BIOL 4640 Study Guide for Exam 1 to be Given on 10/5/26

Filled in from Dr. Rowen's completed lectures only: his Complete notes for Lectures 2 and 4 to 11, Lectures 12, 13 and 14 (student decks with the blanks filled by Claude on 23 and 27 Sep, since his Complete notes are not posted yet), his Fall 2024 and Fall 2025 Exam 1 keys, and the assignments. Answers are in dark text under each of his lines; gray italics after a term give its everyday meaning; [L08 p12] means Lecture 8, slide 12; [L12 s8] means Lecture 12, slide 8; [Key Fa25 Q6] means question 6 of his Fall 2025 paper.

Key: ★ **On his past exams** ✎ **His worked example / model answer** *(gray italics) = plain meaning* [L08 p12] = Lecture 8, slide 12

Topic colors	Culturing, counting, growth	Microscopy	Genome reports and sequencing	Features of genomes
	Genome analysis and metagenomics	Replication and plasmids	Phylogeny and taxonomy	Genome evolution and gene transfer

In person, Closed notes. 60 min. Have a calculator in case you need to do some growth rate calculations.
Will cover topics covered in lecture and in assignments.
Chapter sections listed are from 6th edition of Microbiology: An Evolving Science

Unit 1. Methods used to Study Microorganisms

A. Methods for Culturing Microorganisms (4.3-4.4 in 6th ed)

• **Pure culture**: a single species or strain *(genetic variant of a species)* grown in or on media *(food mix for growing microbes)*. Growing microbes in the lab can be difficult because they vary in required growth conditions; most microbes in nature live in complex multispecies communities, and only about 0.1% of bacteria have been grown in the lab. [L02 p12]

1. Know different Types of media

• **Liquid (broth) medium**: nutrients and minerals dissolved in water, in a tube or flask, sterilized *(all microbes killed)* and covered; often shaken or rotated to keep it mixed and aerated *(supplied with air)*. All cells are subjected to the same conditions; initially food is abundant. [L02 p13]

• **Solid medium**: agar *(seaweed gel that sets solid)* (or another agent) added to broth. Cells placed in or on the medium form colonies *(visible cell piles, each from one cell)*, and cells within a colony experience different conditions. [L02 p14]

• **Complex medium**: nutrient rich but poorly defined (a digest of a protein). Tryptone is a digest of milk protein. [L02 p15, L02 p16]

• **Synthetic (defined) medium**: known composition, specific chemicals added. [L02 p15]

• **Enriched medium**: a complex medium to which additional components are added. [L02 p15]

• **Selective medium**: certain species can grow on it but other species cannot. [L02 p15]

• **Differential medium**: different types can grow, but you can tell them apart by metabolic *(food-processing chemistry)* differences (usually a color difference). [L02 p15]

• **MacConkey agar = complex, selective AND differential**. Complex: contains peptone (digest of gelatin). Selective: bile salts *(gut detergents that block many microbes)*; most Gram negatives *(bacteria with thin wall and outer membrane)* grow, most Gram positives *(bacteria with thick cell wall)* cannot. Differential: lactose fermenters *(microbes that break down milk sugar)* form red colonies, non-fermenters are unstained (whitish). [L02 p17]

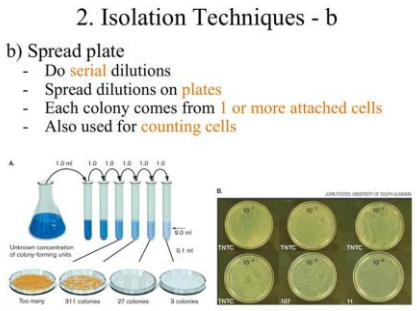
Medium (his Table 4.1)	Type	Key ingredients	Organisms cultured
Lysogeny (Luria Bertani) broth, LB	Complex	Tryptone 10 g, yeast extract 5 g, NaCl 10 g per liter, pH 7	Many Gram negatives and Gram positives (E. coli, Staphylococcus aureus)
M9 medium ***(simple salts plus glucose recipe)***	Defined	Glucose, Na2HPO4, KH2PO4, NH4Cl, NaCl, MgSO4, CaCl2, pH 7	Gram negatives such as E. coli
Sulfur oxidizer medium ***(feeds microbes that burn sulfur)***	Defined	NH4Cl, KH2PO4, MgSO4, CaCl2, elemental sulfur, 5% CO2, pH 3	Acidithiobacillus thiooxidans

Table source: [L02 p16]

2. Isolation techniques

• **Dilution streaking (streak to single colony)**: 1) pick up many cells on a loop *(small wire tool)* or stick; 2) drag it across the agar *(solid gel)* surface; 3) cells are put onto the surface; 4) as the loop is dragged, fewer cells per area are laid down (dilution); 5) change or sterilize the loop or stick; 6) eventually 1 cell (or 1 set of attached cells) is laid down in separate spots; 7) each forms an isolated colony *(cell pile grown from one cell)* separated from the others. [L02 p18, L02 p20]

• **Spread plate**: do serial dilutions *(repeated stepwise dilutions)*, spread the dilutions on plates; each colony comes from 1 or more attached cells. Also used for counting cells. [L02 p21]



Serial 10-fold dilutions plated: countable plates only at the higher dilutions (TNTC = too numerous to count) [L02 p21]

✎ **Getting a pure culture from a mixed sample (his model answer)**: (1) streak the sample to isolation *(spread until colonies sit apart)* on an agar plate, pick an isolated colony, and streak again to isolation or grow it in broth *(liquid growth medium)*; or (2) plate dilutions of the sample onto agar or in molten agar *(melted gel)*, pick an isolated colony, then streak it to isolation or inoculate *(add cells to)* a broth culture. [Key Fa25 Q33]

★ **On his past exams**: Fa25 Q33 (3 pts): describe one procedure to get a pure culture of one type of bacteria from a mixed water sample (streak to isolation, or plate dilutions, then pick an isolated colony and restreak). [Key Fa25 Q33]

B. Methods used to Count Microorganisms –

• **Why count**: to get the concentration of cells in a liquid (cells/ml). You only want to count live cells. Scientists want to use the same number or concentration of cells for an experiment or procedure (e.g. antibiotic susceptibility testing *(checking which drugs stop a microbe)*). [L02 p22]

✎ **His five approaches**: 1) direct count with microscope or machine; 2) viable plate count *(counting colonies from live cells)*; 3) turbidity *(cloudiness)* or optical density (OD); 4) measure a biochemical property; 5) calculation involving concentrations of cells. [L02 p23, L04 p3]

• **Microscope counting chamber (direct count)**: count the cells seen in a very small volume (10^{-8} ml) on a special slide with an inscribed grid. [L02 p24]

• **Live/dead microscope count**: stain live vs dead cells differentially with fluorescent dyes *(dyes that glow under special light)* (green = live, orange/red = dead), view with a fluorescence microscope, count; counting can be automated. [L02 p26]

• **Electronic counting (Coulter counter)**: cells in solution pass through a narrow opening and disrupt electrical conductivity *(how easily current flows)*, and the machine counts them. Not all counters can count bacterial cells (smaller). Does not distinguish live vs dead. [L02 p27]

• **FACS (fluorescence-activated cell sorting)**: cells usually labeled with a fluorescent dye pass through a nozzle one at a time; light hits each cell, a detector records a value. Counts cells, measures a few properties of each cell (e.g. GFP *(green glowing protein)* expression and size), and can sort cells by a property. [L02 p28, L02 p29]

• **Viable plate count**: a) make serial dilutions (usually 10-fold); b) place aliquots *(small measured portions)* on or in agar; c) incubate *(keep warm so cells grow)* until colonies appear; d) count colonies; e) calculate CFU/ml (colony forming units per ml). Two plating methods: **spread plate** (spread on the surface) and **pour plate** (dilution mixed with a small amount of molten agar and poured, so most cells are within the agar). [L02 p30, L02 p31]

• **Clumping lowers the plate count**: bacteria that stay attached after division form one colony per cluster, so CFU/ml is LOWER than the actual number of viable cells. [L02 p32]

• **Turbidity / optical density (OD)**: a spectrophotometer *(light-measuring instrument)* measures cloudiness; cells block passage of light and reduce transmittance *(light passing through)*. Typically absorbance *(light blocked by the sample)* at 550-650 nm *(nanometers, a light wavelength unit)* of a 1 ml sample. Absorbance correlates with **cell mass/ml**, not cell number. To get cells/ml you need a conversion factor (1 OD/ml = y cells/ml) from a standard curve *(reference graph from known samples)* of OD/ml vs cells/ml, and you may need a different factor for each species or strain *(genetic variant of a species)*. [L02 p34, L02 p35]

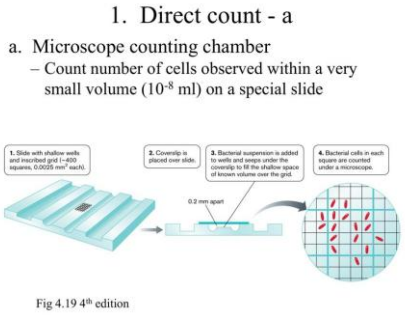
• **Cell size changes the OD conversion**: if strain A cells are twice as large as strain B cells, 1 OD/ml of A has LESS (close to 2-fold less) cells/ml than 1 OD/ml of B, because each A cell blocks about 2X the light. [L02 p36]

• **Biochemical assay**: measure a property that correlates with cell number: protein content, weight of biomass *(total cell material)*, metabolic rate *(speed of cell chemistry)*. [L02 p38]

Advantage and disadvantages of 3 methods described

✎ **His model answer for 'describe a method to count cells in 1 ml' (any one, with equipment)**: 1) direct count: sample on a special slide with boxes, count cells in the boxes, calculate cells/ml; 2) viable plate count: serial dilutions *(repeated stepwise dilutions)*, plate aliquots *(small measured portions)* into molten agar *(melted gel)* plates, count colonies after incubation, calculate CFU/ml *(colony forming units per ml)*; 3) optical density with a spectrophotometer *(light-measuring instrument)*: with a known OD-to-cells/ml conversion factor, take an absorbance reading *(light blocked by the sample)* and convert. [Key Fa24 Q32]

★ **On his past exams**: Fa25 Q18: OD measurement reflects the amount of cell mass in the sample (D). Fa25 Q20: NOT a feature of a direct microscope count = 'counts the cell mass, not the number of cells' (B). Fa24 Q32 (3 pts): describe one counting method with its essential equipment. [Key Fa25 Q18, Key Fa25 Q20, Key Fa24 Q32]



Microscope counting chamber: cells counted in grid squares of a known small volume [L02 p24]

Method	Pros	Cons
Direct count (microscope counting chamber)	Quick (minutes); moderate equipment (special slide + standard microscope); counts all cells in an attached cluster; does not require growing the cells	May count dead cells along with live; needs high cell density (> 10^6/ml); can be inaccurate (only a small aliquot is examined)
Viable plate count *(counting colonies from live cells)*	No expensive equipment; measures only living cells	Takes hours to days (at least 18 hours) for colonies; count can be low due to clumping or attached cells; some species or cells may not grow on solid media
Turbidity / optical density *(cloudiness reading)*	Fast, easy, simple, inexpensive	Indirect measure of cell number (measures cell mass); needs a conversion factor to get cells/ml; inaccurate at low (OD < 0.02) and high (> 1) concentrations (dilute high ones); detects both live and dead cells

Table source: [L02 p25, L02 p33, L02 p37, Key Fa25 Q20]

C. Measuring Growth and Growth cycle

• **Measuring growth**: use the cell counting methods to measure the change in cell number or cell mass over time. Topics: features of growth, batch culture *(one closed flask, no refills)*, continuous culture *(culture fed fresh medium nonstop)*, growth rate, factors affecting growth rate. [L04 p9]

• **Growth curve**: monitor cell number or cell mass (OD) over time; lets you compare the growth of two strains. In his example (Klett units / OD over 35 h), organism A grew faster than B (steeper rise, reaches the plateau *(flat top of the curve)* sooner). [L04 p26, L04 p27]

• **Growth rate is determined in exponential phase**, when all cells are actively growing at their maximal rate for their conditions. [L04 p25]

★ **On his past exams**: Fa25 Q31 (3 pts): draw a typical batch culture growth curve and label the phases (1.5 pts graph, 1.5 pts phases). See paragraph 13. [Key Fa25 Q31]

1. Features of Growth of Microbes

• **Growth usually leads to cell division**: only a small change in cell size (about 2-fold), an increase in cell number; most grow by **binary fission** *(splitting in two)*. [L04 p10]

• **Growth is exponential**: cell number (and usually cell mass) doubles with each round of division; plot on a log scale *(axis in powers of ten)*. [L04 p11]

• **Fast**: most can grow and divide rapidly; some divide in as little as 15-20 minutes; many seem to consume nutrients and multiply as fast as possible. [L04 p12]

• **Very high densities**: millions or billions of cells/ml. **They stop growing when they run out of nutrients.** [L04 p12]

• **Uses of growth rate**: indicator of the effect of conditions on growth; finding optimal growth conditions; indicator of fitness *(ability to survive and multiply)* (compare two mutants *(strains with changed genes)*); indicator of physiological *(cell function)* changes in cells. [L04 p13]

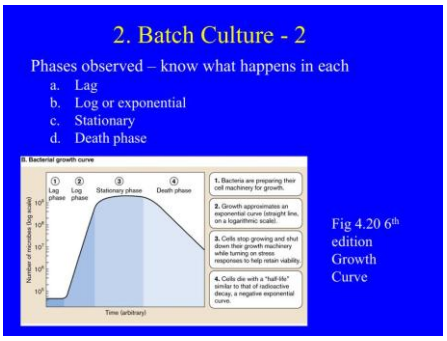
★ **On his past exams**: Fa24 Q23: NOT a feature of bacterial growth = 'most grow by producing 4 daughter cells' (B). Fa25 Q25: NOT a feature = 'a linear increase in cell number' (E); growth is exponential. [Key Fa24 Q23, Key Fa25 Q25]

2. Stages of growth in batch culture vs continuous culture

• **Batch culture**: most common lab method; cells added to liquid media in a container. **Closed system**, no inflow of nutrients. Four phases, and cell physiology can differ in each. [L04 p14]

• **Continuous (steady state) culture**: cells grown in a **chemostat** *(culture vessel fed fresh medium nonstop)*. System is NOT closed: fresh media added and old media plus cells removed. Volume, cell number and nutrient levels kept constant; cells maintained in log phase *(steady doubling stage)*. After reaching steady state, cell number is constant and cells stay in exponential phase with balanced growth *(all cell parts made in step)*. [L04 p16, L04 p20]

• **Apparatus examples**: 500 ml chemostat; now microfluidic devices *(tiny channel chips)* or chemostats with 1 ml chambers; flow cells (media pumped over a surface with attached bacteria) to observe biofilms. **Biofilm**: community of bacterial cells in an extracellular matrix *(slime outside the cells)* attached to a surface bathed by water. [L04 p17, L04 p18, L04 p19]



Batch culture growth curve: lag, log, stationary and death phases (log scale y axis) [L04 p15]

Phase	What happens (his Fig 4.20)
1. Lag *(getting ready to grow)*	Bacteria prepare their cell machinery for growth; little change in number
2. Log (exponential)	Growth approximates an exponential curve (straight line on a log scale); growth rate is measured here
3. Stationary *(holding steady)*	Cells stop growing and shut down their growth machinery while turning on stress responses to help retain viability
4. Death	Cells die with a 'half-life' like radioactive decay (negative exponential curve)

Table source: [L04 p15, L04 p25]

	Batch culture	Continuous culture
System	Closed, no inflow of nutrients	Open: fresh media in, old media and cells out
Vessel	Container / flask of liquid media	Chemostat (500 ml or 1 ml chambers), flow cell
Phases	Lag, log, stationary, death	Held in log (exponential) phase at steady state
Cell number and nutrients	Change over time; growth stops when nutrients run out	Kept constant
Big advantage	Most common, simple	Cells can be maintained in log phase growth (balanced growth)

Table source: [L04 p14, L04 p16, L04 p20, Key Fa24 Q24]

Location in nature	Open or closed
Puddle of water	Closed
Animal intestine	Open (nutrients flow in, waste flows out)
Living on a piece of fruit	Closed
Flowing stream	Open
Bloodstream of an animal	Open
Partially drunk can of pop	Closed

Table source: [L04 p22, L04 p23]

Calculations for Cells/ml, OD and dilution of cultures

• **Cells/ml and OD600/ml are concentrations** (like molarity). To dilute from a high to a lower concentration, solve like a chemistry dilution: **C1 x V1 = C2 x V2**, usually solving for V1 (how much of the concentrated culture to add). [L04 p4, L06 p4]

✎ **His OD dilution example**: stock (overnight) culture OD600 *(cloudiness read with 600 nm light)* = 2.0 OD/ml; you want 5 ml at 0.1 OD/ml. C1 = 2.0, C2 = 0.1, V2 = 5 ml. V1 = (0.1 OD/ml x 5 ml) / 2.0 OD/ml = **0.25 ml = 250 ul** *(microliters, thousandths of a ml)*. [L04 p5, L06 p5]

• **OD to cells/ml**: needs a previously determined conversion factor (measure OD600 and get cells/ml by microscope count or viable plate count *(counting colonies from live cells)*), e.g. 1 OD600/ml = 2 x 10^8 cells/ml. Multiply: OD/ml of sample x (2 x 10^8 cells/ml per OD/ml) = cells/ml. [L04 p6, L06 p6]

✎ **His OD conversion example**: factor 2.0 x 10^8 cells/ml per 1 OD600/ml; OD600 = 0.5. Cells/ml = 0.5 x 2.0 x 10^8 = **1 x 10^8 cells/ml**. Once you have the concentration you can dilute to a specific concentration or add a specific number of cells. [L04 p7, L06 p7]

✎ **Assignment 4 Q1 (his numbers)**: culture A OD600 = 1.8; make 10 ml at 0.01 OD600/ml. V1 = (0.01 x 10 ml) / 1.8 = 0.0556 ml = **about 56 ul**. [A04]

✎ **Assignment 4 Q2 (his numbers)**: make 5 ml at 1 x 10^6 cells/ml from culture B at 5.5 x 10^7 cells/ml. V1 = (1 x 10^6 x 5 ml) / 5.5 x 10^7 = 0.0909 ml = **about 91 ul**. His calculator tip: enter 5.5 x 10^7 as 5.5 [EXP] 7, not the 10. [A04, L06 p3]

