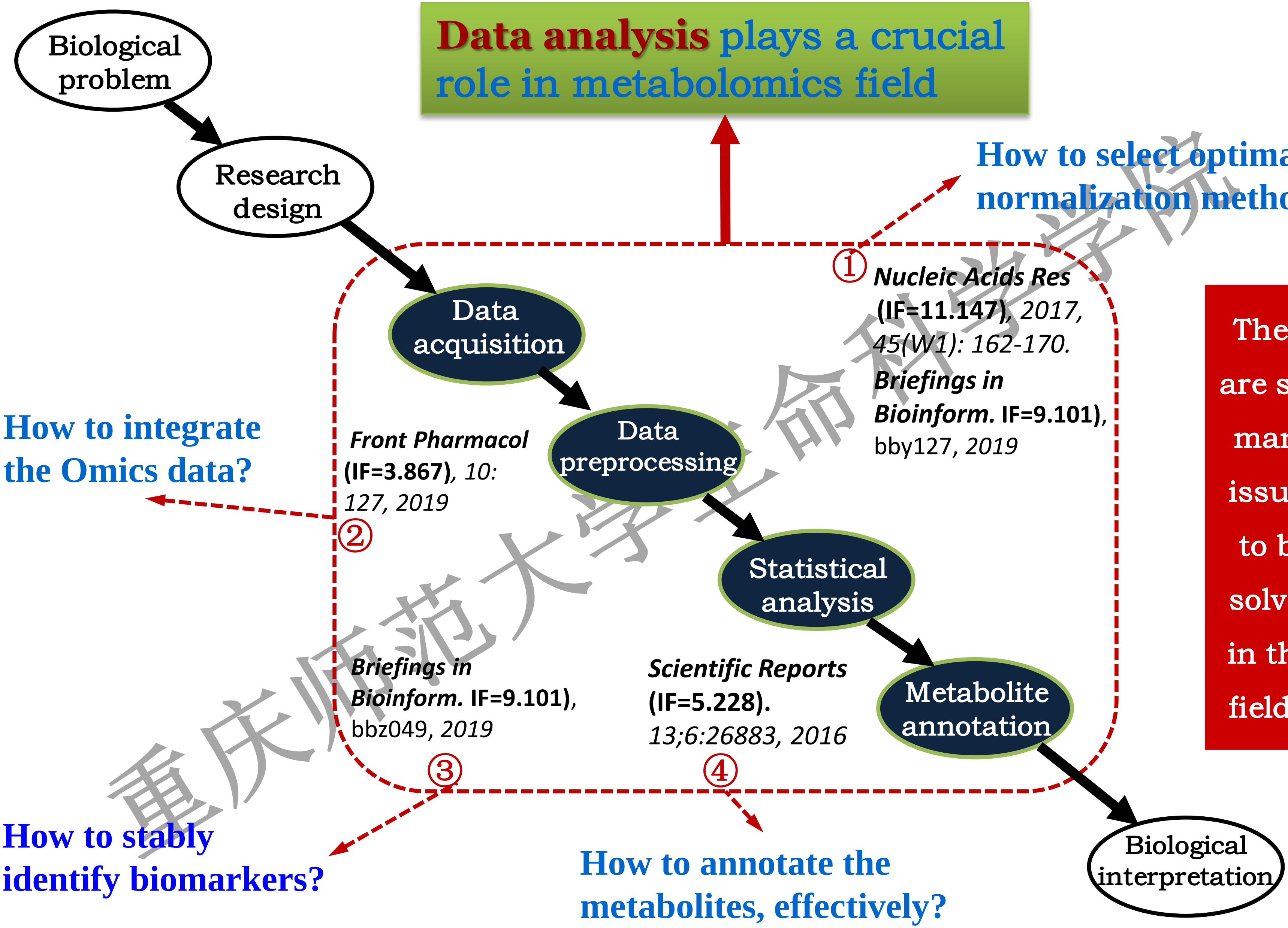
Q: Was computational biology a risky career choice?

A: Yes. But I never looked back, even though at one point I came to feel that computational biology and bioinformatics were viewed as eccentric and unorthodox.

My *Ph.D.* mentor exuded such unwavering optimism and confidence (坚定不移的乐观和信心), however, that it made his lab a great place to work. Altogether, during my *Ph.D.*, I published 10 papers.

Ref: Teichmann, S. & Pain, E. Biology, wet and dry. *Science* **349**, 662 (2015).

Data analysis plays a crucial role in metabolomics field



Biological problem

Research design

Data acquisition

Data preprocessing

Statistical analysis

Metabolite annotation

Biological interpretation

① ***Nucleic Acids Res* (IF=11.147)**, *2017, 45(W1): 162-170.*

***Briefings in Bioinform.* IF=9.101)**, bby127, *2019*

How to select optimal normalization method?

② ***Front Pharmacol* (IF=3.867)**, *10: 127, 2019*

How to integrate the Omics data?

③ ***Briefings in Bioinform.* IF=9.101)**, bbz049, *2019*

How to stably identify biomarkers?

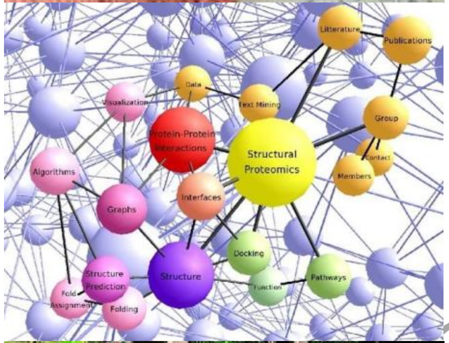
④ ***Scientific Reports* (IF=5.228).** *13;6:26883, 2016*

How to annotate the metabolites, effectively?

There are still many issues to be solved in this field!

Li, B. *et al*. Performance evaluation and online realization of data-driven normalization methods used in LC/MS based untargeted metabolomics analysis. ***Scientific reports*** **6**, 38881 (2016).

PLOS | BIOLOGY



RESEARCH MATTERS

All biology is computational biology

- Here, I argue that computational thinking and techniques are so central to the quest of understanding life that today all biology is computational biology.

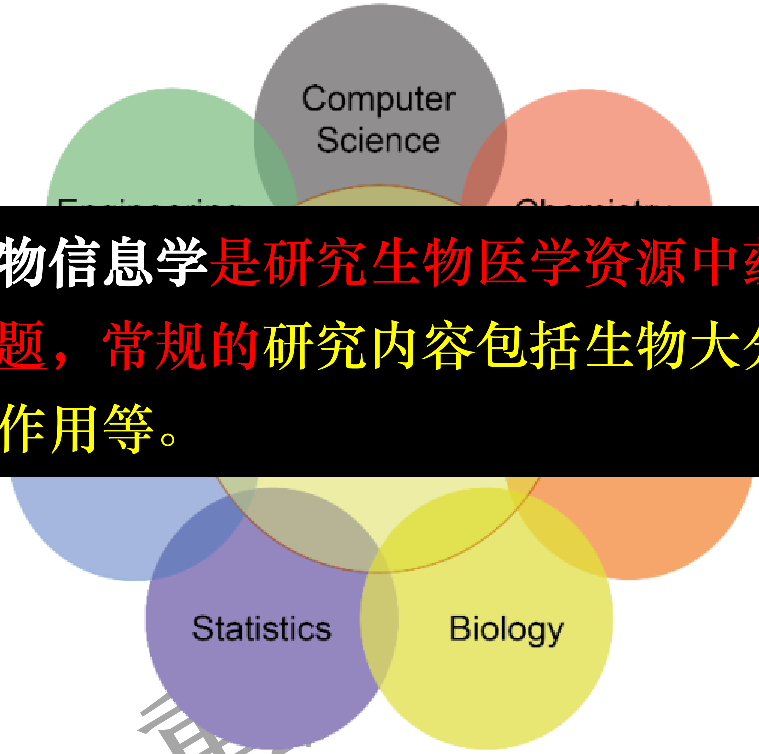
- Computational biology brings order into our understanding of life, it makes biological concepts rigorous and testable, and it provides a reference map that holds together individual insights.

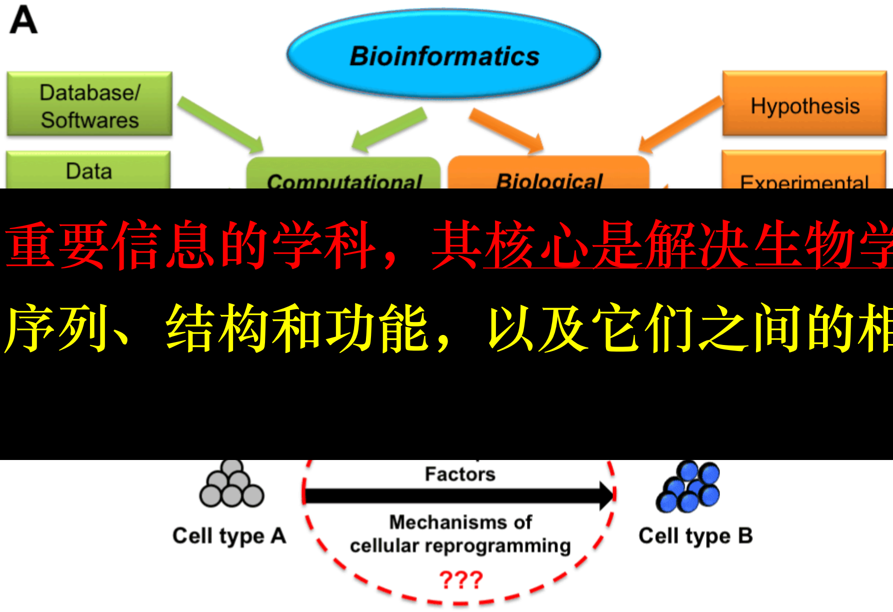
- The next modern synthesis in biology will be driven by mathematical, statistical, and computational methods being absorbed into mainstream biological training, turning biology into a quantitative science.

Ref: Pain, E. Biology, wet and dry. *PLoS Biology* **349**, 662-662 (2017).

第2节：什么是生物信息学

Bioinformatics is an interdisciplinary field, and it mainly develops methods and software tools for understanding biological data.



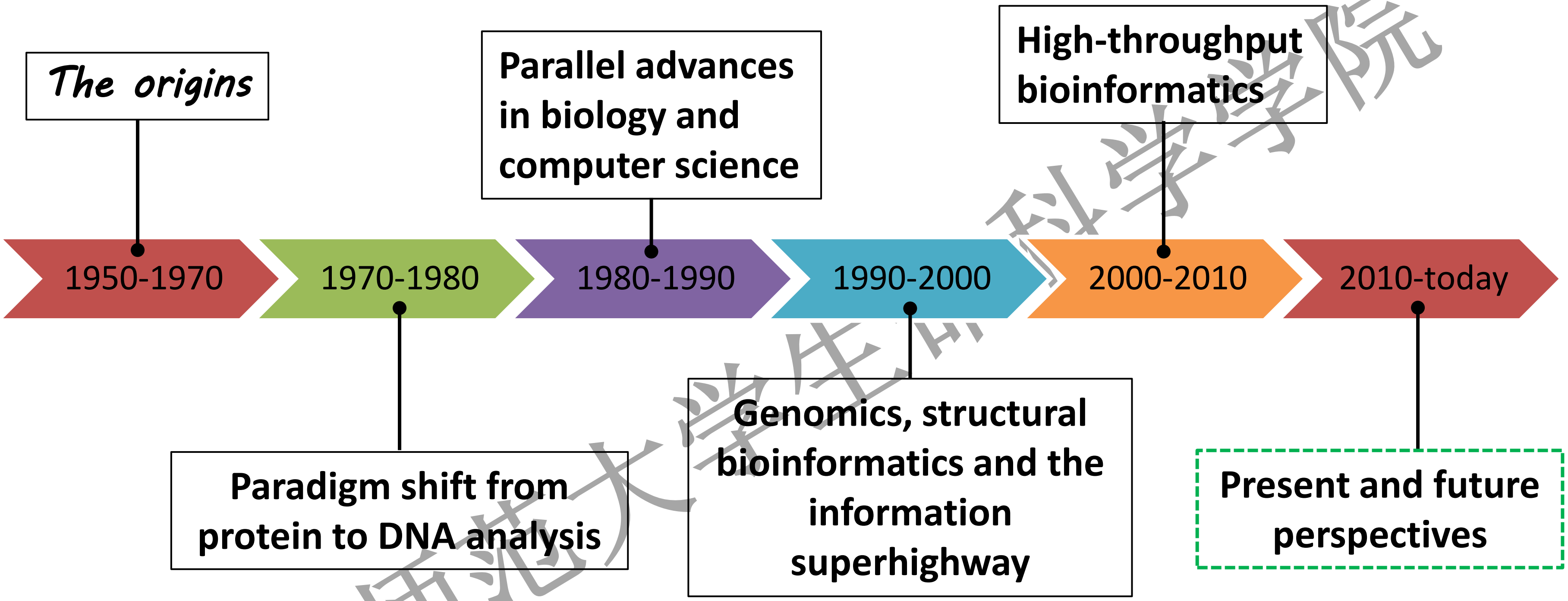


生物信息学是研究生物医学资源中蕴含的重要信息的学科，其核心是解决生物学问题，常规的研究内容包括生物大分子的序列、结构和功能，以及它们之间的相互作用等。

Aims

- ✓ ① To access existing information and to submit new entries
- ✓ ② To develop tools and resources
- ✓ ③ Using these tools to analyze the data and interpret the results

第3节：生物信息学发展简史


The origins
1950-1970
Paradigm shift from protein to DNA analysis
1970-1980
Parallel advances in biology and computer science
1980-1990
Genomics, structural bioinformatics and the information superhighway
1990-2000
High-throughput bioinformatics
2000-2010
Present and future perspectives
2010-today

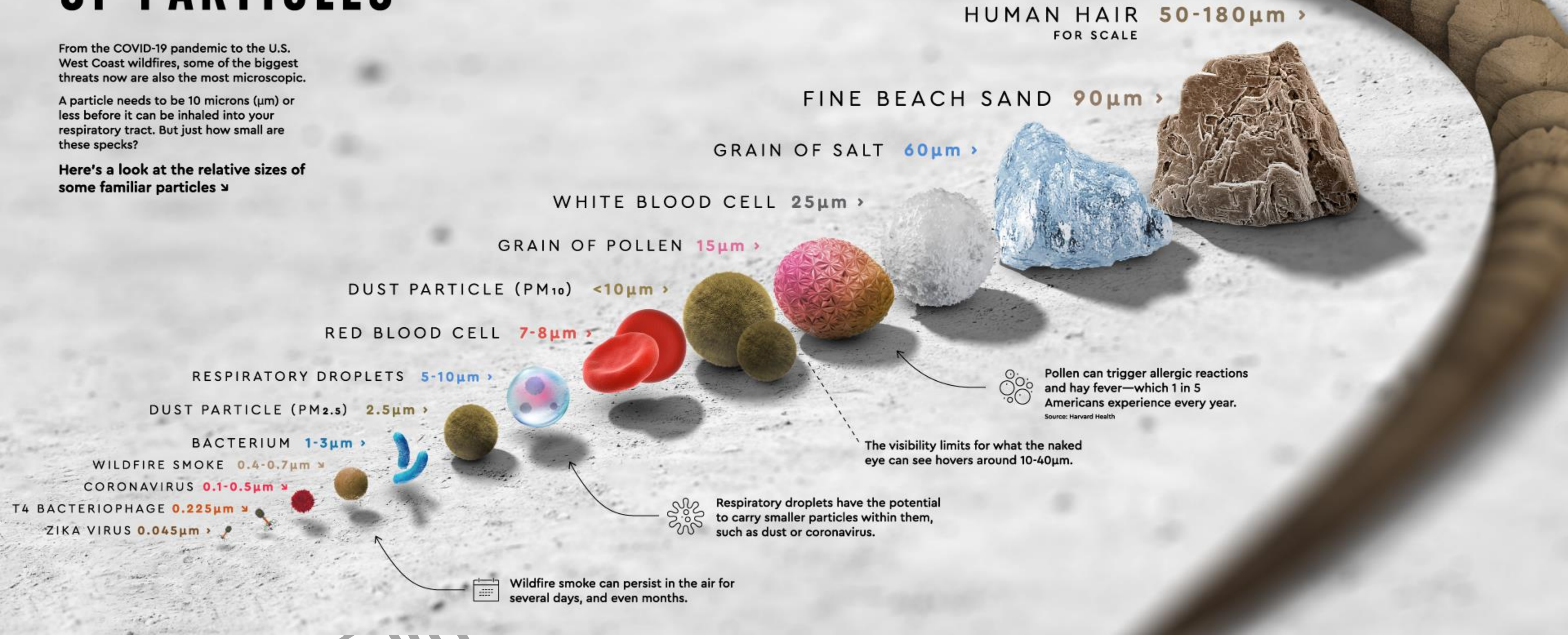
It is easy for researchers to believe that modern bioinformatics are relatively recent, coming to the rescue of NGS data analysis. However, the very beginnings of bioinformatics occurred more than 50 years ago, when desktop computers were still a hypothesis and DNA could not yet be sequenced.

THE RELATIVE SIZE OF PARTICLES

From the COVID-19 pandemic to the U.S. West Coast wildfires, some of the biggest threats now are also the most microscopic.

A particle needs to be 10 microns (µm) or less before it can be inhaled into your respiratory tract. But just how small are these specks?

Here's a look at the relative sizes of some familiar particles ↘



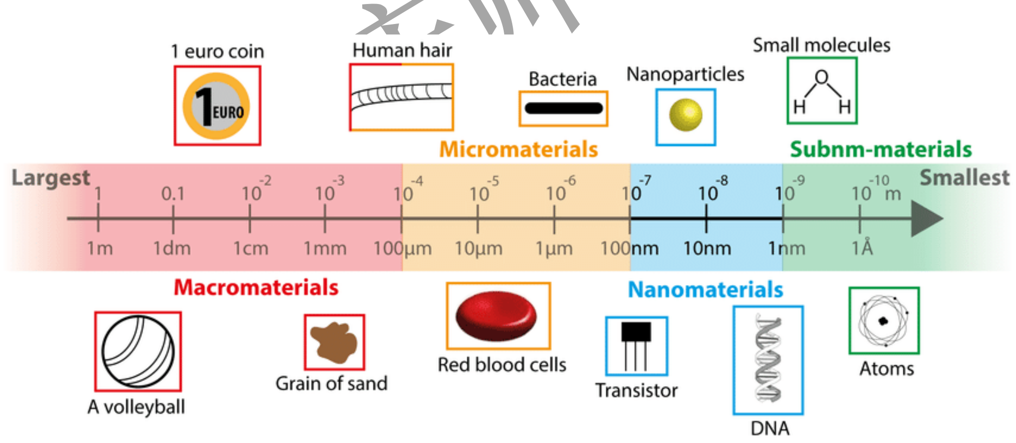
Pollen can trigger allergic reactions and hay fever—which 1 in 5 Americans experience every year.

Source: Harvard Health

The visibility limits for what the naked eye can see hovers around 10-40µm.

Respiratory droplets have the potential to carry smaller particles within them, such as dust or coronavirus.

Wildfire smoke can persist in the air for several days, and even months.



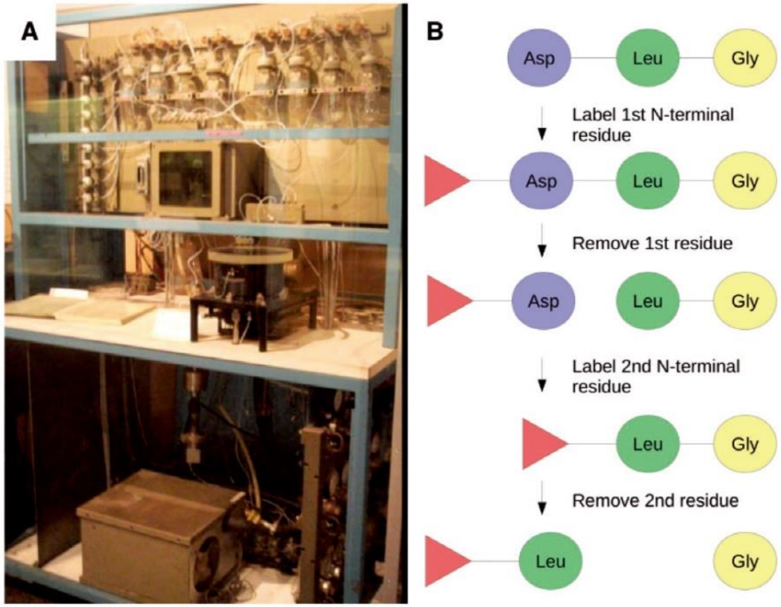
Multi-scales

from ***micro*** *to* ***macro***

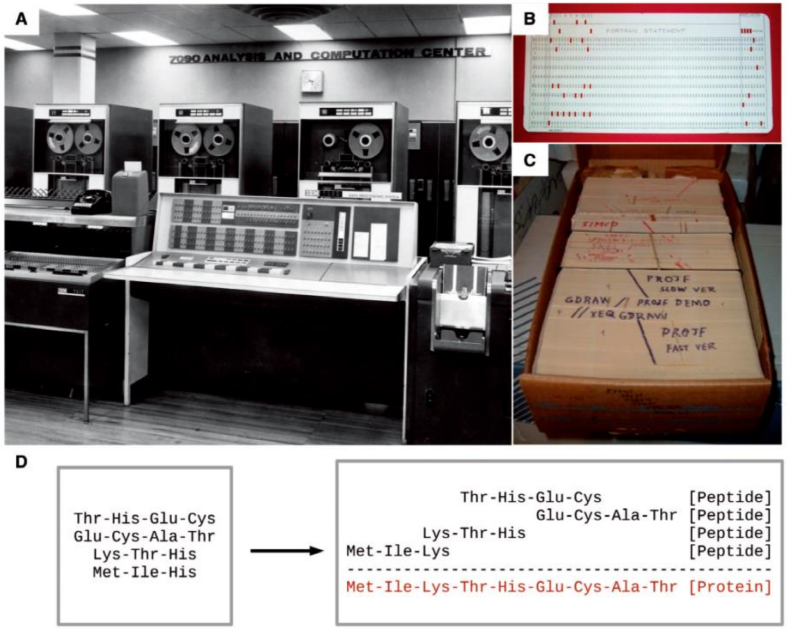
Symbol	Name	Factor	Symbol	Name	Factor
Y	yotta	10^{24}	y	yokto	10^{-24}
Z	zetta	10^{21}	z	zepto	10^{-21}
E	exa	10^{18}	a	atto	10^{-18}
P	peta	10^{15}	f	femto	10^{-15}
T	tera	10^{12}	p	pico	10^{-12}
G	giga	10^{9}	n	nano	10^{-9}
M	mega	10^{6}	µ	micro	10^{-6}
k	kilo	10^{3}	m	milli	10^{-3}
h	hecto	10^{2}	c	centi	10^{-2}
da	deka	10^{1}	d	deci	10^{-1}

❖ 1950~1970: *The origins*

- It did not start with DNA analysis
- *Protein analysis was the starting point*
- Dayhoff: the first bioinformatician
- The computer-assisted genealogy of life (生命谱系)
- A mathematical framework for amino acid substitutions



Edman Sequencing


A
2090 ANALYSIS AND COMPUTATION CENTER
B
C
D
Thr-His-Glu-Cys
Glu-Cys-Ala-Thr
Lys-Thr-His
Met-Ile-His
Thr-His-Glu-Cys [Peptide]
Glu-Cys-Ala-Thr [Peptide]
Lys-Thr-His [Peptide]
Met-Ile-Lys [Peptide]
Met-Ile-Lys-Thr-His-Glu-Cys-Ala-Thr [Protein]

COMPROTEIN

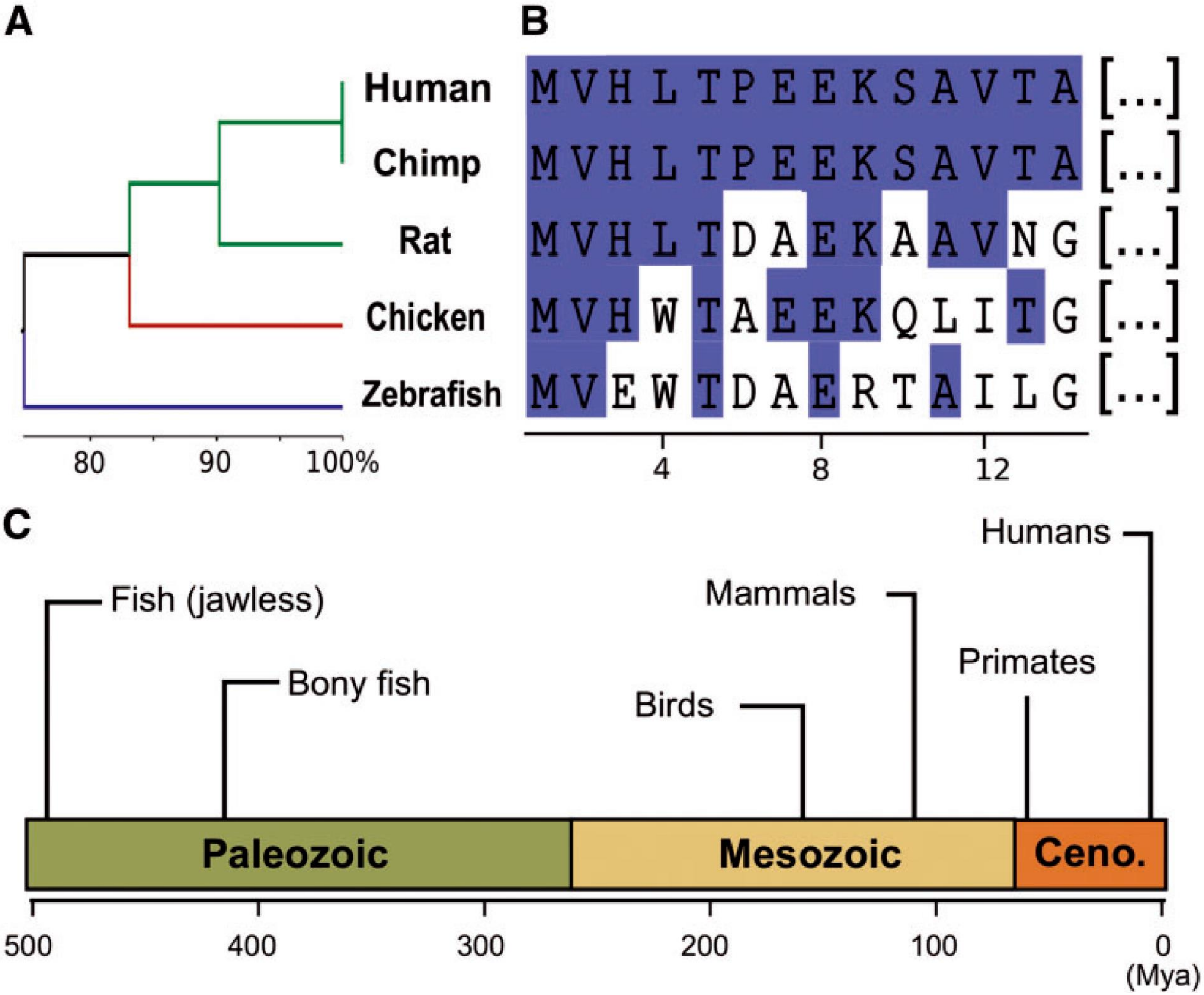
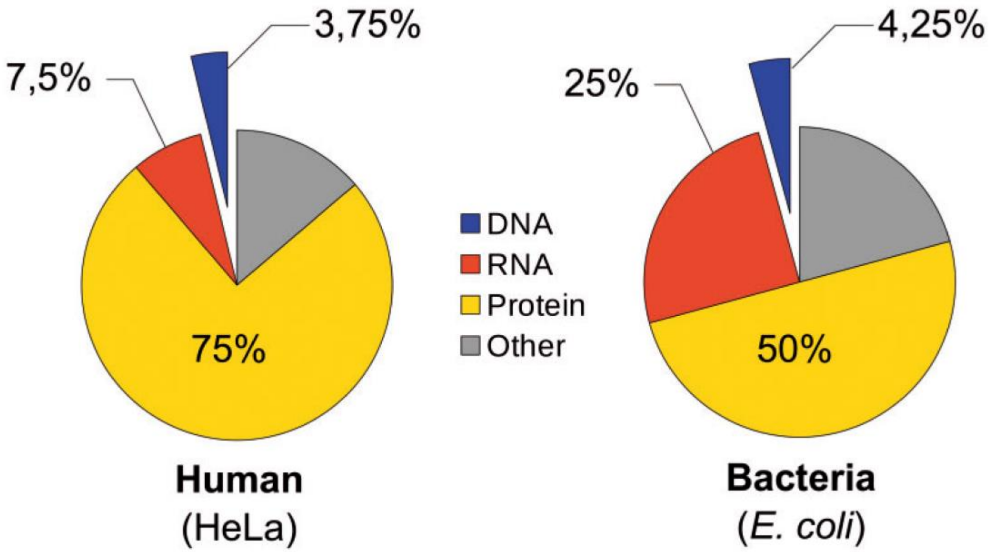


Figure 3. Sequence dissimilarity between orthologous proteins from model organisms correlates with their evolutionary history as evidenced by the fossil record. (**A**) Average distance tree of hemoglobin subunit beta-1 (HBB-1) from human (*Homo sapiens*), chimpanzee (*Pan troglodytes*), rat (*Rattus norvegicus*), chicken (*Gallus gallus*) and zebrafish (*Danio rerio*). (**B**) Alignment view of the first 14 amino acid residues of HBB-1 compared in (A) (residues highlighted in blue are identical to the human HBB-1 sequence). (**C**) Timeline of earliest fossils found for different aquatic and terrestrial animals.

