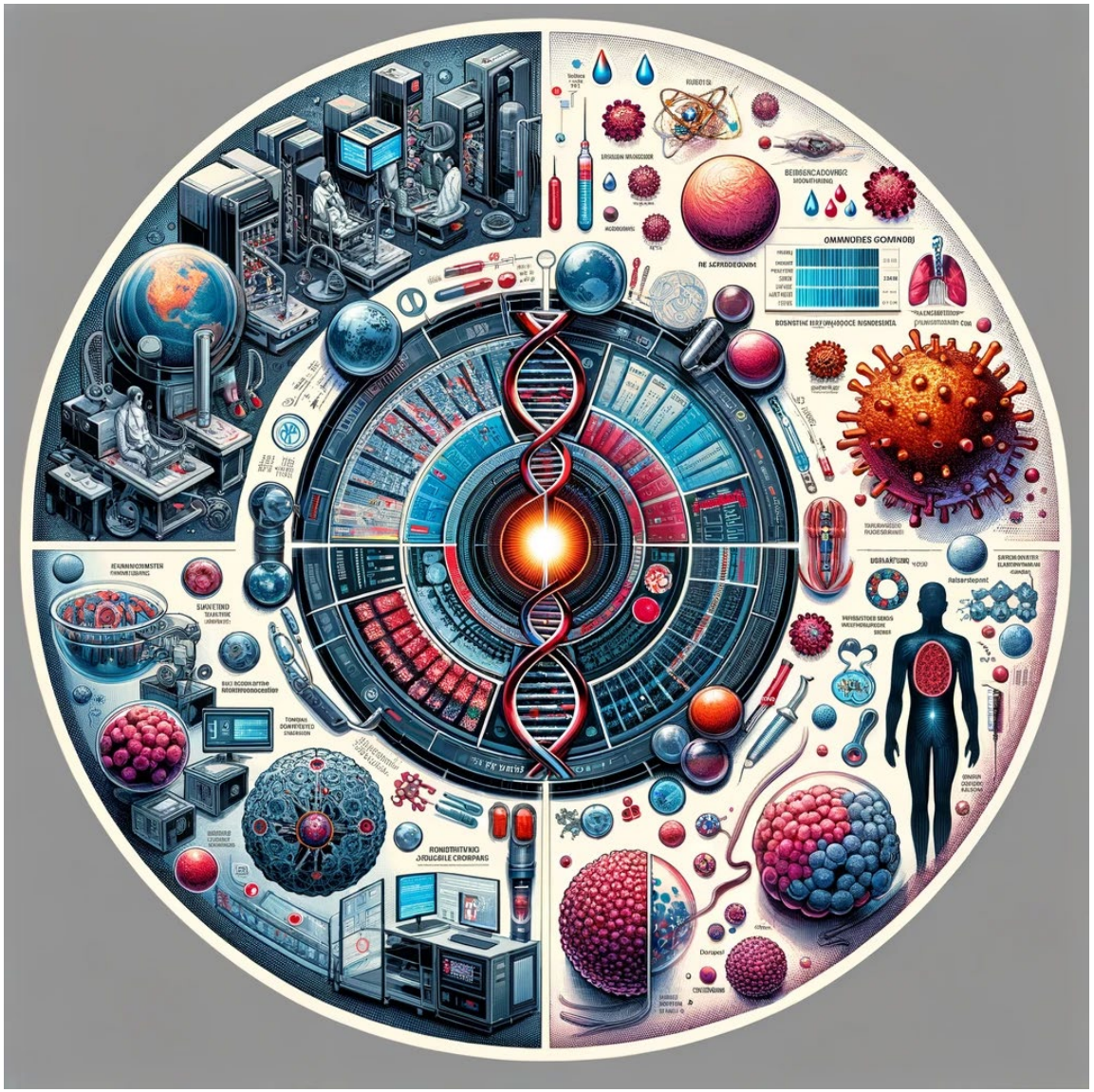
Gene Sequencing



GPT4

What is Sanger Sequencing

DNA Template Preparation: The DNA sequence to be determined is used as a template. This DNA is denatured to create single-stranded templates.
Primer Annealing: A short, single-stranded piece of DNA, known as a primer, which is complementary to a region of the template, is annealed (bound) to the template to initiate DNA synthesis.
Chain Elongation: A special mixture is prepared that includes the normal deoxynucleotide triphosphates (dNTPs) required for DNA synthesis and a small proportion of chain-terminating dideoxynucleotide triphosphates (ddNTPs). The DNA polymerase enzyme extends the primer by adding nucleotides complementary to the template until a ddNTP is incorporated, causing the extension to terminate.
Fragment Separation: The result is a mixture of new DNA strands of varying lengths, each terminating at a different point. These fragments are then separated by size through capillary electrophoresis.
Result Interpretation: The terminal base of each fragment can be identified based on the label of the ddNTP (each of the four ddNTPs is labeled with a different fluorescent dye). By ordering the fragments by size, the DNA sequence can be read from the shortest to the longest fragment, providing the sequence of the template DNA.
While newer sequencing technologies (next-generation sequencing methods) offer faster and cheaper ways to sequence DNA, Sanger sequencing is still used today for its high accuracy, especially for sequencing small DNA fragments and validating DNA sequences obtained by other methods.

GPT4

What is NGS

Here's a basic overview of how NGS works:

Library Preparation:

The first step in NGS is preparing a library of the sample DNA (or RNA, in the case of RNA sequencing). This involves fragmenting the sample DNA into smaller pieces, then attaching adapters to both ends of the fragments. These adapters serve multiple functions: they allow the fragments to bind to the sequencing platform, include sequencing primer binding sites, and may include unique barcodes that enable the multiplexing of different samples in a single sequencing run.

Amplification:

Once the library is prepared, the DNA fragments are amplified, typically using a method like PCR or bridge amplification on a solid surface. This step increases the quantity of DNA, ensuring there is enough signal for detection during sequencing.

Sequencing:

NGS platforms use different methods to sequence the amplified DNA fragments. Common approaches include: Sequencing by Synthesis (Illumina): This is the most widely used method, where DNA polymerase synthesizes a complementary strand to the template, incorporating fluorescently labeled nucleotides that are detected by the sequencer.

Data Analysis:

The raw data from an NGS platform consist of millions to billions of short DNA sequences (reads). These reads are then processed and analyzed using bioinformatics tools. Depending on the application, this may involve aligning the reads to a reference genome, assembling them into longer sequences, identifying variants, or quantifying gene expression levels.

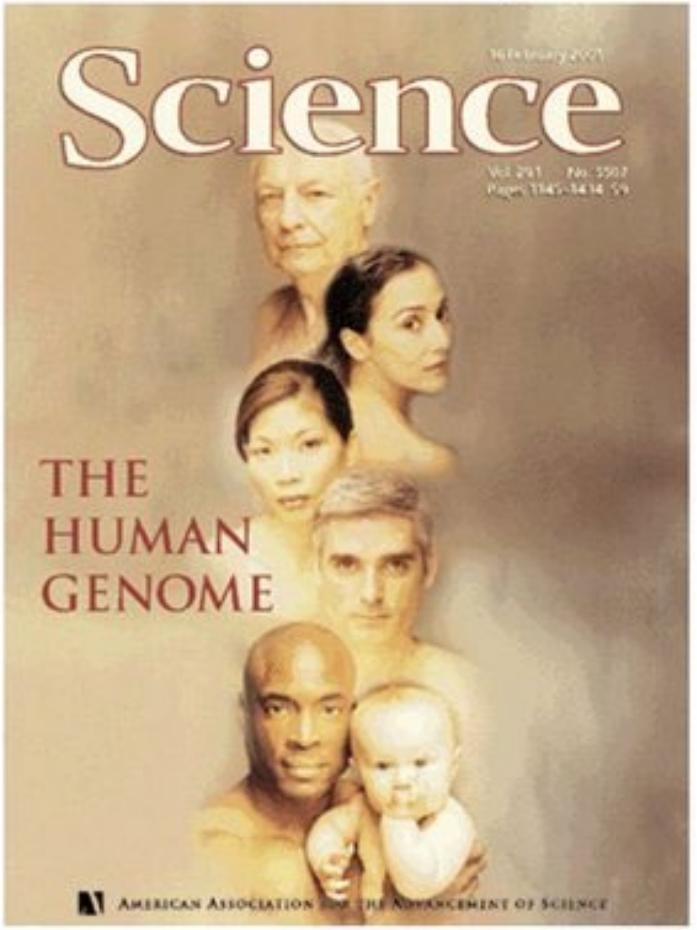
Applications:

NGS can be used for a wide variety of applications, including whole-genome sequencing, targeted sequencing of specific genes or regions, transcriptome analysis (RNA-seq), metagenomics, and epigenomics.

Overall, NGS technologies provide a powerful set of tools for exploring the genetic and molecular basis of life, with applications ranging from basic biological research to clinical diagnostics and personalized medicine.

Genome sequencing

February 2001 - Publication of the first draft of the human genome



The construction of the whole genome sequence for a human during the Human Genome Project (HGP) was a monumental task that involved many steps and utilized a combination of various sequencing techniques, including Sanger sequencing. Here's a simplified overview of the process used to construct the whole human genome sequence:

Sample Collection and Preparation: DNA was extracted from the cells of a small number of donors. The HGP initially used a reference approach, where DNA from a composite of individuals was used to minimize individual genetic variation.

Library Construction: The extracted DNA was randomly fragmented into smaller pieces. These fragments were then cloned into vectors to create a library of DNA fragments. Each vector with an insert of human DNA was maintained in a bacterial host, allowing for amplification and easy retrieval.

Physical Mapping: Before sequencing, it was crucial to organize the DNA fragments. Physical maps were created to understand the arrangement of genes and other sequences on chromosomes. This included creating maps that showed which DNA fragments overlapped, helping to order the fragments.

Sequencing: The DNA fragments in the library were sequenced using the Sanger method. This process involved sequencing many short fragments of DNA (typically 500-800 base pairs in length) and then piecing these sequences together.

Assembly: The sequenced fragments were assembled into longer sequences known as contigs. Computer algorithms compared all the sequenced fragments to find overlapping regions and used these overlaps to join fragments together into continuous sequences.

Gap Closure and Validation: Despite the initial assembly, there were gaps in the sequence where no data were available. Additional targeted sequencing efforts were made to fill these gaps. Moreover, the sequence was checked for errors and validated by resequencing or comparing it against known genetic markers.

Annotation: Once the genome sequence was assembled, scientists worked on identifying and mapping all the genes within the sequence. This involved predicting where genes were located and what functions they might perform, as well as identifying other important sequences such as regulatory regions.

The Human Genome Project was completed in 2003, providing a comprehensive and highly accurate reference sequence of the human genome. This reference has been fundamental to biomedical research, allowing for advances in understanding genetic diseases, human biology, and evolution. Despite the completion of the HGP, the genome sequence continues to be refined and updated as new technologies emerge and our understanding of the genome evolves.

The current version of the human genome reference assembly, GRCh38.p14 (GRCh38), has millions of bases represented by the letter "N," which means that the actual base residing at that location is unknown.

There are also 169 sequences that cannot confidently be ordered or oriented within the assembly, typically owing to their repetitive nature

Until recently, limitations of sequencing technology, primarily that the sequencers could read no more than about 1000 bases at a time,

The HGP opted for a more structured approach. This involved cloning genomic DNA into pieces that could be grown in bacteria (clones) and indexed in 96-well plates. Clones from these libraries were first mapped to chromosome

Identification of Open Reading Frames (ORFs):
Open Reading Frames (ORFs) are sequences of DNA that could potentially encode a protein. They start with a start codon (ATG in DNA or AUG in mRNA) and end with a stop codon (TAA, TAG, or TGA).
Computational tools scan genomic sequences to find ORFs that are long enough to be plausible human genes, typically longer than 100 amino acids.

Gene Prediction Algorithms:
Various computational programs are used to predict the presence of genes. These algorithms analyze the genomic sequence for characteristic features of genes, such as promoter regions, splice sites, exons, introns, and polyadenylation signals.

Comparative Genomics:
Sequences that are conserved across different species are more likely to be functionally important. By comparing the human genome with the genomes of other species, scientists can identify conserved sequences that are likely to be genes. Orthologous genes (genes in different species that evolved from a common ancestral gene) can provide clues about gene location and function.

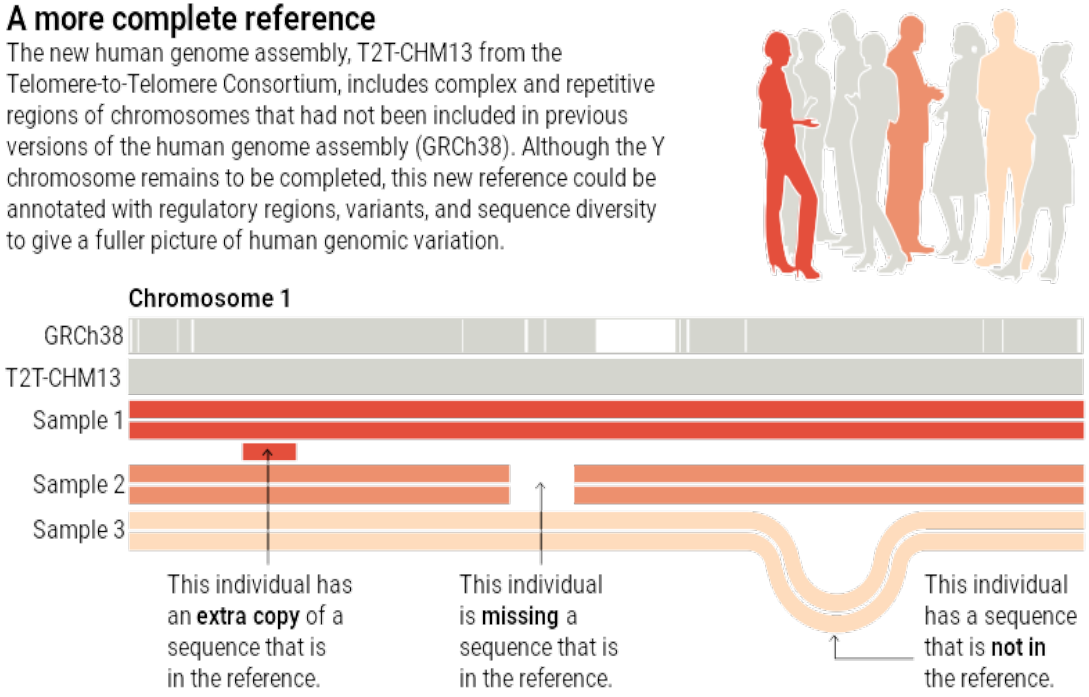
Functional Genomics and Transcriptomics:
RNA sequencing (RNA-seq) provides direct evidence of transcribed regions, helping to identify which parts of the genome are actually being expressed as RNA. Experimental techniques like cDNA cloning and analysis can also validate the presence of a gene by identifying its transcript.

Experimental Validation:
Once a segment is predicted to be a gene, experimental techniques can validate its function. For example, gene knockout or gene editing can be used to assess the impact of removing or altering the gene..

Database and Literature Integration:
Databases like GENCODE, RefSeq, and Ensembl compile information on gene annotations and are continuously updated with new experimental evidence and computational predictions..

Expert Curation:
Despite advances in computational predictions, expert curation is often necessary to resolve ambiguities and integrate diverse types of evidence.
Human curation involves reviewing computational predictions, experimental data, and literature evidence to make the final determination about the presence and function of a gene.
By integrating these approaches, scientists can confidently identify which segments of the human genome are likely to code for proteins, providing a crucial foundation for understanding human biology and disease.

An important attribute of the human reference assembly is that the source DNA was derived from multiple individuals.

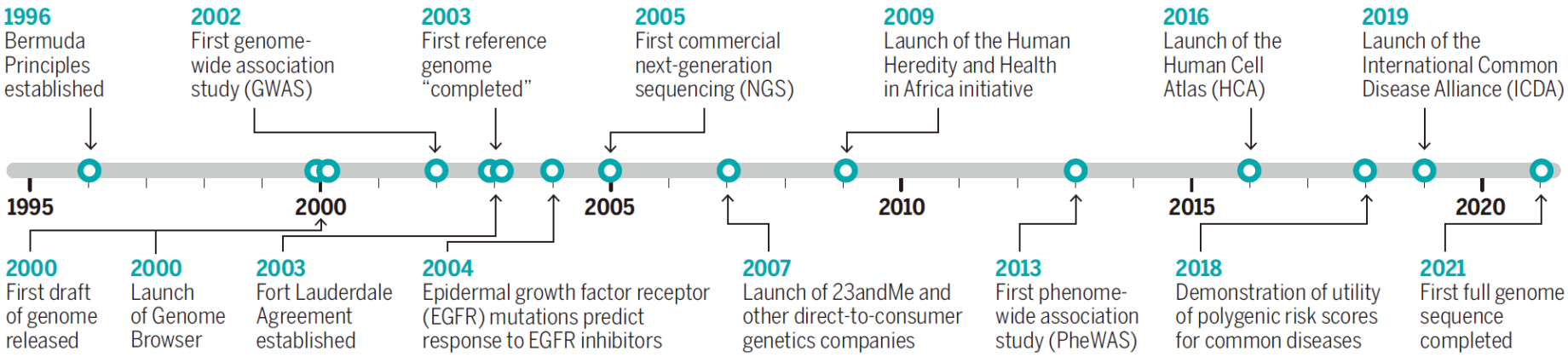


when two clones from different haplotypes of an individual are adjacent in the reference assembly, this can create sequence representations that are not normally found in the population

T2T Consortium, eliminated the problem of allelic diversity by sequencing the genome of a cell line derived from a complete hydatidiform mole (CHM).

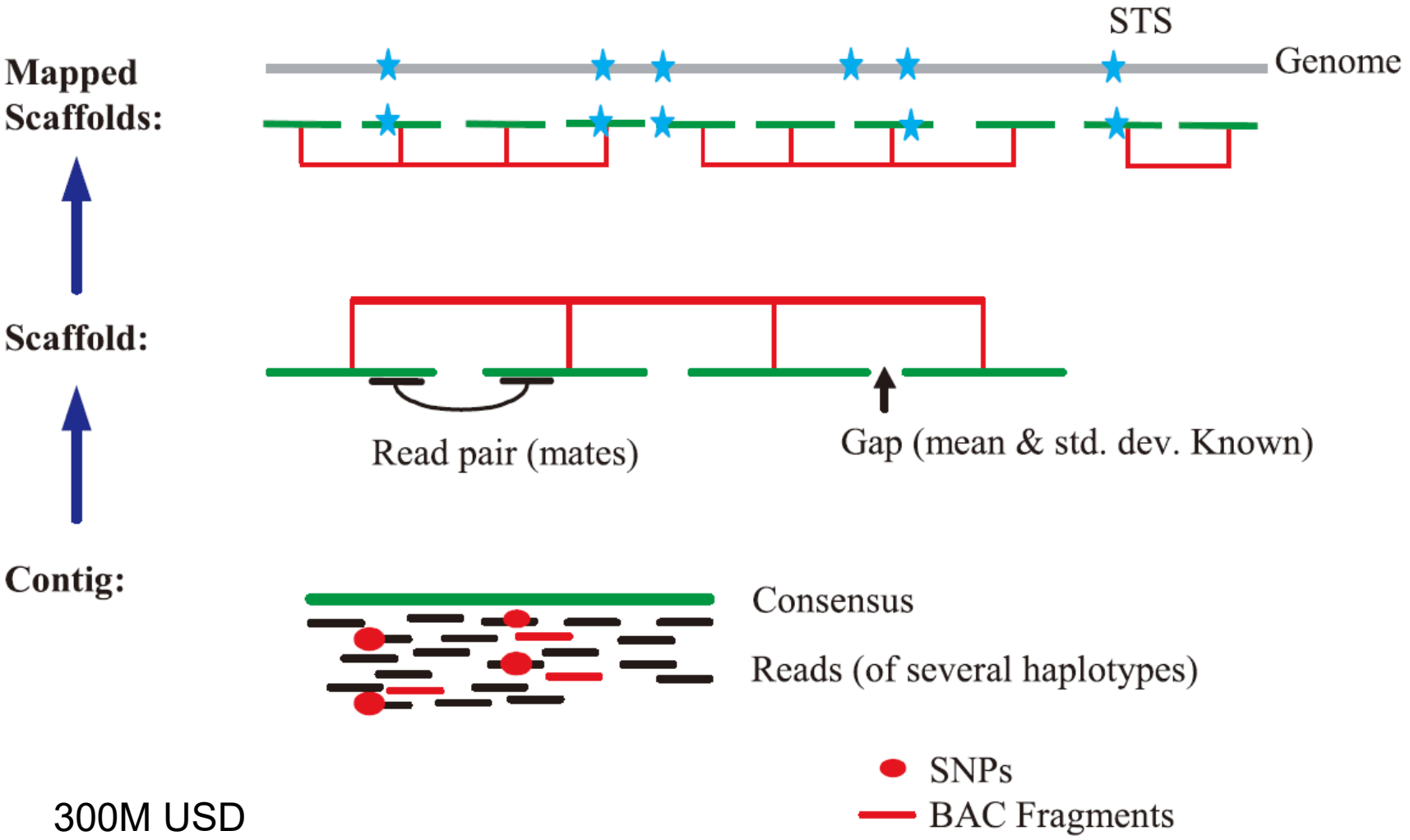
This is duplicated so that the cell contains two copies of the same parental genome

A next-generation human genome sequence, Volume: 376, Issue: 6588, Pages: 34-35, DOI: (10.1126/science.abo5367)



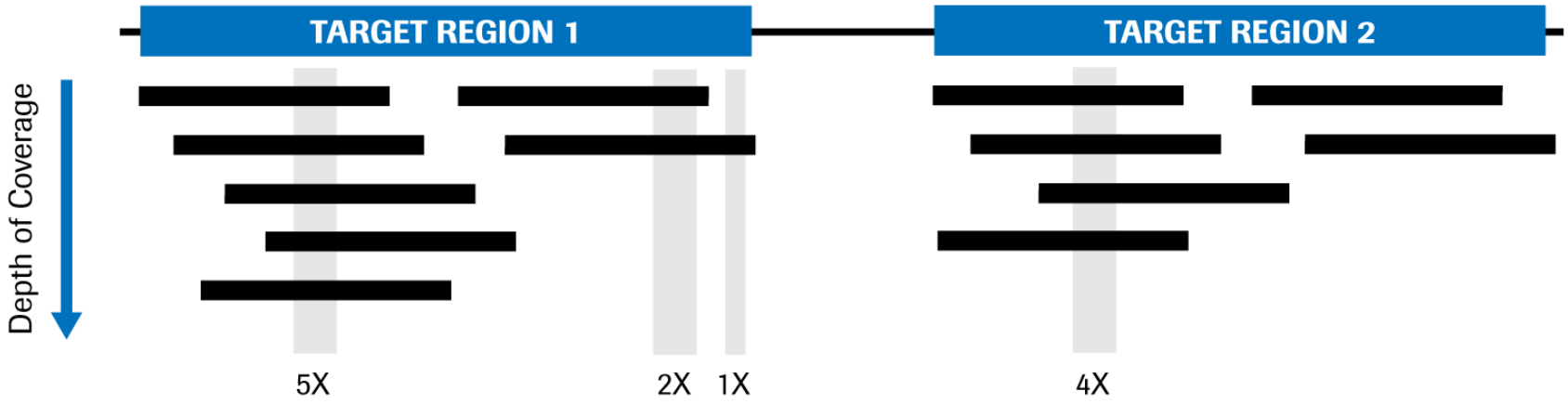
Bermuda Principles

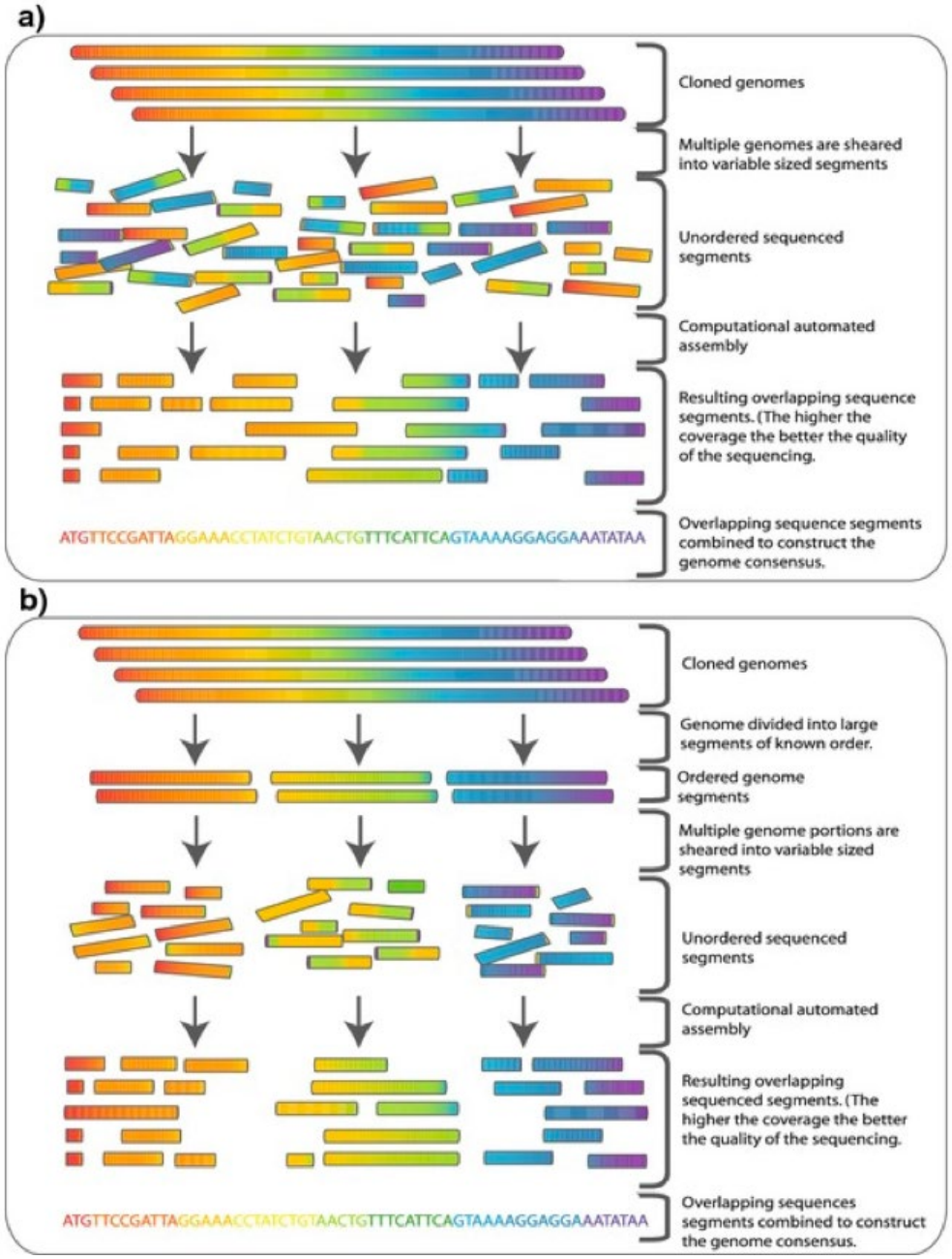
- Automatic release of sequence assemblies larger than 1 kb (preferably within 24 hours).
- Immediate publication of finished annotated sequences.
- Aim to make the entire sequence freely available in the public domain for both research and development in order to maximise benefits to society.



