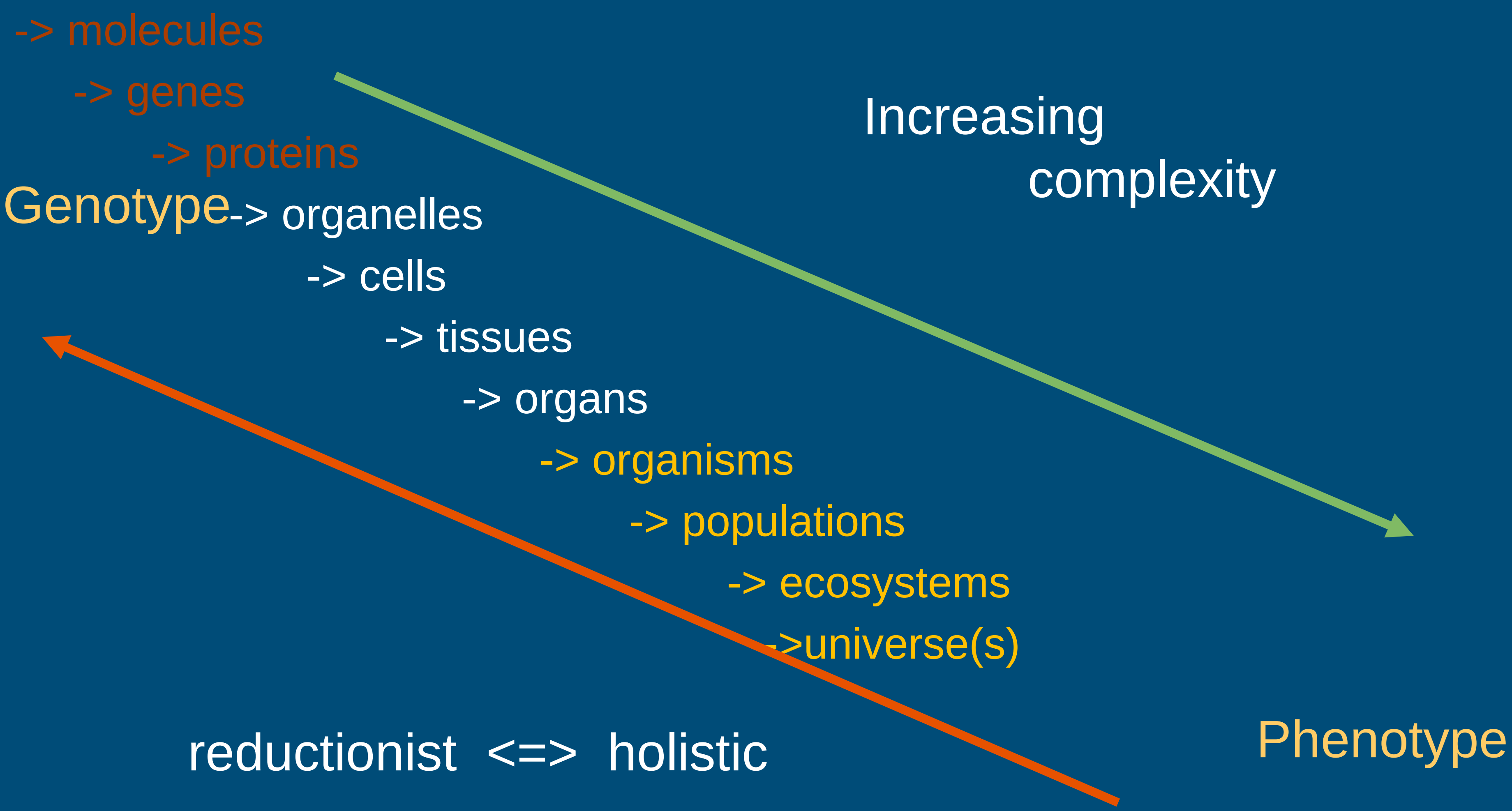
Hierarchy of biological organization



-> molecules

-> genes

-> proteins

Genotype

-> organelles

-> cells

-> tissues

-> organs

-> organisms

-> populations

-> ecosystems

->universe(s)

Increasing complexity

reductionist <=> holistic

Phenotype

What is ‘modern’ biotechnology?

- *In vitro* propagation, cell & tissue culture
  - disease-free, clean, well-defined material
- Molecular markers
  - improved selection; diagnostics
- Genetic engineering
  - recombinant DNA, transgenics; diagnostics
- Omics technologies
  - High throughput data collection; technologies
  - DNA, RNA, protein, metabolites
  - Bioinformatics, computational biology

Prokaryotes <-> Eukaryotes

- (almost) all organisms contain DNA, RNA and protein
- different ways of storing DNA
  - Nucleus in eukaryotes
- complexity versus efficiency ?
  - prokaryotes (*Escherichia coli*):

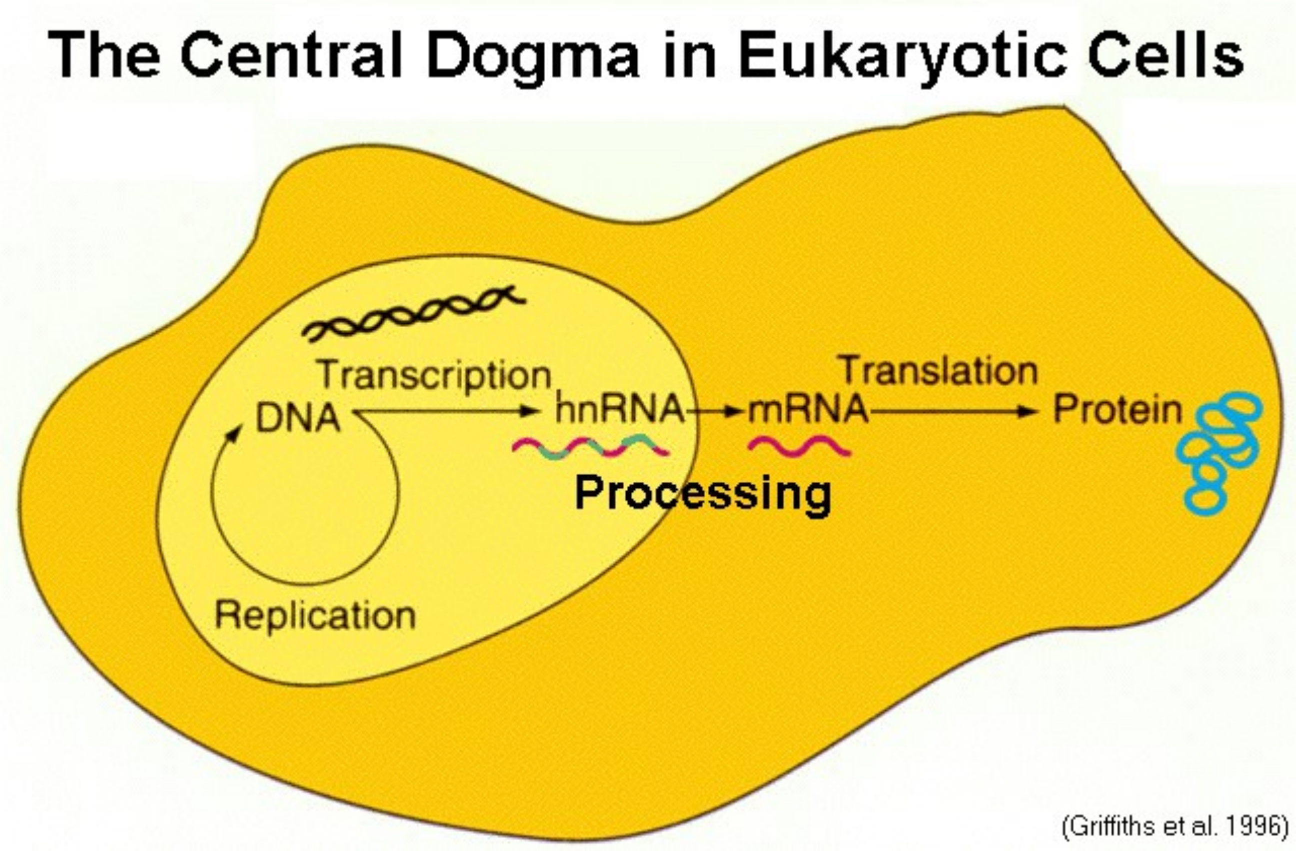
    small, unicellular, efficient
  - eukaryotes (plant, human):

    large, multicellular, subcellular, complex

Differentiation: a conceptual issue

- all cells of an organism contain the same DNA,
- (yet, not all that DNA is identical)
- yet, not all cells **use** the same DNA
- therefore, not all cells look the same
  - differential use of the same genetic information gives different results
  - how is the differential usage organized?
- disease is often caused by errors in or misuse of the genetic material

Biological information transfer



(Griffiths et al. 1996)

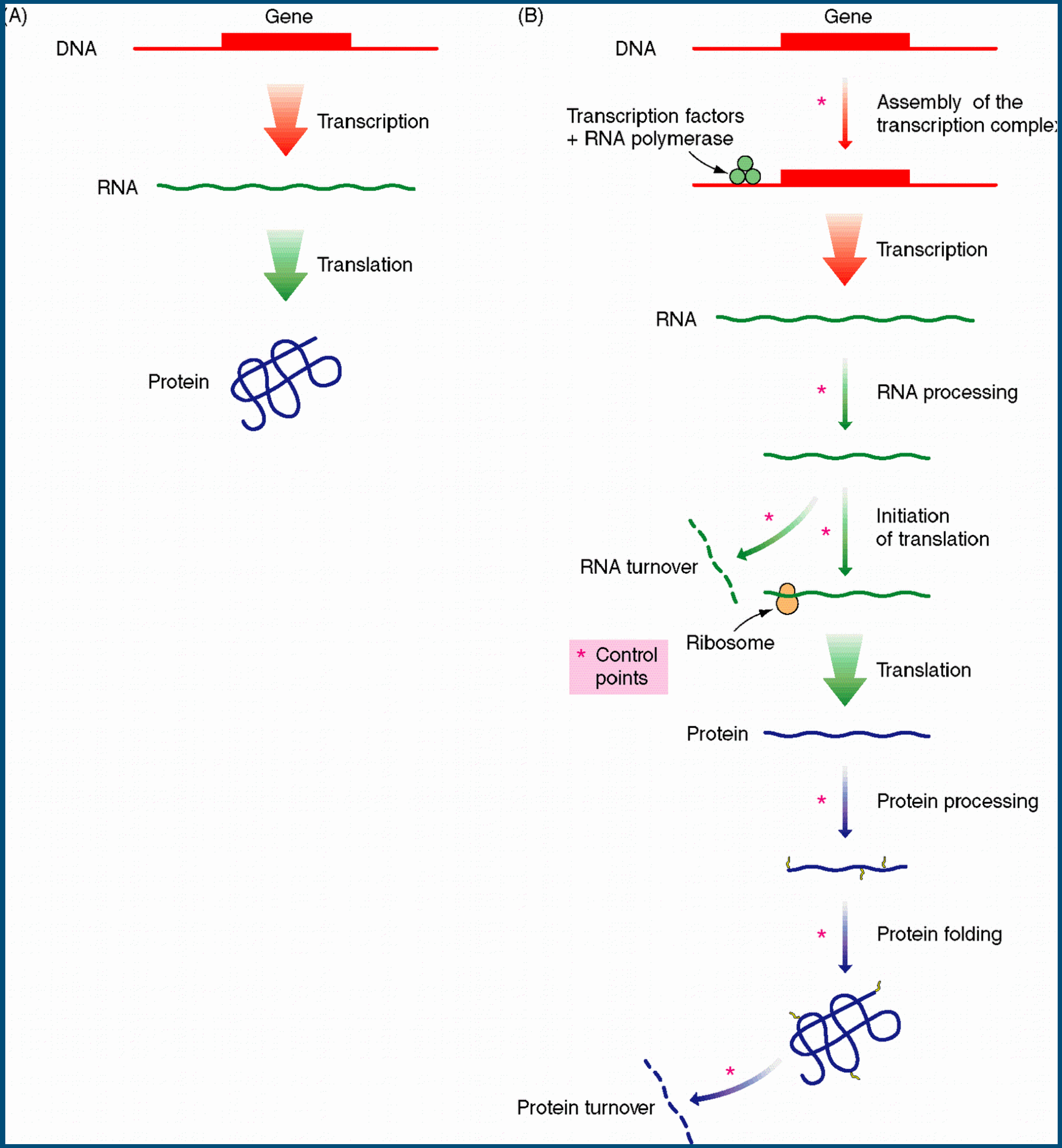
DNA (makes DNA) makes RNA makes protein

(makes metabolites makes action makes phenotype)

