# BASICS ON MOLECULAR BIOLOGY

- Cell – DNA – RNA – protein

- Sequencing methods

    - arising questions for handling the data, making sense of it

    - next two week lectures: sequence alignment and genome
   assembly

Cells

- Fundamental working units of every living system.
- Every organism is composed of one of two radically different types of cells:
  - **prokaryotic** cells
  - **eukaryotic** cells which have DNA inside a **nucleus.**
- **Prokaryotes** and **Eukaryotes** are descended from primitive cells and the results of
 3.5 billion years of evolution.
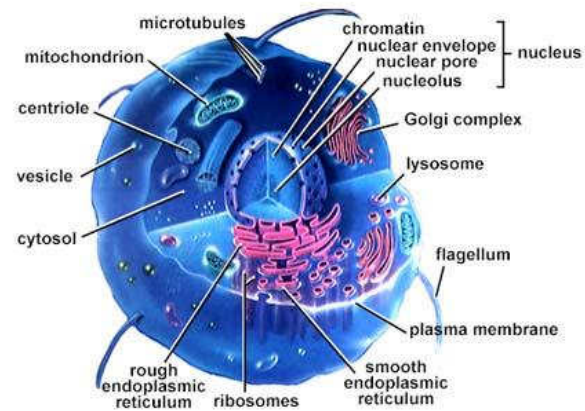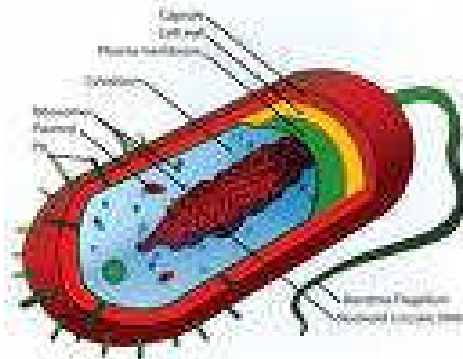
Prokaryotes and Eukaryotes

- According to the most recent evidence, there are three main branches to the tree of life

- Prokaryotes include Archaea ("ancient ones") and bacteria

- Eukaryotes are kingdom Eukarya and includes plants, animals, fungi and certain algae
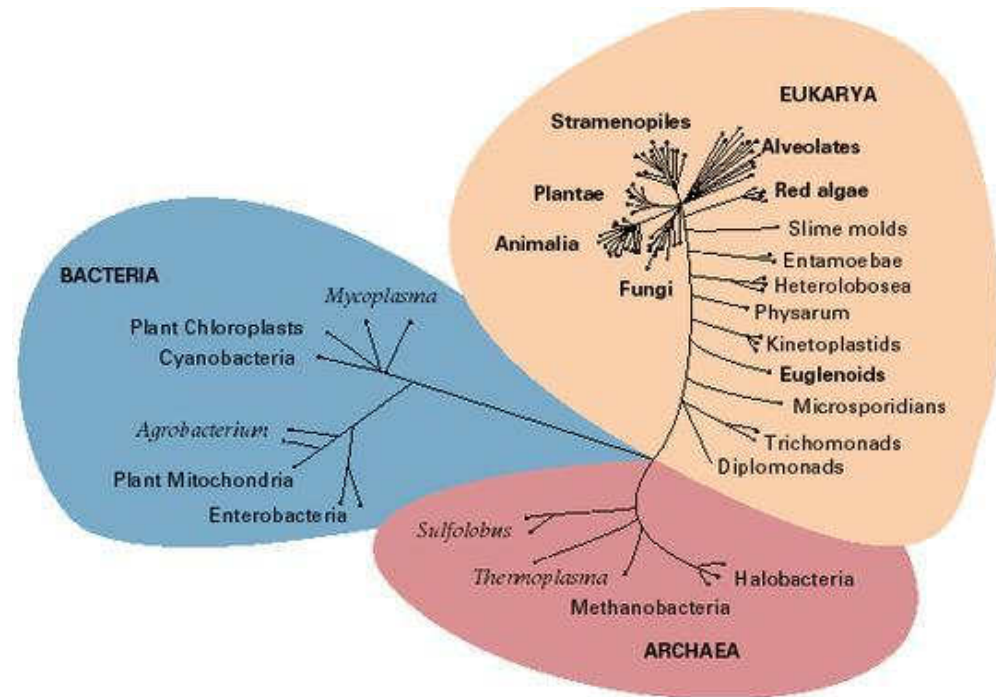
Lecture: Phylogenetic trees, this topic in more detail

All Cells have common Cycles
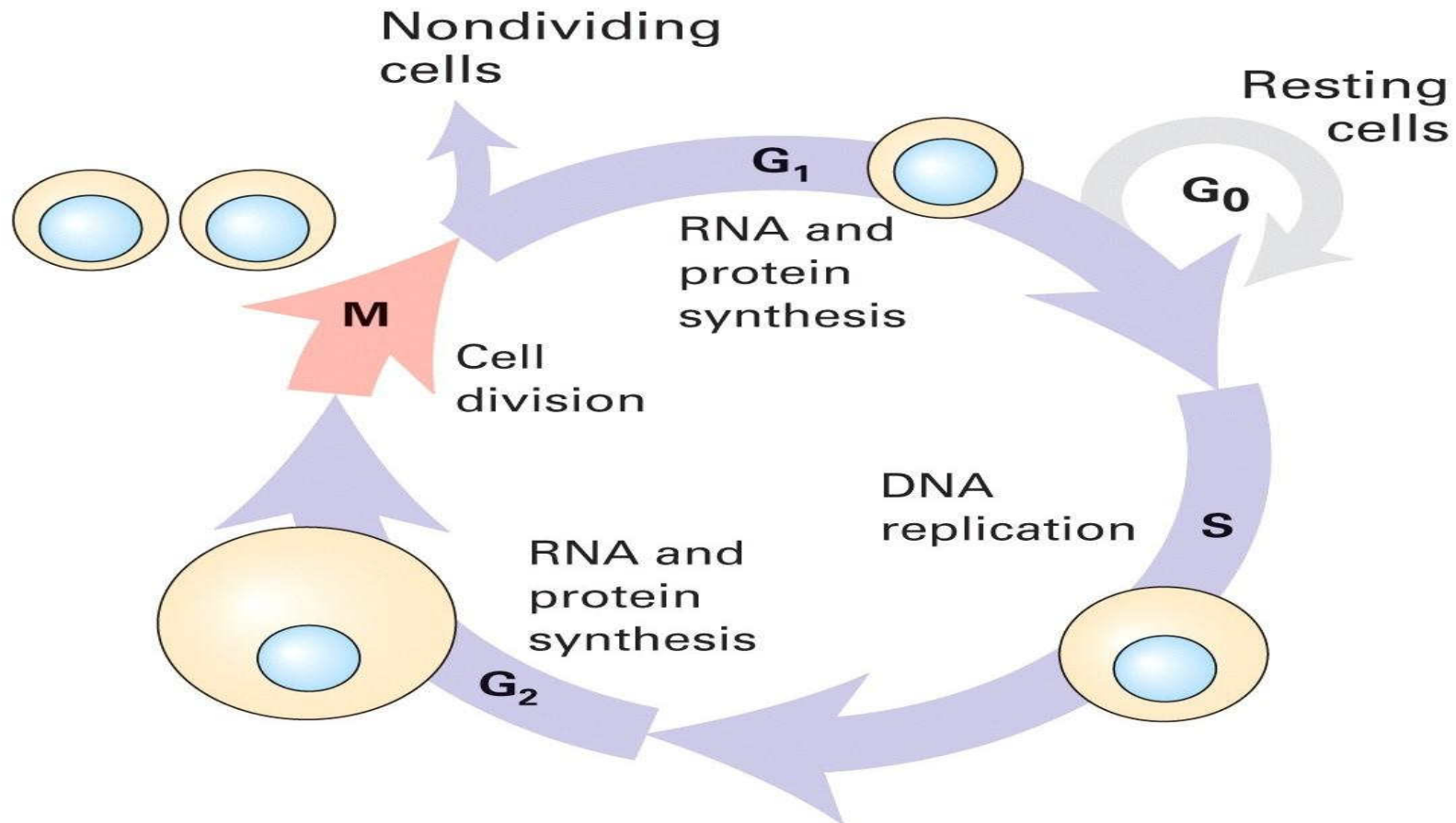

- Born, eat, replicate, and die

Common features of organisms

- Chemical energy is stored in ATP

- Genetic information is encoded by DNA

- Information is transcribed into RNA

- There is a **common triplet genetic code**
  - **some variations are known, however**
- Translation into proteins involves ribosomes

- Shared metabolic pathways

- Similar proteins among diverse groups of organisms

All Life depends on 3 critical molecules

- DNAs (Deoxyribonucleic acid)
  - Hold information on how cell works

- RNAs (Ribonucleic acid)
  - Act to transfer short pieces of information to different parts of cell
  - Provide templates to synthesize into protein