300M USD
2.9 bbp
9 months
5 donors
5.1 folds
Whole-genome shotgun

coverage depth

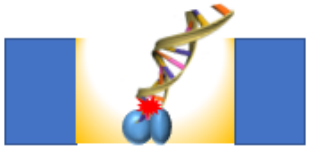



a)
Cloned genomes
Multiple genomes are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequence segments. (The higher the coverage the better the quality of the sequencing.
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA
Overlapping sequence segments combined to construct the genome consensus.
b)
Cloned genomes
Genome divided into large segments of known order.
Ordered genome segments
Multiple genome portions are sheared into variable sized segments
Unordered sequenced segments
Computational automated assembly
Resulting overlapping sequenced segments. (The higher the coverage the better the quality of the sequencing.
ATGTTCCGATTAGGAAACCTATCTGTAACTGTTTCATTCAGTAAAAGGAGGAAATATAA
Overlapping sequences segments combined to construct the genome consensus.

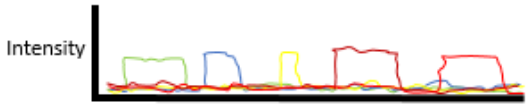
Third Generation Sequencing

PacBio SMRT seq

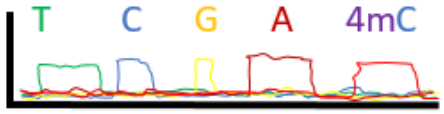
DNA passes thru polymerase in an illuminated volume



Raw output is fluorescent signal of the nucleotide incorporation, specific to each nucleotide

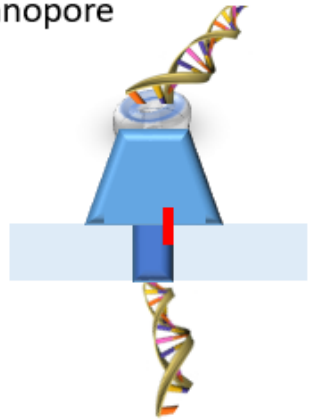


A,C,T,G have known pulse durations, which are used to infer methylated nucleotides

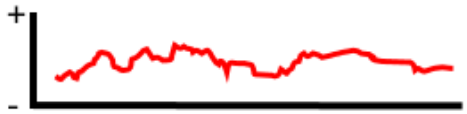


Oxford Nanopore

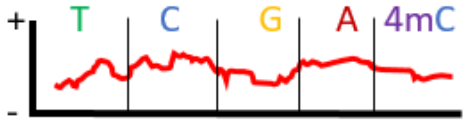
DNA passes thru nanopore



Raw output is electrical signal caused by nucleotide blocking ion flow in nanopore

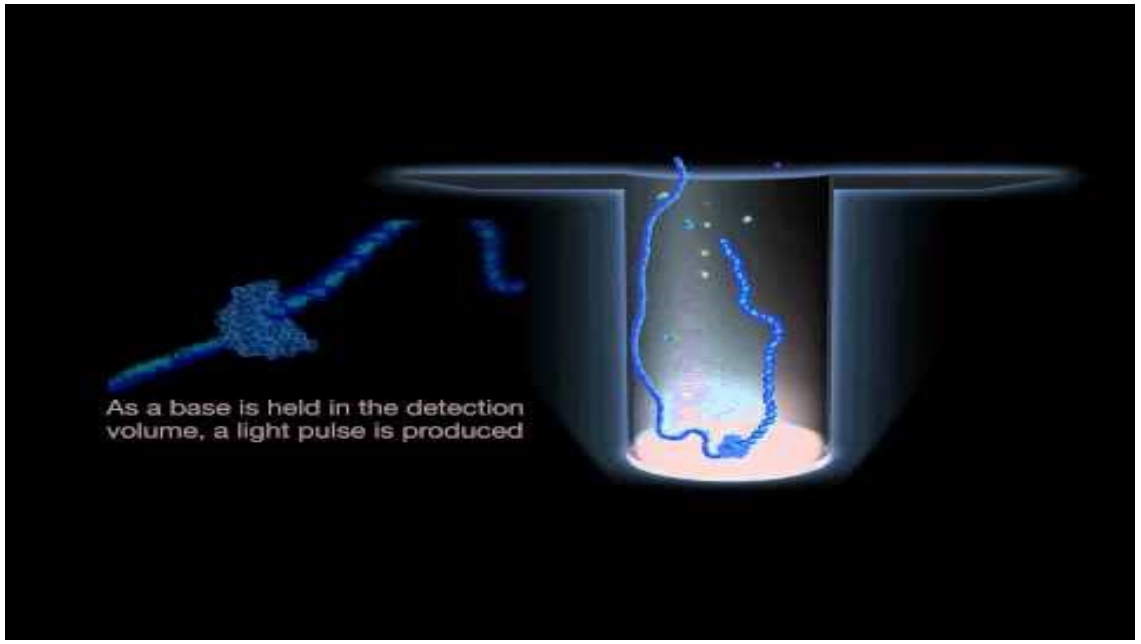


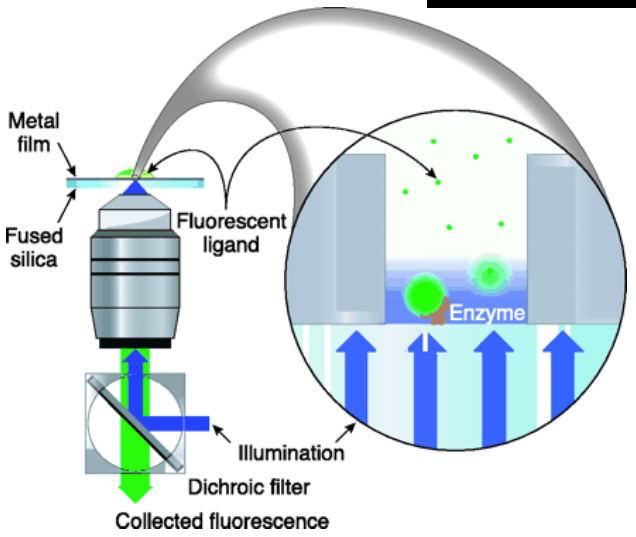
Each nucleotide has a specific electric “signature”



PacBio SMRT Sequencing:
Single Molecule, Real-Time (SMRT) sequencing by Pacific Biosciences uses zero-mode waveguides (ZMWs), tiny wells that allow the observation of a single DNA polymerase molecule as it incorporates nucleotides into a DNA strand.
During sequencing, each of the four DNA bases is attached to a different fluorescent dye. As a DNA polymerase adds a nucleotide to the growing DNA strand inside a ZMW, the incorporated base emits a fluorescent signal, which is detected in real time.
SMRT sequencing enables the reading of very long DNA fragments (**up to 20 kb or more**), which helps in resolving complex genomic regions, identifying large structural variations, and improving genome assembly.
Oxford Nanopore Sequencing:
This technology uses protein nanopores set in an electrically resistant polymer membrane. When a voltage is applied across the membrane, an ionic current flows through the nanopores.
As a DNA or RNA molecule passes through a nanopore, it causes characteristic disruptions in the current. Each type of base (A, C, G, T for DNA; A, C, G, U for RNA) disrupts the current in a unique way, allowing the sequence to be determined.
Nanopore sequencing can process very long strands of DNA or RNA directly, which provides advantages in mapping long repetitive regions and characterizing full-length transcripts in transcriptome studies.
Both technologies have greatly expanded the capabilities of genomic analyses by offering:
Longer Reads: They provide much longer reads than second-generation sequencing, which is crucial for de novo genome assembly, spanning repetitive sequences, and fully characterizing genomic rearrangements.
Single-Molecule Sequencing: Both technologies sequence individual DNA molecules, which helps in detecting base modifications such as methylation directly during the sequencing process, without the need for additional chemical treatments.
Real-Time Data Access: These technologies allow for the monitoring of the sequencing process in real time, enabling rapid access to data and the potential for dynamic experimental adjustments.
Despite these advantages, **third-generation sequencing methods also have limitations, such as higher error rates compared to second-generation sequencin**g. However, these errors are random rather than systematic, which means they can often be overcome with sufficient coverage or by combining with short-read sequencing data for validation and error correction. As the technology continues to evolve, improvements in accuracy, cost, and throughput are expected to further expand the applications of third-generation sequencing.

Zero Mode Waveguide

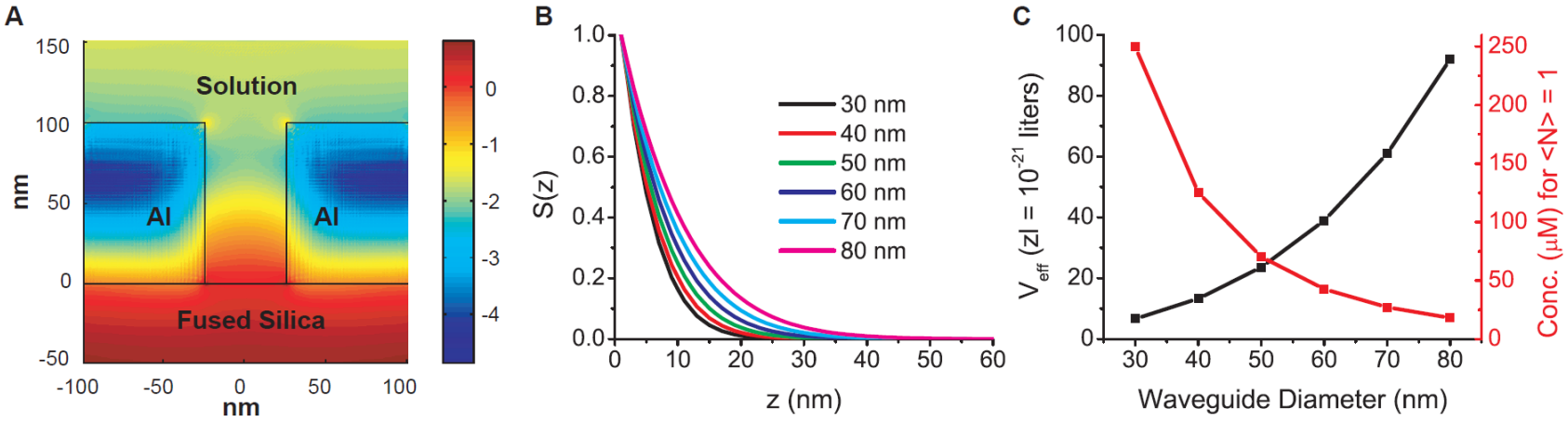




https://www.youtube.com/watch?v=NHCJ8PtYCFc

SCIENCE VOL 299 31 JANUARY 2003

Zero Mode Waveguide



50 x 50 x 10 nm^3 = 2.5 x 10^4 x 10^{-21} cc = 2.5 x 10^{-20} L= 25 zeptoliter

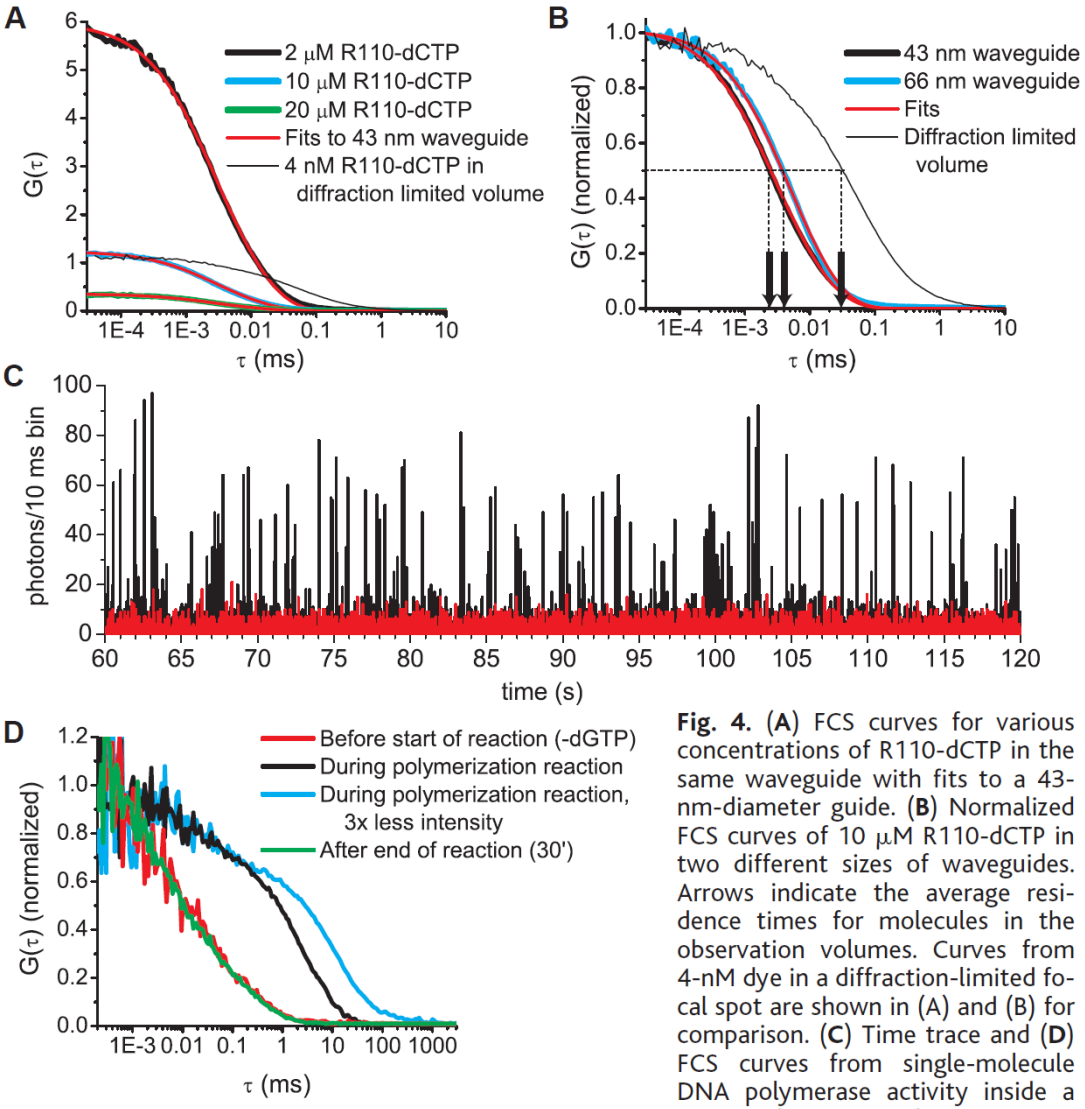
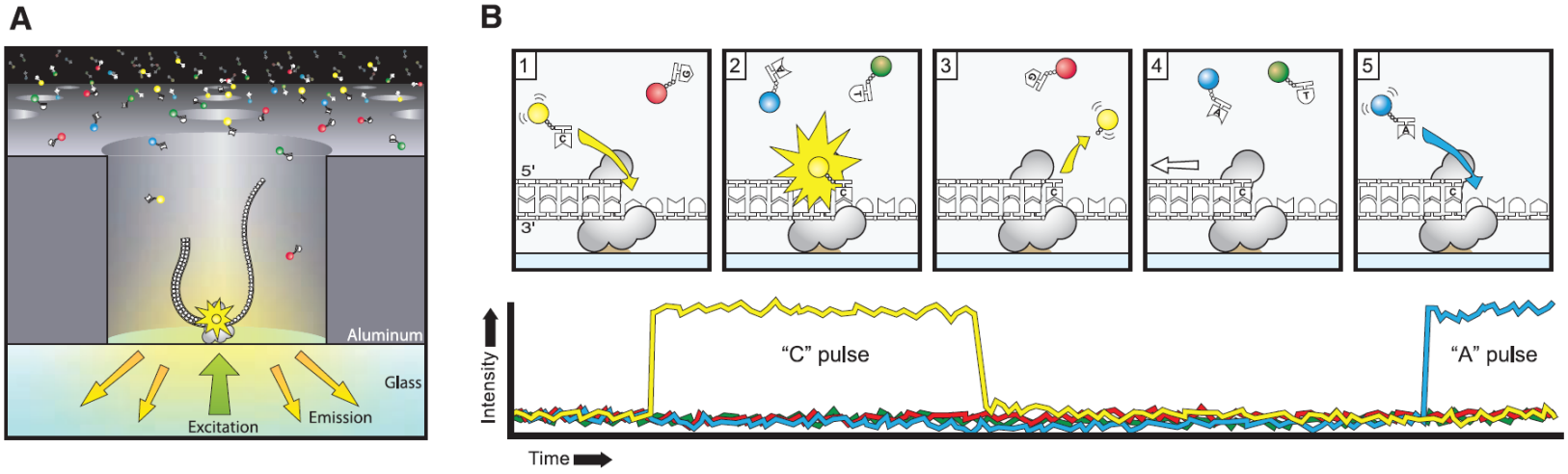


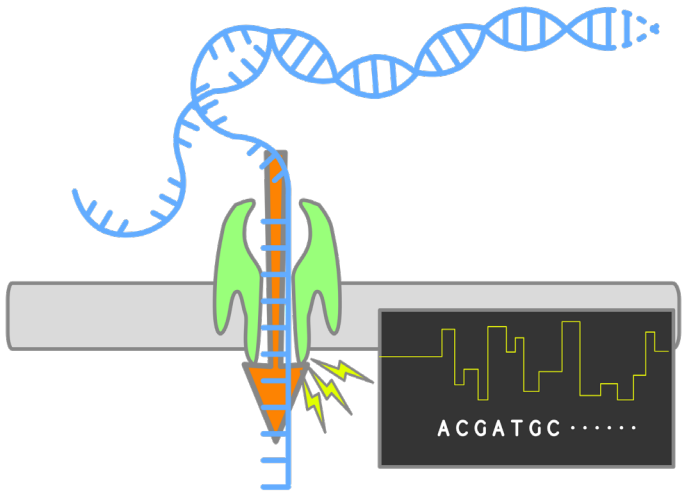
Fig. 4. (**A**) FCS curves for various concentrations of R110-dCTP in the same waveguide with fits to a 43-nm-diameter guide. (**B**) Normalized FCS curves of 10 μM R110-dCTP in two different sizes of waveguides. Arrows indicate the average residence times for molecules in the observation volumes. Curves from 4-nM dye in a diffraction-limited focal spot are shown in (A) and (B) for comparison. (**C**) Time trace and (**D**) FCS curves from single-molecule DNA polymerase activity inside a zero-mode waveguide. Incorporation events and subsequent photobleaching of coumarin-dCTP appear as distinct fluorescence bursts in the black time trace (10-ms time bins). This results in a long-time shoulder in the corresponding FCS curves during polymerization (black and blue curves) in (D). Decreasing the intensity results in slower photobleaching as seen by the longer residence time in the blue curve. The red curves in (C) and (D) are the corresponding negative controls (absence of one native nucleotide) in the same waveguide before initiation of the reaction. The green curve in (D) is the control after the reaction has stopped.

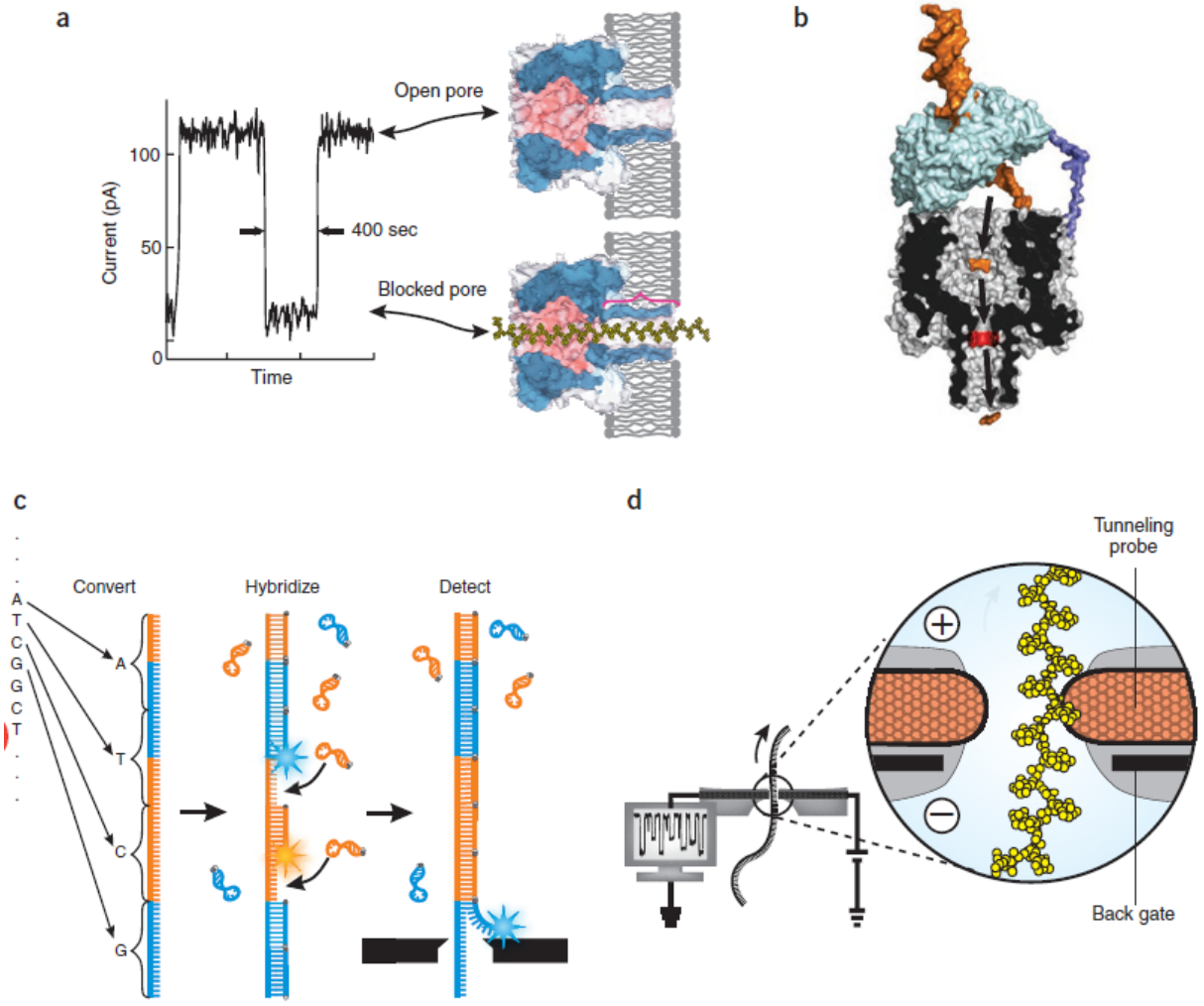
Real-Time DNA Sequencing from Single Polymerase Molecules



SCIENCE VOL 323 2 JANUARY 2009

Nanopore Sequencing




a
Open pore
100
50
0
Current (pA)
400 sec
Blocked pore
Time
b
c
Convert
Hybridize
Detect
A
T
C
G
G
C
T
A
T
C
G
d
Tunneling probe
Back gate

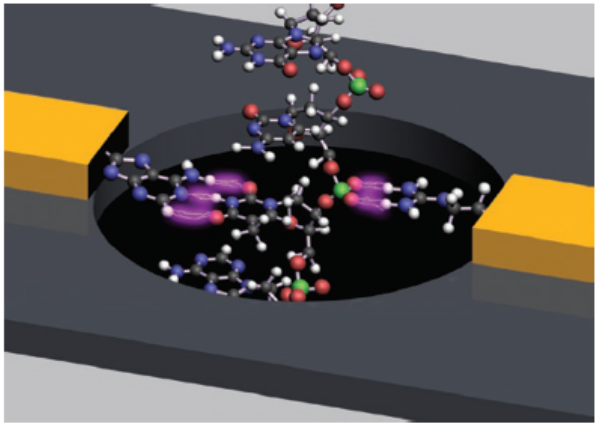
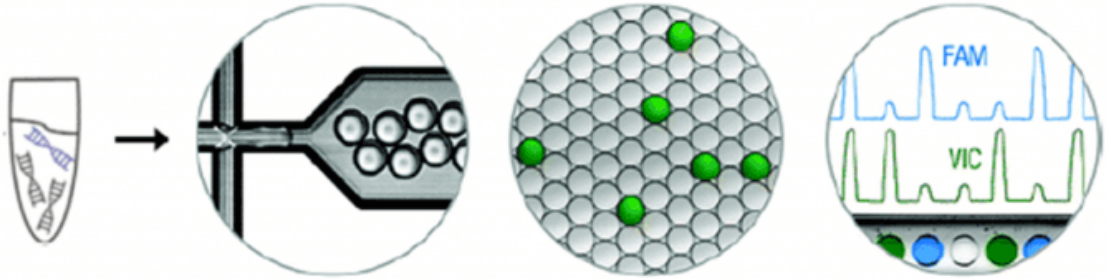


Figure 2 A nanopore reader with chemically functionalized probes. As a strand of DNA emerges from a nanopore, a 'phosphate grabber' on one functionalized electrode and a 'base reader' on the other electrode form hydrogen bonds (light blue ovals) to complete a transverse electrical circuit through each nucleotide as it is translocated through the nanopore.