DNA makes RNA makes protein:

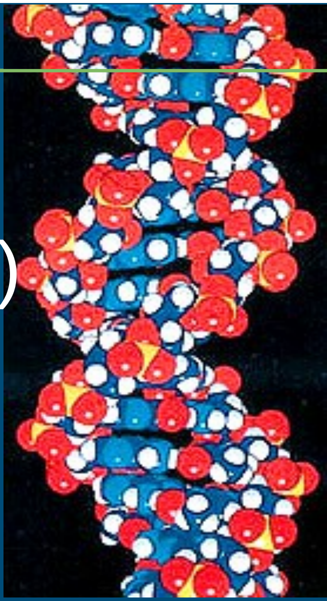
- DNA makes DNA: replication
- DNA makes RNA: transcription
- RNA makes protein: translation
- (protein makes action) ~ enzyme activity
- (genomics/biotechnology/bioinformatics/omics: action makes money)
    - DNA = cooking book, RNA = recipe, protein = dish
    - DNA = chief, RNA = middle management, protein = workforce
    - DNA = hardware, RNA = software, protein = working program)

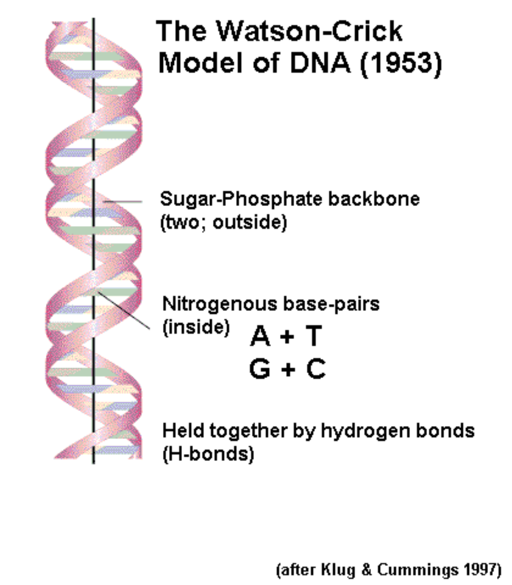
Central dogma is of course (much) more complex

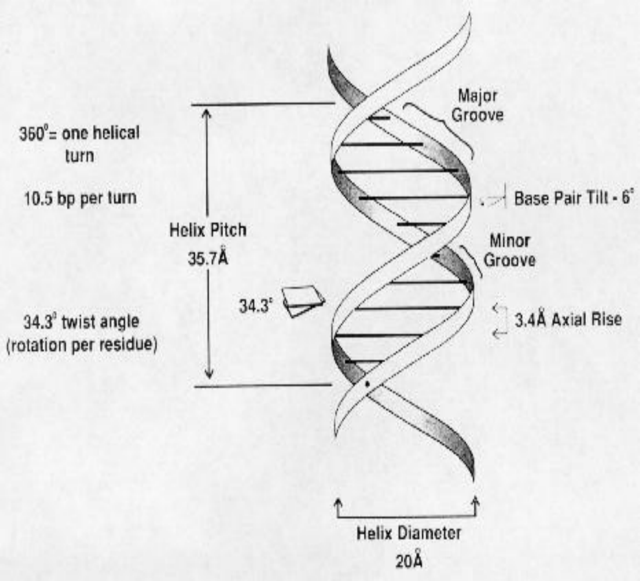


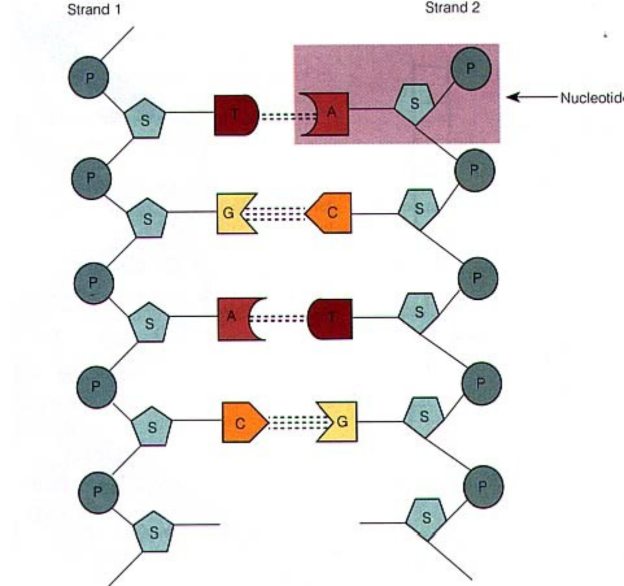
DNA

- basically a salt (compare sodium salt)
- a linear polymer of 4 nucleotides (A,C,G,T)
  - also called bases
- configuration of a double helix
- antiparallel strands (5’ -> 3’)
- occurs in nucleus tightly packed with proteins

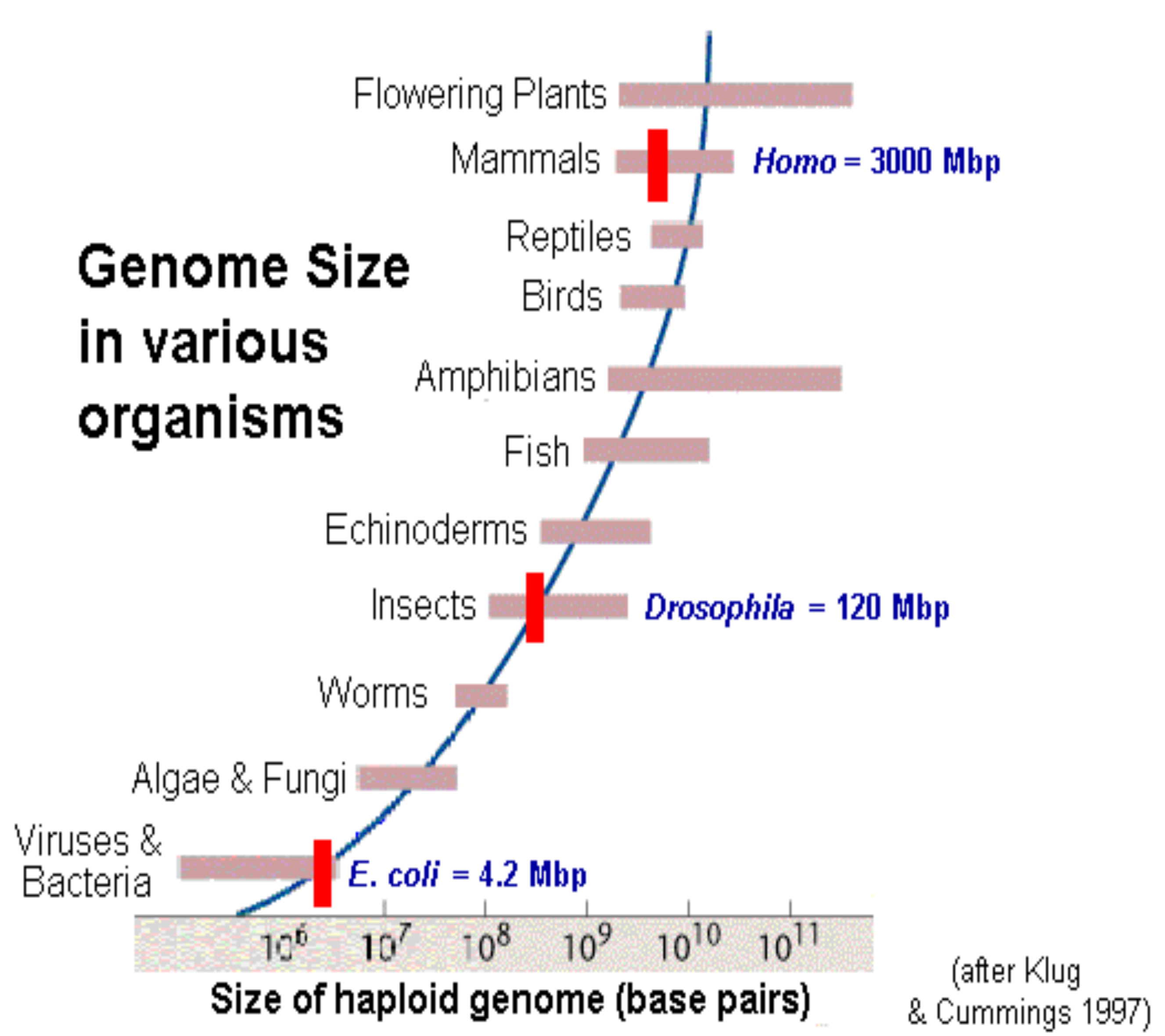



The Watson-Crick Model of DNA (1953)
Sugar-Phosphate backbone (two; outside)
Nitrogenous base-pairs (inside)
A + T
G + C
Held together by hydrogen bonds (H-bonds)
(after Klug & Cummings 1997)