✎ **Assignment 4 Q3 (his numbers)**: add 6 x 10^6 cells; strain C OD600 = 0.6 OD/ml; 1.0 OD600/ml = 2 x 10^8 cells/ml. Step 1: 0.6 x 2 x 10^8 = 1.2 x 10^8 cells/ml. Step 2: cells wanted / concentration = 6 x 10^6 / 1.2 x 10^8 = 0.05 ml = **50 ul**. [A04]

• **CFU/ml** *(colony forming units per ml)* **from a viable plate count**: CFU/ml = colony count x (1 / amount plated) x (1 / dilution). Use only the dilution whose plates have 20-250 colonies, and use the average of that dilution. [A04]

✎ **Assignment 4 Q6 (his data)**: 10-fold serial dilutions, 0.1 ml plated in duplicate *(two plates each)*. 10^{-4} TNTC *(too numerous to count)*; 10^{-5} 276 and 284 (over 250); 10^{-6} 34 and 28 (average 31); 10^{-7} 2 and 5 (under 20). Use 10^{-6}: CFU/ml = 31 x (1/0.1) x (1/10^{-6}) = **3.1 x 10^8 CFU/ml.** [A04]

★ **On his past exams**: Fa24 Q32 (3 pts): the OD route works only if you already know the OD-to-cells/ml conversion factor. Fa25 Q18: OD reflects cell mass. [Key Fa24 Q32, Key Fa25 Q18]

3. Growth rate calculations – know symbols and how to calculate

• **Growth rate**: a quantitative measure of how fast a microbe grows, often reported as the rate of doubling of cell number or cell mass. Used to compare conditions or strains and as a measure of fitness *(ability to survive and multiply)*. Measure it in **exponential phase** when all cells grow at their maximal rate. [L04 p25]

• **Where n comes from**: in log phase *(steady doubling stage)* Nt = N0 x 2^n, so n = log2(Nt/N0). In base 10: log Nt = log N0 + n x log 2, and log 2 = 0.301, so n = (1/0.301) x (log Nt - log N0) = 3.3 x log10(Nt/N0). [L04 p30, L06 p8]

• **Direction**: shorter (smaller) generation time g = faster growth. Larger growth rate constant k = faster growth. [L04 p31, L04 p32, L04 p33]

✎ **His k example**: n = 2 generations, t = 4 hours: k = 2/4 = **0.5 generations/hour** (so g = 2 hours). [L04 p33, L06 p11]

Symbol	Meaning	Formula
N0	Starting cell number, cell mass or OD (time 0)	measured
Nt	Ending cell number, cell mass or OD	measured
t	Time (usually hours)	measured
n	Number of generations (doublings)	n = log(Nt/N0) / 0.301 = 3.3 x log(Nt/N0)
g	Generation (doubling) time: time for the population to double	g = t / n = t x 0.301 / log(Nt/N0)
k	Growth rate constant: generations per unit time (usually per hour), reciprocal of g	k = n / t = 1 / g

Table source: [L04 p29, L04 p30, L04 p31, L04 p33]

I will give the basic equations ($g = t/n \quad k = n/t \quad n = \log(N_t/N_0)/0.301$)

✎ **Worked example, his Fa25 key (Q32)**: N0 = 1.4 x 10^7 cells/ml, Nt = 5.5 x 10^7 cells/ml, t = 90 min = **1.5 hr** (convert to hours). n = log(5.5 x 10^7 / 1.4 x 10^7) / 0.301 = log(3.9) / 0.301 = 0.59 / 0.301 = **1.95 generations**. k = n / t = 1.95 / 1.5 hr = **1.3 gen/hr** *(generations per hour)*. (g = t / n = 1.5 / 1.95 = about 0.77 hr, not asked on the key.) [Key Fa25 Q32, L04 p31]

✎ **Worked example, his Fa24 key (Q30)**: N0 = 1.2 x 10^7 cells/ml, Nt = 4.8 x 10^7 cells/ml, t = 3 hr. n = log(4.8 x 10^7 / 1.2 x 10^7) / 0.301 = (7.68 - 7.07) / 0.301 = **2 generations** (the ratio is 4 = two doublings). k = 2.0 / 3 hr = **0.66 gen/hr** on his key (0.67 rounded). g = 3 / 2 = 1.5 hr. [Key Fa24 Q30, L04 p31]

✎ **Assignment 4 Q4 (his prompt)**: one cell to 1 x 10^6 cells (about a visible colony), dividing every 30 min. n = 3.3 x log(10^6/1) = 3.3 x 6 = **about 20 generations**; t = n x g = 20 x 0.5 hr = **10 hours**. [A04, L04 p30]

✎ **Assignment 4 Q5 (his data), t = 1.5 hr**: Culture A 1.0 x 10^7 to 2.5 x 10^7: n = 3.3 x log(2.5) = 1.31; g = 1.5/1.31 = 1.14 hr; k = 1.31/1.5 = 0.88 gen/hr. Culture B 2.0 x 10^7 to 8.0 x 10^7: n = 3.3 x log(4) = 1.99 (2 doublings); g = 1.5/1.99 = 0.75 hr; k = 1.99/1.5 = 1.33 gen/hr. **B grows faster** (smaller g, larger k). [A04, L04 p30, L04 p31, L04 p33]

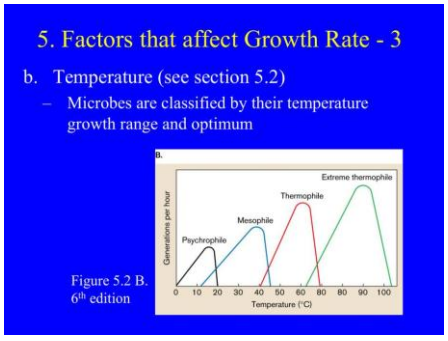
• **Exam traps**: convert minutes to hours before k (90 min = 1.5 hr); n uses the RATIO Nt/N0 (the 10^7 cancels); enter exponents with the EXP key. [Key Fa25 Q32, L06 p3, A04]

★ **On his past exams**: Fa25 Q32 (4 pts): n = 1.95 gen, k = 1.3 gen/hr. Fa24 Q30 (4 pts): n = 2 gen, k = 0.66 gen/hr. The equations g = t/n, k = n/t, n = log(Nt/N0)/0.301 are printed on the exam. [Key Fa25 Q32, Key Fa24 Q30]

4. Factors that affect maximal Growth rate

• **Growth rate is mainly regulated by environmental conditions**: a) availability of nutrients (type and level); b) temperature; c) oxygen availability; d) pressure; e) osmotic (salt) levels; f) pH. These parameters are also used to classify and describe microbes. [L04 p34, L04 p35]

• **Temperature**: microbes grow only within a certain range and fastest at their optimum *(best temperature)*. Minimum is set by membrane gelling *(membrane stiffening in the cold)* (transport too slow); at the optimum enzymatic reactions *(protein-driven chemistry)* run at their maximal rate; above it, proteins denature *(lose their shape)* and stop functioning (collapse of the membrane, thermal lysis). Microbes are classified by their growth range and optimum; the curve rises slowly to the optimum then drops steeply. [L04 p28, L04 p36, L04 p37]



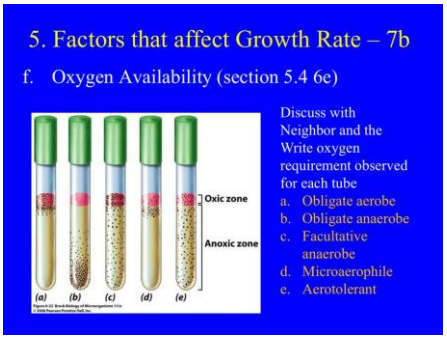
Growth rate (generations per hour) vs temperature for psychrophile, mesophile, thermophile, extreme thermophile [L04 p36]

• **Nutrients change the maximal rate**: on his Fig 5.2A (1000/T vs k, left = hotter), growth in rich media is faster than in minimal media *(bare-bones food mix)* at the same temperature. [L04 p37]

• **Oxygen classes**: (obligate) aerobe requires O2; **facultative anaerobe** grows with or without O2; (obligate) anaerobe grows only without O2; microaerophile grows in low O2; aerotolerant anaerobe does not use O2 but grows in its presence. [L04 p38]

• **Standing test tube (no agitation)**: O2 is highest at the top and falls with depth. Growth pattern: aerobe at the top; anaerobe at the bottom; facultative throughout (densest at top); microaerophile in a band below the top; aerotolerant evenly throughout. [L04 p39, L04 p40, L04 p41]

★ **On his past exams**: Fa24 Q21: grows with or without oxygen and uses it when available = facultative anaerobe (C). [Key Fa24 Q21]



Oxygen requirement read from growth position in standing tubes (answers a-e) [L04 p41]

Parameter	Classes (optimal growth condition, his Table 5.1)
Temperature	Hyperthermophile (above 80 C); thermophile (50-80 C); mesophile (15-45 C); psychrophile (below 15 C)
pH *(acidity scale)*	Alkaliphile (above pH 9); neutralophile (pH 5-8); acidophile (below pH 3)
Osmolarity *(dissolved salt level)*	Halophile (high salt, > 2 M NaCl); halotolerant (high salt not required, can grow up to 2 M NaCl)
Oxygen	Strict aerobe (only in O2); facultative (with or without O2); microaerophile (only small amounts of O2); strict anaerobe (only without O2)
Pressure	Barophile (> 380 atm); barotolerant (10-500 atm)

Table source: [L04 p35]

D. Methods Used to Observe Microbes – Microscopy (2.1-2.6 of 6th ed)

• **Why a microscope**: most microbes are too small for the naked eye. The human eye cannot resolve *(tell two close points apart)* below **150 µm** *(micrometers: millionths of a meter)*. [L05 p3]

• **Three microscope technologies**: (a) light, (b) electron *(beam of electrons, not light)*, (c) atomic force *(a tiny tip traces the surface)*. [L05 p4, L05 p30]

• **What a microscope improves**: **magnification** (apparent size), **resolution** (smallest distance between two objects that can still be distinguished), **detection** (ability to see an object at all). [L05 p5]

• **Resolution matters more than magnification**: his slide shows two images at the same magnification, one made with light and one with electrons; the electron image shows far more because its resolution is better. [L05 p6]

Microbe	Size (his numbers)
Eukaryotic microbes *(nucleus-containing microbes)*	10-100 µm wide (1 µm = 1 micron = 10^-6 m)
Bacteria	0.4-10 µm
Viruses	0.01-0.1 µm, i.e. 10-100 nm

Table source: [L05 p3]

1. Features of Light microscopy

• **Standard light microscopy**: uses light to see objects. Typical maximum magnification **about 2,000X**. Limit of resolution *(smallest gap still seen as two)* **0.2 µm (200 nm)**. [L05 p7]

• **What you can and cannot see**: the shape of bacterial cells (about 1 µm) but not the details of their structure; some organelles *(membrane-bound compartments inside cells)* inside eukaryotic *(nucleus-containing)* cells (most over 10 µm). [L05 p7]

• **Biggest factor limiting resolution with light**: the width of the light beam (wavelength). [L05 p7]

• **Bright-field appearance**: cells usually appear dark on a white to gray background. This is the microscope used to view Gram-stained *(dyed by the Gram method)* slides. [L05 p9, L05 p28, A03]

• **Parts of the compound light microscope**: eyepiece (ocular lens), objective lens *(lens nearest the sample)*, stage *(platform holding the slide)*, **condenser** (concentrates light), diaphragm *(adjustable opening that controls light)*, light source. [L05 p10]

• **Total magnification** = power of the objective (4-100X) x power of the eyepiece (10X). [L05 p10]

• **Shape terms to know**: **coccus** = sphere, **bacillus** = rod, **spiral**. Some grow in clusters or filaments *(long threads)* (details under Staining below). [L05 p8]

★ **On his past exams**: Fa25 Q21 (Gram-stained slides are viewed with bright-field light microscopy); Fa24 Q26 and Fa25 Q26 (read shape from the stained slide: spheres = coccus, rods = bacillus). [Key Fa25 Q21, Key Fa24 Q26, Key Fa25 Q26]

2. Factors limiting Resolution with light

• **a) Light is a wave, and the width of the wave limits resolution**: the wave must fit between two items to resolve *(tell apart)* them. Wavelength *(distance between wave peaks)* of light 0.4-0.7 µm, so the limit of resolution is 0.2 µm. [L05 p11, L05 p12]
• **b) White light is a mix of waves of different lengths** (0.4-0.7 µm). [L05 p11]
• **c) Light is refracted (bent)**: happens when light goes from water or glass into air, and different wavelengths refract differently. This is the problem immersion oil *(oil between lens and slide)* fixes. [L05 p13, L05 p16]
• **d) Lack of contrast**: many microbes look translucent *(nearly see-through)* in ordinary bright-field *(lit from below, cells dark on light)* microscopy. Phase contrast *(lens system that boosts contrast)* gives improved contrast of the same cells. [L05 p14]

★ **On his past exams**: Fa24 Q25 and Fa25 Q24 both ask which light problem oil corrects: refraction (light is bent). Wrong choices are the other limits: wavelength width, mixed wavelengths, translucent cells lacking contrast. [Key Fa24 Q25, Key Fa25 Q24]

3. Advances in Microscopy –

• **Advances in light microscopy (his list)**: (a) use of oil, (b) additional types of light microscopy (phase contrast, differential interference, dark field, fluorescent, confocal, super-resolution), (c) staining. [L05 p15]
• **Immersion oil**: placed between the glass slide and the microscope lens. Use oil above **400X**. It **reduces refraction** *(bending of light)*. [L05 p16]

★ **On his past exams**: oil reduces refraction (Fa24 Q25, Fa25 Q24); pink spheres after Gram stain = Gram-negative *(thin wall, loses the purple)* coccus (Fa24 Q26); purple rods about 2 µm = Gram-positive *(thick wall, keeps the purple)* bacillus (Fa25 Q26); view Gram stains with bright-field (Fa25 Q21); to see the location and shape of an internal compartment in bacteria use confocal fluorescent *(laser imaging of stacked focal planes)*, super-resolution *(sharper-than-normal light imaging)* or TEM *(electron beam through thin slices)* (SEM also accepted) (Fa24 Q33, short answer, 3 pts). [Key Fa24 Q25, Key Fa25 Q24, Key Fa24 Q26, Key Fa25 Q26, Key Fa25 Q21, Key Fa24 Q33]

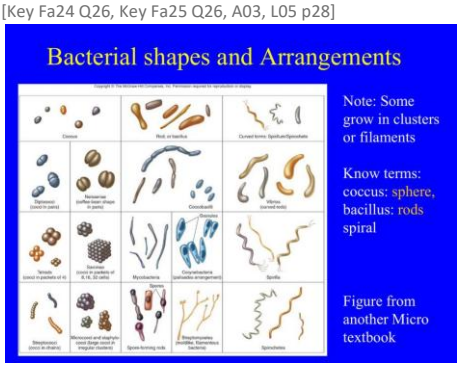
Different types of Light microscopy

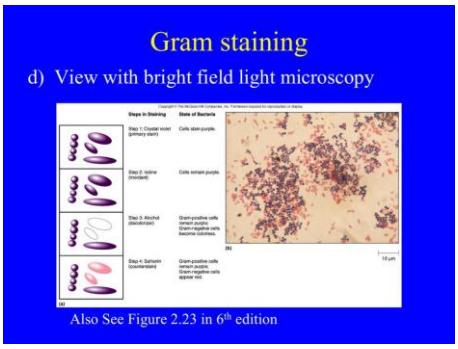
• **Phase contrast**: uses a special lens (special optical system) that improves contrast so live, unstained cells can be observed. [L05 p17, L05 p14, A03]
• **Differential interference microscopy (DIC)**: gives a 3D look; like phase contrast, it improves contrast to allow observation of live cells. [L05 p17]
• **Dark-field microscopy**: special type of light microscope; organisms appear **bright on a dark (black) background** (opposite of bright field). [L05 p18, A03]
• **Fluorescent, confocal and super-resolution** are also light microscopy types (see the next two lines). [L05 p15]

✎ **His panel exercise (protozoan cysts that cause diarrhea)**: panels A, C, F = stained bright-field (colored cells on light background); B, D, G = phase contrast (gray, unstained, high contrast); E, H = fluorescent (glowing on dark background). [L05 p38, L05 p39, A03]

Staining – Gram stain, also common shapes and arrangements

• **Staining**: chemicals that bind to cellular structures to give them color. [L05 p25]
• **Gram stain**: a differential stain that distinguishes bacteria by the nature of their cell wall (cell envelope). [L05 p26]
• **Gram stain steps**: (a) prepare a smear *(thin film of cells on a slide)*, (b) fix *(kill and stick cells to the slide)* with methanol or heat, (c) stain: 1) **crystal violet** (primary stain, all cells purple), 2) **iodine** (mordant, complex with crystal violet increases retention, cells stay purple), 3) **decolorize with ethanol** (hold slide at an angle, 15-20 sec; Gram-positive stay purple, Gram-negative become colorless), 4) **counterstain with safranin** (Gram-negative turn red/pink). Wash with water after each step. (d) View with bright-field light microscopy. [L05 p26, L05 p27, L05 p28]
• **Reading the result**: **purple = Gram-positive**, **red/pink = Gram-negative**. [L05 p28, L05 p29]
• **Exam pattern**: stain color gives the Gram type, shape gives the name. Pink spheres about 1 µm = **Gram-negative coccus** (Fa24 Q26). Purple rods about 2 µm = **Gram-positive bacillus** (Fa25 Q26). Purple circles = Gram-positive coccus (A03 Q11). [Key Fa24 Q26, Key Fa25 Q26, A03, L05 p28]





Bacterial shapes and arrangements: coccus = sphere, bacillus = rod, spiral [L05 p08]

Gram stain: crystal violet, iodine, alcohol, safranin; purple Gram-positive cocci and red Gram-negative rods [L05 p28]

Stain type	What it is	His examples
Simple	Uses only one stain	Crystal violet stain of E. coli; methylene blue stain of Corynebacterium
Differential	Two or more stains that react differently	Gram stain; acid-fast (red = acid-fast, blue = non-acid-fast)
Special	Stains a specific cellular structure or type of cell	Flagellum; endospore (spores red, vegetative cells blue); India ink capsule stain
Fluorescent	Uses a fluorophore	Fluorescent microscopy

Table source: [L05 p25, L05 p29]