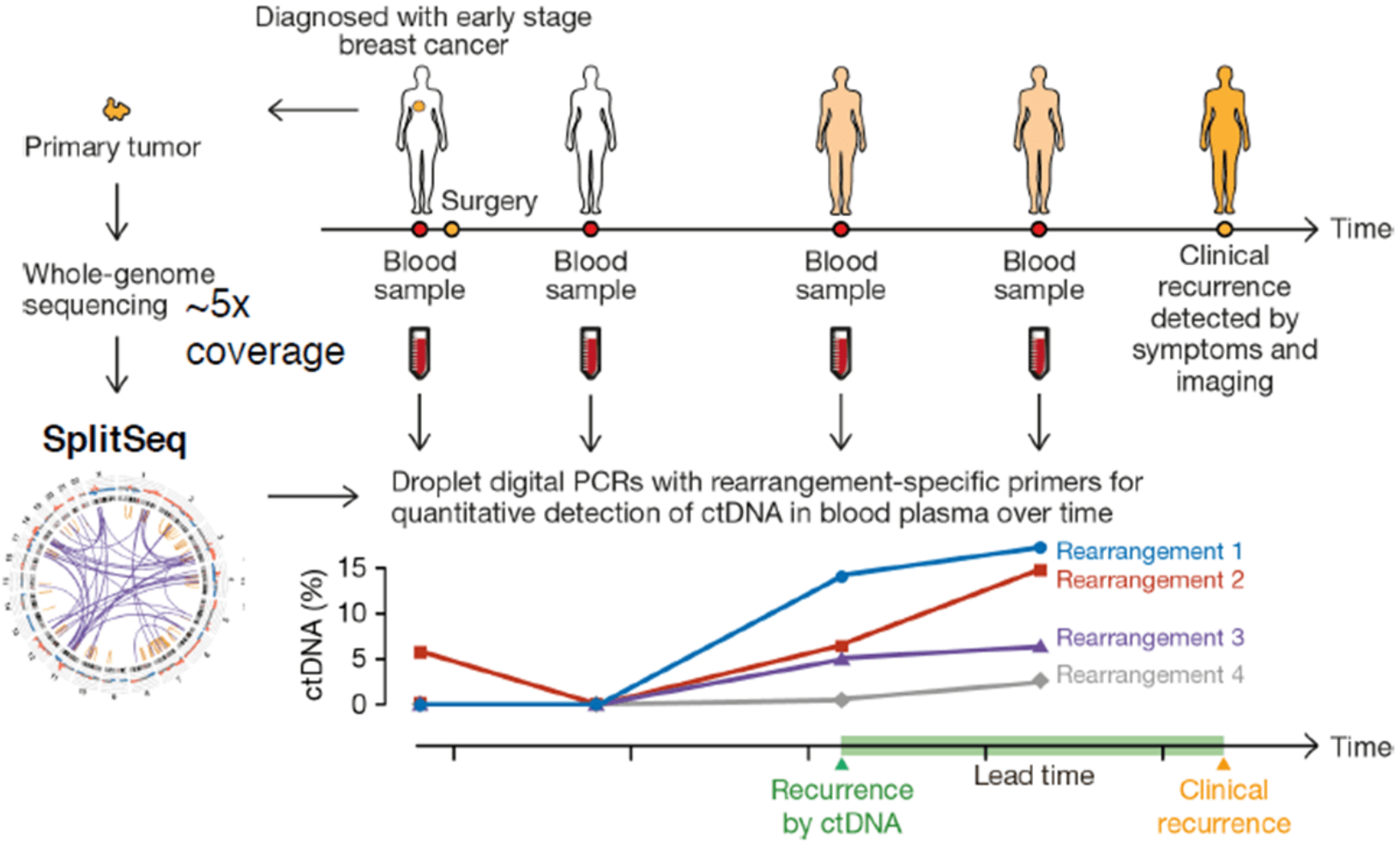
Digital PCR

Droplet digital PCR



Bio-Rad QX100

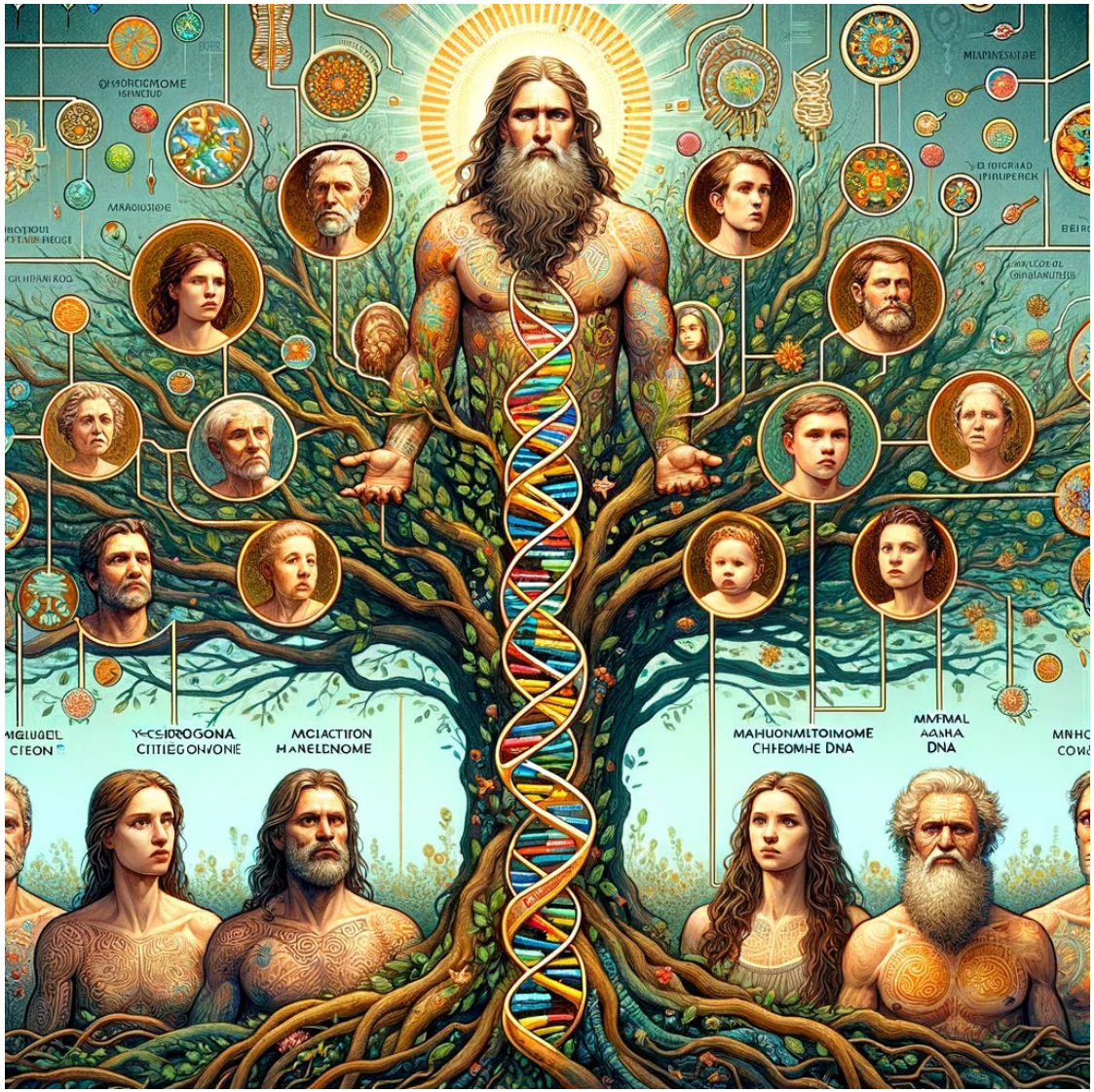
1. MAKE	2. CYCLE	3. READ	
Sample is partitioned into 20,000 droplets	Run PCR cycles in all droplets simultaneously	Measure fluorescence intensity in each droplet	Calculate concentration from number of positive droplets


Diagnosed with early stage breast cancer
Primary tumor
Surgery
Time
Whole-genome sequencing
~5x coverage
Blood sample
Blood sample
Blood sample
Blood sample
Clinical recurrence detected by symptoms and imaging
SplitSeq
Droplet digital PCRs with rearrangement-specific primers for quantitative detection of ctDNA in blood plasma over time
ctDNA (%)
15
10
5
0
Rearrangement 1
Rearrangement 2
Rearrangement 3
Rearrangement 4
Time
Lead time
Recurrence by ctDNA
Clinical recurrence

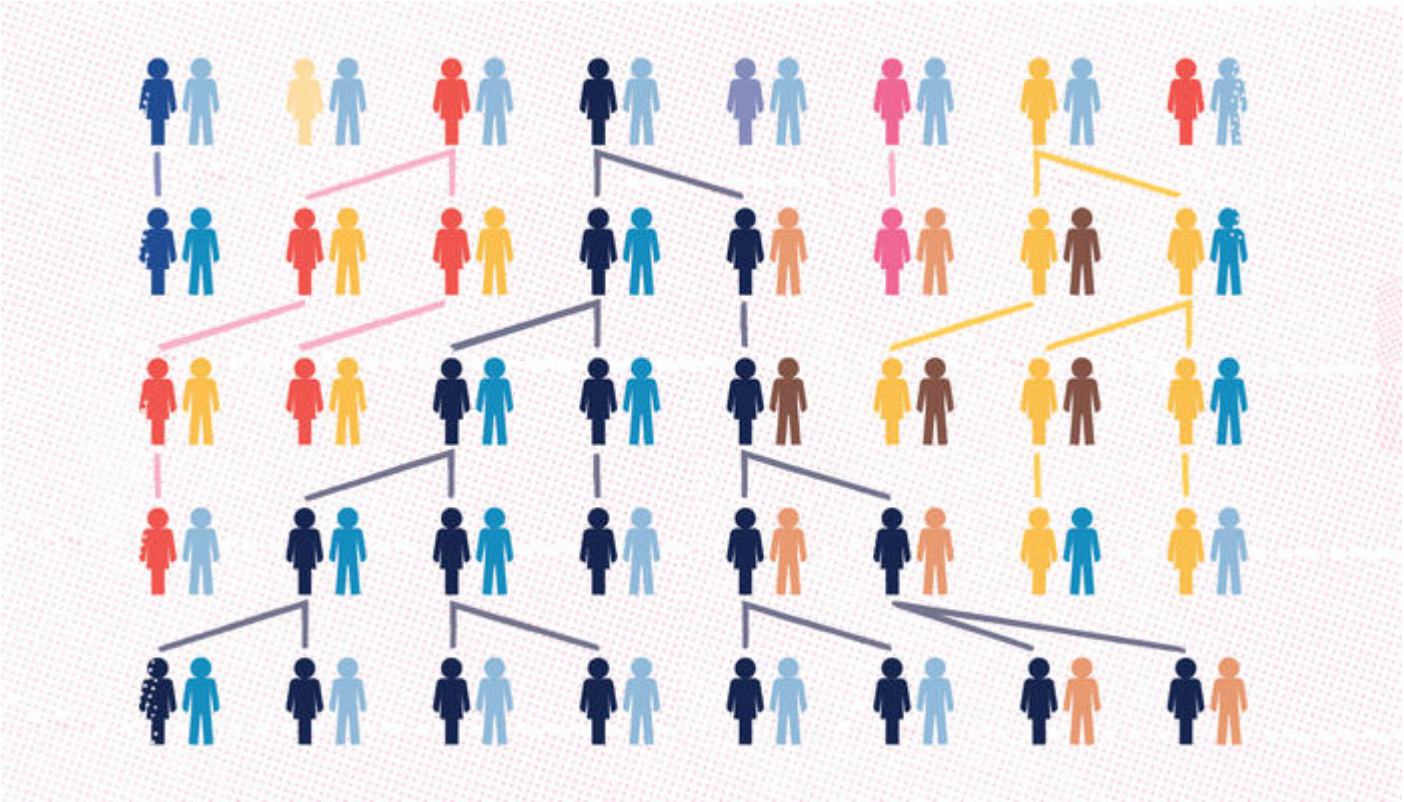
Currently, the cost of sequencing a whole human genome has significantly decreased, making it more accessible for both research and clinical applications. As of the latest data, a human genome can now be sequenced for as **low as $600.** This dramatic reduction in cost, from thousands of dollars to a few hundred, is a result of advancements in sequencing technologies, particularly through innovations by companies like Illumina, which have improved the efficiency and reduced the costs of sequencing.

While specific details on the exact speed of sequencing were not provided in the sources I accessed, the advancements in next-generation sequencing technologies have made it possible to sequence whole genomes **in a matter of days,** depending on the sequencing platform and the depth of coverage required. This rapid turnaround is crucial for applications in clinical diagnostics and personalized medicine, where timely results can significantly impact patient care and treatment decisions.

DNA Sequencing for Lineage Tracing



GPT4



How are we all related to one woman/female ancestor?

The piece of Eve's DNA that has been traced comes from something called the mitochondria – it's in all human cells. There is DNA inside the mitochondria, and it's this DNA that has been traced back all the way to Eve - some 150-200,000 years ago, and is of African origin (sub-Sahara), probably from Ethiopia or Kenya.

How come our direct maternal lineage does not go back further than Eve?

The species Homo sapiens is older than 200,000 years but - this is the hard part – 'Eve' is dated back to Africa about 150 to 200 thousand years ago.

We know this because analysis of variation in living humans, plus the mutation rate in mitochondrial DNA, permits us to calculate when the ancestor of all living mitochondrial DNA chromosomes was alive.

Prior to 'Eve', there were actually still many females living, but all those lines (other than Eve's own direct ancestor) have died out. Some will have died after a hundred, some thousands, some tens of thousands of years.

Remember, for the lineage to survive to today, a female must have a fertile daughter, who in turn has a daughter over 8,000 generations.

The average mutation rate was estimated to be approximately 2.5 x 10(-8) mutations per nucleotide site or 175 mutations per diploid genome per generation. Rates of mutation for both transitions and transversions at CpG dinucleotides are one order of magnitude higher than mutation rates at other sites. Single nucleotide substitutions are 10 times more frequent than length mutations. Comparison of rates of evolution for X-linked and autosomal pseudogenes suggests that the male mutation rate is 4 times the female mutation rate, but provides no evidence for a reduction in mutation rate that is specific to the X chromosome.

What are “mutation markers” and how do they trace back to Eve through our DNA?

A mutation is any change in the DNA sequence. These alterations or mutations occur by chance and are, therefore, random.

If a mutation arises on the Y chromosome, or on our mitochondrial DNA, and it does not kill us, it will be passed to the offspring.

So these mutations can be traced back through the generations by analysing the Y chromosome ancestry of an individual’s direct father’s line, and therefore can only be performed on males.

Mitochondrial DNA is the equivalent for tracking the direct maternal ancestry as it is passed through females. In this way, we can trace the direct maternal and paternal ancestries of males.

When a new mutation occurs and ‘branches off’ from the line, it’s called a branch.

A mutation, which became known as the M branch, arose around 70,000 years ago and there have been very many further mutations of people belonging to the M branch over time, so that today there are many, many sub-branches within the ancient M branch.

All these sub branches share the common M ancestor: a woman in who lived approximately 70 thousand years ago.

How we can trace Genghis Khan's DNA through the Y Chromosome?

Analysis of Y chromosomes in Asian men showed the spread of a Y lineage that can be explained by a historical fact – the spread of the Mongol empire under Genghis Khan and his descendants.

This particular Y lineage was found from the Caspian Sea to the Pacific and accounted for approximately 8% of the Y chromosomes in the region.

The time to the most recent common ancestor of these Y chromosomes was estimated to be approximately 1000 years and Mongolia was the clear source of the Y chromosome. This data is consistent with the hypothesis that they represent the Y chromosome of Genghis Khan, his immediate male relatives, and all their descendants.

The Y chromosome is passed from father to son virtually unchanged, except for occasional mutations. By comparing the Y chromosome sequences of different males, scientists can build a paternal lineage tree.
The most recent common ancestor (MRCA) in this context is often referred to as "Y-chromosomal Adam." This individual is not the first or only male human alive at his time but is the most recent male from whom all living humans' Y chromosomes are descended.
By analyzing the mutation rates and the differences in the Y chromosomes among diverse populations, scientists can estimate the time back to this common ancestor.
Mitochondrial DNA Sequencing for Maternal Lineage:
Mitochondrial DNA is passed from mothers to their children without recombination, making it a stable marker for tracing maternal lineage.
Similar to the Y-chromosome analysis, scientists can compare mtDNA sequences across different individuals to construct a maternal lineage tree.
The MRCA traced through mtDNA is often called "Mitochondrial Eve." Like Y-chromosomal Adam, Mitochondrial Eve is the most recent woman from whom all living humans inherited their mtDNA, not the only woman alive at her time.Differences and mutations in mtDNA across various populations help estimate the timeline back to Mitochondrial Eve.
Data Analysis and Phylogenetic Trees:
Sequencing data from Y chromosomes or mtDNA is analyzed to identify specific haplogroups, which are groups of similar haplotypes that share a common ancestor with a single-nucleotide polymorphism (SNP) mutation.
Phylogenetic trees can be constructed based on these haplogroups to visualize the relationships and divergences among different lineages, tracing back to common ancestors.
Population and Evolutionary Studies:
These genetic analyses are complemented by studies of human migration, archaeological findings, and historical records to provide context for the genetic data.
By understanding the geographical distribution of haplogroups, researchers can infer migration patterns and historical events that shaped human genetic diversity.
Interdisciplinary Collaboration:
Determining common ancestors and understanding human evolutionary history is an interdisciplinary effort involving genetics, archaeology, anthropology, and history, among other fields.
By using gene sequencing in this manner, scientists can uncover details about human ancestry, trace lineage back thousands of years, and gain insights into the migration and interaction of ancient human populations.

The mutation rate in the human genome is a critical parameter for understanding genetic variation and evolution. It refers to the frequency at which new mutations occur in the genome over generations. The mutation rate can vary depending on the type of mutation and the genomic context, but here are some general figures:

Overall Mutation Rate: Recent estimates suggest that the human genome mutation rate is about **0.5 x 10^-9 per base pair per year**. This means that each base pair has a 0.5 in a billion chance of mutating each year.

Germline Mutation Rate: The germline mutation rate is particularly important for understanding heredity and evolution. For humans, it is estimated that about **50 to 100 new mutations occur in the genome of each individual per generation**. Given that there are **about 20 years per human generation**, this translates to approximately 2.5 to 5 x 10^-8 mutations per base pair per generation.

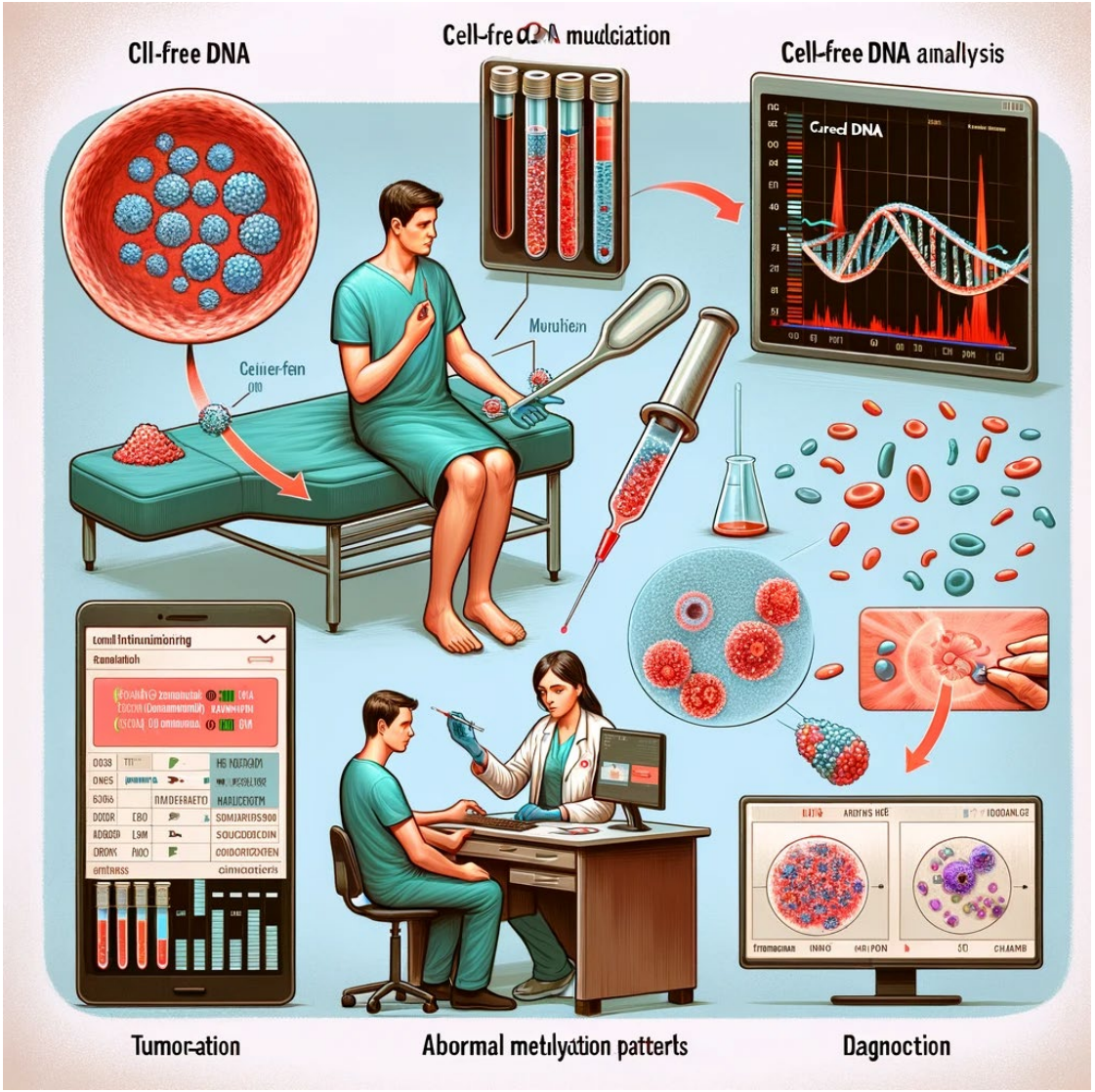
Somatic Mutation Rate: Somatic mutations are those that occur in non-reproductive cells, and they can lead to cancer and other diseases but are not passed on to offspring. The somatic mutation rate is higher than the germline mutation rate and can vary widely depending on the cell type and the individual's age and environmental exposures.

Variation Across the Genome: The mutation rate is not uniform across the human genome. It can be influenced by factors such as the local sequence context, the presence of mutational hotspots, and the activity of DNA repair mechanisms. For instance, repetitive regions of the genome tend to have higher mutation rates.

Understanding these mutation rates is crucial for fields like human genetics, evolutionary biology, and cancer research, as they help in constructing phylogenetic trees, estimating divergence times between species, and understanding the mechanisms underlying genetic diseases.

Liquid Biopsy

Cell Free DNA



Cell-free DNA (cfDNA) are small fragments of DNA that are released into the bloodstream from all cells in the body, including normal cells and cancer cells. When these fragments originate from cancer cells, they are specifically termed circulating tumor DNA (ctDNA). The use of cfDNA for cancer diagnosis and monitoring leverages the detection and analysis of ctDNA to provide insights into the genetic mutations and alterations present in the tumor.

Sample Collection and Preparation:

A blood sample is drawn from the patient. This minimally invasive procedure is often referred to as a **"liquid biopsy."**

DNA Sequencing:

The extracted cfDNA is sequenced using high-throughput sequencing technologies, often referred to as next-generation sequencing (NGS). This allows for the detection of specific genetic alterations that are characteristic of different types of cancer.

Advanced sequencing techniques can identify various genetic changes, including single nucleotide variations, insertions, deletions, copy number variations, and even larger chromosomal rearrangements.

Data Analysis:

The sequencing data is analyzed to identify the presence of ctDNA among the normal cfDNA. Bioinformatics tools are used to distinguish between the background noise of normal cfDNA and the cancer-specific alterations. The presence and amount of ctDNA can provide valuable information about the tumor's genetic profile, tumor burden, and response to treatment.

Clinical Applications:

Early Detection and Diagnosis: In some cases, ctDNA can be detected before clinical symptoms of cancer appear, offering a potential tool for early detection.

Treatment Selection: By identifying specific mutations in ctDNA, clinicians can tailor cancer treatment to the individual's tumor profile, selecting therapies that are more likely to be effective based on the genetic alterations present.

Monitoring Response to Treatment: Changes in the levels of ctDNA can indicate how well a patient is responding to treatment. A decrease in ctDNA levels can suggest that the tumor is responding to therapy, while an increase might indicate progression or recurrence.

Detection of Minimal Residual Disease and Relapse: After treatment, ctDNA monitoring can help detect minimal residual disease or early signs of relapse, potentially before they are detectable by imaging studies.

Cell Free DNA (cfDNA)

TABLE 2 Cell-free (cf)DNA concentrations and tumour response according to response evaluation criteria in solid tumours (RECIST) criteria

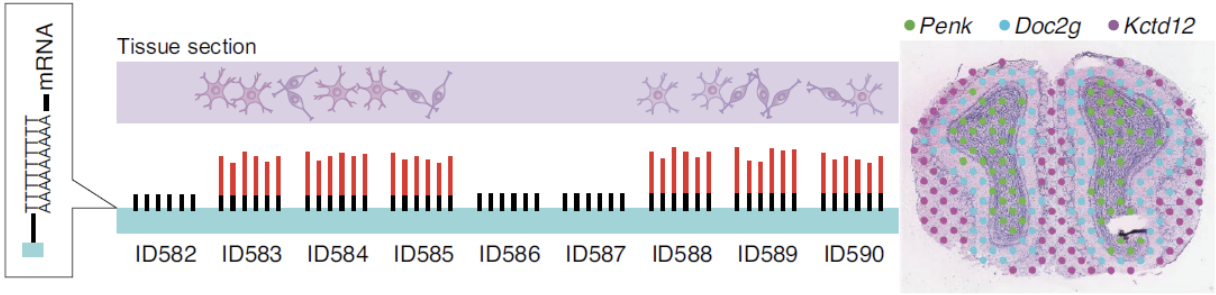
	Progressive disease	Stable disease	Partial response	p-value
Baseline concentration ng·mL^{-1}	23.88 (35.84)	32.83 (37.32)	26.79 (28.98)	0.358
Post-chemotherapy concentration ng·mL^{-1}	24.16 (21.66)	28.61 (37.92)	30.72 (61.33)	0.358
Difference between post-chemotherapy and baseline concentration ng·mL^{-1}	−0.22 (27.52)	−2.01 (28.63)	−0.56 (41.95)	0.473
Variation in concentration %	−0.01 (1.04)	−0.08 (0.92)	−0.02 (1.80)	0.402

Data are presented as median (interquartile range), unless otherwise stated.

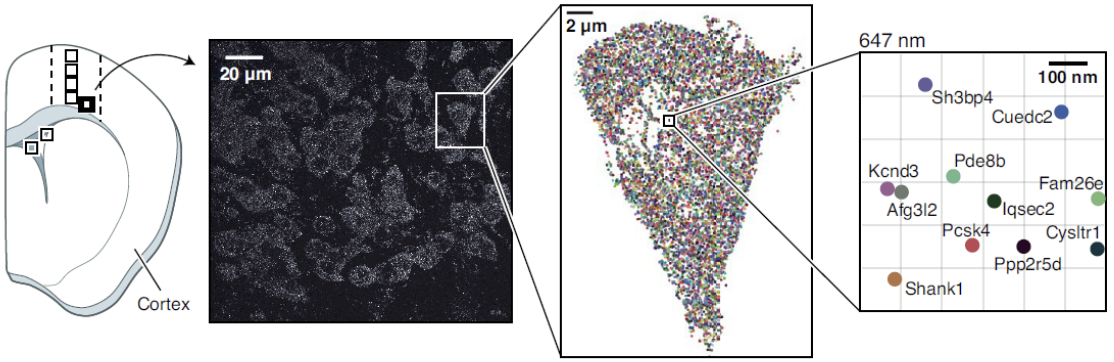
0.01%-90% circulating tumor DNA (ctDNA)

Method of the Year: spatially resolved transcriptomics

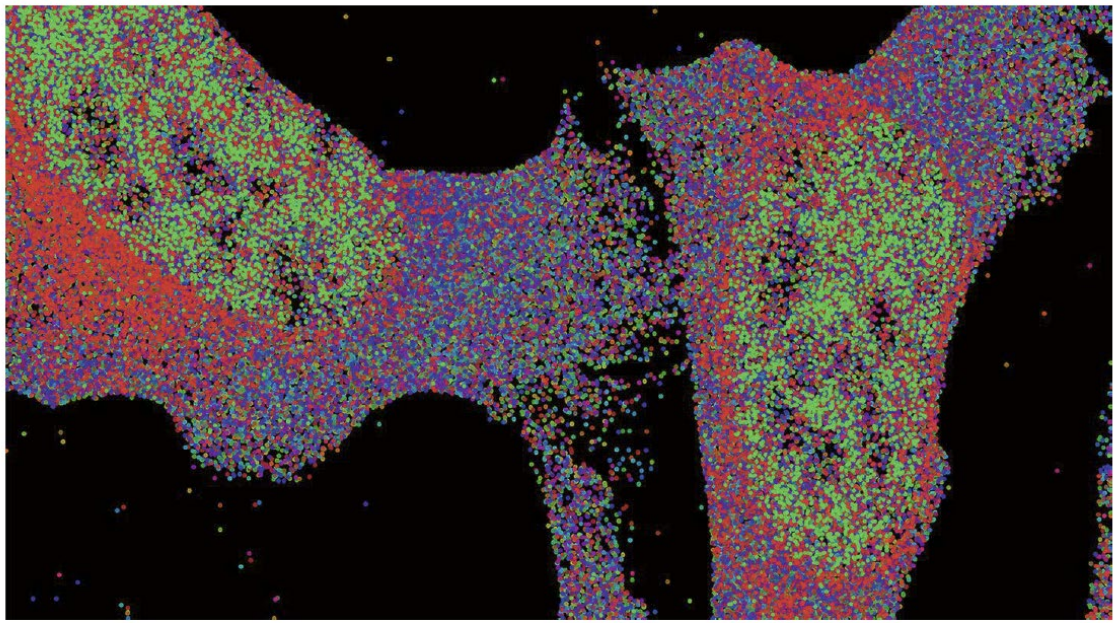
Nature Methods has crowned spatially resolved transcriptomics Method of the Year 2020.