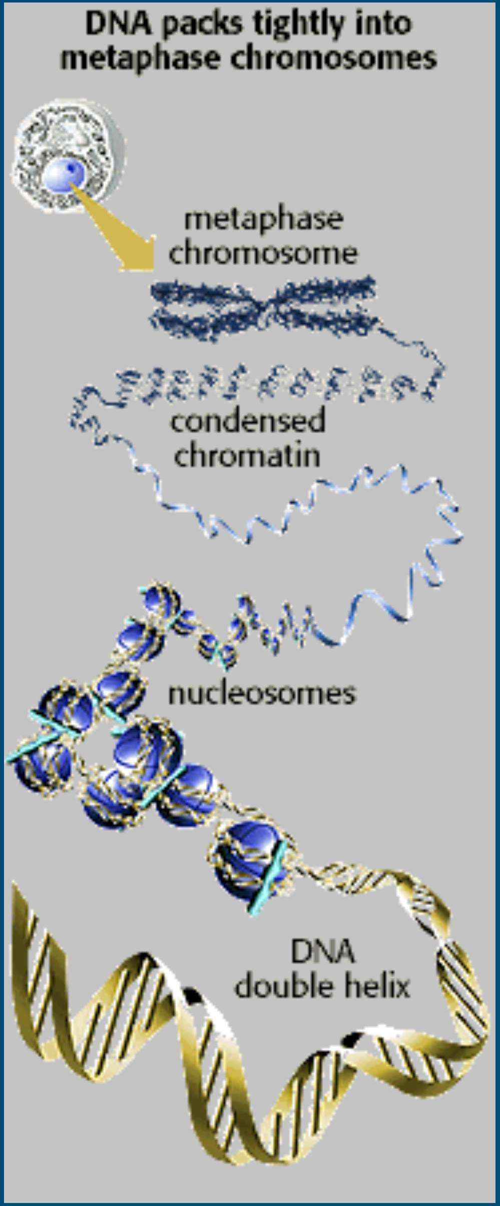


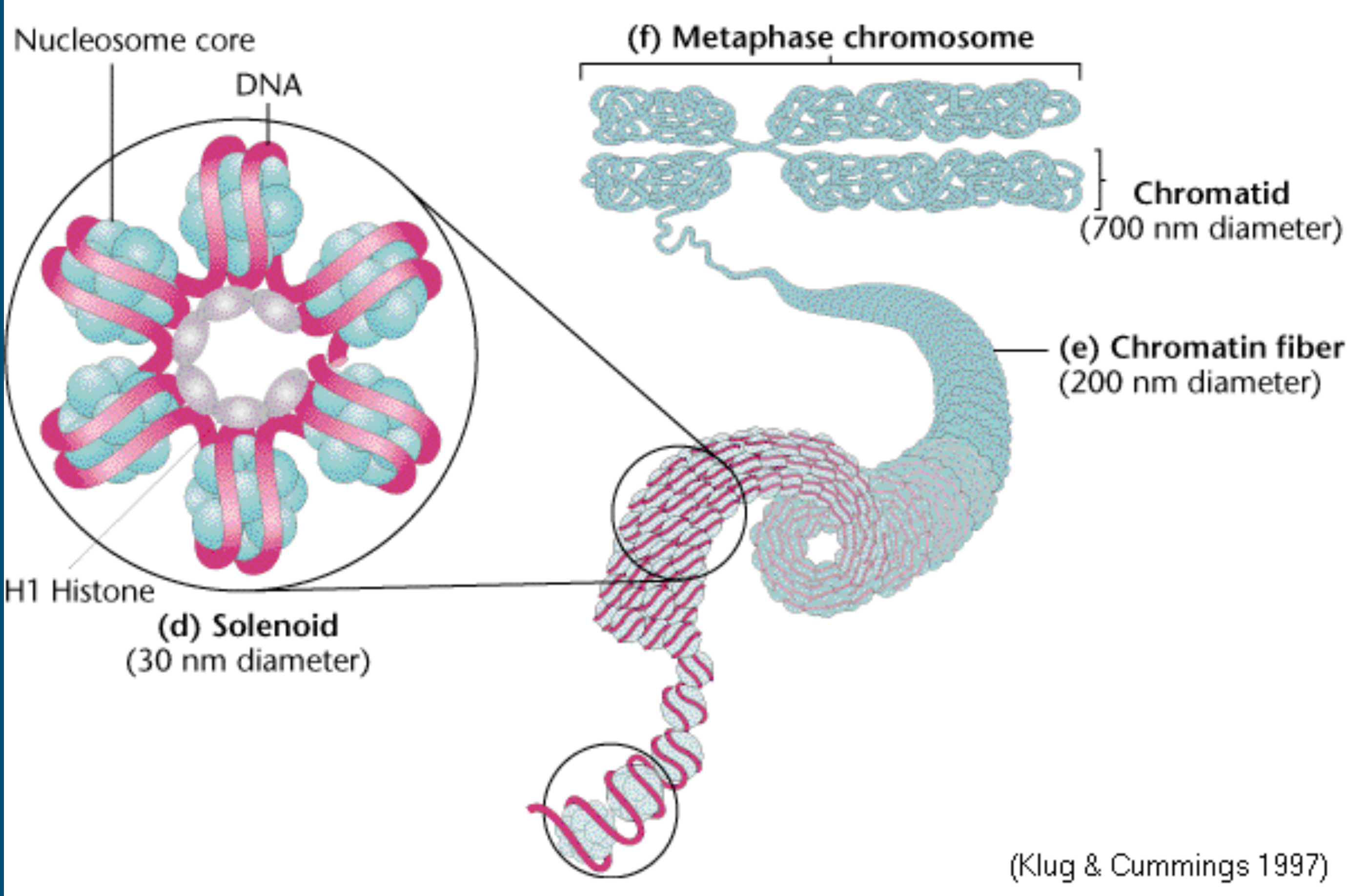
Genome sizes



(after Klug
& Cummings 1997)

DNA condensation

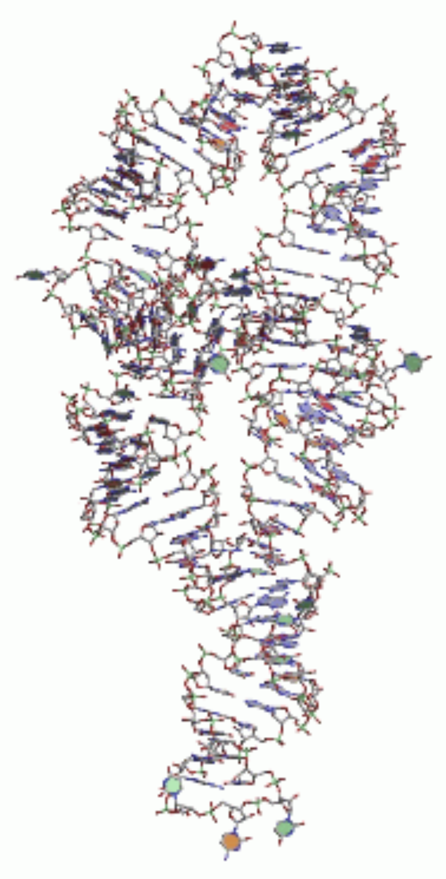




(Klug & Cummings 1997)

RNA

- basically a salt (compare sodium salt)
- ribose in stead of deoxyribose
- a linear polymer of 4 nucleotides (A,C,G,U)
- single stranded (no double helix)
- various intermolecular structures possible
- different forms with different functions
- thought to be the “origin”

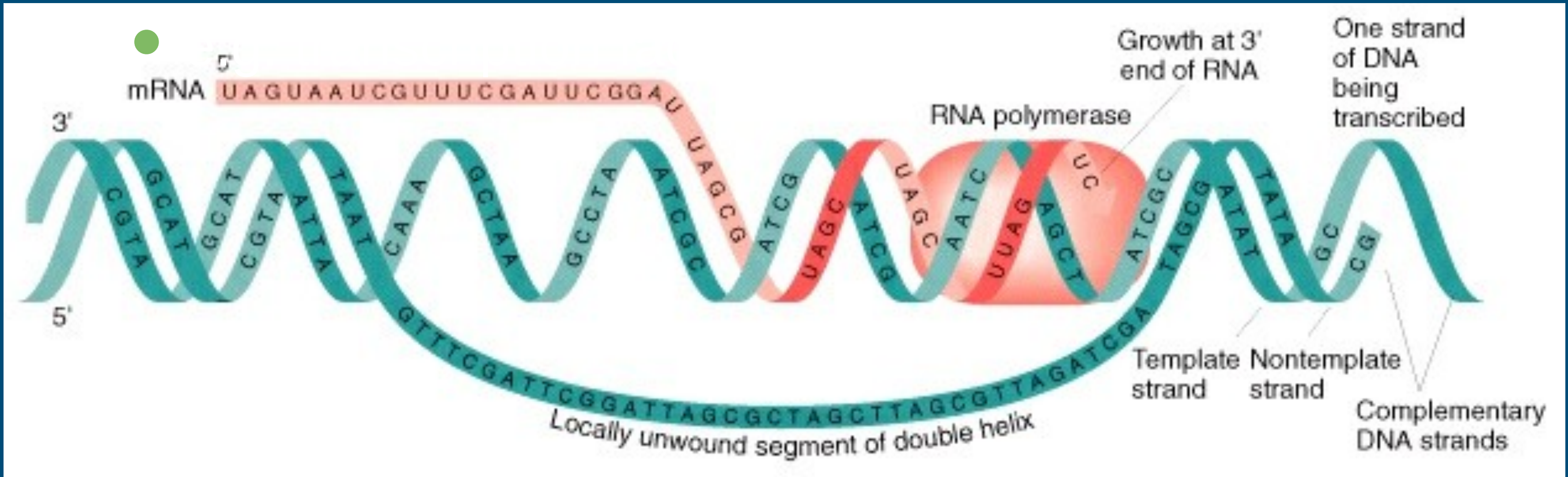


RNA types

- mRNA: messenger RNA
  - gives protein
- rRNA: ribosomal RNA
  - participates in making protein
- tRNA: transfer RNA
  - participates in making protein
- sn/scRNA: small nuclear/small cytoplasmic RNA
  - presumed regulatory functions
  - microRNA; siRNA

DNA transcription: DNA makes RNA

- by RNA polymerase
- in nucleus
- in 5’ -> 3’ direction only
- on DNA template
- requires start and stop signals in template
- involves numerous other factors and proteins



DNA Transcription

The **messenger RNA transcript** is equivalent to the sense strand of the DNA

5' - **G T A A T C C T C** - 3' sense (coding) strand
3' - **C A T T A G G A G** - 5' antisense (template) strand
ppp 5'- **G U A A U C C U C** - 3'OH messenger RNA

=> Direction of transcription =>

DNA transcription

- start signal: promoter
  - binds RNA polymerase and transcription factors
  - determines transcriptional regulation:

    is the RNA made,how much is made, where is it made?
- stop signal: termination of transcription

In eukaryotes:

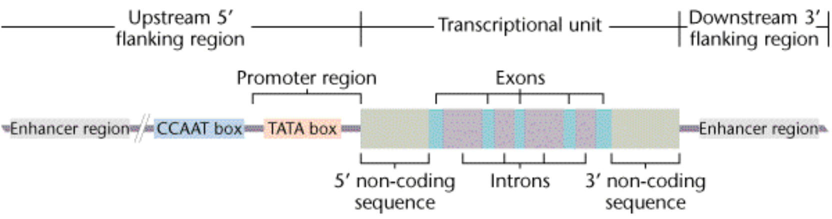
- transcribed RNA is further modified
  - 5’ cap, poly-A tail
  - Splicing phenomena

RNA bioinformatics

- EST databases
  - expressed sequence tag = sequenced cDNA (=mRNA)
  - Deep sequencing
- Expression databases
  - Microarray, MPSS, other
- Non-coding RNA databases
  - microRNA, 16S RNA, a.o.
- Splicing databases
  - Alternative splicing

Genes of eukaryotes are split

- DNA not co-linear with mature RNA, but longer
  - primary transcript is ~ as long as the DNA
- RNA undergoes further modification
  - in which parts of the RNA are removed
- modification is called: splicing
  - the removed parts are called ‘introns’ or ‘intervening sequences’
  - introns may have a function
- intron splice sites are conserved
- various mechanisms exist for splicing