- Proteins
  - Form enzymes that send signals to other cells and regulate gene activity
  - Form body's major components
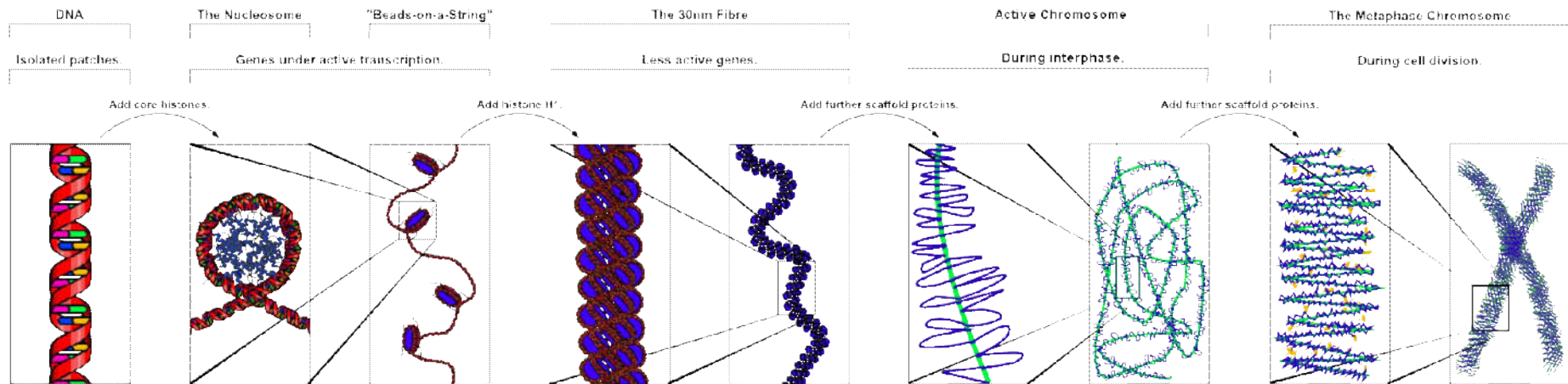
DNA structure
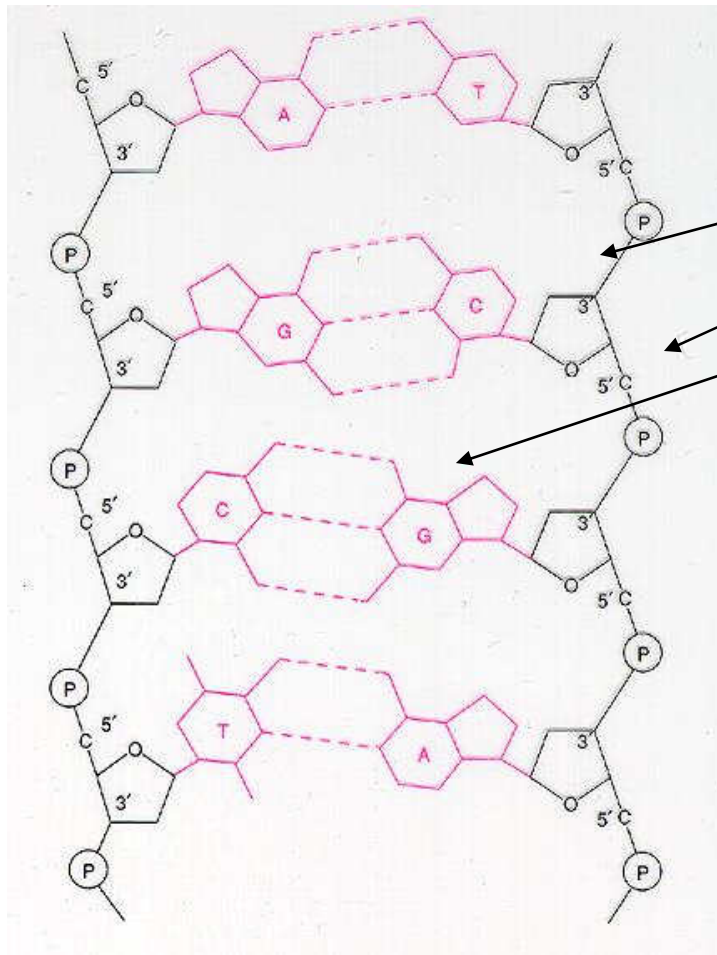
- DNA has a double helix structure which is composed of
  - sugar molecule
  - phosphate group
  - and a base (A,C,G,T)

- By convention, we read DNA strings in direction of transcription: from 5' end to 3' end

  5' ATTTAGGCC 3'
 3' TAAATCCGG 5'

DNA is contained in chromosomes

□ In eukaryotes, DNA is packed into linear chromosomes

□ In prokaryotes, DNA is usually contained in a single, circular chromosome

http://en.wikipedia.org/wiki/Image:Chromatin_Structures.png

Human chromosomes

- Somatic cells (cells in all, except the germline, tissues) in humans have 2 pairs of 22 chromosomes + XX (female) or XY (male) = total of 46 chromosomes

- Germline cells have 22 chromosomes + either X or Y = total of 23 chromosomes
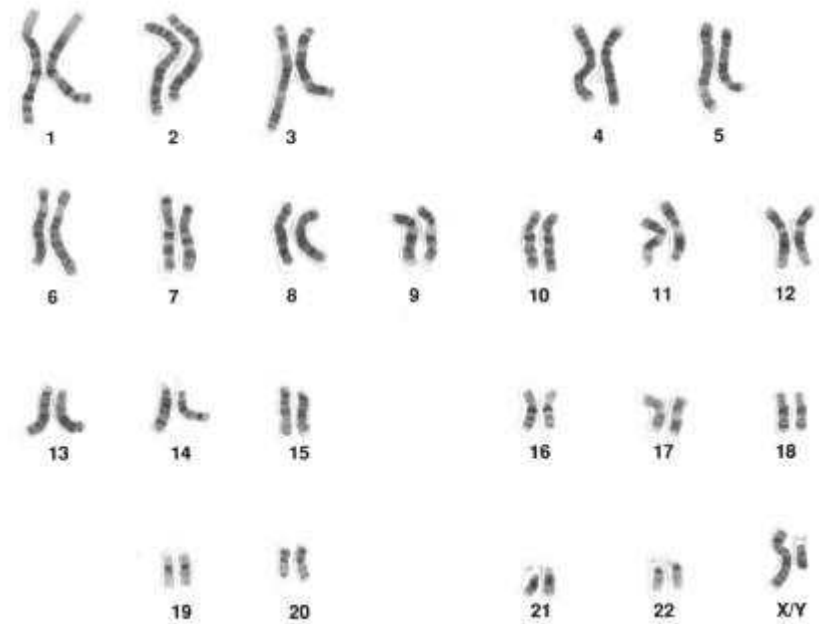
Karyogram of human male using Giemsa staining
(http://en.wikipedia.org/wiki/Karyotype)

RNA

- RNA is similar to DNA chemically. It is usually only a single strand. T(hyamine) is replaced by U(racil)

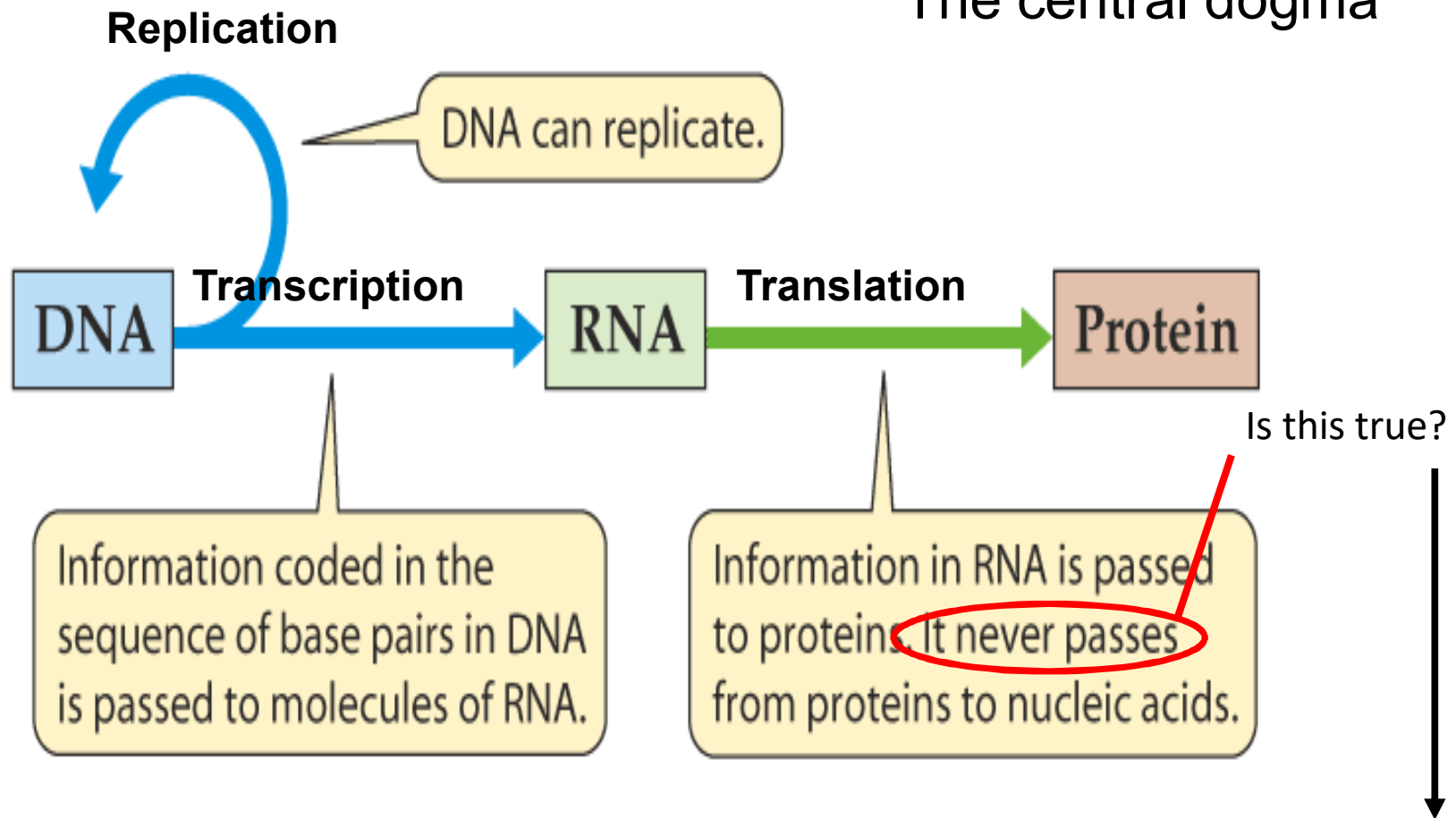- Several types of RNA exist for different functions in the cell.