Researchers in Sweden developed an approach in which fixed, stained tissue is imaged, permeabilized and the mRNAs attach to an array of barcoded oligos. The RNAs are reverse-transcribed; the cDNAs are sequenced and yield spatially resolved transcriptomic information. Credit: Adapted with permission from ref. [4], AAAS



Spatial techniques help with atlas-building by localizing expressed genes. Here, seqFISH+ was used to measure 10,000 genes in mouse cortex. Credit: Cai lab, Caltech, I. Strazhnik; adapted with permission from ref. [6], Springer Nature.



With MERFISH, the Zhuang lab captured the expression of 10,050 genes in individual cells in human cancer cells. RNA molecules from different genes are shown in different colors. Credit: X. Zhuang laboratory, Harvard U./HHMI

Spatial Biology

Spatial Biology is an interdisciplinary research field that focuses on the study of biological systems in relation to their spatial organization and distribution at various scales, ranging from molecular to cellular, tissue, and organism levels. This field combines advanced imaging techniques, computational methods, and molecular biology approaches to investigate the spatial arrangement of biomolecules, cells, and tissues, and how these spatial relationships influence biological functions, processes, and interactions.

GPT4-4

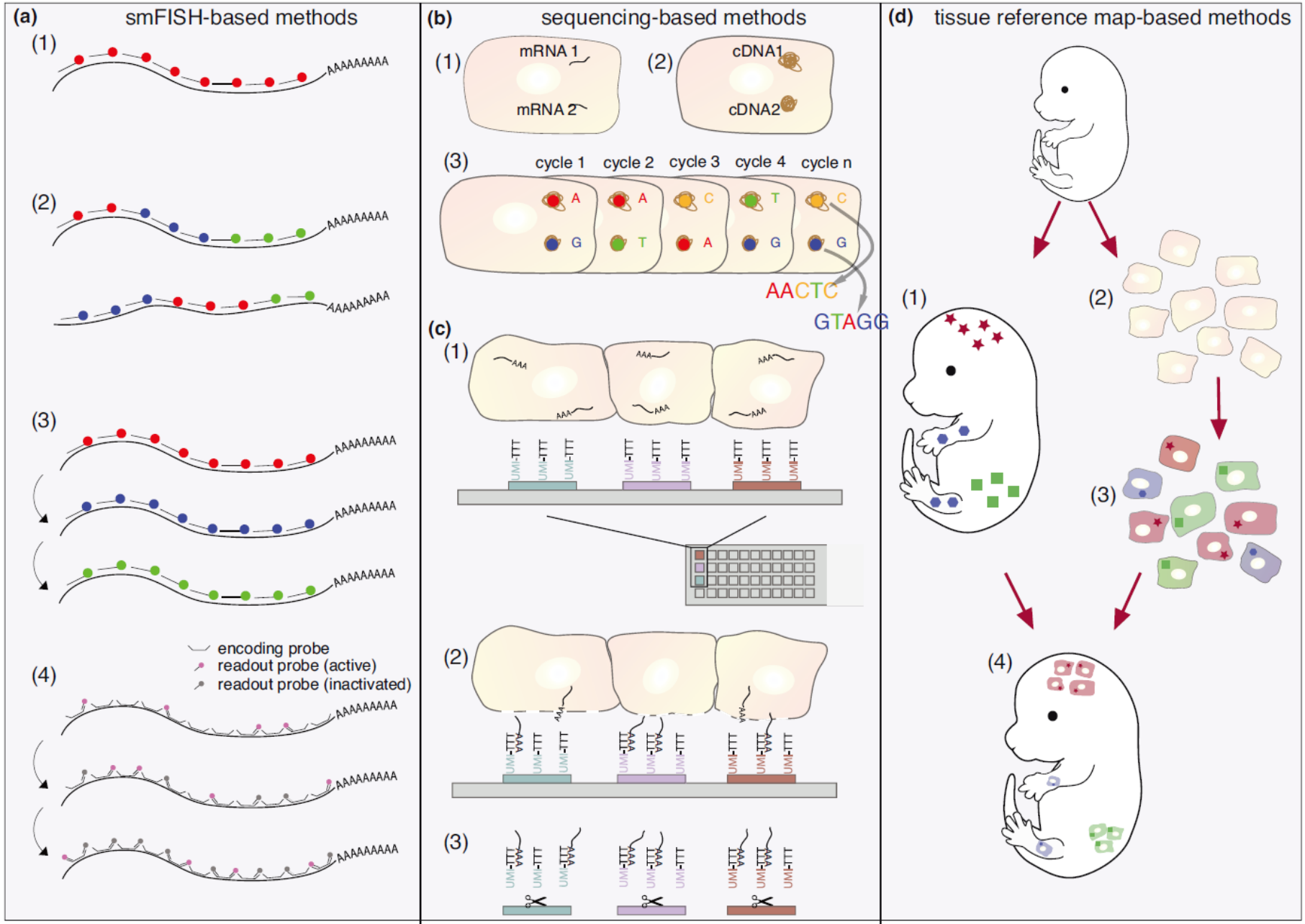
Applications of Spatial Biology

Drug Discovery: Spatial biology can be used to identify drug targets and understand their spatial organization, expression, and function within cells and tissues. This can facilitate the development of drugs that target specific cellular components, pathways, or interactions, leading to more effective and targeted therapies.

Diagnostics: Spatial biology can help identify disease-specific spatial patterns or biomarkers, which can be used for the early detection, diagnosis, and monitoring of diseases such as cancer, neurological disorders, and infectious diseases.

Personalized Medicine: Understanding the spatial organization of cells and tissues in individual patients can lead to the development of personalized treatment strategies tailored to their unique biological context. Spatial biology can inform decisions regarding drug selection, dosage, and administration, as well as the identification of patient subpopulations that may respond differently to specific therapies.

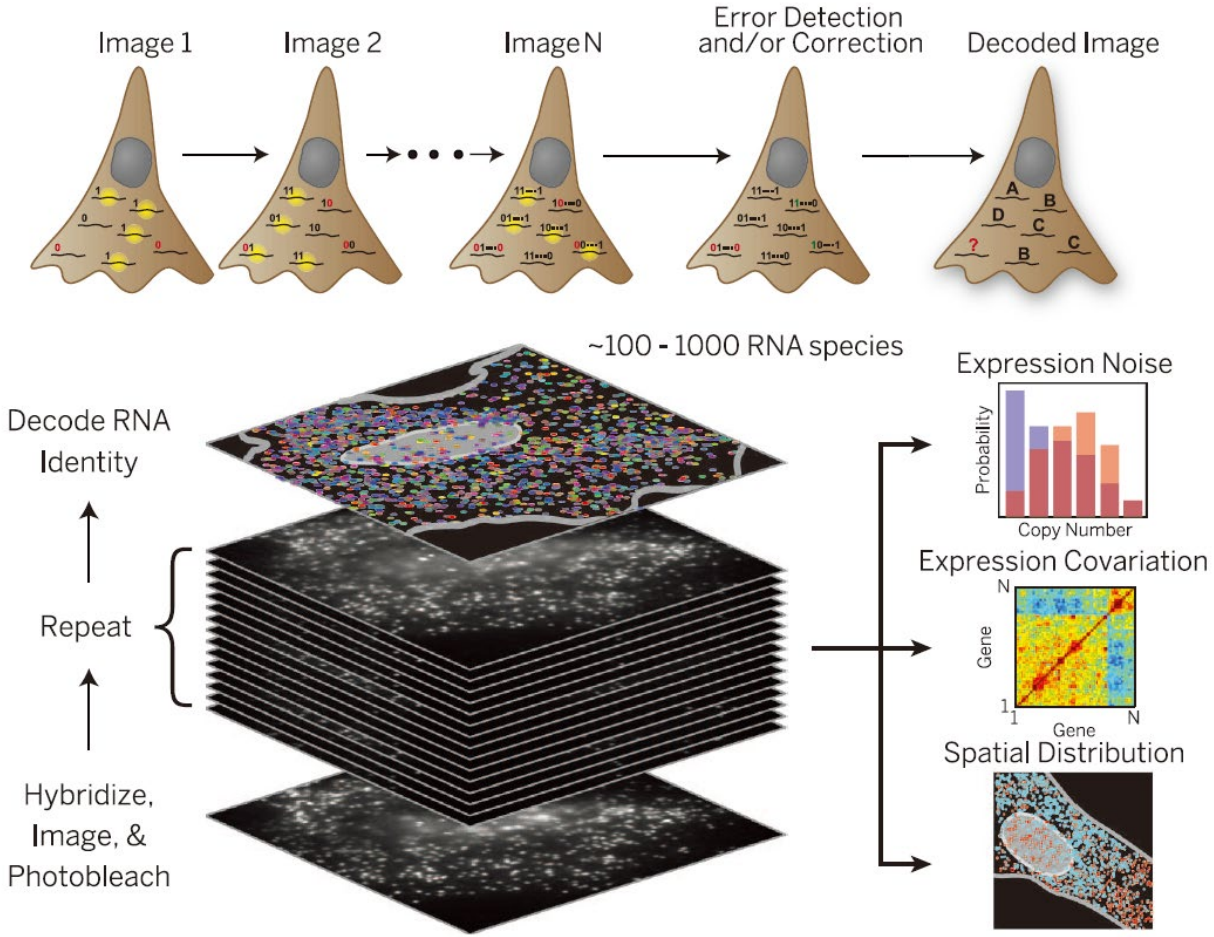
Tissue Engineering and Regenerative Medicine: Spatial biology provides insights into tissue organization, development, and repair, which can be applied to tissue engineering and regenerative medicine. This information can guide the design of artificial tissues and organs, as well as the development of strategies to promote tissue regeneration and repair following injury or disease.


(a)
smFISH-based methods
(1)
AAAAAAAAA
(2)
AAAAAAAAA
AAAAAAAAA
(3)
AAAAAAAAA
AAAAAAAAA
AAAAAAAAA
(4)
encoding probe
readout probe (active)
readout probe (inactivated)
AAAAAAAAA
AAAAAAAAA
AAAAAAAAA
(b)
sequencing-based methods
(1)
mRNA 1
mRNA 2
(2)
cDNA1
cDNA2
(3)
cycle 1
cycle 2
cycle 3
cycle 4
cycle n
A
A
C
T
C
G
T
A
G
G
AACTC
GTAGG
(c)
(1)
UMI-TTT
(2)
(3)
(d)
tissue reference map-based methods
(1)
(2)
(3)
(4)

Spatially resolved, highly multiplexed RNA profiling in single cells

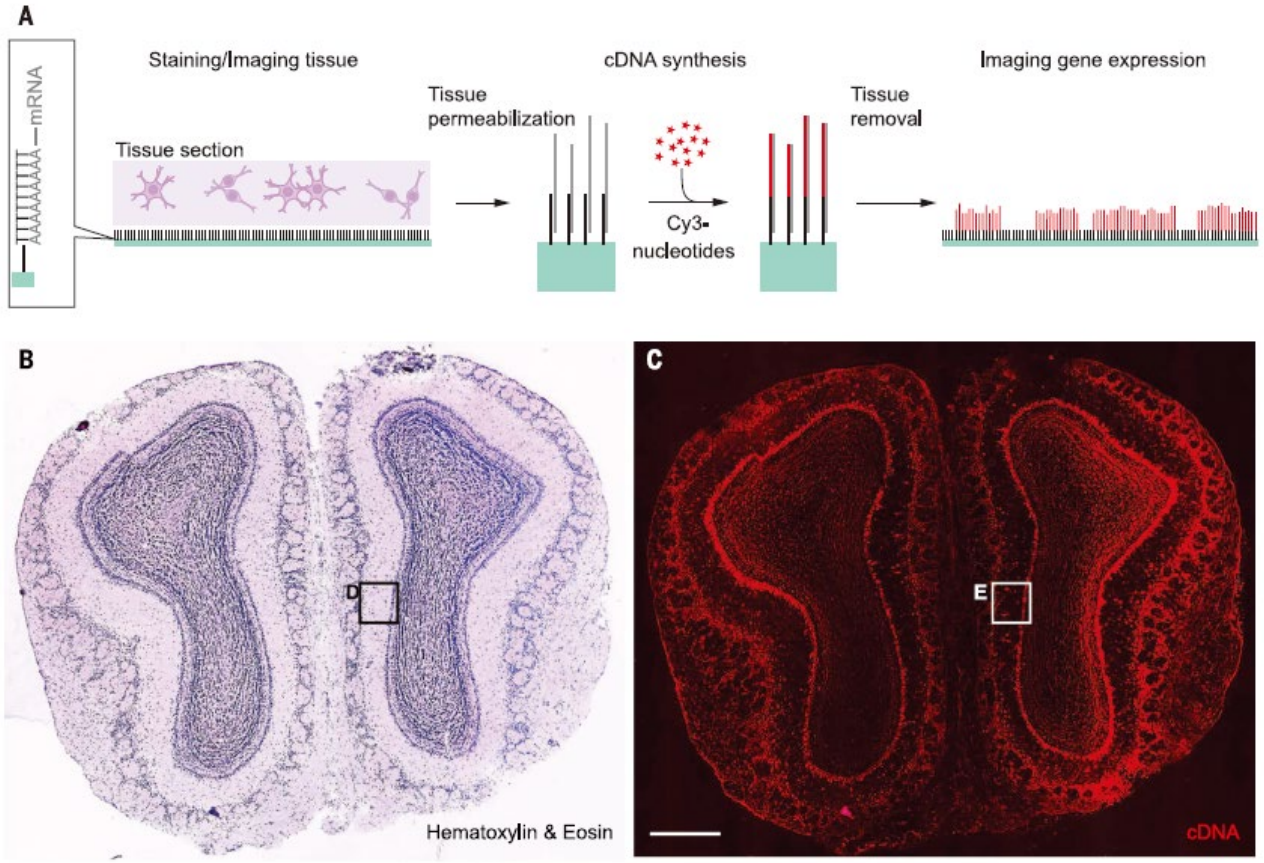
Kok Hao Chen,[1*] Alistair N. Boettiger,[1*] Jeffrey R. Moffitt,[1*]
Siyuan Wang,[1] Xiaowei Zhuang[1,2†]

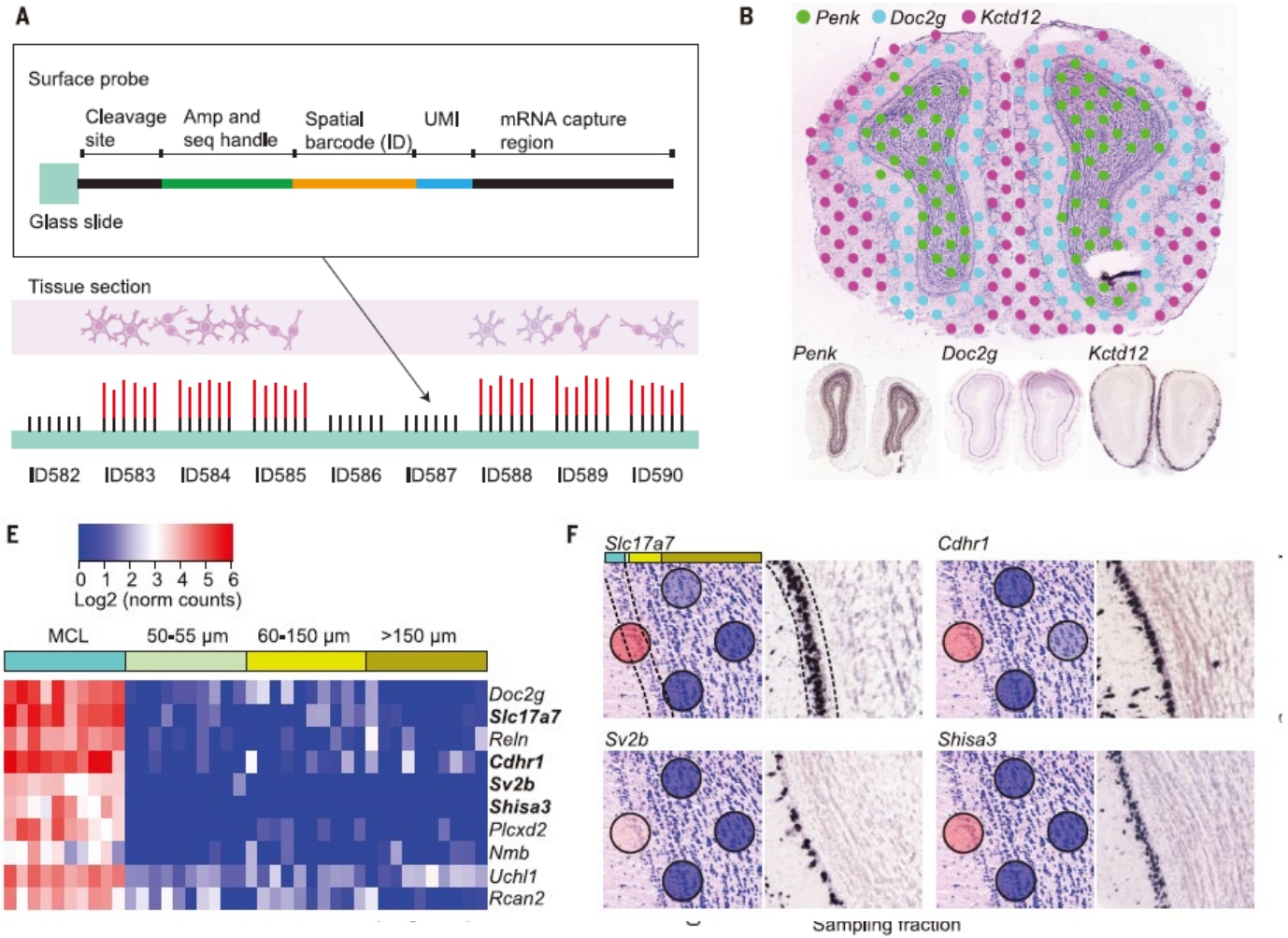
412 24 APRIL 2015 • VOL 348 ISSUE 6233 sciencemag.org SCIENCE



30 nucleotides long, which includes a 20-nucleotide sequence complementary to the target RNA, a 6- to 10-nucleotide readout sequence for the fluorescent readout probe, and a spacer region.

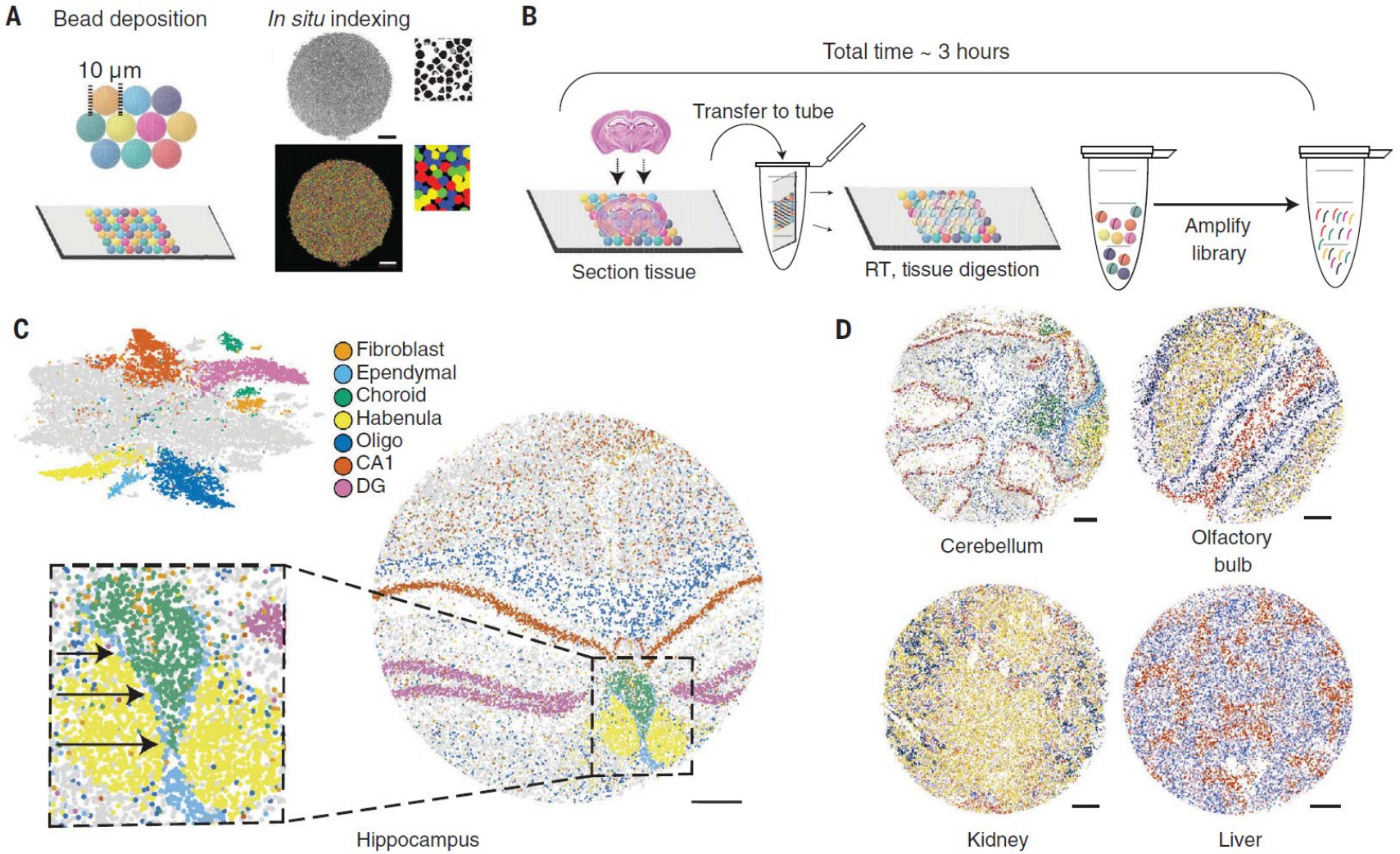
Visualization and analysis of gene expression in tissue sections by spatial transcriptomics



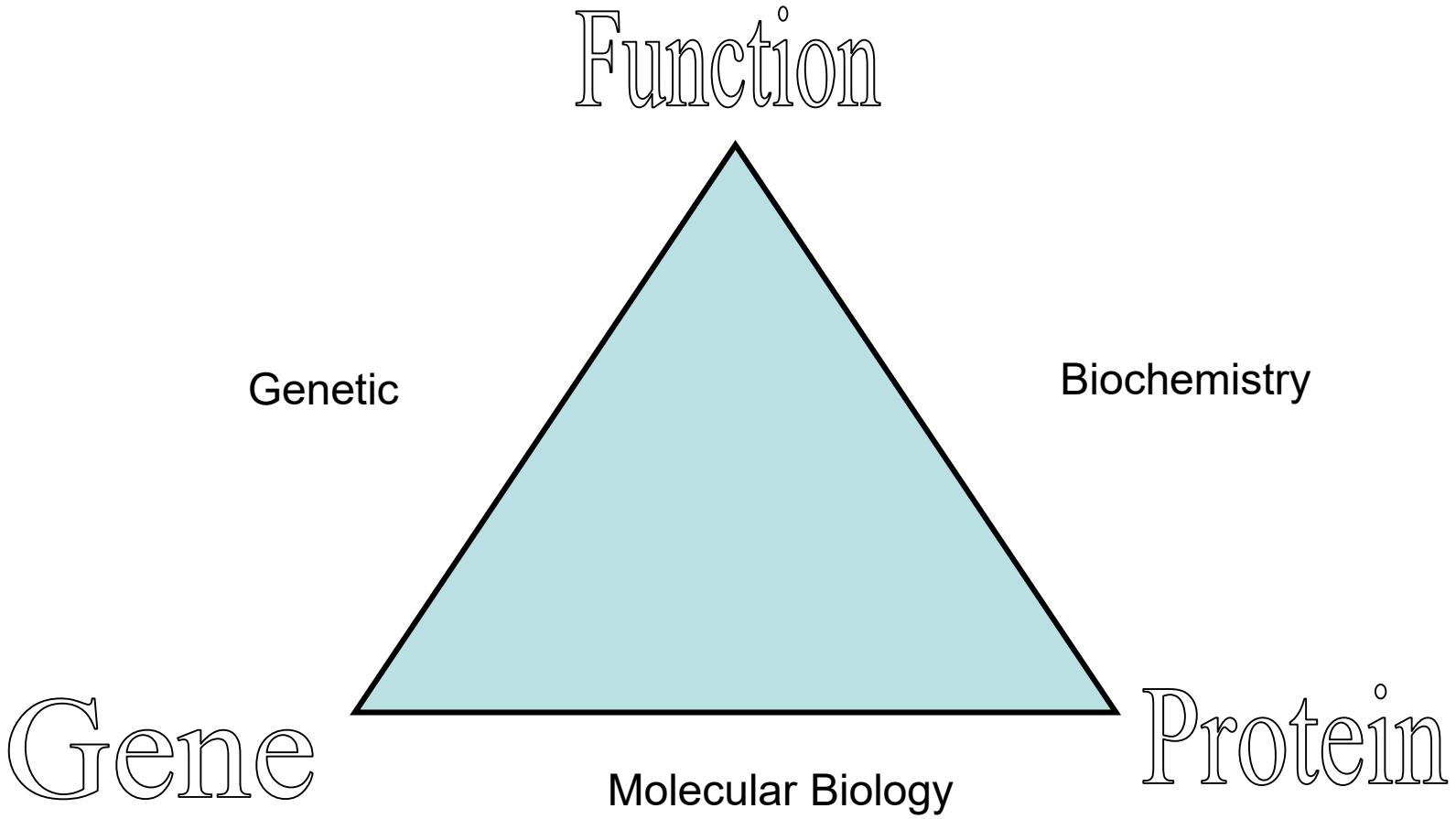

A
Surface probe
Cleavage site
Amp and seq handle
Spatial barcode (ID)
UMI
mRNA capture region
Glass slide
Tissue section
ID582
ID583
ID584
ID585
ID586
ID587
ID588
ID589
ID590
B
Penk
Doc2g
Kctd12
E
0 1 2 3 4 5 6
Log2 (norm counts)
MCL
50-55 µm
60-150 µm
>150 µm
Doc2g
Slc17a7
Reln
Cdhr1
Sv2b
Shisa3
Plcxd2
Nmb
Uchl1
Rcan2
F
Slc17a7
Cdhr1
Sv2b
Shisa3
Sampling fraction