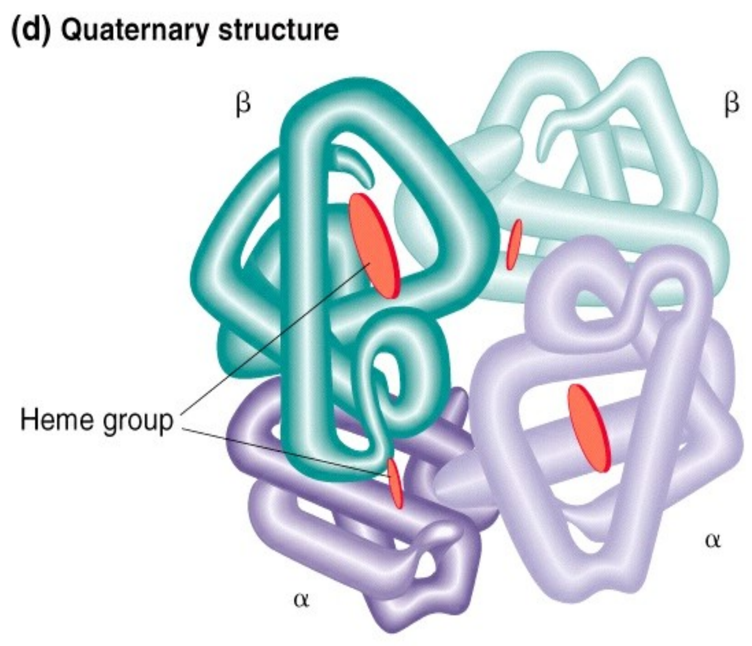
(Klug & Cummings 1997)

Epigenetics

- Code ‘on top’ of the DNA code
- DNA methylation
- Histone modification: “Histone code”
- Role being eludicated
    - Differences between cells
    - Communication with the environment
    - Disease development
- “Lamarck’s last laugh?”
    - evolution

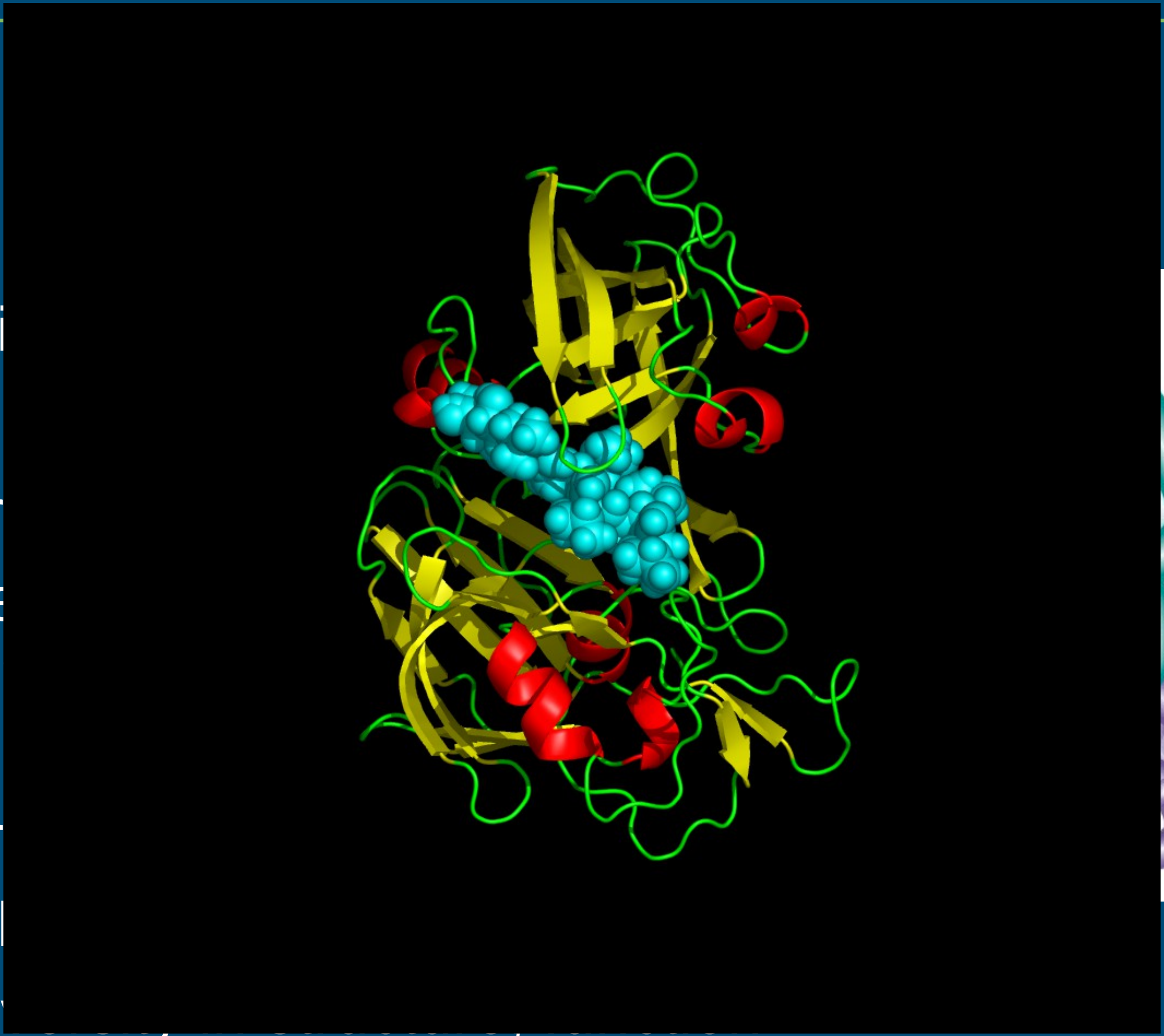
Protein

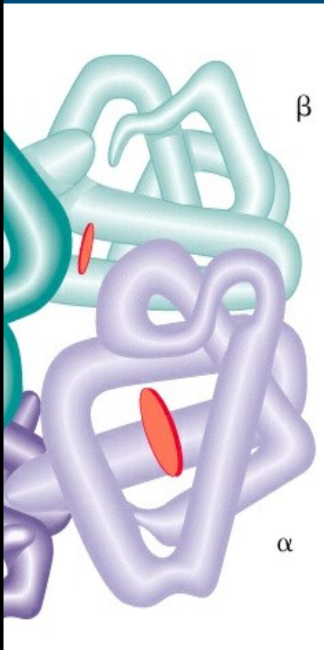
- a linear polymer
- made up of amino acids (> 20 different ones)
  - peptide bond
- various secondary and tertiary structures
  - protein folding largely determines activity
- often multimeric: quaternary structure
- many chemical modifications possible
- large diversity in structure, function and chemical characteristics


(d) Quaternary structure

Protein

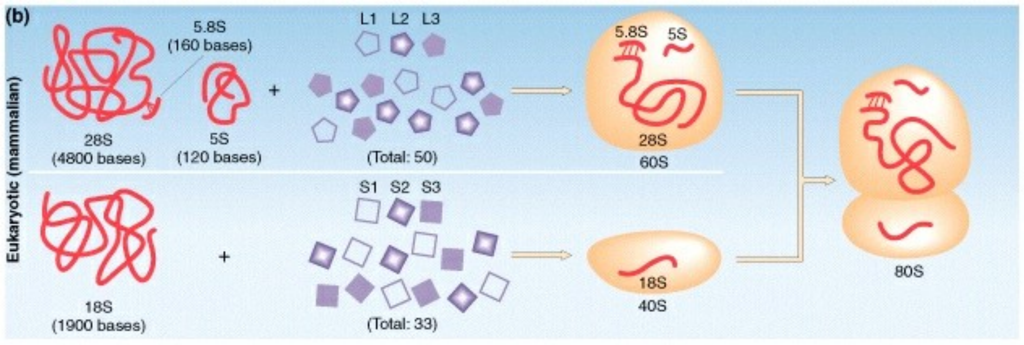
- a linear
- made u
  - pepti
- various structur
  - prote dete
- often m structur
- many cl
- large di and chemical characteristics

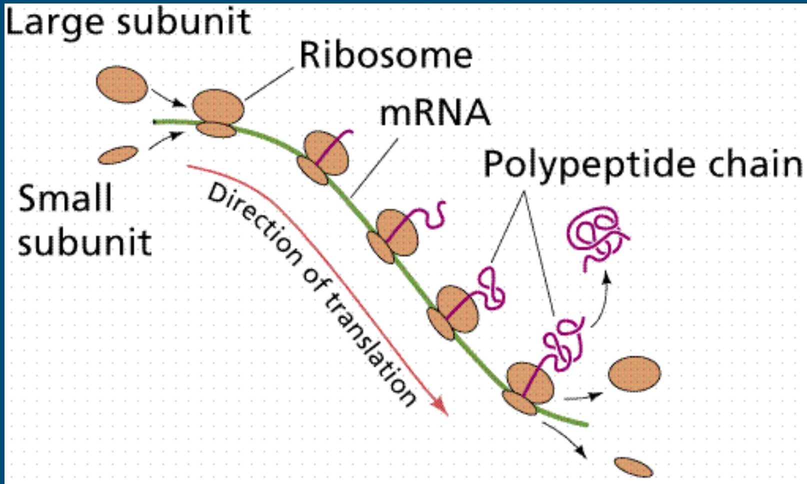




RNA translation: RNA makes protein

- in cytoplasm
- joint action of rRNA, tRNA and many protein factors
  - rRNA + proteins forms ribosome
  - tRNA + amino acid generates the encoded protein sequence
- ribosome moves along the mRNA
  - recognizes triplets: genetic code
  - recruits tRNA
  - starts/checks/corrects/ terminates process

