



**VIỆN CÔNG NGHỆ THÔNG TIN
VÀ TRUYỀN THÔNG**



Ứng dụng toán học và AI trong công nghệ sinh học thế hệ mới

TS. Đỗ Văn Hoàn

Trung tâm Toán ứng dụng và Tin học, Viện CNTT-TT

Học viện Kỹ thuật Quân sự

Bacterial pandemics in history

Mycobacterium tuberculosis: infected and killed human 9,000 years ago

- Remains a global threat, with an estimated 16-33 million cases and 200,000 deaths annually

Yersinia pestis:

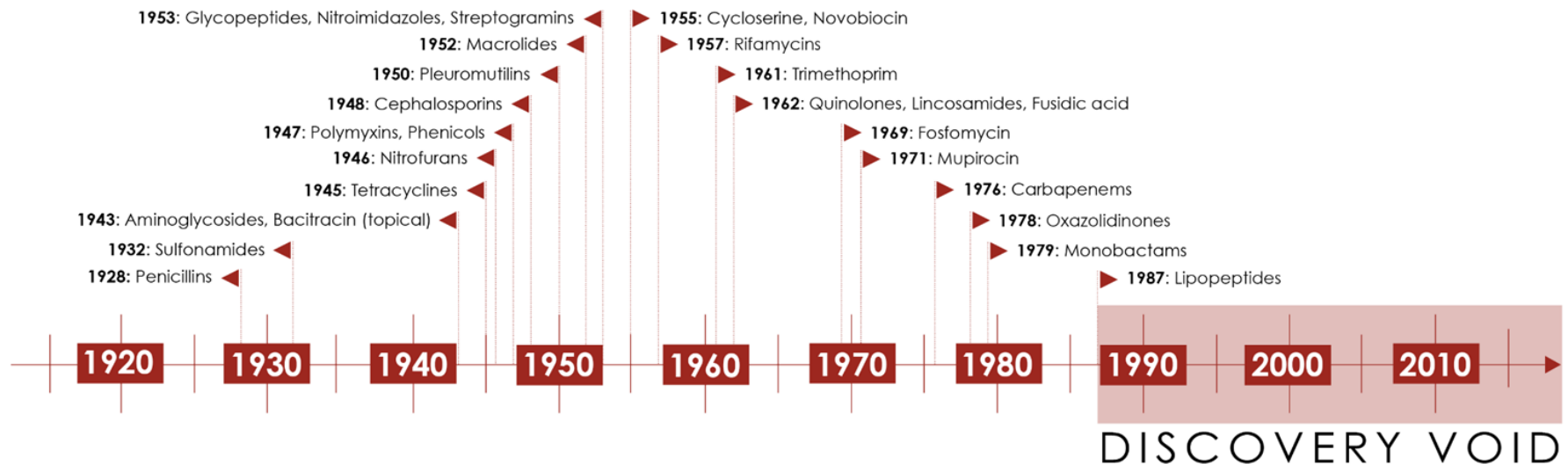
- Plague of Justinian 6th century: killed half of European population
- Black Death (14th century): reduced world population by a third

Salmonella enterica aka typhus:

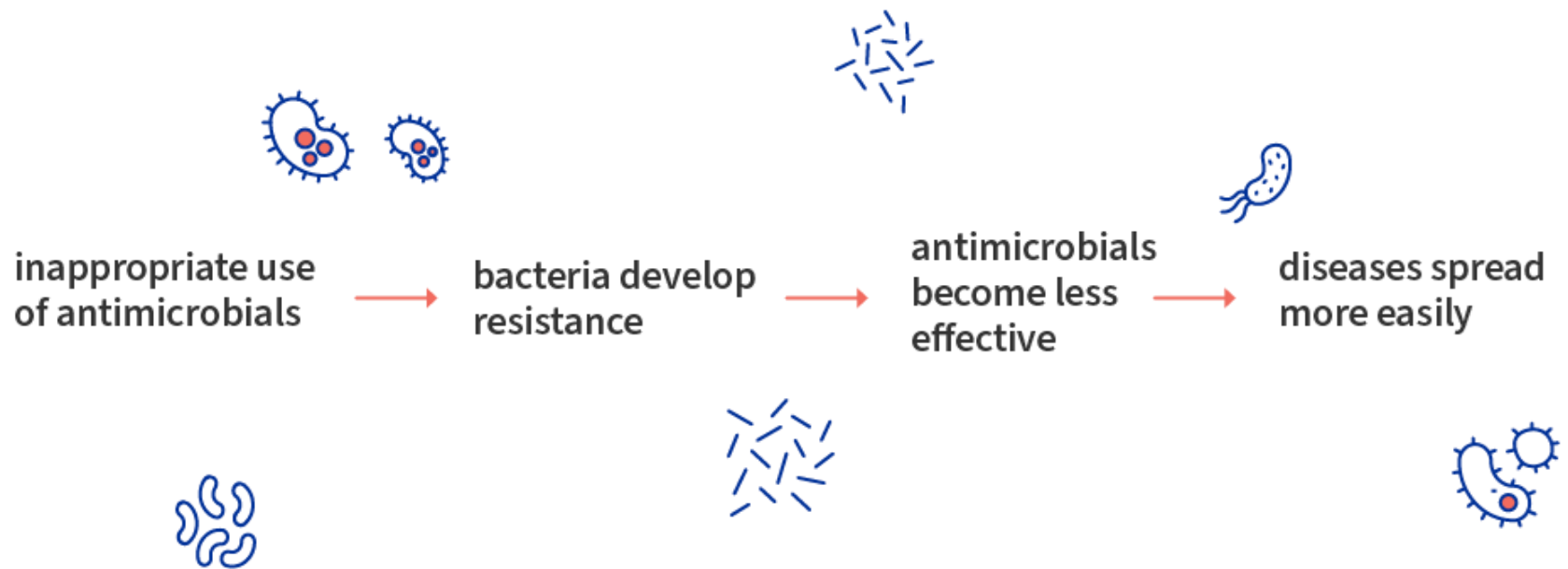
- Plague of Athens, killing 1/4 army, led to the collapse of the first democratic society
- Together with smallpox and measles, wiped 90% of American native population in 16th century

The Antibiotic: A Turning Point in History

- Antibiotic was discovered in 1930s-1940s
 - Nobel prize to Flemming, Florey and Chain saves more lives than all others combined.
- Increased life expectancy significantly, eg. 48 to 79 years in US
- 1950s and 1960s are the golden age of antibiotics discovery, but no new classes since 1970s



Antibiotics resistance



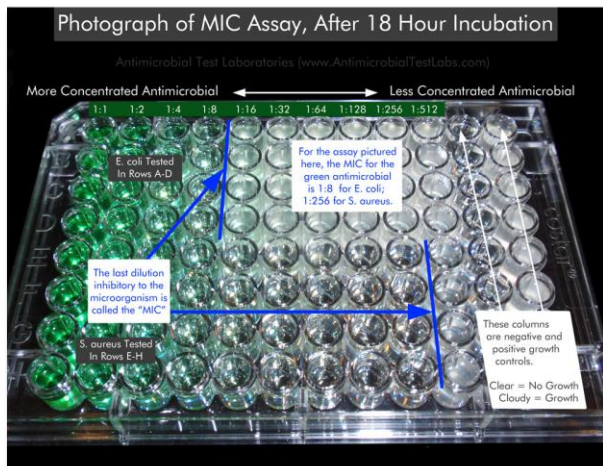
leads to a minimum of 700,000 deaths each year!

**A need of powerful tools to
fight against antimicrobial
resistance (AMR)**

Antibiotics resistance prediction

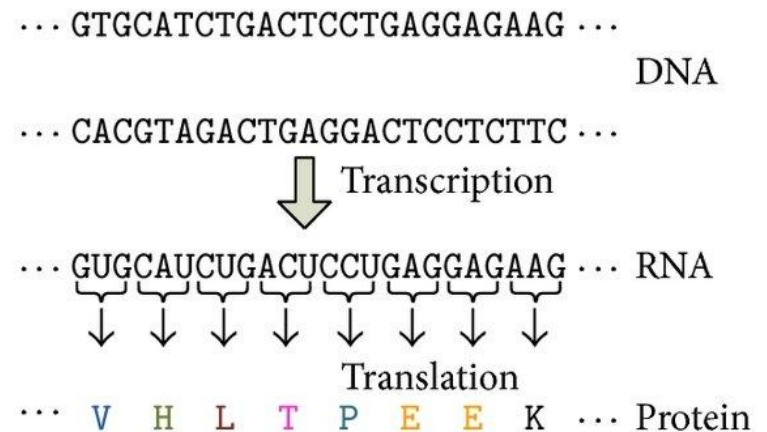
- **Problem:** Predict whether a bacterial strain will exhibit resistance or susceptibility to specific antibiotics.
- Crucial in healthcare and microbiology as it aids in determining appropriate treatment strategies for bacterial infections.
- Predictive models for antibiotic resistance utilize various features, such as **phenotypic data**, **genomic information**, and environmental factors.