❖ 1970~1980: Paradigm shift from protein to DNA analysis

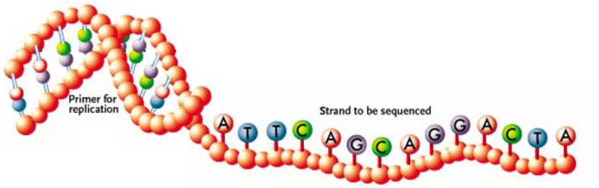
- Deciphering of the DNA language: the genetic code
- *Cost-efficient reading of DNA*
- Using DNA sequences in phylogenetic inference



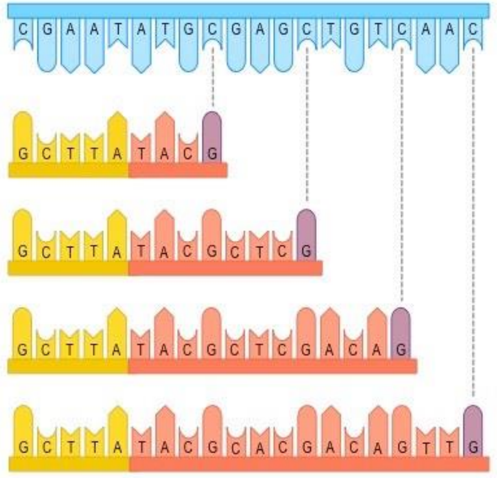
DNA is the least abundant macromolecular cell component that can be sequenced.

DNA replication,

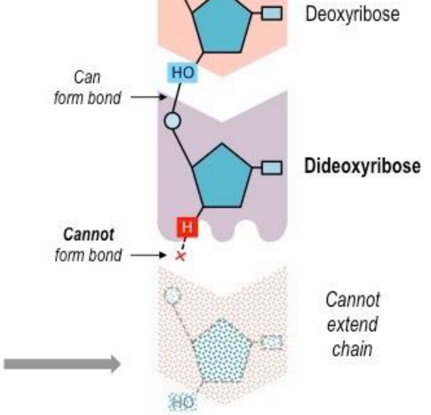
in vivo (体内) and *in vitro* (体外)



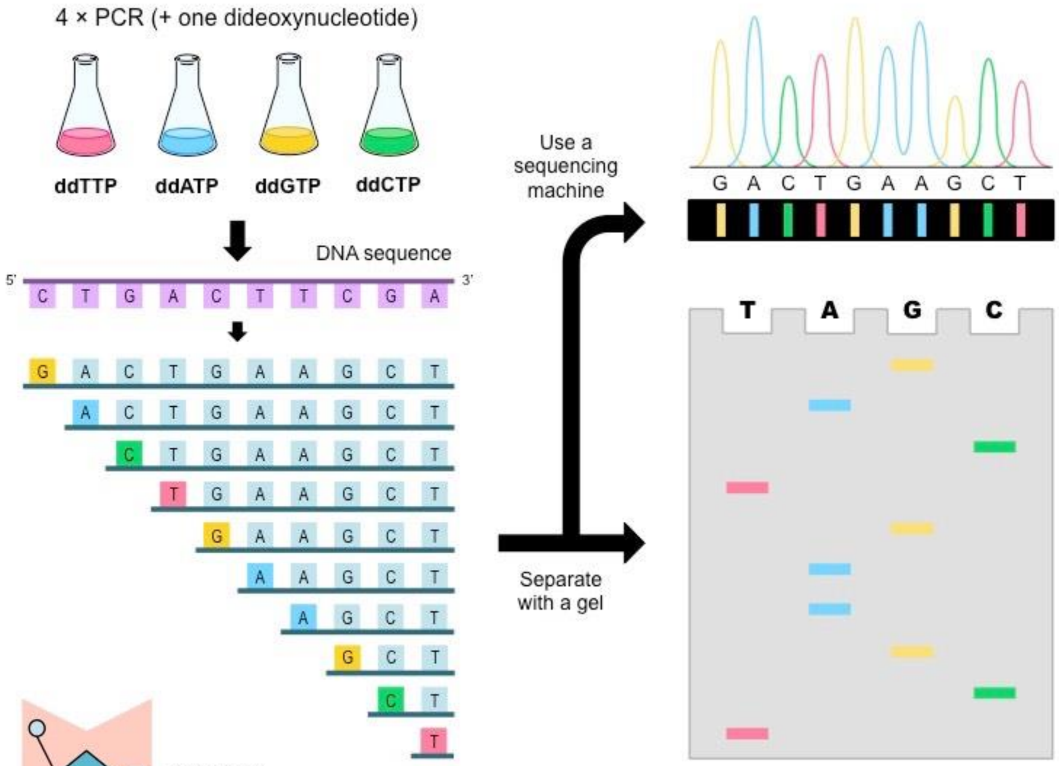
Take replicating *in vitro* as example

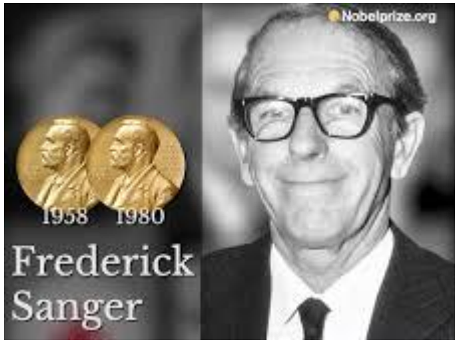

C G A A T A T G C G A G C T G T C A A C
G C T T A T A C G
G C T T A T A C G C T C G
G C T T A T A C G C T C G A C A G
G C T T A T A C G C A C G A C A G T T G

Sequence terminates when the ddNTP is incorporated
Fragment lengths reflect base position in sequence



Chain termination by dideoxynucleotides





双脱氧终止法测序原理

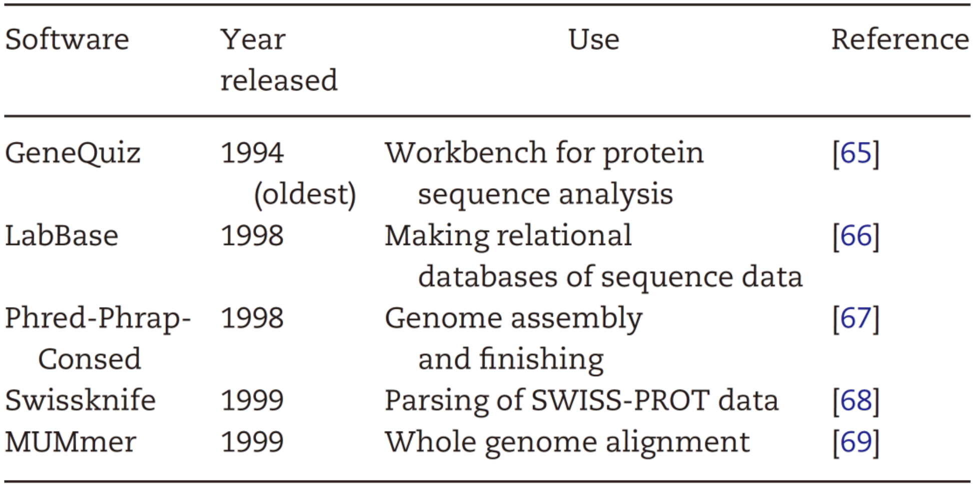
❖ 1980~1990: Parallel advances in biology and computer science

- Molecular methods to target and amplify specific genes
- *Access to computers and specialized software*
- Bioinformatics and the free software movement
- Desktop computers and new programming languages

DEC ODO-8, 1965
a 'minicomputer' fairly had the dimensions and weight of a small household refrigerator

DEC VAX-11/780 Minicomputer. From right to left: The computer module, two tape storage units, a monitor and a terminal. The GCG software package was initially designed to run on this computer.

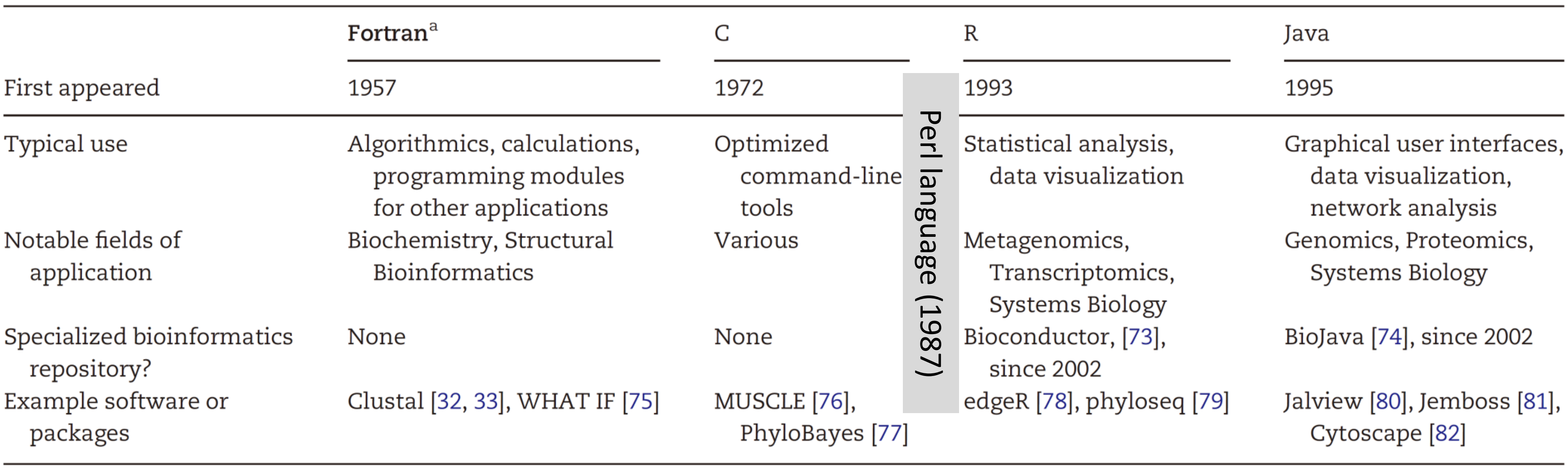
HP-9000 desktop workstation running the Unix-based system HP-UX. Image: Thomas Schanz//CC-BY-SA 3.0.



Software	Year released	Use	Reference
GeneQuiz	1994 (oldest)	Workbench for protein sequence analysis	[65]
LabBase	1998	Making relational databases of sequence data	[66]
Phred-Phrap-Consed	1998	Genome assembly and finishing	[67]
Swissknife	1999	Parsing of SWISS-PROT data	[68]
MUMmer	1999	Whole genome alignment	[69]

PubMed Key: (perl bioinformatics) AND (“1987”[Date-Publication]: “2000” [Date-Publication]).

Table 3. Notable nonscripting and/or statistical programming languages used in bioinformatics



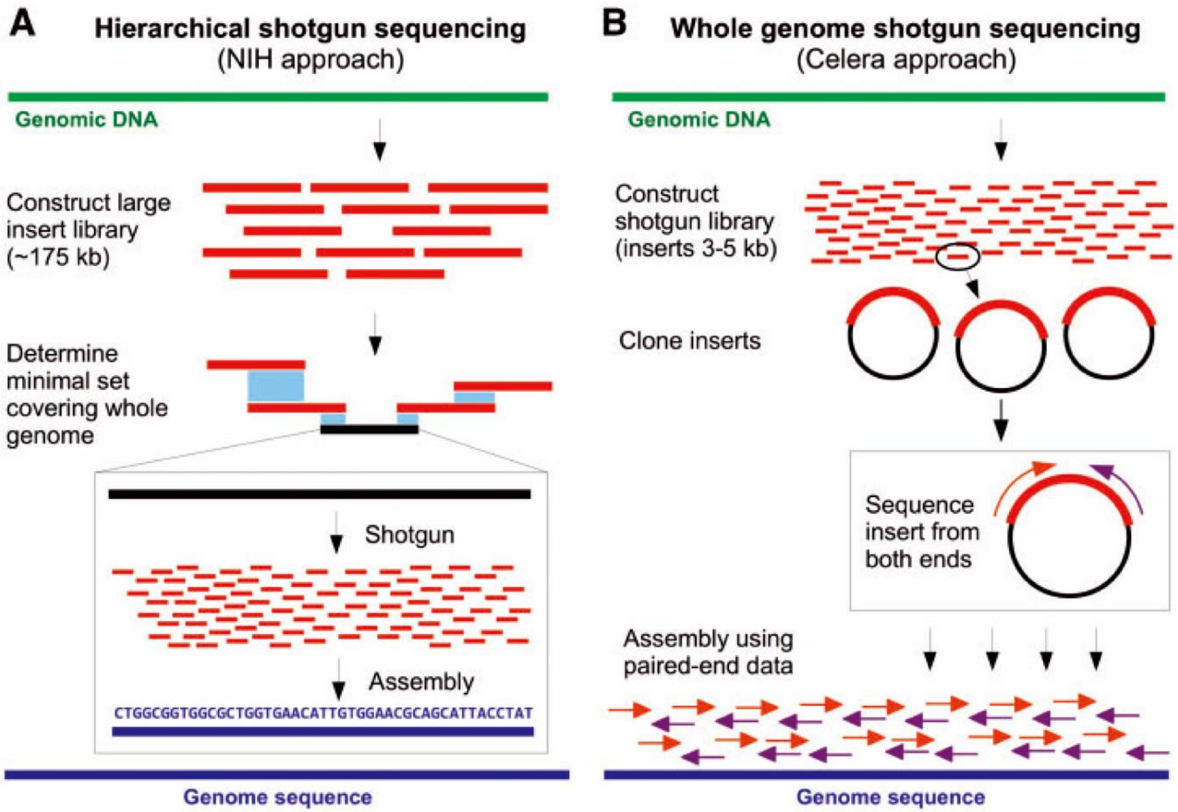
	Fortran[a]	C	R	Java
First appeared	1957	1972	1993	1995
Typical use	Algorithmics, calculations, programming modules for other applications	Optimized command-line tools	Statistical analysis, data visualization	Graphical user interfaces, data visualization, network analysis
Notable fields of application	Biochemistry, Structural Bioinformatics	Various	Metagenomics, Transcriptomics, Systems Biology	Genomics, Proteomics, Systems Biology
Specialized bioinformatics repository?	None	None	Bioconductor, [73], since 2002	BioJava [74], since 2002
Example software or packages	Clustal [32, 33], WHAT IF [75]	MUSCLE [76], PhyloBayes [77]	edgeR [78], phyloseq [79]	Jalview [80], Jemboss [81], Cytoscape [82]

Perl language (1987)

❖ 1990~2000: *Genomics, bioinformatics and the information superhighway*

In 1995, the first complete genome sequencing of a free-living organism (*Haemophilus influenzae*) was sequenced. However, the turning point that started the genomic era, as we know it actually, was the publication of the human genome at the beginning of the 21st century.

- Dawn of the genomics era
- *Bioinformatics online*
- *Beyond sequence analysis: structural bioinformatics*

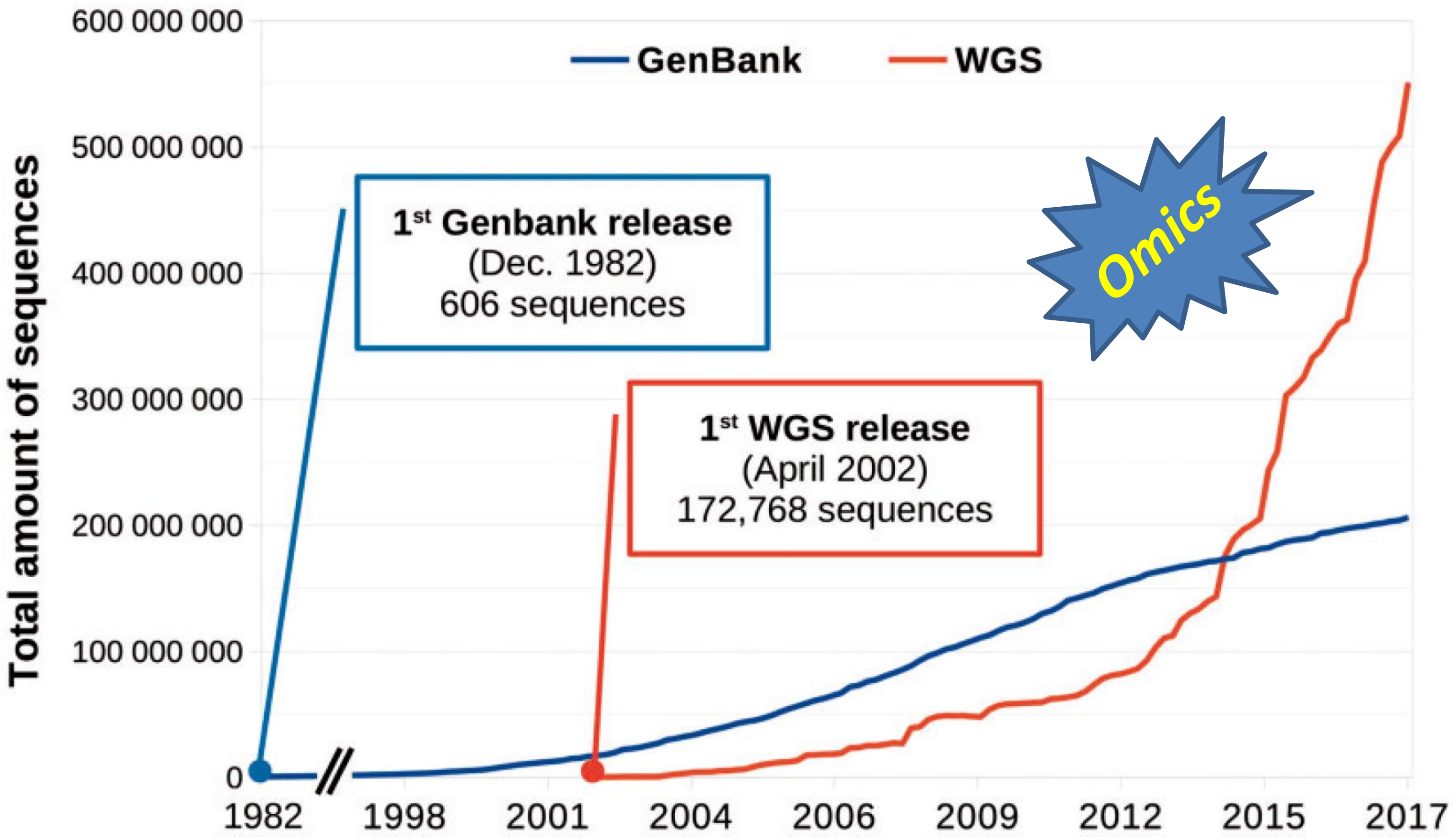


A
Hierarchical shotgun sequencing
(NIH approach)
Genomic DNA
Construct large insert library (~175 kb)
Determine minimal set covering whole genome
Shotgun
Assembly
CTGGCGGTGGCGCTGGTGAACATTGTGGAACGCAGCATTACCTAT
Genome sequence
B
Whole genome shotgun sequencing
(Celera approach)
Genomic DNA
Construct shotgun library (inserts 3-5 kb)
Clone inserts
Sequence insert from both ends
Assembly using paired-end data
Genome sequence

Hierarchical shotgun sequencing **versus** ***whole genome shotgun sequencing*****. Both approaches respectively exemplified the methodological rivalry between the public (*NIH, A*) and private (*Celera, B*) efforts to sequence the human genome.**

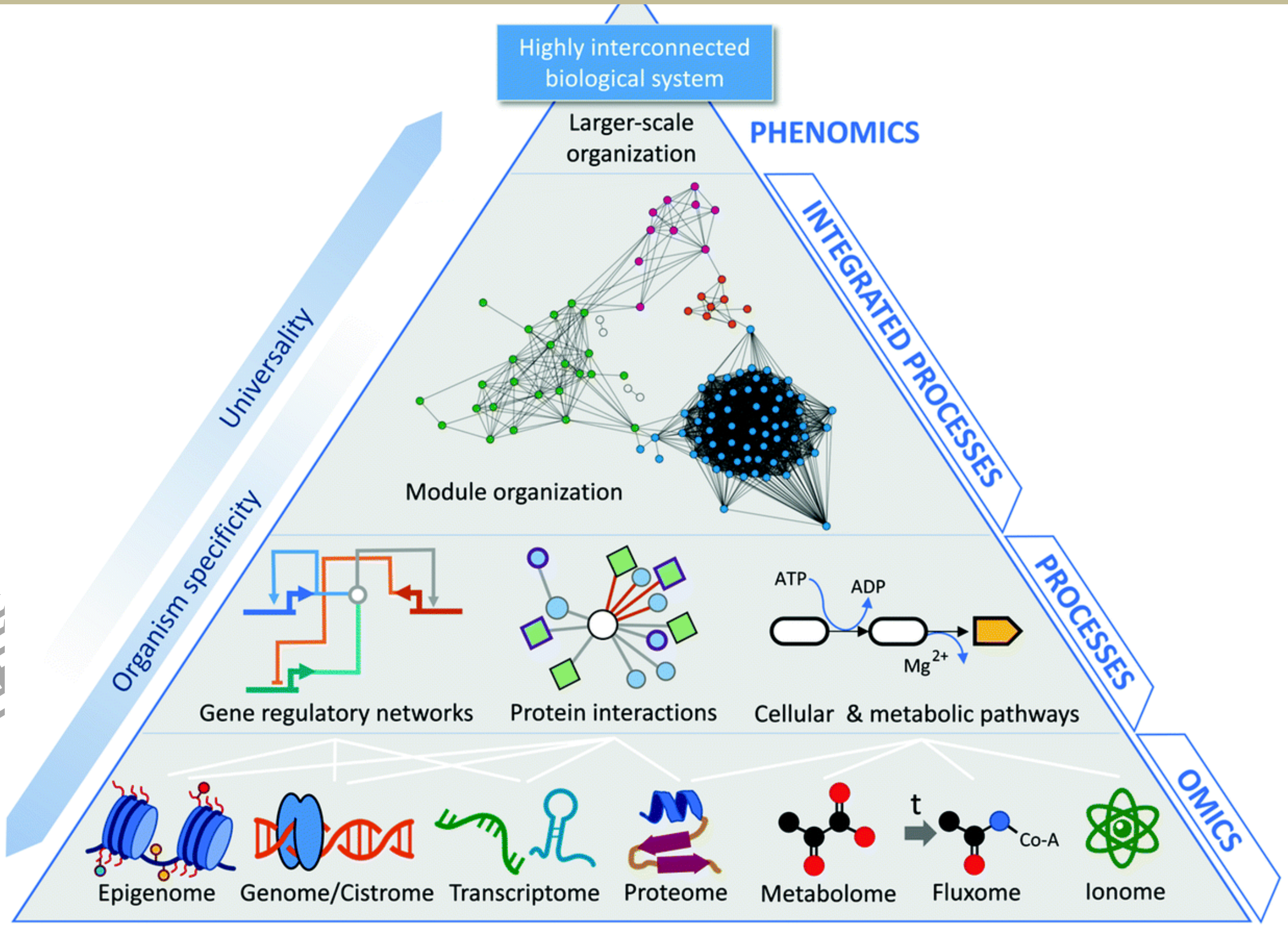
❖ 2000~2010: High-throughput bioinformatics

- Second-generation sequencing
- *Biological Big Data*
- *High-performance bioinformatics and collaborative computing*



❖ 2010~Today: *Present and future perspectives*

- ◆ Clearly defining the bioinformatician profession
- ◆ Is the term ‘bioinformatics’ now obsolete (过时的)?
- ◆ *Towards modeling life as a whole: systems biology*



Indeed, the use of computers has become ubiquitous in biology, as well as in most natural sciences (physics, chemistry, mathematics, cryptography, etc.),

but interestingly,

only biology has a specific term to refer to the use of computers in this discipline *('bioinformatics')*.

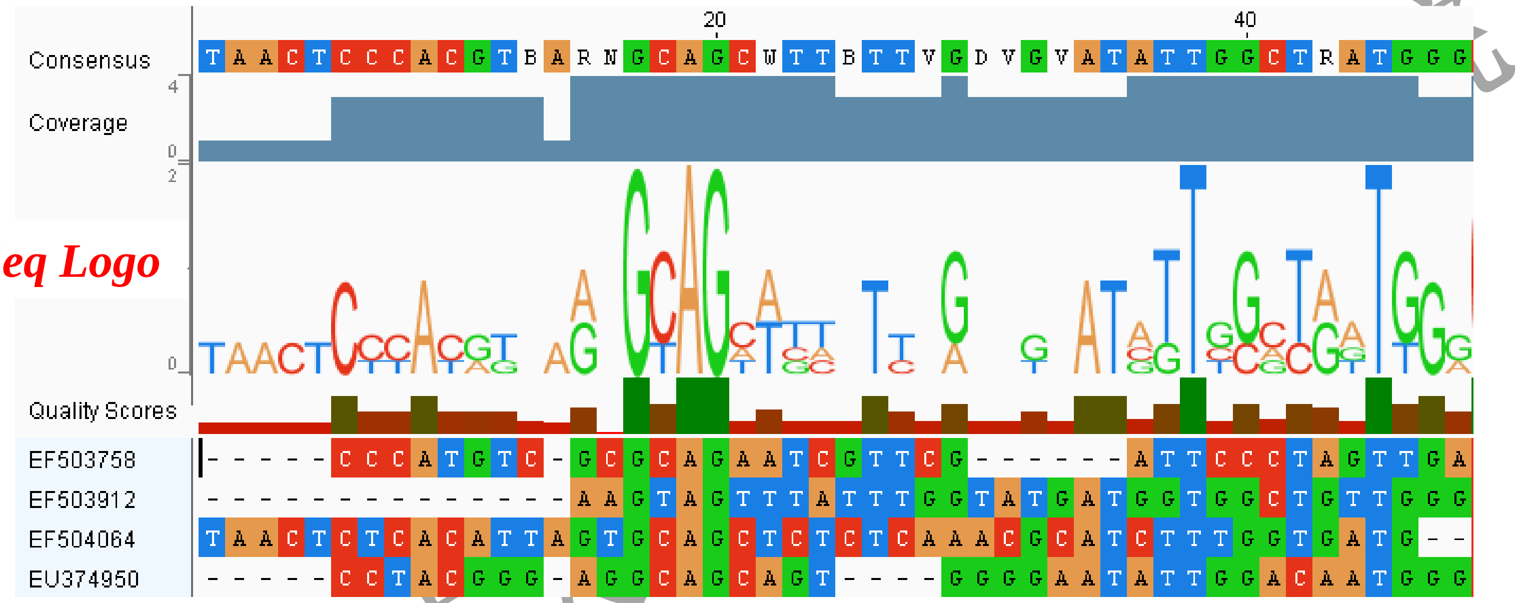
Why is that so?