tRNA linear and 3D view: http://www.cgl.ucsf.edu/home/glasfeld/tutorial/trna/trna.gif
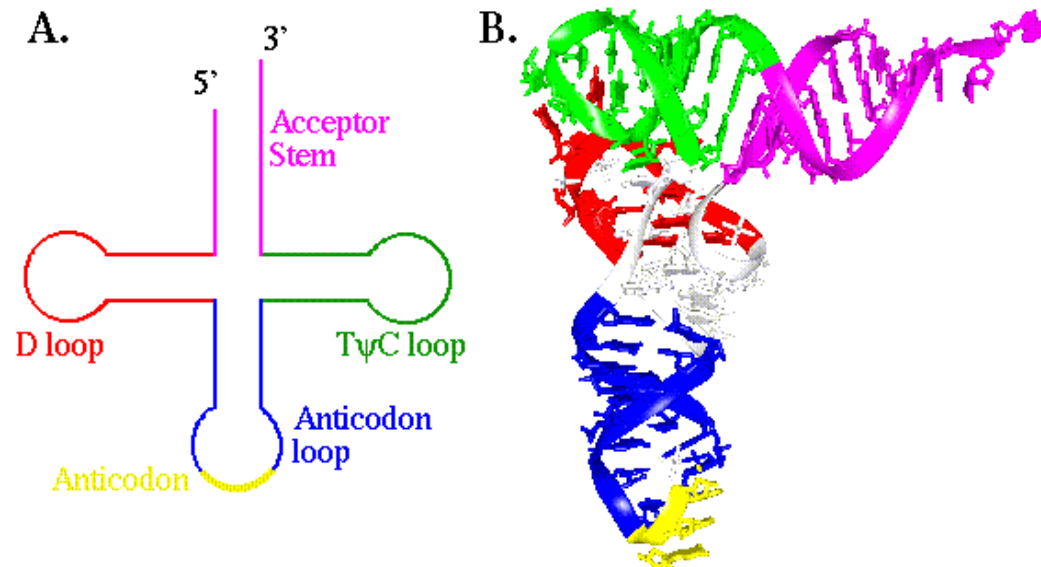
DNA, RNA, and the Flow of Information

Proteins

- Proteins are polypeptides (strings of amino acid residues)

- Represented using strings of letters from an alphabet of 20: AEGLV…WKKLAG

- Typical length 50…1000 residues

Urease enzyme from Helicobacter pylori

Amino acids

How DNA/RNA codes for protein?

- DNA alphabet contains four letters but must specify protein, or polypeptide sequence of 20 letters.

- Trinucleotides (triplets) allow $4^3 = 64$ possible trinucleotides

- Triplets are also called *codons*

Proteins

- 20 different **amino acids**
  - different chemical properties cause the protein chains to fold up into specific three-dimensional structures that define their particular functions in the cell.

- Proteins do all essential work for the cell
  - build cellular structures
  - digest nutrients
  - execute metabolic functions
  - mediate information flow within a cell and among cellular communities.

- Proteins work together with other proteins or nucleic acids as "molecular machines"
  - structures that fit together and function in highly specific, lock-and-key ways.

Genes

- "A gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products"

- A DNA segment whose information is expressed either as an RNA molecule or protein

(translation)   (folding)

MSG …

augagugga …

(transcription)

5' … a t g a g t g g a … 3'

3' … t a c t c a c c t … 5'