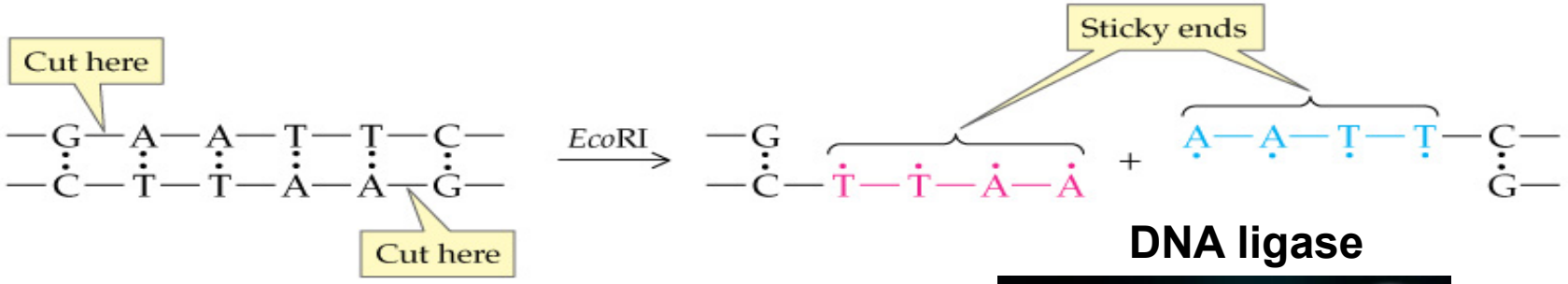
Rodriques *et al.*, *Science* **363**, 1463–1467 (2019)

Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution

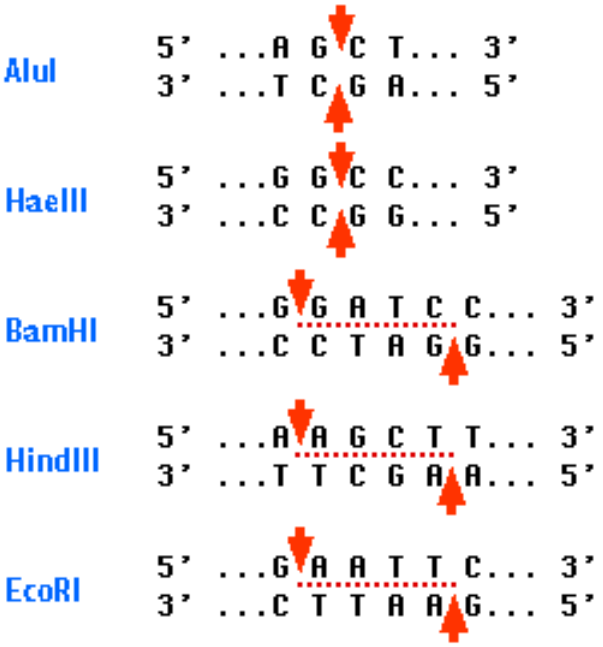


Recombinant DNA





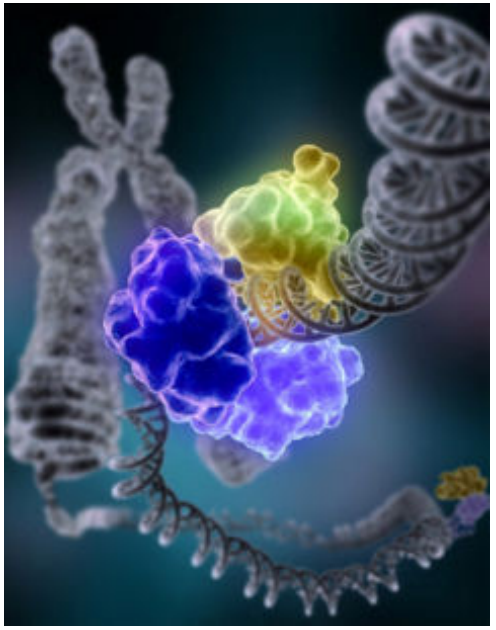
Restriction Enzyme

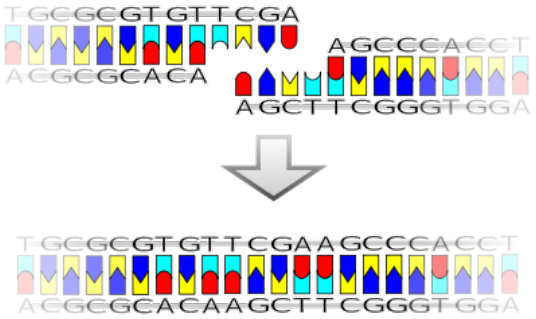


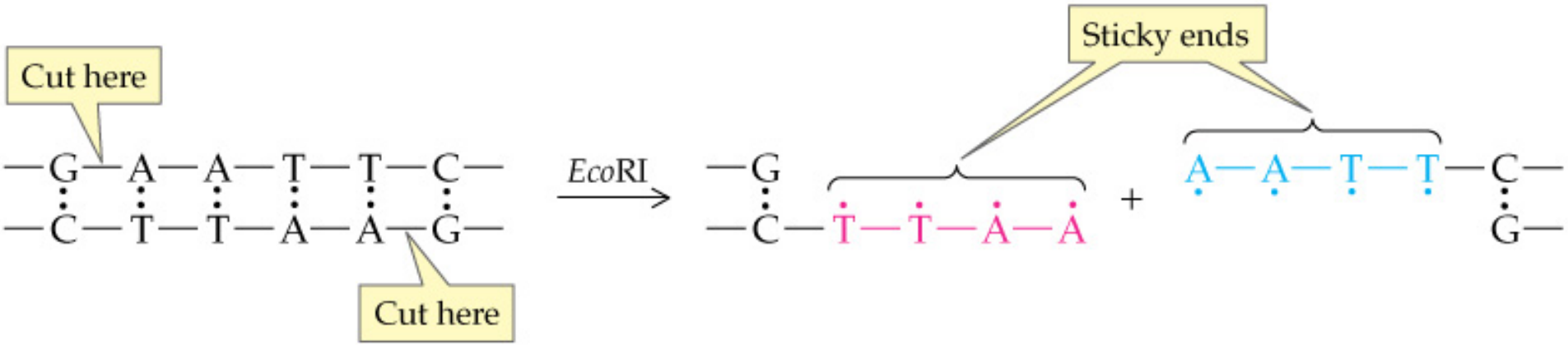
AluI and **HaeIII** produce blunt ends

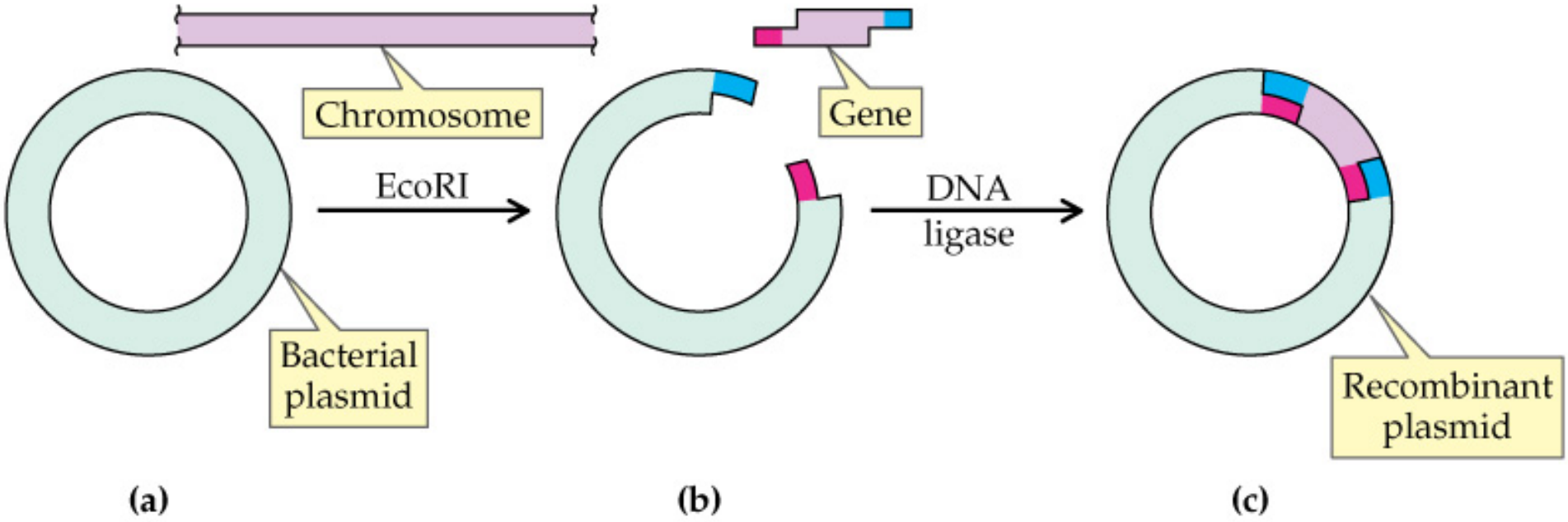
BamHI **HindIII** and **EcoRI** produce "sticky" ends

DNA ligase



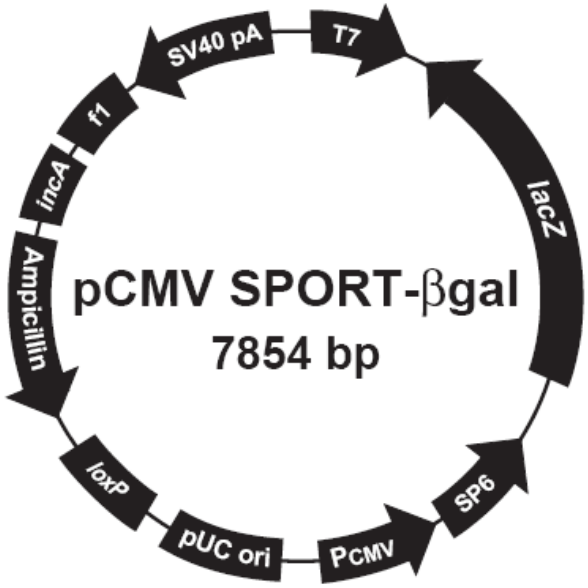


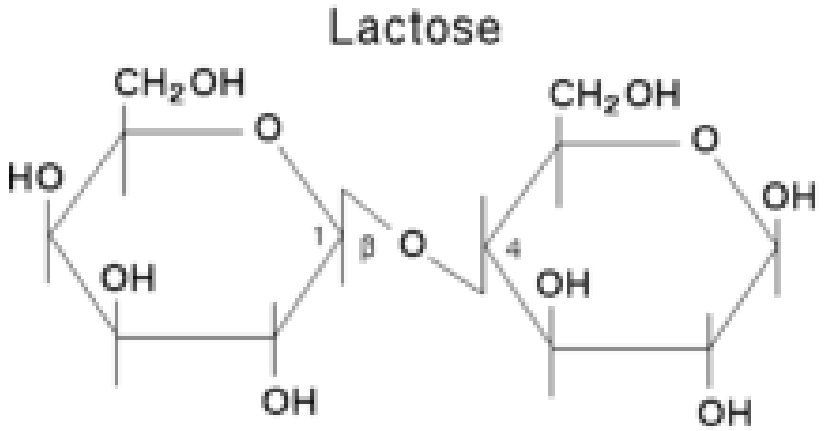

Cut here
G A A T T C
C T T A A G
Cut here
EcoRI
Sticky ends
G
C T T A A
+
A A T T C
G


Chromosome
Gene
EcoRI
DNA
ligase
Bacterial
plasmid
Recombinant
plasmid
(a)
(b)
(c)

β-Galactosidase

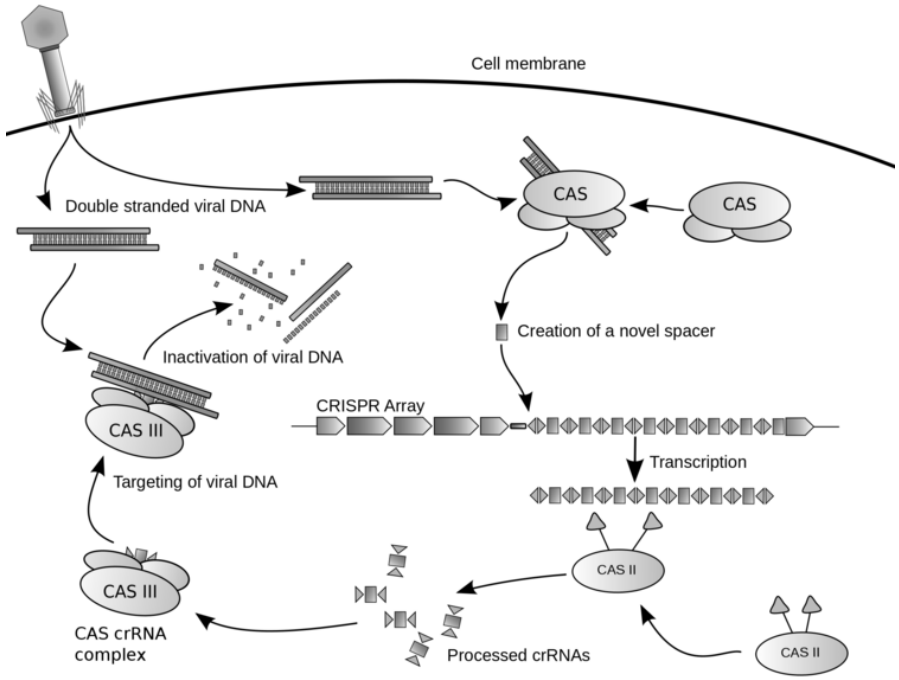
The enzyme that splits lactose into glucose and galactose. Coded by a gene (lacZ) in the lac operon of Escherichia coli.



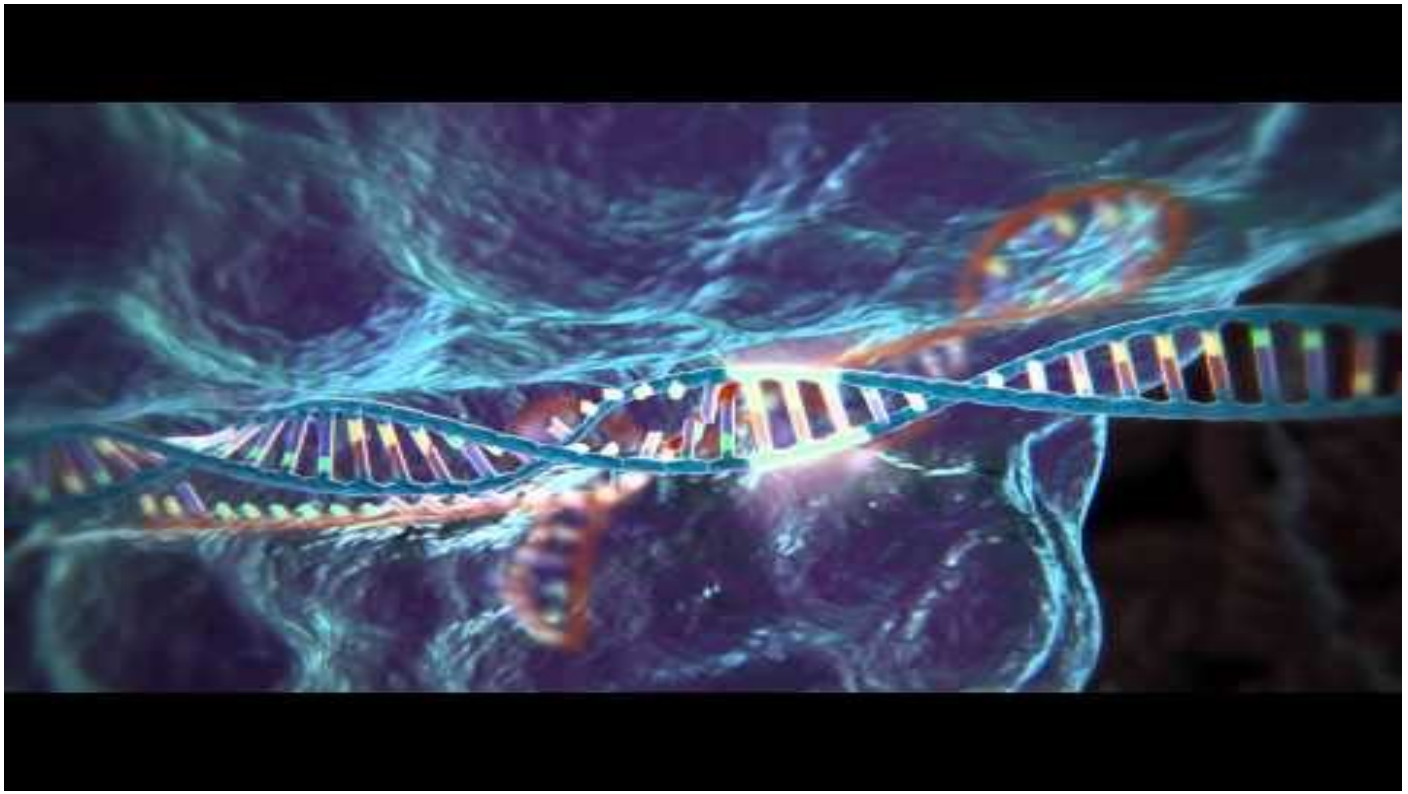


PUC is a family of plasmids that have an ampicillin resistance gene and more importantly a *lacZ* gene. A functional lacZ gene will produce the protein β - galactosidase. Bacterial colonies in which β - galactosidase is produced, will form blue colonies in the presence of the substrate 5 - bromo - 4 - chloro - 3 - indolyl - b - D - galactoside or as it is more commonly referred to, X-gal.

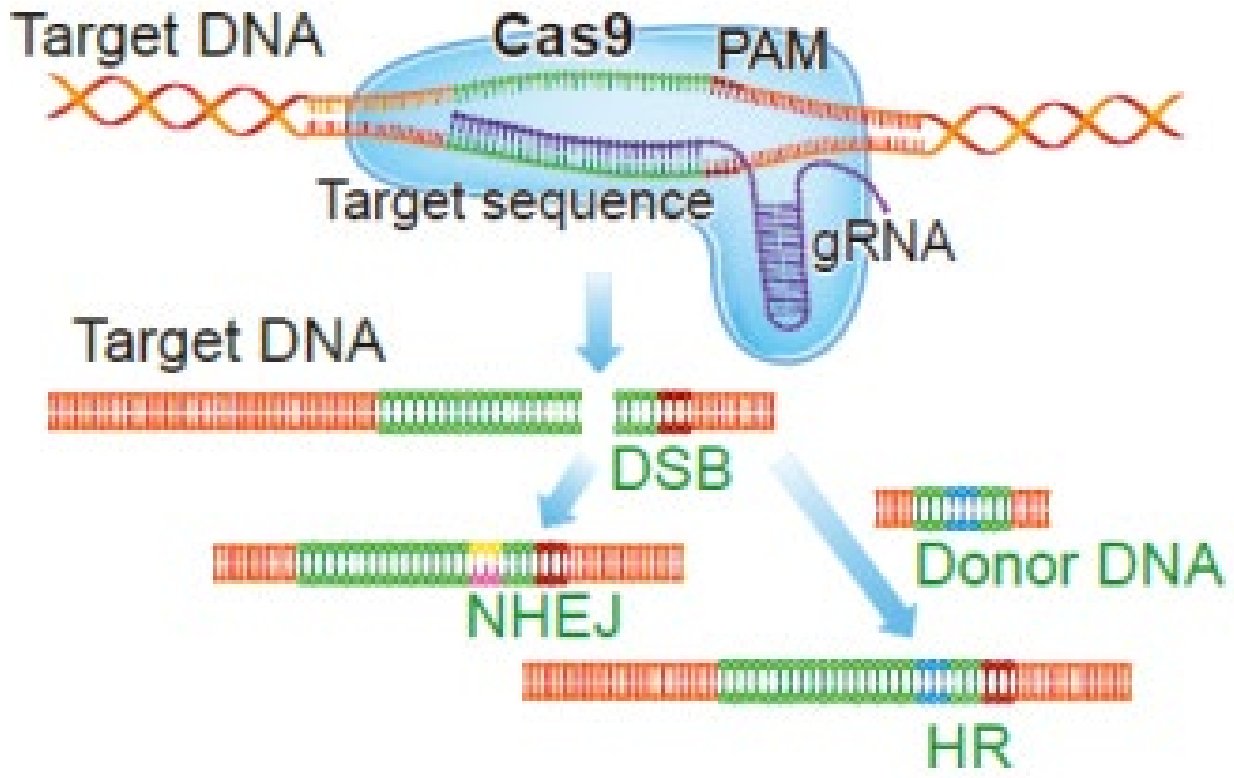
CRISPRs (**clustered regularly interspaced short palindromic repeats**) are segments of prokaryotic DNA containin g short repetitions of base sequences. Each repetition is followed by short segments of "spacer DNA" from previous exposures to a bacterial virus or



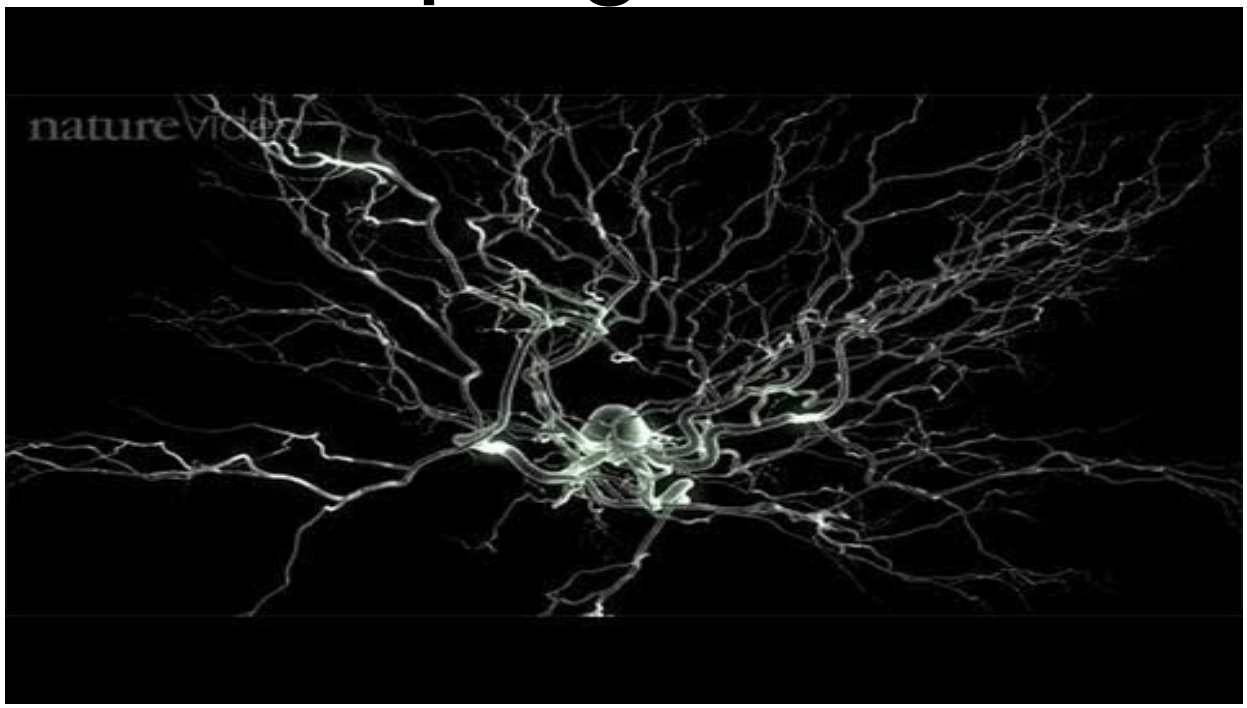
CRISPR CAS9

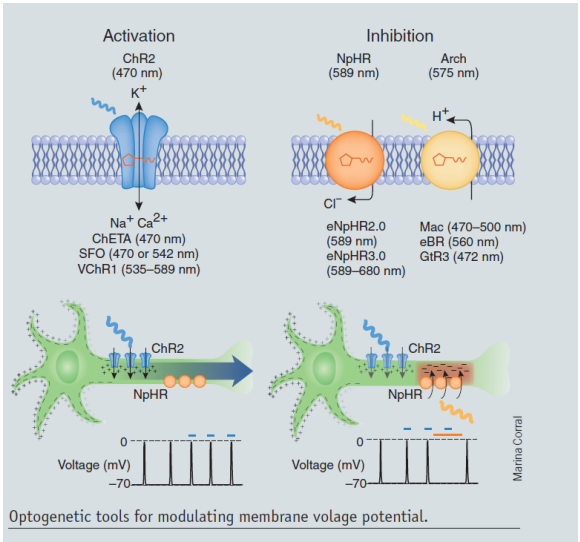


CRISPRs (**clustered regularly interspaced short palindromic repeats**) are segments of prokaryotic DNA containing short repetitions of base sequences. Each repetition is followed by short segments of "spacer DNA" from previous exposures to a bacterial virus


Target DNA
Cas9
PAM
Target sequence
gRNA
Target DNA
DSB
NHEJ
Donor DNA
HR

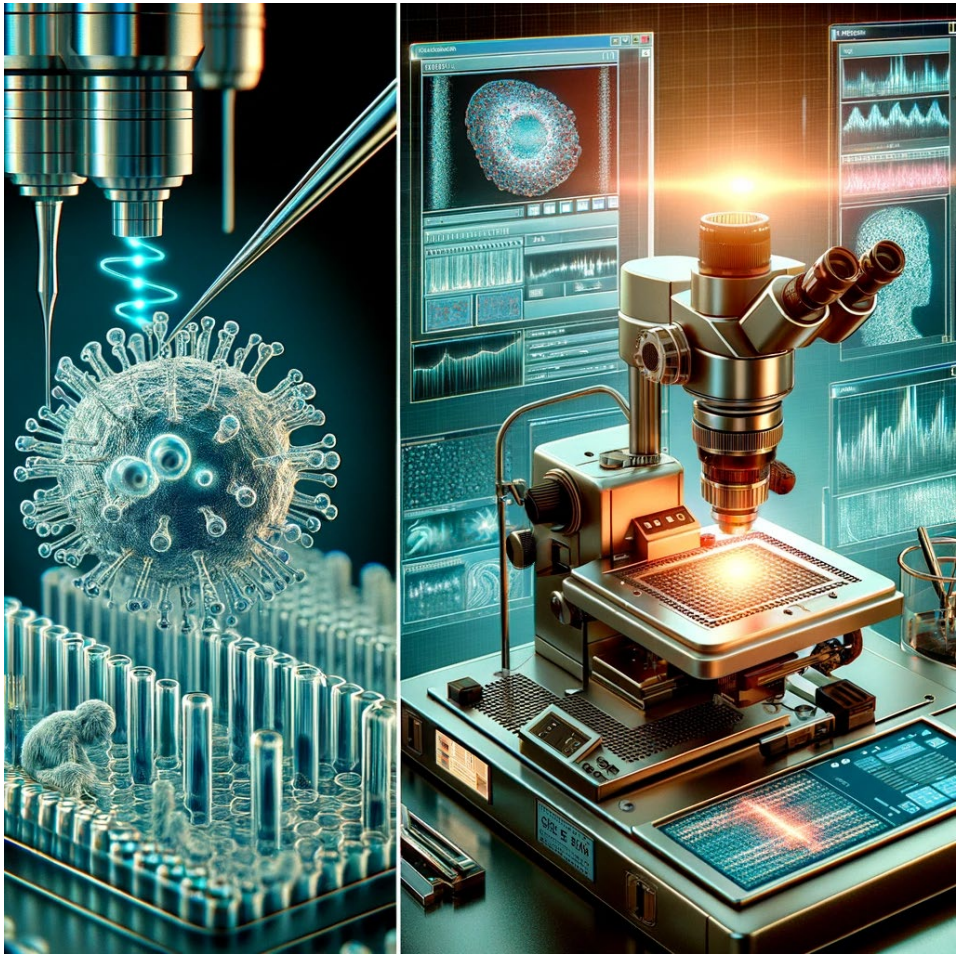
Optogenetics





Optogenetic tools for modulating membrane volage potential.