Pros: Accurate
Cons: slow and expensive



Experiment method: MIC test

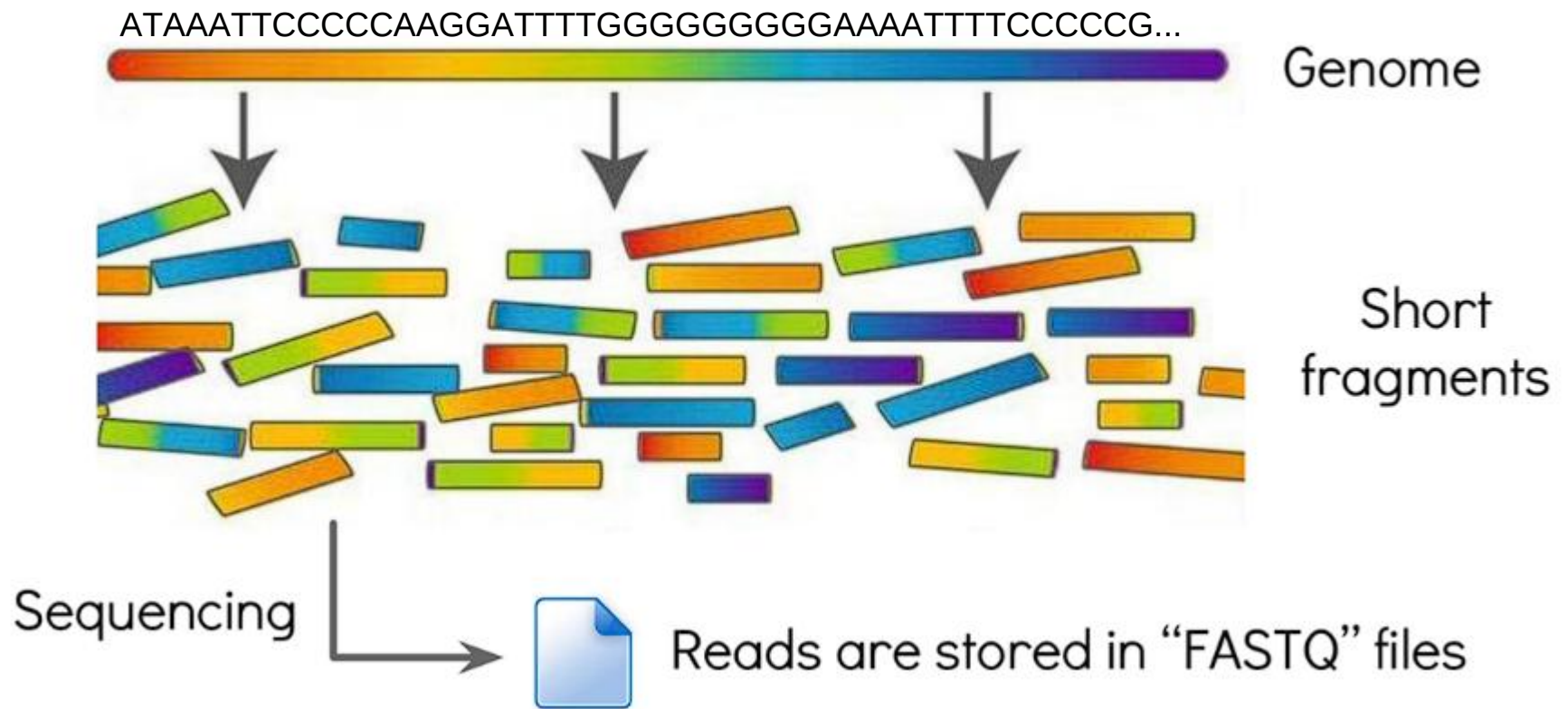
Pros: Fast, effective
Cons: Need of training data



Computational method: Use genomic info

DNA Sequencing

Is the process of determining the order of nucleotides in DNA

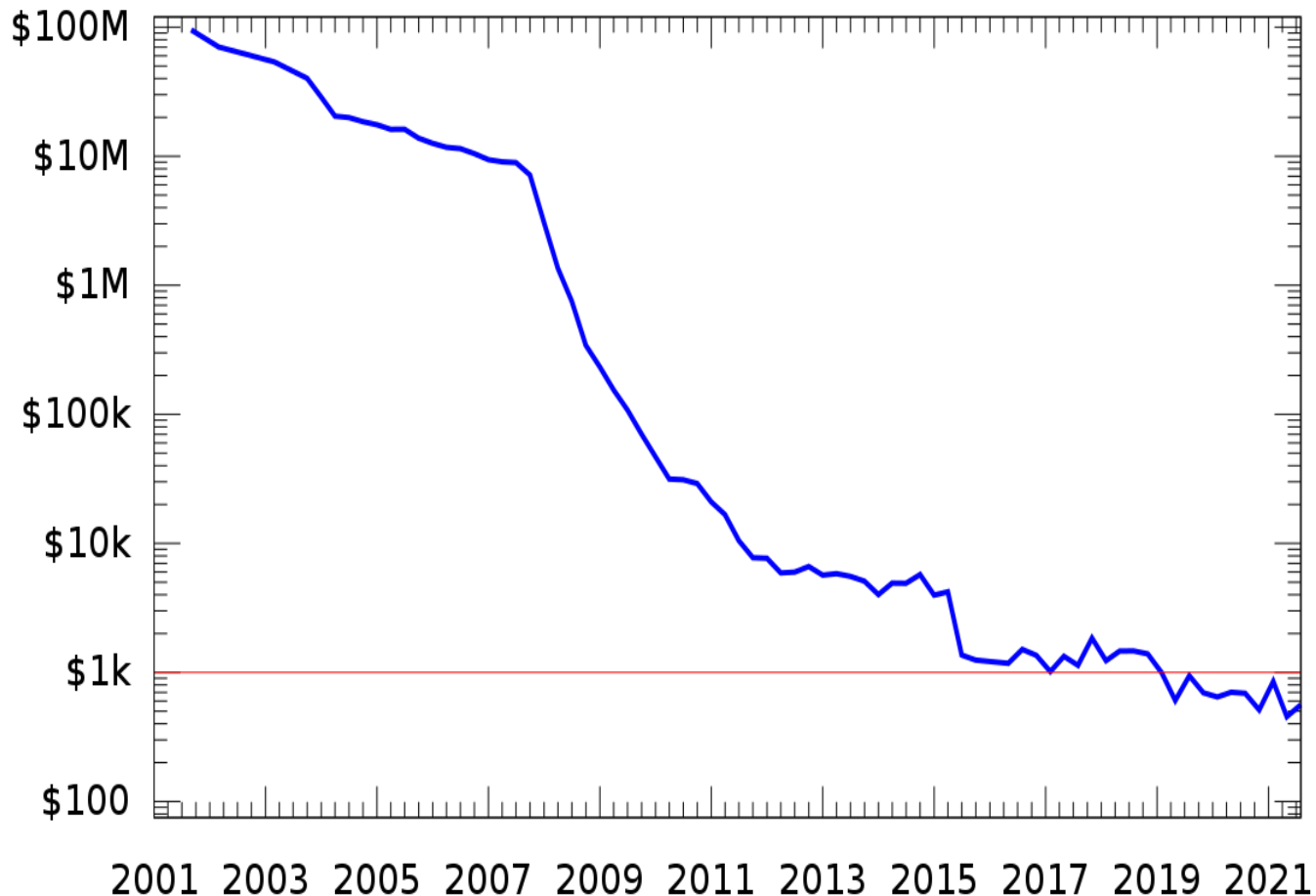


A brief history of DNA sequencing

- Robert Holley, an American biochemist, sequenced the first tRNA in 1965, for which he was awarded the Nobel Prize in 1986.
- In the 1970s, Fredrick Sanger and his collaborators were working on an alternative DNA sequencing method. The “chain termination method,” developed in 1977, uses radio labeled and partially digested oligonucleotides to analyze pieces of the molecule.
- Later known as the Sanger method, it was widely used throughout the 2000s and earned Sanger a Nobel Prize in 1980.

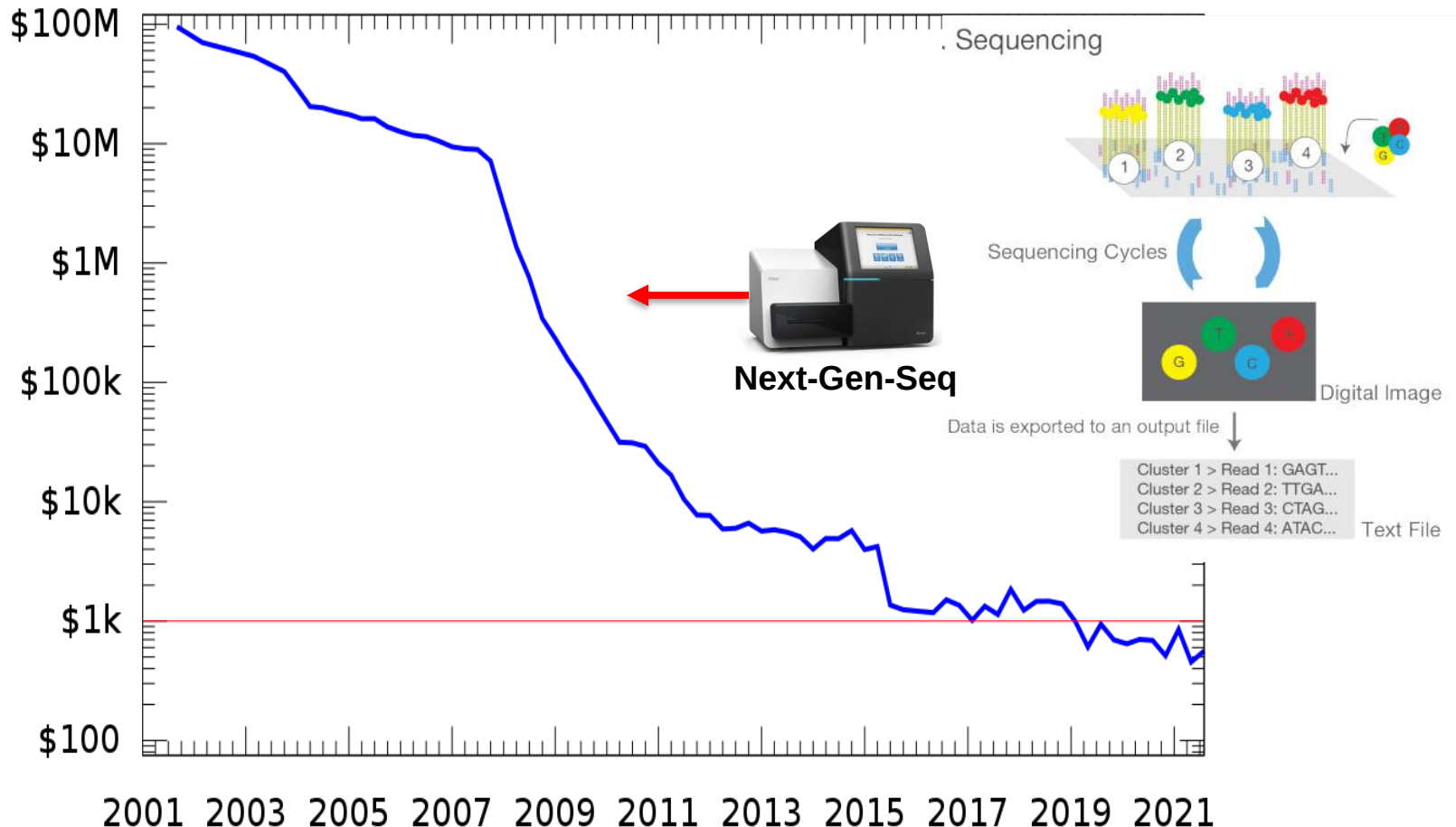
A brief history of DNA sequencing

Cost to sequence a human genome (USD)