http://fold.it

16
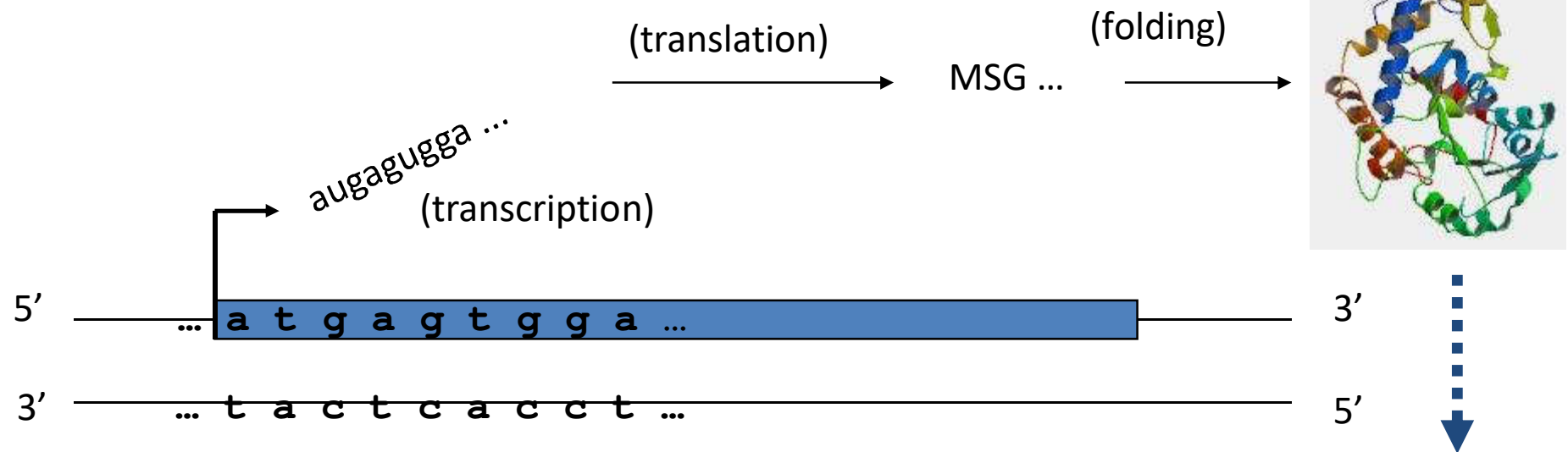
Genes & alleles

- A gene can have different variants
- The variants of the same gene are called *alleles*
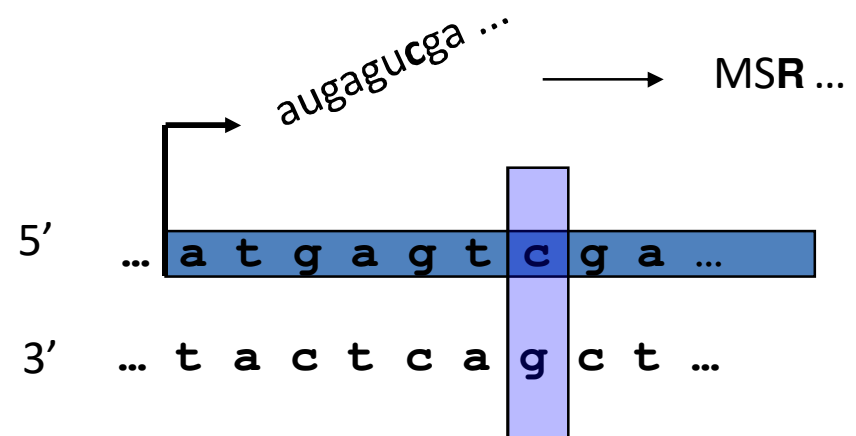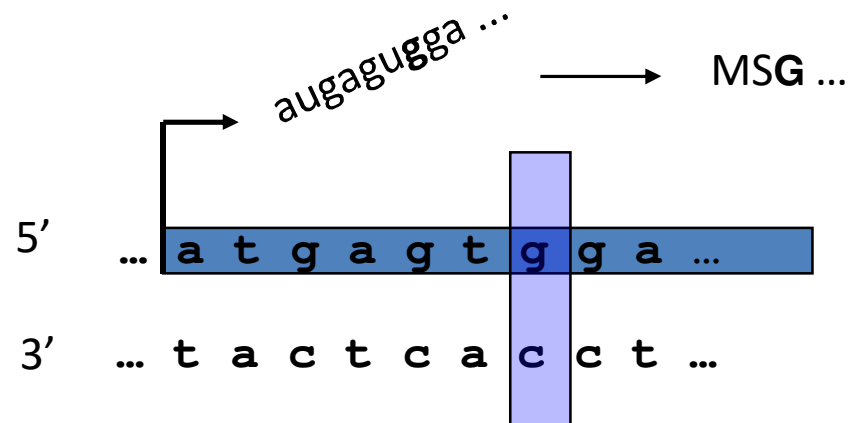
Genes can be found on both strands
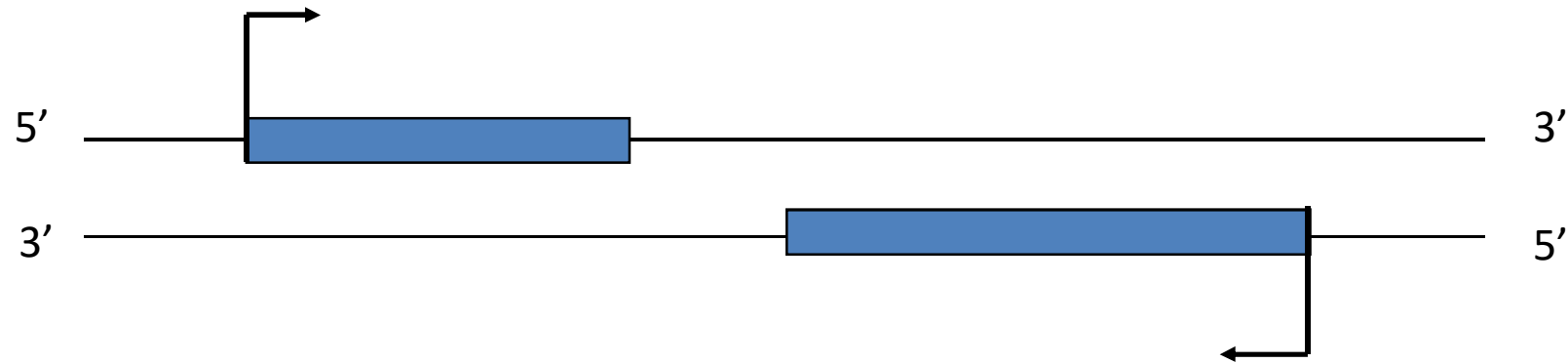
Exons and introns & splicing
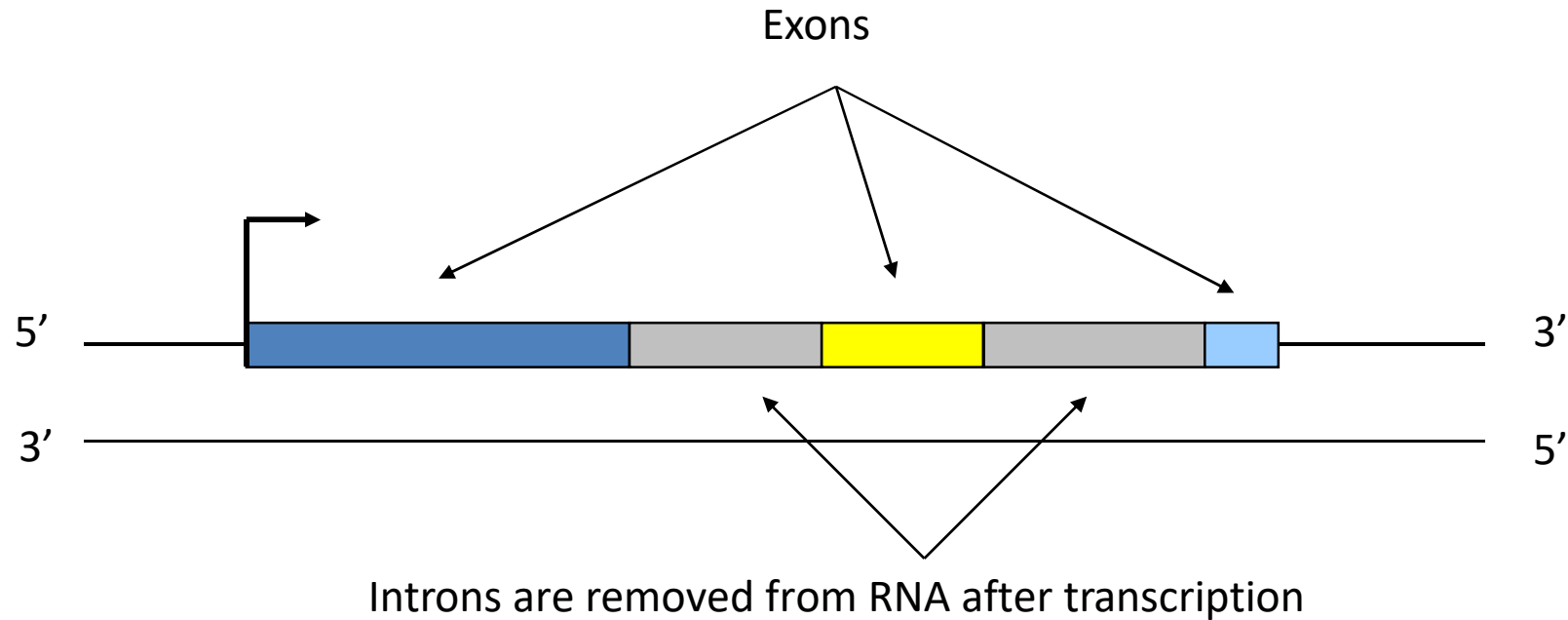
Alternative splicing

Different *splice variants* may be generated
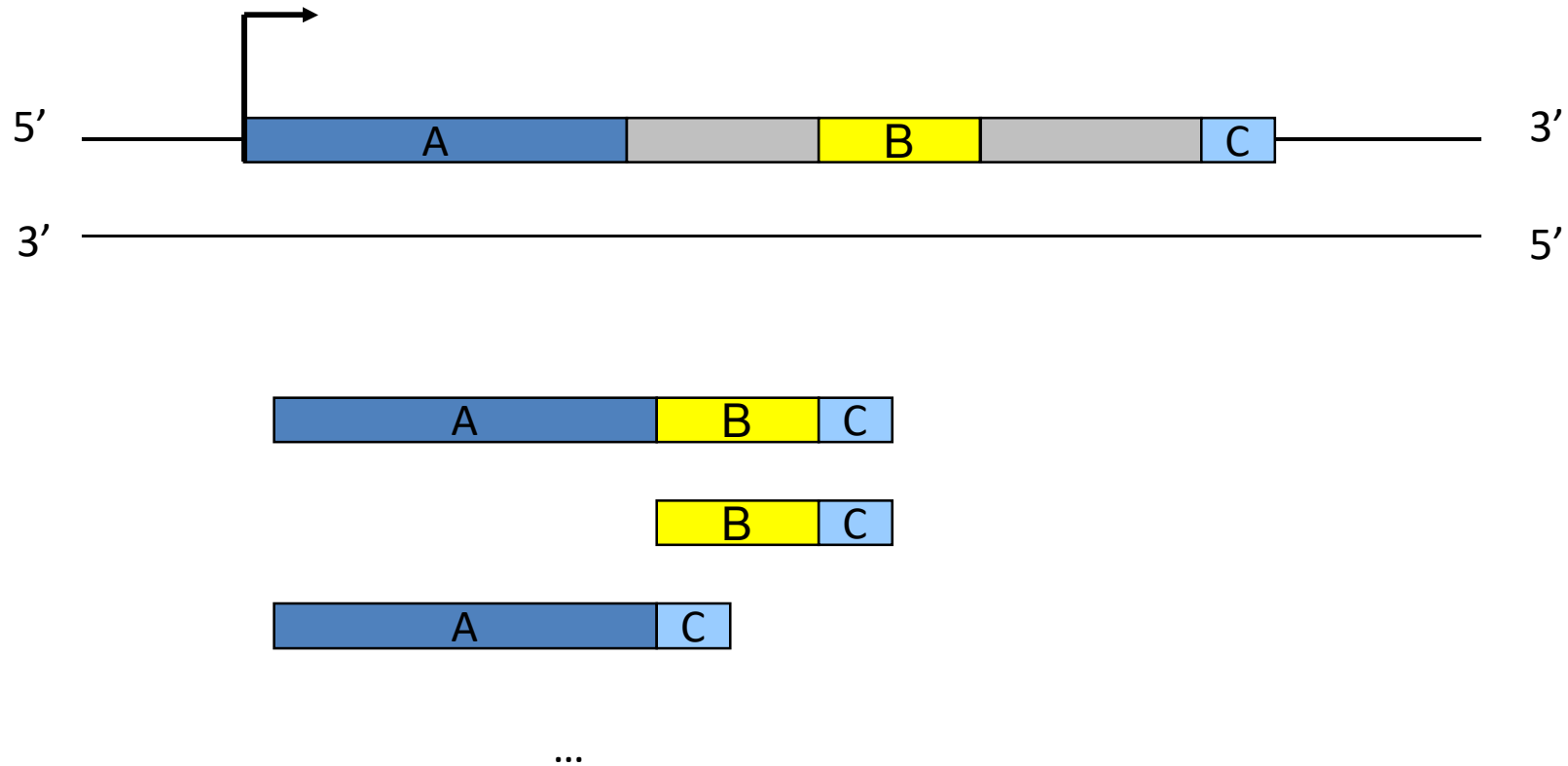
DNA and continuum of life….