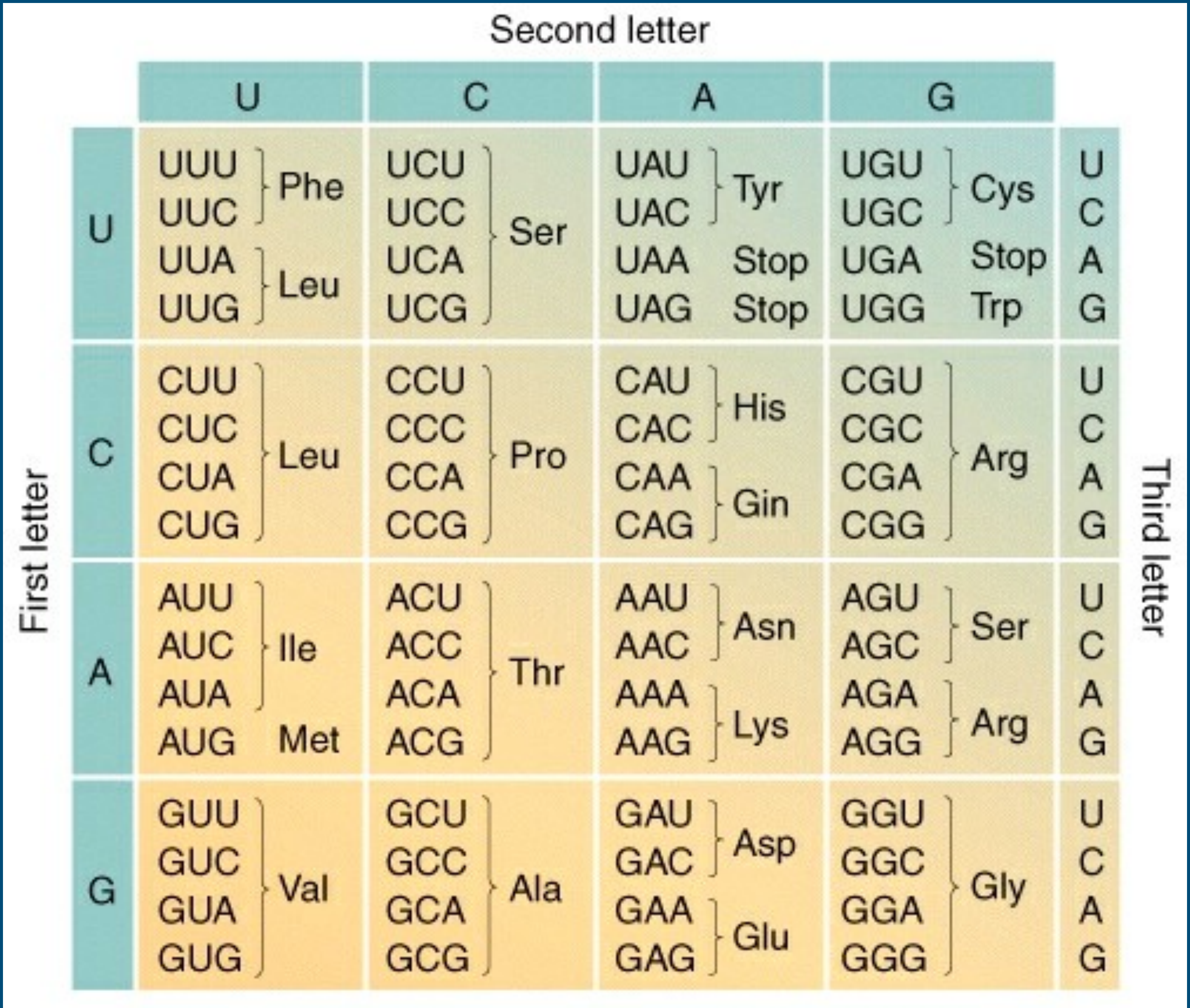




Genetic code

- **three** RNA bases determine **one** amino acid
- is universal
  - With small exceptions
- is degenerate (64 codons for 20 amino acids)
- specific triplets for start and stop of translation
- codon usage differs between organisms and organelles: Choose your codon table wisely
  - ORF finding in prokaryotes
  - Gene analysis of plant chloroplast DNA
  - Mammalian genes

Genetic code



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

Errors in translation

- THE BIG CAT ATE THE FAT RAT

single deletion (frame shift):

- THE BIG ATA TET HEF ATR AT

single deletion plus addition

- THE BIG ATA ATE THE FAT RAT

New ‘sentence’ is gibberish or a changed protein

Protein processing

- proteolytic cleavage
- chemical modification
  - acetylation, methylation, hydroxylation, glycosylation etc.
- active transport to place of work
  - to ER etc: secretion pathway by extra signal peptide
  - to organelles: complex signal peptides
  - to nucleus: nuclear targetting
- turnover (synthesis <-> breakdown)

Restriction endonucleases (enzymes)

- mostly of microbial origin; protective role in microbes (against viruses)
- recognize and digest a specific sequence in DNA
- recognition sequence is usually palindromic
- often results in staggered DNA ends
- with DNA ligase important for all recombinant DNA applications
- many, many are now commercially available

Protein bioinformatics

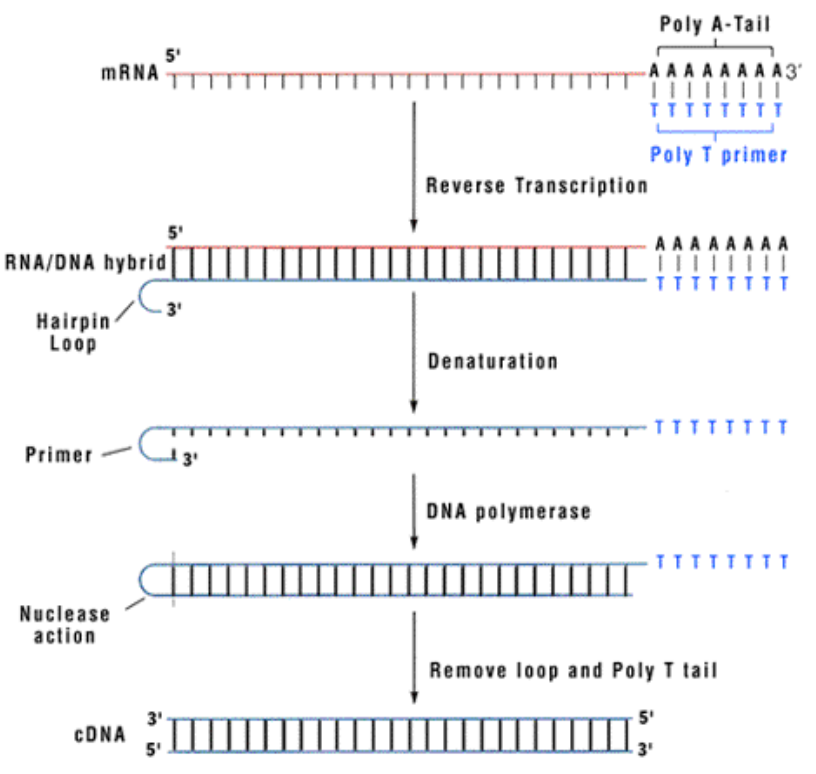
- proteome databases
  - all proteins encoded by a genome
- codon usage databases
- structural databases
  - structure/structure function
- interaction databases
  - partners in action
  - networks
- integrated databases
- many, many more

Metabolites

- Enormous chemical diversity of metabolites
  - Notably in plants
- Methodology being developed
  - Targeted versus non targeted approaches
  - Expensive equipment
- Metabolomics
  - All metabolites of an organism
  - Relationship metabolites and phenotype
    - Resistance
    - Health
    - Value compounds
    - Etc. etc.

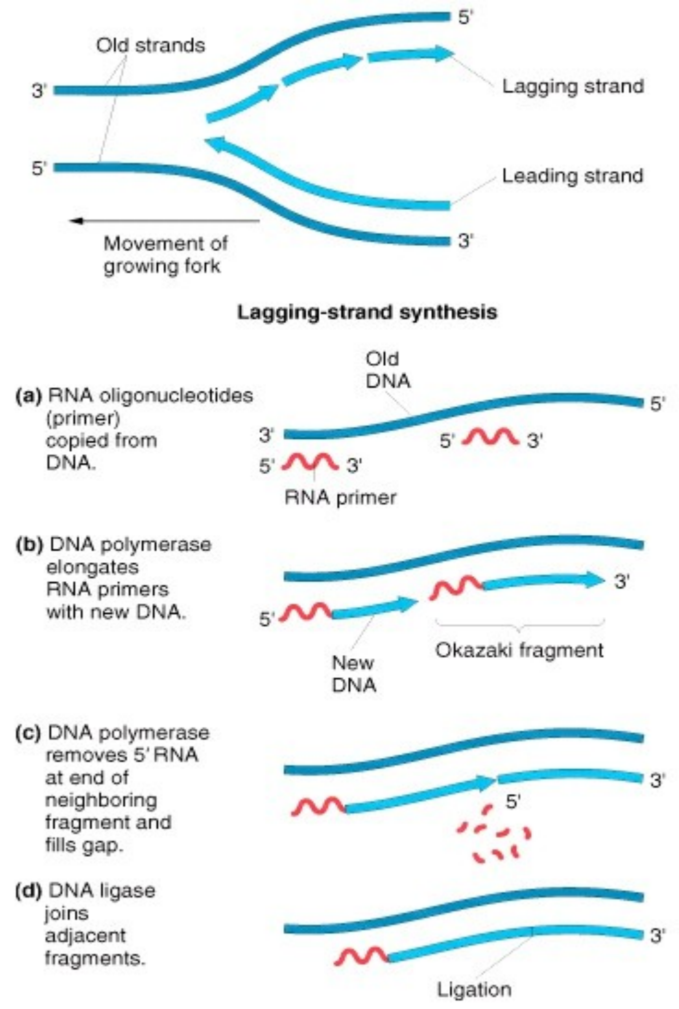
Reversed transcription: RNA makes DNA

- strategy of RNA viruses
- by reversed transcriptase
- in 5’ -> 3’ direction only
- on RNA template
- requires primer
- reversed transcriptase is used *in vitro* for converting mRNA to its DNA form (cDNA)



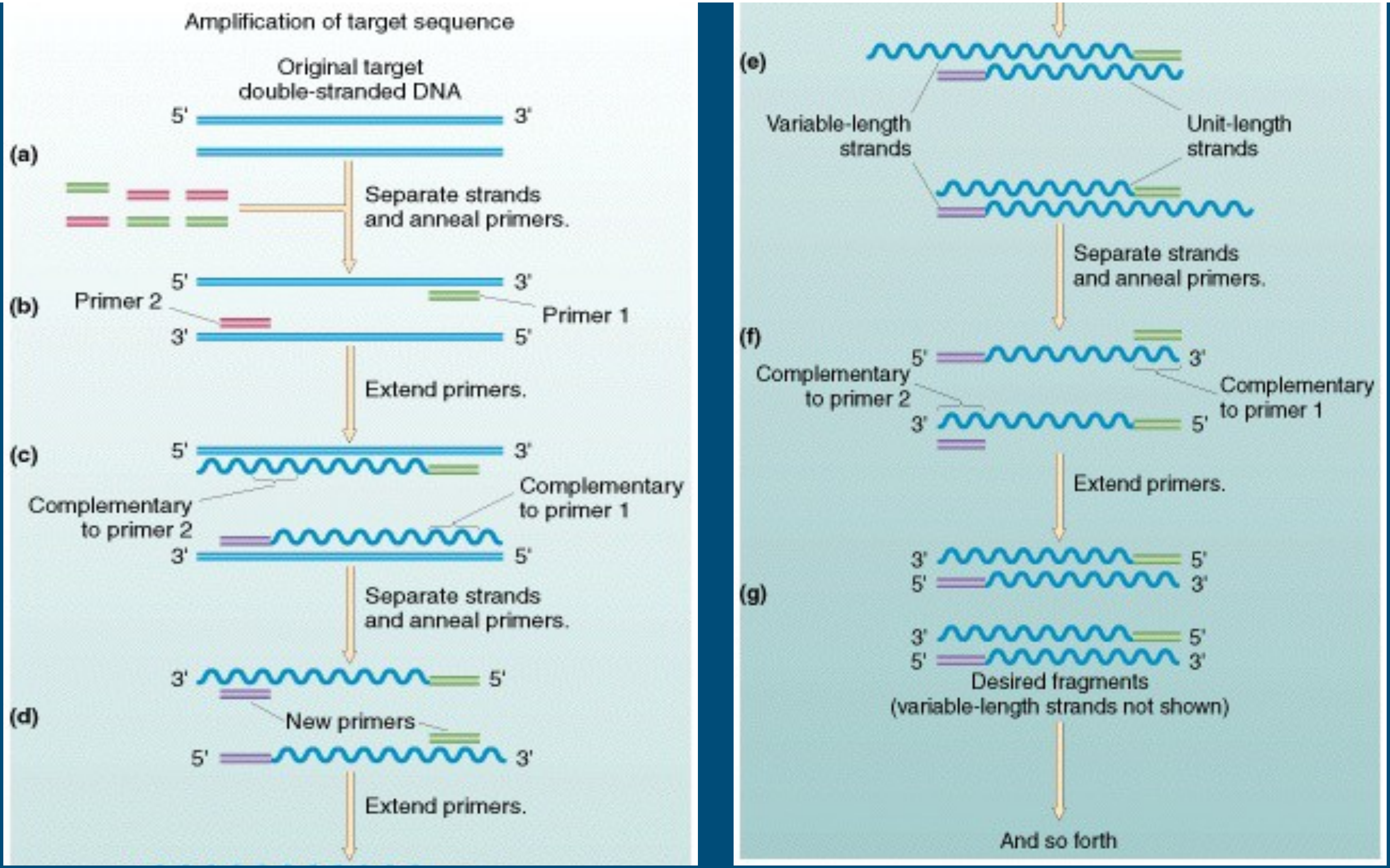
DNA replication: DNA makes DNA

- by DNA polymerase
- in 5’ -> 3’ direction only
- in nucleus
- requires beginning (= primer)
- involves numerous other factors and proteins
  - e.g. DNA ligase
- Applications:
  - PCR
  - DNA sequencing
  - cloning