A brief history of DNA sequencing

Cost to sequence a human genome (USD)



Next Generation Sequencing

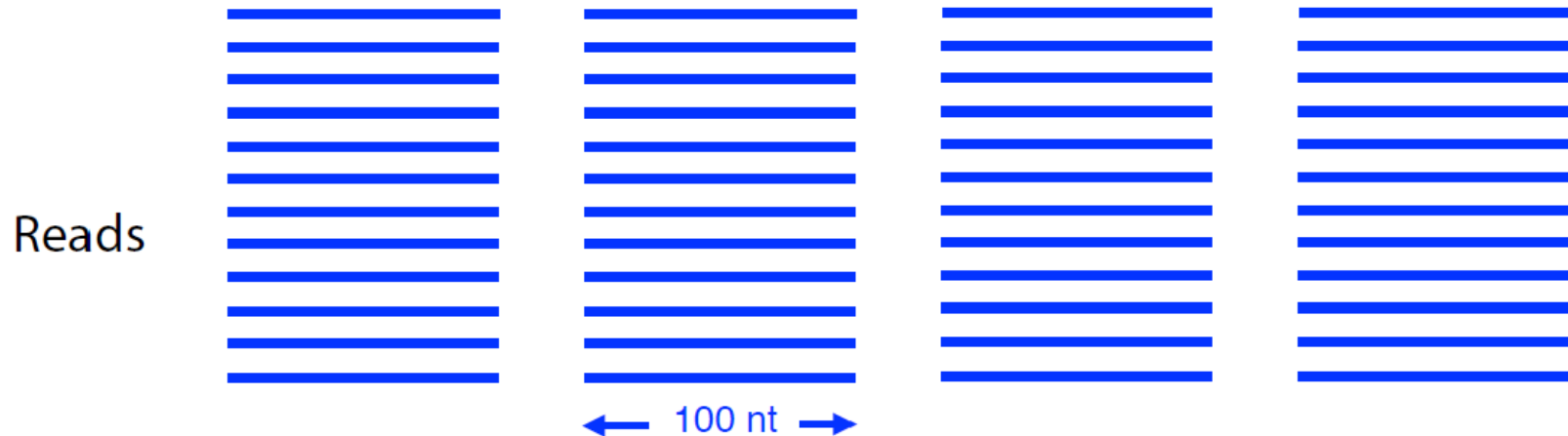
Reads

GTATGCACGCGATAG	TATGTCGCAGTATCT	CACCCTATGTCGCAG	GAGACGCTGGAGCCG
TAGCATTGCGAGACG	GGTATGCACGCGATA	TGGAGCCGGAGCACC	CGCTGGAGCCGGAGC
TGTCTTTGATTCCTG	CGCGATAGCATTGCG	GCATTGCGAGACGCT	CCTATGTCGCAGTAT
GACGCTGGAGCCGGA	GCACCCTATGTCGCA	GTATCTGTCTTTGAT	CCTCATCCTATTATT
TATCGCACCTACGTT	CAATATTCGATCATG	GATCACAGGTCTATC	ACCCTATTAACCACT
CACGGGAGCTCTCCA	TGCATTTGGTATTTT	CGTCTGGGGGGTATG	CACGCGATAGCATTG
GTATGCACGCGATAG	ACCTACGTTCAATAT	TATTTATCGCACCTA	CCACTCACGGGAGCT
GCGAGACGCTGGAGC	CTATCACCCCTATTAA	CTGTCTTTGATTCCT	ACTCACGGGAGCTCT
CCTACGTTCAATATT	GCACCTACGTTCAAT	GTCTGGGGGGTATGC	AGCCGGAGCACCTA
GACGCTGGAGCCGGA	GCACCCTATGTCGCA	GTATCTGTCTTTGAT	CCTCATCCTATTATT
TATCGCACCTACGTT	CAATATTCGATCATG	GATCACAGGTCTATC	ACCCTATTAACCACT
CACGGGAGCTCTCCA	TGCATTTGGTATTTT	CGTCTGGGGGGTATG	CACGCGATAGCATTG

Your genome

CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCCTATGTCGCAGTATCTGTCTTTGATTCCTG