Single Cell Analysis



Single-cell analysis refers to the study of individual cells isolated from tissues in a mixed population, which provides a detailed examination of the cellular differences and a deeper understanding of the biological function and complexity at the single-cell level. This approach is crucial in various fields of biological sciences and medicine, including developmental biology, neuroscience, immunology, and cancer research. Below are key aspects and methodologies of single-cell analysis:

Cell Isolation: The first step in single-cell analysis is to isolate individual cells from a tissue or cell culture. Various techniques can be used for this purpose, including flow cytometry, microfluidics, laser capture microdissection, and manual cell picking.

Single-Cell Sequencing: Once individual cells are isolated, their genomic, transcriptomic, or epigenomic content can be analyzed. Single-cell RNA sequencing (scRNA-seq) is one of the most common methods, allowing researchers to measure gene expression levels in individual cells. This provides insights into the cellular heterogeneity within a tissue and can identify distinct cell types and states, even within seemingly homogeneous cell populations.

Single-Cell Genomics and Epigenomics: Beyond transcriptomics, single-cell DNA sequencing can reveal genomic variations at the single-cell level, such as mutations or copy number variations. Single-cell epigenomics, including methods like single-cell ATAC-seq, can assess chromatin accessibility and other epigenetic features, offering clues about the regulatory mechanisms driving gene expression in individual cells.

Proteomics and Metabolomics: Advanced techniques also enable the analysis of proteins and metabolites at the single-cell level. These approaches can provide functional data that complements genomic and transcriptomic information, offering a more comprehensive view of the cell's state and activity.

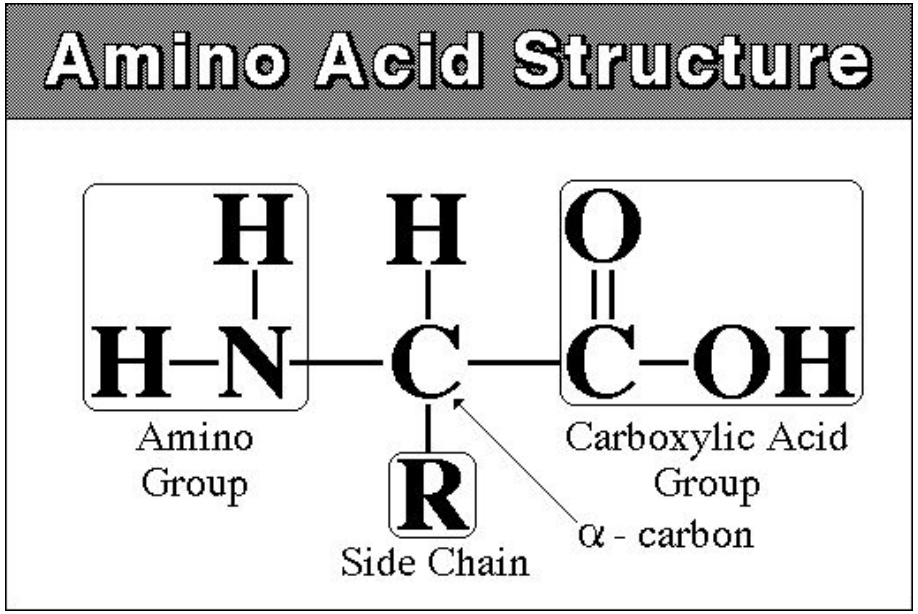
Spatial Analysis: Recent advancements have enabled the spatial characterization of cells within their native tissue contexts. Techniques like spatial transcriptomics and imaging-based methods allow researchers to understand not only the individual cell profiles but also their spatial organization and interactions within the tissue.

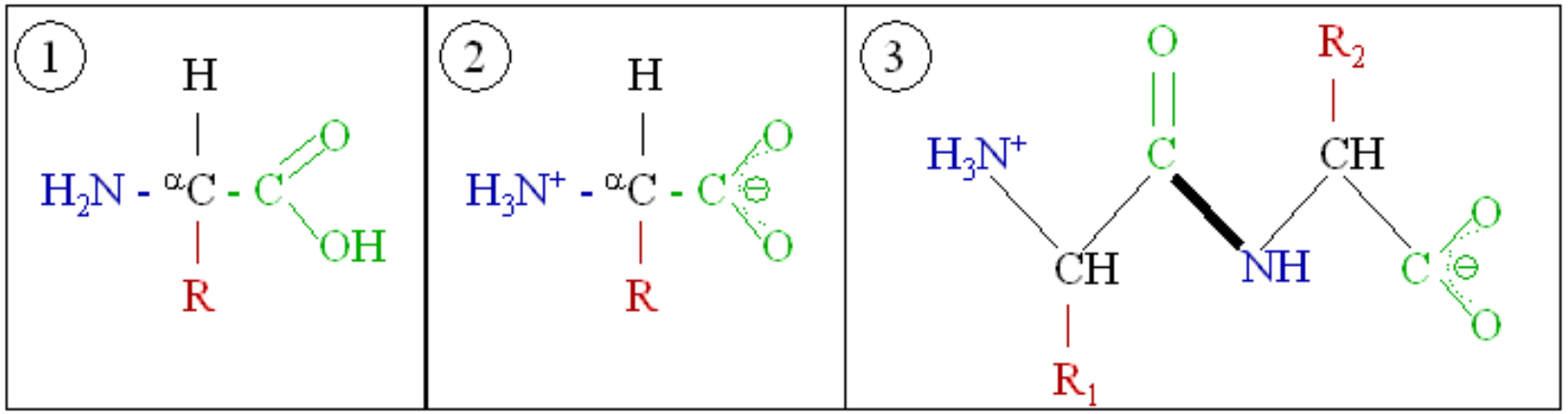
Data Analysis and Integration: Single-cell datasets are typically large and complex, requiring advanced computational tools for analysis and interpretation. Bioinformatics and data analysis methods are employed to cluster cells into distinct groups, identify marker genes, analyze differential expression, and infer developmental trajectories or cell-cell interaction networks.

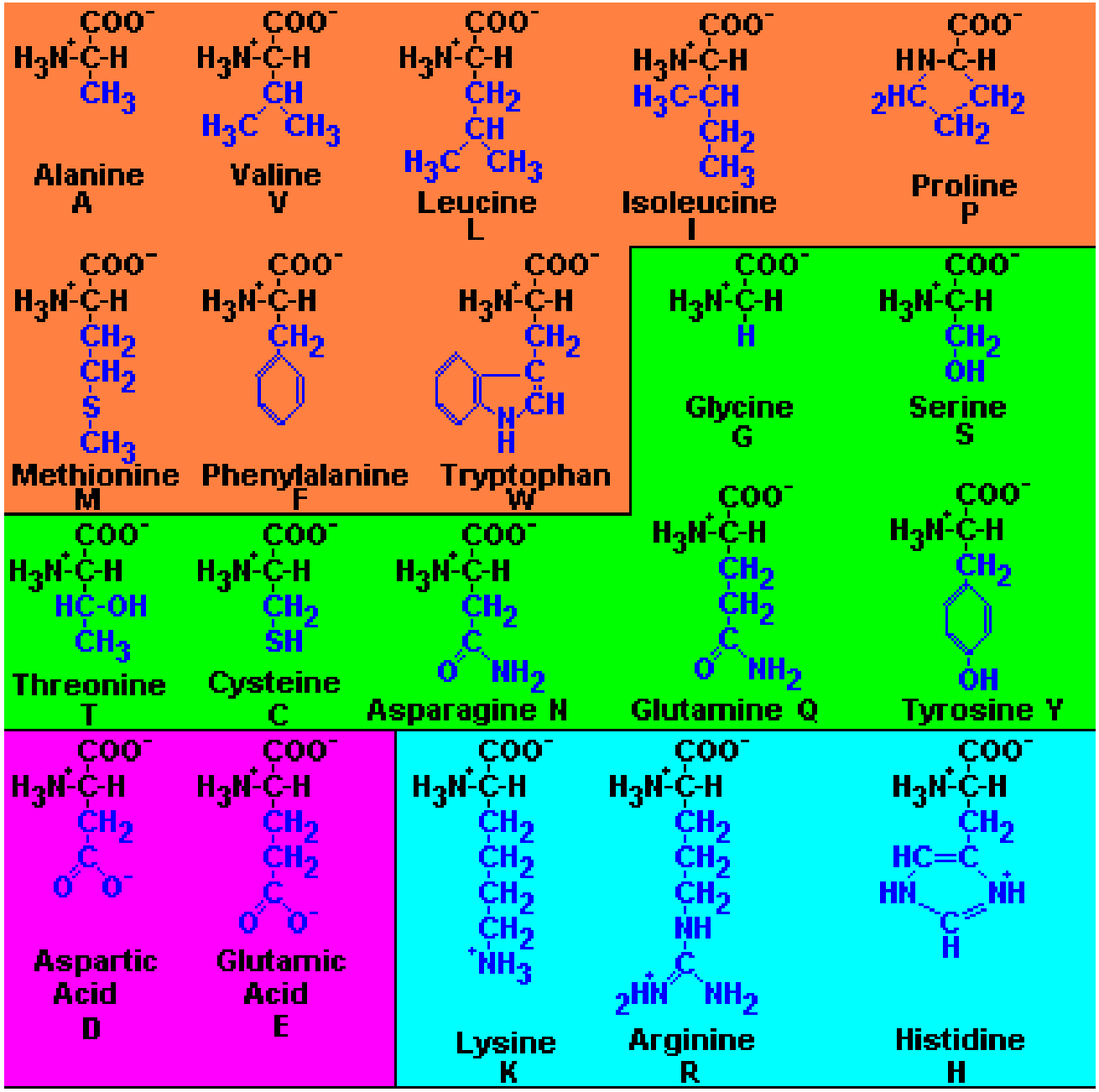
Applications: Single-cell analysis has numerous applications across various fields of biology and medicine. For example, in cancer research, it can reveal the heterogeneity within tumors, identify rare cancer stem cells, or uncover mechanisms of drug resistance. In developmental biology, it can elucidate cell lineage relationships and developmental pathways. In immunology, it helps in characterizing immune cell diversity and responses.

Quality Control and Data Preprocessing:
Quality control (QC) is crucial to remove low-quality cells and genes that could distort the analysis. This step typically involves filtering out cells with extremely high or low total gene counts or high mitochondrial gene expression, which might indicate dead or damaged cells.
Dimensionality Reduction:
High-dimensional single-cell data are often reduced to lower dimensions for visualization and further analysis. Techniques such as PCA (Principal Component Analysis), t-SNE (t-Distributed Stochastic Neighbor Embedding), and UMAP (Uniform Manifold Approximation and Projection) are popular for this purpose.
Dimensionality reduction serves to highlight the inherent structure of the data, revealing clustering of cells that might correspond to different cell types or states.
Clustering:
Once the data are in a reduced dimension space, clustering algorithms can identify groups of cells with similar expression profiles. These clusters often correspond to different cell types or subtypes within the sample.
Various clustering algorithms can be used, such as k-means, hierarchical clustering, or graph-based clustering methods (e.g., Louvain algorithm). The choice of algorithm and parameters (e.g., resolution in Louvain) can significantly affect the results, so it may require optimization based on the data.
Differential Expression Analysis:
To characterize the identified clusters and infer their biological significance, differential expression analysis is performed to identify genes that are significantly up- or down-regulated between clusters.
This step can highlight marker genes that define cell types or states and provide insights into the biological processes and pathways active in different cell populations.
Annotation and Interpretation:
The next step is to annotate the identified cell clusters with known cell types, which can be done by comparing the expression of marker genes in each cluster with known profiles from the literature or reference datasets.
Further biological interpretation can involve pathway analysis, gene set enrichment analysis, or network analysis to understand the functions and interactions of the differentially expressed genes.
Integration and Comparison:
If analyzing data from multiple samples or conditions, integration methods can be used to align datasets, allowing for comparative analysis and identification of condition-specific responses or cell types.
Techniques for data integration include canonical correlation analysis, mutual nearest neighbors, or specialized tools like Seurat or Scanpy's integration methods.
Trajectory Inference:

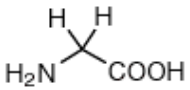
Amino Acid


Amino Acid Structure
H
H
O
H–N
C
C–OH
Amino Group
Carboxylic Acid Group
R
α - carbon
Side Chain

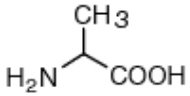

1
H
H_2N - $^{\alpha}C$ - C
O
OH
R
2
H
H_3N^+ - $^{\alpha}C$ - C
O
O
R
3
O
R_2
H_3N^+
C
CH
CH
NH
C
O
O
R_1


Alanine
A
Valine
V
Leucine
L
Isoleucine
I
Proline
P
Methionine
M
Phenylalanine
F
Tryptophan
W
Glycine
G
Serine
S
Threonine
T
Cysteine
C
Asparagine N
Glutamine Q
Tyrosine Y
Aspartic
Acid
D
Glutamic
Acid
E
Lysine
K
Arginine
R
Histidine
H

Small

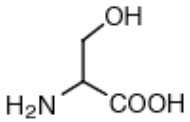


Glycine (Gly, G)
MW: 57.05

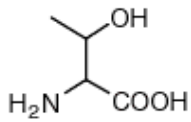


Alanine (Ala, A)
MW: 71.09

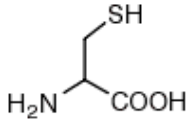
Nucleophilic



Serine (Ser, S)
MW: 87.08, pK_a ~ 16

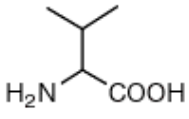


Threonine (Thr, T)
MW: 101.11, pK_a ~ 16

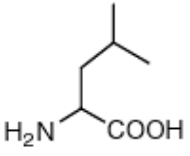


Cysteine (Cys, C)
MW: 103.15, pK_a = 8.35

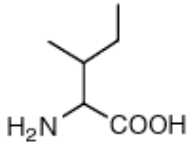
Hydrophobic



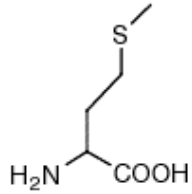
Valine (Val, V)
MW: 99.14



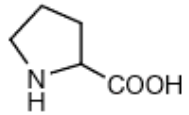
Leucine (Leu, L)
MW: 113.16



Isoleucine (Ile, I)
MW: 113.16

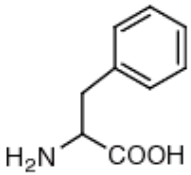


Methionine (Met, M)
MW: 131.19

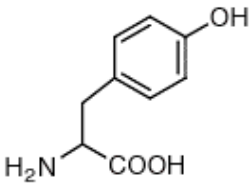


Proline (Pro, P)
MW: 97.12

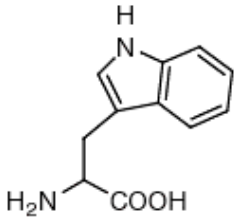
Aromatic



Phenylalanine (Phe, F)
MW: 147.18

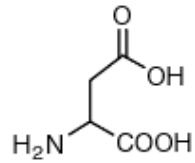


Tyrosine (Tyr, Y)
MW: 163.18

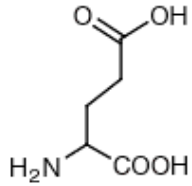


Tryptophan (Trp, W)
MW: 186.21

Acidic

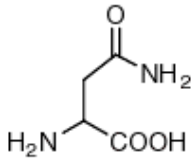


Aspartic Acid (Asp, D)
MW: 115.09, pK_a = 3.9

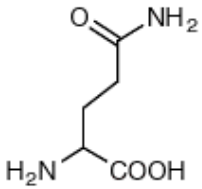


Glutamic Acid (Glu, E)
MW: 129.12, pK_a = 4.07

Amide

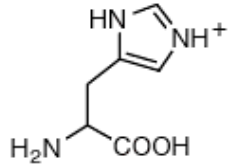


Asparagine (Asn, N)
MW: 114.11

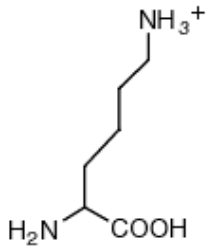


Glutamine (Gln, Q)
MW: 128.14

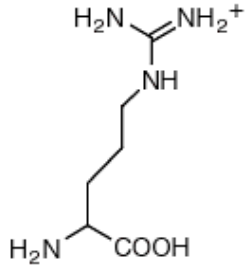
Basic



Histidine (His, H)
MW: 137.14, pK_a = 6.04



Lysine (Lys, K)
MW: 128.17, pK_a = 10.79

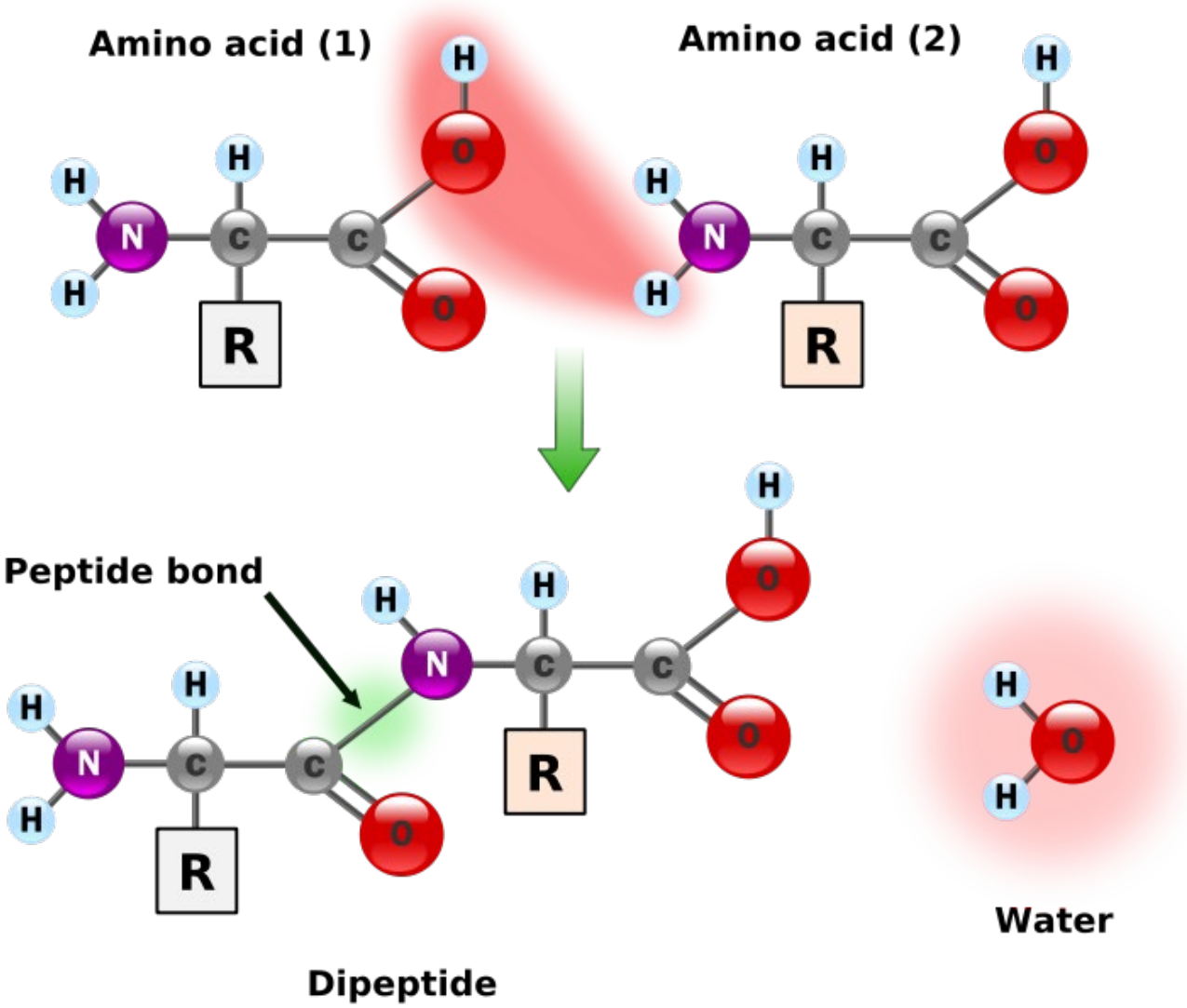


Arginine (Arg, R)
MW: 156.19, pK_a = 12.48

Protein Structure and Function

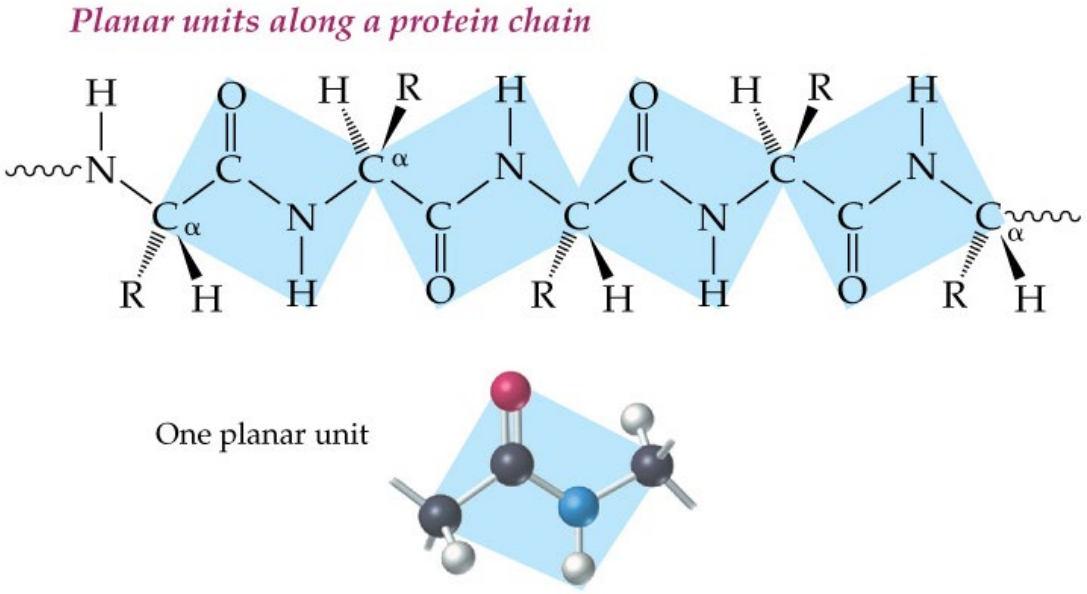
- Proteins are polymers of amino acids.
- Each amino acids in a protein contains a amino group, -NH_2, a carboxyl group, -COOH, and an R group, all bonded to the central carbon atom. The R group may be a hydrocarbon or they may contain functional group.
- All amino acids present in a proteins are *α-amino acids* in which the amino group is bonded to the carbon next to the carboxyl group.
- Two or more amino acids can join together by forming amide bond, which is known as a ***peptide bond*** when they occur in proteins.

Peptide bond



Primary Protein Structure

- Primary structure of a proteins is the sequence of amino acids connected by peptide bonds. Along the backbone of the proteins is a chain of alternating peptide bonds and α-carbons and the amino acid side chains are connected to these

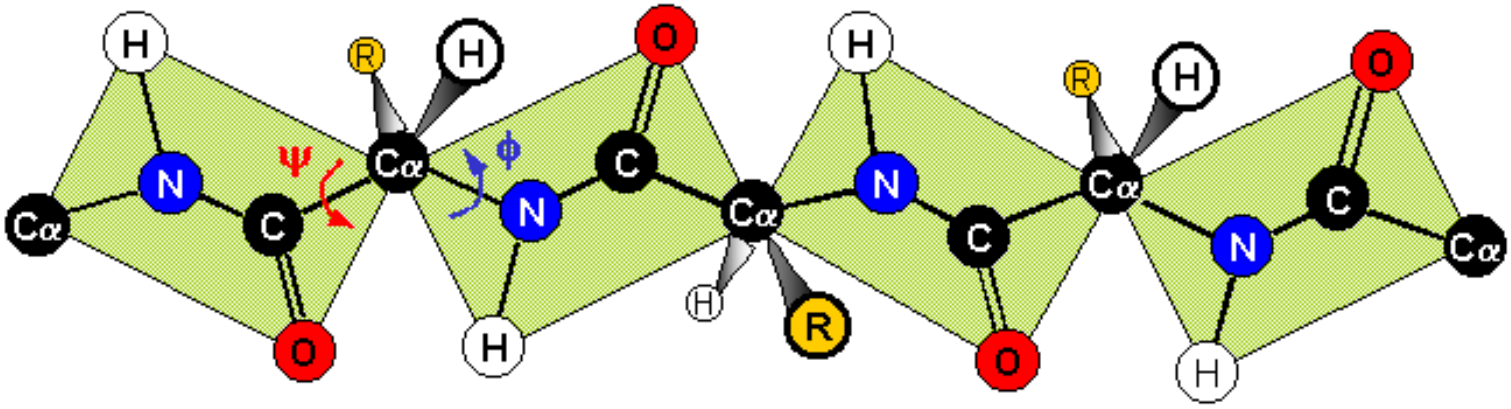


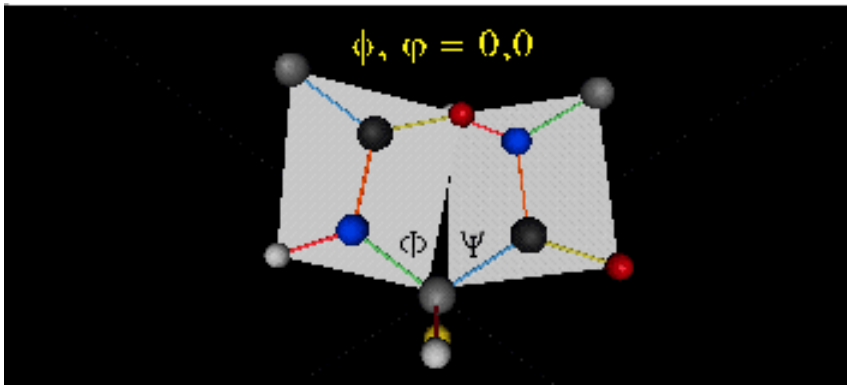
- By convention, peptides and proteins are always written with the amino terminal amino acid (N-terminal) on the left and carboxyl-terminal amino acid (C-terminal) on the right.