第4节：生物信息学在生命科学中的应用

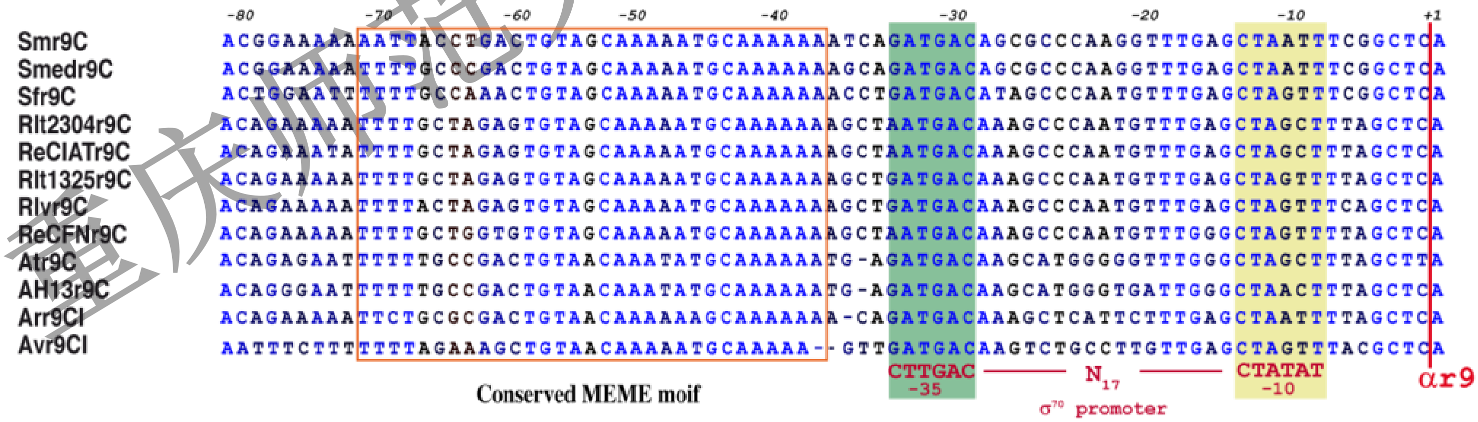
重要知识点

- ✓ 序列比对、装配
- ✓ 基于预测
- ✓ 多态性
- ✓ RNA表达分析
- ✓ 分子进化
- ✓ 结构预测
- ✓ 分子间相互作用

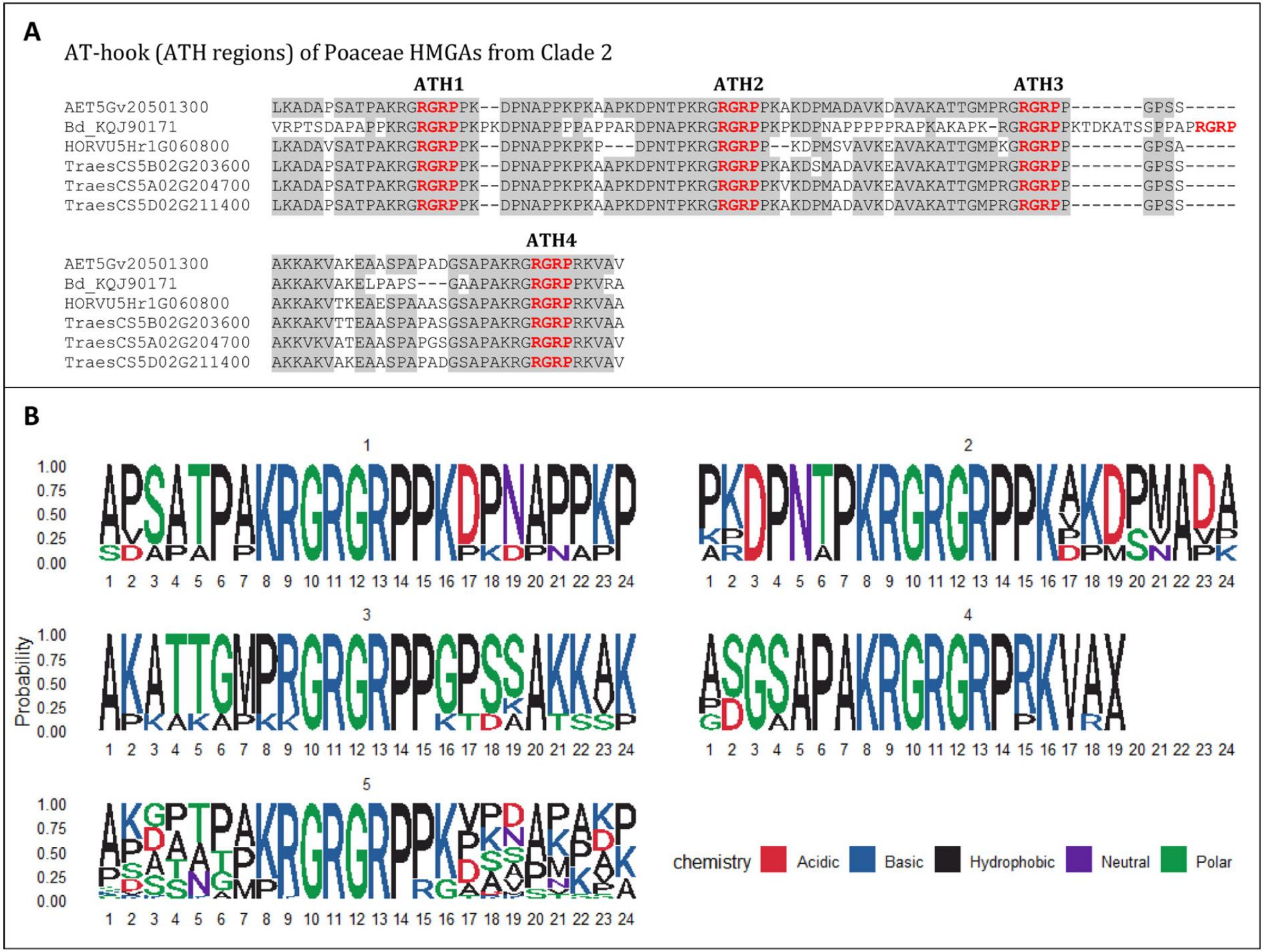
❖ 4.1 序列比对



Seq Logo



意义： 寻找保守区，酶切位点，重要基序，进化分析等



A

AT-hook (ATH regions) of Poaceae HMGAs from Clade 2

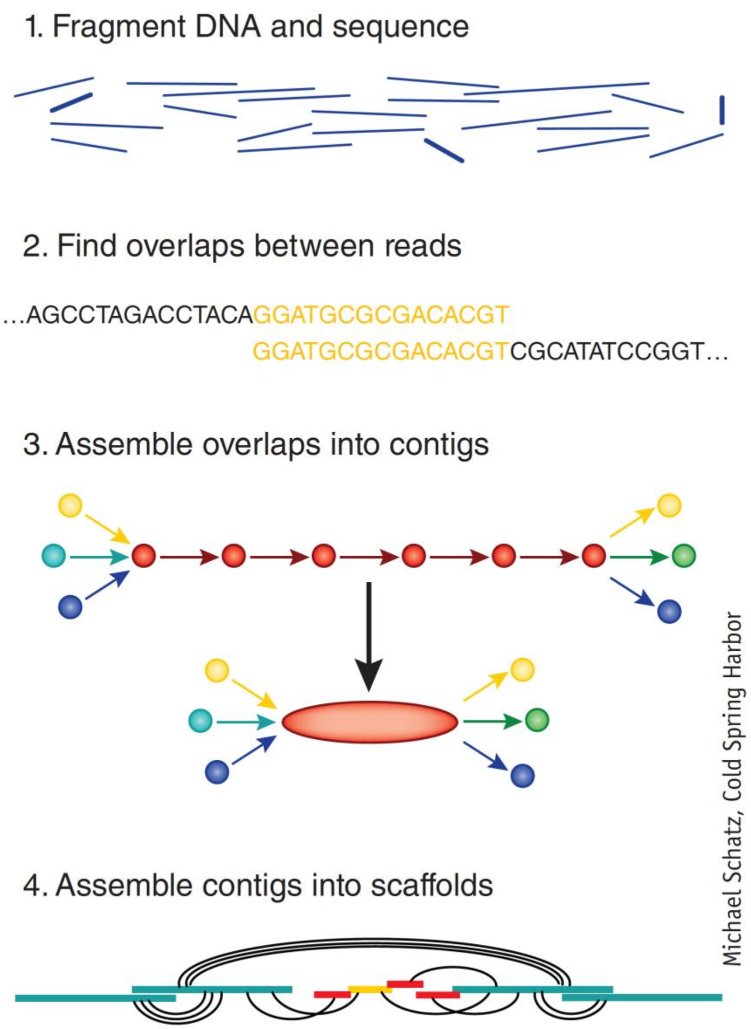
```
                                    ATH1                            ATH2                            ATH3
AET5Gv20501300       LKADAPSATPAKRGRGRPPK--DPNAPPKPKAAPKDPNTPKRGRGRPPKAKDPMADAVKDAVAKATTGMPRGRGRPP-------GPSS-----
Bd_KQJ90171          VRPTSDAPAPPKRGRGRPPKPKDPNAPPPPAPPARDPNAPKRGRGRPPKPKDPNAPPPPPRAPKAKAPK-RGRGRPPKTDKATSSPPAPRGRP
HORVU5Hr1G060800     LKADAVSATPAKRGRGRPPK--DPNAPPKPKP---DPNTPKRGRGRPP--KDPMSVAVKEAVAKATTGMPKGRGRPP-------GPSA-----
TraesCS5B02G203600   LKADAPSATPAKRGRGRPPK--DPNAPPKPKAAPKDPNTPKRGRGRPPKAKDSMADAVKEAVAKATTGMPRGRGRPP-------GPSS-----
TraesCS5A02G204700   LKADAPSATPAKRGRGRPPK--DPNAPPKPKAAPKDPNTPKRGRGRPPKVKDPMADAVKEAVAKATTGMPRGRGRPP-------GPSS-----
TraesCS5D02G211400   LKADAPSATPAKRGRGRPPK--DPNAPPKPKAAPKDPNTPKRGRGRPPKAKDPMADAVKDAVAKATTGMPRGRGRPP-------GPSS-----

                                          ATH4
AET5Gv20501300       AKKAKVAKEAASPAPADGSAPAKRGRGRPRKVAV
Bd_KQJ90171          AKKAKVAKELPAPS---GAAPAKRGRGRPPKVRA
HORVU5Hr1G060800     AKKAKVTKEAESPAAASGSAPAKRGRGRPRKVAA
TraesCS5B02G203600   AKKAKVTTEAASPAPASGSAPAKRGRGRPRKVAA
TraesCS5A02G204700   AKKVKVATEAASPAPGSGSAPAKRGRGRPRKVAV
TraesCS5D02G211400   AKKAKVAKEAASPAPADGSAPAKRGRGRPRKVAV
```

B

Ref: Negi A P, *et al*. ***Computational biology and chemistry***, 2020, 87: 107306.

❖ 4.2 序列装配（组装）

Ref: Baker M. ***Nature methods***, 2012, 9(4): 333-337.

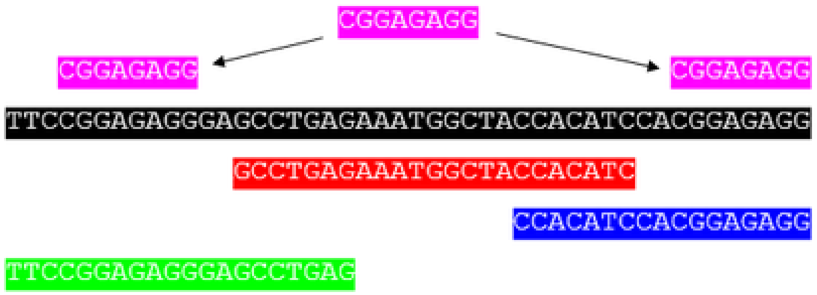

1. Fragment DNA and sequence
2. Find overlaps between reads
...AGCCTAGACCTACAGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGT...
3. Assemble overlaps into contigs
4. Assemble contigs into scaffolds
Michael Schatz, Cold Spring Harbor

Genome assembly stitches together a genome from short sequenced pieces of DNA.

De novo assembly

- **Genome Assembly** - Create new reference ‘from scratch’
- Examine reads for overlapping sequence
- **Contig** - longer assembled sequence from short reads
- **Scaffold** - assembled contigs
- **Chromosome** - assembled scaffolds
- Assembly from short reads is hard

Basic Principle


CGGAGAGG

CGGAGAGG

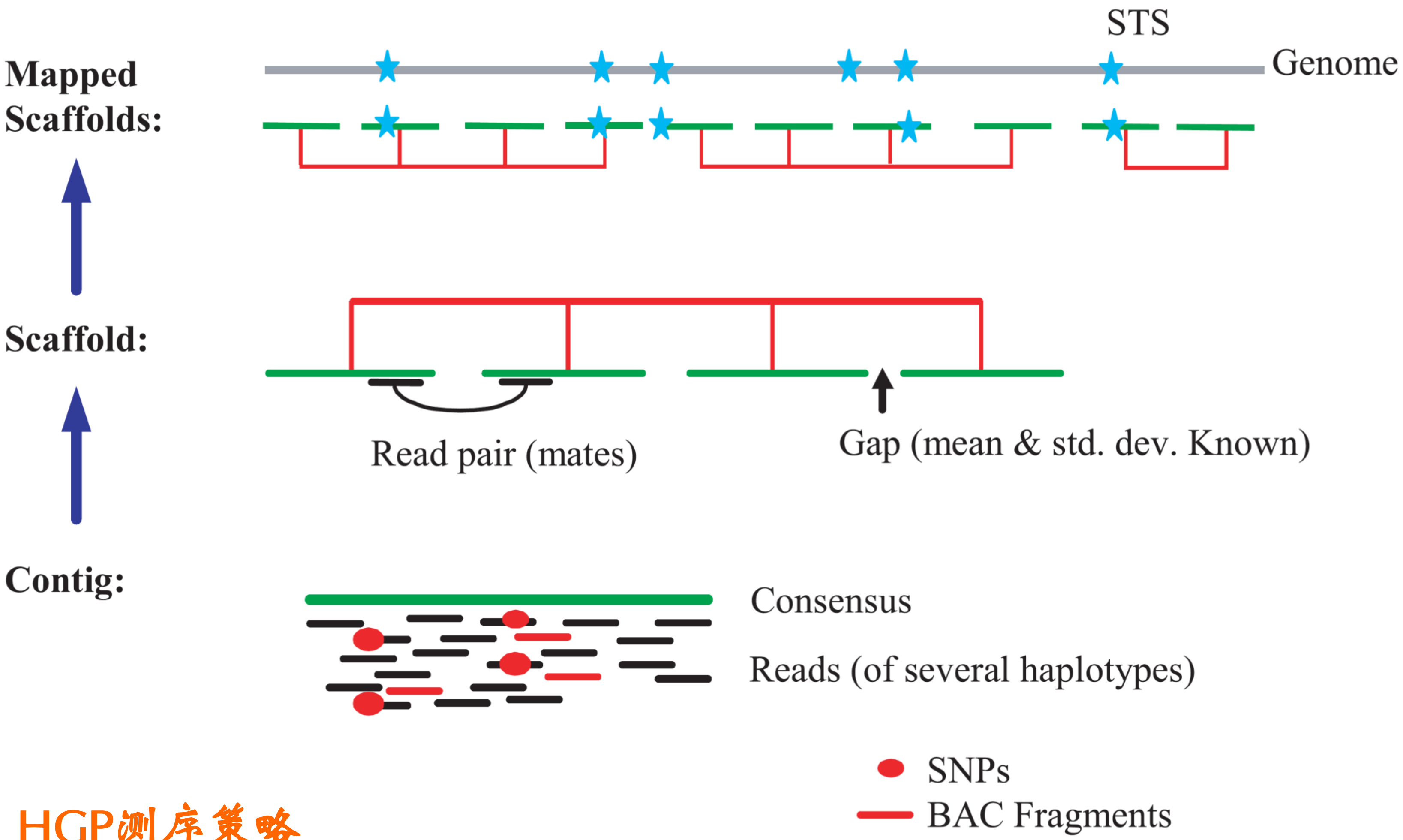
CGGAGAGG

TTCCGGAGAGGGAGCCTGAGAAATGGCTACCACATCCACGGAGAGG

GCCTGAGAAATGGCTACCACATC

CCACATCCACGGAGAGG

TTCCGGAGAGGGAGCCTGAG

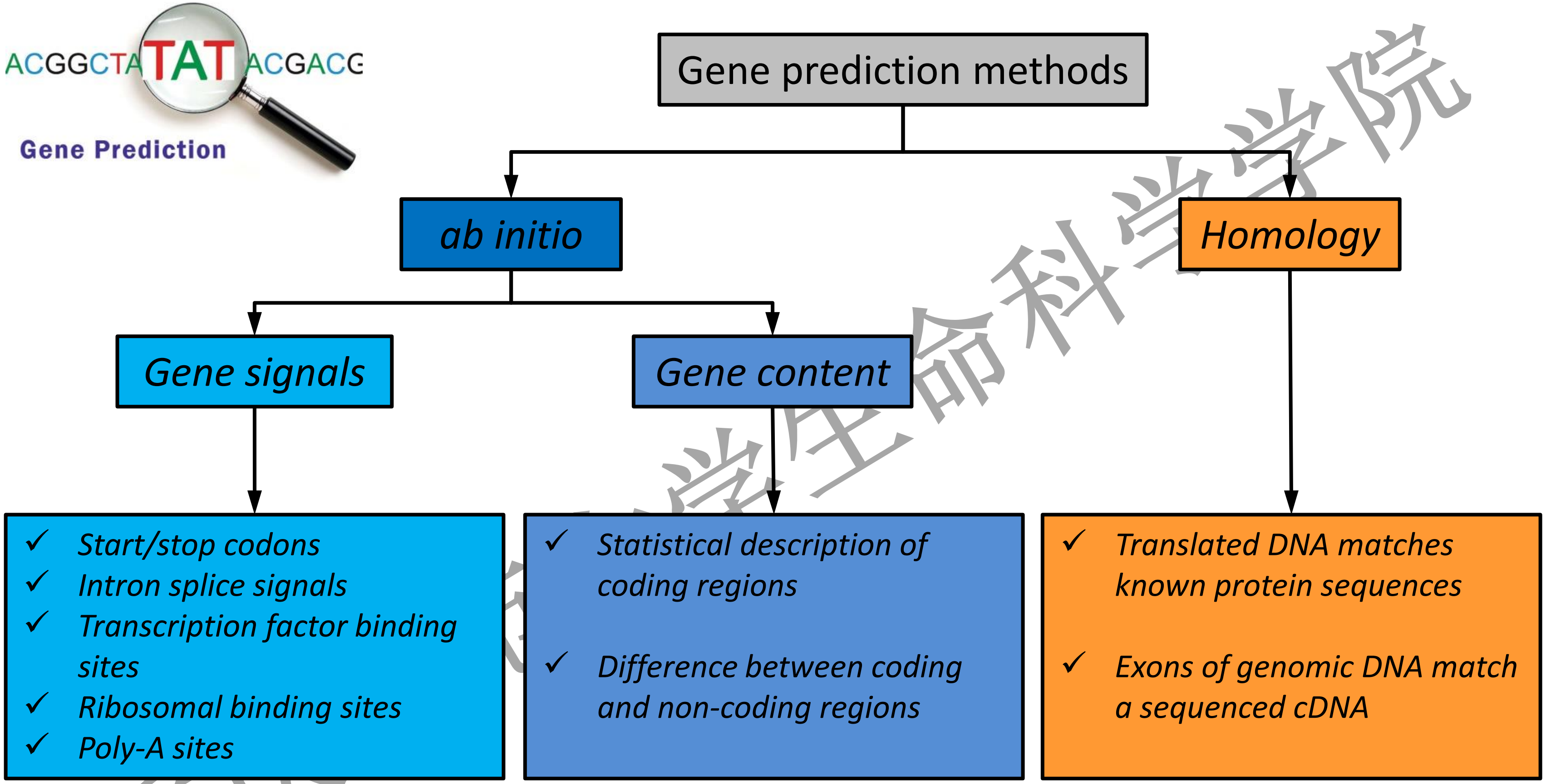


HGP测序策略

Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

（read/村 — Contig/镇 — Scaffold/县 — Chromosome/省 — Genome/国家）

❖ 4.3 基因识别（预测）



- ✓ Start/stop codons
- ✓ Intron splice signals
- ✓ Transcription factor binding sites
- ✓ Ribosomal binding sites
- ✓ Poly-A sites

- ✓ Statistical description of coding regions
- ✓ Difference between coding and non-coding regions

- ✓ Translated DNA matches known protein sequences
- ✓ Exons of genomic DNA match a sequenced cDNA

Statistical approaches

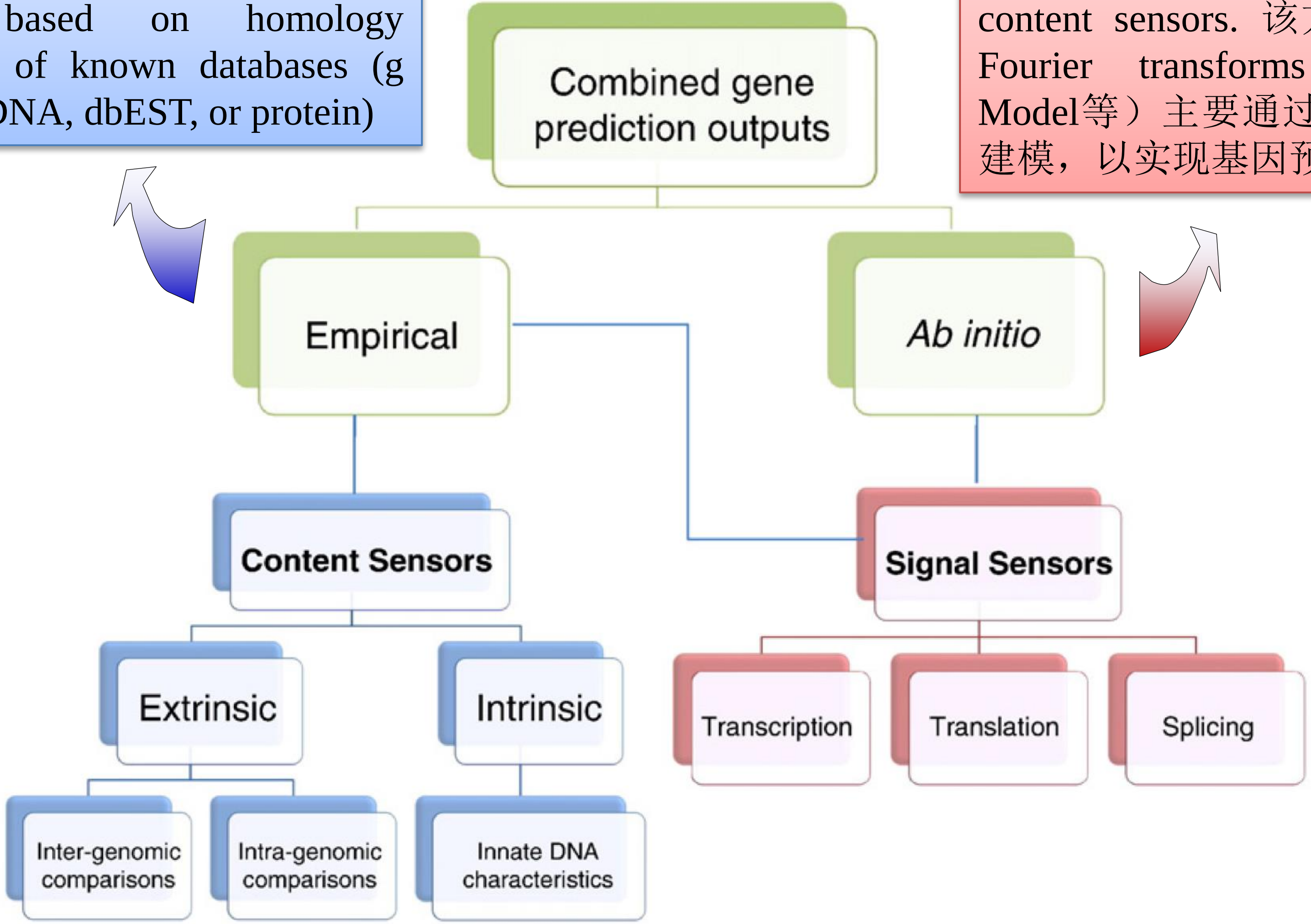
- *Exploit statistical characteristics of coding regions and non-coding regions and other knowledge about gens*
- *Can be potentially detect new genes*
- *May not be reliable*

Similarity approaches

- *Exploit fact that many genes are conserved across species*
- *Can be highly reliable*
- *Good at finding known genes*

Empirical gene predictors, also referred to as **sequence similarity based gene-finders**, identify genes based on homology searches of known databases (g DNA, cDNA, dbEST, or protein)

Ab initio (or *de novo*) gene-finders rely on sequence information afforded by both signal and content sensors. 该方法（如NN、Fourier transforms 、 Markov Model等）主要通过对基因结构建模，以实现基因预测。



Combined gene prediction outputs

Empirical

Ab initio

Content Sensors

Signal Sensors

Extrinsic

Intrinsic

Transcription

Translation

Splicing

Inter-genomic comparisons

Intra-genomic comparisons

Innate DNA characteristics

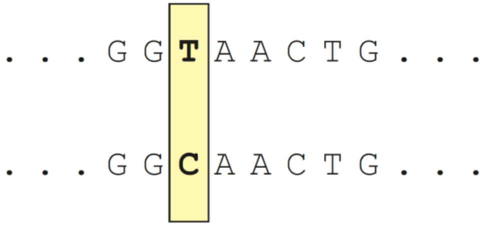
Fig. Schematic overview of eukaryote gene prediction methods and the underling sensors routinely used to locate genes in genomic sequences

Ref: Sleator R D. ***Gene***, 2010, 461(1-2): 1-4.

❖ 4.4 基因多态性分析（如SNP）

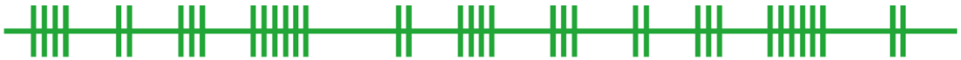
What is an SNP?

Different people can have a
different nucleotide or base at
a given location on a chromosome



What is an SNP map?

Location of SNPs
on human DNA

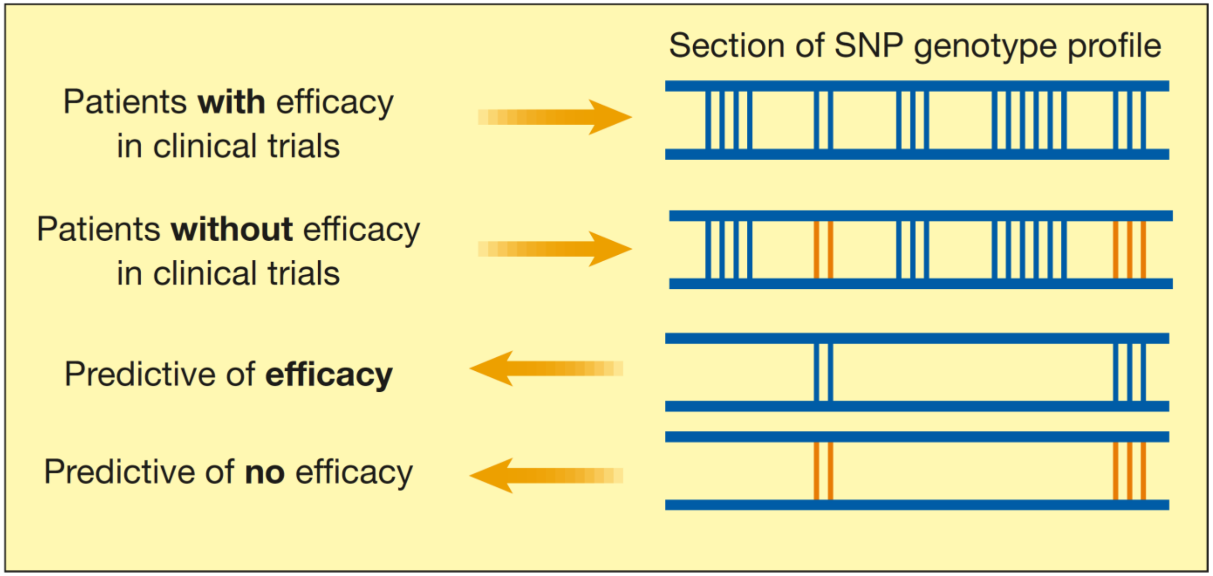


Human
DNA

How can an SNP map be used to predict medicine response?