Term	Meaning
Coccus	Sphere
Bacillus	Rod
Curved forms	Spirillum / spirochete; vibrios = curved rods
Diplococci	Cocci in pairs (Neisseriae: coffee-bean shape in pairs)
Streptococci	Cocci in chains
Tetrads	Cocci in packets of 4
Sarcinae	Cocci in packets of 8, 16, 32 cells
Micrococci and staphylococci	Large cocci in irregular clusters
Corynebacteria ***(club-shaped rods)***	Palisades arrangement
Streptomycetes	Moldlike, filamentous bacteria
Spore-forming rods	Rods carrying spores

Table source: [L05 p8]

Fluorescence microscopy – confocal microscopy

- **Fluorescent microscopy**: stain with a fluorescent dye (**fluorophore**), illuminate with a specific wavelength *(color)* of light, detect at a different wavelength. You see fluorescent color on a dark background. [L05 p19]
- **Excitation and emission**: each fluorophore absorbs light (excitation) at one wavelength and emits light at a second, **longer** wavelength. Many fluorophores exist (his chart: DAPI, CFP, GFP, FITC, YFP, rhodamine B, Cy3, Alexa 568, RFP, Cy5). [L05 p20]
- **Protein localization in live cells**: (1) express a hybrid protein *(two proteins fused into one)* with **GFP** (green fluorescent protein) attached (e.g. a microtubule-associated protein tagged with GFP), or (2) attach a fluorophore to an **antibody** *(immune protein that binds one target)* or chemical that binds specific structures (e.g. fluorescent antibodies against the cytoskeleton). [L05 p21]
- **Also used for counting**: live-dead microscope count stains live and dead cells differently with fluorescent dyes (green live, orange/red dead) and is viewed with a fluorescent microscope. [L02 p26]
- **Confocal microscopy**: advanced fluorescent microscopy that uses **lasers** and can view multiple planes. Takes multiple images at different depths and a computer stacks them into a **3D image** (his example: 3D shape of bacterial biofilms). [L05 p22, L05 p23]

> ★ **On his past exams**: confocal fluorescent microscopy is a keyed answer for seeing the location and shape of an internal compartment in bacterial cells (Fa24 Q33). [Key Fa24 Q33]

Super resolution microscopy

- **Super-resolution light microscopy**: improves the resolution of fluorescent light microscopy by **about 10-fold**. There are a couple of different technologies. [L05 p24]
- **STORM (one way)**: only some of the fluorescent probes *(glowing tag molecules)* emit light at any one time; take multiple images and then overlay them (combines many images of individual fluorescent molecules). [L05 p24, A03]

> ★ **On his past exams**: super-resolution microscopy is a keyed answer for seeing the location and shape of an internal compartment in bacteria (Fa24 Q33). [Key Fa24 Q33]

4. Electron microscopy – types and uses

- **Electron microscopy (EM)**: uses a beam of electrons instead of light. Better resolution because electrons have a much smaller wavelength (**0.004 nm**) than light (**400-700 nm**). Shows internal structures of cells and virus particles. Only works on **fixed** *(killed and preserved)* cells, more work and more expensive. Many types. [L05 p31]
- **Transmission EM (TEM)**: resolution **1 nm**, **1,000,000X**. Specimens stained with heavy atoms so structures are electron dense *(blocks electrons, looks dark)*; electrons pass through the sample. Used to study cell structures and the shape of virus particles. [L05 p32, L05 p34]
- **TEM preparation methods**: (a) **positive staining**: cut into thin slices, stain with a heavy metal that binds certain material (dark areas); can also use a specific antibody with a gold particle attached. (b) **metal shadowing** *(metal sprayed at an angle to cast shadows)*. (c) **negative staining** *(stain the background, not the object)* (e.g. influenza virus particles). (d)

freeze fracture *(freeze, then crack the sample open)* (his interpretation drawing shows proteins sitting in the phospholipid layers). [L05 p32, L05 p33, L05 p34]

• **Scanning EM (SEM)**: resolution limited to about **10 nm**, **100,000X**. Cell surface coated with a heavy metal, electron beam scans the surface, scattered electrons are collected to make a **3D image of the surface** (his examples: colorized E. coli, macrophage). [L05 p35]

• **Cryo-electron microscopy (cryo-EM)**: flash freeze the sample so structures keep their shape; **no metal staining**. Tomography *(combining many tilted views into 3D)* gives 3D images of the shape of protein complexes and virus particles (his example: bacterial flagellar motors). [L05 p36]

• **Atomic force (scanning probe) microscopy**: a tip moves across the surface, is deflected *(pushed aside)* by objects on the slide, and this builds the image. Visualizes the shape of whole **live** cells or large molecules (protein complexes); surface topography *(surface shape)* without staining. [L05 p37, A03]

★ **On his past exams**: Fa25 Q22 (view viruses and fine internal structures of cells with electron microscopy); Fa24 Q33 (TEM keyed for an internal compartment; SEM also accepted "as now can be used that way"). [Key Fa25 Q22, Key Fa24 Q33]

Microscope	What it shows	How contrast / the image is made	Resolution, magnification	Typical use
Bright-field light *(ordinary lit-from-below view)*	Cell shape, not bacterial detail	Cells dark on white to gray background; stains add color	0.2 µm; about 2,000X max	Gram-stained slides
Phase contrast	Live, unstained cells	Special lens improves contrast	Light limit	Watching live cells without stain
Differential interference (DIC)	Live cells with 3D look	Improves contrast	Light limit	Live cells
Dark-field	Organisms bright on dark background	Special light microscope	Light limit	Not given beyond bright-on-dark viewing
Fluorescent	Specific structures or proteins	Fluorophore excited at one wavelength emits a longer one; color on dark background	Light limit	Protein localization (GFP, antibodies); live-dead counts
Confocal	Multiple planes stacked into 3D	Fluorescence plus lasers and optical sections	Light limit	3D shape of biofilms; internal compartments
Super-resolution (e.g. STORM)	Finer fluorescent detail	Only some probes emit at once; many images overlaid	About 10-fold better than fluorescence	Internal compartments
TEM	Internal structures, virus shape	Heavy metal staining, electrons pass through thin sections	1 nm; 1,000,000X	Cell structure, viruses; fixed cells only
SEM	3D cell surface	Heavy metal coat, beam scans, scattered electrons collected	About 10 nm; 100,000X	Surfaces of cells
Cryo-EM	3D shape of protein complexes and viruses	Flash frozen, no metal stain, tomography	Not given	Natural-shape 3D reconstructions
Atomic force (scanning probe)	Shape of whole live cells, large molecules	Tip deflected by objects as it moves across surface	Not given	Live cells, protein complexes, no stain

Table source: [L05 p7, L05 p9, L05 p17, L05 p18, L05 p19, L05 p21, L05 p22, L05 p23, L05 p24, L05 p28, L05 p31, L05 p32, L05 p35, L05 p36, L05 p37, L02 p26, A03]

Unit 2. Bacterial Genomes and Genetics

A. Genome Reports –

• **Genome report**: a published article reporting a genome sequence. Read it for two things: how the genome was sequenced, and what properties of the genome were reported and what they mean. [L06 p13]

• **Basic steps in sequencing a genome**: (a) extract DNA, (b) fragment, (c) tag ends and create a library *(collection of DNA pieces)* of fragments, (d) may amplify *(make many copies)*, (e) sequence millions of different fragments, getting a read of bases on one strand from each, (f) assemble reads into contigs *(continuous stretches of assembled sequence)*, (g) annotate (predict genes). [L07 p5]

• **Assembling reads into contigs**: a computer links reads by looking for overlaps between the ends of reads, building contigs and ultimately chromosomes. A draft-quality *(unfinished, gaps remain)* assembly may have multiple contigs and thus many gaps; more sequencing is needed to eliminate all gaps (his example: a partial assembly with 55 scaffolds or contigs). [L07 p33, L07 p18, L07 p34]

★ **On his past exams**: Fa25 Q9 (genome in four scaffolds: tell chromosomes from plasmids because chromosomes carry essential housekeeping genes *(genes for basic survival)* and plasmids carry auxiliary genes *(extra, optional genes)*, not by size); Fa24 Q16 (3000 kb piece with many essential genes = chromosome, 1000 kb with a few essential genes = secondary chromosome, 200 kb with none = plasmid). [Key Fa25 Q9, Key Fa24 Q16]

know terms and sequencing technology used

• **Paper 1, Virgivacillus massiliensis (Illumina)**: gDNA *(genomic DNA: whole-cell DNA)* sequenced on the Illumina MiSeq (a next-generation sequencing-by-synthesis machine) using the mate pair strategy *(reads from ends of long pieces)*. [L07 p6]
• **Mate pair strategy**: newer strategy; fragments are bigger, so the reads from each end are farther apart. Often paired with traditional short-insert paired-end reads *(both ends of each piece read)*. Steps: DNA fragmentation, end labeling and circularization *(joining ends into a ring)*, fragmentation of the circular DNA, affinity purification (only biotin-labeled fragments kept), adapter ligation *(gluing on short known DNA ends)*, sequencing, data analysis. Useful for detecting large structural changes (inversions, duplications, translocations, deletions). [L07 p10, L07 p11, L07 p12]
• **Fragment sizes in the paper**: tagmented *(cut and tagged in one step)* fragments 1 to 10 kb (no size selection, only 14 ng circularized); circularized DNA mechanically sheared *(broken by physical force)* to an average of 696 bp *(base pairs)* (Covaris). **Bioanalyzer LabChip**: an electrophoresis system that measures fragment size. [L07 p12, L07 p13]
• **2 x 251 bp**: a read of about 251 bp from each end (2x) of a fragment; a machine setting. The run took a single 42-hour run. [L07 p14]
• **Run output (his worked numbers)**: 4.7 Gb total (4,700,000,000 bp) from all projects on the run; cluster density *(clusters per square millimeter)* 488K/mm2; 97.2% of clusters passed quality control (9,590,000 clusters). **Index representation** 11.16% = the share of reads that were Virgivacillus (run shared with 11 other projects). 4.7 billion x 11.16% = 0.52 billion bp (520,000,000 bp) of Virgivacillus; 9,590,000 clusters x 11.16% = 1,070,244 clusters x 2 reads per cluster = 2,140,488 reads. [L07 p15]
• **Trimming (Trimmomatic)**: each read carries some adapter sequence that must be trimmed off; the adapter can also include ID information so you know which sample the DNA came from. [L07 p16]
• **Assembly**: reads assembled with SPAdes *(assembly program)*; contigs combined with SSPACE and Opera *(programs that order contigs)*, helped by GapFiller *(program that fills gaps)*. **Contig** = continuous sequence. Final draft genome = seven contigs, so only finished to draft quality *(unfinished, gaps remain)*. [L07 p17]
• **Paper 2, Klebsiella pneumoniae KP36 (PacBio)**: sequenced on the PacBio RS II *(long-read sequencing machine)*, a third-generation *(newest, long-read type)* single-molecule real time (SMRT) machine. A 20-kb library *(collection of prepared DNA pieces)* was built; 189,711 reads with a mean read length of 7,718 bp. [L07 p19, L07 p21]
• **Circular contigs**: one short contig (23,273 bp) was discarded because its assembly was problematic. In each of the remaining four contigs the head segment was almost identical to the tail segment, showing the molecule is circular; the redundant segments were removed. Result: one circular chromosome and three circular plasmids. [L07 p22]
• **NCBI genome report terms** *(US gene database center)*: **Assembly** = name for the product of assembling the sequencing reactions (the built genome). **GenBank** = identifying number of the submitted sequence. **RefSeq** = NCBI's curated version *(checked by experts)*; if no RefSeq name is listed, assembly and annotation *(gene labeling)* were not checked. **Modifier** = qualifier (strain, isolate, cultivar, etc.) that distinguishes this genome from others of the same species. [L08 p35]
• **More NCBI terms**: **Annotation** = whether a gene annotation report exists and who made it (NCBI RefSeq pipeline or the submitter). **Level** = completeness of sequencing. **Release date** = when the assembly was first made public. **WGS accession** = master accession for Whole Genome Shotgun *(sequencing random pieces of the whole genome)* assemblies. **Scaffolds** = the large pieces generated by assembly; can equal chromosomes plus plasmids. **Reference sequence** (green check) = curated, non-redundant *(no duplicate entries)*, validated baseline genome for a species. [L08 p37, L08 p40]

✎ **His own summary of NCBI levels** (best to worst as he lists them): Chromosome (0 to a couple small gaps), Complete (near complete, a few small gaps), Contig (multiple contigs, some larger gaps), Scaffold (lower quality, more gaps and contigs). Note this slide says Complete has a few small gaps, while the NCBI definition on L08 p38 says Complete is gapless. [L08 p39, L08 p38]

• **JGI-IMG levels** *(US genome institute database)*: **Finished (F)** = high quality from more passes or depth *(times each base was read)*, few errors, usually contiguous molecules. **Permanent Draft (P)** = fewer passes, more errors, usually gaps (multiple contigs), order may be wrong. **Draft (D)** = many gaps and errors, many contigs, may still be in progress. [L08 p41]

	Virgivacillus (Illumina MiSeq)	Klebsiella (PacBio SMRT)
Generation	Next generation	Third generation
Number of reads	2,140,488 (much more)	189,711 (much fewer)
Read length	251 bp	7,718 bp mean (much longer)

Table source: [L07 p27]

NCBI assembly level	Meaning (NCBI definition)
Complete genome	All chromosomes gapless (runs of 9 or fewer ambiguous bases), no unplaced scaffolds, all expected chromosomes present
Chromosome	Sequence for one or more chromosomes; may have gaps; may also have unplaced sequences
Scaffold	Contigs connected across gaps of 10 or more bases; all sequences unplaced
Contig	No gaps in any sequence, but all unplaced; includes gapless long-read ultra-contigs

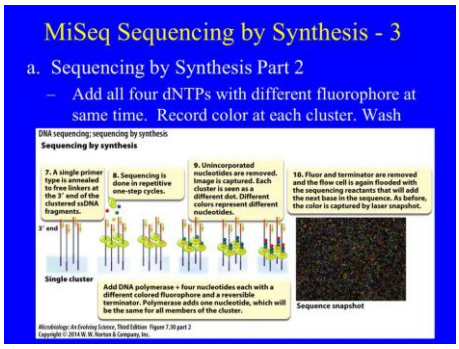
Table source: [L08 p38]

B. Sequencing methodologies – Know advantages/disadvantages of each (A3.6 and notes)

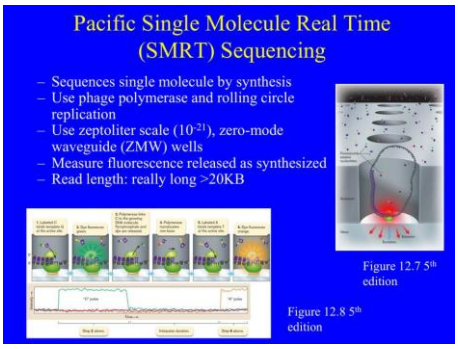
• **Key features to compare**: sample prep *(preparing DNA for the machine)* required, read length, accuracy, cost per base, run time, number of samples per run. [L07 p23]

✎ **Assignment 6 properties (answers come from his tables)**: longest reads = SMRT *(PacBio long-read method)*; shortest reads = Illumina *(short-read machine)*; high accuracy (over 99%) = Sanger *(older chain-termination method)* and Illumina; high to moderate throughput *(amount sequenced per run)* (millions of reads per run) = Illumina and SMRT; runs take hours (under 24 h) = Sanger and SMRT; low cost per billion bases (under $150) = Illumina and SMRT; minimal sample prep and smallest amount of DNA = SMRT. [A06, L07 p25, L07 p26, Key Fa25 Q15]

★ **On his past exams**: Fa25 Q4 (NOT an advantage of next-gen over Sanger: longer reads; next-gen reads are shorter. Listed advantages: automation of 1.6 million reactions at a time, less DNA, faster, no cloning of each piece into a plasmid); Fa25 Q15 (SMRT feature: minimal sample prep, needs only a very small amount of DNA). [Key Fa25 Q4, Key Fa25 Q15]



Illumina sequencing by synthesis: all four labeled nucleotides added, one base per cycle, color imaged at each cluster [L07 p09]



PacBio SMRT: one polymerase in a zero-mode waveguide well, fluorescence read as each base is added [L07 p20]

Feature	Sanger (dideoxy, chain termination)	Illumina (sequencing by synthesis)	PacBio SMRT
Generation	First	Next (second)	Third
Read length	400 to 900 bp (400 to 1000 bp on his Sanger slide; about 1,000 bases in Table 12.1)	Short: 50 to 600 bp by model (MiSeq 50 to 600 bp); 400 bases or less in Table 12.1	Longest: 30,000 bp (N50), maximum over 100,000 bp; over 20 kb on his SMRT slide
Accuracy (single read)	99.9%	99.9% (Phred30)	87% raw read (12% error per read); over 99% with circular consensus
Reads per run	1 to 96	1 to 25 million (MiniSeq/MiSeq) up to 3 billion (HiSeq X)	4,000,000 per Sequel 2 SMRT cell (moderate)
Time per run	20 minutes to 3 hours	1 to 11 days	30 minutes to 20 hours
Cost per 1 billion bases	$2,400,000	$5 to $150	$7.2 to $43.3
Advantages	High accuracy, long individual reads, useful for many applications; cheap per run (can do just one)	High sequence yield and throughput (up to 1,800 Gb); low cost per read	Longest reads; fast; minimal sample prep; detects DNA methylation (4mC, 5mC, 6mA)
Disadvantages	More expensive; impractical for large projects; needs time-consuming plasmid cloning or PCR; low throughput, high cost per read	Equipment very expensive; requires high concentrations of DNA; short reads so assembly a challenge; high cost per run	Low single-read accuracy; moderate throughput; equipment very expensive; high cost per run
Ideal uses	A few PCR products or plasmids	Diversity studies (metagenomics), RNAseq, low-complexity genomes	Metagenomics, high-complexity genomes, methylomics

Table source: [L07 p24, L07 p25, L07 p26, L07 p28, L07 p31]

1. Sanger dideoxy

• **Sanger dideoxy (chain termination)**: first-generation method. Still used for sequencing individual plasmids *(small extra DNA circles)* and PCR products *(DNA copies made by a copying reaction)*; was used for some of the first genomes. Accuracy very high; cost per base high; cost per run cheap (can do just one); 1 to 96 reads per run. [L07 p28]

• **Prep (clone genome fragments into a plasmid)**: (1) break genomic DNA into small pieces by mechanical shearing *(breaking by physical force)* (1000 to 2000 bp), (2) clone fragments into a plasmid vector *(plasmid used to carry DNA)*, (3) introduce plasmids into bacterial cells, (4) purify plasmid from different colonies *(cell piles grown from one cell)* (each colony = one plasmid), (5) sequence one at a time. [L07 p29]