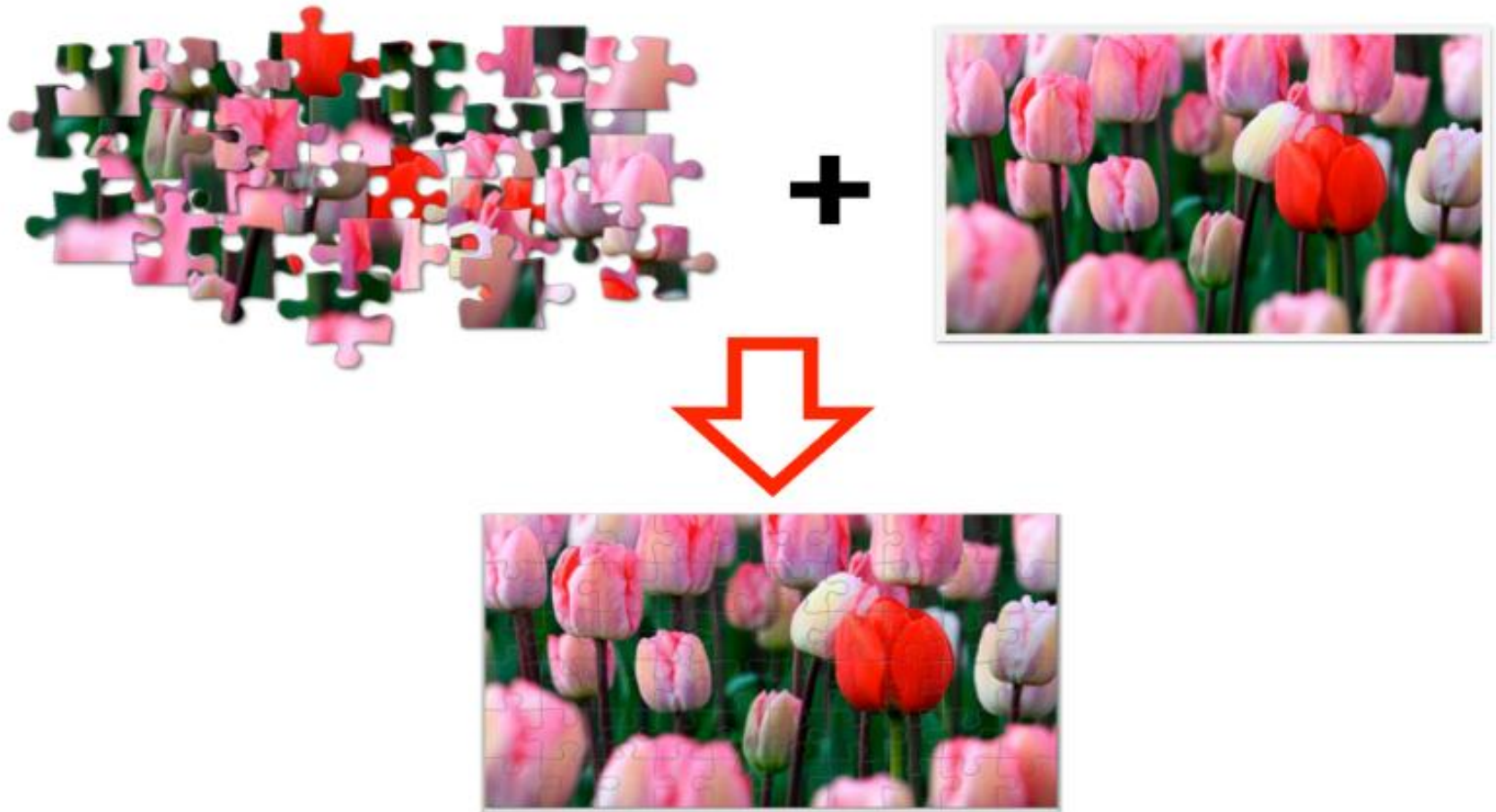
Next Generation Sequencing



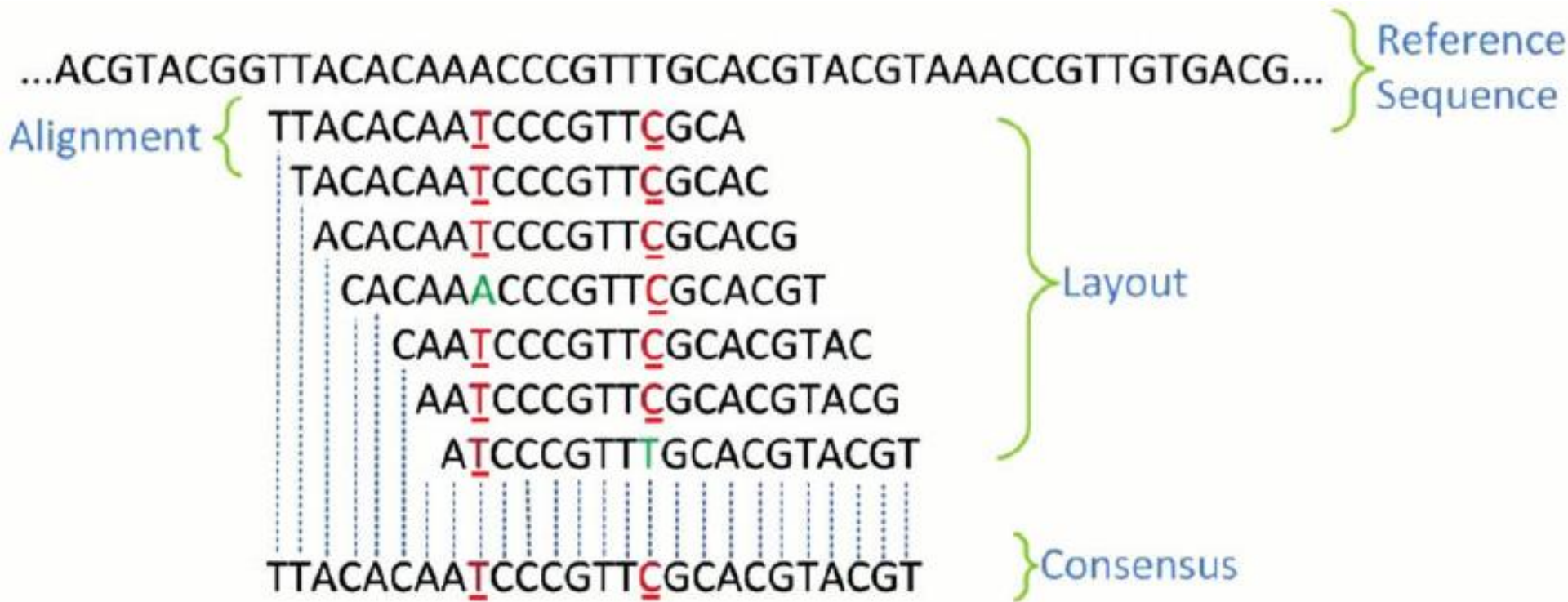
Your genome



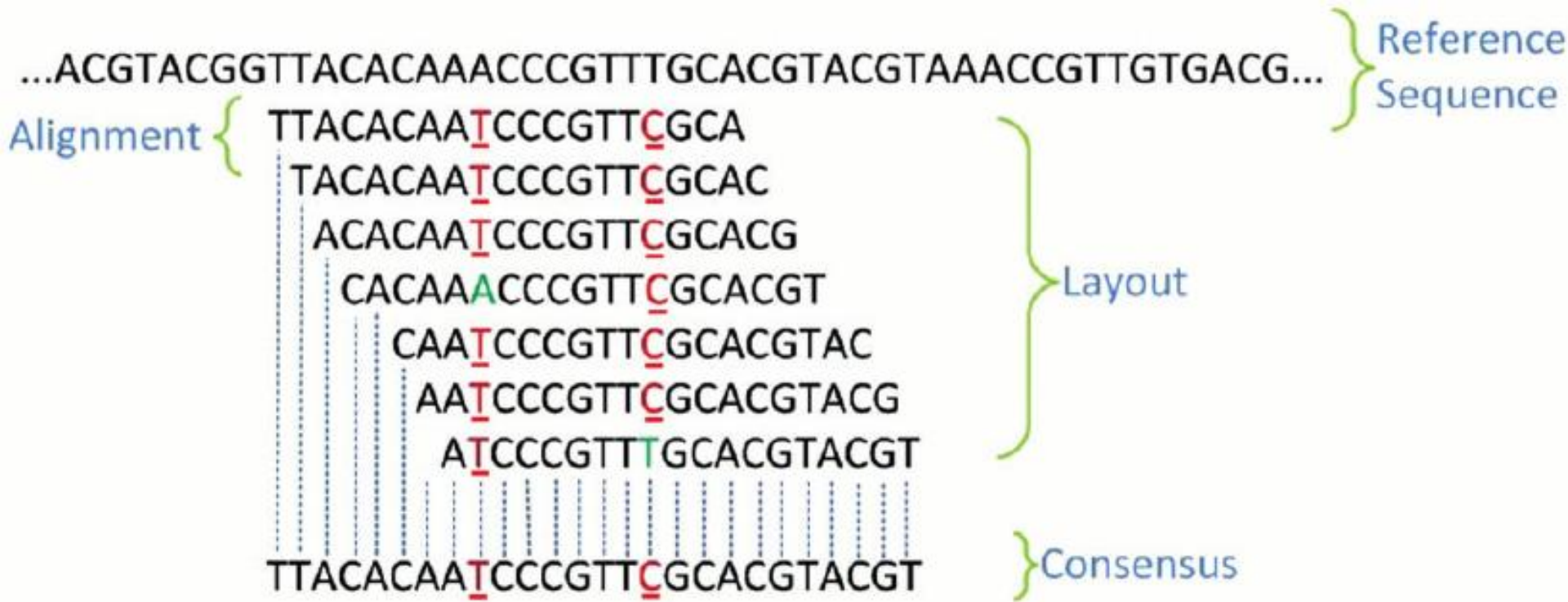
Next Generation Sequencing



Next Generation Sequencing



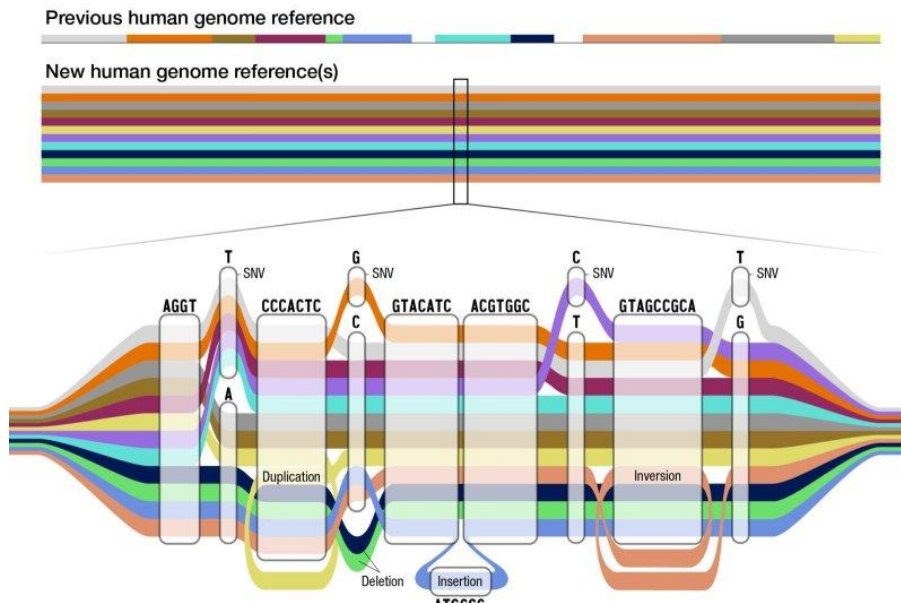
Next Generation Sequencing



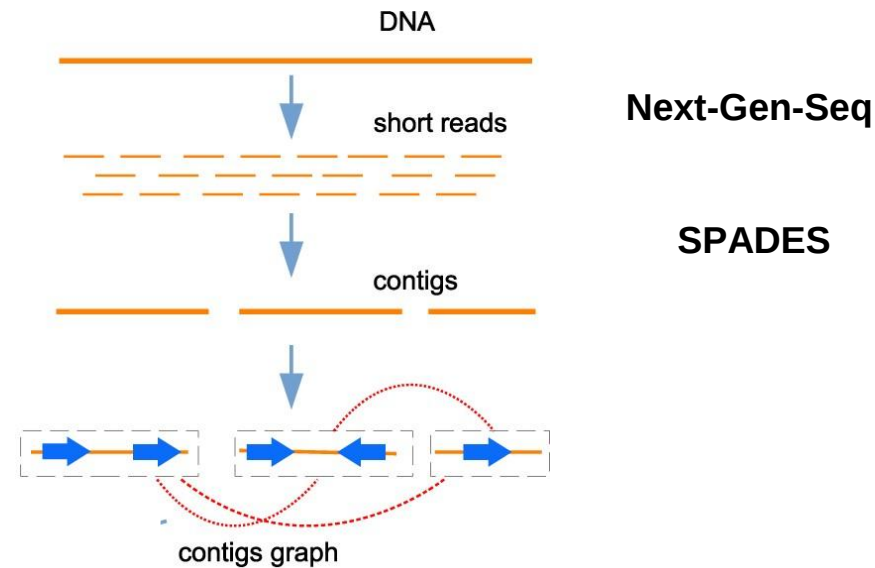
Limitations of using a single reference genome.

- The bacterial genomes are quite dissimilar, compared to human genome.
- Different bacterial strains can have unique genes and mutations not present in the reference genome, affecting accuracy in identifying pathogenicity, resistance, and other traits.