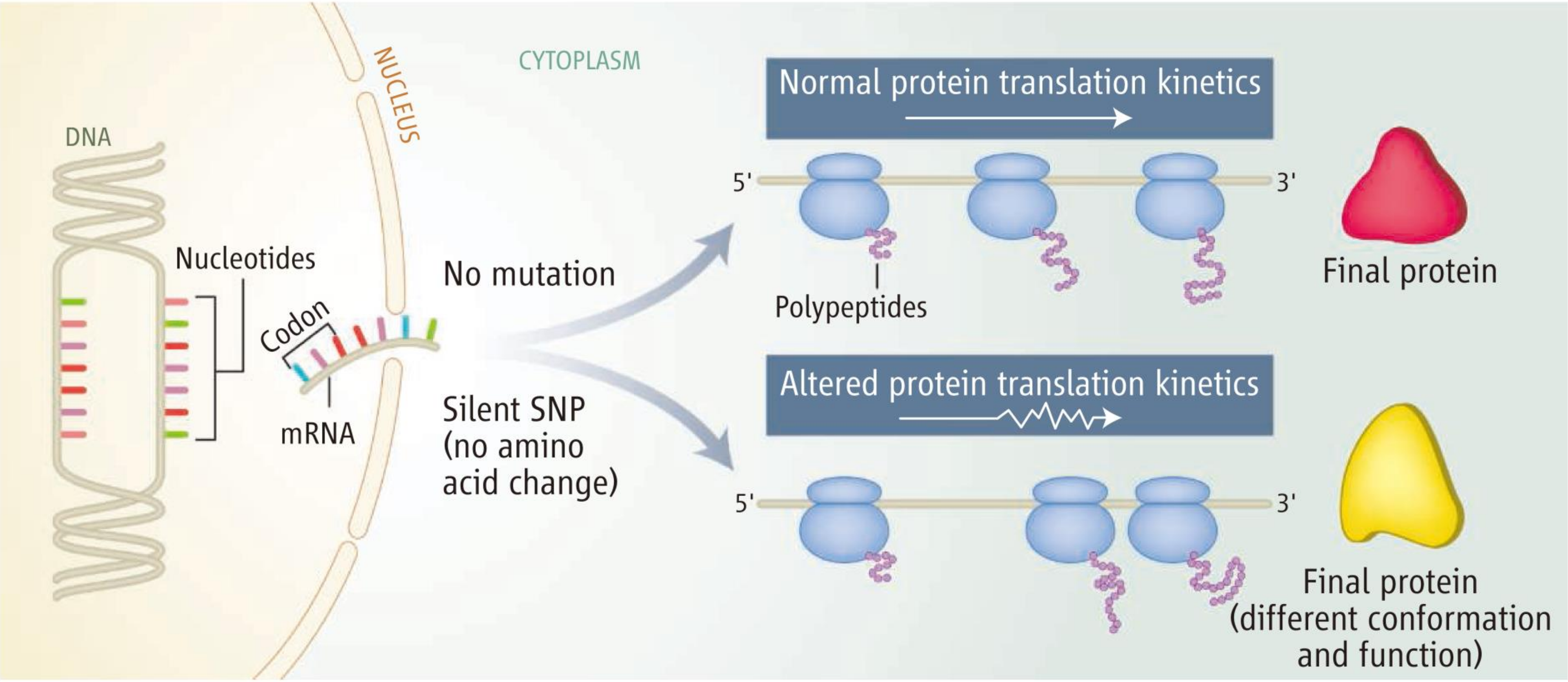
Notes:

- ***Once in every 1,000 nucleotides*** *on average;*
- *There are roughly* ***4~5 million SNPs in a person's genome****;*
- *Scientists have found* ***more than 100 million SNPs in populations*** *around the world.*


Section of SNP genotype profile
Patients **with** efficacy in clinical trials
Patients **without** efficacy in clinical trials
Predictive of **efficacy**
Predictive of **no** efficacy

Roses A D. Pharmacogenetics and the practice of medicine[J]. Nature, 2000, 405(6788): 857-865.

➢ *SNP and Protein*


CYTOPLASM
NUCLEUS
DNA
Nucleotides
Codon
mRNA
No mutation
Silent SNP (no amino acid change)
Normal protein translation kinetics
5'
3'
Polypeptides
Final protein
Altered protein translation kinetics
5'
3'
Final protein (different conformation and function)

Translation kinetics and protein folding. Unaffected translation kinetics results in a correctly folded protein. Abnormal translation kinetics, caused by the ribosome moving faster or slower through certain mRNA regions, can produce a different final protein conformation. Abnormal kinetics may arise from a silent single nucleotide polymorphism (SNP) in a gene that creates a codon synonymous to the wild-type codon. However, this synonymous codon substitution may lead to different kinetics of mRNA (protein) translation, thus yielding a protein with a different final structure and function.

Komar A A. SNPs, silent but not invisible[J]. Science, 2006.

❖ 4.5 细胞内RNA表达分析

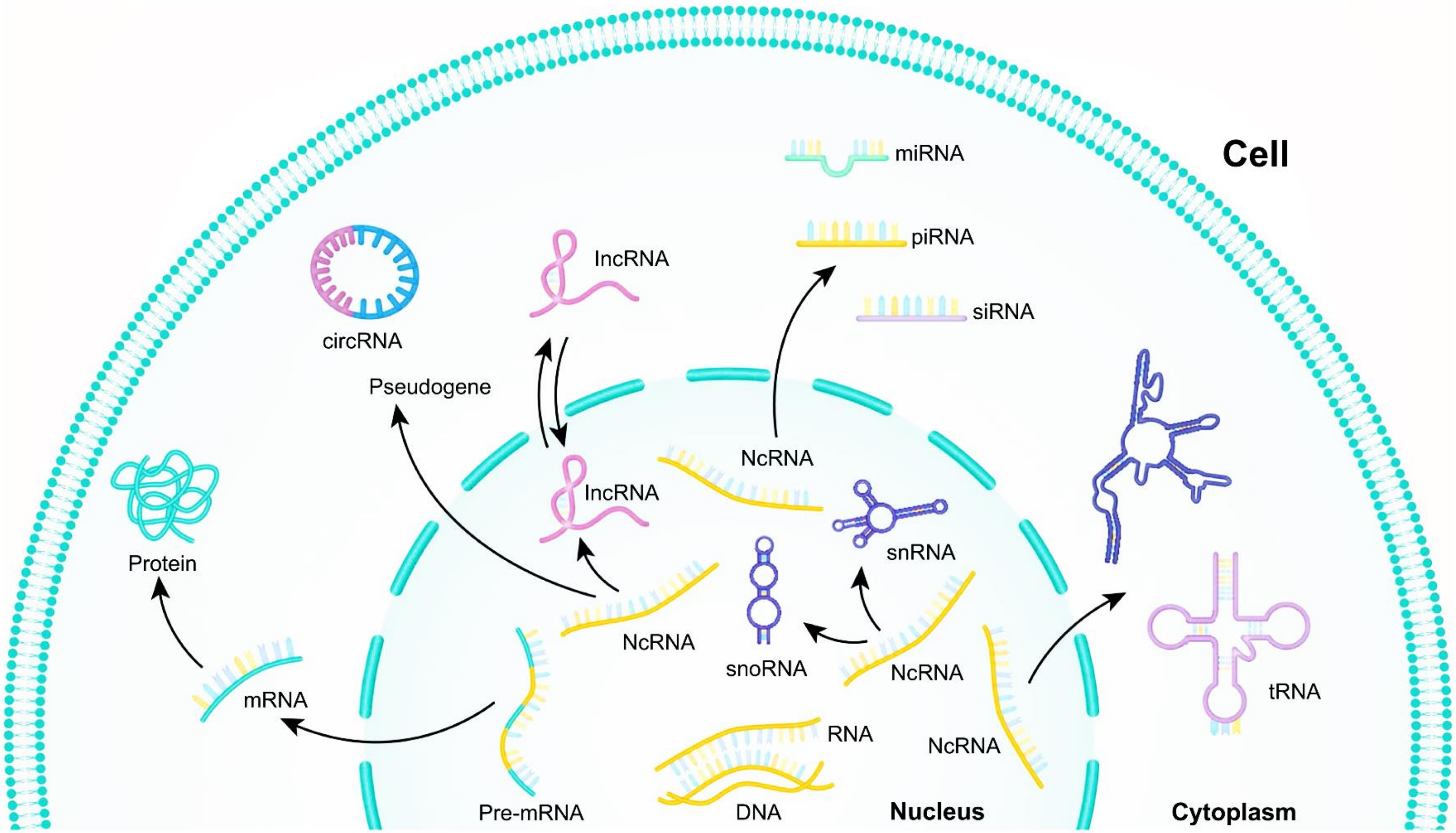
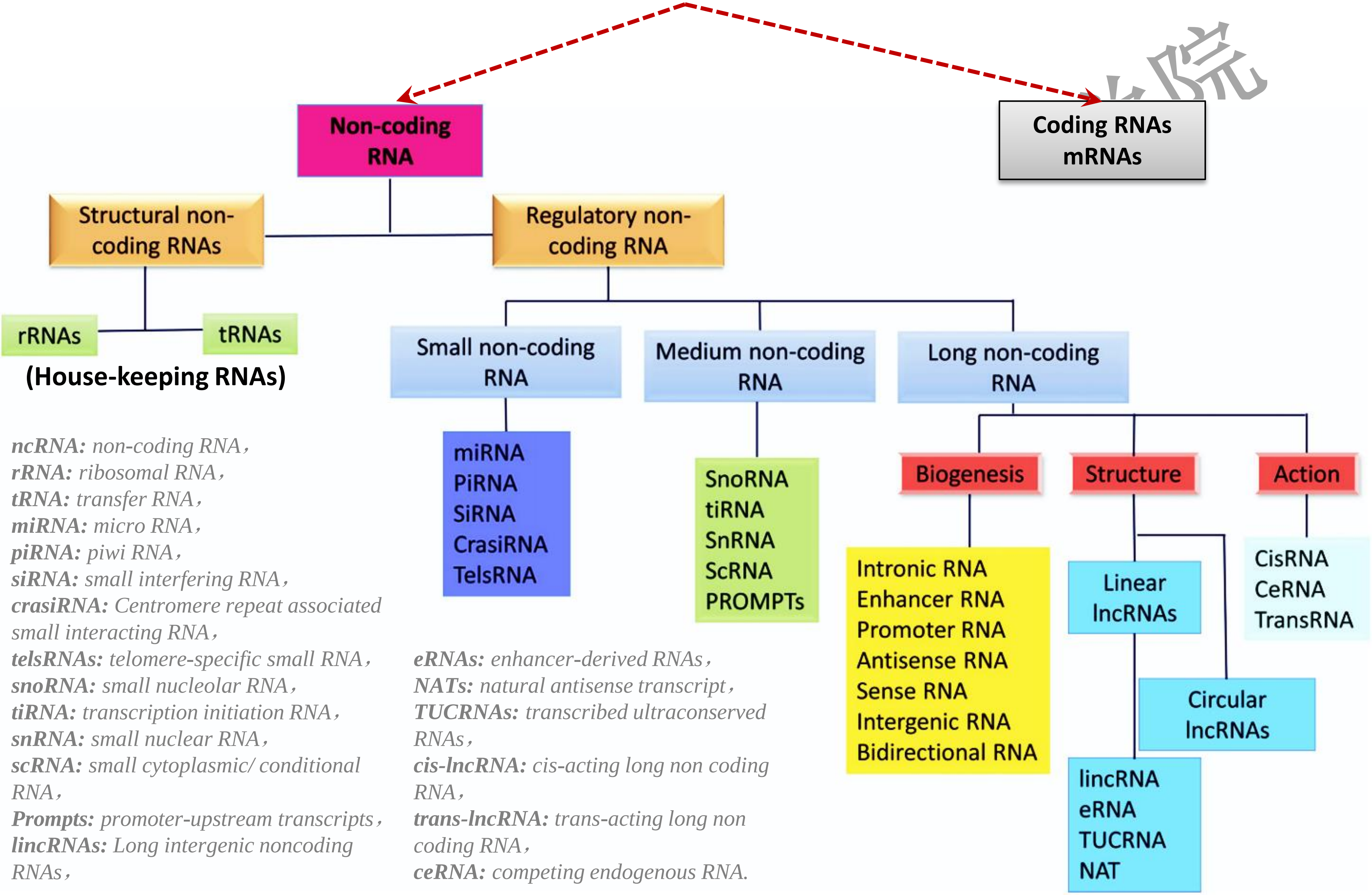


Fig. 1 The classification of coding and noncoding RNA. Eukaryotic mRNA molecules are usually composed of small segments of the original gene and are generated by a process of cleavage and rejoining from an original precursor RNA (pre-mRNA) molecule, which is an exact copy of the gene. Noncoding RNA (ncRNA) mainly include long non-coding RNA (lncRNA), microRNA (miRNA), pseudogene, circular RNA (circRNA), small interfering RNA (siRNA), piwi-interacting RNA (piRNA), small nucleolar RNA (snoRNA), small nuclear RNA (snRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA)

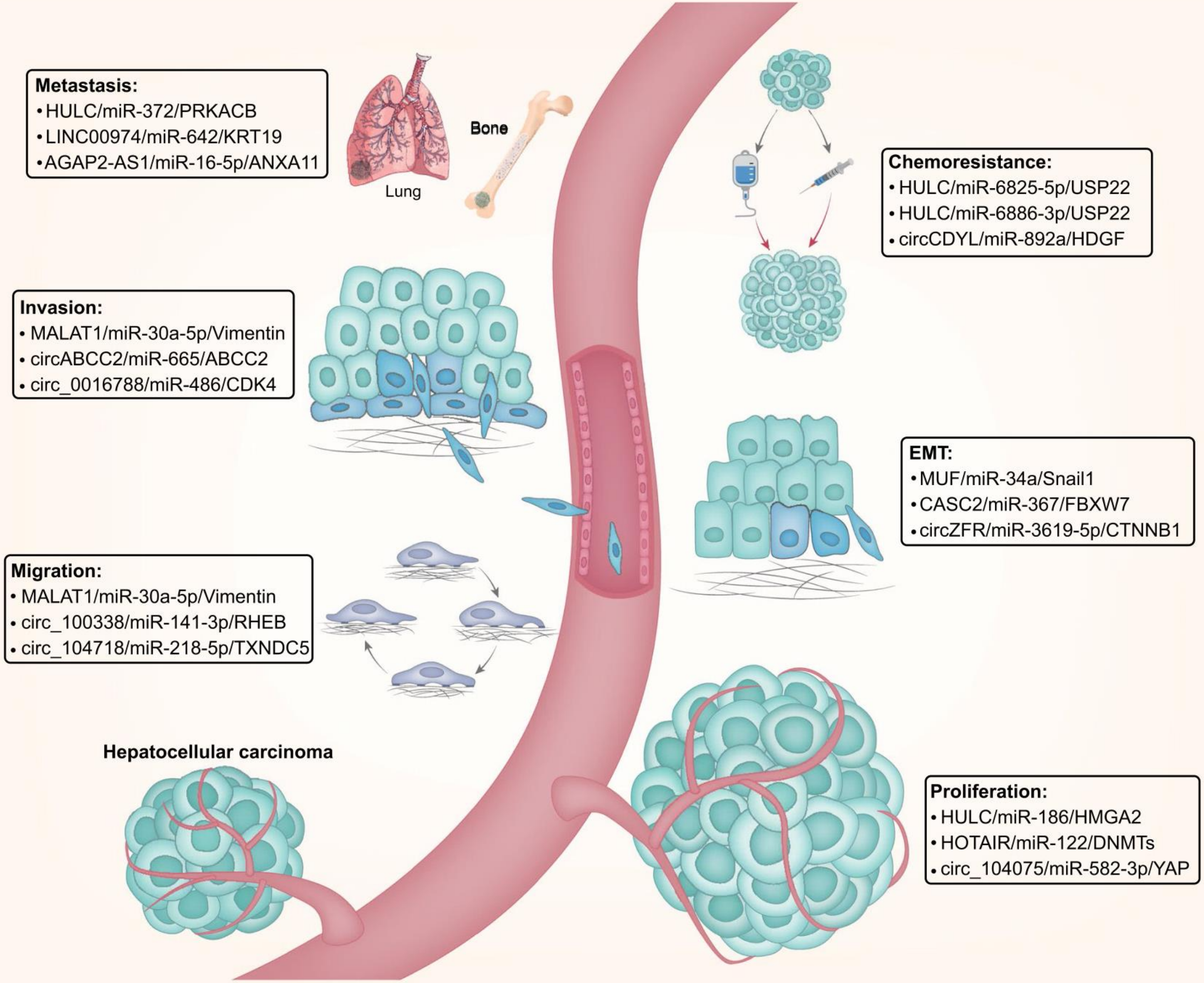
Xu G, Xu W Y, Xiao Y, et al. The emerging roles of non-coding competing endogenous RNA in hepatocellular carcinoma[J]. Cancer Cell International, 2020, 20(1): 1-21.

The World of RNAs



ncRNA: *non-coding RNA，*
rRNA: *ribosomal RNA，*
tRNA: *transfer RNA，*
miRNA: *micro RNA，*
piRNA: *piwi RNA，*
siRNA: *small interfering RNA，*
crasiRNA: *Centromere repeat associated small interacting RNA，*
telsRNAs: *telomere-specific small RNA，*
snoRNA: *small nucleolar RNA，*
tiRNA: *transcription initiation RNA，*
snRNA: *small nuclear RNA，*
scRNA: *small cytoplasmic/ conditional RNA，*
Prompts: *promoter-upstream transcripts，*
lincRNAs: *Long intergenic noncoding RNAs，*

eRNAs: *enhancer-derived RNAs，*
NATs: *natural antisense transcript，*
TUCRNAs: *transcribed ultraconserved RNAs，*
cis-lncRNA: *cis-acting long non coding RNA，*
trans-lncRNA: *trans-acting long non coding RNA，*
ceRNA: *competing endogenous RNA.*



Metastasis:

- HULC/miR-372/PRKACB
- LINC00974/miR-642/KRT19
- AGAP2-AS1/miR-16-5p/ANXA11

Chemoresistance:

- HULC/miR-6825-5p/USP22
- HULC/miR-6886-3p/USP22
- circCDYL/miR-892a/HDGF

Invasion:

- MALAT1/miR-30a-5p/Vimentin
- circABCC2/miR-665/ABCC2
- circ_0016788/miR-486/CDK4

EMT:

- MUF/miR-34a/Snail1
- CASC2/miR-367/FBXW7
- circZFR/miR-3619-5p/CTNNB1

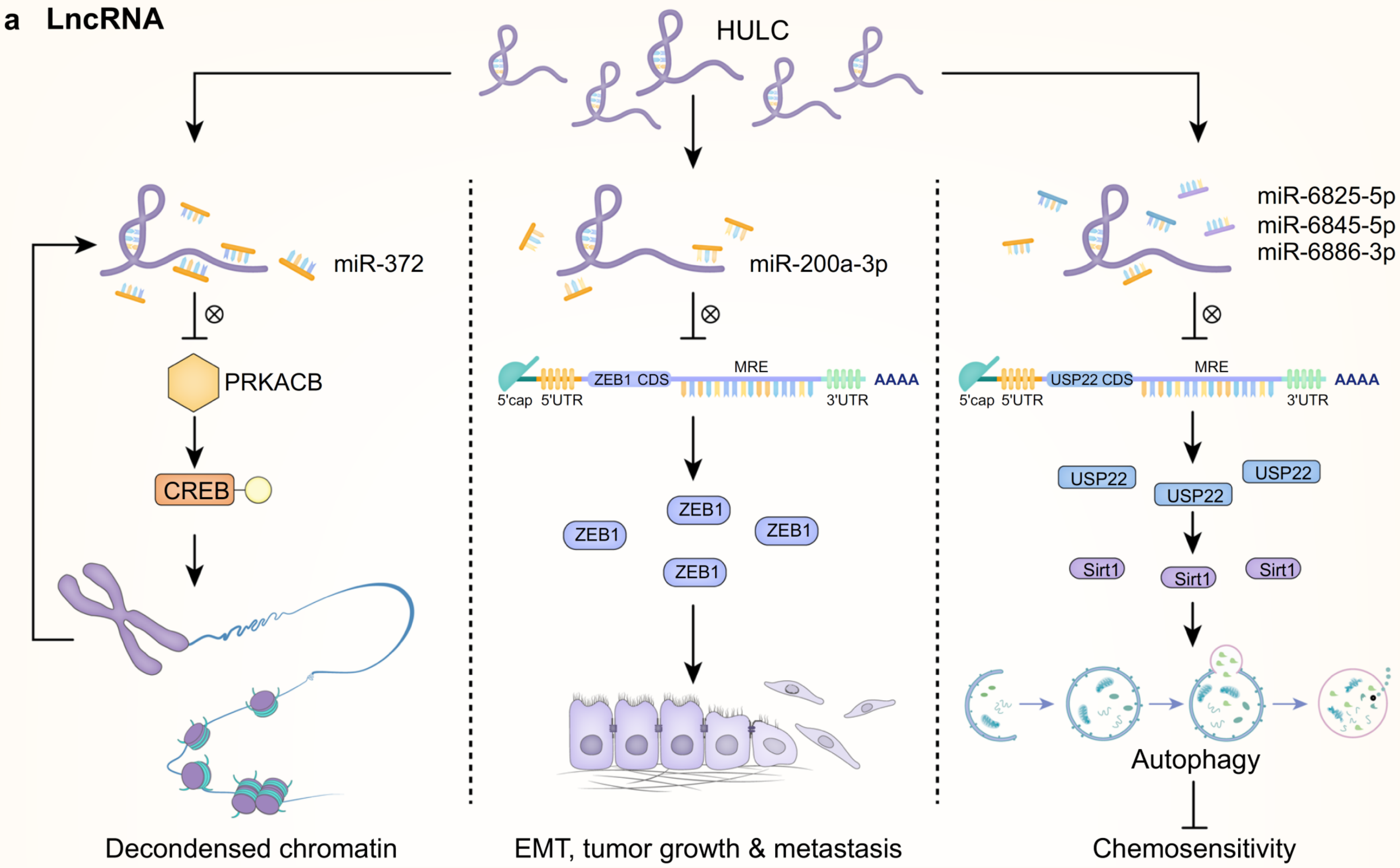
Migration:

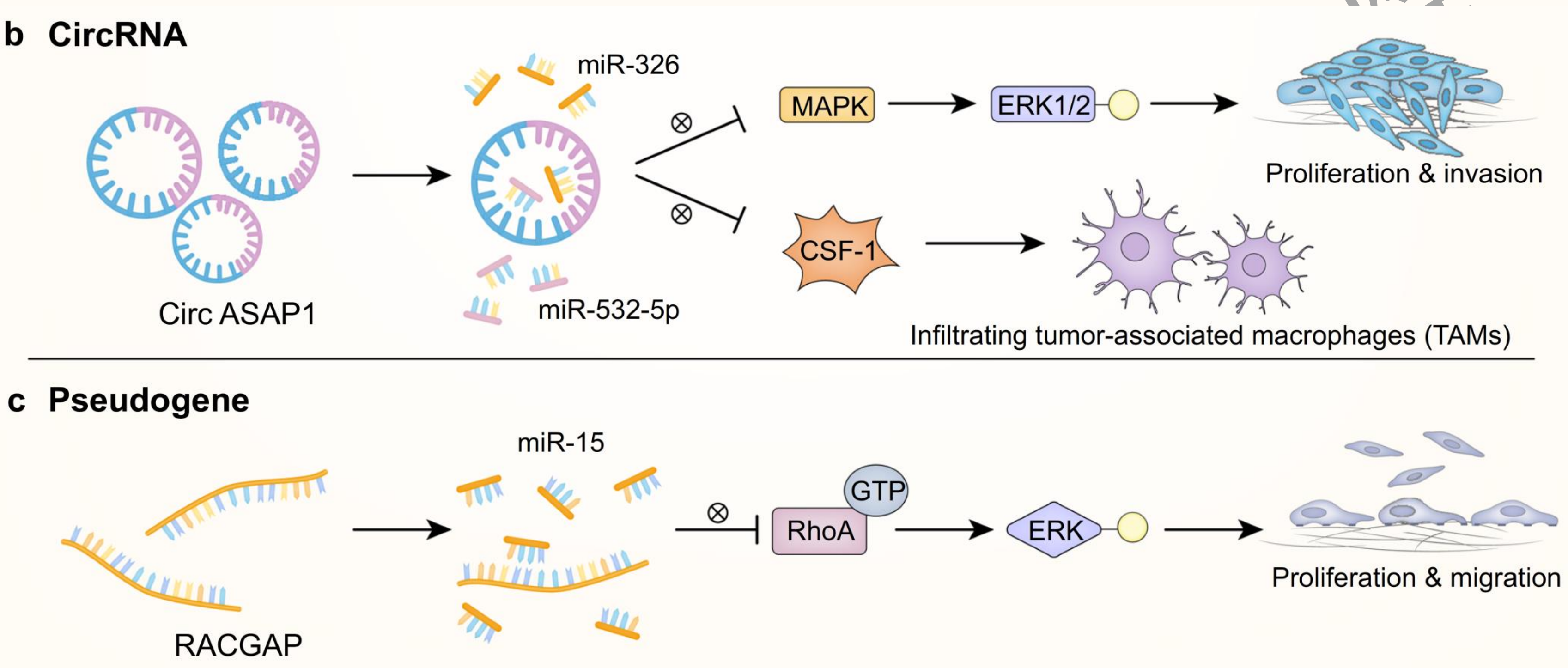
- MALAT1/miR-30a-5p/Vimentin
- circ_100338/miR-141-3p/RHEB
- circ_104718/miR-218-5p/TXNDC5

Proliferation:

- HULC/miR-186/HMGA2
- HOTAIR/miR-122/DNMTs
- circ_104075/miR-582-3p/YAP

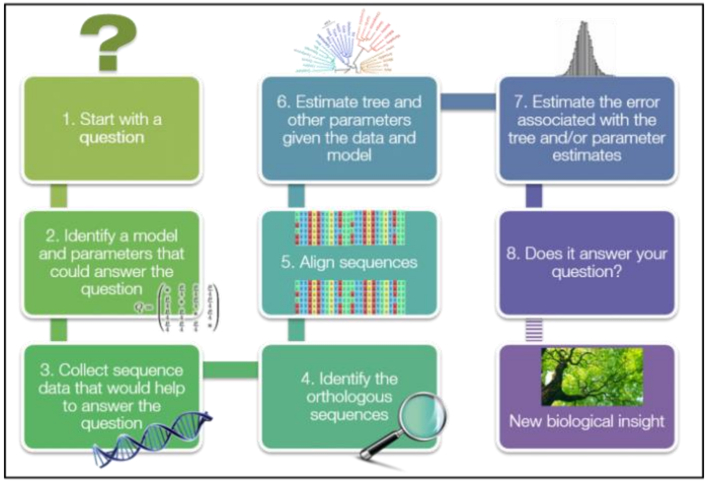
Fig. Summary of ncRNAs act as ceRNAs mediated function in hepatocellular carcinoma progression.


a LncRNA
HULC
miR-372
PRKACB
CREB
Decondensed chromatin
miR-200a-3p
5'cap
5'UTR
ZEB1 CDS
MRE
3'UTR
AAAA
ZEB1
ZEB1
ZEB1
ZEB1
EMT, tumor growth & metastasis
miR-6825-5p
miR-6845-5p
miR-6886-3p
5'cap
5'UTR
USP22 CDS
MRE
3'UTR
AAAA
USP22
USP22
USP22
Sirt1
Sirt1
Sirt1
Autophagy
Chemosensitivity



Xu G, Xu W Y, Xiao Y, et al. The emerging roles of non-coding competing endogenous RNA in hepatocellular carcinoma[J]. Cancer Cell International, 2020, 20(1): 1-21.