• **Reaction**: add DNA polymerase, primer *(short starter DNA strand)*, dNTPs *(normal DNA building blocks)* and ddNTPs *(chain-stopping DNA building blocks)* (usually fluorescently labeled). New strands are synthesized complementary to the template *(strand being copied)* starting at the primer; a dideoxynucleotide causes termination, giving strands of different lengths. [L07 p30]

• **Read-out**: separate fragments by size (length) by gel electrophoresis *(sorting DNA by size in gel)*; a laser reads the color as fragments run through the gel; a computer converts the color peaks into DNA sequence (chromatogram). Read length 400 to 1000 bp; run time a couple hours. [L07 p31, L07 p32]

2. Next Generation - Solexa Illumia sequencing by synthesis –

• **Illumina (Solexa) sequencing by synthesis**: next-generation method (MiSeq in the Virgivacillus paper). Steps: (1) fragment DNA (usually under 1 kb), (2) add adapters *(short known DNA ends)*, (3) amplify individual DNAs by bridge PCR *(copying on a surface)* to form a cluster, (4) sequence thousands at once, detecting the fluorescence *(glowing light)* of each base added, (5) assemble sequences. [L07 p7, L06 p18]

• **Each cycle**: add DNA polymerase plus all four dNTPs *(DNA building blocks)* at the same time, each with a different fluorophore *(glowing dye)* and a reversible terminator *(removable stop cap)*; polymerase adds one nucleotide, the same for every member of a cluster. Wash away unincorporated *(not added)* nucleotides, record the color at each cluster, remove fluor and terminator, repeat. [L07 p9]

• **Paired reads**: 2 x 251 bp = about 251 bp read from each end of a fragment. Adapters carry sample ID (index) so many samples can share one run; adapter sequence is trimmed from reads before assembly. [L07 p14, L07 p15, L07 p16]

• **Pros and cons**: high accuracy (99.9%), high throughput *(amount sequenced per run)* (up to 3 billion reads per run), low cost ($5 to $150 per billion bases); but short reads (50 to 600 bp) make assembly a challenge, runs take 1 to 11 days, equipment is very expensive and it requires high concentrations of DNA. [L07 p25, L07 p26]

3. Third Generation- SMRT

• **PacBio Single Molecule Real Time (SMRT) sequencing**: third generation. Sequences a single molecule by synthesis; uses a phage polymerase *(copying enzyme from a virus)* and rolling circle replication *(copying around a circle repeatedly)*; zeptoliter-scale (10^-21) zero-mode waveguide (ZMW) wells *(tiny light-trapping holes)*; measures the fluorescence released as each base is synthesized. Read length really long, over 20 kb *(thousand bases)*. [L06 p19, L07 p20]

• **Pros and cons**: longest reads (30,000 bp N50, maximum over 100,000 bp), fast (30 minutes to 20 hours), minimal sample prep, detects DNA methylation *(chemical tags on DNA bases)*, low cost per billion bases ($7.2 to $43.3); but low single-read accuracy (87%, 12% error per read; over 99% with circular consensus), moderate throughput *(amount read per run)* (4 million reads per SMRT cell), very expensive equipment, high cost per run. [L07 p25, L07 p26]

• **In the Klebsiella paper**: 189,711 reads averaging 7,718 bp, far fewer but far longer reads than the Illumina paper (2,140,488 reads of 251 bp). The assembled contigs *(continuous assembled stretches)* had near-identical head and tail segments, showing circular molecules (1 chromosome + 3 plasmids). [L07 p21, L07 p22, L07 p27]

C. Features of Microbial Genomes (7.1-7.2, 7.5, and 6.3, notes and Assignment 7)

• **Features reported in a genome report**: 1 genome size, 2 nature of genetic material, 3 %GC *(share of G and C bases)*, 4 number of chromosomes, 5 geometry *(shape: circle or line)* of chromosomes, 6 number of plasmids *(small extra DNA molecules)* and secondary chromosomes, 7 number of genes (CDS and total), 8 functional RNA genes, 9 genes with functional prediction, 10 genes assigned to COGs *(protein family groups)*, 11 coding regions and % protein coding, 12 organization of genes, 13 codon usage *(which codons are favored)*. [L08 p6, L08 p9, L08 p13, L08 p14, L08 p15, L08 p18, L08 p19]

• **Context matters**: genome reports give many numbers that have little meaning alone; you need the range of values and what is typical for similar organisms (the point of Assignment 7 Part A). [L08 p22, L08 p23]

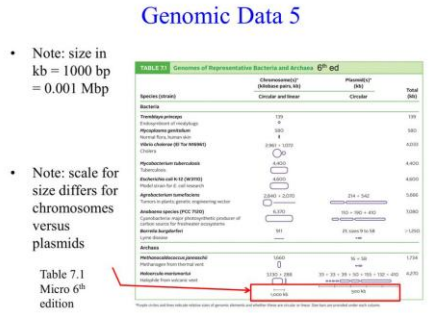


Table 7.1: chromosomes and plasmids of representative bacteria and archaea [L08 p29]

★ **On his past exams**: Fa25 Q6 (NOT observed in bacteria: more than 10 chromosomes; observed: wide %GC range, linear chromosomes, multiple plasmids, plasmids as big as some chromosomes); Fa25 Q8 (region with different %GC = recently acquired by horizontal transfer); Fa25 Q10 (high %GC species prefer G/C-rich codons, low %GC species prefer codons with fewer G/C); Fa25 Q16 (extensive genome size overlap: archaea and bacteria); Fa25 SA30 (bacterial vs eukaryotic genomes: smaller, fewer chromosomes, more likely to have and can have multiple plasmids, fewer genes, usually circular, higher % protein coding, wider %GC range); Fa24 Q9 (NOT a difference: eukaryotes having more plasmids); Fa24 Q14 (property varying least across bacteria: number of chromosomes); Fa24 Q17 (most overlap across all four groups: %GC). [Key Fa25 Q6, Key Fa25 Q8, Key Fa25 Q10, Key Fa25 Q16, Key Fa25 Q30, Key Fa24 Q9, Key Fa24 Q14, Key Fa24 Q17]

Attribute (Virgivacillus, Table 4)	Value	% of total
Genome size (bp)	4,353,177	100
DNA coding region (bp)	3,754,518	86.25
DNA G + C (bp)	1,605,022	36.87
Total genes	4519	100

rRNA *(ribosome RNA)*	5	0.11
tRNA *(amino acid carrier RNA)*	42	0.93
tmRNA *(rescues stalled ribosomes)*	0	0
miscRNA *(other noncoding RNAs)*	78	1.73
Protein-coding genes	4394	86.25
Genes with function prediction	3287	74.81
Genes assigned to COGs	4291	97.66

Table source: [L08 p5]

1. Terms to know: genome, chromosome, secondary chromosome, plasmid, gene, %GC, codon usage

• **Genome**: all the genetic material of an organism or virus: chromosomes plus extrachromosomal DNA (plasmids). [L06 p14, L08 p4]

• **Chromosome**: DNA carrying genes passed down to all offspring. In bacteria typically a single double-stranded DNA molecule, usually a closed circle, containing the essential genetic instructions. [L06 p14, L08 p4]

• **Secondary chromosome (chromid)**: the smaller chromosome(s) in cells with multiple chromosomes. Contains at least one essential gene *(gene needed to survive)* (high so far is 5). Thought to have evolved from a plasmid that acquired an essential gene from the primary chromosome; replication similar to plasmids in those studied so far. Example: Vibrio cholerae 2,961 + 1,072 kb *(thousand base pairs)*. [L08 p11, L08 p12]

• **Plasmid**: extrachromosomal DNA carrying auxiliary *(extra, helper)* or non-essential genes; found in some bacteria and fungi (also archaea, some mitochondria and chloroplasts). Usually smaller than chromosomes (under 100 kb) but some as large as 500 kb. Now distinguished from chromosomes by the function of the genes carried (non-essential, ancillary) and by independent replication, not by size. [L08 p4, L08 p9, L08 p10]

• **Genome size**: reported in base pairs, usually for chromosomal DNA only; total genome = all chromosomes plus plasmid DNA; for eukaryotes haploid *(one set of chromosomes)* or total. Units vary: kbp = 1 x 10^3, Mbp = 1 x 10^6, Gbp = 1 x 10^9 bp. [L08 p6, L08 p26]

• **Nature of genetic material**: double-stranded DNA for all living organisms; varies in viruses. **Number of chromosomes**: bacteria can have more than 1, including secondary chromosomes. **Geometry**: linear or circular. [L08 p6, L08 p9]

• **Gene**: a sequence of nucleotides that encodes a functional product (protein or RNA). [L08 p4]

• **Number of genes (predicted)**: usually excludes genes of viral and transposable elements *(jumping DNA pieces)*. **Protein-coding genes (CDS, coding sequence)** = regions predicted to code for protein. **Total genes** = CDS plus genes for RNAs (tRNA, rRNA, misc/noncoding RNA, tmRNA). [L08 p13]

• **tmRNA** (transfer-messenger RNA, 10Sa RNA or SsrA): acts as both tRNA and mRNA to rescue stalled ribosomes *(protein-making machines)*. **misc RNA / noncoding RNA**: may be regulatory, such as microRNA *(tiny regulating RNA)* or short noncoding RNAs. [L08 p13]

• **Genes with functional prediction**: number (and %) of predicted genes whose product function could be predicted; small for some bacteria (under 50%). **COGs** (clusters of orthologous groups) = protein families; % of genes matching a known family. [L08 p14]

• **% protein coding**: bases predicted to code for protein as % of total. Very high in bacteria, over 85%, with few introns *(non-coding gene interruptions)*, transposons *(jumping genes)* or repetitive elements (L13: 80 to 90%). Human: protein-coding regions only about 1.2% of the genome; introns 20 to 35%, transposable elements or remnants about 45%. [L08 p15, L08 p16, L08 p17, L13 s7]

• **Gene organization in bacteria**: most genes in operons *(genes under one on/off switch)*, several genes transcribed on the same (polycistronic) mRNA; introns rare. [L08 p18]

• **%GC (GC content)**: percent of total bases that are G or C. Varies between species of bacteria and eukaryotes, is a characteristic of a species, and is used to classify bacteria. A region whose %GC differs from the rest of the chromosome was recently acquired horizontally *(from another species)* (e.g. a pathogenicity island). [L08 p6, L08 p8]

• **Codon usage**: the frequency with which each of the codons *(three-base code words)* for the same amino acid appears in the coding regions of an organism. Conclusion: codon usage can vary considerably within and between species, and the codons used correlate with %GC. [L08 p19, L08 p21]

• **Reading the table**: high-GC P. aeruginosa almost never uses A/U-rich UUA (0.3) and favors CUC and CUG, which is why Fa25 Q10 answer B is correct. [L08 p20, Key Fa25 Q10]

Leucine codon (per 1000 codons)	E. coli (51.8% GC)	P. aeruginosa (67.1% GC)
UUA	13.9	0.3
UUG	13.7	8.7
CUU	11.0	3.1
CUC	11.0	27.8

CUA	3.9	1.4
CUG	52.8	82.9

Table source: [L08 p20]

2. How do the properties of bacterial genomes compare to viral & eukaryotic genomes?

• **Size overlaps**: the archaeal *(of archaea, non-bacterial microbes)* range falls inside the bacterial range (Fa25 Q16 answer A); the largest viruses (2.8 Mbp) are bigger than the smallest bacteria (0.16 to 0.58 Mbp); the largest bacteria (13 Mbp) are about the size of yeast (12 Mbp). [L08 p26, Key Fa25 Q16]

• **%GC has the most overlap** *(share of G plus C bases)*: all four groups share the middle values and differ only at the lowest and highest ends (Fa24 Q17). The %GC values are not on his slides; Assignment 7 asked you to find them. [Key Fa24 Q17, L08 p23]

• **Gene number vs size**: in bacteria, archaea and viruses gene number rises with genome size (E. coli 4.6 Mbp, 4,300 genes; S. cellulosum 13 Mbp, 9,400 genes). Not in eukaryotes: human 3.2 Gbp *(billion base pairs)* has 21,000 genes, fewer than poplar (46,000) or wheat (95,000); onion (15,000 Mb) has only 8 haploid *(one chromosome set)* chromosomes vs human (3,000 Mb) 23. [L08 p26, L08 p30]

• **Bacterial vs eukaryotic genomes (his key answer)**: bacterial genomes are generally smaller, have fewer chromosomes, are more likely to have plasmids *(small extra DNA circles)* and can have multiple plasmids, have fewer genes, usually circular chromosomes, higher % protein coding, and a wider range of %GC. [Key Fa25 Q30]

• **Why genes are easier to find in bacteria**: no introns *(non-coding gene interruptions)* and less repetitive DNA and pseudogenes *(broken, nonworking gene copies)*; the human genome is about 45% transposable elements *(jumping DNA pieces)* or remnants, about 5% pseudogenes, about 10% simple tandem repeats *(back-to-back repeated sequences)*. [L09 p15, L08 p17]

Feature	Bacteria	Archaea	Eukaryotes	Viruses
Genome size	139 kb (Tremblaya) or 160 kbp (C. ruddii) to 13 Mbp (Sorangium cellulosum)	490 kbp (Nanoarchaeum) to about 5.8 Mb (Methanosarcina bar chart)	12 Mbp (yeast) to 150 Gbp (Paris japonica)	1.7 kb (hepatitis D) to 2.8 Mbp (Pandoravirus)
Number of genes	182 (C. ruddii) to 9,400 (S. cellulosum)	550 (N. equitans) to 2,000 (P. furiosus)	4,800 (S. pombe) to 95,000 (wheat)	1 (hepatitis D) to 2,500 (Pandoravirus)
Chromosomes	1 to 2 (V. cholerae, A. tumefaciens, R. sphaeroides have 2)	1 to 3 (Halobacterium NRC1 has 3)	3 (S. pombe) to 48 (chimp, 2n); haploid 4 to 39 in Table 5.3	1 dsDNA up to 8 ssRNA segments (influenza)
Nature of material	dsDNA	dsDNA	dsDNA	DNA or RNA, single or double stranded
Geometry *(shape)*	Mostly circular; linear in Borrelia and Streptomyces	Circular	Linear (telomeres on each end)	Linear or circular
Plasmids	0 to 21 different (Borrelia)	Up to 7 (Haloarcula)	Some fungi	
% protein coding	About 88 to 89%		Yeasts 60 to 70%, human 1.2%	

Table source: [L08 p25, L08 p26, L08 p27, L08 p28, L08 p29, L08 p30, L08 p31, L08 p32, L10 p33]

3. How do the properties of bacterial genomes vary between species?

• **Lifestyle affects size**: free-living bacteria have larger genomes, obligate symbionts *(partners that cannot live alone)* smaller (e.g. Wolbachia has a small genome because it is a parasite of insects and worms). [L08 p7, L09 p35]

• **Size and gene number vary over a wide range**: 139 kb (Tremblaya, endosymbiont of mealybugs) to 13 Mbp (Sorangium); 182 to 9,400 genes. **%GC** *(share of G plus C bases)* **and codon usage** *(which codons are favored)* vary between species but are characteristic of each species. [L08 p29, L08 p26, L08 p6, L08 p21]

• **Chromosome number varies least**: bacteria have 1 to 2 chromosomes (Fa24 Q14), sometimes secondary chromosomes *(smaller extra chromosomes)*; more than 10 has never been observed (Fa25 Q6). Plasmid content varies more: 0 to 21 plasmids *(small extra DNA circles)* (Borrelia has 21), and presence varies even within a species. [L08 p31, L11 p17, L11 p20, Key Fa24 Q14, Key Fa25 Q6]

• **% genes with functional prediction** also varies: small for some bacteria (under 50%). [L08 p14]

Lifestyle	Free-living	Recent or facultative pathogen	Obligate symbiont or pathogen
Genome size	Large (5 to 10 MB)	Intermediate (2 to 5 MB)	Small (0.5 to 1.5 MB)
Pseudogenes *(dead gene copies)*	Few	Many	Rare
Lateral gene transfer (LGT)	Frequent	Frequent to rare	Rare to none
Selfish genetic elements *(DNA that copies itself)*	Few	Common	Rare
Genome organization	Stable or unstable	Unstable	Stable

Table source: [L08 p7]

D. Genomic Sequence Analysis

• **Genomic sequence analysis** (one area of bioinformatics): (1) determine basic properties (size, number of chromosomes and plasmids, %GC); (2) annotate *(label genes and features)* the genome: (a) identify genes, (b) predict function of genes or gene products, (c) identify other components (transposable elements, lysogenic viruses); (3) determine how genomes and species have changed and evolved. [L09 p4]

• **Evolution from sequence**: a region with a different %GC suggests it was recently acquired by horizontal transfer *(gene passing between species)* from another species. Comparing to other strains *(genetic variants of a species)* (E. coli O157, E. coli K12 and Shigella): grey lines = shared regions, colored boxes = differences. [L09 p37, L09 p38]

★ **On his past exams**: Fa24 Q5 (ORF search: six reading frames encoding over 30 amino acids); Fa25 SA28 (identify genes: ORFs in six frames over 30 aa from AUG, ribosome binding site and transcription signals, similarity to proteins in other bacteria); Fa24 SA27a (predict function: search for similar proteins in other bacteria); Fa24 SA27b (why function cannot be predicted: similar proteins have no known function, gene prediction wrong, sequence error); Fa25 Q17 (BLASTp 95% query cover, E 3.0E-150, replication protein: highly likely a protein and likely involved in replication); Fa25 Q8 (different %GC region = horizontal transfer). [Key Fa24 Q5, Key Fa25 Q28, Key Fa24 Q27, Key Fa25 Q17, Key Fa25 Q8]

1. Gene Annotation (9.5 6e)

• **Gene annotation**: adding explanation to the sequence, first by finding the genes. The main clue is the ability to code for a protein: an open reading frame *(stretch that could code a protein)*. [L09 p4, L09 p6]

• **Easier in bacteria**: no introns *(non-coding gene interruptions)* and less repetitive DNA and pseudogenes *(broken, nonworking gene copies)*, so genes are densely packed (see the 250,000 bp E. coli segment and the predicted genes on the two Vibrio cholerae chromosomes). [L09 p15, L09 p16, L09 p17]

Search for Genes – open reading frame search – how, details

• **Six reading frames**: a DNA sequence can be read in 3 frames on each strand, 6 in total, but only one of the possible frames is used within a given sequence. [L09 p8, L09 p9]