- Prokaryotes are typically haploid: they have a single (circular) chromosome

- DNA is usually inherited vertically (parent to daughter)

- Inheritance is clonal
  - Descendants are faithful copies of an ancestral DNA
  - Variation is introduced via mutations, transposable elements, and horizontal transfer of DNA
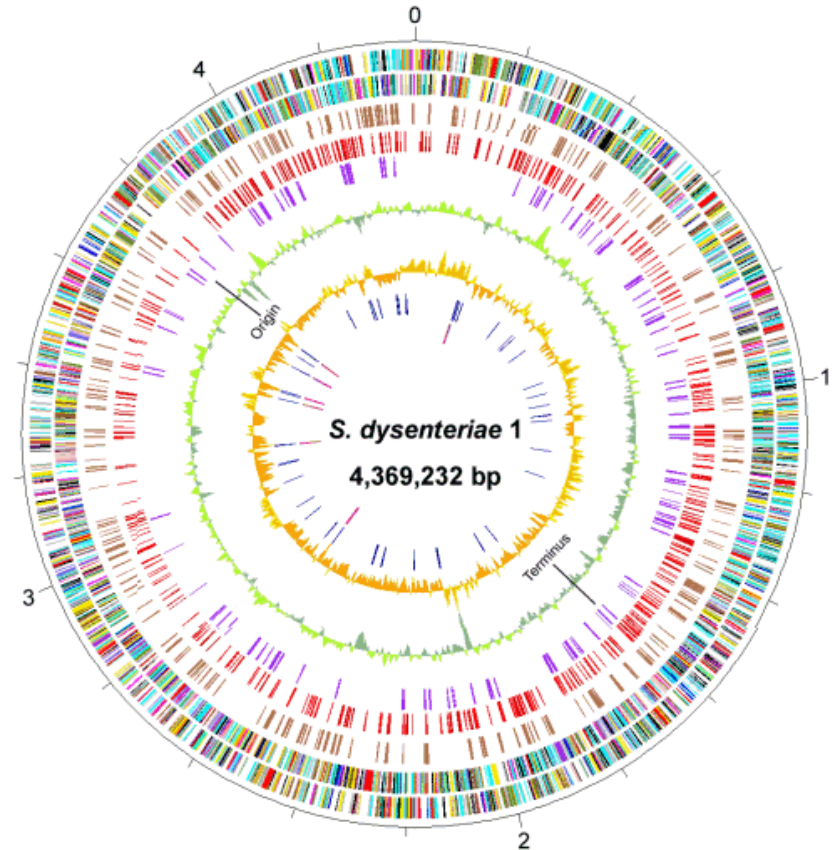
Chromosome map of *S. dysenteriae*, the nine rings describe different properties of the genome
http://www.mgc.ac.cn/ShiBASE/circular_Sd197.htm

21

Biological string manipulation

- Point mutation: substitution of a base
    - …ACG**G**CT… => …ACG**C**CT…
- Deletion: removal of one or more contiguous bases (substring)
    - …TT**GA**TCA… => …TTTCA…
- Insertion: insertion of a substring
    - …GGCTAG… => …GG**TCAAC**TAG…

▶ Lecture: Sequence alignment
Lecture: Genome rearrangements

Genome sequencing & assembly

- DNA sequencing
  - How do we obtain DNA sequence information from organisms?

- Genome assembly
  - What is needed to put together DNA sequence information from sequencing?

- First statement of sequence assembly problem:
  - Peltola, Söderlund, Tarhio, Ukkonen: Algorithms for some string matching problems arising in molecular genetics. Proc. 9th IFIP World Computer Congress, 1983

Recovery of shredded newspaper
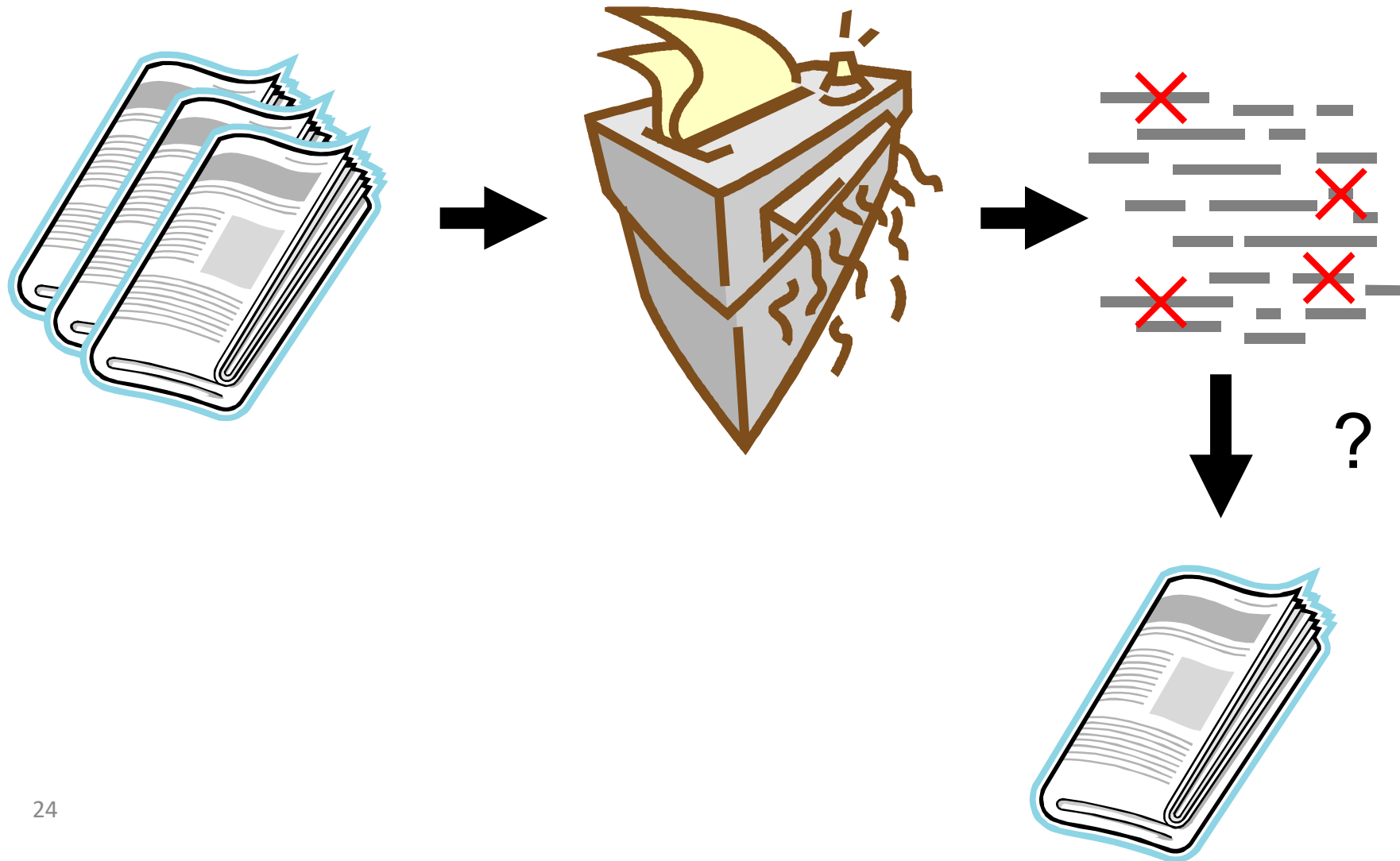
DNA sequencing

- DNA sequencing: resolving a nucleotide sequence (whole-genome or less)
- Many different methods developed
  - Maxam-Gilbert method (1977)
  - Sanger method (1977)
  - High-throughput methods, "next-generation" methods

Sanger sequencing: sequencing by synthesis

- A sequencing technique developed by 1977
- Also called *dideoxy sequencing*
- A *DNA polymerase* is an enzyme
  that catalyzes DNA synthesis
- DNA polymerase needs a *primer*
- Synthesis proceeds always in 5'->3' direction
- In Sanger sequencing, chain-terminating
- dideoxynucleoside triphosphates (ddXTPs) are employed
  - ddATP, ddCTP, ddGTP, ddTTP
    lack the 3'-OH tail of dXTPs
- A mixture of dXTPs with small amount of ddXTPs
  is given to DNA polymerase with DNA template and primer
- ddXTPs are given fluorescent labels
- When DNA polymerase encounters a ddXTP, the synthesis
  cannot proceed
- The process yields copied sequences of different lengths
- Each sequence is terminated by a labeled ddXTP
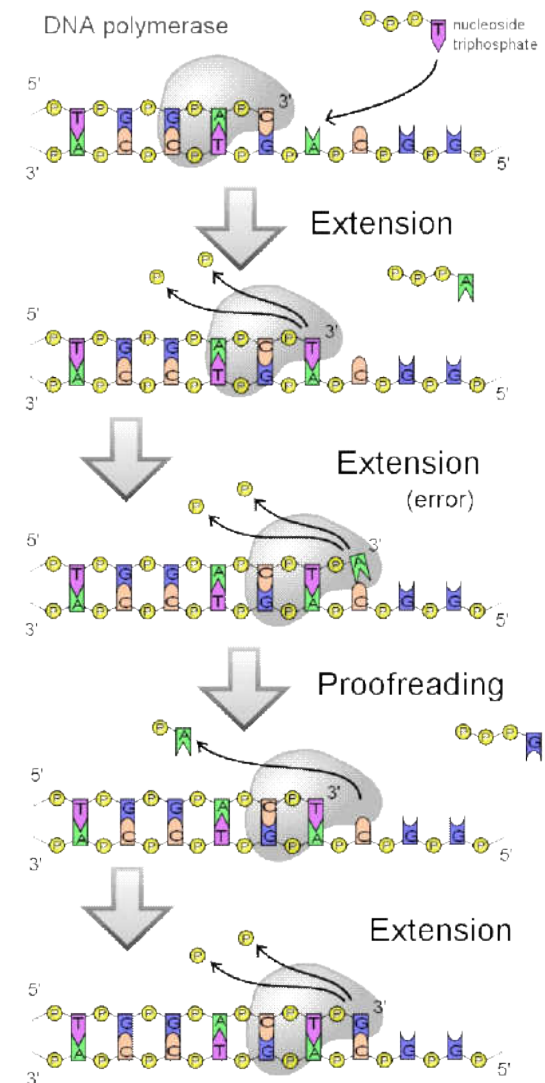
Determining the sequence

- Sequences are sorted according to length by capillary electrophoresis

- Fluorescent signals corresponding to labels are registered

- *Base calling*: identifying which base corresponds to each position in a read
  - Non-trivial problem!

AAACAACTTCGTAAGTATA
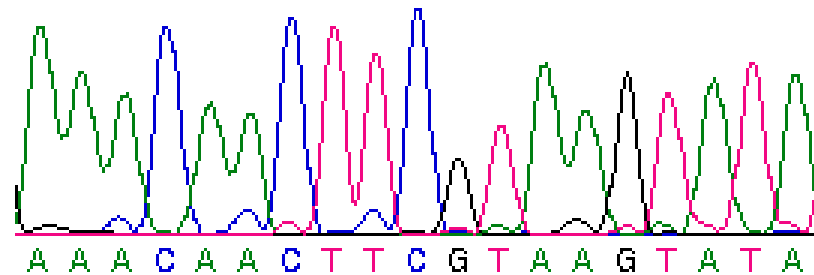
Output sequences from base calling are called **reads**

Reads are short!

- Modern Sanger sequencers can produce quality reads up to ~750 bases[1]
  - Instruments provide you with a quality file for bases in reads, in addition to actual sequence data
- Compare the read length against the size of the human genome (2.9×10^9 bases)

- Reads have to be **assembled**!

Problems

- Sanger sequencing error rate per base varies from 1% to 3%[1]
- **Repeats in DNA**
  - For example, ~300 base longs *Alu* sequence repeated is over million times in human genome
  - Repeats occur in different scales
- What happens if repeat length is longer than read length?

- **Shortest superstring problem**
  - Find the shortest string that "explains" the reads
  - *Given a set of strings (reads), find a shortest string that contains all of them*

Sequence assembly and combination locks

- What is common with sequence assembly and opening keypad locks?
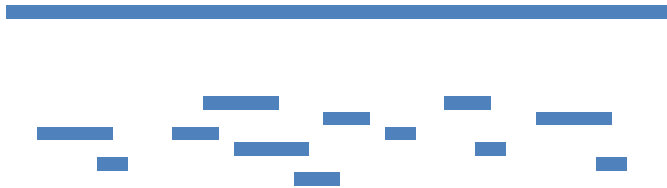
Whole-genome shotgun sequence

- *Whole-genome shotgun sequence assembly* starts with a large sample of genomic DNA

  1. Sample is randomly partitioned into *inserts* of length > 500 bases
  2. Inserts are multiplied by cloning them into *a vector* which is used to infect bacteria
  3. DNA is collected from bacteria and sequenced
  4. Reads are assembled

Assembly of reads with Overlap-Layout-Consensus algorithm

- Overlap
  - Finding potentially overlapping reads
- Layout
  - Finding the order of reads along DNA
- Consensus (Multiple alignment)
  - Deriving the DNA sequence from the layout

- Next, the method is described at a very abstract level, skipping a lot of details

Finding overlaps

acggagtcc
    agtccgcgctt

- First, pairwise overlap alignment of reads is resolved

- Reads can be from either DNA strand: The *reverse complement* r* of each read r has to be considered

r_1

5' … a t g a g t g g a … 3'

3' … t a c t c a c c t … 5'

r_2
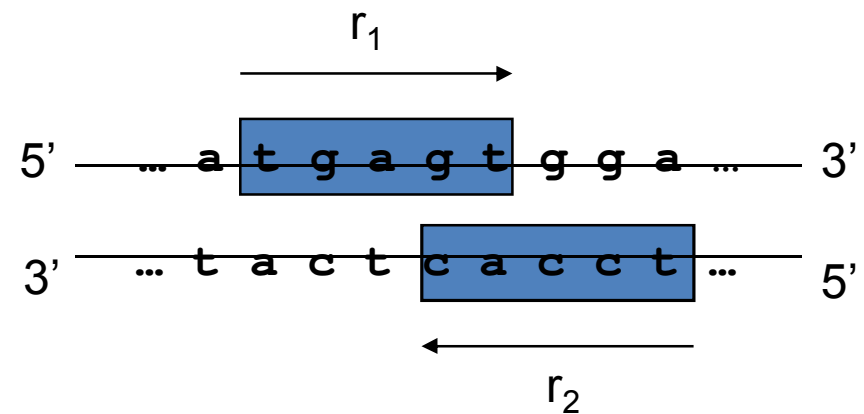
r_1: tgagt, r_1^*: actca
r_2: tccac, r_2^*: gtgga

Example sequence to assemble

5' – C**AGCGCGCT**GCGTGACGAGTCTGACAAAGACGGTATGCGCATCG
TGATTGAAGTGAAACGCG**ATGCGGTC**GGTCGGTGAAGTTGTGCT - 3'

- ## 20 reads:

#	Read	Read*	#	Read	Read*
1	CATCGTCA	TCACGATG	11	GGTCGGTG	CACCGACC
2	CGGTGAAG	CTTCACCG	12	ATCGTGAT	ATCACGAT
3	TATGCGCA	TGCGCATA	13	GCGCTGCG	CGCAGCGC
4	GACGAGTC	GACTCGTC	14	GCATCGTG	CACGATGC
5	CTGACAAA	TTTGTCAG	15	**AGCGCGCT**	**AGCGCGCT**
6	ATGCGCAT	ATGCGCAT	16	GAAGTTGT	ACAACTTC
7	**ATGCGGTC**	GACCGCAT	17	AGTGAAAC	GTTTCACT
8	CTGCGTGA	TCACGCAG	18	ACGCGATG	CATCGCGT
9	GCGTGACG	CGTCACGC	19	GCGCATCG	CGATGCGC
10	GTCGGTGA	TCACCGAC	20	AAGTGAAA	TTTCACTT

Finding overlaps

- Overlap between two reads can be found with a *dynamic programming* algorithm
  - Errors can be taken into account

- Dynamic programming will be discussed more during the next two weeks

- Overlap scores stored into the overlap matrix
  - Entries (i, j) below the diagonal denote overlap of read r_i and r_j^*