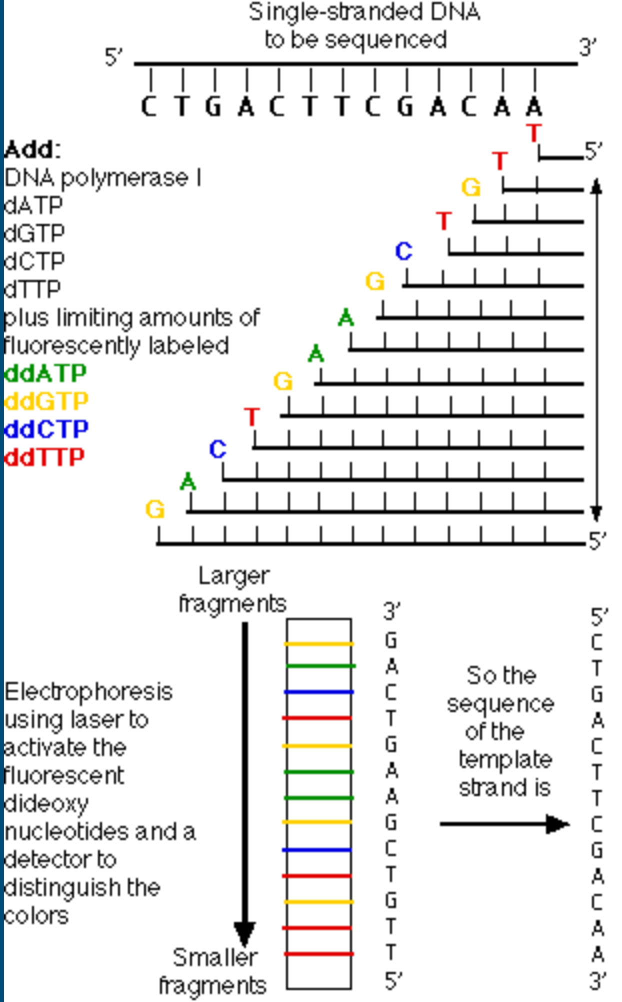

Old strands
5'
3'
Lagging strand
5'
Leading strand
3'
Movement of growing fork
Lagging-strand synthesis
Old DNA
(a) RNA oligonucleotides (primer) copied from DNA.
3'
5'
5'
3'
5'
3'
RNA primer
(b) DNA polymerase elongates RNA primers with new DNA.
3'
5'
New DNA
Okazaki fragment
(c) DNA polymerase removes 5' RNA at end of neighboring fragment and fills gap.
3'
5'
(d) DNA ligase joins adjacent fragments.
3'
Ligation

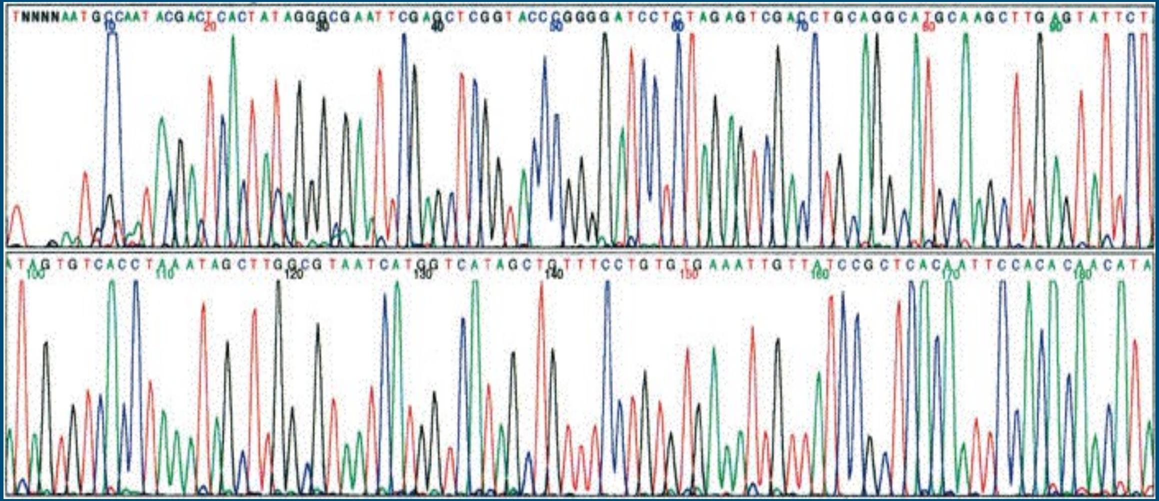
Polymerase Chain Reaction (PCR) (1992)

- exponential amplification of DNA
- uses DNA polymerase and two opposing primers
- rounds of DNA denaturation and DNA synthesis
  - smart tric: DNA polymerase from thermophilic organism (e.g. *Thermus aquaticus* -> Taq polymerase)
- many, many applications
  - detection; mapping; mutagenesis; gene isolation; sequencing; cloning


Amplification of target sequence
Original target double-stranded DNA
5'
3'
(a)
Separate strands and anneal primers.
5'
3'
(b)
Primer 2
Primer 1
3'
5'
Extend primers.
5'
3'
(c)
Complementary to primer 2
Complementary to primer 1
3'
5'
Separate strands and anneal primers.
3'
5'
(d)
New primers
5'
3'
Extend primers.
(e)
Variable-length strands
Unit-length strands
Separate strands and anneal primers.
(f)
5'
3'
Complementary to primer 2
Complementary to primer 1
3'
5'
Extend primers.
3'
5'
5'
3'
(g)
3'
5'
5'
3'
Desired fragments (variable-length strands not shown)
And so forth