❖ 4.6 分子进化分析



1. Start with a question

2. Identify a model and parameters that could answer the question

3. Collect sequence data that would help to answer the question

4. Identify the orthologous sequences

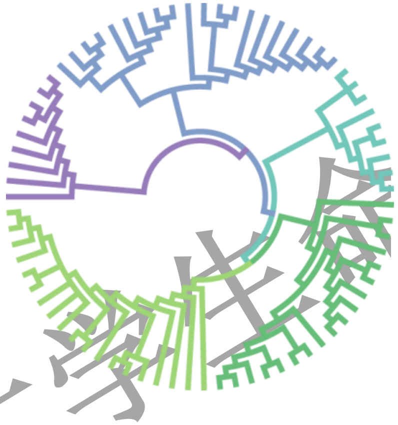
5. Align sequences

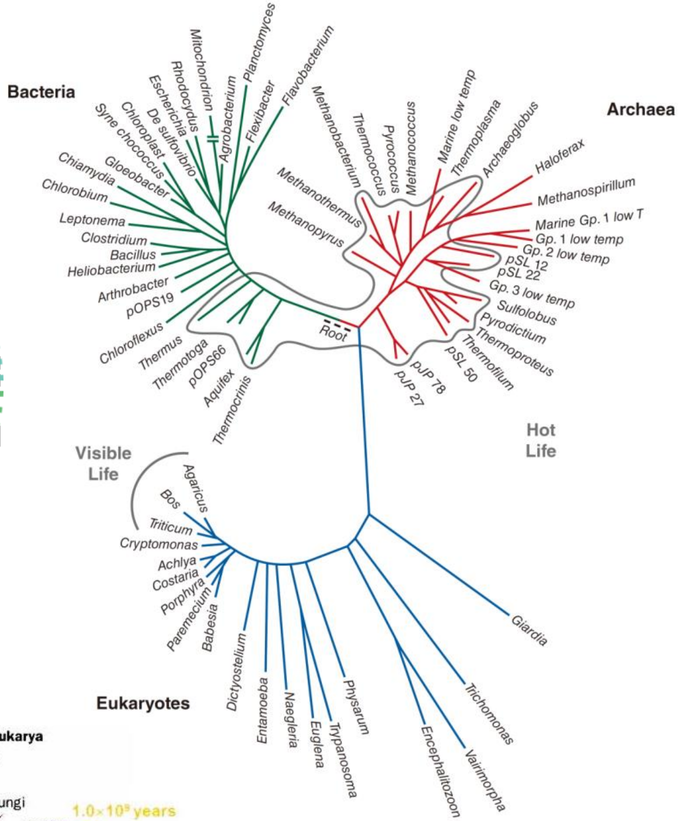
6. Estimate tree and other parameters given the data and model

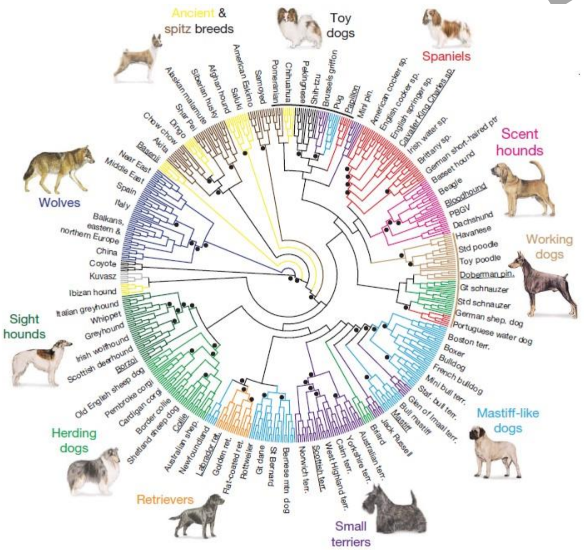
7. Estimate the error associated with the tree and/or parameter estimates

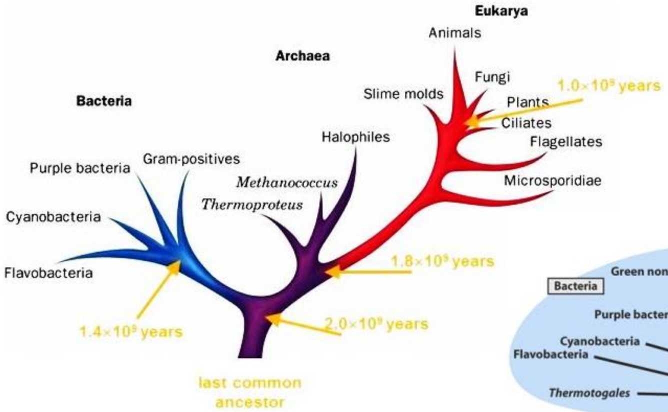
8. Does it answer your question?

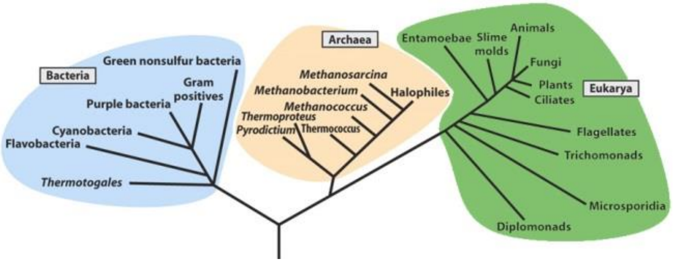
New biological insight

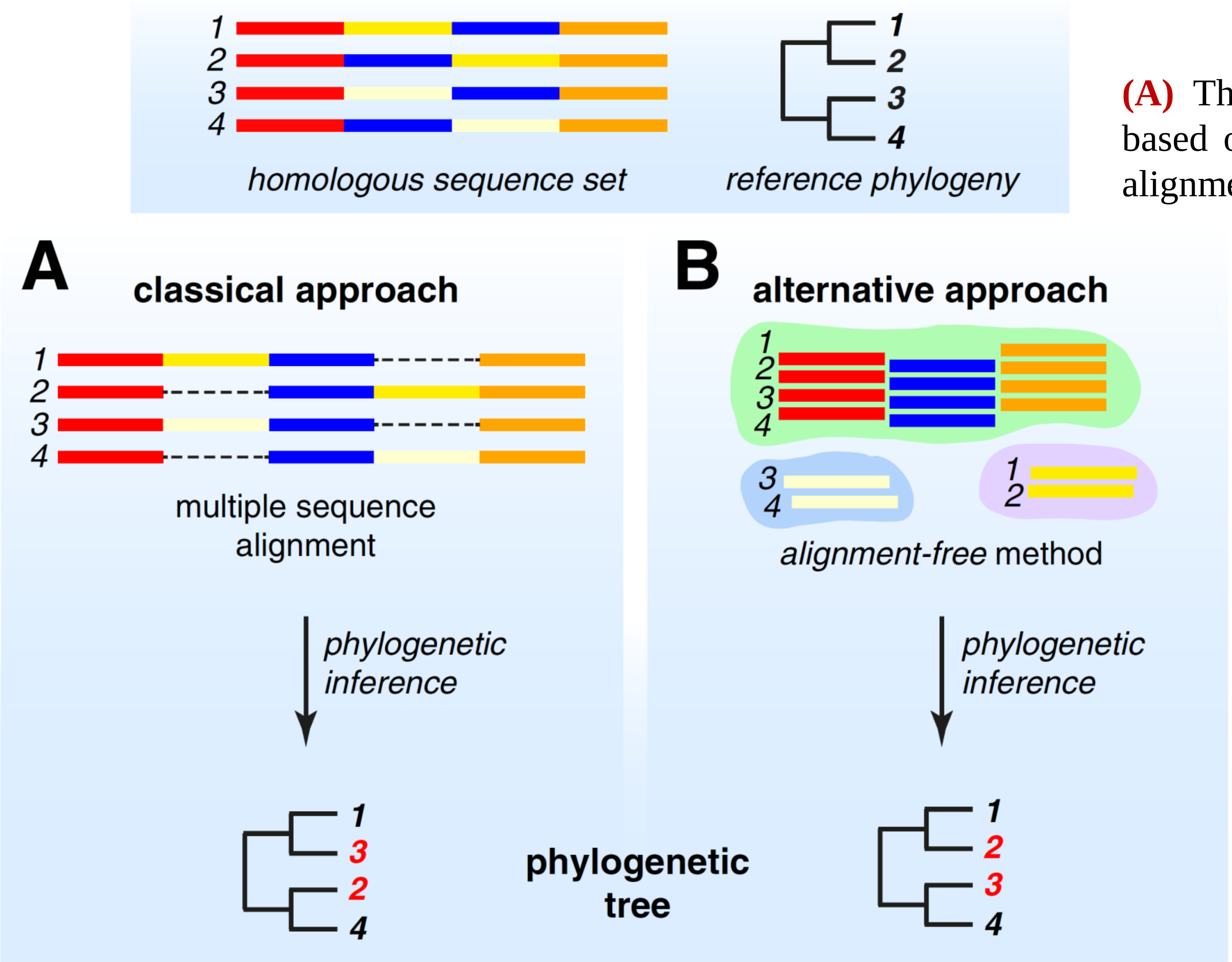












(A) The classical approach based on multiple sequence alignment.

(B) An alternative approach based on alignment-free methods, for a simple analysis example of homologous sequences 1, 2, 3 and 4, with a known phylogeny as a reference (shown on top)

Fig. Simplified workflow of phylogenomic approaches, **A** and **B**.

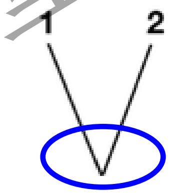
DNA Sequences

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
Sequence 1	A	C	T	G	C	C	G	A	T	C	G	C	A	T	C	G	T	A	C	G	Pig
Sequence 2	A	C	T	G	C	C	G	T	T	C	G	C	A	T	C	G	T	A	C	G	Dog
Sequence 3	A	C	A	G	C	C	G	-	T	C	G	C	A	T	C	G	T	A	C	G	Chicken
Sequence 4	A	C	T	G	G	C	G	A	T	C	G	C	T	T	C	G	T	T	C	G	Fish
Sequence 5	A	C	T	G	C	G	G	A	A	C	G	G	A	T	C	G	A	A	C	G	Sheep

❶

	1	2	3	4	5
1		0.05	0.1	0.15	0.2
2			0.1	0.2	0.25
3				0.25	0.3
4					0.25
5					

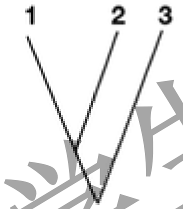
❷



❸

	1-2	3	4	5
1-2		0.1 1/2*(0.1+0.1)	0.175	0.225
3			0.25	0.3
4				0.25
5				

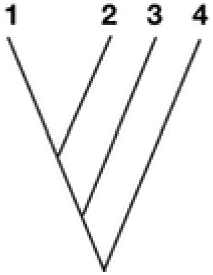
❹



❺

	1-3	4	5
1-3		0.2125 1/2*(0.25+0.175)	0.2625
4			0.25
5			

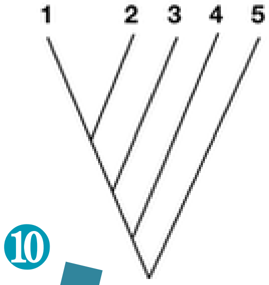
❻



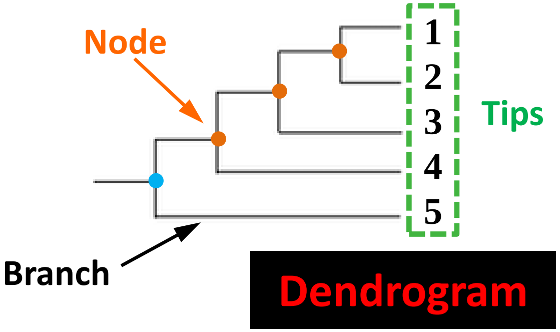
❼

❽ ...

❾

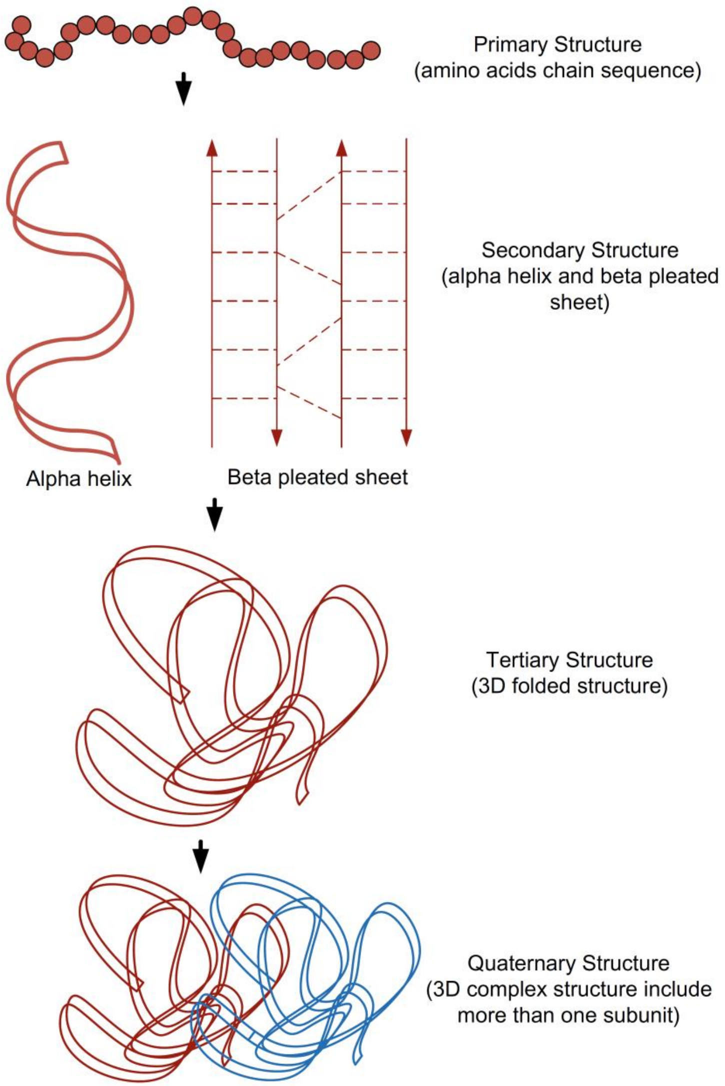


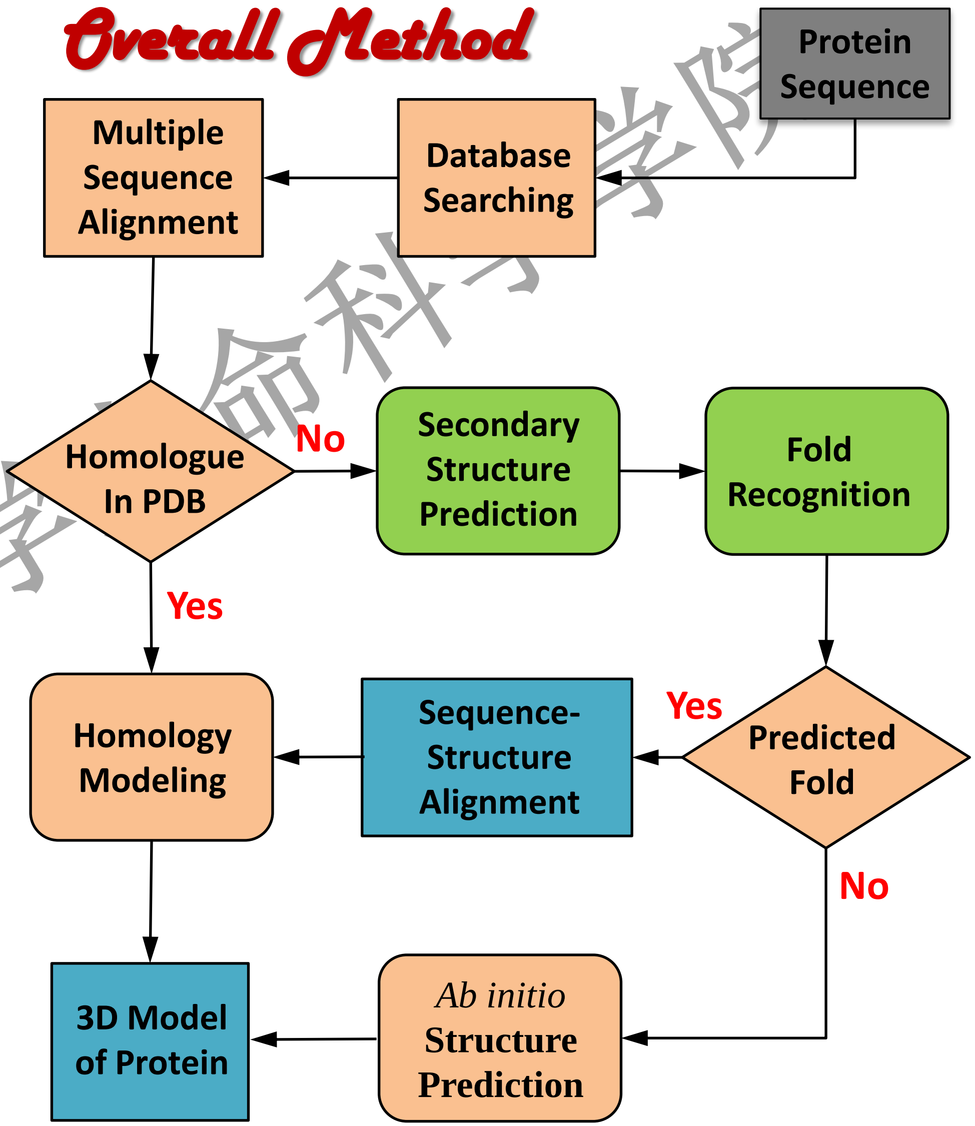
❿



Dendrogram

❖ 4.7 结构预测


Primary Structure
(amino acids chain sequence)
Secondary Structure
(alpha helix and beta pleated sheet)
Alpha helix
Beta pleated sheet
Tertiary Structure
(3D folded structure)
Quaternary Structure
(3D complex structure include more than one subunit)


Overall Method
Protein Sequence
Database Searching
Multiple Sequence Alignment
Homologue In PDB
No
Secondary Structure Prediction
Fold Recognition
Yes
Homology Modeling
Sequence-Structure Alignment
Yes
Predicted Fold
No
3D Model of Protein
Ab initio Structure Prediction

Homology modeling in drug discovery: current trends and applications

Protein with unknown 3D structure

↓

Identification of a homologous protein with experimental structure

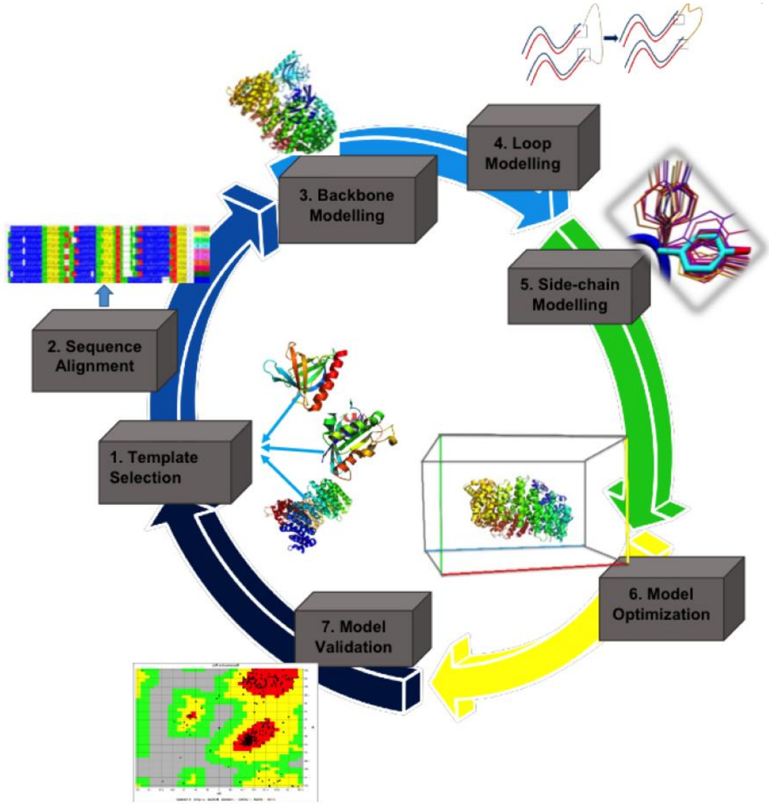
↓

Target/template sequence alignment

↓

Model building
Refinement & validation

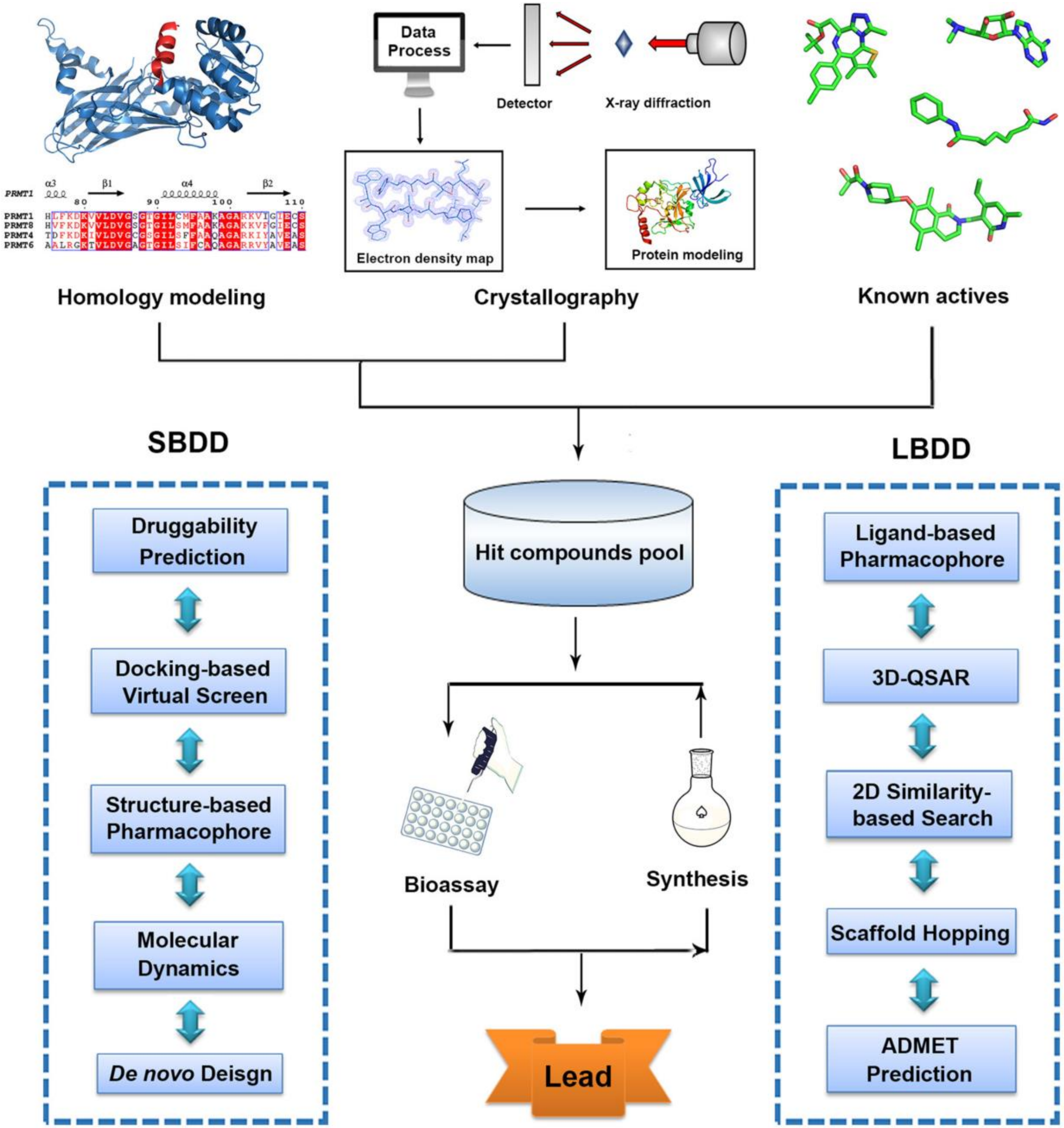
↓



Applications to drug discovery }

- Study of protein function and mechanism
- Assessment of target druggability
- Design of mutagenesis experiments
- High-throughput docking
- Lead identification and optimization

Refs: Munsamy G, et al. *Letters in Drug Design & Discovery*, 2017, 14(9): 1099-1111.
Cavasotto C N, *et al*. Drug discovery today, 2009, 14(13-14): 676-683.