Pangenome: graph reference

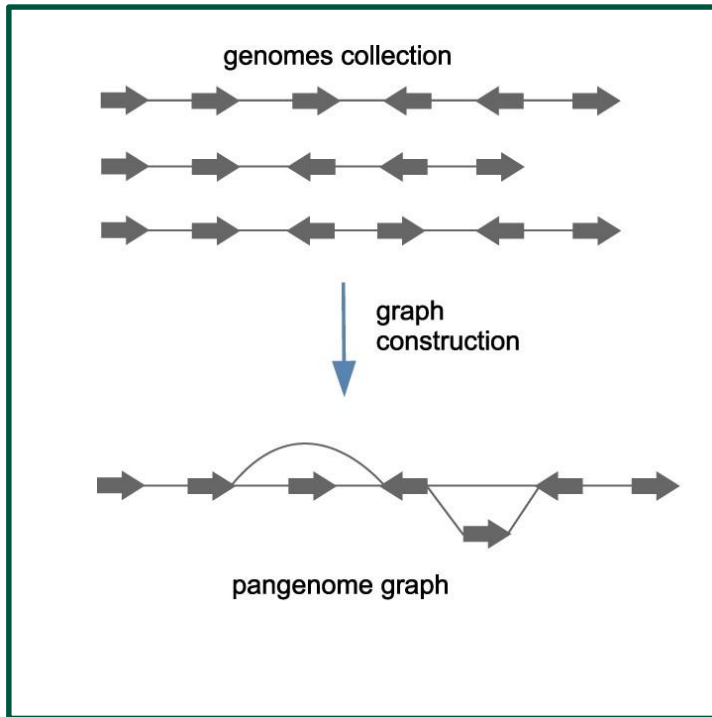


Pasa (Pangenome-based Scaffolding)

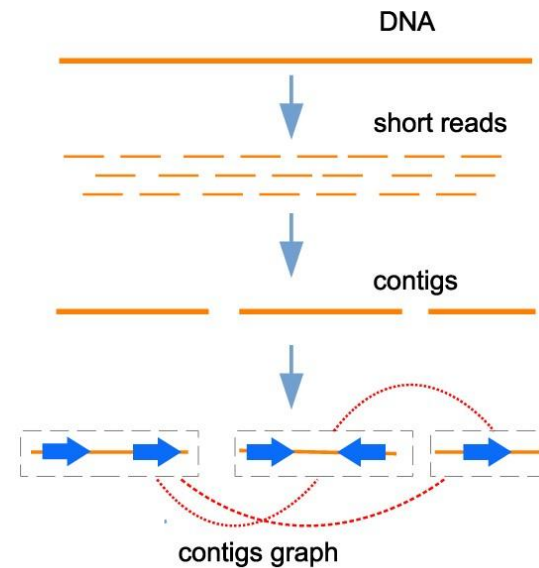


Our goal: **Reconstruct bacterial genomes** from **short DNA segments** (produced by NGS) using the bacterial **pangenome graph**.

Pasa (Pangenome-based Scaffolding)



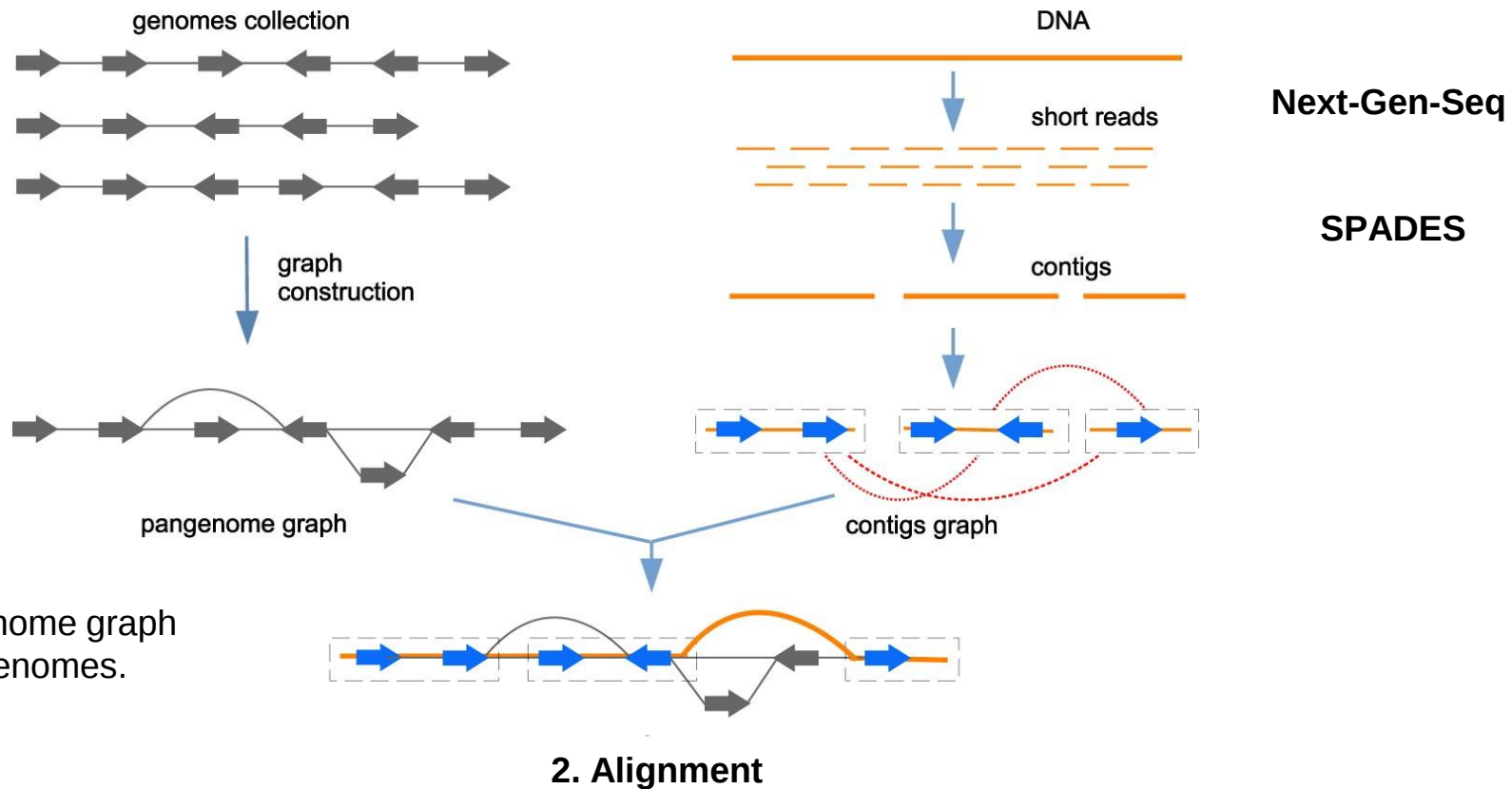
**1. Build a pangenome graph
from reference genomes**



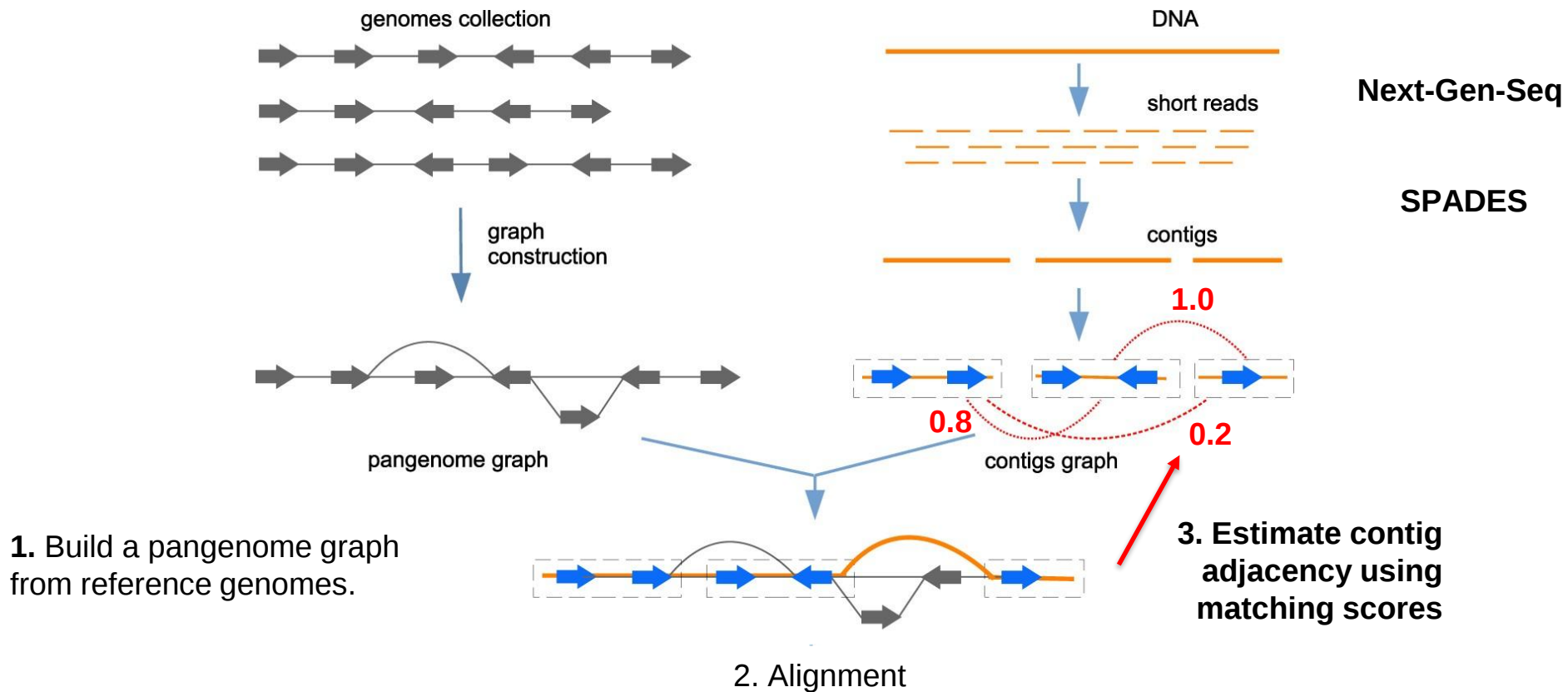
Next-Gen-Seq

SPADES

Pasa (Pangenome-based Scaffolding)

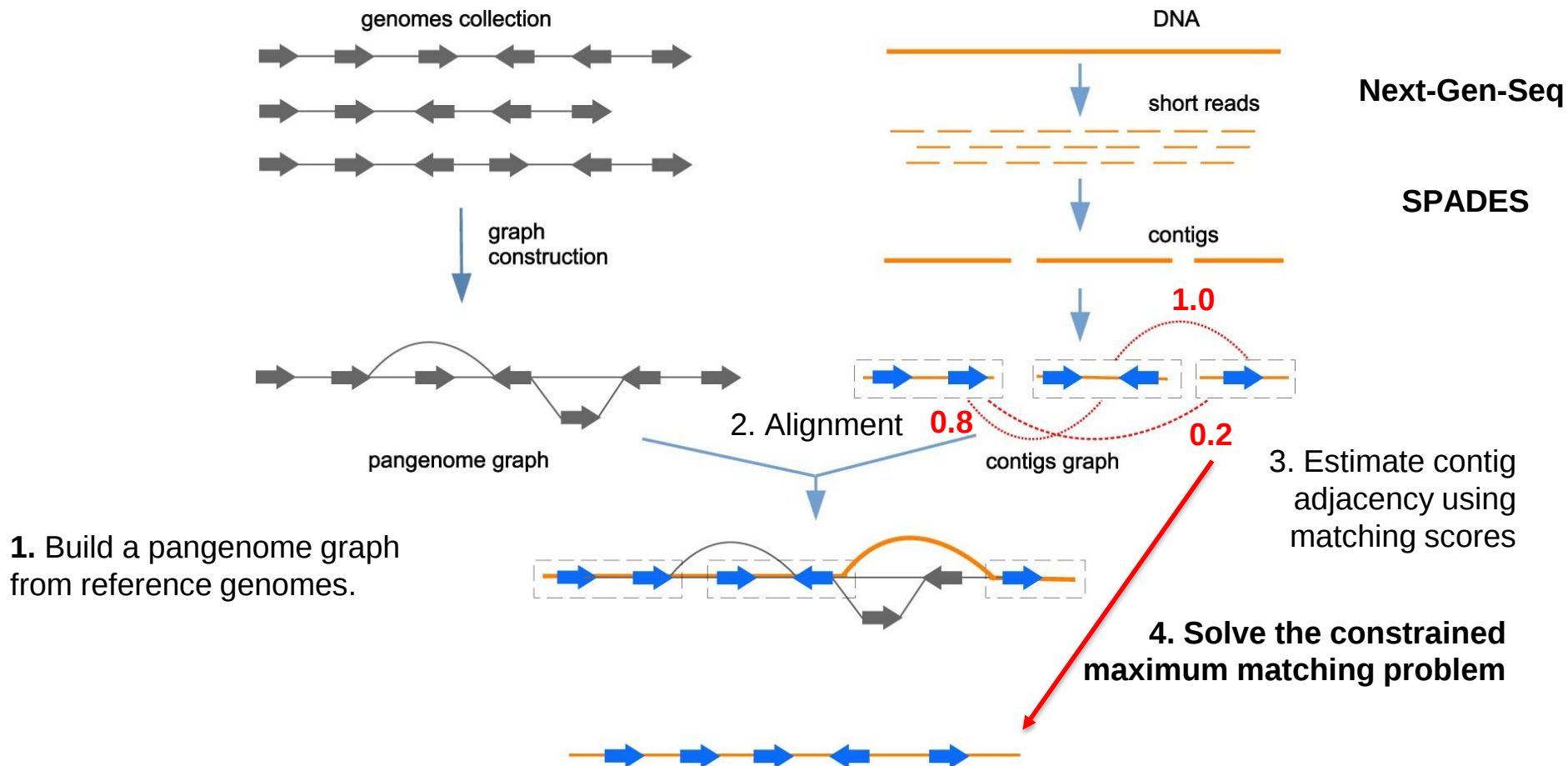


Pasa (Pangenome-based Scaffolding)

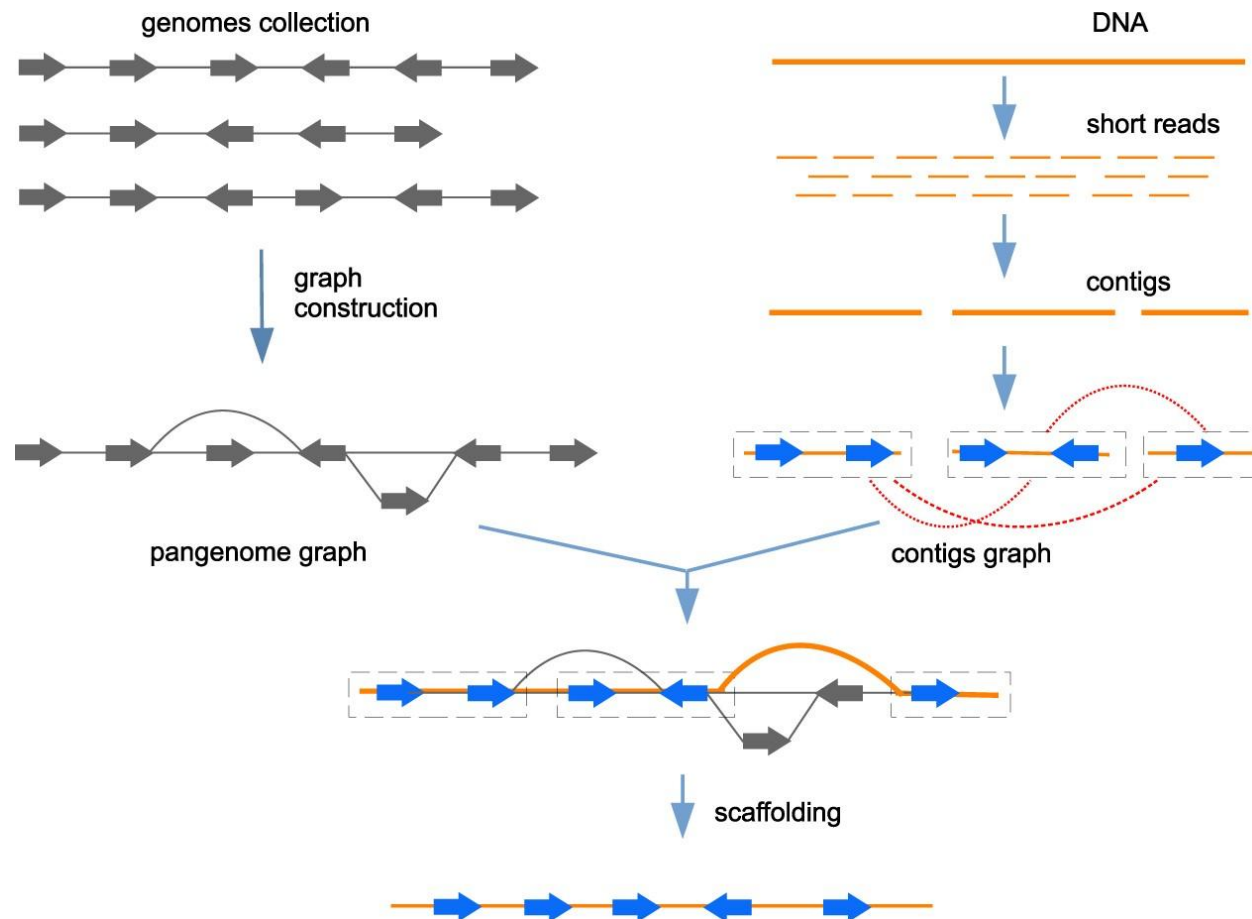


"The matching score measures the likelihood of two contigs being adjacent"

Pasa (Pangenome-based Scaffolding)

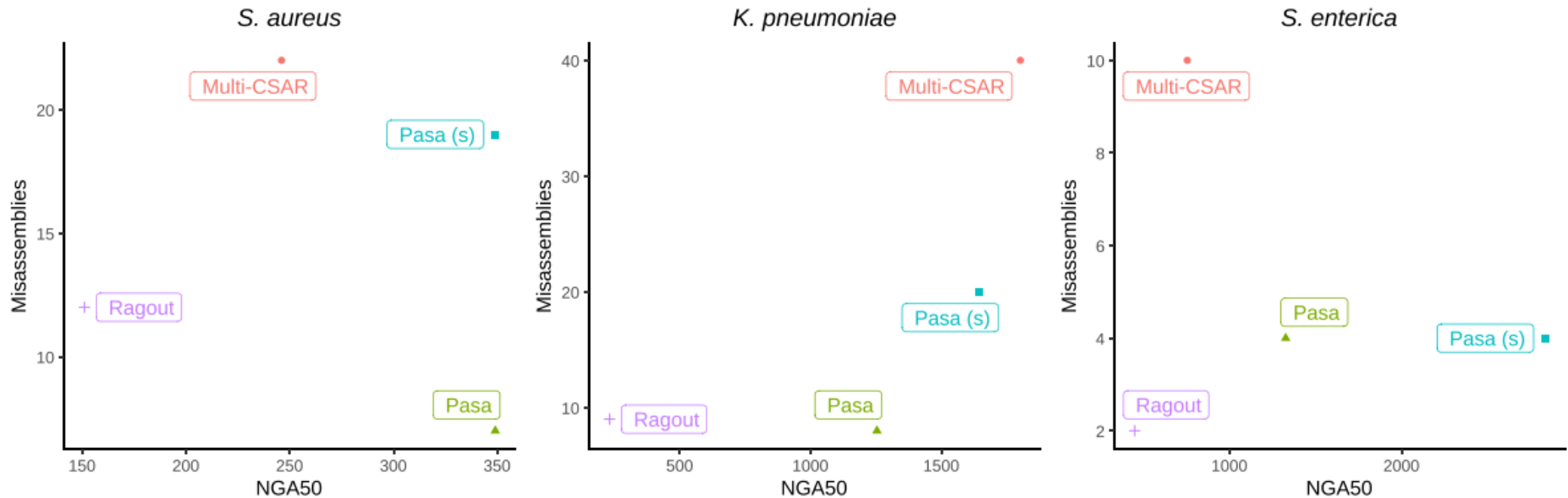


Pasa (Pangenome-based Scaffolding)