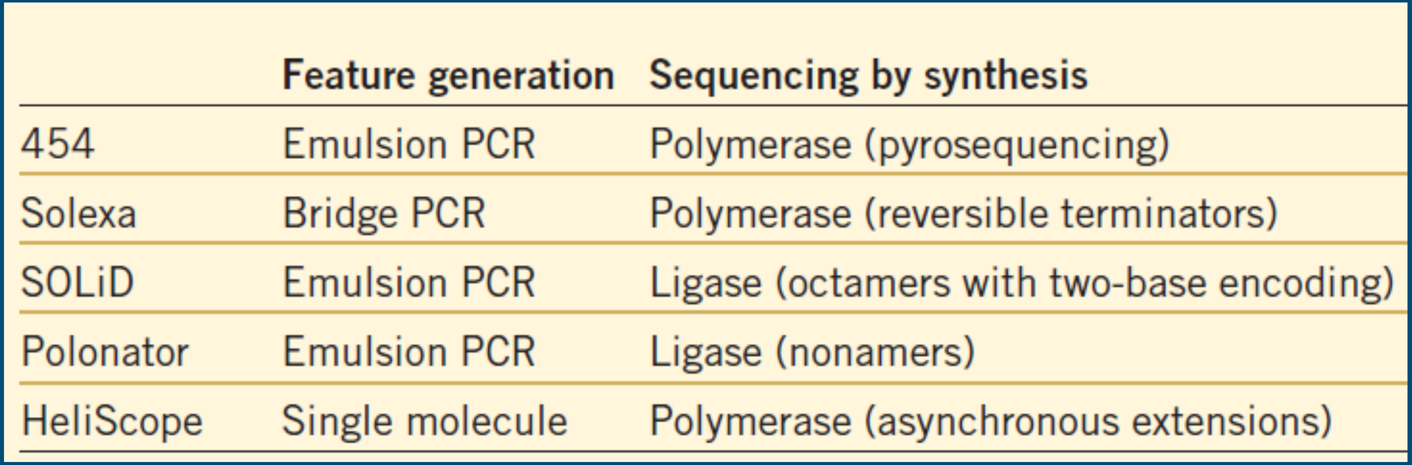
DNA sequencing



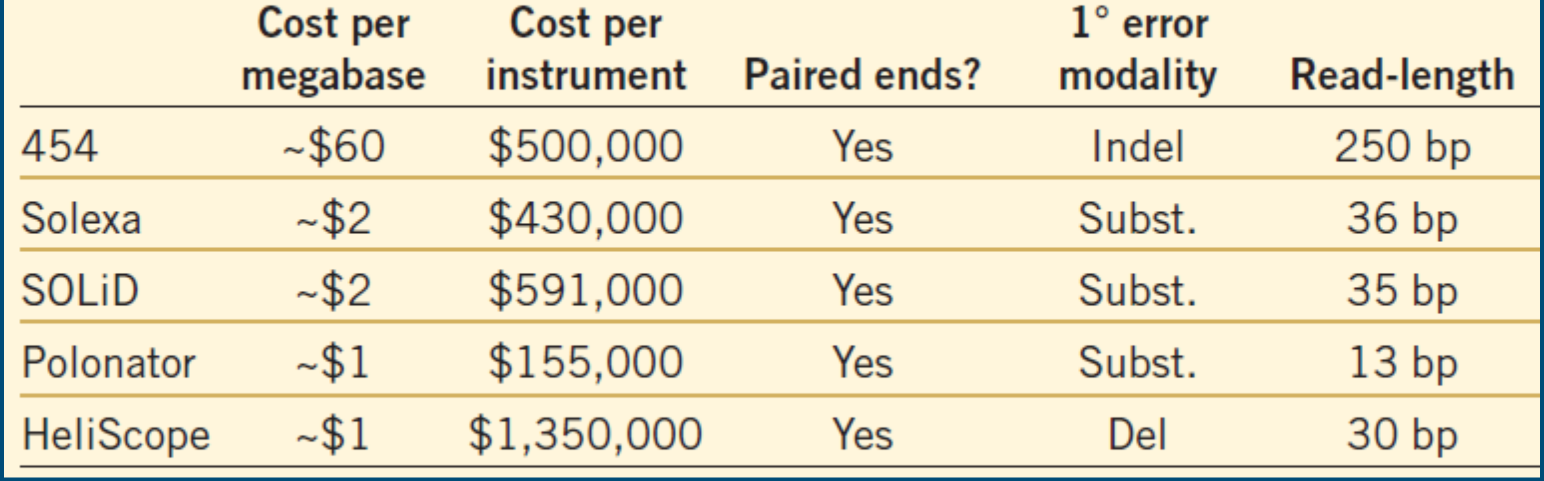


This is how it used to be done…

Next-gen: Comparison of existing methods

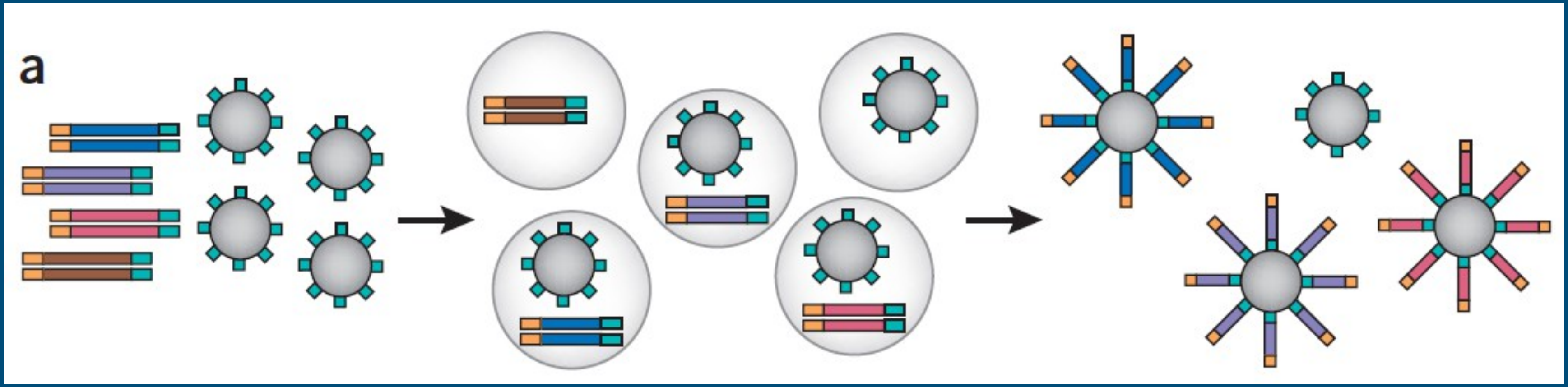


	Feature generation	Sequencing by synthesis
454	Emulsion PCR	Polymerase (pyrosequencing)
Solexa	Bridge PCR	Polymerase (reversible terminators)
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)
Polonator	Emulsion PCR	Ligase (nonamers)
HeliScope	Single molecule	Polymerase (asynchronous extensions)



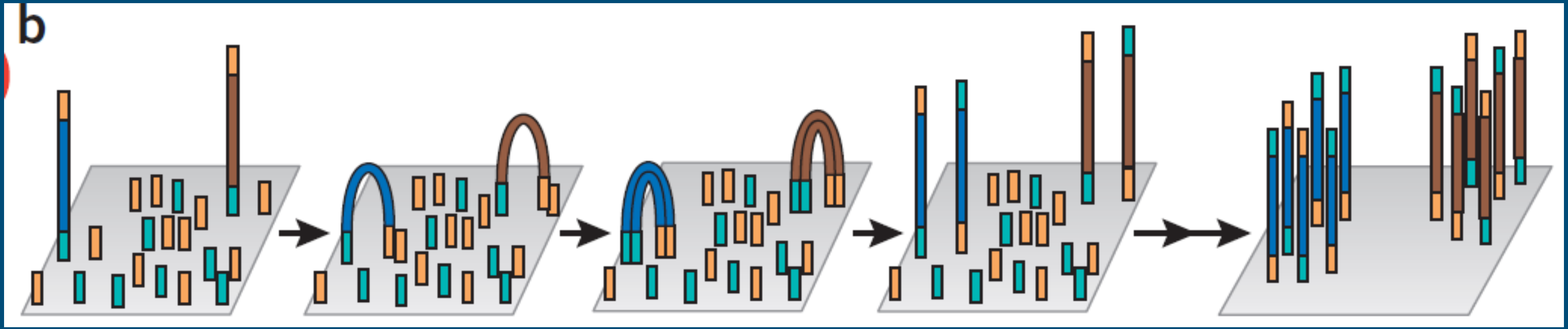
	Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length
454	~$60	$500,000	Yes	Indel	250 bp
Solexa	~$2	$430,000	Yes	Subst.	36 bp
SOLiD	~$2	$591,000	Yes	Subst.	35 bp
Polonator	~$1	$155,000	Yes	Subst.	13 bp
HeliScope	~$1	$1,350,000	Yes	Del	30 bp

Next-gen sequencing: Emulsion PCR



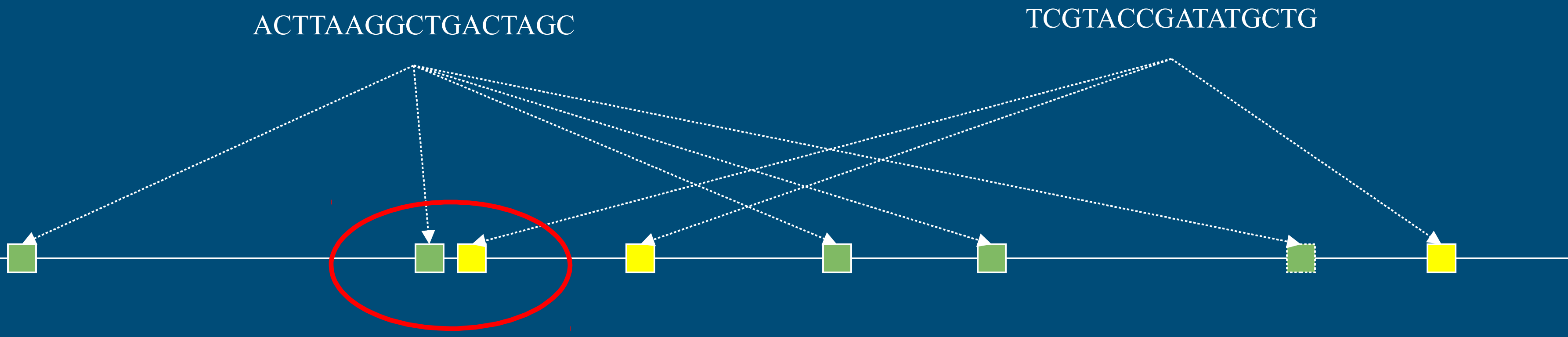
- Fragments, with adaptors, are PCR amplified within a water drop in oil.
- One primer is attached to the surface of a bead.
- Used by 454, Polonator and SOLiD.

Next-gen sequencing: Bridge PCR



- DNA fragments are flanked with adaptors.
- A flat surface coated with two types of primers, corresponding to the adaptors.
- Amplification proceeds in cycles, with one end of each bridge tethered to the surface.
- Used by Solexa.

Read length and pairing



- Short reads are problematic, because short sequences do not map uniquely to the genome.
- Solution #1: Get longer reads.
- Solution #2: Get paired reads.

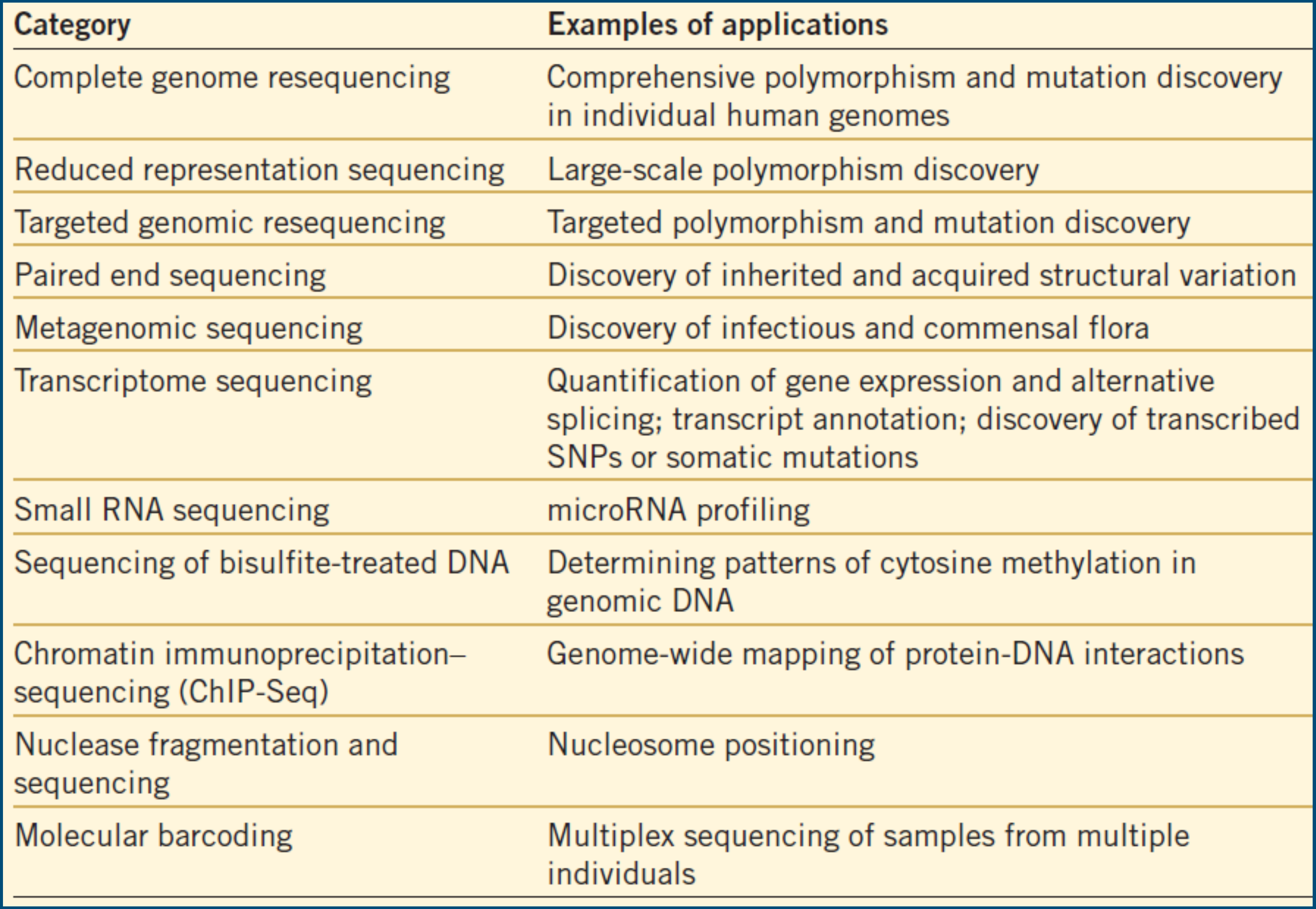
Third generation

- Nanopore sequencing
  - Nucleic acids driven through a nanopore.
  - Differences in conductance of pore provide readout.
- Real-time monitoring of PCR activity
  - Read-out by fluorescence resonance energy transfer between polymerase and nucleotides or
  - Waveguides allow direct observation of polymerase and fluorescently labeled nucleotides