Data Process

Detector

X-ray diffraction

Electron density map

Protein modeling

Homology modeling

Crystallography

Known actives

SBDD

Druggability Prediction

Docking-based Virtual Screen

Structure-based Pharmacophore

Molecular Dynamics

De novo Deisgn

Hit compounds pool

Bioassay

Synthesis

Lead

LBDD

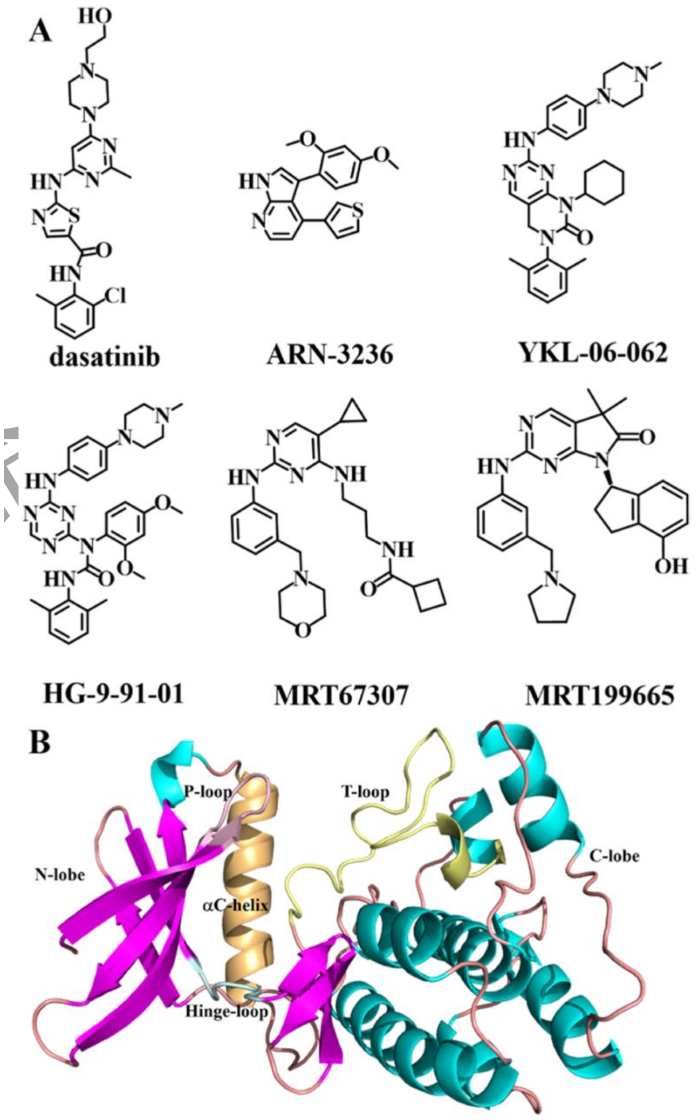
Ligand-based Pharmacophore

3D-QSAR

2D Similarity-based Search

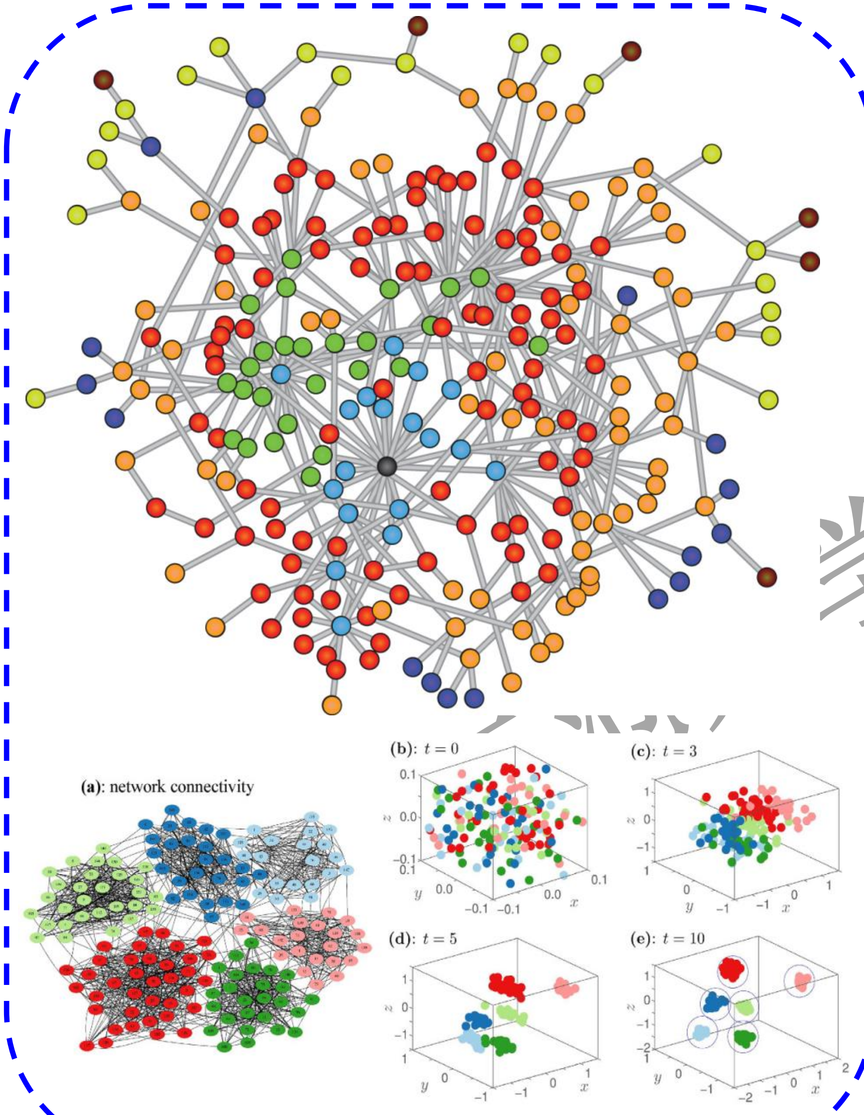
Scaffold Hopping

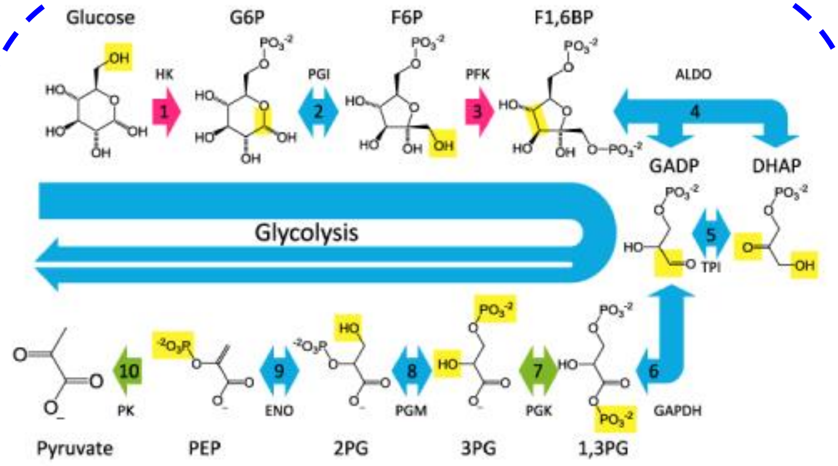
ADMET Prediction

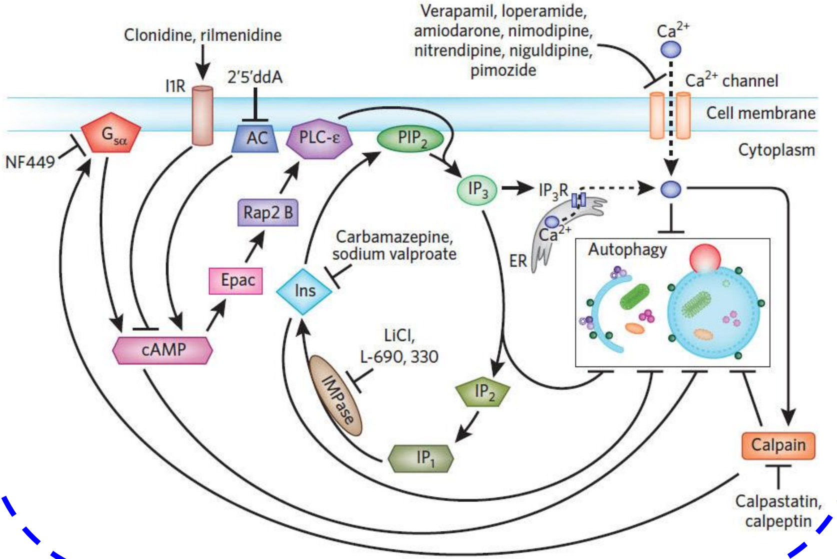


Traditional workflow of structure-based drug design (SBDD) and ligand-based drug design (LBDD).

❖ 4.8 分子间相互作用

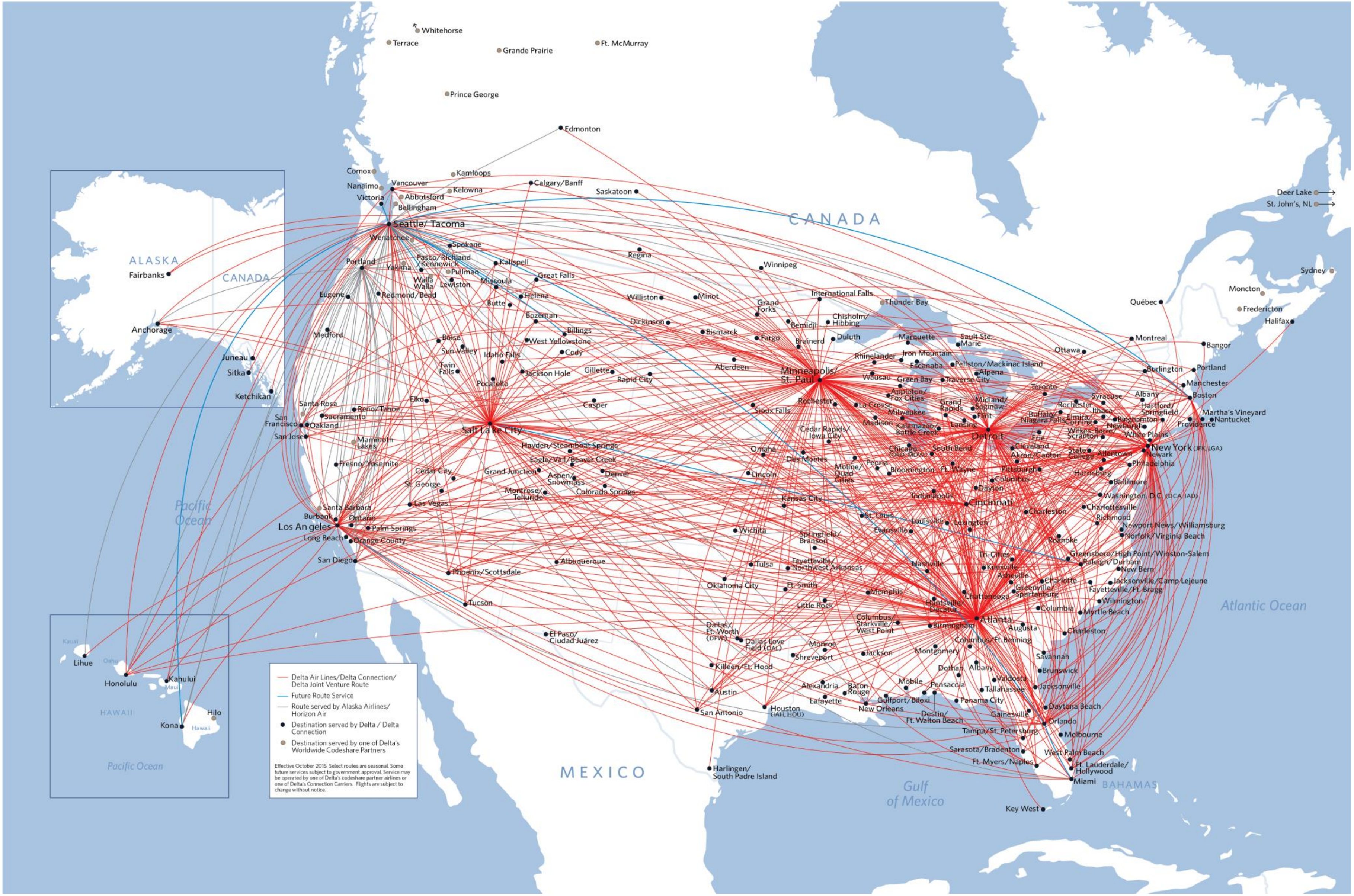






网络分析与通路分析

★美国及其周边重要航空枢纽

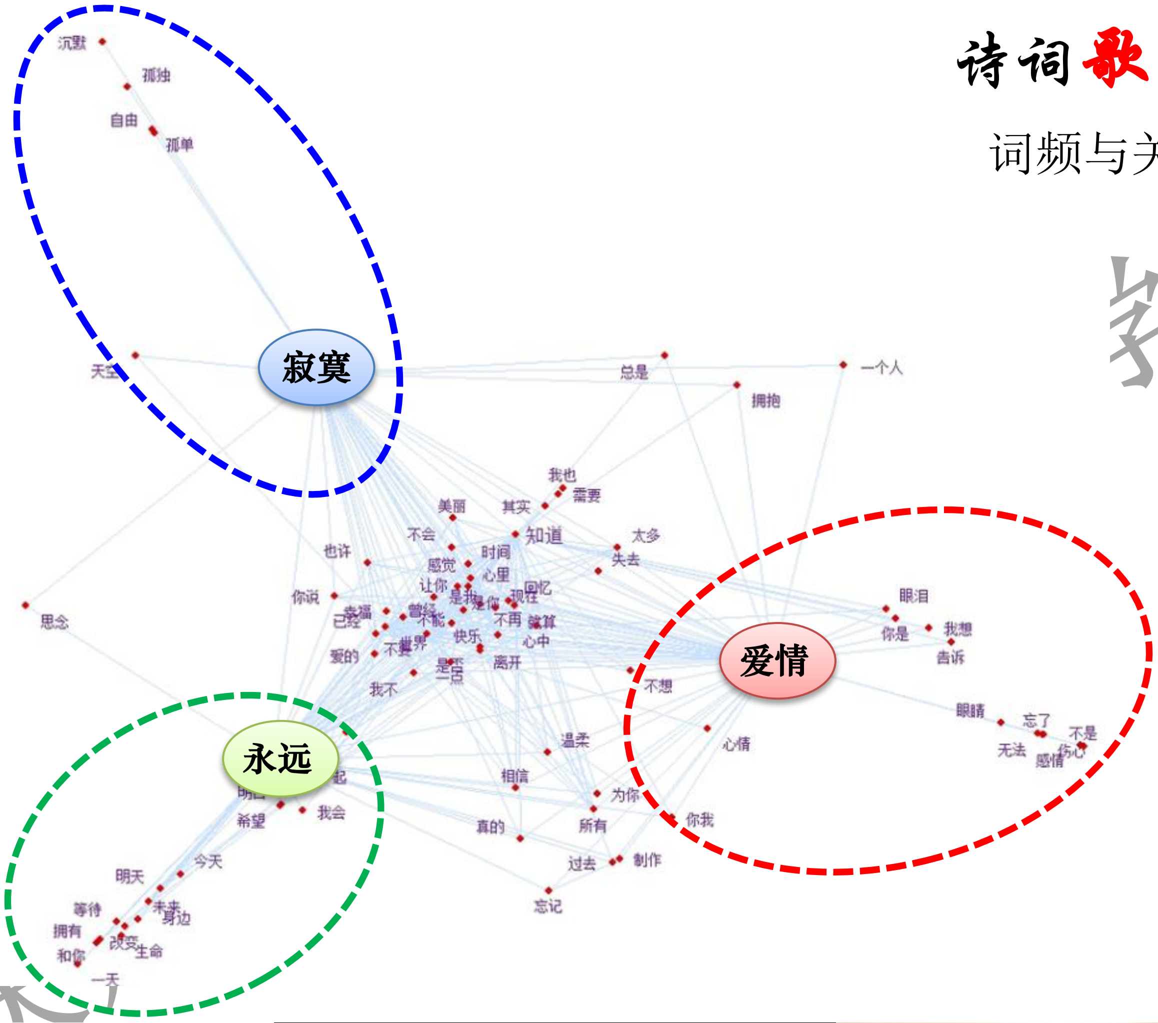


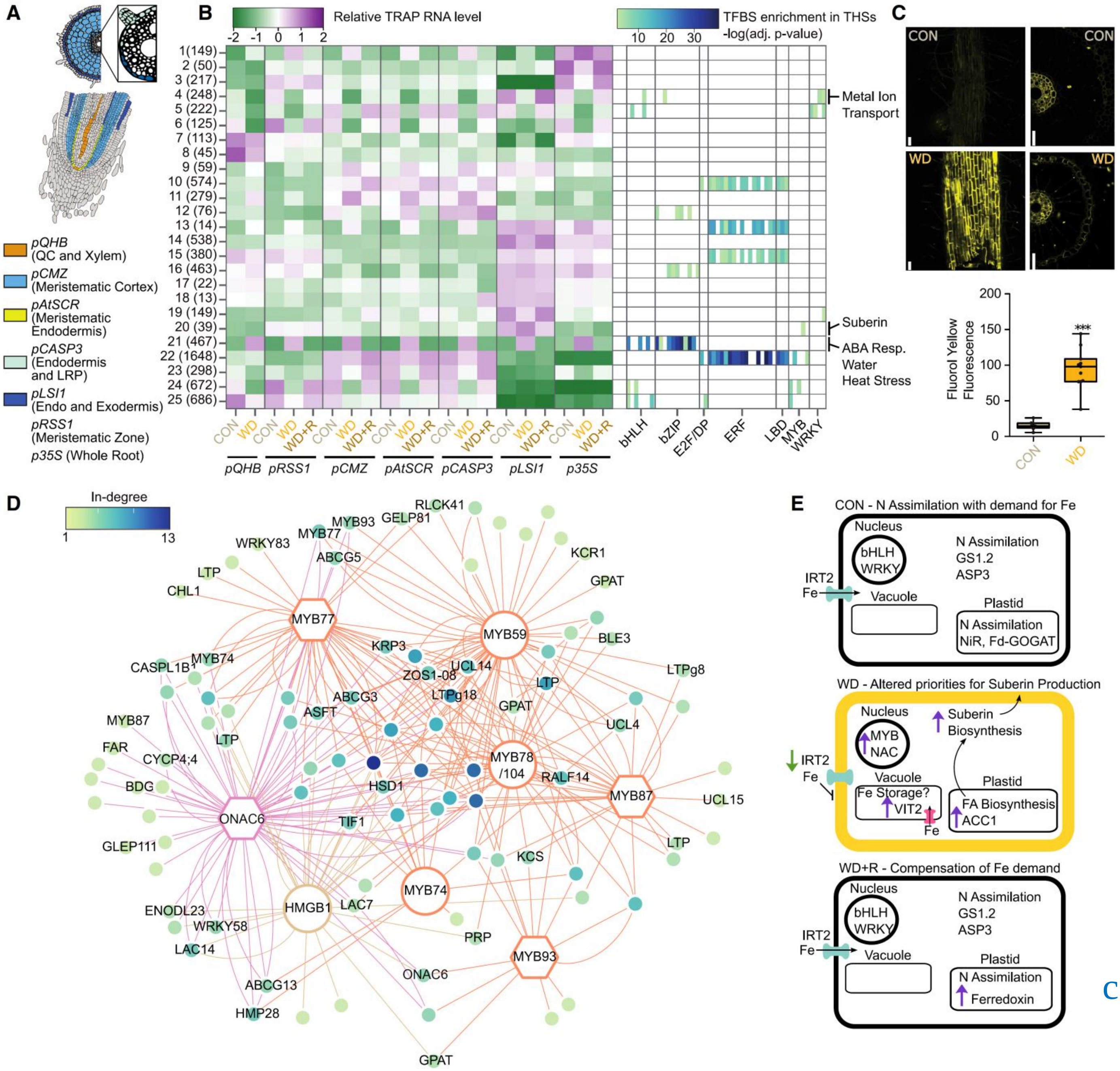
Effective October 2015. Select routes are seasonal. Some future services subject to government approval. Service may be operated by one of Delta's codeshare partner airlines or one of Delta's Connection Carriers. Flights are subject to change without notice.

诗词歌赋中的人生

词频与关键词网络分析

- 22996首歌曲
- 844位歌手




A
pQHB (QC and Xylem)
pCMZ (Meristematic Cortex)
pAtSCR (Meristematic Endodermis)
pCASP3 (Endodermis and LRP)
pLSI1 (Endo and Exodermis)
pRSS1 (Meristematic Zone)
p35S (Whole Root)
B
Relative TRAP RNA level
-2 -1 0 1 2
TFBS enrichment in THSs -log(adj. p-value)
10 20 30
1(149)
2 (50)
3 (217)
4 (248)
5 (222)
6 (125)
7 (113)
8 (45)
9 (59)
10 (574)
11 (279)
12 (76)
13 (14)
14 (538)
15 (380)
16 (463)
17 (22)
18 (13)
19 (149)
20 (39)
21 (467)
22 (1648)
23 (298)
24 (672)
25 (686)
Metal Ion Transport
Suberin
ABA Resp.
Water
Heat Stress
CON WD CON WD WD+R CON WD WD+R CON WD WD+R CON WD WD+R CON WD WD+R CON WD WD+R
pQHB pRSS1 pCMZ pAtSCR pCASP3 pLSI1 p35S
bHLH bZIP E2F/DP ERF LBD MYB WRKY
C
CON
CON
WD
WD
Fluorol Yellow Fluorescence
200
150
100
50
0
***
CON WD
D
In-degree
1 13
RLCK41 MYB93 GELP81 MYB77 WRKY83 ABCG5 KCR1 LTP GPAT CHL1
MYB77 MYB59 BLE3 KRP3 CASPL1B1 MYB74 ZOS1-08 UCL14 LTPg8 LTP
LTPg18 ABCG3 ASFT GPAT MYB87 UCL4 LTP FAR CYCP4;4 MYB78 /104
RALF14 HSD1 BDG MYB87 UCL15 ONAC6 TIF1 GLEP111 LTP KCS
MYB74 ENODL23 HMGB1 LAC7 WRKY58 PRP LAC14 MYB93 ONAC6 ABCG13 HMP28 GPAT
E
CON - N Assimilation with demand for Fe
Nucleus
bHLH
WRKY
N Assimilation
GS1.2
ASP3
IRT2
Fe
Vacuole
Plastid
N Assimilation
NiR, Fd-GOGAT
WD - Altered priorities for Suberin Production
Nucleus
MYB
NAC
Suberin Biosynthesis
IRT2
Fe
Vacuole
Fe Storage?
VIT2
Fe
Plastid
FA Biosynthesis
ACC1
WD+R - Compensation of Fe demand
Nucleus
bHLH
WRKY
N Assimilation
GS1.2
ASP3
IRT2
Fe
Vacuole
Plastid
N Assimilation
Ferredoxin

Ref:
Reynoso *et al.*, 2022, Developmental Cell 57, 1177–1192

Fig. Responses to water deficit and recovery across root domains reveal dynamic patterns of translatomes and contrasting metabolic pathways

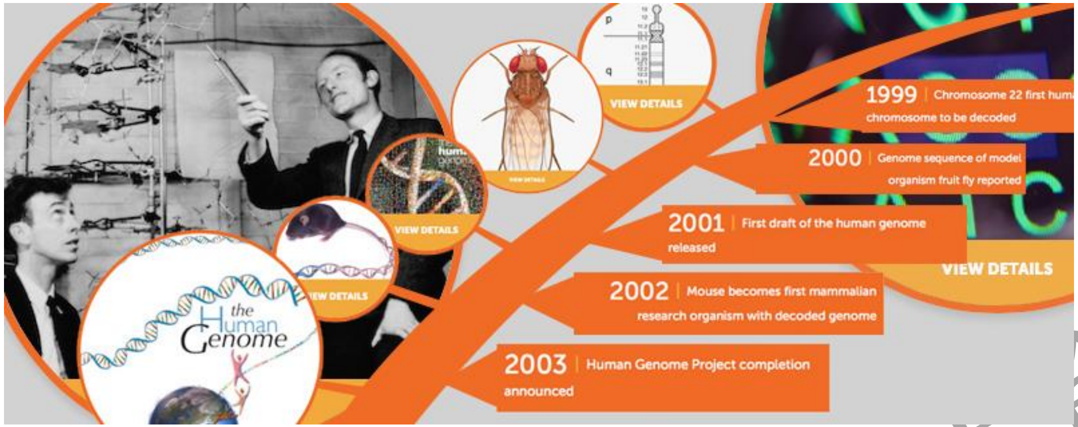
第5节：大数据时代的生物信息学

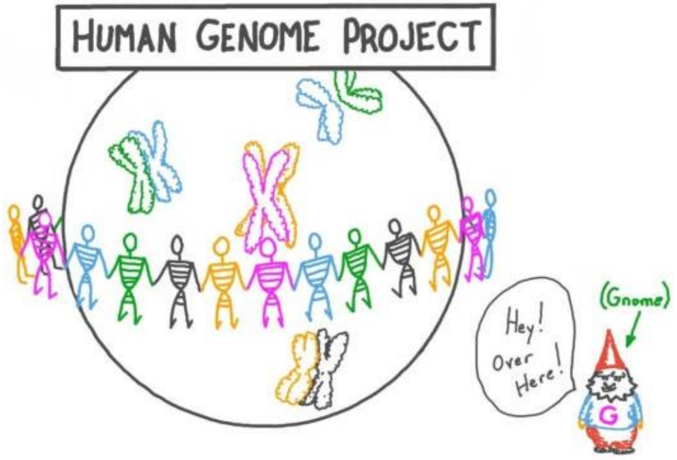
重要知识点

- ✓ 人类基因组计划
- ✓ 生物组学
- ✓ 生物大数据
- ✓ 知识发现

❖ 5.1 人类基因组计划

Ref: Collins F S, *et al*. ***Science***, 2003, 5617: 286-290.


the Human Genome
VIEW DETAILS
VIEW DETAILS
VIEW DETAILS
VIEW DETAILS
1999 | Chromosome 22 first hum
chromosome to be decoded
2000 | Genome sequence of model
organism fruit fly reported
2001 | First draft of the human genome
released
2002 | Mouse becomes first mammalian
research organism with decoded genome
2003 | Human Genome Project completion
announced



Area	Goal	Achieved	Date
Genetic map 遗传图	2- to 5-cM resolution map (600 to 1,500 markers)	1-cM resolution map (3,000 markers)	September 1994
Physical map 物理图	30,000 sequence-tagged sites (STSs)	52,000 STSs	October 1998
DNA sequence 序列图	95% of gene-containing part of human sequence finished to 99.99% accuracy	>98% of gene-containing part of human sequence finished to 99.99% accuracy	April 2003
Capacity and cost of finished sequence	Sequence 500 Mb/year at <$0.25 per finished base	Sequence >1,400 Mb/year at <$0.09 per finished base	November 2002
Human sequence variation	100,000 mapped human SNPs	3.7 million mapped human SNPs	February 2003
Gene identification 基因图	Full-length human cDNAs	15,000 full-length human cDNAs	March 2003
Model organisms	Complete sequences of *E. coli*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*	Finished sequences of *E. coli*, *S. cerevisiae*, *C. elegans*, *D. melanogaster*, plus whole-genome drafts of several others, including *C. briggsae*, *D. pseudoobscura*, mouse, and rat	April 2003
Functional analysis	Develop genomic-scale technologies	High-throughput oligonucleotide synthesis	1994
		DNA microarrays	1996
		Normalized and subtracted cDNA libraries	1996
		Eukaryotic, whole-genome knockouts (yeast)	1999
		Scale-up of two-hybrid mapping	2002

The human genome project was initiated in 1990 in order to sequence the whole genetic content of the human genome and other species to know genes and their functions.

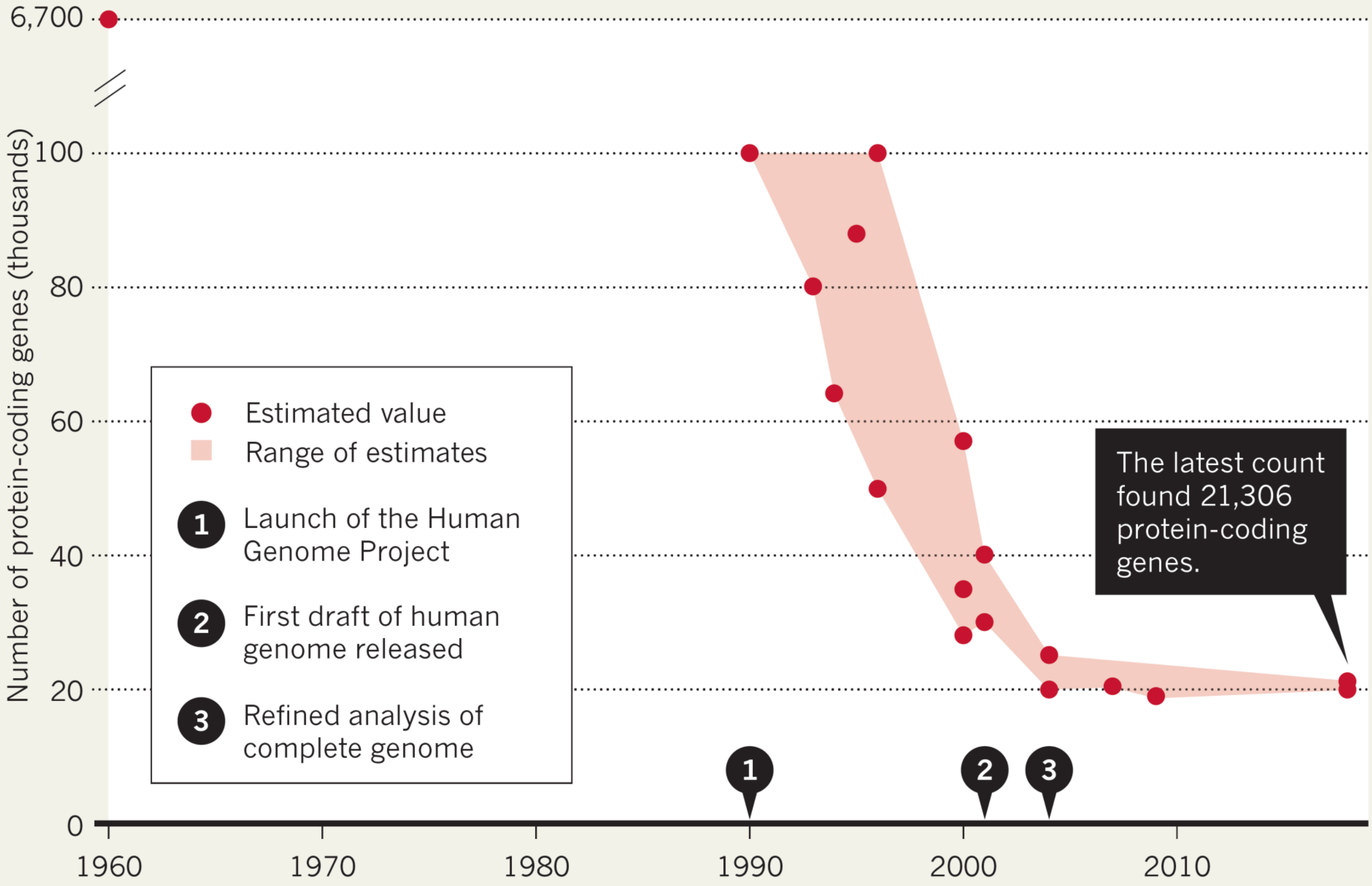
Top 8 genes

TP53
TNF
EGFR
IL6
VEGFA
APOE
TGFB1
MTHFR

对人类基因组的认识越来越清晰

GENE TALLY

Scientists still don’t agree on how many protein-coding genes the human genome holds, but the range of their estimates has narrowed in recent years.