Overlap(1, 6) = 3

6 ATGCGCAT

1 CATCGTCA

12 ATCGTGAT
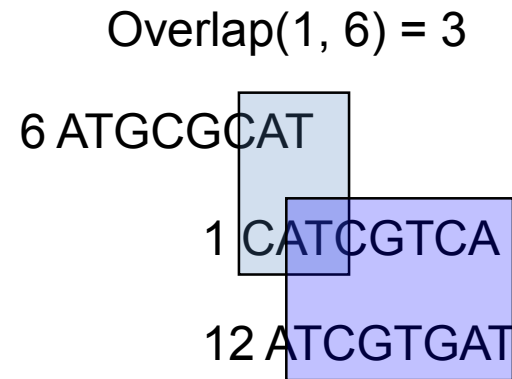
Overlap(1, 12) = 7

	6	12
1	3	7

Finding layout & consensus

- Method extends the assembly *greedily* by choosing the best overlaps

- Both orientations are considered

- Sequence is extended as far as possible

Ambiguous bases

14   GCATCGTG
1   CATCGTGA
12   ATCGTGAT
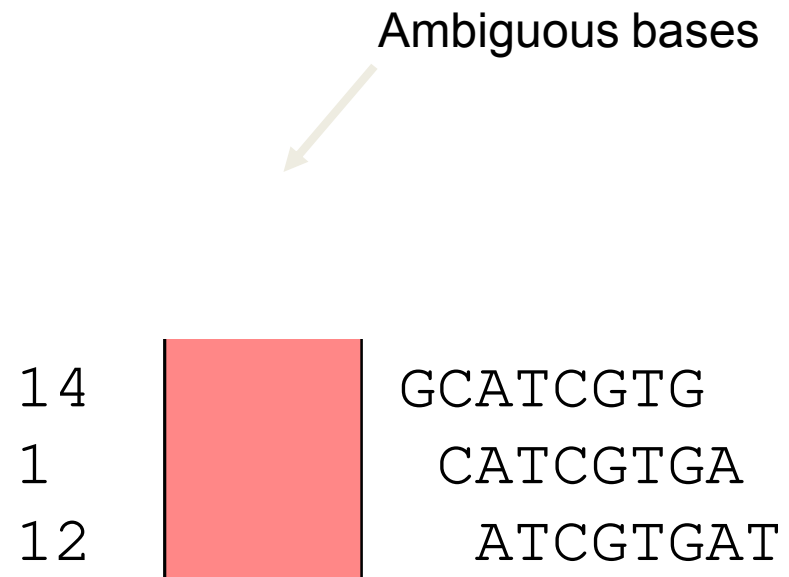
consensus sequence

Finding layout & consensus

- We move on to next best overlaps and extend the sequence from there

- The method stops when there are no more overlaps to consider

- A number of **contigs** is produced

- Contig stands for contiguous sequence, resulting from merging reads

```
2               CGGTGAAG
10             GTCGGTGA
11            GGTCGGTG
7      ATGCGGTC
_____
       ATGCGGTCGGTGAAG
```

Whole-genome shotgun sequencing: summary

Original genome sequence

Reads

Non-overlapping read

Overlapping reads => Contig
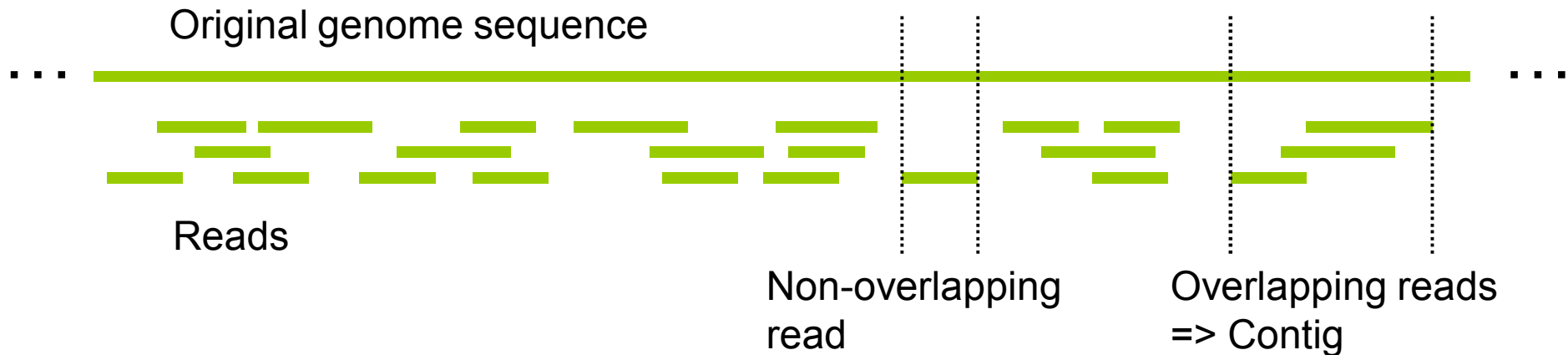
- Ordering of the reads is initially unknown

- Overlaps resolved by aligning the reads

- In a 3×10^9 bp genome with 500 bp reads and 5x coverage, there are ~10^7 reads and ~$10^7(10^7-1)/2 = $ ~5×10^{13} pairwise sequence comparisons

Repeats in DNA and genome assembly

Two instances of the same repeat
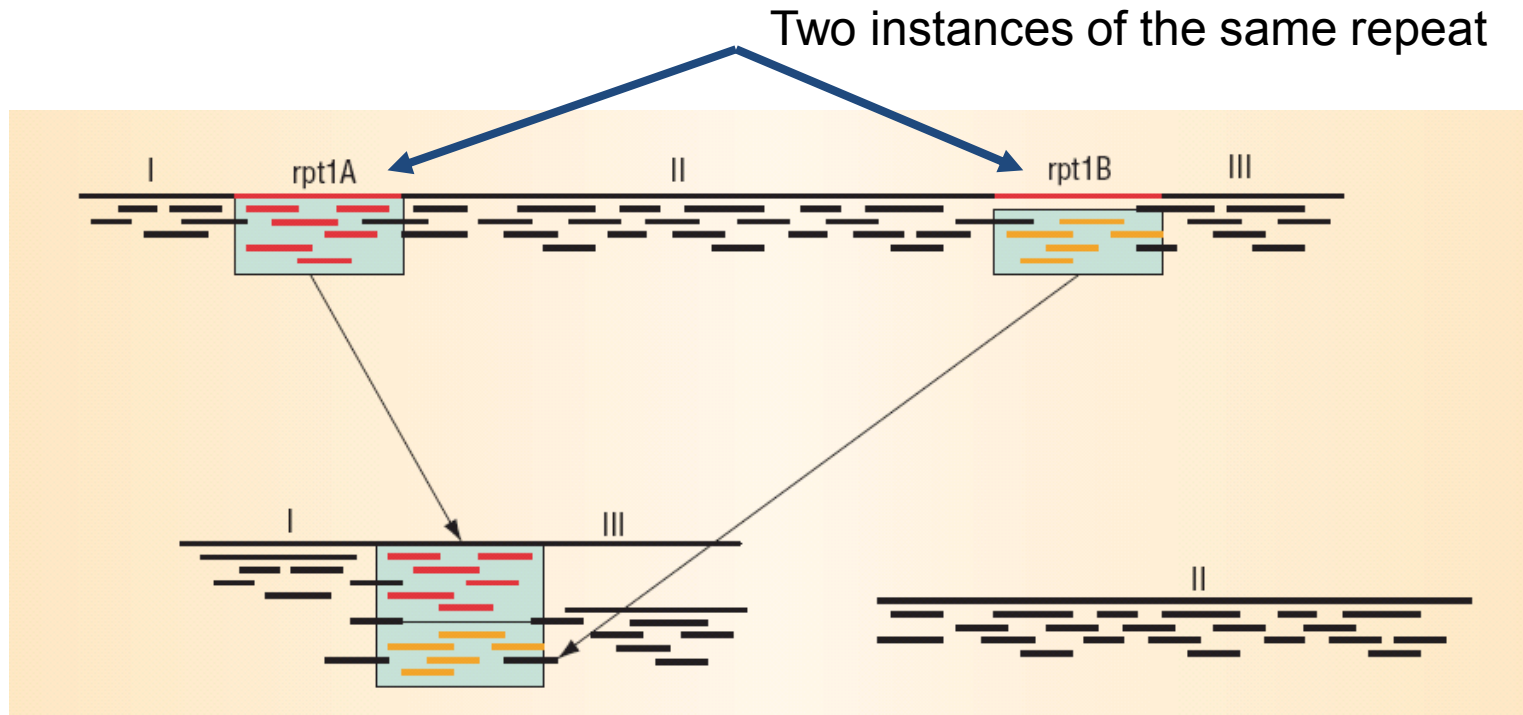
Figure 2. Repeat sequence. The top represents the correct layout of three DNA sequences. The bottom shows a repeat collapsed in a misassembly.