• **How an ORF search works** *(ORF = open reading frame)*: (a) determine the amino acids all six frames could produce; (b) find start and stop codons *(signals to begin and end a protein)* and look for frames able to code for over 30 amino acids; (c) non-coding frames hit a stop codon on average every 100 bp (33 amino acids), so short reading frames can usually be rejected. [L09 p10, L09 p11, L09 p12]

• **Supporting evidence**: (d) look for sequences needed for translation (ribosome binding site) and transcription (promoter); (e) a similar gene or protein found in other organisms; (f) in eukaryotes, work backwards from RNA transcripts and find them in the genome. [L09 p14]

✎ **His model answer (Fa25 SA28)**: look for ORFs in all six frames able to make a peptide *(short protein chain)* of over 30 amino acids from an AUG *(start codon)*; then look for a ribosome binding site (Shine-Dalgarno) upstream *(just before the gene)* and transcriptional regulatory sequences; then check whether the predicted amino acid sequence is similar to a protein in a different bacterium. [Key Fa25 Q28]

• **NCBI ORF Finder settings (Assignment 8)**: minimal ORF length 300 nucleotides = 100 amino acids is the default for bacteria; start codon ATG only or ATG plus alternatives (in E. coli 17% of genes begin with GTG or TTG); option to ignore nested ORFs *(ORFs inside larger ORFs)*. Lowering the minimal length admits more short ORFs, most of which should be rejected. [A08, L09 p12]

2. Gene Function prediction – how

• **Goal**: predict the function of each predicted gene or ORF based solely on sequence data. [L09 p18]

• **1) Analyze properties of the predicted amino acid sequence**, e.g. hydrophobicity *(how water-repelling it is)*. [L09 p18]

• **2) Search for proteins with similar sequence (BLASTp)** to find homologs with known function in other organisms. **Homologs**: genes derived from the same ancestor gene, likely to have the same function. BLASTp: paste the amino acid sequence, select non-redundant *(no duplicate entries)* protein sequences, click BLAST; read the Description of the top hits for probable function (one of the top hits needs a predicted function). [L09 p18, L09 p19, L09 p20]

• **3) Conserved domains, motifs or protein families** *(protein parts kept unchanged over evolution)*: Conserved Domain Database (CDD), Pfam *(protein family database)*, Protein Data Bank *(protein structure database)*; the function of each domain is usually known. In NCBI: Graphic Summary tab, click the conserved domain covering the query, read the Conserved domains and Protein classification sections. [L09 p24, L09 p25, L09 p26, L09 p27]

• **4) Categorize gene function or role** (transport, transcription, metabolism); report number and % in each category and compare to similar organisms (pie charts, color-coded genome maps). Problem: some genes fall in the unknown or hypothetical *(predicted, function unknown)* category. [L09 p28, L09 p30, L09 p31]

• **6) Predict metabolic ability**: use annotations to predict metabolic pathways *(chains of chemical reactions)* and transport systems, e.g. Helicobacter pylori, Wolbachia (small genome, parasite of insects and worms), Treponema pallidum (syphilis). [L09 p33, L09 p34, L09 p35, L09 p36]

✎ **His key (Fa24 SA27a)**: the scientist would search for proteins in other bacteria that are similar to the protein predicted to be produced by the gene. [Key Fa24 Q27]

Interpreting Blast search results: e value, query score,

• **E value (expectation value)**: the chance the similarity is due to chance and not to homology *(shared ancestry)*, so a smaller number means greater similarity. 0.0 = identical; e-10 to e-150 suggests some to high similarity. It is written as a negative exponent: 1e-90 means 1 x 10^-90, a very small number and highly similar proteins. Should be less than 1e-10. [L09 p22, A08]

• **Alignment view**: also aligns conservative substitutions (same type of amino acid), marked with + in the middle line. [L09 p23]

• **Interpreting**: a real protein-coding gene should have similar proteins in other organisms; an ORF *(open reading frame: possible protein stretch)* that does not really encode a protein gives no similar proteins or only weak similarities. If the match is strong, read the Description of the top hits for function, or open the top hit's Graphics to see a protein family. [A08, L09 p20]

✎ **His exam case (Fa25 Q17)**: top hit with 95% query coverage *(share of your sequence matched)*, E value *(odds the match is chance)* 3.0E-150, described as a replication protein in another species: the ORF is highly likely to encode a protein, likely one involved in replication (answer C). [Key Fa25 Q17]

BLAST value	Meaning	Strong match
Max score	Highest alignment (bit) score for a segment of query vs subject	Higher is stronger (over 1000 in A08)
Total score	Sum of scores of all aligned segments between query and subject	Higher
Query score / query cover	Percentage of your query sequence that aligns with the subject (1 to 100%)	Higher, near 100%
E value	Number of hits expected by chance with a similar score	Lower; 0.0 = identical; should be under 1e-10
Max ident / % identity	Proportion of amino acids (or nucleotides) exactly the same in the aligned region	Higher

Table source: [L09 p21, A08]

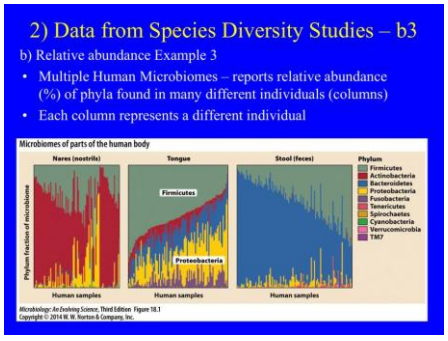
Limitations

• **Limitations of function prediction**: can usually only predict function if the function of similar proteins is known; otherwise the annotation *(gene label)* says unknown, hypothetical or conserved hypothetical *(seen in many species, function unknown)*. Predictions are often only of a general type or class (transporter, kinase) without the specific molecule it acts on. Many different names are used for similar proteins. [L09 p32, L09 p28]

✎ **His key (Fa24 SA27b)**: function cannot be hypothesized when (1) similar proteins in other bacteria have no predicted function, (2) the gene prediction is wrong (not really a gene), or (3) there is an error in the sequence data. [Key Fa24 Q27]

E. Metagenomics – terms and data presentation

• **Metagenomics**: study of genetic material recovered directly from environmental samples. Used to determine which microbial species are present and their abundance (oceans, soil, human microbiomes), can detect unculturable *(cannot be grown in lab)* bacteria, can examine the genes present to analyze the metabolic abilities of a community, and to sequence ancient remains (mammoths). Uses second or third generation sequencing. [L10 p4]



Relative abundance of phyla in nares, tongue and stool; each column is one person [L10 p17]

• **Basic steps**: collect environmental sample; may filter for the type wanted; break open cells and isolate DNA; fragment and construct library *(collection of prepared DNA pieces)*; sequence; assemble; interpret. [L10 p5]

• **Species diversity and abundance studies** (most common report): identify bacteria and fungi by the small subunit rRNA gene (SSU rDNA): 16S rRNA gene for bacteria, 18S for fungi; it has highly variable and highly conserved regions. Method a: PCR the rRNA genes with universal primers *(starters matching all bacteria)* to the conserved regions, then sequence many products. Method b: shotgun, sequence random pieces of all DNA (2nd or 3rd generation), use reads matching rRNA genes to identify bacteria, estimate abundance from the number of reads per species. Concern: whether rare microbes are detected and whether results reflect the biome *(living community site)*. [L10 p10, L10 p11, L10 p13]

• **Limitations**: some organisms lost by filtering; DNA of some organisms not isolated; universal SSU primers may not be universal or may amplify other organisms; much sequencing needed for low-abundance species; bacteria have multiple rRNA gene copies that can differ within one organism; viruses are harder to collect and sequence. [L10 p30]

★ **On his past exams**: Fa24 SA29 (information reported from a microbiome study: types and relative abundance of bacteria at phylum or genus level per individual and site; species present and abundance; predicted metabolic ability;

correlations of bacteria with disease vs healthy); Fa25 Q14 (HMP sampled over 240 healthy humans at 15 to 18 sites). [Key Fa24 Q29, Key Fa25 Q14]

Data reported	How presented	His example answer
Total number of species	Rarefaction curve (species observed vs sequencing effort)	More sequencing gives a gradual increase that levels off; shows if more sequencing is worthwhile
Relative abundance	Stacked bar graphs (each column a sample or individual)	Most abundant order in air: Caulobacterales
Relative abundance	Bubble chart; dot size = abundance	Most abundant genus in throat: Streptococcus
Alpha diversity	Diversity within a sample, usually number of species (species richness)	
Beta diversity	Differences in community composition between samples	
Gene function categories	Bar graphs of predicted gene function per sample	Gut has more amino acid and energy genes than ocean; ocean samples showed not much difference
Genome assembly	Full or partial genomes assembled from the sample	Partial genome of strain APb

Table source: [L10 p14, L10 p19, L10 p23, L10 p25, L10 p26, L10 p27, L10 p28, L10 p29]

Results of Human microbiome Study (Assignment 7 Part C)

• **The paper**: Structure, function and diversity of the healthy human microbiome, Human Microbiome Project Consortium, Nature 2012, 486:207-14. Goal: characterize the normal microbial flora *(resident microbes)* of healthy humans. [A08]

• **Subjects and sites**: 242 healthy adults (129 men, 113 women). Women sampled at 18 body habitats, men at 15 (the difference is 3 vaginal sites). Sites: 9 oral, 4 skin, 1 nasal (anterior nares), 1 stool, 3 vaginal. His exam answer: over 240 healthy humans at 15 to 18 sites. [A08, Key Fa25 Q14]

• **Two sequencing technologies**: 454 pyrosequencing *(light-based sequencing method)* of 16S rRNA genes to identify which bacteria are present; Illumina paired-end shotgun metagenomic sequencing *(both ends of random pieces of all DNA)* to identify function (gene families and metabolic pathways). [A08, L07 p24]

• **Alpha diversity** *(variety within one sample)*: highest in the oral sites (saliva among the highest), lowest in the vagina. **Beta diversity** *(difference between samples)*: highest at the skin (antecubital fossae), meaning skin communities differ most from person to person; saliva is rich within a person but similar between people. [A08, L10 p19]

• **Carriage varies** *(how much each person carries)*: Streptococcus on the tongue (Fig 4a) varies a great deal between individuals, both in total amount and in which species are present. [A08]

• **Host phenotypes** *(observable traits)* **correlated with composition**: ethnicity/race (strongest), and vaginal pH (higher pH with less Lactobacillus). [A08]

• **Why it mattered**: it defined a baseline for the healthy microbiome that disease studies compare against; no single taxon *(named group of organisms)* was found in everyone, but metabolic pathways *(chains of chemical reactions)* were stable across individuals despite the variation in which organisms were present. [A08]

• **16S identification practice** *(ribosome RNA gene used for ID)*: BLAST a 16S rRNA gene sequence at NCBI to name an isolate *(single cultured strain)* and report its phylogeny *(evolutionary family tree)* (Assignment 8 Part C). [A08]

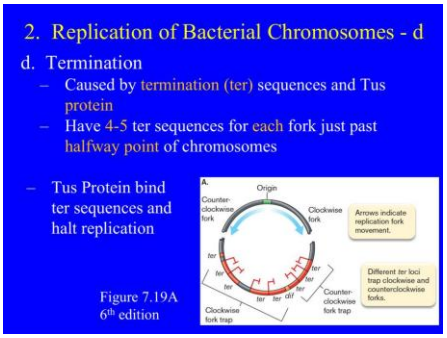
Body site	Most abundant phylum (Fig 2a)	Genus (Fig 3b)	Species (Fig 3a)
Anterior nares (his example row)	Actinobacteria	Corynebacterium	Propionibacterium acnes
Buccal mucosa ***(inner cheek lining)***	Firmicutes	Streptococcus	Streptococcus mitis
Stool	Bacteroidetes	Bacteroides	Prevotella copri
Posterior fornix ***(deep vaginal recess)***	Firmicutes	Lactobacillus	Lactobacillus crispatus

Table source: [A08]

F. Replication of Bacterial chromosomes and Plasmids (7.3 and 7.4 6 ed.)

• **The questions he asks**: how bacteria start and stop replication of chromosomes and plasmids (not the mechanism of synthesis), how chromosomes and plasmids are segregated *(separated)* so every daughter cell gets each, and how this differs from eukaryotes. [L11 p4]

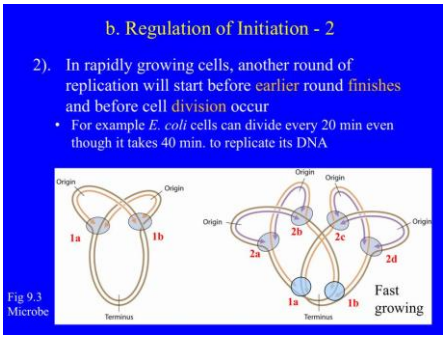
• **Linear bacterial chromosomes** have the end problem *(ends cannot be fully copied)*; two solutions: (a) **hairpin ends (Borrelia)**: the two strands are covalently *(by strong chemical bonds)* joined at each end; telomere resolvase (ResT) cuts and rearranges replicated DNA to regenerate the hairpins, so there is no free 5' end *(strand starting end)* to fill. (b) **Terminal proteins (Streptomyces)**: proteins covalently attached to the 5' ends serve



One origin, forks moving both ways, ter sites bound by Tus trapping each fork past the halfway point [L11 p11]

as primers *(starters for copying)* to fill in the missing terminal DNA (protein-primed end patching). [L11 p15]

★ **On his past exams**: Fa24 Q4 (NOT a plasmid feature: replication always tied to the chromosome); Fa24 Q22 (DnaA acts in initiation and is thought to tie replication to growth rate); Fa24 SA31 (draw a circle with a single origin and multiple termini on the opposite side; replication goes both ways from the origin with both strands copied until the forks hit the termini; another round can start before the first finishes); Fa25 Q19 (doubling every 30 min with 60 min replication: two complete copies each halfway replicated, so 4 copies of regions near the origin); Fa25 Q23 (one origin and multiple termination sequences, several for each fork). [Key Fa24 Q4, Key Fa24 Q22, Key Fa24 Q31, Key Fa25 Q19, Key Fa25 Q23]



Fast growth: a new round starts at the origins before the first round finishes [L11 p06]

	Eukaryotic chromosomes	Bacterial chromosomes	Archaeal chromosomes
Location		Cytosol	
Shape	Linear	Mostly circular (some linear)	Mostly circular
Origins per chromosome ***(replication start sites)***	Several	One	1 to several (example with 3)
Direction	Both directions from each origin	Both directions (bidirectional)	Bidirectional from each
End problem	Yes: cannot replicate both strands to the end; solved by repetitive telomeres bound by proteins	None for circles, but replicated circles are interlinked	
Segregation ***(splitting copies into daughter cells)***	Chromosomes condense, mitotic spindle attaches, pulled to opposite ends, cell divides	No mitotic spindle, DNA not condensed; origins move to opposite sides	
Enzymes		Synthesis very similar to eukaryotes	More like eukaryotic enzymes

Table source: [L10 p33, L10 p34, L11 p5, L11 p9, L11 p12, L11 p16]

1. Bacterial chromosomes:

• **Bacterial chromosome replication basics**: DNA located in the cytosol *(cell fluid)*; most bacterial and archaeal chromosomes are circular, which eliminates the linear end problem but creates another (interlinked daughter circles). One origin *(replication start site)* per chromosome; replication goes in both directions; initiation *(start of copying)* is regulated by the energy state of the cell. [L11 p5]

• **Mechanism of synthesis** (initiation and elongation by DNA polymerase and accessory proteins) is very similar to eukaryotes; he does not test it here. [L11 p9, L11 p4]

a. Sequences elements needed for initiation and termination

• **Initiation element: the origin (oriC)**, one per chromosome, with three 13 bp repeats and four 9 bp repeats. DnaA *(replication starter protein)* + ATP binds the origin, then HU *(DNA-bending protein)* + ATP, then DnaB helicase *(DNA-unwinding enzyme)* (with DnaC) is loaded, then priming *(laying down starter RNA)* and replication. GATC sites in the origin are methylated *(chemically tagged with methyl groups)*. [L10 p36]

• **Termination elements: ter sequences plus Tus protein** *(replication stop sites and their binding protein)*. There are 4 to 5 ter sequences for each fork *(Y-shaped copying point)*, just past the halfway point of the chromosome; Tus binds ter and halts replication. Different ter loci form a clockwise fork trap and a counterclockwise fork trap. [L11 p11]

• **Exam answer**: one origin and multiple termination sequences *(replication stop sites)*, several for each replication fork (Fa25 Q23 C). Drawing (Fa24 SA31): a circle with a single origin and multiple termini on the opposite side; two forks move in opposite directions until they hit the termini. [Key Fa25 Q23, Key Fa24 Q31]

b. Role of DnaA and link to cell growth/division

• **Initiation is regulated by energy levels in the cell**: for a while after replication the origin *(replication start site)* is inactive because its DNA is hemimethylated *(methyl tag on one strand only)*. If cells are growing, levels of DnaA and ATP rise and the origin DNA becomes (fully) methylated, allowing a new initiation. [L10 p36]

• **Link to growth**: DnaA *(replication starter protein)* acts in initiation and is hypothesized to tie replication to growth rate (Fa24 Q22 answer E). In rapidly growing cells a new round starts before the earlier round finishes and before the cell divides. [Key Fa24 Q22, L11 p6]

c. Method of Chromosome segregation

• **Different from eukaryotes**: no mitotic spindle *(fiber machine pulling chromosomes apart)*, DNA not condensed *(tightly packed)*; not totally understood. [L11 p12]
• **Step a, resolve interlinking** *(untangle joined rings)* of the replicated circles (see below). **Step b, segregation** *(splitting copies to each side)*: the origin appears attached to an unidentified structure (membrane?) in the center of the cell; the replisome *(copying machinery)* appears to stay in the center; the origins then move to opposite sides of the cell. The mechanism is disputed and may vary; sometimes a filament *(protein fiber)* may help. [L11 p13, L11 p14]
• **Microscope evidence** (origins stained green, termini red): right after division, two origins and one terminus; just before division, 4 origins and 1 terminus. [L11 p7, L11 p8]

d. Issues/challenges with Replication of chromosomes in bacteria

• **Challenges**: (1) replicated circular chromosomes end up interlinked; (2) in fast-growing cells division occurs faster than one round of replication; (3) the few linear bacterial chromosomes must solve the end problem *(ends cannot be fully copied)* (hairpin ends, terminal proteins). [L11 p5, L11 p13, L11 p6, L11 p15]

i. Duplicated Circular chromosomes being interlinked

• **Interlinked chromosomes**: replicated circular chromosomes are interlinked like linked rings. Resolved two ways: (1) **topoisomerase** *(enzyme that cuts and rejoins DNA)* cuts both strands of one chromosome; (2) **XerCD** recombination (repair) enzymes cause recombination and resolution. [L11 p13]

ii. Cell division occurring faster than replication time

• **Division faster than replication**: in rapidly growing cells another round of replication starts before the earlier round finishes and before division. E. coli can divide every 20 min even though it takes 40 min to replicate its DNA. [L11 p6]
• **Consequence**: just before division 4 origins *(replication start sites)* and 1 terminus *(replication end point)* are seen, so each daughter cell receives partially replicated chromosomes. [L11 p8]

✎ **Worked exam case (Fa25 Q19)**: doubling every 30 min, 60 min to replicate. Just before cytokinesis *(cell splitting in two)* there are two complete copies of the chromosome, each already halfway replicated, so 4 copies of regions near the origin (answer D). [Key Fa25 Q19]

2. Plasmids –

• **Plasmids**: extrachromosomal *(outside the chromosome)* DNA that must replicate and be inherited on its own; section 6 of his replication lecture covers features, replication and inheritance. [L11 p17]

a. features, how differ from chromosomes

• **Features**: (1) extrachromosomal *(outside the chromosome)*, carrying genes that are not essential, supplementary function (antibiotic or metal resistance, conjugation sex pili); (2) frequent in bacteria and archaea (one study: 40% of bacteria), presence varies within a species, and they can be transferred between strains *(genetic variants)* and species. [L11 p17]
• **(3) Size 1 to 2000 kb** *(thousand base pairs)* vs chromosomes 120 to 16,000 kb: most plasmids are smaller but the ranges can overlap. **(4) Number of different plasmids**: can be several (over 20); Borrelia has 21. [L11 p19, L11 p20]
• **(5) Usually circular and negatively supercoiled** *(underwound, twisted on itself)***; (6) most replicate autonomously** *(on their own)***, not tied to chromosome replication; (7) copies per cell vary: most low copy (1 to 10), some high copy (50 to 700); (8) a single origin of replication (ori); (9) replicate only in a specific host range** *(species it can live in)***.** [L11 p21]
• **How to tell a plasmid from a chromosome**: by the function of its genes (essential or not) and independent replication, not by size, since the size ranges overlap. [L08 p10, Key Fa25 Q9, Key Fa24 Q16]

✎ **His plasmid size numbers differ between slides**: L08 p10 says usually under 100 kb with some as large as 500 kb; L11 p19 says 1 to 2000 kb. L08 p9 says plasmids occur in up to about 50% of bacterial species; L11 p17 says one study found 40%. [L08 p9, L08 p10, L11 p17, L11 p19]

	Chromosome	Secondary chromosome	Plasmid
Genes	Essential genes	At least one essential gene	Non-essential, ancillary
Size	120 to 16,000 kb	Smaller chromosome of a multi-chromosome cell (e.g. 1,072 kb in V. cholerae)	1 to 2000 kb, usually under 100 kb
Replication	One origin, bidirectional	Similar to plasmids	Autonomous, single ori; bidirectional or rolling circle
Copies	1 (more near origin in fast growth)		1 to 10, some 50 to 700

Table source: [L08 p10, L08 p11, L08 p12, L11 p5, L11 p19, L11 p21, L11 p22]

b. Methods of replication

• **Two replication mechanisms**: (1) bidirectional *(both ways)* from the origin, like circular chromosomes; (2) **rolling circle**: replication goes in one direction and synthesizes one strand, then the complementary strand is synthesized. [L11 p22]

• **Control**: a single ori *(replication start site)*; most replicate autonomously *(on their own)*, not tied to chromosome replication (Fa24 Q4: replication always tied to the chromosome is NOT a plasmid feature), though for some, replication can be tied to chromosomal replication as an inheritance mechanism. [L11 p21, L11 p23, Key Fa24 Q4]

c. Methods of segregation/retention

• **Inheritance mechanisms**: (1) many copies, rely on chance; (2) replication tied to chromosomal replication; (3) Par filament partitioning system *(protein fiber that splits plasmids)*; (4) carry genes that enable survival (drug resistance); (5) toxin-antitoxin addiction module *(poison plus antidote gene pair)*. [L11 p23]

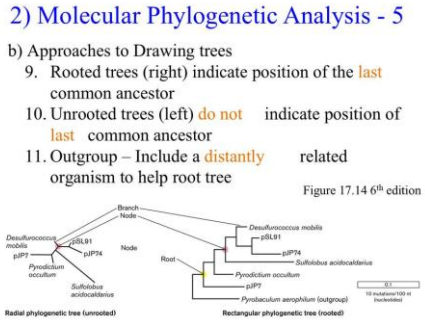
• **Par partitioning**: the plasmid encodes ParR, which binds a plasmid sequence (parC), and ParM, which forms a filament that pushes plasmids to opposite ends of the cell. [L11 p24]

• **Toxin-antitoxin addiction module**: the plasmid encodes a long-lived toxin that kills the cell if the plasmid is lost, and a short-lived antitoxin that blocks the toxin while its levels are high enough (cell still has the plasmid). Example: Phd antitoxin inhibits Doc toxin; after plasmid loss a protease *(protein-cutting enzyme)* degrades Phd and Doc causes cell death. Some TA modules are on the chromosome; their role is not known (perhaps bacterial programmed cell death). [L11 p25, L11 p26]

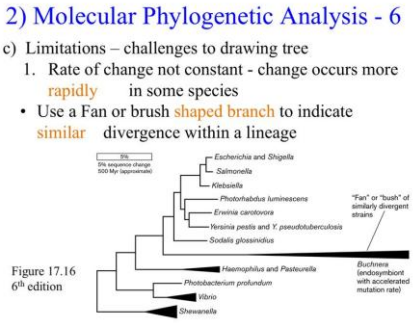
• **Exam link**: a plasmid may carry a toxin-antitoxin module (true option in Fa24 Q4). [Key Fa24 Q4]

G. Bacterial Phylogeny and Taxonomy (Chapter 17.5, 17.3 and 18.1 6e)

• **Taxonomy**: the scientific field that deals with naming, defining and classifying groups of organisms based on shared characteristics. It includes (1) nomenclature *(naming rules)*, (2) classification (usually based on phylogeny, evolutionary relationships), (3) identification, (4) the current bacterial phylogeny classification. [L11 p29, L12 s5]



Rooted (rectangular) vs unrooted (radial) tree: branch, node, root and outgroup (Figure 17.14) [L12 s19]



Fan or brush shaped branches show similar divergence within a lineage; Buchnera has an accelerated mutation rate (Figure 17.16) [L12 s20]

1. Classification – groupings

• **Classification**: placement of different kinds of life into categories, usually based on phylogeny (evolutionary relationships). [L11 p29]

• Bacteria use the same classification system as other organisms, with a **strain** *(one lineage within a species)* name or designation added after the species designation. [L11 p31]

• Subclass and suborder also exist between the main ranks (e.g. Actinobacteridae, Streptomycineae). [L11 p31]

• **Two types of features used for bacteria**: (1) phenotypic traits *(observable features)* and ecological niche (cell wall, metabolism; where it lives, what it infects, what disease it causes); (2) genetic relatedness, an estimate of phylogeny from similarity of DNA sequences such as the 16S rRNA gene *(small ribosome subunit RNA gene)*. Genetic relatedness is used more frequently. [L11 p37, L12 s6]

Rank	Long-studied taxon	Less-studied taxon	Uncultivated sample
Domain *(top-level group of life)*	Bacteria	Archaea	Bacteria
Division (phylum)	Actinobacteria	Euryarchaeota	Proteobacteria
Class	Actinobacteria	Methanococci	Alphaproteobacteria
Order	Actinomycetales	Methanococcales	Rickettsiales
Family	Streptomycetaceae	Methanocaldococcaceae	SAR11 cluster
Genus	Streptomyces	Methanocaldococcus	Pelagibacter
Species *(group of closely related species)*	S. coelicolor (1908)	M. jannaschii (1984)	P. ubique (2002)

Table source: [L11 p31]

• criteria used for a new species

• **Animals and plants**: compare features, morphology *(shape and structure)* and properties, compare DNA; same species if they can interbreed (more complicated than that). [L11 p35]

• **Why bacteria are harder**: few morphological properties, no sexual reproduction, so a species is harder to define (new species vs just a new strain). [L11 p36]

✎ **His Fa24 key uses 97% for 16S**: "look for a similarity less than 97%" for the 16S gene, and orthologs *(same gene in different species)* less than 95%. The lecture slides say 95% (and call it the genus cutoff). Know both. [Key Fa24 Q28, L11 p39]

• **16S (SSU) rRNA comparison**: compare % identity of the best alignment *(lining sequences up side by side)* of the small subunit rRNA gene (16S in bacteria, 18S in eukaryotes). Some regions are highly conserved *(barely changed over evolution)*, so it changes more slowly than most orthologs: better for distantly related species; book says only good to genus level. [L11 p40, L12 s8]

• **Orthologs**: genes in different species derived from a common ancestor gene; mutations arise as speciation occurs; any gene works if all share the common ancestor; better for closely related species. **Paralogs**: genes duplicated within a species, each can evolve different functions. [L11 p41, L12 s9]

• **DNA-DNA hybridization**: denature *(separate the two strands)* and shear *(break into pieces)* genomic DNA of one strain and attach to a membrane; denature and shear the second genomic DNA and label it (fluorometric or radioactivity); add to the membrane, let hybridize, lower temperature to anneal *(strands pair back up)*; determine the % that stuck (annealed). [L11 p42, L11 p43, L12 s11]

• **Model answer, two procedures for a new species**: PCR *(method that makes millions of DNA copies)* amplify and sequence the 16S rRNA gene and compare to known species; compare other genes (orthologs); sequence the entire genome and compare; hybridize its genome to its most similar species; show it came from a different habitat. [Key Fa24 Q28]

Test	Same as nearest species if	Novel if	What it tells you
SSU (16S) rRNA identity *(match of small ribosome RNA gene)*	>= 95%	< 95%	Different genus
ANI of ortholog genes *(average nucleotide identity: DNA match percent)*	>= 95%	< 95%	Different species
Ecotype (habitat)	Similar habitat	Different habitat	Supports new species
DNA-DNA hybridization (old) *(how well two genomes pair up)*	> 70%	< 70%	Old method

Table source: [L11 p38, L11 p39, L12 s7]

2. Nomenclature - rules

• **Nomenclature**: the naming of groups of organisms. Overseen by the **International Committee on Systematics of Prokaryotes (ICSP)**; new species or strain names must be approved by it, usually by publishing a description of the organism in the *International Journal of Systematic and Evolutionary Microbiology*. [L11 p29, L11 p30]

• **Terms for new or poorly characterized isolates** *(strains grown from a sample)*: "unclassified" or "uncultured" for minimally characterized bacteria that may not have been grown in pure culture *(growth of one species only)* (Uncultured actinomycete); an environment description plus a number for uncharacterized environmental isolates (Oil degrading bacterium AOB1); **Candidatus** *(Latin: candidate)* Genus species for a proposed or provisional name of a likely novel isolate awaiting acceptance (Candidatus Nostocoida limicola). [L11 p30, L11 p33]

Level	Capitalize?	Italicize?	Example
Genus	Yes	Yes	Escherichia
Species	No (lower case)	Yes	coli
Strain	As written (letters and numbers)	No	K12
Family or higher	Yes	No	Enterobacteriaceae
Informal group	No	No	enterics

Table source: [L11 p32]

3. Molecular Phylogenetic analysis (17.3)

• **Classifying by phylogeny**: classify by evolutionary relationships; draw phylogenetic trees *(family trees of species)* with clades *(groups sharing one ancestor)* to show divergence (ancestry). [L12 s12]

• **Molecular phylogeny**: use DNA sequence information to derive the evolutionary relationship of species or strains; compare sequences and use the number of differences (mutations) to infer the relationship. [L12 s13]

- ### Molecular clock – concept, assumptions

• **Molecular clock concept**: DNA changes accumulate over time at a steady rate, so more differences = more time since divergence. [L12 s13]

• **Assumptions**: (1) the gene chosen has retained the same function in both organisms; (2) generation time *(time for one cell division)* is about the same; (3) mutation rate remains constant among species and across generations; (4) changes accumulate over time due to random errors in replication (mutations), passed by vertical transmission *(parent to offspring)*. [L12 s14]

• **Wrong-assumption traps on his exams**: "the gene should have different functions across organisms" is NOT an assumption (Fa24 Q13); "generation time must be different for all organisms" is NOT an assumption (Fa25 Q7). "The gene was transmitted vertically from a common ancestor" IS one. [Key Fa24 Q13, Key Fa25 Q7]
• **Using the clock**: align sequences to homologous regions *(matching stretches from a shared ancestor)* (fewest mismatches), count differences, use divergence to propose a tree. His example: similarity 27/29 = 93% gives 7% divergence, 26/29 = 90% gives 10%, 24/29 = 83% gives 17%; the tree joins (1) and (2) at 7%, adds (3) at 10% and (4) at 17%. [L12 s15]
• **Clock violated, deep-branching thermophiles** *(heat-loving microbes)*: they look very old (low identity to all other bacteria), but the divergence time is thought to be overestimated because of a faster mutation rate (rate not constant) and a shorter generation time (more replications, more mutations). [L12 s62]

- ## Drawing phylogenetic trees and their interpretation

• **Building the tree**: use a mathematical analysis to find the best or most probable tree. **Maximum parsimony**: the "best fit" tree, the one requiring the fewest mutations or changes to fit the data (multiple trees could be generated). **Maximum likelihood**: the most probable (likely) tree based on the DNA sequences. [L12 s16]
• **Reading the tree**: branches represent lineages *(lines of descent)*; nodes are the common ancestor before divergence; branch length to a node = number of mutations (time) from the ancestor. Scale bar example: 0.1 = 10 mutations/100 nt *(nucleotides: DNA letters)*. [L12 s17]
• **Tree shapes**: rectangular (Fig 17.14 right) or radial (left). [L12 s18]

✎ **His Classification of Bacteria tree (Fig 18.2, 4e)**: rectangular tree, rooted *(anchored to a shared ancestor)* using Archaea and Eukarya, phyla *(major groups)* in blue, based on 16S rDNA *(gene for small ribosome RNA)*. Archaea trees were rooted with Bacteria and Eukarya. [L12 s46, L12 s51]

- ## Know the terms: nodes, clades, common ancestors, outgroup, rooted

• **Exam wording**: a tree that includes one distantly related organism is a **rooted tree** (Fa24 Q18); "rooted" means the tree shows the position of the last common ancestor of all organisms by including a distantly related organism (Fa25 Q11). [Key Fa24 Q18, Key Fa25 Q11]

Term	Meaning
Branch	A lineage
Node	Common ancestor of the organisms before they diverged
Common ancestor	Shown by a node; the root is the last common ancestor of all
Clade	Monophyletic group: related organisms sharing a common ancestor not shared by other groups
Rooted tree	Shows the position of the last common ancestor
Unrooted tree	Does not show the position of the last common ancestor
Outgroup	A distantly related organism included to help root the tree (e.g. Pyrobaculum aerophilum in Fig 17.14)
Branch length	Number of mutations (time) from the ancestor

Table source: [L12 s12, L12 s17, L12 s19]

- ## Limitations/challenges to Molecular phylogenetic analysis

• **1. Rate of change not constant**: change is faster in some species (e.g. Buchnera, an endosymbiont with an accelerated mutation rate). A **fan or brush shaped branch** shows similar divergence within a lineage *(line of descent)*. Scale in Fig 17.16: 5% sequence change is about 500 Myr *(million years)*. [L12 s20]
• **2. Horizontal transfer**: prevalent in some bacteria, which get DNA from other species and add the genes to their own genome. [L12 s21]
• **3. Genome evolution is both vertical and horizontal**: some genes are gained or lost by horizontal transfer. [L12 s22]
• **4. The gene used affects the result**: if strain A shares an ancestor with B but replaced its gene 15 with gene 15 from species L, a gene-15 tree wrongly shows A closer to L than to B. Solutions: use informational genes *(genes for making RNA and protein)* (e.g. 16S rRNA) passed vertically; use many genes or whole genomes. [L12 s23]
• **Deep-branching thermophiles** *(heat lovers that split off early)* show the clock problem: faster mutation rate and shorter generation time make the divergence time estimate too high. [L12 s62]

Gene type	Encodes	Interactions	Horizontal transfer
Informational (better for trees)	Products essential for transcription and translation	Interact with many cell components	Less likely
Operational	Metabolism, stress response, pathogenicity	Function relatively independently	More likely

Table source: [L12 s24]

- ## Methods used for Identification of bacteria

• **Practical identification** combines genetic and phenotypic *(observable)* traits. [L12 s26]

• **Dichotomous key**: a series of yes/no decisions narrows the possible species; most traits are phenotypic (morphology, cell size, motility); his example key goes down to genus. [L12 s26, L12 s27]
• **Probabilistic indicator** (used more often now): a battery of biochemical tests is run at once on the isolate *(the cultured strain being tested)*; needs a predefined database of known bacteria; the probabilities of all results (positive or negative) are multiplied to give a score for each species in the database, and the top-scoring species is the best identification. Numbers in the table = probability a member of that species gives that result. [L12 s28, L12 s29]
• **Genotypic identification**: compare the 16S rRNA gene sequence to known bacteria (Assignment 8 Part C used a 16S rRNA sequence to identify a bacterium), plus orthologs *(same gene in different species)*, ANI *(average nucleotide identity: genome match percent)* and DNA-DNA hybridization. [L09 p39, L11 p37, L11 p40]