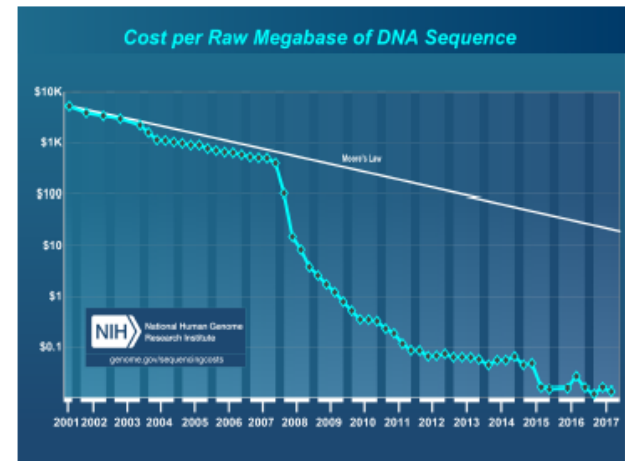
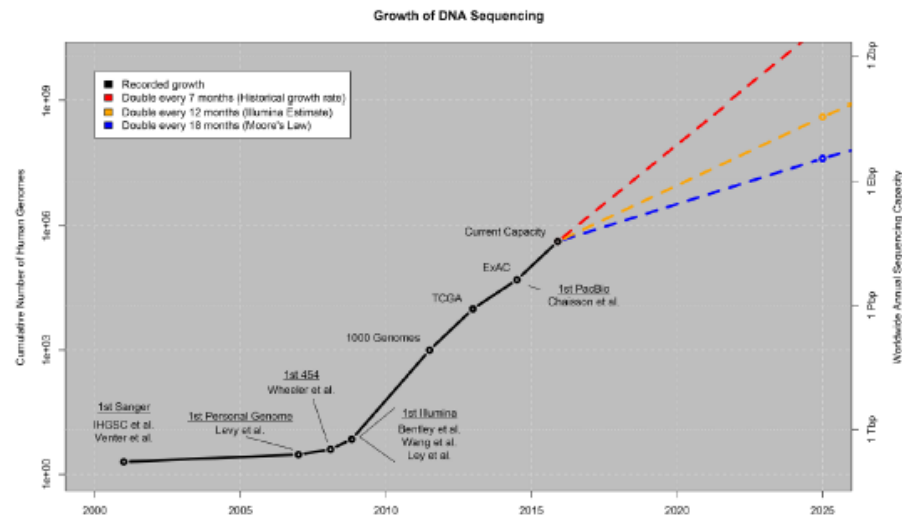
Van Hoan Do, et al., Pasa: leveraging population pangenome graph to scaffold prokaryote genome assemblies, *Nucleic Acids Research*, 2024.

Pasa vs competing methods



Van Hoan Do, et al., Pasa: leveraging population pangenome graph to scaffold prokaryote genome assemblies, *Nucleic Acids Research*, 2024.

Genomics is the biggest data source!



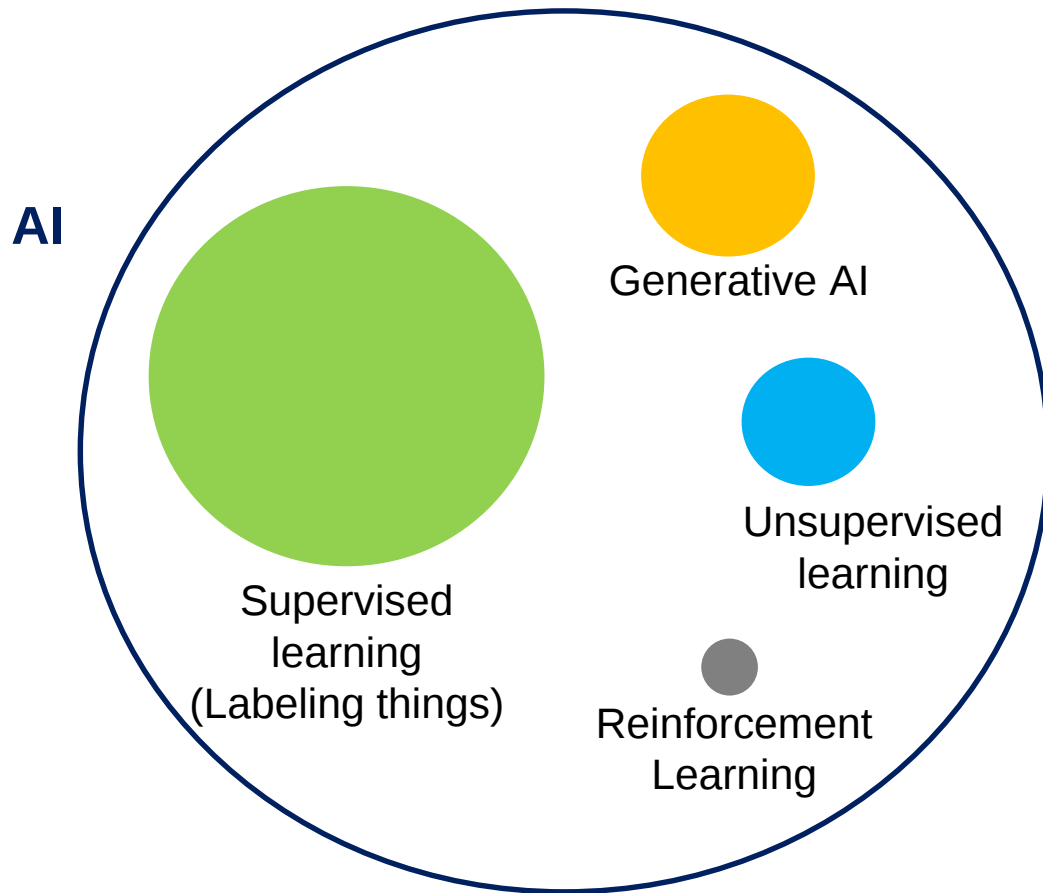
Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

Genomics will be the most demanding domain by 2025!

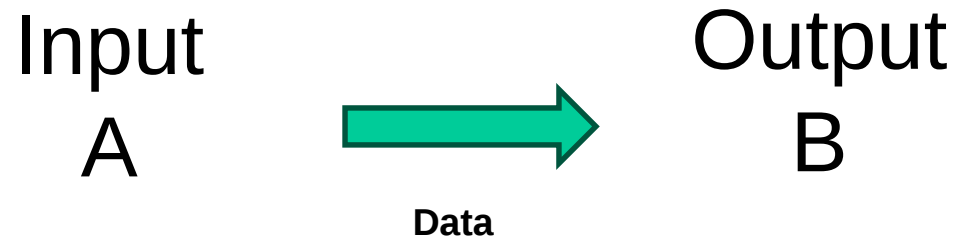
Stephens Z et al, Big Data: Astronomical or Genomical? PLOS Biology, 2015

Wetterstrand KA. DNA Sequencing Costs. www.genome.gov/sequencingcostsdata.

Data driven research tools



Supervised learning



Antibiotic resistance prognosis

chr11:5246500-5248500 (reverse strand):

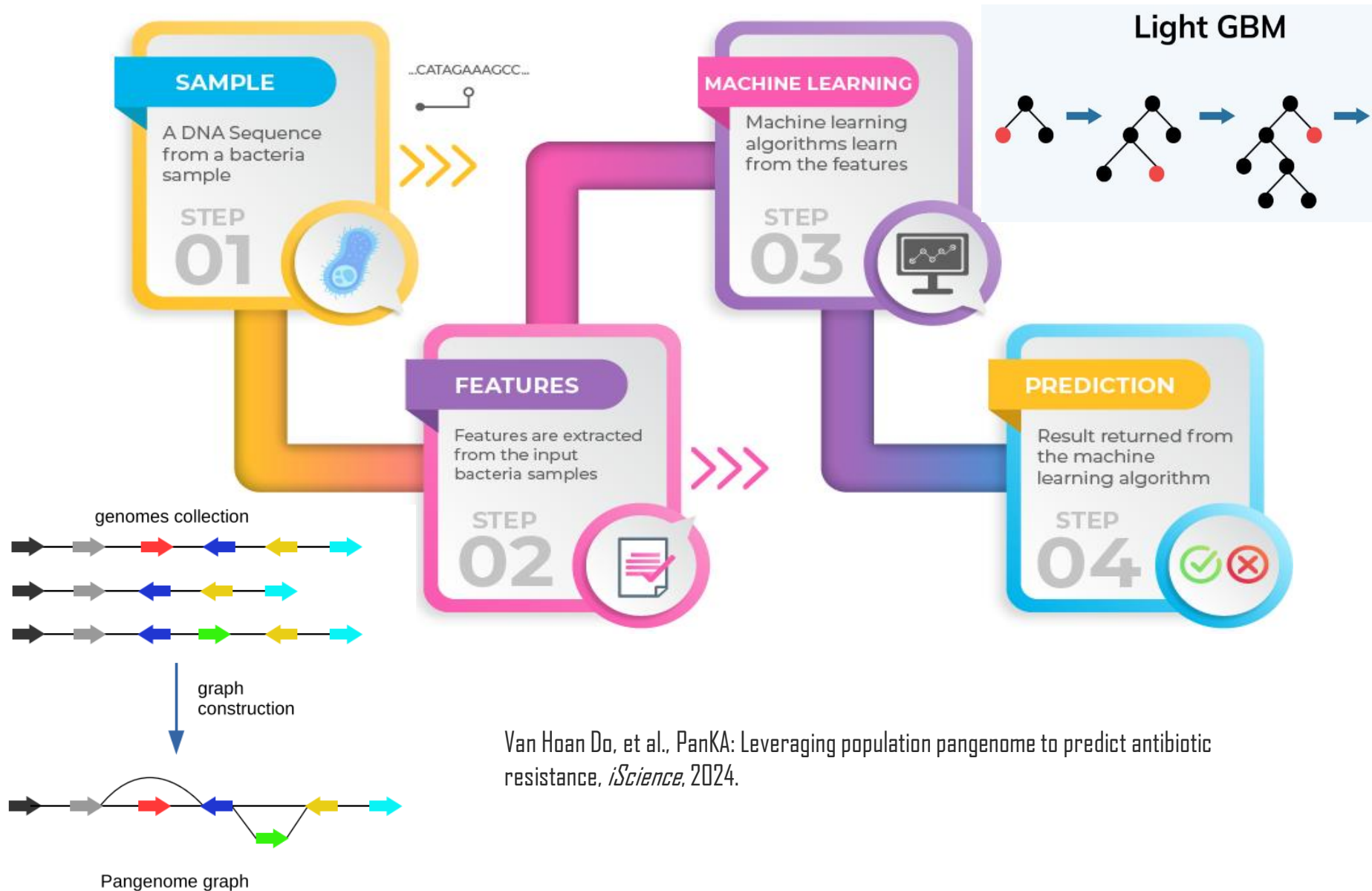
```
ATATCTTAGAGGGAGGGCTGAGGGTTTGAAGTCCAACCTCTAAGCCAGTGCCAGAAGAGCCAAGGACAGGTACGGCTGTC
ATCACTTAGACCTCACCTGTGGAGCCACACCCTAGGGTTGGCCAATCTACTCCCAGGAGCAGGGAGGGCAGGAGCCAGG
GCTGGGCATAAAAGTCAGGGCAGAGCCATCTATTGCTTACATTTGCTTCTGACACAACCTGTGTTCACTAGCAACCTCAAA
CAGACACCATGGTGATCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTT
GGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGA
CAGAGAAGACTCTTGGGTTTCTGATAGGCACTGACTCTCTGCTATTGGTCTATTTTCCACCCTTAGGCTGCTGGTG
GTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGGCAACCCTAAGGT
GAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCA
CACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAAGCTTCAGGGTGAGTCTATGGGACGCTTGATGTTTT
CTTCCCTTCTTTTCTATGGTTAAGTTCATGTCATAGGAAGGGGATAAGTAACAGGGTACAGTTTAGAATGGGAAACAG
ACGAATGATTGCATCAGTGGAAGTCTCAGGATCGTTTTAGTTTCTTTTATTTGCTGTTTATAACAATTGTTTTCTTT
GTTTAATTCTTGCTTTCTTTTTTTTCTTCTCCGCAATTTTACTATTATACTTAATGCCTTAACATTGTGTATAACAAA
AGGAAATATCTCTGAGATACATTAAAGTAACTTAAAAAAAACCTTTACACAGTCTGCCTAGTACATTACTATTTGGAATAT
ATGTGTGCTTATTTGCATATTCATAATCTCCCTACTTTATTTTCTTTTATTTTAAATTGATACATAATCATTATACATAT
TTATGGGTTAAAGTGTAATGTTTTAATATGTGTACATATTGACCAAATCAGGGTAATTTTGCAATTTGTAATTTAAAA
AATGCTTTCTTCTTTAATATACTTTTTGTTTATCTTATTTCTAATACTTTCCCTAATCTCTTCTTTCAGGGCAATAA
TGATACAATGTATCATGCCTCTTTGCACCATCTAAAGAATAACAGTGATAATTTCTGGGTTAAGGCAATAGCAATATCT
CTGCATATAAATATTTCTGCATATAAATTGTAAGTATGTAAGAGGTTTTCATATTGCTAATAGCAGCTACAATCCAGCTA
CCATTCTGCTTTTATTTTATGGTTGGGATAAGGCTGGATTATTCTGAGTCCAAGCTAGGCCCTTTGCTAATCATGTTCA
TACCTCTTATCTTCTCCACAGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACC
CCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCACAAGTATCACTAAGCTCGCTT
```