The latest count found 21,306 protein-coding genes.

Ref: Willyard C. New human gene tally reignites debate[J]. *Nature*, 2018, 558(7710): 354-356.

★中国在行动

杨焕明院士（前排右二）

Ref: X. L. Wang, *et al*. ***Protein & Cell***, 2018, 9(4): 317-321

RECOLLECTION

The international Human Genome Project (HGP) and China’s contribution

Xiaoling Wang, Zhi Xia, Chao Chen, Huanming Yang✉

BGI-China, Shenzhen 518083, China
✉ Correspondence: yanghuanming@genomics.cn (H. Yang)

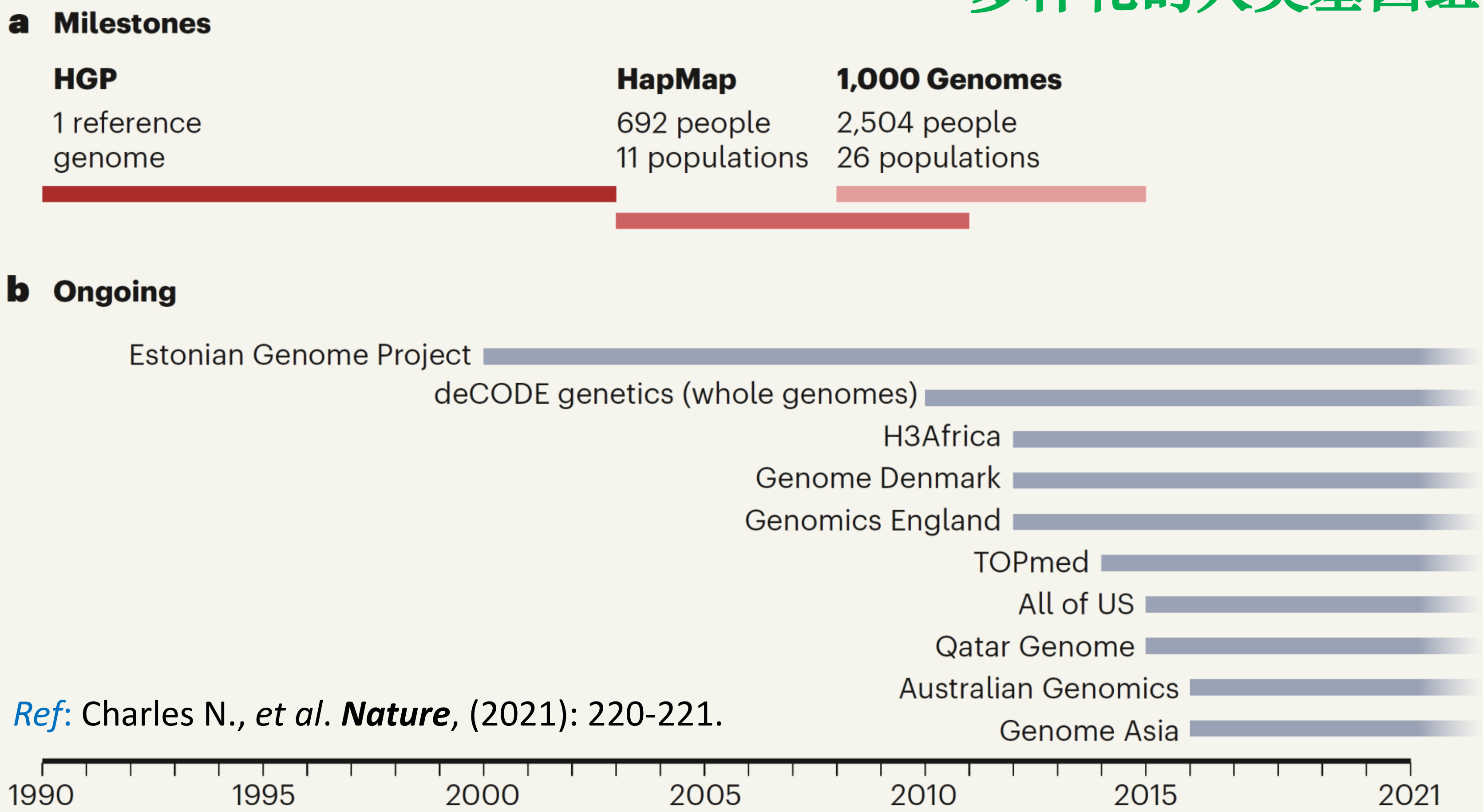
截止目前，人类基因组被明确的基因数目如下：

Locus Group	Locus Type	Count
protein-coding gene (19193) 19206 ↑	gene with protein product	19193
non-coding RNA (8581) 8906 ↑	RNA, Y	4
	RNA, cluster	119
	RNA, long non-coding	5243
	RNA, micro	1912
	RNA, misc	30
	RNA, ribosomal	60
	RNA, small nuclear	50
	RNA, small nucleolar	568
	RNA, transfer	591
	RNA, vault	4
Pseudogene (13908) 14008 ↑	T cell receptor pseudogene	36
	immunoglobulin pseudogene	203
	pseudogene	13669
Other (1035) 1004 ↑	T cell receptor gene	201
	complex locus constituent	29
	endogenous retrovirus	109
	fragile site	116
	immunoglobulin gene	229
	protocadherin	39
	readthrough	138
	region	38
	unknown	128
	virus integration site	8
Total Approved (42717) 43124 ↑		

Last update: 17/08/21
Newly update: 21/08/22

https://www.genenames.org/download/statistics-and-files/

多样化的人类基因组

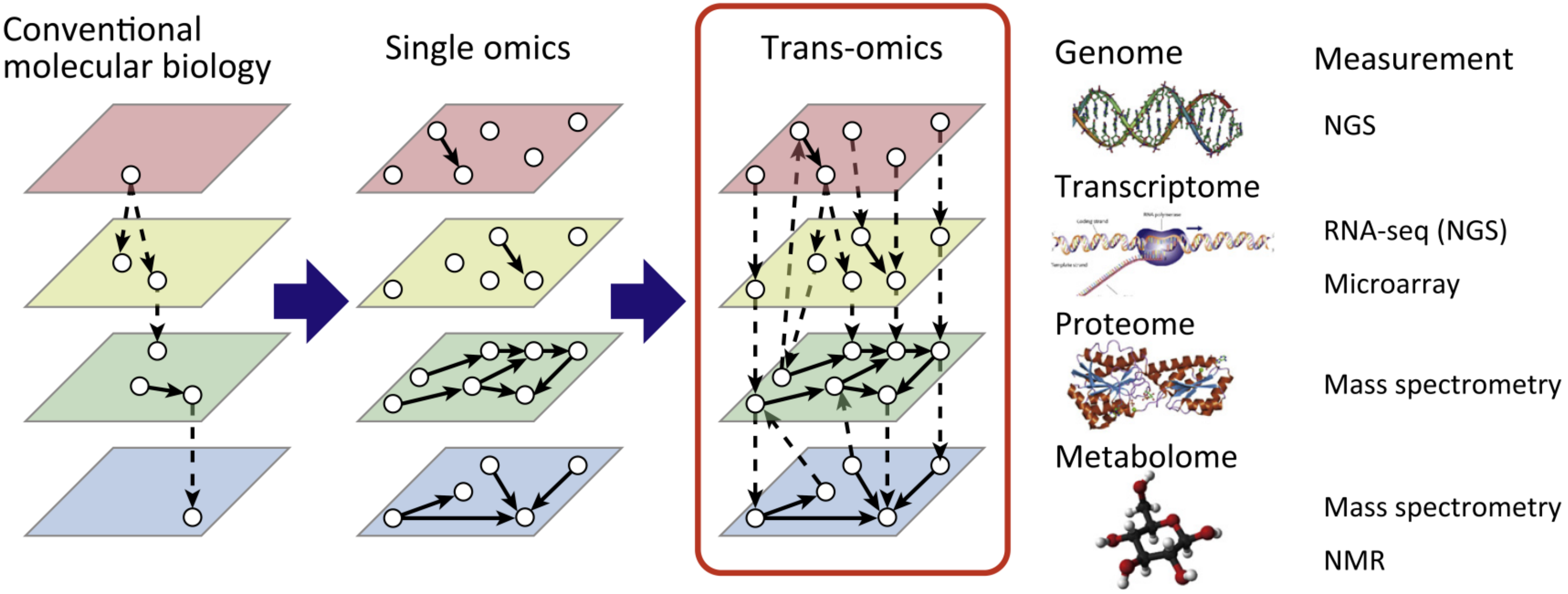


Ref: Charles N., *et al*. ***Nature***, (2021): 220-221.

***Increasing diversity in genomics*.** A) The **Human Genome Project (HGP)** was established in 1990 and completed in 2003, with the first draft of the human genome published in 2001. Since then, collaborative efforts have resulted in the analysis of large numbers of genomes from increasingly diverse populations. Milestones of note include the International **HapMap Project** and **the 1000 Genomes Project**. b) Today, there are many ongoing projects to sequence populations around the world.

❖ 5.2 生物组学 (Omics)

Ref: Yugi K, *et al*. Trends in biotechnology, 2016, 34(4): 276-290.



Conventional molecular biology

Single omics

Trans-omics

Genome

Transcriptome

Proteome

Metabolome

Measurement

NGS

RNA-seq (NGS)

Microarray

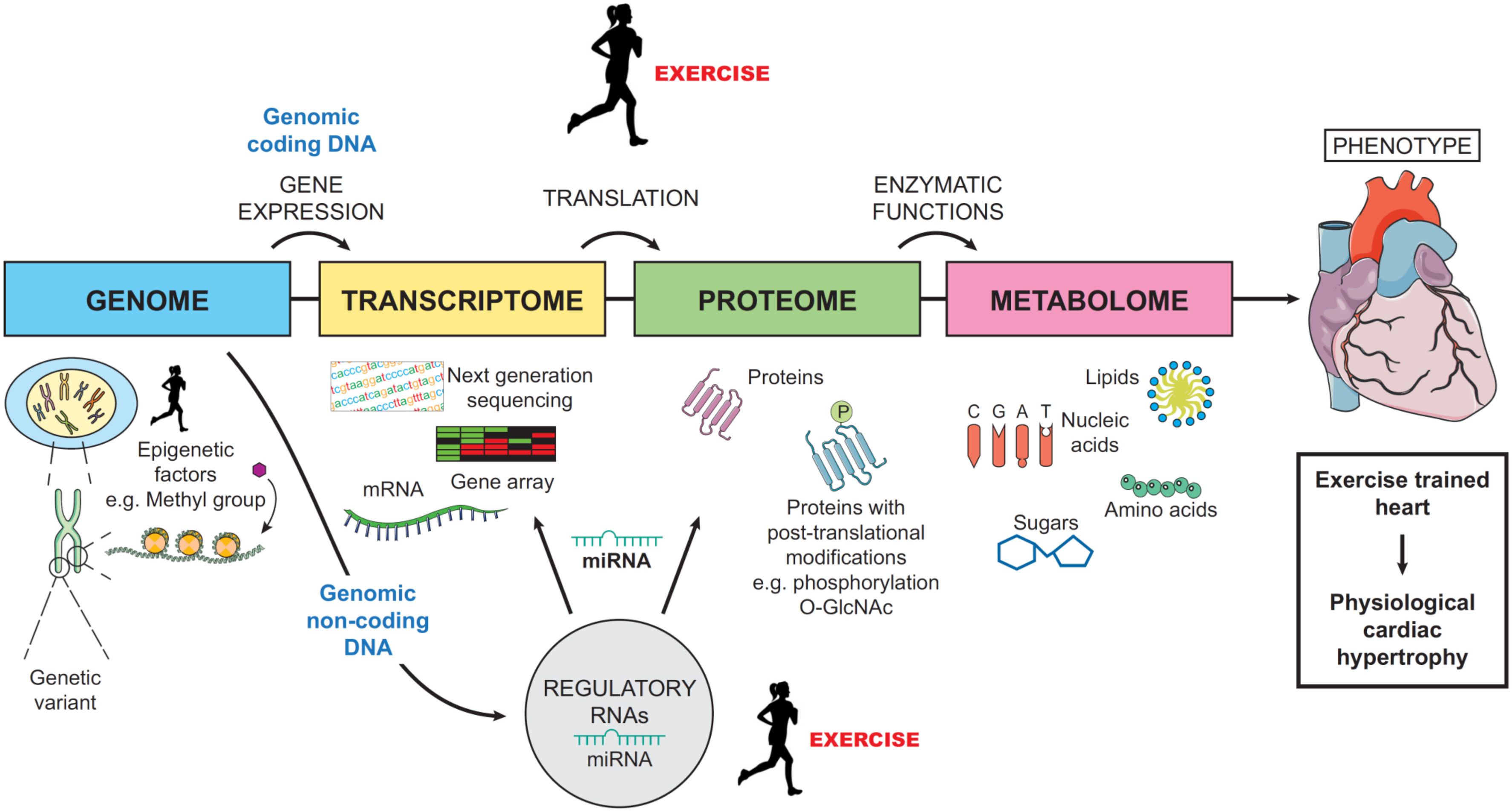
Mass spectrometry

Mass spectrometry

NMR

- ✓ **What can happen?** Human genome contains roughly 3 billion nucleotides and just under 20,000 protein-coding genes - an estimated 1% of the genome's total length.
- ✓ **What appears to be happening?** Approximately 360,000 mRNA molecules are present in a single mammalian cell, made up of about 12,000 (14,000 for human) different transcripts with a typical length of around 2 kb. Some mRNAs comprise 3% of the mRNA pool whereas others account for less than 0.1%. These rare or low-abundance mRNAs may have a copy number of only 5~15 molecules per cell.
- ✓ **What makes it happen?** Human body contains 80,000~400,000 proteins in proteome, while A typical cell holds 42 million protein molecules, scientists reveal.
- ✓ **What actually happens?** HMDB collects detailed information ~3100 metabolites found in human urine along with 4651 metabolites found in human serum.

Ref: Bernardo B C, *et al*. *Physiol. Reviews*, 2018, 98(1): 419-475.



EXERCISE

Genomic coding DNA

GENE EXPRESSION

TRANSLATION

ENZYMATIC FUNCTIONS

PHENOTYPE

GENOME

TRANSCRIPTOME

PROTEOME

METABOLOME

Epigenetic factors e.g. Methyl group

Genetic variant

Next generation sequencing

Gene array

mRNA

miRNA

Genomic non-coding DNA

REGULATORY RNAs

miRNA

EXERCISE

Proteins

Proteins with post-translational modifications e.g. phosphorylation O-GlcNAc

Lipids

C G A T

Nucleic acids

Amino acids

Sugars

Exercise trained heart

Physiological cardiac hypertrophy

The use of multi-omics platforms to identify novel mechanisms and uncover exercise signatures. Integrating data from multi-omics systems to understand genetic variants and epigenetic marks, gene expression and miRNAs, proteins, and metabolites during exercise to define molecular pathways of exercise.

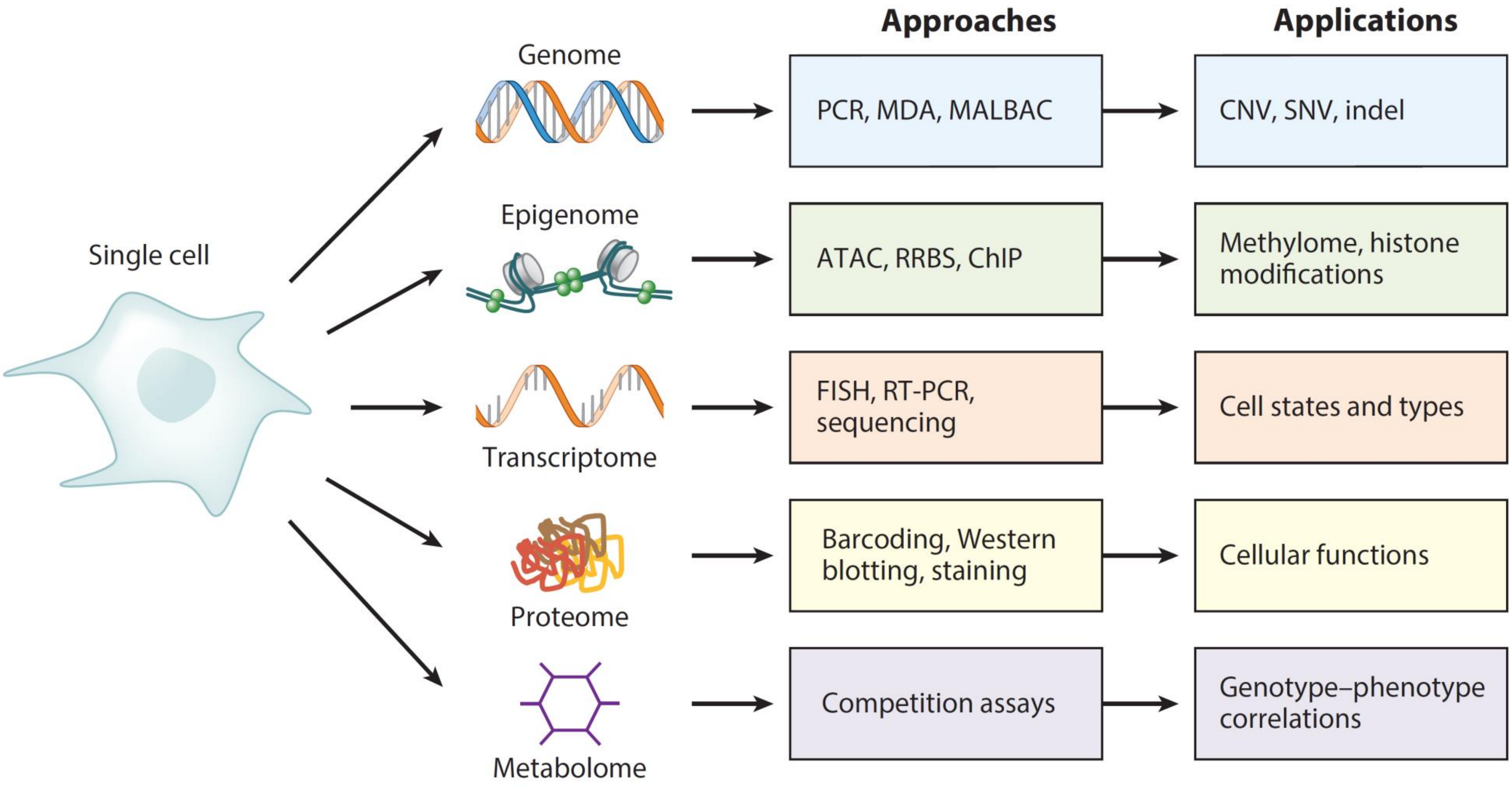
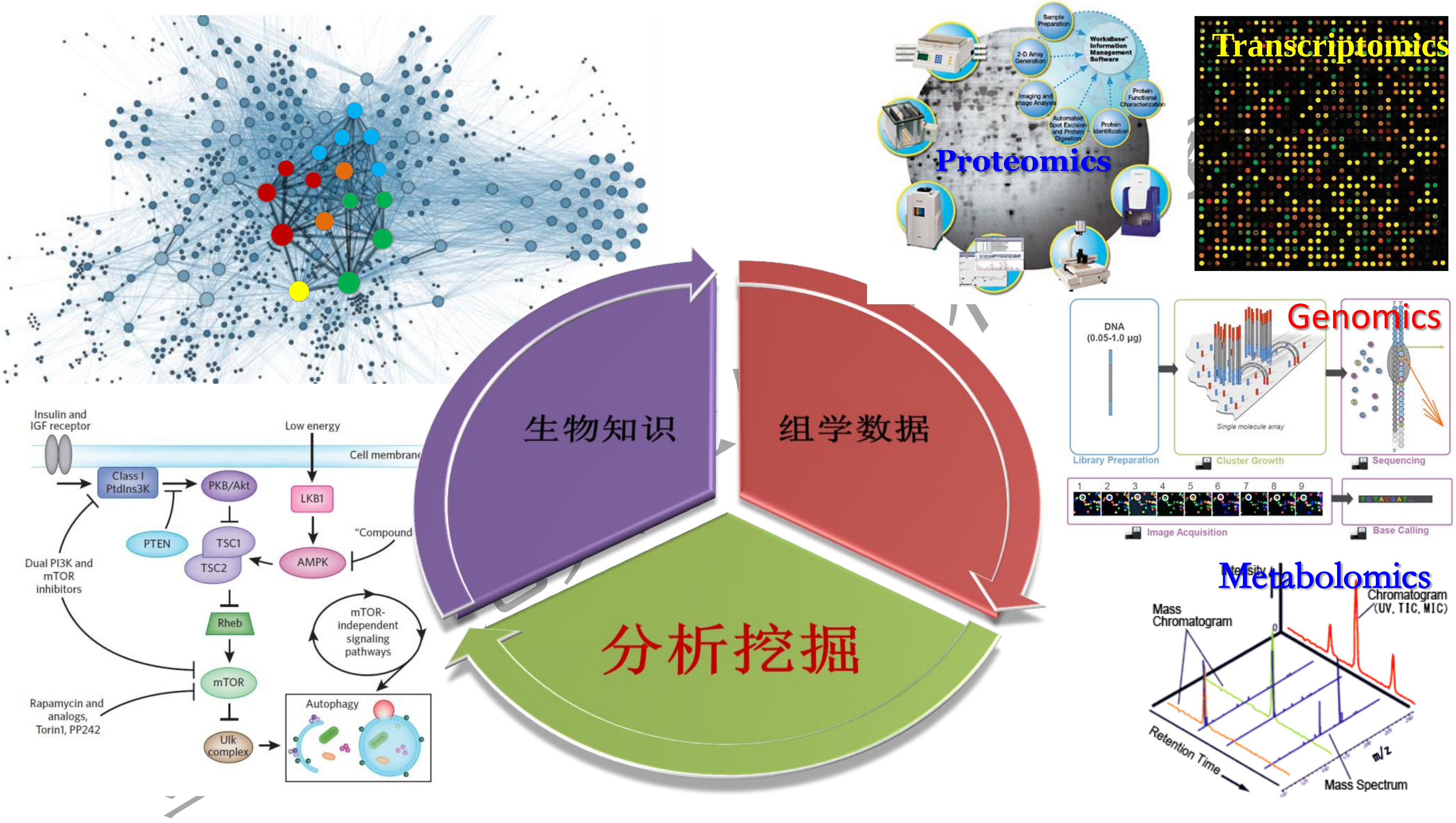


Figure 1

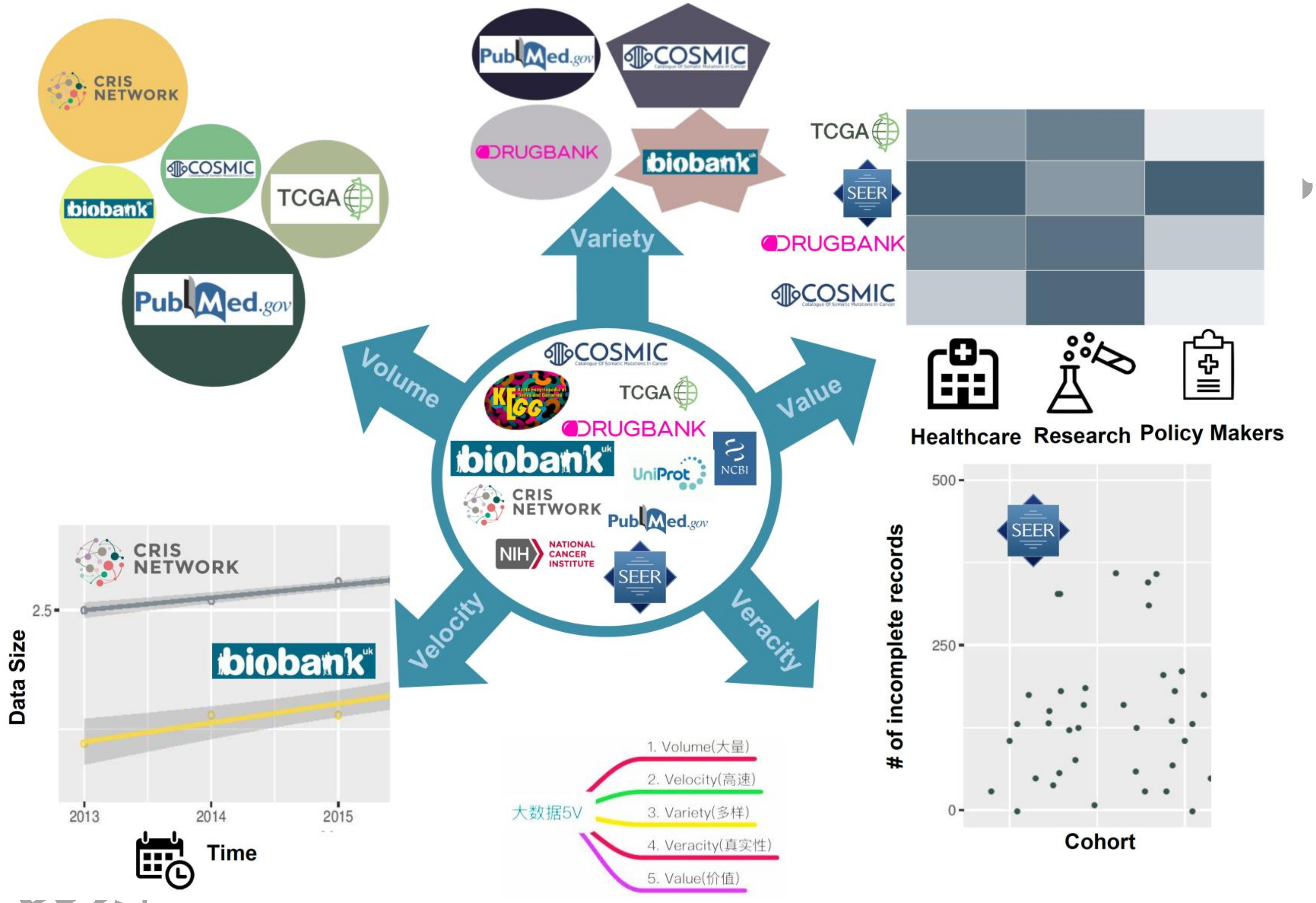
Ref: Deng Y, *et al*. ***Annual review of biomedical engineering***, 2019, 21: 365-393.

Overview of approaches and applications in single-cell omics measurement. Abbreviations: ATAC, assay for transposase-accessible chromatin; ChIP, chromatin immunoprecipitation; CNV, copy number variation; FISH, fluorescence in situ hybridization; indel, insertion/deletion; MALBAC, multiple annealing and loop-based amplification cycling; MDA, multiple displacement amplification; PCR, polymerase chain reaction; RRBS, reduced-representation bisulfite sequencing; RT-PCR, reverse transcription polymerase chain reaction; seq, sequencing; SNV, single-nucleotide variant.