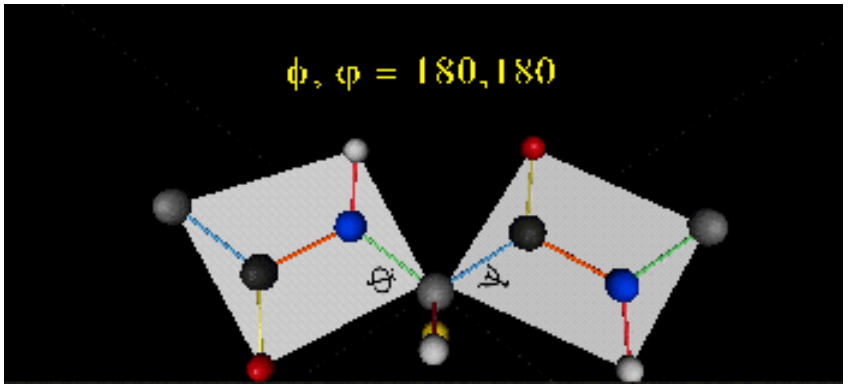
Secondary Protein Structure

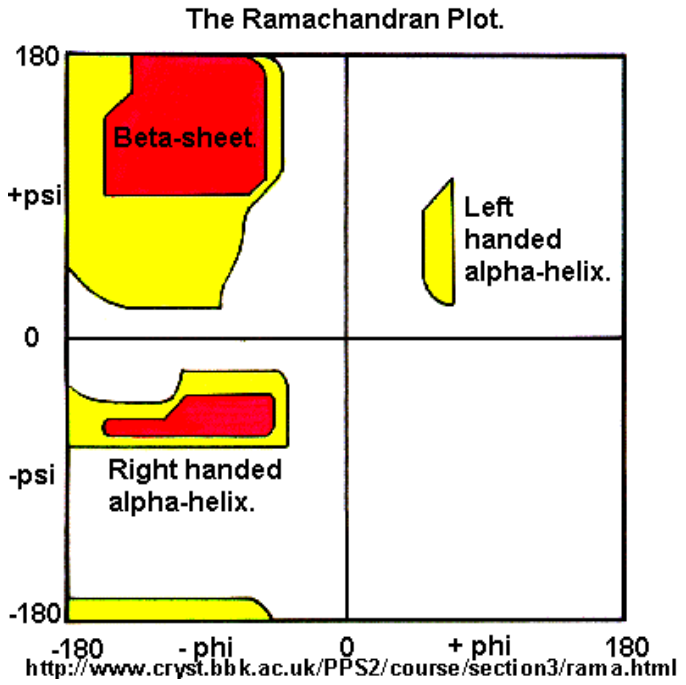
- Secondary structure of a protein is the arrangement of polypeptide backbone of the protein in space. The secondary structure includes two kinds of repeating pattern known as the *α-helix* and *β-sheet*.
- Hydrogen bonding between backbone atoms are responsible for both of these secondary structures.

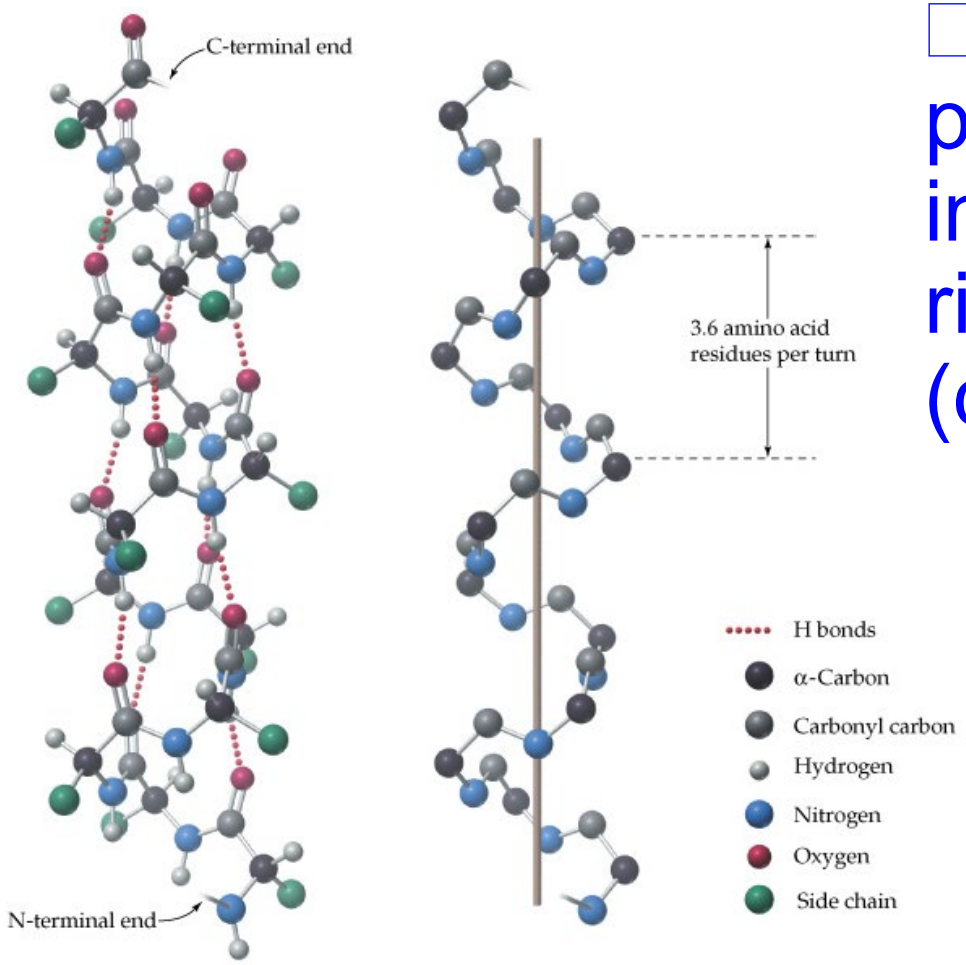
FULLY EXTENDED POLYPEPTIDE CHAIN





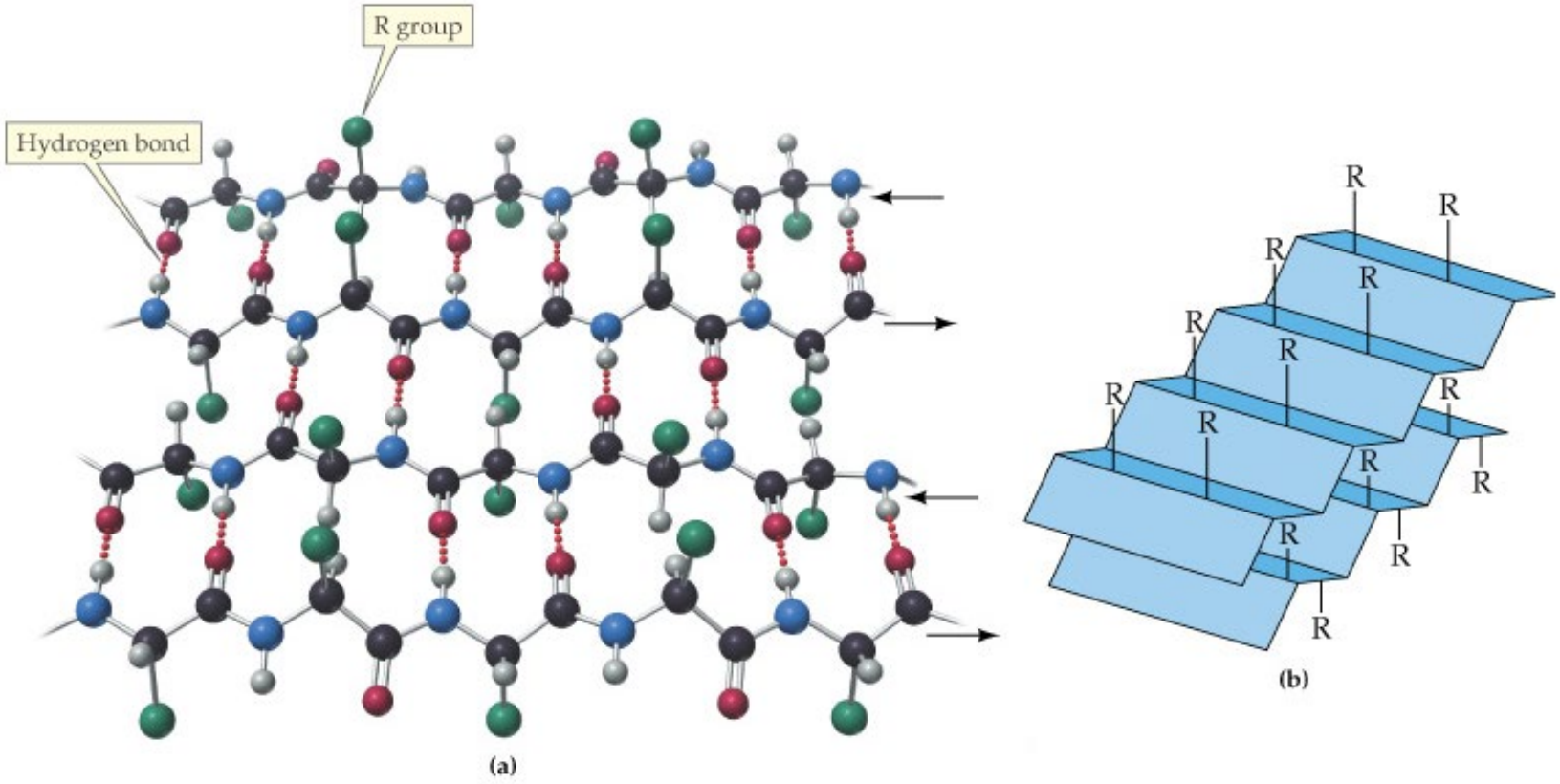




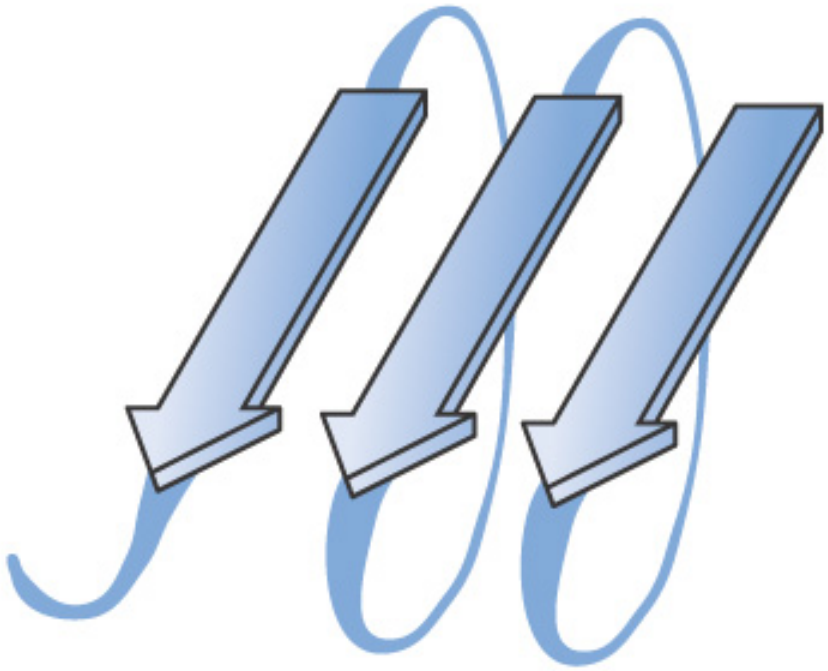


- ***α-Helix:*** A single protein chain coiled in a spiral with a right-handed (clockwise) twist.

β-Sheet: The polypeptide chain is held in place by hydrogen bonds between pairs of peptide units along neighboring backbone segments.



α helix



β sheet

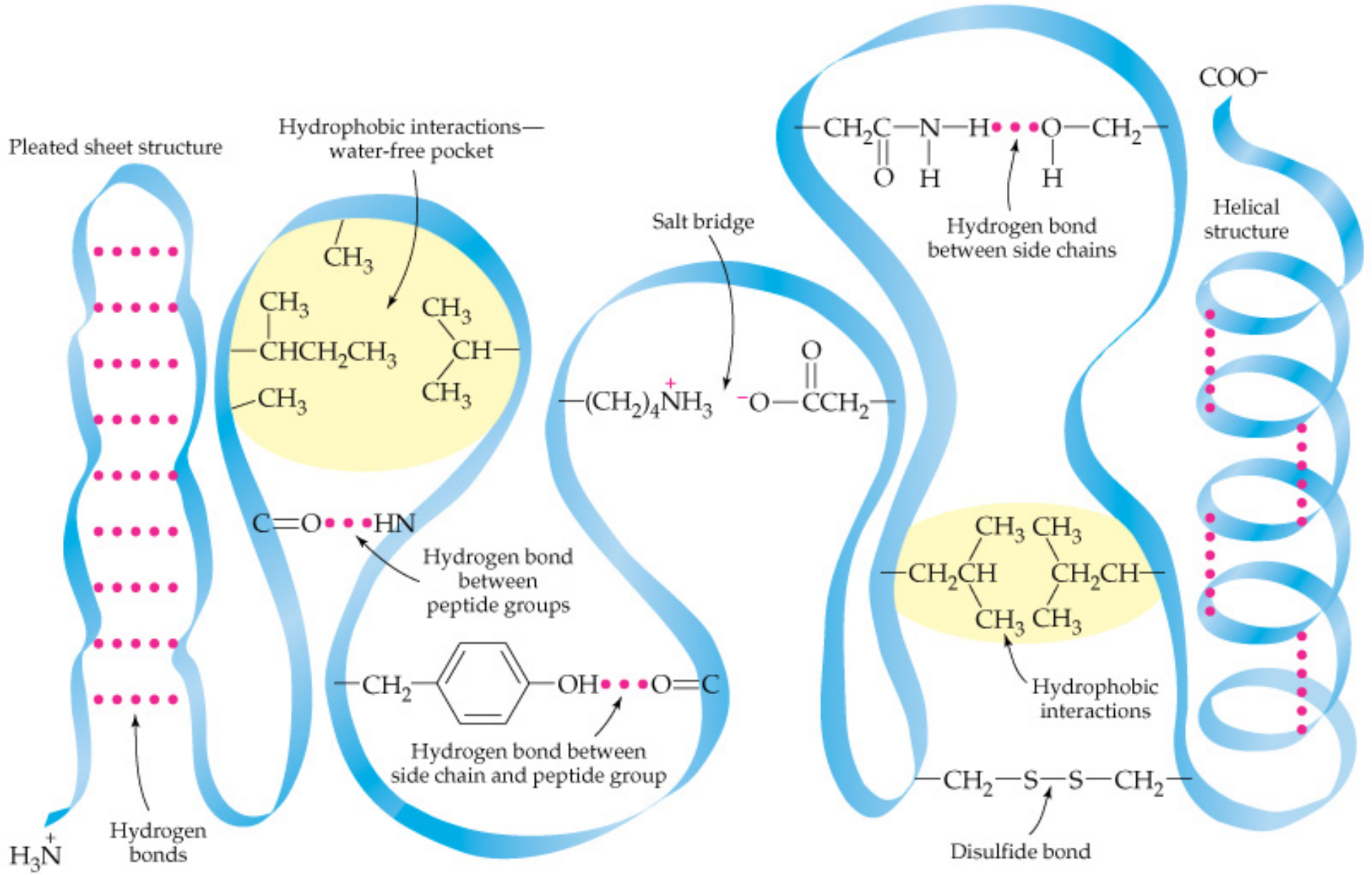
Tertiary Protein Structure

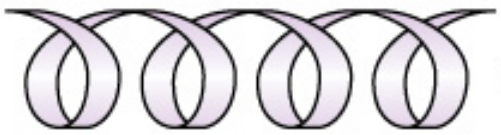
- ***Tertiary Structure of a proteins*** The overall three dimensional shape that results from the folding of a protein chain. Tertiary structure depends mainly on attractions of amino acid side chains that are far apart along the same backbone. Non-covalent interactions and disulfide covalent bonds govern tertiary structure.
- A protein with the shape in which it exist naturally in living organisms is known as a *native protein.*

Shape-Determining Interactions in Proteins

- The essential structure-function relationship for each protein depends on the polypeptide chain being held in its necessary shape by the interactions of atoms in the side chains.

- Protein shape determining interactions are summarized below:
- Hydrogen bond between neighboring backbone segments.
- Hydrogen bonds of side chains with each other or with backbone atoms.
- Ionic attractions between side chain groups or salt bridge.
- Hydrophobic interactions between side chain groups.
- Covalent sulfur-sulfur bonds.


Pleated sheet structure
Hydrophobic interactions—
water-free pocket
Salt bridge
Hydrogen bond
between side chains
Helical
structure
Hydrogen bond
between
peptide groups
Hydrogen bond between
side chain and peptide group
Hydrophobic
interactions
Hydrogen
bonds
Disulfide bond

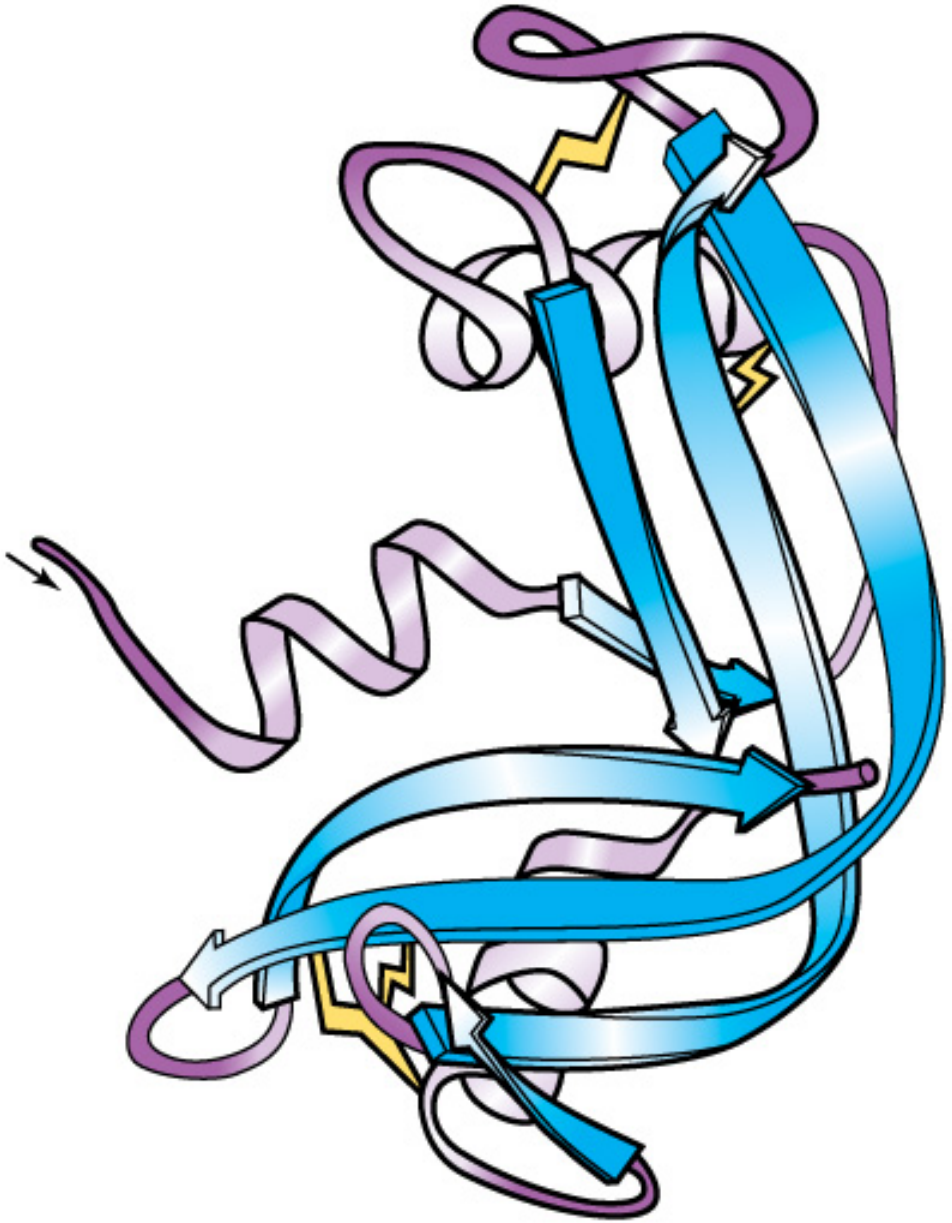


α Helix

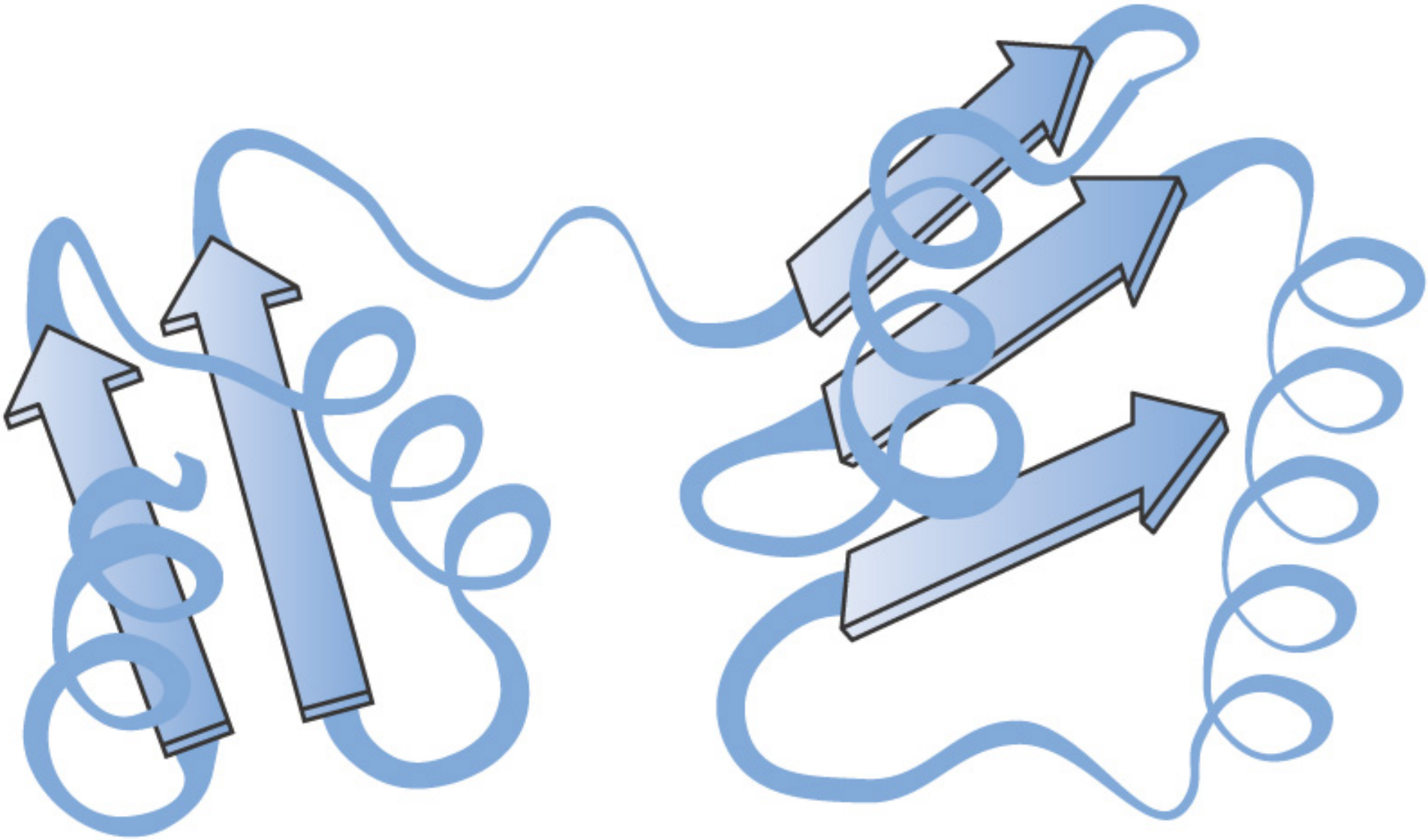
β Sheet

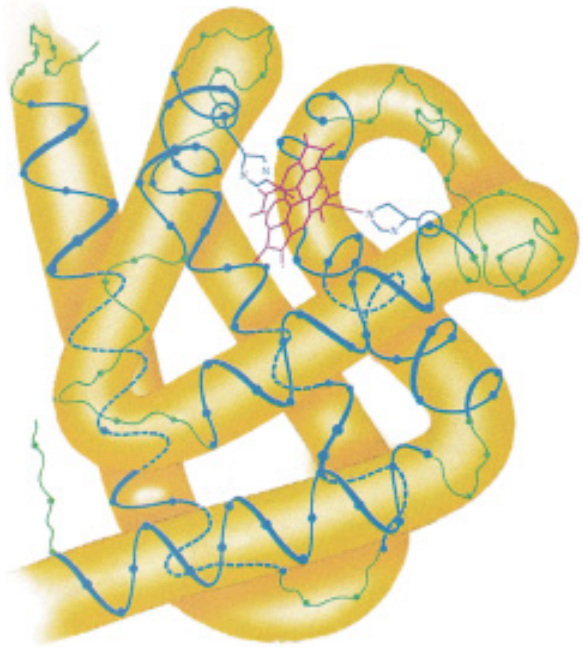
Connecting loop

—S—S— links

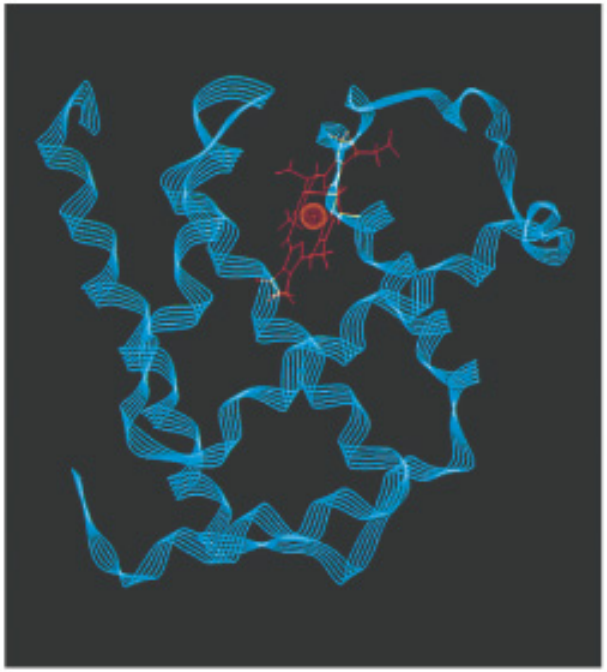


Ribonuclease

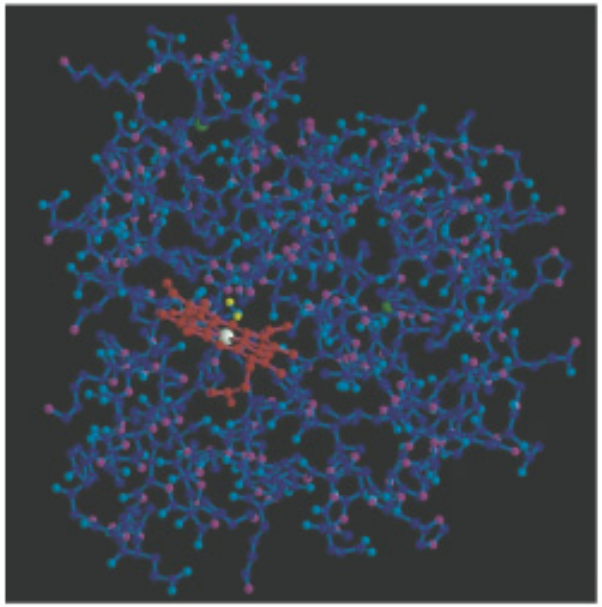




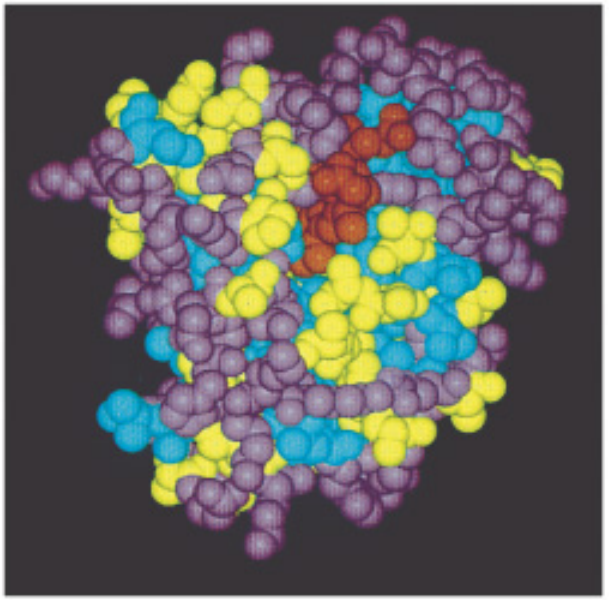
(a)