**Machine Learning
(supervised learning)**

Antibiotic resistance
Genetic diseases (cancer)
Plasmids detection

Antibiotic resistance prognosis



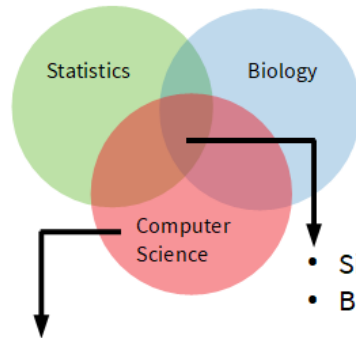
Van Hoan Do, et al., PanKA: Leveraging population pangenome to predict antibiotic resistance, *iScience*, 2024.

Discover resistance mechanisms

Antibiotics	Feature ID	Variant Type	Gene Location	Gene	Gene Function					
AMC	DRKRLISLG	AMRKmer	203-213	oleI	OleI family self-immunity macrolide glycosyltransferase					
AMC	FAMAHIVLT	AMRKmer	369-379	emrB_1	Colistin resistance protein EmrB					
AMC	SARSQRQLLQ	AMRKmer	195-205	blaTEM_1	class A broad-spectrum beta-lactamase TEM-1					
AMC	YAGNPARLLR	AMRKmer	138-148	fdtC	dTDP-3-amino-3%2C6-dideoxy-alpha-D-galactopyranose 3-N-acetyltransferase					
AMC	AQYQPVILEA	AMRKmer	190-200	aadA5	ANT(3'')-Ia family aminoglycoside nucleotidyltransferase AadA5					
AMC	DIVVALLQEK	AMRKmer	380-390	ileS	Isoleucine--tRNA ligase					
AMC	ugpB_2	PAGene	1-671	ugpB_2	sn-glycerol-3-phosphate-binding periplasmic protein UgpB					
AMC	ERNLKEEIKR	AMRKmer	114-124	atoC	Regulatory protein AtoC					
AMC	KNWLGKTTEH	AMRKmer	18-28	ddpF	putative D%2CD-dipeptide transport ATP-binding protein DdpF					
AMC	IsoA	PAGene	1-346	IsoA	mRNA endoribonuclease LsoA					
AMP	blaTEM_1	PAGene	1-290	blaTEM_1	class A broad-spectrum beta-lactamase TEM-1					
AMP	NGKPQDVLTQ	AMRKmer	139-149	hmuV	Hemin import ATP-binding protein HmuV					
AMP	DRKRLISLG	AMRKmer	203-213	oleI	OleI family self-immunity macrolide glycosyltransferase					
AMP	DIVVALLQEK	AMRKmer	380-390	ileS	Isoleucine--tRNA ligase					
AMP	ICYATAMAVL	AMRKmer	373-383	floR	chloramphenicol/lorfenicol efflux MFS transporter FloR					
AMP	pqiC@549019	SNP	51	pqiC	Intermembrane transport lipoprotein PqiC					
AMP	FLSSRFRDCL	AMRKmer	108-118	qnrVC7	quinolone resistance pentapeptide repeat protein QnrVC7					
AMP	groups_1014	PAGene	1-152	groups_1014	Unannotated					
AMP	ubiH@357015	SNP	36	ubiH	2-octaprenyl-6-methoxyphenol hydroxylase					
AMP	NWISITLGDS	AMRKmer	103-113	fdtC	dTDP-3-amino-3%2C6-dideoxy-alpha-D-galactopyranose 3-N-acetyltransferase					
AMX	YAGNPARLLR	AMRKmer	138-148	fdtC	dTDP-3-amino-3%2C6-dideoxy-alpha-D-galactopyranose 3-N-acetyltransferase					
AMX	ELRFVGDKCQ	AMRKmer	358-368	cusS	Sensor histidine kinase CusS					
AMX	RVRELDQIRN	AMRKmer	208-218	hlyB	Alpha-hemolysin translocation ATP-binding protein HlyB					
AMX	AGSVTITLTF	AMRKmer	24-34	oqx28_2	multidrug efflux RND transporter permease subunit Oqx28					
AMX	DRKRLISLG	AMRKmer	203-213	oleI	OleI family self-immunity macrolide glycosyltransferase					
AMX	DIVVALLQEK	AMRKmer	380-390	ileS	Isoleucine--tRNA ligase					
AMX	TSSLTNKEVD	AMRKmer	125-135	alsC	D-allose transport system permease protein AlsC					
AMX	rluF@92150	SNP	64	rluF	Dual-specificity RNA pseudouridine synthase RluF					
AMX	groups_4@1611	SNP	2458	groups_4	Unannotated					
AMX	NGKPQDVLTQ	AMRKmer	139-149	hmuV	Hemin import ATP-binding protein HmuV					

Outlooks

Computational biology



- Machine Learning
- Deep Learning
- AI

- Single cell technologies
- Big Data

Science

Current Issue First release papers Archive About ▾ Submit ma

HOME > SCIENCE > VOL. 375, NO. 6584 > A SINGLE-CELL ATLAS OF THE NORMAL AND MALFORMED HUMAN BRAIN VASCULATURE

RESEARCH ARTICLE | NEUROGENOMICS

f X t in e s o e

A single-cell atlas of the normal and malformed human brain vasculature

nature

Article Published: 18 November 2020

A molecular cell atlas of the human lung from single-cell RNA sequencing

Building “Google Maps” for the human body

Published: 30 December 2013

Method of the Year 2013

Nature Methods 11, 1 (2014) | [Cite this article](#)

28k Accesses | 34 Citations | 126 Altmetric | [Metrics](#)

Methods to sequence the DNA and RNA of single cells are poised to transform many areas of biology and medicine.

Editorial | [Published: 06 January 2020](#)

Method of the Year 2019: Single-cell multimodal omics

Nature Methods 17, 1 (2020) | [Cite this article](#)

Editorial | [Published: 06 January 2021](#)

Method of the Year 2020: spatially resolved transcriptomics

Nature Methods 18, 1 (2021) | [Cite this article](#)

42k Accesses | 64 Citations | 241 Altmetric | [Metrics](#)

[nature](#) > [nature biotechnology](#) > [news](#) > article

News | [Published: 13 September 2024](#)

Foundation models build on ChatGPT tech to learn the fundamental language of biology

Building “ChatGPT”: language of molecular biology

Thank you for your attention!

References:

- [1] Van Hoan Do, et al., Pasa: leveraging population pangenome graph to scaffold prokaryote genome assemblies, *Nucleic Acids Research*, 2024.
- [2] Van Hoan Do, et al., PanKA: Leveraging population pangenome to predict antibiotic resistance, *iScience*, 2024.

Tools (Python):

<https://github.com/amromics/pasa>

<https://github.com/amromics/panka>