❖ 5.3 生物大数据与知识发现



生物大数据 $\xrightarrow{\text{数据分析}}$ 生物学知识

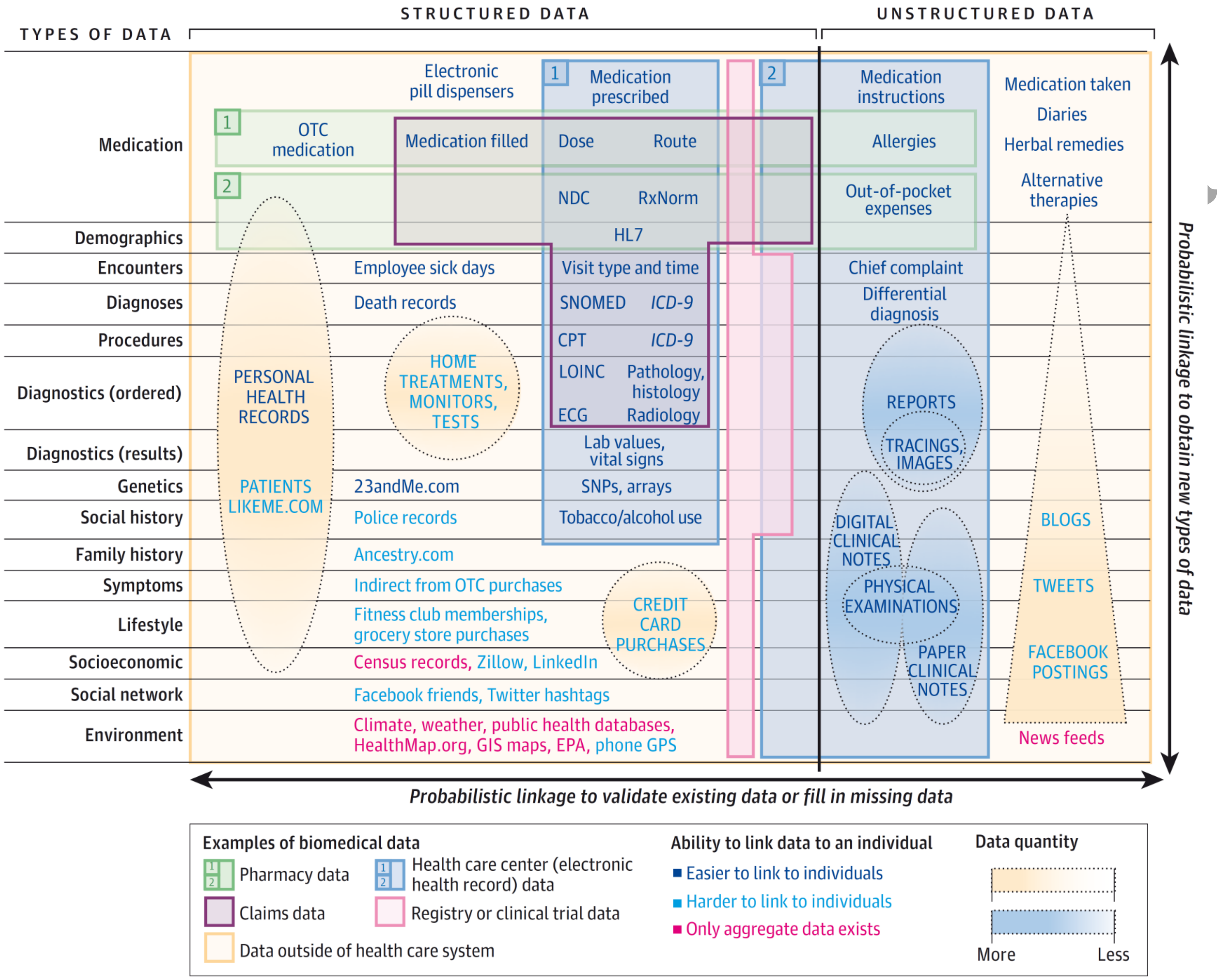


Big Biomedical Data. The **5Vs model** is utilized to characterize the very nature of big biomedical data. As observed, the dominant big data dimensions, i.e., volume, velocity, variety, veracity, and value, are present in existing biomedical datasets.

Ref: Vidal, et al. Current Trends in Semantic Web Technologies. Springer, Cham, 2019. 25-56.

□ 大数据的3V~8V

大容量	✓ While volume is by no means the only component that makes Big Data "big," it is certainly a primary feature.
快速化	✓ Big Data technology allows databases to process, analyse, and configure data while it is being generated – sometimes within milliseconds.
多样化	✓ Big Data is typically comprised of combinations of structured, unstructured, and semi-structured data.
真实性	✓ Big Data, it's only valuable if it is accurate, relevant, and timely.
价值化	✓ Without question, the results that come from Big Data analysis are often fascinating and unexpected.



TYPES OF DATA

STRUCTURED DATA

UNSTRUCTURED DATA

Medication

Electronic pill dispensers

1 Medication prescribed

2 Medication instructions

Medication taken

Diaries

Herbal remedies

Alternative therapies

1 OTC medication

Medication filled

Dose

Route

Allergies

2

NDC

RxNorm

Out-of-pocket expenses

HL7

Demographics

Encounters

Employee sick days

Visit type and time

Chief complaint

Diagnoses

Death records

SNOMED

ICD-9

Differential diagnosis

Procedures

CPT

ICD-9

Diagnostics (ordered)

PERSONAL HEALTH RECORDS

HOME TREATMENTS, MONITORS, TESTS

LOINC

Pathology, histology

ECG

Radiology

REPORTS

Diagnostics (results)

Lab values, vital signs

TRACINGS, IMAGES

Genetics

PATIENTS LIKEME.COM

23andMe.com

SNPs, arrays

Social history

Police records

Tobacco/alcohol use

DIGITAL CLINICAL NOTES

BLOGS

Family history

Ancestry.com

Symptoms

Indirect from OTC purchases

PHYSICAL EXAMINATIONS

TWEETS

Lifestyle

Fitness club memberships, grocery store purchases

CREDIT CARD PURCHASES

Socioeconomic

Census records, Zillow, LinkedIn

PAPER CLINICAL NOTES

FACEBOOK POSTINGS

Social network

Facebook friends, Twitter hashtags

Environment

Climate, weather, public health databases, HealthMap.org, GIS maps, EPA, phone GPS

News feeds

Probabilistic linkage to obtain new types of data

Probabilistic linkage to validate existing data or fill in missing data

Examples of biomedical data

1 2 Pharmacy data

1 2 Health care center (electronic health record) data

Claims data

Registry or clinical trial data

Data outside of health care system

Ability to link data to an individual

■ Easier to link to individuals

■ Harder to link to individuals

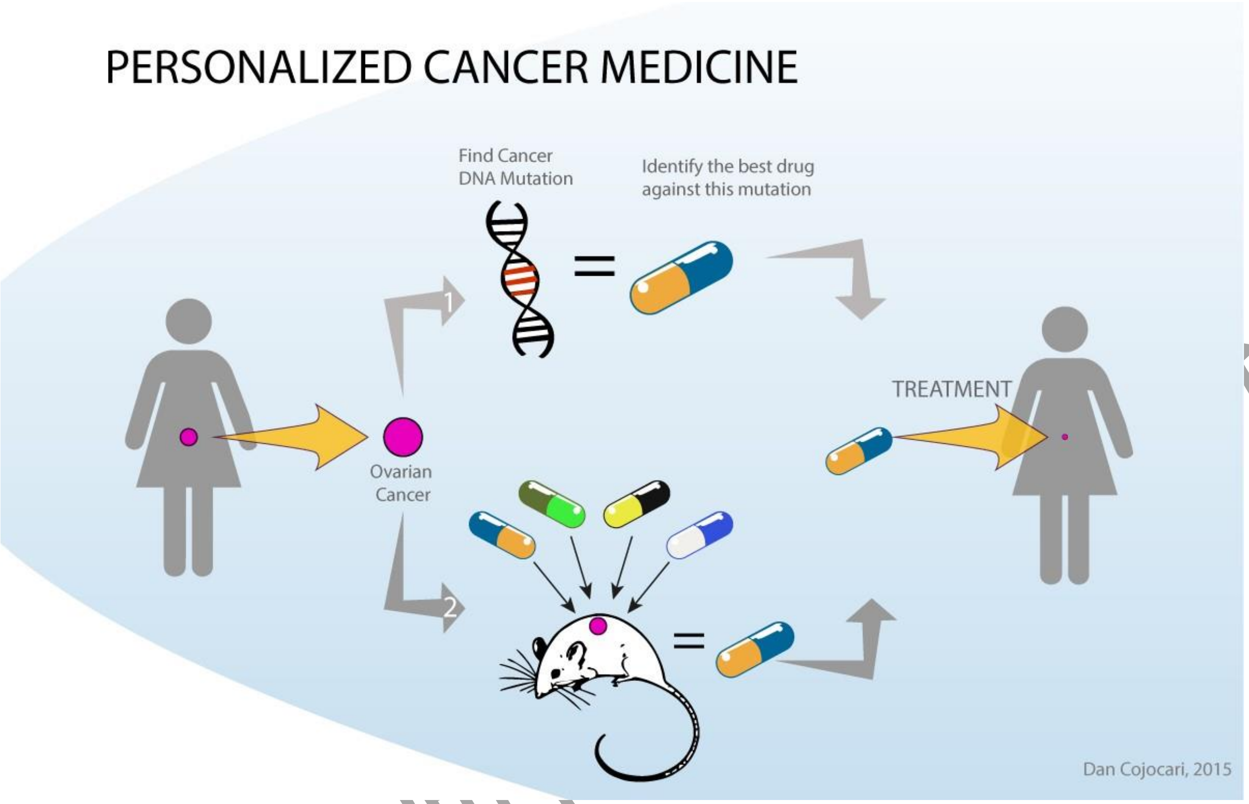
■ Only aggregate data exists

Data quantity

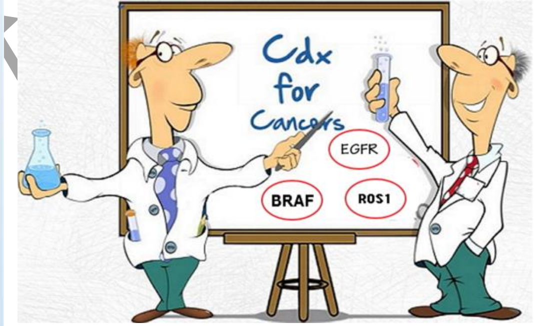
More

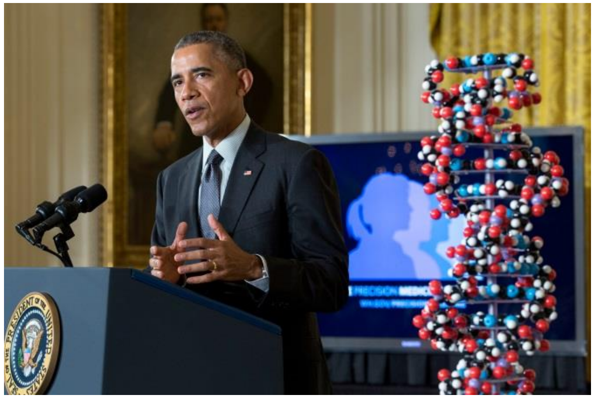
Less

Fig：潜在的高价值信息来源的图谱，可能与个人联系在一起用于医疗保健


PERSONALIZED CANCER MEDICINE
Find Cancer DNA Mutation
Identify the best drug against this mutation
1
=
Ovarian Cancer
TREATMENT
2
=
Dan Cojocari, 2015

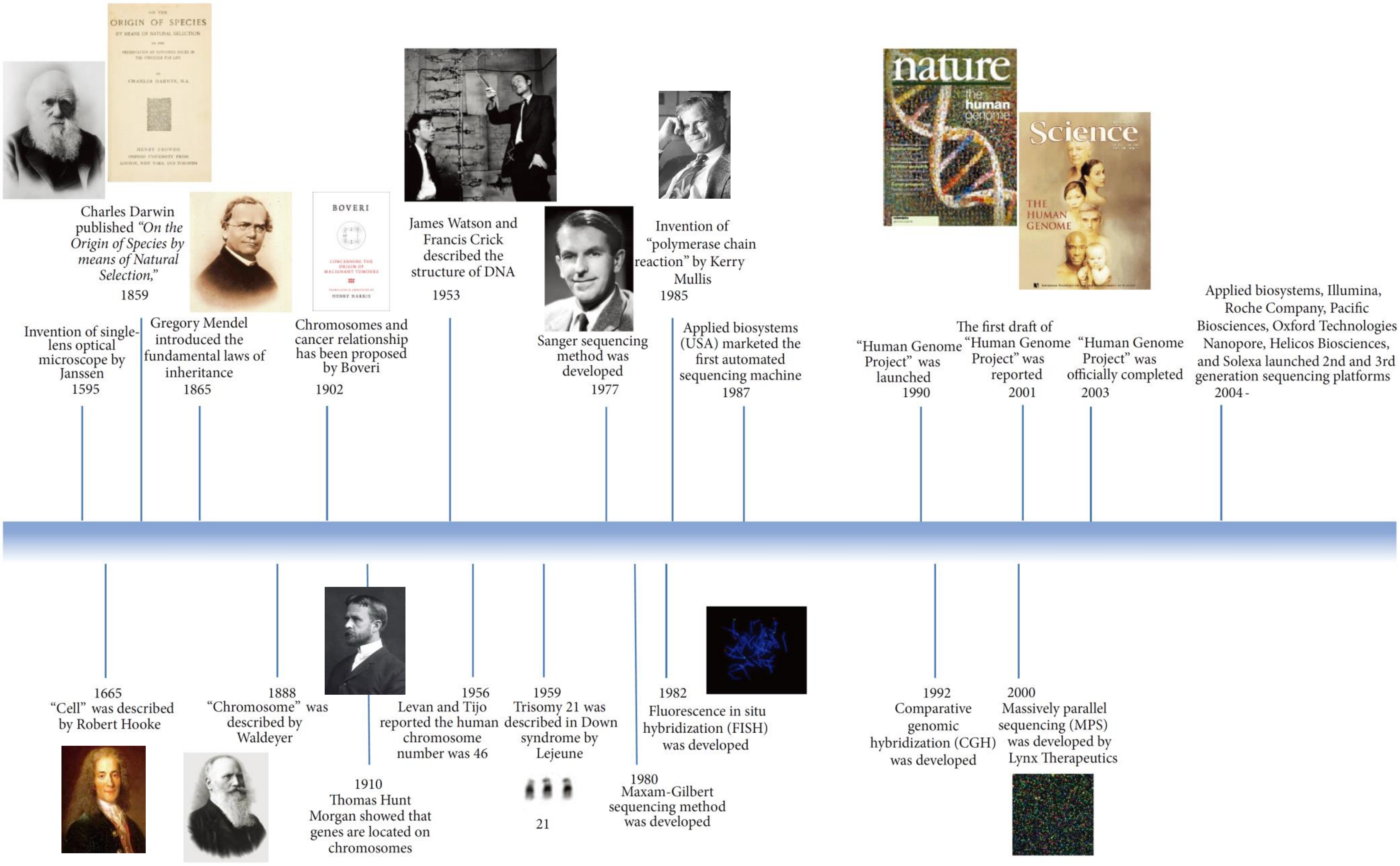
precision medicine for me





个性化医疗与精准医学

第6节：总结与展望



Invention of single-lens optical microscope by Janssen
1595

Charles Darwin published *"On the Origin of Species by means of Natural Selection,"*
1859

Gregory Mendel introduced the fundamental laws of inheritance
1865

Chromosomes and cancer relationship has been proposed by Boveri
1902

James Watson and Francis Crick described the structure of DNA
1953

Sanger sequencing method was developed
1977

Invention of "polymerase chain reaction" by Kerry Mullis
1985

Applied biosystems (USA) marketed the first automated sequencing machine
1987

"Human Genome Project" was launched
1990

The first draft of "Human Genome Project" was reported
2001

"Human Genome Project" was officially completed
2003

Applied biosystems, Illumina, Roche Company, Pacific Biosciences, Oxford Technologies Nanopore, Helicos Biosciences, and Solexa launched 2nd and 3rd generation sequencing platforms
2004 -

1665
"Cell" was described by Robert Hooke

1888
"Chromosome" was described by Waldeyer

1910
Thomas Hunt Morgan showed that genes are located on chromosomes

1956
Levan and Tijo reported the human chromosome number was 46

1959
Trisomy 21 was described in Down syndrome by Lejeune

1980
Maxam-Gilbert sequencing method was developed

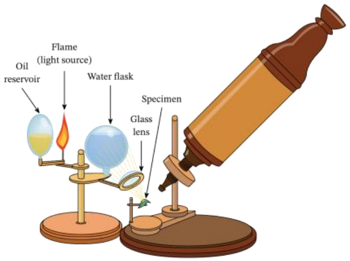
1982
Fluorescence in situ hybridization (FISH) was developed

1992
Comparative genomic hybridization (CGH) was developed

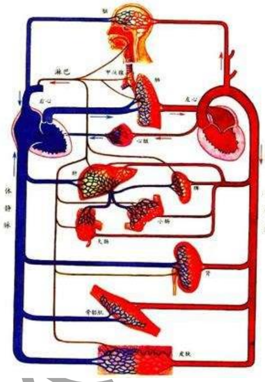
2000
Massively parallel sequencing (MPS) was developed by Lynx Therapeutics

❖ 6.1 生物学发展简史及特点

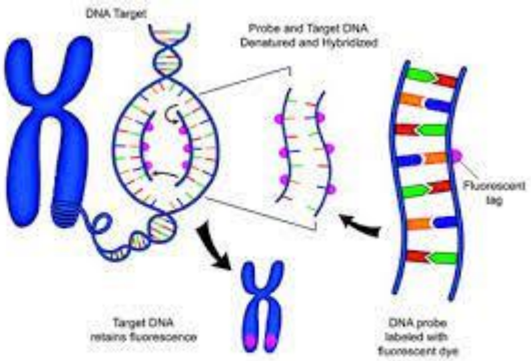
- 观察描述



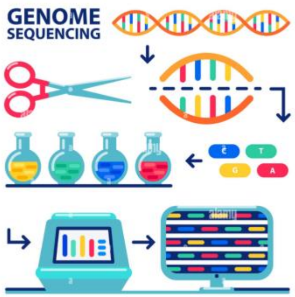
- 实验方法



- 分子生物学方法

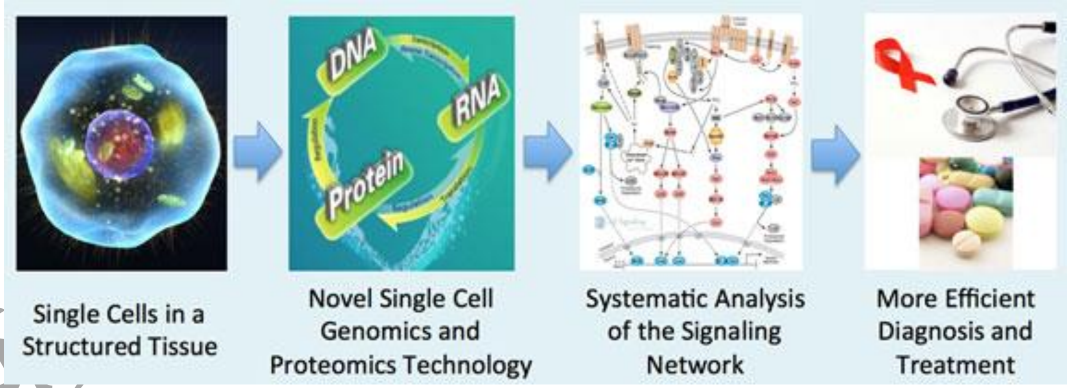


- 计算机辅助法、系统论



◆宏观到微观

◆单一到交叉



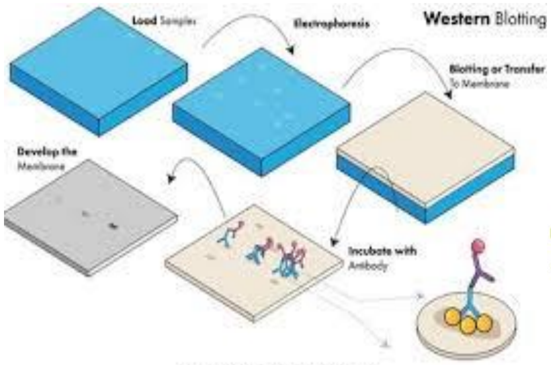
Single Cells in a Structured Tissue

Novel Single Cell Genomics and Proteomics Technology

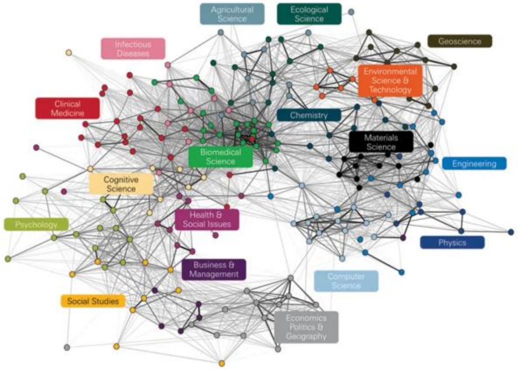
Systematic Analysis of the Signaling Network

More Efficient Diagnosis and Treatment

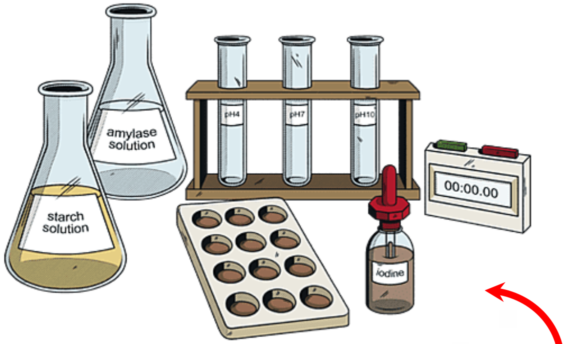
◆实验到计算

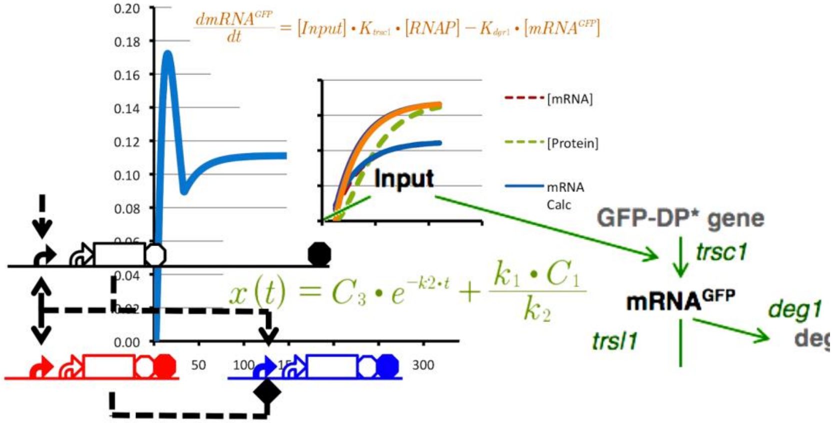


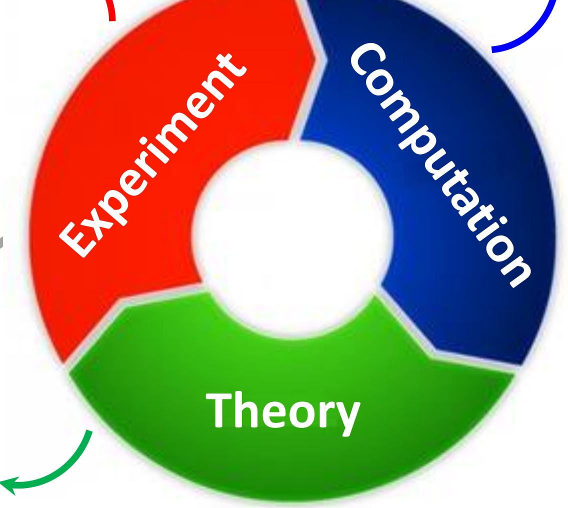
shutterstock.com · 1951327312

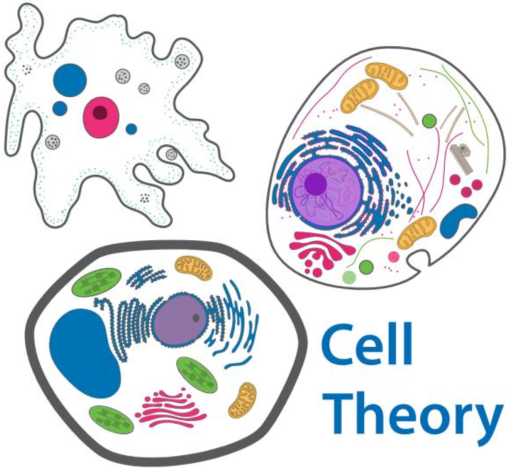


❖ 6.2 生物学知识的来源各异





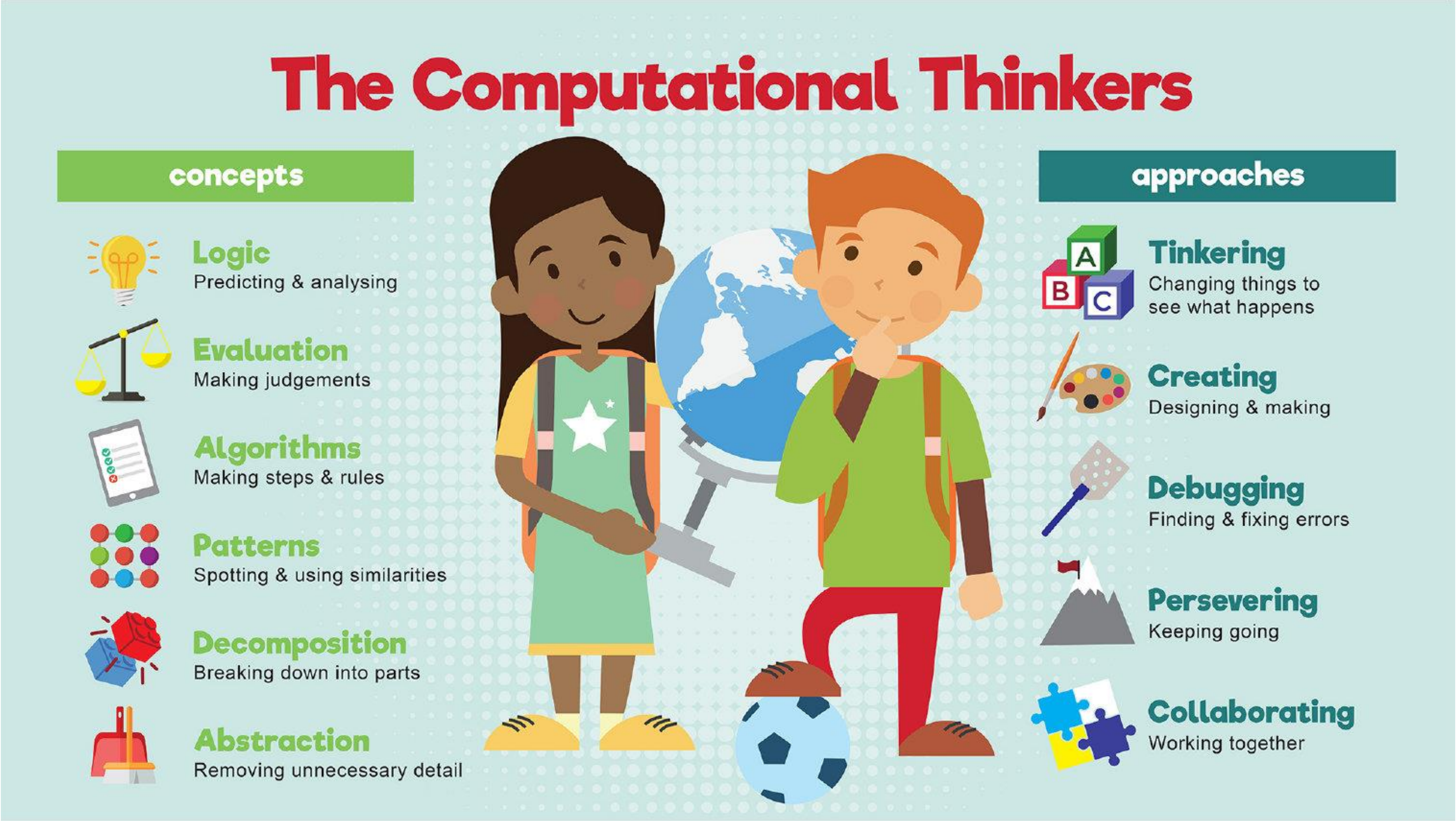




- ➢ 实验生物学
- ➢ 理论生物学
- ➢ 计算生物学

❖ 6.3 计算思维

运用计算机科学的基本理念，进行问题求解、系统设计及理解人类行为。即一种运用计算机科学的基本理念来解决问题的思考方式。



We are all computational thinkers here!

重庆师范大学

CHONG QING NORMAL UNIVERSIYT

Thanks for your attention!

Acknowledgement

College of Life Sciences, Chongqing Normal University

2022, Chongqing of P. R. C