4. Features commonly used to group bacteria into their largest groupings (phyla)

• **Current system** is based on phenotypic *(observable)* and genetic information. Life has three domains: Bacteria, Archaea *(single-celled microbes distinct from bacteria)*, Eukarya *(organisms whose cells have a nucleus)*. Archaea share some traits with bacteria and some with eukaryotes. [L12 s30, L12 s31]
• **Features used to separate bacteria into phyla** *(major groups)* (Table A2.2 / 18.1): cell envelope or Gram stain *(dye test that sorts cell walls)* (Gram-positive: multiple peptidoglycan layers cross-linked by teichoic acids; Gram-negative: outer membrane with LPS; some lack peptidoglycan); %GC *(share of G and C bases)* (Firmicutes low GC, Actinobacteria high GC); shape (Spirochetes narrow coiled cells with axial filaments); growth temperature (deep-branching thermophiles, Aquificae 70-95 C); nutrition and energy source (Cyanobacteria oxygenic photoautotrophs, lithotrophs, heterotrophs); motility *(ability to move)*; deep branching. [L12 s36, L12 s37, L12 s39, Key Fa25 Q29]
• **Terms**: autotrophs fix CO2; heterotrophs use preformed organic molecules; oligotroph grows best at low organic nutrient levels; phototrophs use light; chemotrophs use chemicals (redox); lithotrophs use inorganic electron donors (H2O, iron, sulfur, nitrogen); organotrophs use organic molecules (sugars); **deep branching** = split from other lineages long ago; high or low %GC genome. [L12 s41, L12 s42, L12 s43]
• **New in the 5th and 6th editions: Candidate Phyla Radiation (CPR)**: 30-100 phyla known only by DNA sequence, no cultured isolates, ultrasmall cells that pass a 0.2 um filter, probably obligate symbionts *(cannot live without a host partner)*. [L12 s47]
• **Archaea phyla (5e/6e)**: TACK superphylum *(group of related phyla)* with Crenarchaeota (widest temperature range) and Thaumarchaeota (mesophilic heterotrophs, sulfur and ammonia oxidizers); Euryarchaeota (greatest range of metabolism: methanogens, halophiles, acidophiles, alkalinophiles); Asgard (vent thermophiles); DPANN (obligate symbionts). [L12 s61]
• **NOT a phylum feature**: number of chromosomes (Fa24 Q10). [Key Fa24 Q10]

★ **On his past exams**: Fa24 Q8 (genus and species are italicized); Fa25 Q5 (format Borrelia burgdorferi ZS7, family Spirochaetaceae: italic genus and species only); Fa24 Q28 (two procedures for a new species: 16S under 97%, orthologs, whole genome, hybridization, habitat); Fa24 Q13 and Fa25 Q7 (NOT clock assumptions: gene with different functions; different generation times); Fa24 Q15 (fan or brush shaped branch); Fa24 Q18 (distantly related organism makes a rooted tree); Fa25 Q11 (rooted = shows last common ancestor); Fa24 Q10 (number of chromosomes NOT a phylum feature); Fa25 Q29 (phylum features: envelope, %GC, shape, thermophily, energy source, motility, deep branching). [Key Fa24 Q8, Key Fa25 Q5, Key Fa24 Q28, Key Fa24 Q13, Key Fa25 Q7, Key Fa24 Q15, Key Fa24 Q18, Key Fa25 Q11, Key Fa24 Q10, Key Fa25 Q29]

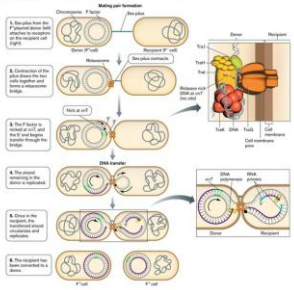
Trait	Bacteria	Archaea	Eukarya
DNA chromosome	Circular (usually)	Circular (usually)	Linear
DNA organization	Nucleoid	Nucleoid	Nucleus with membrane
Introns *(noncoding stretches inside genes)*	Rare	Common	Common
RNA polymerase	Bacterial	Eukaryotic homologs	Eukaryotic homologs
Translation initiator *(first amino acid of a protein)*	Formylmethionine	Methionine	Methionine
Methanogenesis *(making methane)*	No	Yes	No
Thermophilic growth	Up to 95 C	Up to 120 C	Up to 80 C
Membrane lipids	Ester-linked fatty acids	Ether-linked isoprenoids	Ester-linked fatty acids
Pathogens	Many	None	Many

Table source: [L12 s31]

H. Bacterial Genome Evolution (Chapter 9)

a) Conjugative Plasmids - 3

i. F plasmid conjugation in *E. coli* (Gm Neg)
Figure 9.3
4th edition



F plasmid conjugation: pilus, relaxosome, nick at oriT, single-strand transfer, both cells F+ (Figure 9.3, 4e) [L13 s26]

a). Generalized Transduction - 1

1. Phage infects bacterial cell.
2. Synthesis of new phage components and breakdown of chromosomal DNA.
3. Phage DNA and occasionally chromosomal DNA (gray) is packaged into phage heads.
4. Assembly of phages is completed.

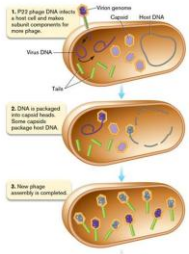


Figure 9.22 6th edition top

Generalized transduction: phage breaks down host DNA and some capsids package host DNA (Figure 9.22) [L13 s40]

1. How do bacterial genomes change?

• **Sources of genome change**: mutations from errors in replication; mutations from chemical alterations of bases; chromosomal rearrangements *(large pieces moved or flipped)*; insertion of transposable *(able to jump positions)* and viral elements; getting DNA from other organisms (gene transfer); sexual reproduction (eukaryotes only). [L13 s4]

• **Model answer, three mechanisms and effect**: (1) lateral exchange (conjugation, transformation, transduction): the cell picks up DNA from another cell and can incorporate it; (2) base change by mutation: one base pair altered; (3) transposition *(jumping DNA moving)* into a new site: disrupts the new site; (4) chromosomal rearrangement, deletion or duplication: varies. [Key Fa25 Q27]

• **Genome is a mosaic**: most DNA is inherited vertically with the species' characteristic %GC; some pieces came by horizontal transfer from other species (pathogenicity or genomic islands) and may have a different %GC and codon usage *(preferred three-letter DNA words)*; a few transposable elements and integrated viruses; rearrangements, usually inversions *(flipped segments)*. [L13 s8]

• **Two gene pools**: core genome (essential functions: ribosomes, cell envelope, key metabolic pathways, DNA replication) vs flexible or accessory pool (non-essential, ancillary, frequently acquired horizontally, varies between strains: genomic islands, bacteriophages, plasmids, integrons, transposons; encode pathogenicity, antibiotic resistance, secretion, symbiosis, degradation). [L13 s12]

• Horizontal gene transfer plays a more prominent role in the evolution of bacteria than of eukaryotes. [Key Fa24 Q12]

Change	Acquired
Mutations	Vertically
Chromosomal rearrangements	Vertically
Gene transfers	Horizontally
Viral elements	Horizontally

Table source: [L13 s5]

2. Gene Transfer or Exchanges – lateral or horizontal gene exchanges (9.3 6e)

• **Horizontal (lateral) gene transfer**: transfer of DNA from one cell (donor) to another cell (recipient); generates genetic diversity in bacteria; not part of reproduction. [L13 s13]

• **Three major types**: conjugation (conjugative plasmid transfer, Hfr), transduction (generalized, specialized), transformation (natural competence, artificially induced competence); plus other mechanisms. [L13 s17]

• **Other mechanisms**: nanotubes *(tiny membrane tubes)* between cells; membrane vesicles *(small membrane bubbles)* that may contain DNA; gene transfer agents (protein similar to phage shell protein forming virus-like particles). [L13 s56]

a. How are they recognized in genomes?

• **Different %GC** *(share of G and C bases)*: a region with a %GC different from the rest of the genome suggests recent acquisition by horizontal transfer from a different species (e.g. pathogenicity islands). [L13 s10, L08 p8, L09 p37, Key Fa25 Q8]

• **Different codon usage** *(preferred three-letter DNA words)* in the acquired piece. [L13 s8]

• **Genomic islands** *(inserted blocks of foreign genes)* found inserted into genomes in whole-genome comparisons (compare to other strains or close species, e.g. E. coli O157, E. coli K12, Shigella: grey = shared, colored = differences). [L13 s11, L09 p38]

• Example: in pathogenic *(disease-causing)* E. coli O157:H7 the red genes are hypothesized to have been acquired by horizontal transfer. [L13 s9]

b. What are its features? How is it different from sexual reproduction?

• **Features of gene transfer**: (1) generates genetic diversity; (2) not part of reproduction; (3) one way, donor to recipient; (4) transfers a small amount of DNA, not a complete copy of the chromosome; (5) rare, only a few cells in a million pick up donor DNA; (6) random, the recipient usually gets a random piece; (7) usually between closely related cells (same or close species); (8) the transferred gene usually gets expressed, since there is only one version of each chromosome so donor alleles *(gene versions)* are expressed; (9) may require replacement of the chromosomal gene by homologous recombination *(swapping in matching DNA)*; (10) evidence says it occurs frequently in bacteria. [L13 s13, L13 s14, L13 s15, L13 s16]

• **How it differs from sexual reproduction** (eukaryotes): not part of reproduction, only one direction (donor to recipient), and only a small part of a chromosome moves, never a full copy. [L13 s4, L13 s13, Key Fa24 Q1]

• **Traps**: the recipient does NOT receive a complete chromosome copy (Fa24 Q1); the donor is NOT always alive and active (Fa25 Q1, transformation uses DNA from dead cells). [Key Fa24 Q1, Key Fa25 Q1]

c. For each know features and steps, and source of donor DNA

• **Features that vary between mechanisms**: donor alive or dead; direct contact needed or not; type of DNA (plasmid, phage, chromosomal, or a combination); pili *(hair-like surface tubes)* or bacteriophage *(virus that infects bacteria)* involved; whether DNA can enter some eukaryotic cells; whether the DNA replaces the recipient gene by homologous recombination *(swapping in matching DNA)*. [L13 s18]

Mechanism	Donor alive or dead	Contact needed	DNA transferred	Pilus or phage	Recombination needed	Into eukaryotes
Conjugative (F) plasmid *(plasmid that transfers itself)*	Alive	Yes, direct	Plasmid only (circular, no chromosomal DNA)	Pilus (T4SS)	No, plasmid stays separate	Yes, some (Agrobacterium Ti plasmid into plants)
Hfr conjugation *(high frequency of recombination)*	Alive	Yes, direct	Chromosomal DNA, starting in the integrated F	Pilus	Yes, homologous (double crossover)	Not stated
Generalized transduction *(phage carries random host DNA)*	Dead (cell lyses)	No	Random chromosomal DNA instead of phage DNA	Phage	Yes	Not stated
Specialized transduction *(phage carries DNA beside its site)*	Dead (prophage excises and goes lytic)	No	Phage DNA plus nearby chromosomal DNA	Phage	Yes	Not stated
Transformation *(uptake of free DNA)*	Dead (or DNA secreted)	No	Free naked DNA (chromosomal)	Neither	Yes, must be homologous	Artificial: yeast, other eukaryotes

Table source: [L13 s20, L13 s32, L13 s36, L13 s39, L13 s41, L13 s42, L13 s46, L13 s51, Key Fa24 Q3, Key Fa25 Q3]

1) Conjugation

• **Conjugation**: direct transfer of DNA from a live donor cell to a live recipient; requires contact; usually transfers a plasmid *(small extra DNA circle)*; often through a protein channel (pili); can reach many species of bacteria and some eukaryotic cells. [L13 s20]

• Documented in 1946 by Joshua Lederberg (Nobel prize): mixed two strains with different mutations and got a strain with no mutations. [L13 s21]

• Seen in many Gram negatives *(bacteria with an outer membrane)*, some Gram positives *(bacteria with a thick wall)*, a few Archaea *(microbes distinct from bacteria)*. Gram negatives usually use a special pilus and a plasmid; Gram positives use a different mechanism. [L13 s22]

• **Types**: conjugative plasmids in Gm neg (plasmid encodes the pili for transfer); Hfr (transfers chromosomal DNA); Gm+ mechanism; transfer from bacteria to eukaryotic cells. [L13 s23]

• **Gram-positive conjugation** (Enterococcus faecalis): no sex pili; plasmid involved; regulated by a quorum sensing *(cell density signaling)* pheromone *(signal molecule)*. The recipient makes pheromone, which stimulates the donor (has plasmid) to make aggregation substance *(sticky surface protein)*; cells stick together and the plasmid is transferred. [L13 s34, L13 s35]

• **Into eukaryotes**: Agrobacterium tumefaciens transfers the Ti plasmid *(tumor-inducing plasmid)* into plant cells; it integrates into their chromosome and codes for growth factors, causing crown gall tumor *(lumpy plant tumor)*. Now used to genetically modify plants. [L13 s36]

• Conjugation = donor transfers plasmid DNA via a pilus, or a little chromosomal DNA along with plasmid DNA via a pilus (answer A and B). [Key Fa24 Q7]

• Steps in F plasmid conjugation and Mobilizable plasmids

• **F plasmid conjugation steps (E. coli)**: (1) the F+ donor *(has the F plasmid)* synthesizes and assembles a pilus (components secreted by the Type 4 secretion system) that attaches to the F- recipient *(lacks the F plasmid)*; (2) the pilus retracts, drawing cells together and forming a relaxosome bridge *(protein complex that starts transfer)*; (3) relaxase nicks the F plasmid at oriT *(origin of transfer site)* and one strand transfers into the recipient; (4) donor and

recipient each make the complementary strand, regenerating double-stranded plasmid in both; (5) both cells express plasmid genes, make pilus, and are F+. [L13 s26, L13 s27]

• **Mobilizable plasmids**: can be transferred only if a conjugative plasmid is present; they lack the genes for pili and T4SS *(type 4 secretion system: export channel)*. [L13 s29]

• **Triparental conjugation**: a helper strain with a conjugative plasmid transfers a mobilizable plasmid into a recipient, creating a recipient with the mobilizable plasmid. [L13 s30]

• F plasmid conjugation: pilus contact, a circular extrachromosomal piece moves and no chromosomal DNA (Fa25 Q3); the plasmid stays separate and does not recombine into the recipient chromosome (Fa24 Q3). [Key Fa25 Q3, Key Fa24 Q3]

• Features of Conjugative Plasmids

• **Conjugative plasmids**: encode the proteins needed for conjugation: pili, T4SS *(type 4 secretion system: export channel)*, relaxase *(strand-cutting enzyme)*, and oriT. Most relatively large (~100 kb); 13 tra genes *(transfer genes)* make the pili, T4SS and other transfer factors; oriT = sequence required for transfer; carry IS (insertion sequence) elements. [L13 s24, L13 s28]

• Belong to different **incompatibility groups**, which prevent transfer to cells with the same type of plasmid. Many carry antibiotic resistance genes. Best studied: F plasmid in Gm neg *(Gram-negative bacteria)*. [L13 s24]

• Plasmids carry non-essential genes such as antibiotic or metal resistance and conjugation sex pili, and can be transferred between strains and species. [L11 p17]

• Steps in Hfr conjugation

• **Hfr = high frequency of recombination**: moves chromosomal genes by conjugation at a relatively high frequency compared with transformation or transduction; needs a strain with the F plasmid integrated into its chromosome; used to map genes. [L13 s31]

• **Steps**: (1) strain has F in its chromosome; (2) attaches to an F- cell (F pilus); (3) nick in the F DNA; (4) transfer starts (T4SS moves one strand, second strand made in recipient); (5) portions of the chromosome transfer, amount varies; genes close to F transfer most often; (6) donor DNA exchanges alleles *(gene versions)* into the recipient chromosome by homologous recombination (double crossover), the rest of the linear DNA is lost. [L13 s32]

• **Mapping**: stop transfer at different times by agitating *(shaking)* the culture; the entire E. coli chromosome takes 100 minutes. [L13 s33]

2). Transduction

• **Transduction**: mediated by bacteriophages (bacterial DNA viruses) that pick up bacterial chromosomal DNA instead of phage DNA and carry it to a recipient; requires homologous recombination (replacement of recipient DNA). Two mechanisms: generalized and specialized. [L13 s37]

• Generalized transduction = phage particle carrying bacterial chromosomal DNA instead of phage DNA (Fa25 Q2). Specialized transduction = a small portion of chromosomal DNA transferred along with phage DNA (Fa24 Q2). [Key Fa25 Q2, Key Fa24 Q2]

• Features and steps in generalized and specialized transduction

• **Generalized steps**: (1) phage infects; (2) new phage parts made and chromosome broken down; (3) phage DNA and occasionally chromosomal DNA packaged into heads; (4) assembly completed; (5) cell lyses *(bursts open)* and releases phages; (6) a few phages inject donor chromosomal DNA into a different cell; (7) it recombines into the chromosome. [L13 s40, L13 s41]

• **Specialized steps** (lambda, Fig 9.23): (1) prophage *(phage DNA inserted in the chromosome)* pops out taking nearby chromosomal DNA (gal); (2) phage plus chromosomal DNA go into the phage head; (3) particle injects into recipient; (4) DNA recombines into the chromosome. [L13 s42]

Step	Generalized	Specialized
Start	Phage infects cell (lytic)	Prophage already integrated in chromosome (similar to Hfr)
DNA picked up	Host chromosome broken down; occasionally a random chromosomal piece packaged into a phage head	Phage pops out and takes nearby chromosomal DNA with it (aberrant excision)
Particle carries	Chromosomal DNA only	Phage DNA plus chromosomal DNA
Release	Cell lyses, phages released	Phage head injects into recipient
In recipient	Injected DNA recombines into chromosome	Injected DNA recombines into chromosome (integrates at attB)

Table source: [L13 s40, L13 s41, L13 s42]

• Normal life cycle of Bacteriophages

• **Bacteriophages**: viruses of bacteria; most double-stranded DNA; complex structure; inject DNA through the cell wall into the cytoplasm. [L13 s38]
• **Two options**: lytic (done by all phages) and lysogenic *(quietly inserts into host DNA)* (option for temperate phages). Lytic: (1) attach and inject DNA; (2) phage DNA circularizes; (3) make viral nucleic acid and proteins, assemble virions *(virus particles)*; (4) cell lyses, releasing virions. Lysogenic: (5) viral DNA integrates into the chromosome (prophage); (6) lysogenic bacterium can grow normally; (7) on occasion the prophage excises *(cuts itself out)* and starts a lytic cycle. [L13 s39]

3) Transformation - competence

• **Transformation**: a cell takes up free, "naked" DNA from its environment; the DNA can replace similar chromosomal genes by homologous recombination *(swapping in matching DNA)*. Discovered before DNA was known to be the genetic material. [L13 s43, Key Fa24 Q6]
• **Griffith 1928**: smooth (capsule) cells = disease; rough (mutant, no capsule) = no disease; heat-killed S = no disease; heat-killed S + live R = disease. **Avery, MacLeod and McCarty 1944** used transformation to help show DNA is the genetic material. [L13 s44, L13 s45]
• **Natural competence**: cells naturally develop the ability to take up DNA in response to conditions; shown in over 80 species of bacteria and some Archaea *(microbes distinct from bacteria)*. **Artificial competence**: made in the lab by specific treatment, used for genetic engineering; works in bacteria, yeast, other eukaryotes. [L13 s46, L13 s47]
• **Bacillus subtilis**: competence develops when nutrients are limiting and cell density is high; 10-20% of the population become competent on entering stationary phase *(growth has stopped)* (others sporulate); induces a DNA uptake system of 20-30 proteins; density sensed by quorum sensing *(cell density signaling)*. In Gm+ bacteria (Streptococcus) competence is regulated by the level of competence factor *(secreted signal peptide)*. [L13 s48, L13 s49, L13 s50]
• **Artificial competence methods**: chemical (Ca2+ or Li+ cations interact with DNA and membrane) and electroporation (electrical shock). [L13 s53, L13 s54]

• Features and Steps in process

• **Natural transformation steps**: (1) donor cell dies and releases DNA, or secretes DNA; (2) competent *(able to take up DNA)* recipient binds free DNA; (3) takes in a single strand; (4) recombines with homologous chromosomal DNA, replacing one strand; (5) mismatches *(wrongly paired bases)* repaired, going either to the new sequence or back to the old one. [L13 s51, L13 s52]

• What is the source of the donor DNA?