(b)



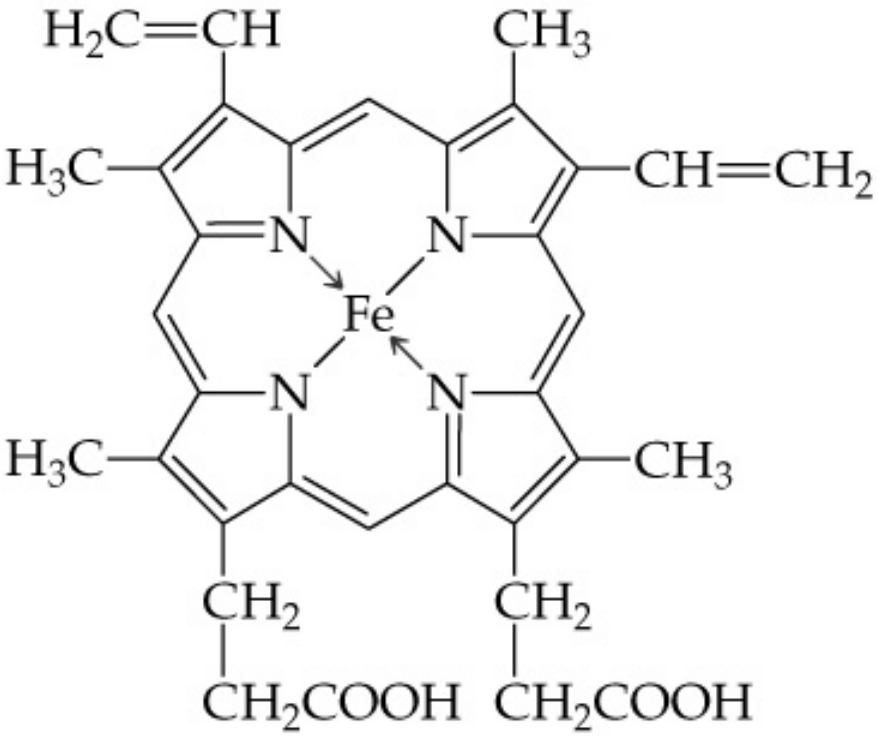
(c)



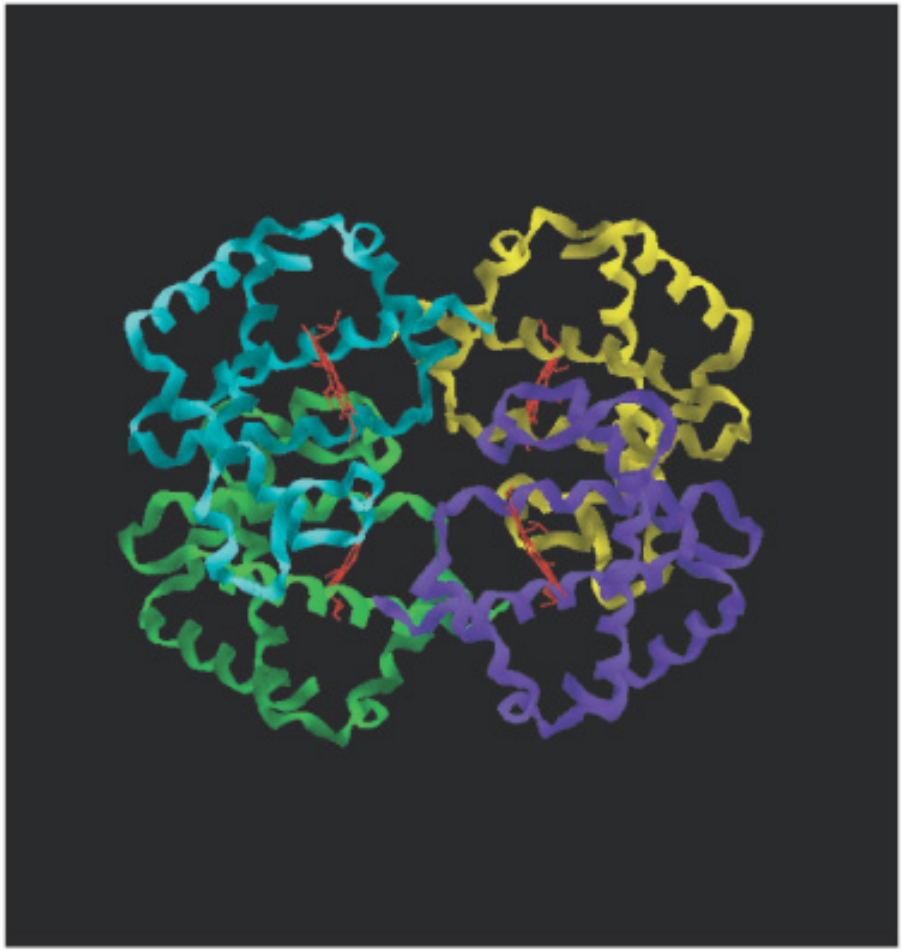
(d)

Quaternary Protein Structure

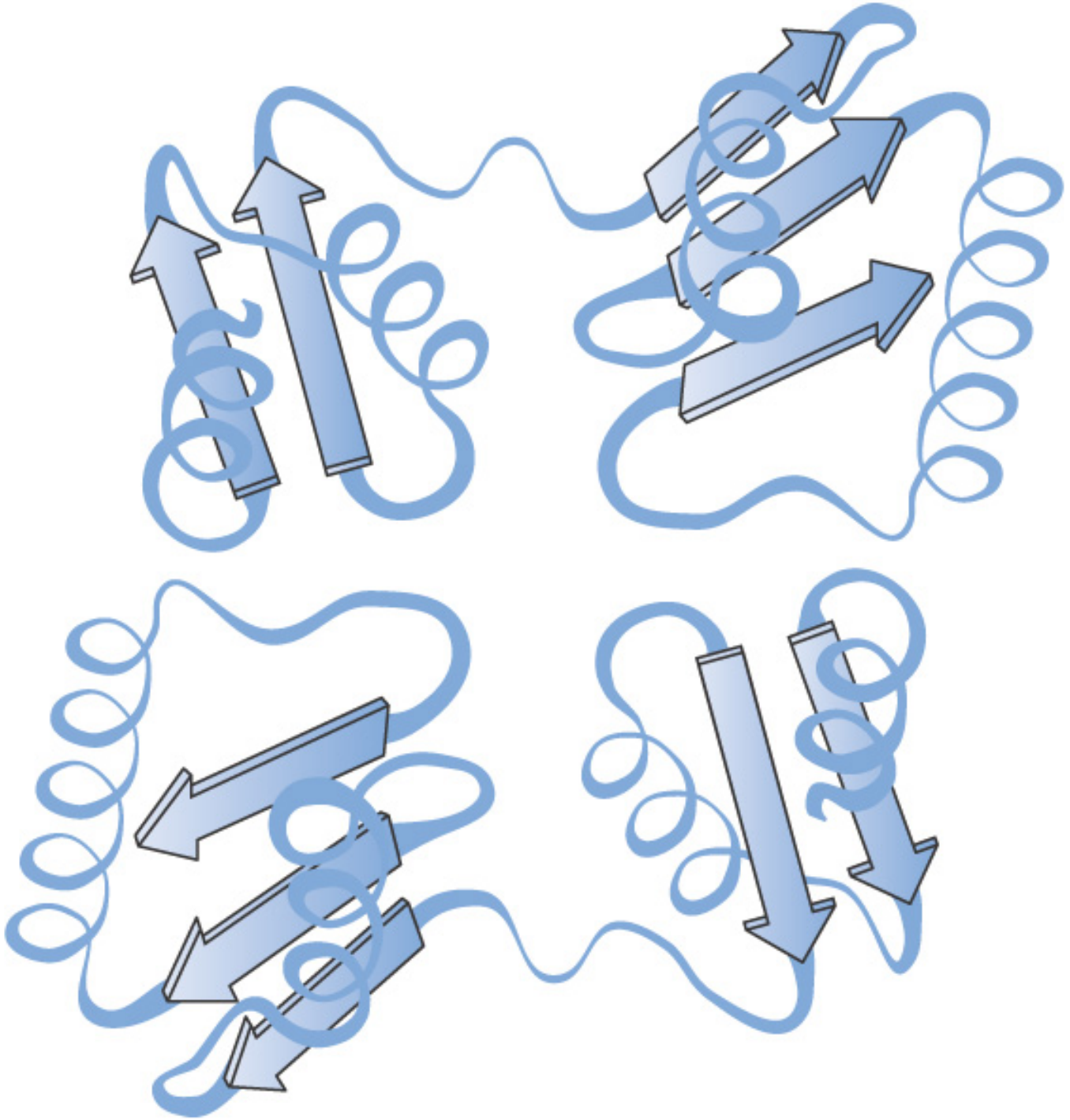
- ***Quaternary protein structure***: The way in which two or more polypeptide sub-units associate to form a single three-dimensional protein unit. Non-covalent forces are responsible for quaternary structure essential to the function of proteins.

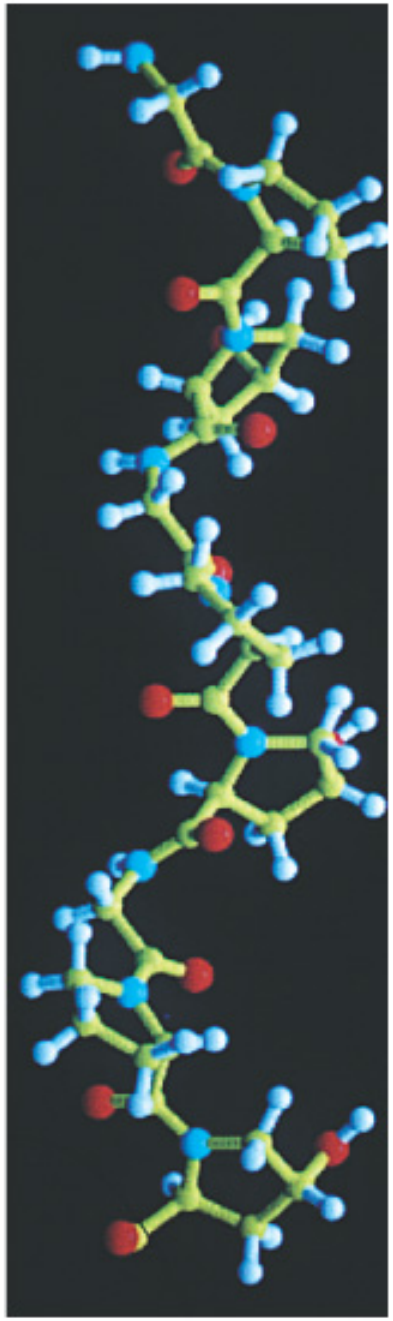


(a)

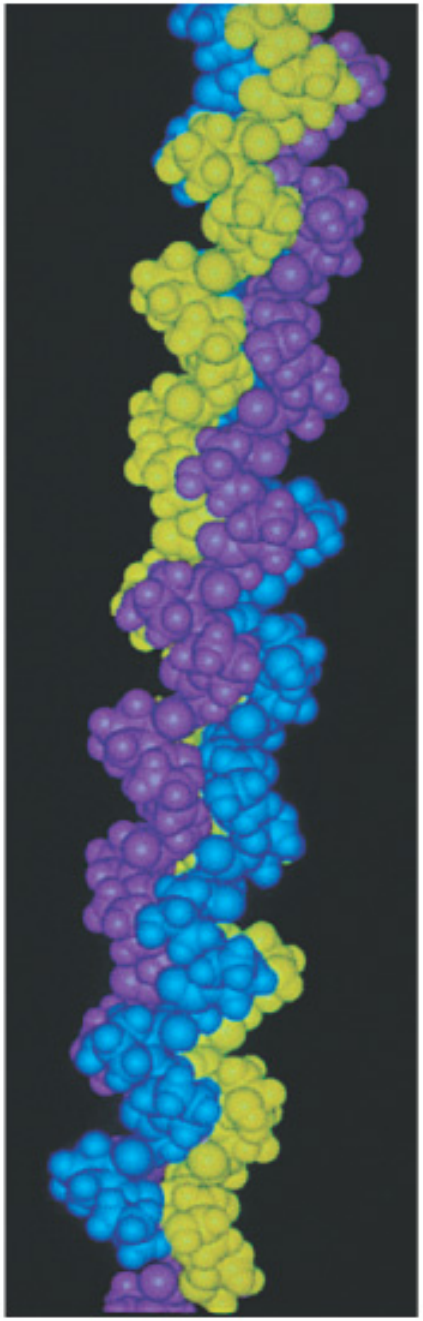


(b)

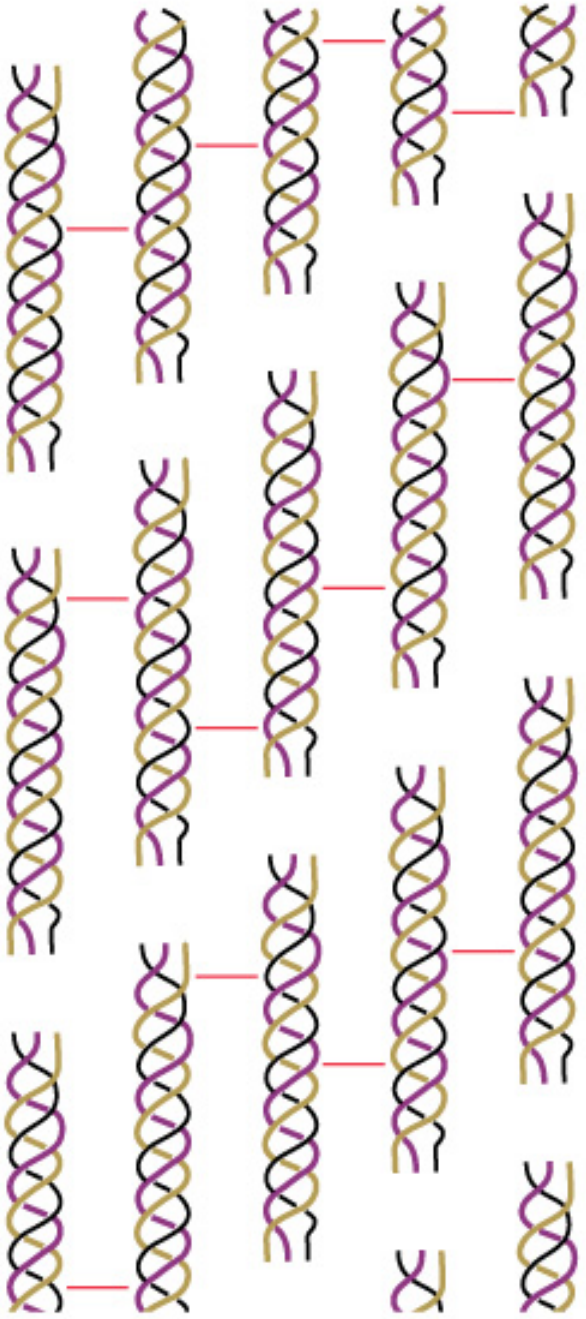




(a)

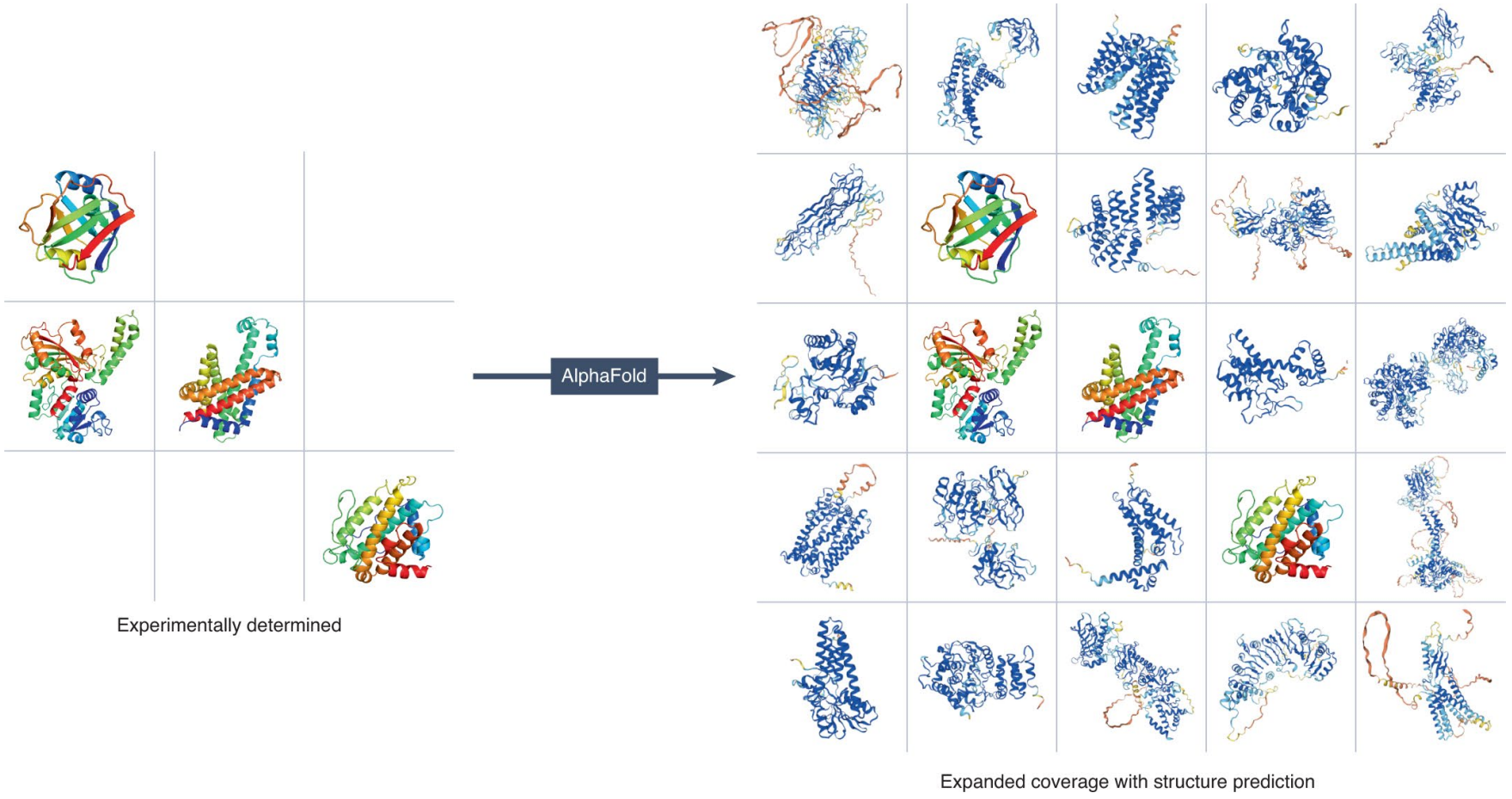


(b)

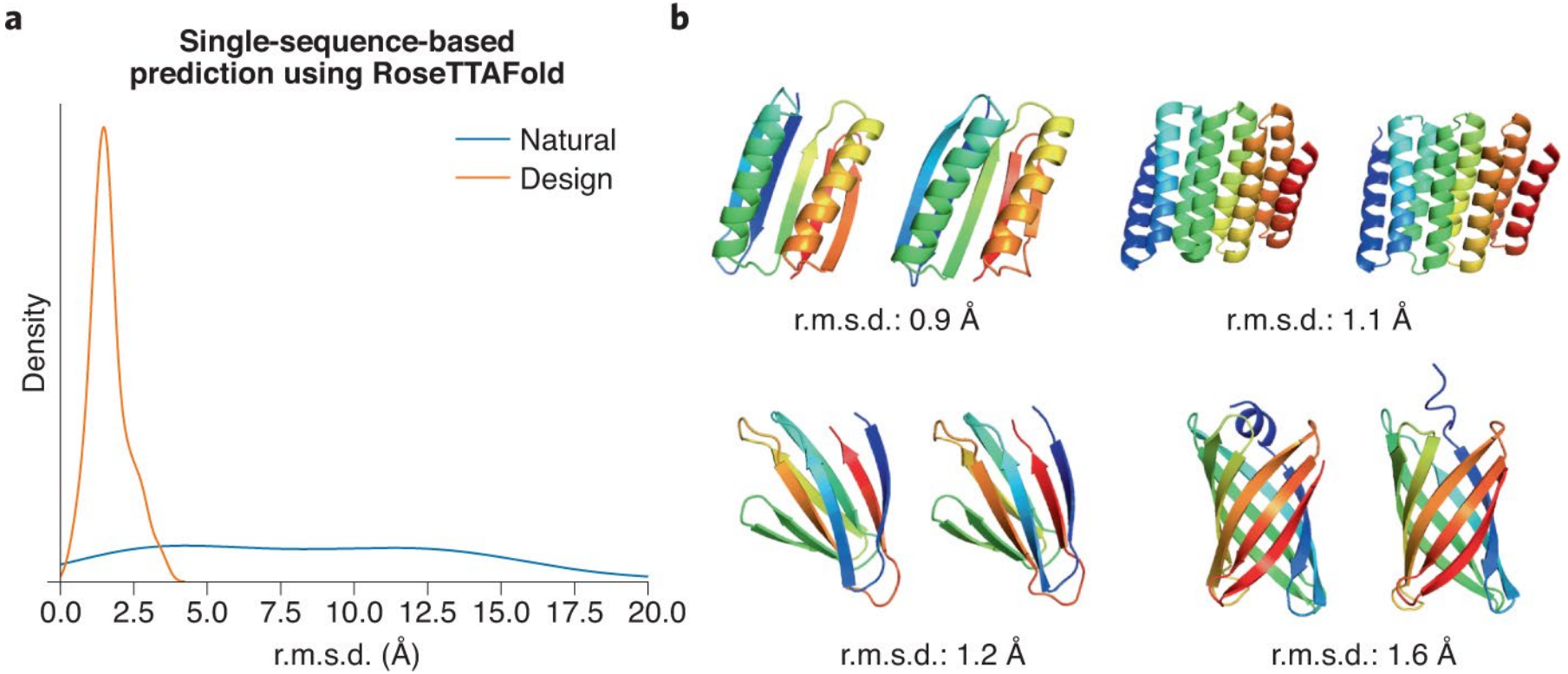


(c)

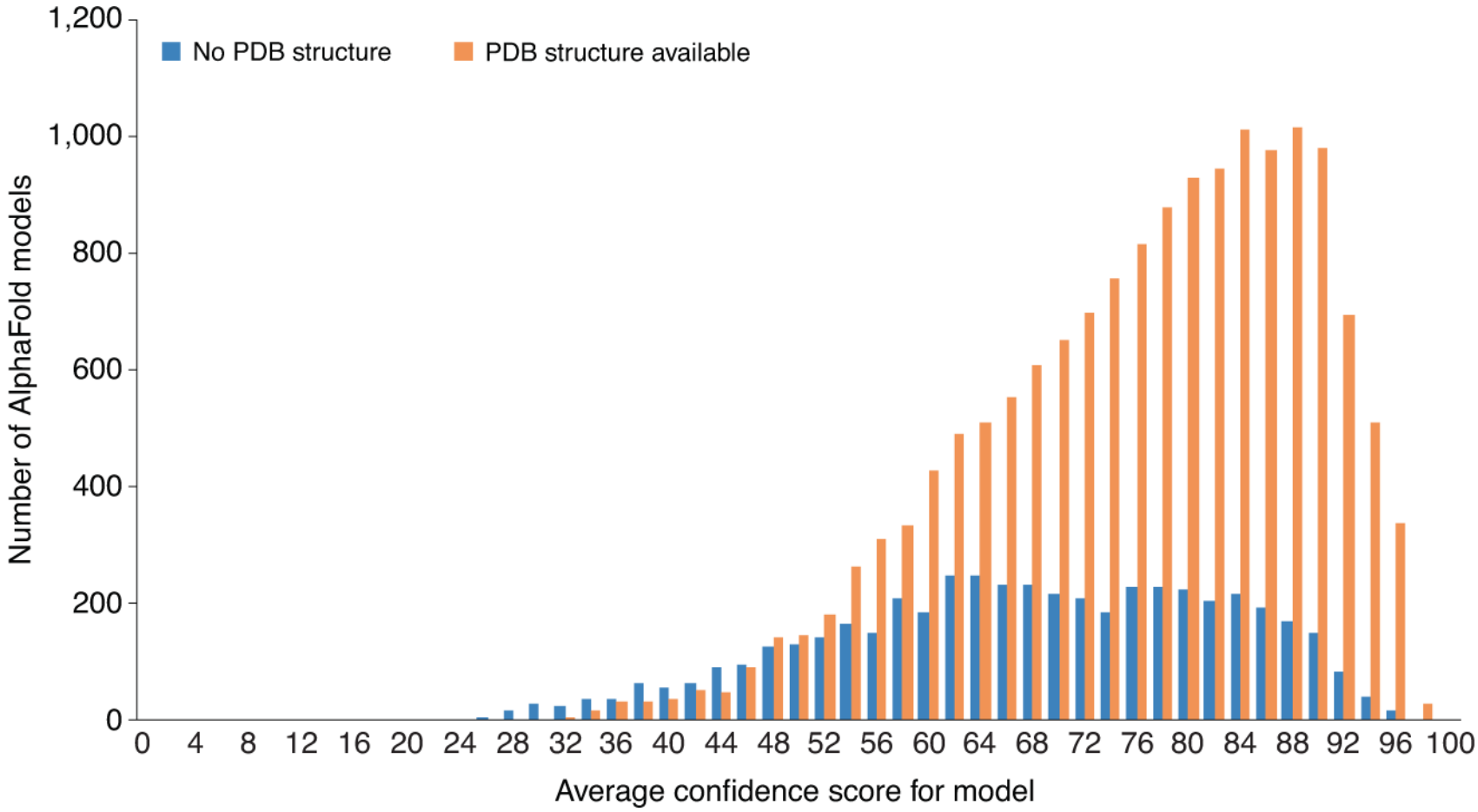
Protein structure predictions to atomic accuracy with AlphaFold



RoseTTAFold accurately predicts structures of de-novo-designed proteins from their amino acid sequences.



Distribution of average confidence scores for AlphaFold2 models of human proteins with and without homologs available in the PDB.



>200 M protein structure prediction

The number of entries at resolutions better than 6 Å released by the Electron Microscopy Data Bank per year from 2012 to 2021