Repeats in DNA cause problems in sequence assembly

- Recap: if repeat length exceeds read length, we might not get the correct assembly

- This is a problem especially in eukaryotes
  - ~3.1% of genome consists of repeats in Drosophila, **~45%** in human

- Possible solutions

1. Increase read length – feasible?
2. Divide genome into smaller parts, with known order, and sequence parts individually

"Divide and conquer" sequencing approaches: BAC-by-BAC

Whole-genome shotgun sequencing

Genome

Divide-and-conquer

Genome

BAC library

41
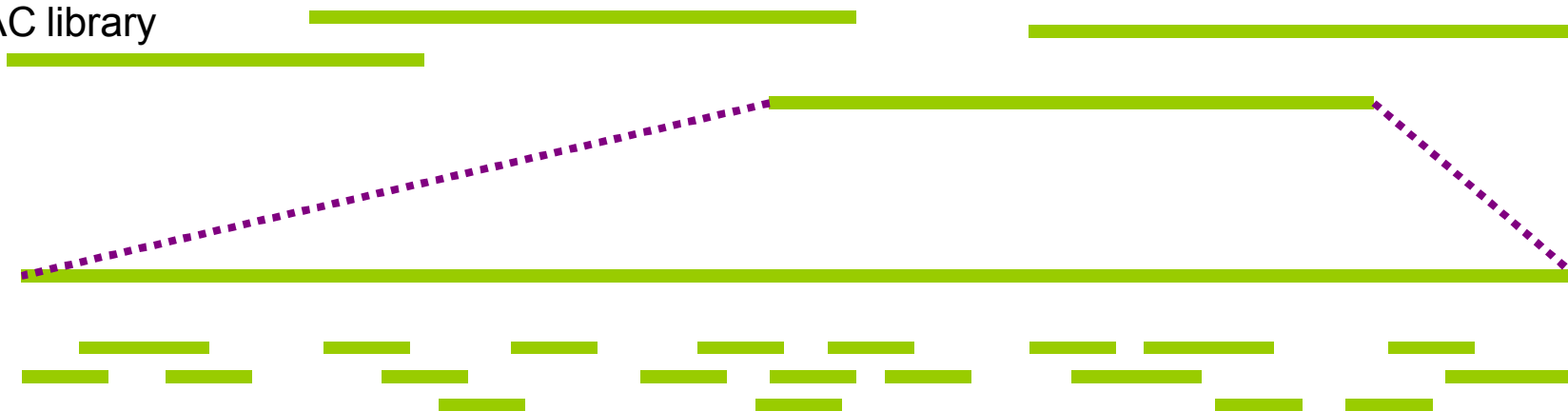
BAC-by-BAC sequencing

- Each BAC (Bacterial Artificial Chromosome) is about 150 kbp

- Covering the human genome requires ~30000 BACs

- BACs shotgun-sequenced separately

  – Number of repeats in each BAC is **significantly smaller** than in the whole genome...

  – ...needs **much more manual work** compared to whole-genome shotgun sequencing

Hybrid method

- Divide-and-conquer and whole-genome shotgun approaches can be combined
  - Obtain high coverage from whole-genome shotgun sequencing for short contigs
  - Generate of a set of BAC contigs with low coverage
  - Use BAC contigs to "bin" short contigs to correct places
- This approach was used to sequence the brown Norway rat genome in 2004

First whole-genome shotgun sequencing project: *Drosophila melanogaster*

- Fruit fly is a common *model organism* in biological studies
- Whole-genome assembly reported in Eugene Myers, *et al.*, A Whole-Genome Assembly of *Drosophila*, *Science* 24, 2000
- Genome size 120 Mbp

http://en.wikipedia.org/wiki/Drosophila_melanogaster

Sequencing of the Human Genome

- The (draft) human genome was published in 2001

- Two efforts:
  - Human Genome Project (public consortium)
  - Celera (private company)

- HGP: BAC-by-BAC approach

- Celera: whole-genome shotgun sequencing

HGP: Nature 15 February 2001
Vol 409 Number 6822
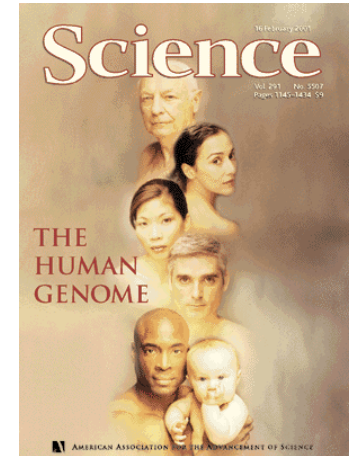
Celera: Science 16 February 2001
Vol 291, Issue 5507

Sequencing of the Human Genome

- The (draft) human genome was published in 2001

- Two efforts:
  - Human Genome Project (public consortium)
  - Celera (private company)

- HGP: BAC-by-BAC approach

- Celera: whole-genome shotgun sequencing

HGP: Nature 15 February 2001
Vol 409 Number 6822
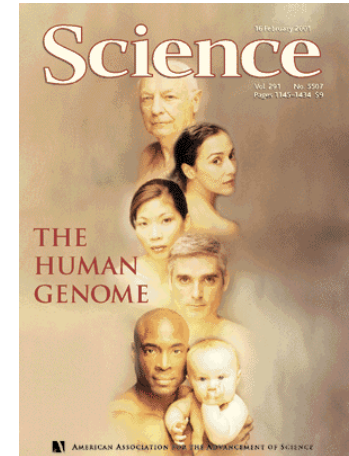
Celera: Science 16 February 2001
Vol 291, Issue 5507

Next-gen sequencing: 454

- Sanger sequencing is the prominent first-generation sequencing method
- Many new sequencing methods are emerging

- Genome Sequencer FLX (454 Life Science / Roche)
    - >100 Mb / 7.5 h run
    - Read length 250-300 bp
    - >99.5% accuracy / base in a single run
    - >99.99% accuracy / base in consensus

a

DNA library preparation

4.5 hours

Ligation

Selection
(isolate AB
fragments
only)

- Genome fragmented by nebulization
- No cloning; no colony picking
- sstDNA library created with adaptors
- A/B fragments selected using avidin-biotin purification

gDNA →→→ sstDNA library

b

Emulsion PCR

8 hours

Anneal sstDNA to an excess of DNA capture beads

Emulsify beads and PCR reagents in water-in-oil microreactors

Clonal amplification occurs inside microreactors

Break microreactors and enrich for DNA-positive beads

sstDNA library →→→ Bead-amplified sstDNA library

c

Sequencing

7.5 hours

- Well diameter: average of 44 μm
- 400,000 reads obtained in parallel
- A single cloned amplified sstDNA bead is deposited per well

Amplified sstDNA library beads →→→ Quality filtered bases

Mardis ER. 2008.
Annu. Rev. Genomics Hum. Genet. 9:387–402

The method used by the Roche/454 sequencer to amplify single-stranded DNA copies from a fragment library on agarose beads.
A mixture of DNA fragments with agarose beads containing complementary oligonucleotides to the adapters at the fragment ends are mixed in an approximately 1:1 ratio.
The mixture is encapsulated by vigorous vortexing into aqueous micelles that contain PCR reactants surrounded by oil, and pipetted into a 96-well microtiter plate for PCR amplification. The resulting beads are decorated with approximately 1 million copies of the original single-stranded fragment, which provides sufficient signal strength during the pyrosequencing reaction that follows to detect and record nucleotide incorporation events.

sstDNA, single-stranded template DNA.

Next-gen sequencing: Illumina Solexa

- Illumina / Solexa Genome Analyzer
  - Read length 35 - 50 bp
  - 1-2 Gb / 3-6 day run
  - > 98.5% accuracy / base in a single run
  - 99.99% accuracy / consensus with 3x coverage

a

Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Attach DNA to surface
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Denature the double stranded molecules

b

First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers, and DNA polymerase enzyme to the flow cell.

Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

Before initiating the next chemistry cycle
The blocked 3' terminus and the fluorophore from each incorporated base are removed.

Sequence read over multiple chemistry cycles
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at a time.

Mardis ER. 2008.
Annu. Rev. Genomics Hum. Genet. 9:387–402

The Illumina sequencing-by-synthesis approach. Cluster strands created by bridge amplification are primed and all four fluorescently labeled, 3′-OH blocked nucleotides are added to the flow cell with DNA polymerase. The cluster strands are extended by one nucleotide. Following the incorporation step, the unused nucleotides and DNA polymerase molecules are washed away, a scan buffer is added to the flow cell, and the optics system scans each lane of the flow cell by imaging units called tiles. Once imaging is completed, chemicals that effect cleavage of the fluorescent labels and the 3′-OH blocking groups are added to the flow cell, which prepares the cluster strands for another round of fluorescent nucleotide incorporation.

Next-gen sequencing: SOLiD

- SOLiD
    - Read length 25-30 bp
    - 1-2 Gb / 5-10 day run
    - >99.94% accuracy / base
    - >99.999% accuracy / consensus with 15x coverage

The ligase-mediated sequencing approach of the Applied Biosystems SOLiD sequencer. In a manner similar to Roche/454 emulsion PCR amplification, DNA fragments for SOLiD sequencing are amplified on the surfaces of 1-μm magnetic beads to provide sufficient signal during the sequencing reactions, and are then deposited onto a flow cell slide. Ligase-mediated sequencing begins by annealing a primer to the shared adapter sequences on each amplified fragment, and then DNA ligase is provided along with specific fluorescent-labeled 8mers, whose 4th and 5th bases are encoded by the attached fluorescent group. Each ligation step is followed by fluorescence detection, after which a regeneration step removes bases from the ligated 8mer (including the fluorescent group) and concomitantly prepares the extended primer for another round of ligation. (*b*) Principles of two-base encoding. Because each fluorescent group on a ligated 8mer identifies a two-base combination, the resulting sequence reads can be screened for base-calling errors versus true polymorphisms versus single base deletions by aligning the individual reads to a known high-quality reference sequence.