• **Source of donor DNA**: released by cells that die (randomly or by programmed cell death) or secreted. How often it happens in nature is a knowledge gap. [L13 s51, L13 s55]
• **Why transformation evolved** (open question): DNA repair (template to fix a damaged chromosome), DNA as food (C, N, P source), or gaining new genes (diversity, new functions). [L13 s55]

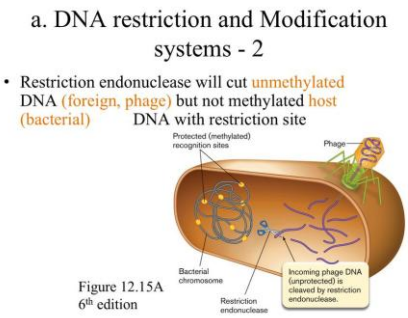
★ **On his past exams**: Fa24 Q1 (NOT a feature: complete copy of donor chromosome); Fa25 Q1 (NOT true: donor always alive); Fa24 Q2 (specialized transduction moves chromosomal DNA with phage DNA); Fa25 Q2 (generalized transduction carries chromosomal instead of phage DNA); Fa24 Q3 (F plasmid does not recombine); Fa25 Q3 (F plasmid: pilus, circular DNA, no chromosomal DNA); Fa24 Q6 (transformation takes up free DNA); Fa24 Q7 (conjugation: A and B); Fa24 Q12 (horizontal transfer more prominent in bacteria); Fa25 Q8 (different %GC = recent horizontal transfer); Fa25 Q27 (three mechanisms that change the chromosome). [Key Fa24 Q1, Key Fa25 Q1, Key Fa24 Q2, Key Fa25 Q2, Key Fa24 Q3, Key Fa25 Q3, Key Fa24 Q6, Key Fa24 Q7, Key Fa24 Q12, Key Fa25 Q8, Key Fa25 Q27]

3. Bacterial viral defensive systems (12.3 6e)

• **Two bacterial viral defense systems**: DNA restriction endonucleases with modification systems, and the CRISPR cassette system. Bacteria use both to try to protect themselves from bacteriophages *(viruses of bacteria)*. [L14 s19, Key Fa25 Q12]

a. Restriction endonucleases and modification systems

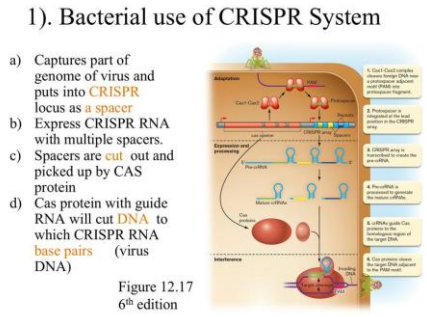
• **Restriction-modification system**: two activities. The **restriction endonuclease** *(DNA-cutting enzyme)* cuts DNA in or near a specific recognition sequence. The **modification enzyme** methylates *(adds methyl groups to)* bases in the recognition sequence, which blocks cutting. Some have both activities in the same protein. [L14 s20]
• **Result**: the restriction endonuclease cuts unmethylated (foreign, phage) DNA but not methylated host (bacterial) DNA that carries the same restriction site. His example: EcoRI cuts GAATTC, making staggered nicks in the phosphodiester backbone *(sugar-phosphate chain)*; the methylase adds methyl groups on adenines, protecting the site. [L14 s20, L14 s21]
• Restriction-modification is also listed as a feature of the flexible gene pool *(optional, variable genes)* (Table 9.4). [L13 s12]



Phage DNA is cut; methylated host sites are protected (Fig 12.15A) [L14 s21]

b. CRISPR-Cas9 systems

• **CRISPR** = Clustered Regularly Interspaced Short Palindromic Repeats. Bacteria use it to generate RNAs (guide RNAs) that mediate a defense against a particular virus (phage). [L14 s22, Key Fa24 Q19]
• **CRISPR locus**: cas genes plus an array of spacer *(short pieces copied from viruses)* sequences between repeat sequences. [L14 s23]
• **How bacteria use it**: (a) capture part of the virus genome and put it into the CRISPR locus as a spacer; (b) express CRISPR RNA with multiple spacers; (c) spacers are cut out and picked up by a Cas protein *(CRISPR-associated nuclease)*; (d) the Cas protein with its guide RNA cuts the DNA to which the CRISPR RNA base pairs (the virus DNA). [L14 s24]
• **Genome editing with CRISPR-Cas9**: bacterial Cas9 protein plus an engineered guide RNA introduce a double-strand break at a specific site, in a test tube or in live cells, in many prokaryotic and eukaryotic organisms. Uses: knock out a gene, introduce a mutation, introduce a gene into the genome, fix a disease gene. [L14 s22, L14 s25]



Bacterial use of CRISPR: adaptation, processing, interference (Fig 12.17) [L14 s24]

★ **On his past exams**: Fa24 Q19 (CRISPR-Cas9 protects against viruses; it does NOT mediate horizontal transfer, promote mutations in stationary phase, or transfer conjugative plasmids); Fa25 Q12 (restriction enzymes and CRISPR protect against bacteriophages; they do NOT recombine transformed DNA, move DNA in conjugation, or move transposons). [Key Fa24 Q19, Key Fa25 Q12]

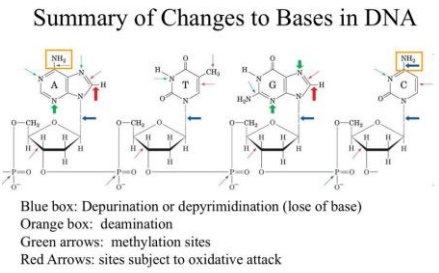
4. Mutations and DNA Repair (9.1-9.2 6e)

• **Mutation**: a heritable change in a nucleotide sequence; it can be passed on to daughter cells = vertical transmission *(parent cell to offspring)*. [L14 s27, L13 s5]
• **Mutation rate** = rate at which changes appear minus the rate of repair. A change becomes a mutation only if it is not repaired before transmission to offspring. The effect varies. [L14 s27]
• **Replication error rates**: bacteria under normal conditions make one error per 10^8 to 10^9 bases synthesized; bacterial genomes are about 10^6 to 10^7 bp, so one change occurs every 100 to 10,000 replications, and cultures quickly reach 10^7 to 10^9 cells/ml. Viral polymerases are more error prone; eukaryotic polymerases are less error prone. Error rate correlates with genome size, so an organism wants a low but not zero error rate. [L14 s29]
• **Stationary phase and DNA damage**: in some bacteria the error rate is higher in cells entering stationary phase *(growth has stopped)* or responding to DNA damage. Cells induce the error-prone DNA Pol IV and V, part of the SOS response *(emergency DNA-damage program)*, which allows replication through some damage. Hypothesis: the higher error rate increases genetic diversity. [L14 s30]
• **DNA repair on his slides**: after transformation, mismatches are repaired to either the new or the old sequence; in the SOS response, error-prone polymerases replicate through damage. Lecture 14 shows no other repair pathway. [L14 s13, L14 s30]

a. Causes

• **1) Errors by DNA pol during replication**: (a) insertion of a wrong nucleotide, from a DNA polymerase error or because a base on the template strand was modified so its base pairing is altered; (b) pol incorporates altered or damaged bases (modified bases or base analogs *(look-alike bases)* in the free nucleotide pool); (c) insertion or deletion of a nucleotide on the strand being replicated; (d) duplication or deletion of a portion of the chromosome (chromosomal rearrangement). [L14 s28]

• **2a) Depurination or depyrimidination** (spontaneous): hydrolytic *(water-driven)* removal of the base from a nucleotide in DNA, leaving an apurinic or apyrimidinic (AP) site. Depurination (A or G) is more frequent. The polymerase stalls or inserts a random nucleotide opposite the AP site. It also impairs sequencing of ancient DNA. [L14 s33, L14 s34]



His summary of the sites where bases in DNA are changed [L14 s40]

• **2b) Deamination** (spontaneous): hydrolytic loss of an amine group *(NH2, nitrogen group)* from a base. It occurs spontaneously but is also induced by alkaline solutions, elevated temperatures and certain chemicals (nitrous acid, nitrates). [L14 s35, L14 s36]

• **3) DNA-damaging chemicals**: **alkylating agents** add a methyl or ethyl group (nitrosoguanidine, which his slide spells "nitrosoquanidine"; mustard gas; EMS): at O6 of guanine (G-C to G*-T) and O4 of thymine (T-A to T*-G). **Deaminating agents** replace amine groups with oxygen (nitrous acid, nitrosamines). **Reactive oxygen species** *(damaging oxygen molecules)* (hydroxyl radicals OH, peroxide H_2O_2, superoxide O_2^-) modify bases at many different sites, remove hydrogen protons, and alter base pairing or block replication. [L14 s37, L14 s38]

• **4) Radiation**: UV light forms pyrimidine dimers *(two neighboring C or T bases fused)* that prevent transcription and replication of the strand; ionizing radiation *(X-rays, gamma rays)* causes double-strand breaks and forms hydroxyl radicals. [L14 s39]

• **Summary figure of changes to bases**: blue box = depurination or depyrimidination (loss of base); orange box = deamination; green arrows = methylation sites; red arrows = sites subject to oxidative attack. [L14 s40]

• **DNA alterations that promote mutations** (his key): loss of a base from a nucleotide; loss of an amino group from a base; alteration of bases by reactive oxygen species; linking of two adjacent pyrimidine bases. NOT one: adding an acetyl group *(small two-carbon chemical group)* to the sugar. [Key Fa24 Q20]

Base	Deaminated to	Base pair change
Cytosine	Uracil	C-G to U-A
Adenine	Hypoxanthine (inosine)	A-T to I-C
Guanine	Xanthine	G-C to X-C
5-methylcytosine	Thymine	C-G to T-A

Table source: [L14 s35]

Mutagenic agent	Example	Effect
Base analog	caffeine, 5-bromouracil	look-alike base put in during replication: point mutation
Alkylating agent	nitrosoguanidine	adds an alkyl (e.g. methyl) group, incorrect pairing: point mutation
Deaminating agent	nitrous acid, nitrates, nitrites	removes an amino group: point mutation
Acridine derivative	acridine dyes, quinacrine	inserts between base pairs, distorts helix: frameshift
UV rays	n/a	links adjacent pyrimidines (thymine dimer), impairs replication; lethal if not repaired
X-rays, gamma rays	n/a	ionize molecules to free radicals that break DNA; lethal if not repaired

Table source: [L14 s41]

b. Types of changes observed

• **Chromosomal rearrangements**: duplication and deletion both result from homologous recombination *(swapping between matching DNA)* between direct repeat sequences on the chromosome copies present during replication; one chromosome gets the duplicated region and the other loses it. Inversion of part of the chromosome also occurs (an inversion was found in two strains of E. coli O157:H7). [L14 s31, L14 s32]

• **Types of mutation** (Fig 9.2): missense point mutation (T-to-A transversion turns TTT Phe into TTA Leu); nonsense point mutation (C-to-G transversion turns TCA Ser into a TGA stop); insertion frameshift (inserting AT shifts the reading frame and changes all downstream amino acids); deletion frameshift (removing two nucleotides); inversion (flipping a DNA sequence changes the amino acids); duplication (a DNA sequence is copied). [L14 s42]

• **Effect of mutations**: varies with the mutagenic agent *(mutation-causing agent)* (Table 9.1, under Causes above). [L14 s41, L14 s43]

• **Silent (synonymous) mutations**: change the codon but not the amino acid. Hypothesized to have no effect (neutral), but a recent study (Shen et al., Nature 2022, yeast genes) suggests they can have strong effects on gene expression level and can affect mRNA levels. [L14 s43]

• **Types of change on his key**: a base change alters one base pair; chromosomal rearrangement, deletion or duplication; rearrangements seen are usually inversions *(flipped segments)*. [Key Fa25 Q27, L13 s8]

> ★ **On his past exams**: Fa24 Q20 (acetyl on the sugar is NOT a mutagenic alteration); Fa25 Q27 (base change alters one base pair); Fa24 Q12 (errors in replication and chemical alterations of bases are NOT the changes more prominent in bacteria than eukaryotes; horizontal transfer is). [Key Fa24 Q20, Key Fa25 Q27, Key Fa24 Q12]

5. Transposons and Mobile Genetic Elements (9.4 6e)

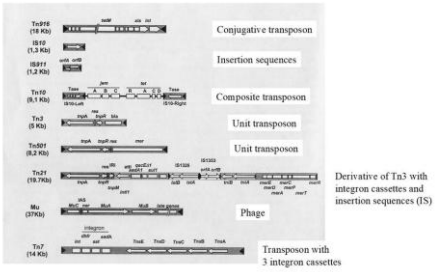
• **Transposable elements** (mobile genetic elements): two main mechanisms of transposition; several types (insertion sequences, composite transposons, complex transposons, conjugative transposons); usually scattered throughout the genome; also used for research. [L14 s44]

• Insertion of transposable and viral elements is a source of genome change; viral elements are acquired horizontally *(from another cell, not a parent)*. Bacterial genomes contain a few transposable elements (few transposons overall, >85% protein coding), unlike the human genome (~45% transposable elements or remnants). [L13 s4, L13 s5, L13 s8, L08 p16, L08 p17]

a. What are the different types and how do they vary

• His types figure also shows unit transposons, a phage, a transposon with 3 integron cassettes *(gene-capture units)*, and a derivative of Tn3 with integron cassettes and insertion sequences. Earlier pages: IS elements sit on conjugative plasmids; transposons and integrons in the flexible gene pool encode transposases *(enzymes that move transposons)* and integrases *(enzymes that insert DNA)*. [L14 s47, L13 s28, L13 s12]

b. Types of Transposable Elements



Types of transposable elements [L14 s47]

Type	What it carries	Key feature
Insertion sequence (IS)	only the genes needed for mobility, usually just a transposase	simplest; 700-2500 nt; short inverted repeats of 10-40 bp at each end
Composite transposon	two IS flanking one to several genes (often antibiotic resistance or metabolic)	often only the outermost inverted repeats work
Complex transposon	transposase + resolvase (helps resolve integration), often antibiotic resistance	resolvase gene sets it apart
Conjugative transposon	genes for conjugation	jumps out, circularizes, makes pili, transfers, integrates
Second conjugative type	only genes for integration (integrase) and an oriT	needs a conjugative plasmid to transfer, then integrates

Table source: [L14 s47, L14 s48, L14 s49, L14 s50, L14 s51, L14 s52]

b. Insertion sequences, complex transposon

• **Insertion sequences (IS)**: simplest type of transposable or mobile element; 700-2500 nucleotides long; carry only the genes needed for mobility, usually just a transposase; short inverted repeats *(same sequence reversed on the other end)* of 10-40 bp at each end; insert at specific sequences called target or attachment (att) sites; large variety, but distinct families have been identified (his example IS2 with its tnp gene). [L14 s48]

• **Composite transposon**: two insertion sequences flank one to several genes, often antibiotic resistance or metabolic genes; often only the outermost inverted repeats are functional. [L14 s49]

• **Complex transposon**: has a resolvase gene (helps resolve integration) in addition to a transposase, and often an antibiotic resistance gene. [L14 s50]

• **Conjugative transposon**: contains genes for conjugation and can be transferred to another cell by conjugation: it jumps out of the chromosome and circularizes, makes conjugative pili, is transferred, and integrates into both

chromosomes (SXT, Fig 9.30). **Another type** carries only genes for integration and an oriT *(origin of transfer)*; it needs a conjugative plasmid to make conjugation happen, then is transferred and integrates (Fig 9.31). [L14 s51, L14 s52]

• Movement of an insertion sequence from one site to another = transposition of a transposon (not horizontal transfer, conjugation or transformation). [Key Fa24 Q11]

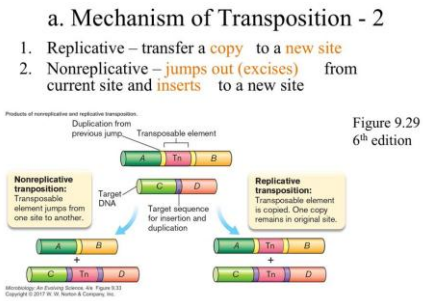
c. Mechanisms of insertion or jumping- transposition

• **Transposition steps**: cut at the target sequence, move into the target site, fill in; integrates at a specific short (2-6 bp) sequence (att). [L14 s45]

• **Replicative** transposition transfers a copy to a new site (one copy stays in the original site). **Nonreplicative** transposition jumps out (excises) from the current site and inserts at a new site. [L14 s46]

• **How a transposon moves** (his key, answer both): it copies itself and inserts the copy at a new site, or it jumps out of its current location and inserts into a new site. Transposition into a new site disrupts that site. [Key Fa25 Q13, Key Fa25 Q27]

• **Uses in research**: transposases and integrases (enzymes and sequences) are used for vector *(DNA carrier)* construction; transposons that integrate somewhat randomly are used to create random mutations (mutants). [L14 s53]



Nonreplicative vs replicative transposition (Fig 9.29) [L14 s46]

★ **On his past exams**: Fa24 Q11 (IS movement = transposition); Fa25 Q13 (copy and insert, or jump out and insert: A and B); Fa25 Q27 (transposition disrupts the new site). [Key Fa24 Q11, Key Fa25 Q13, Key Fa25 Q27]