Analysis tasks

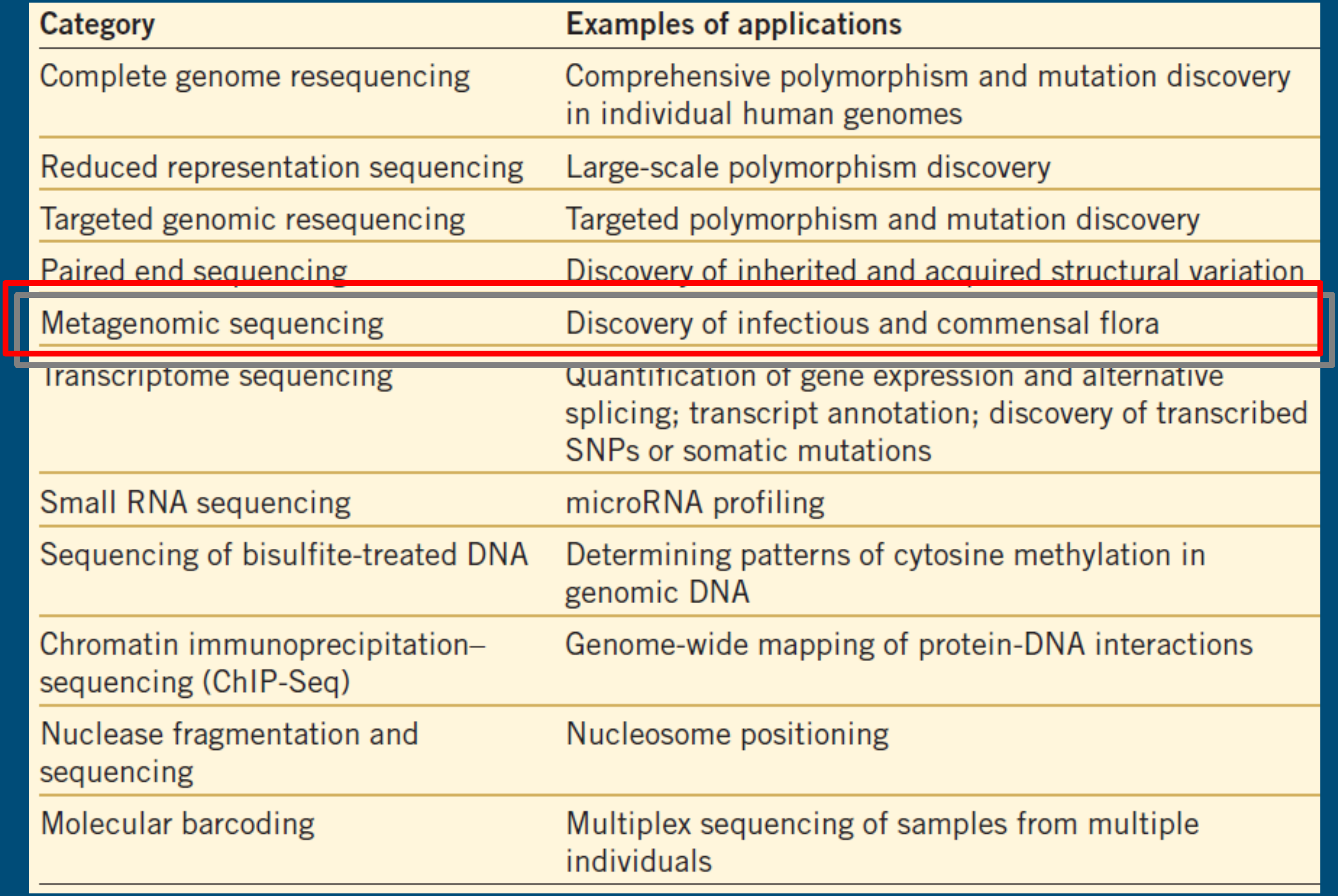
- Base calling / polymorphism detection
- Mapping to a reference genome
    - Expression studies!
    - SNP identification / GWAS studies
- *De novo* or assisted genome assembly
    - Annotation!
- Metagenomics

Next-generation sequencing: Applications



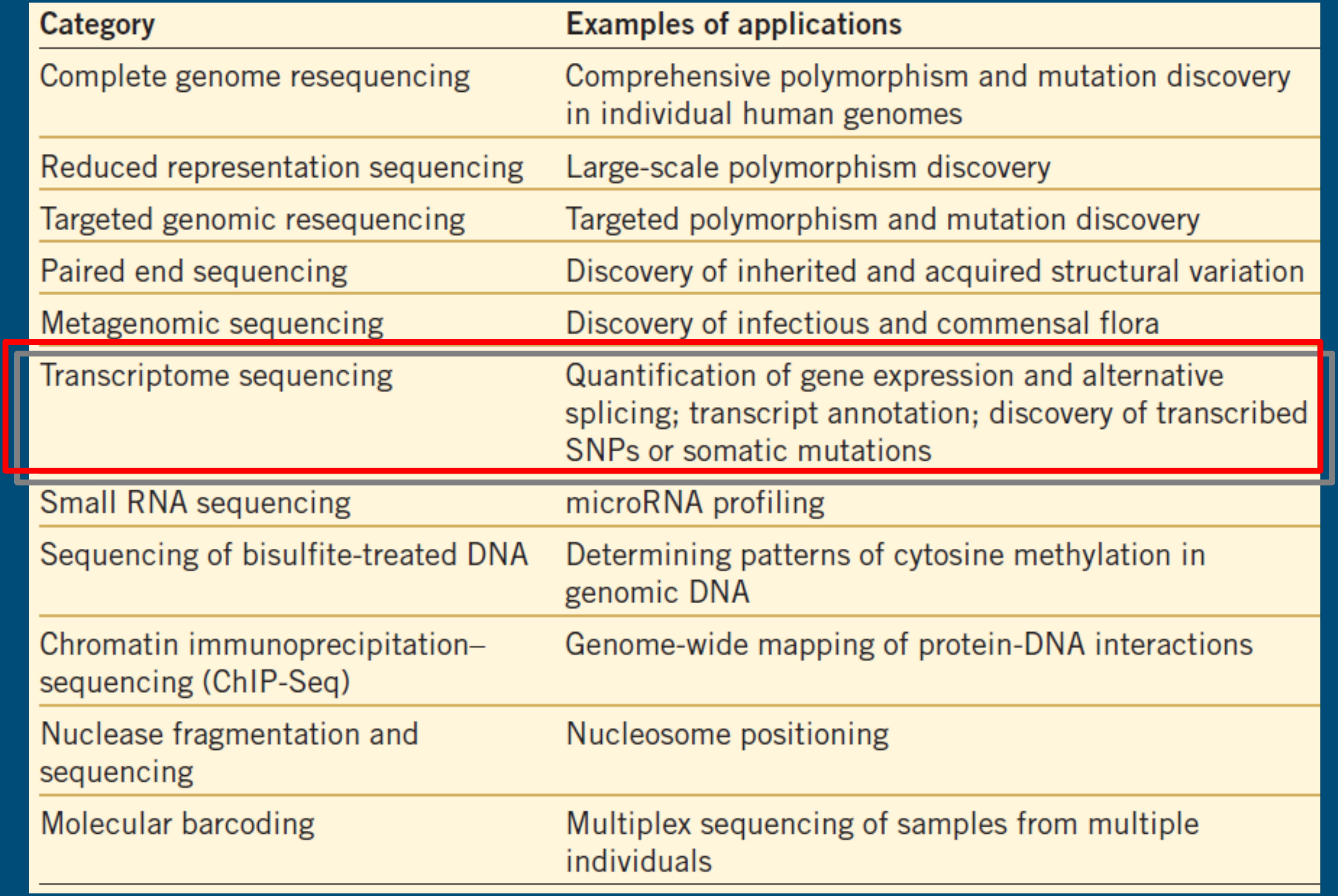
Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

Next-generation sequencing: Applications



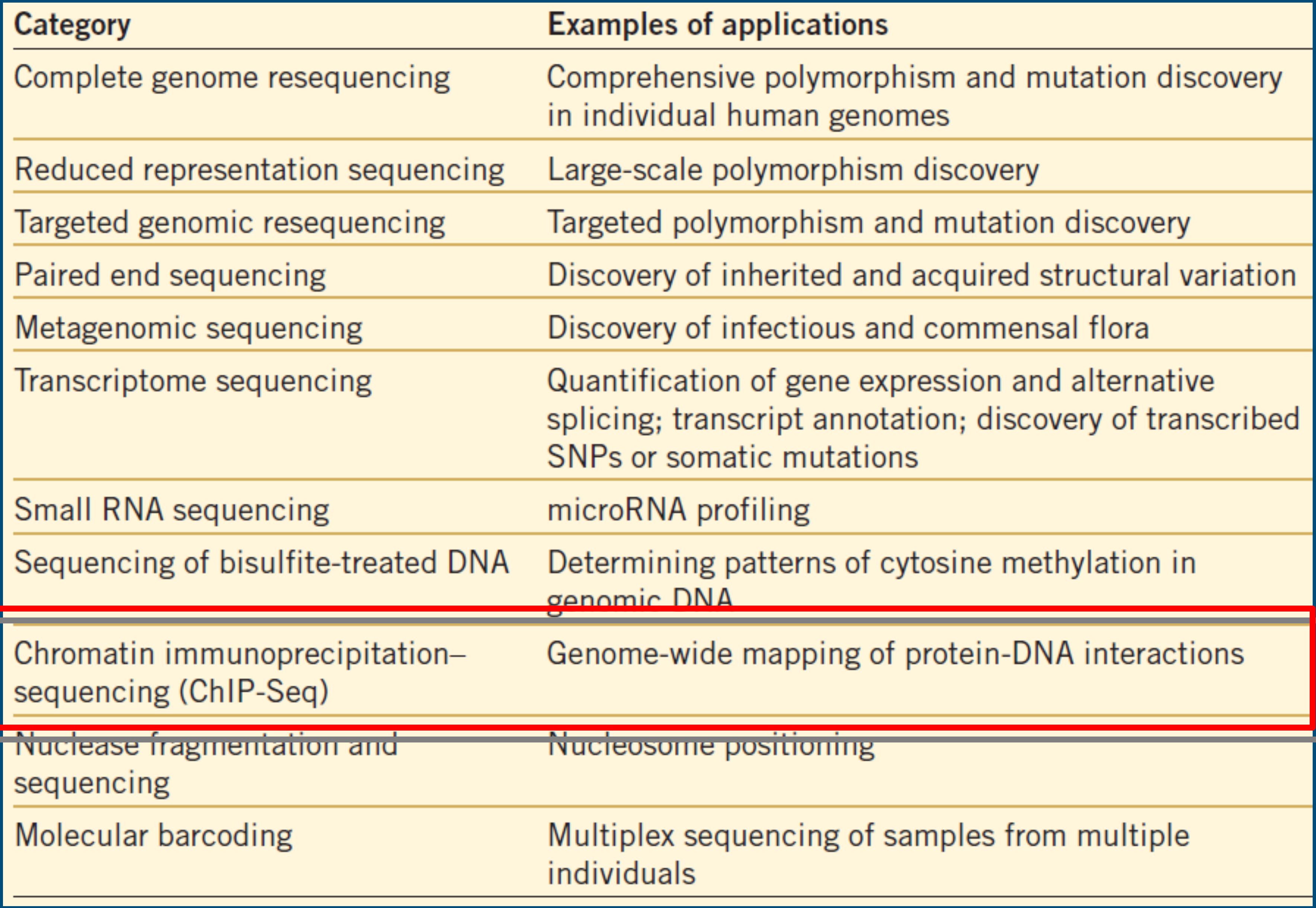
Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

Next-generation sequencing: Applications



Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

Next-generation sequencing: Applications



Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals