Reading and Regulating the Genome

Gene Regulation · Splicing · Catalytic RNA · Translation & the Genetic Code

Sources: Lewin's Genes XII (2018) · Watson, Molecular Biology of the Gene 7e (2013) · Mobile DNA III (Craig, Chandler, Gellert & Lambowitz eds., ASM Press, 2015)

Genes, Genomas y Proteomas · Tema 04

From Sequence to Functional Protein

- **Regulation —** decide which genes are transcribed, when, and how much.
- **Splicing —** edit the primary transcript; one gene can yield many mRNAs.
- **Catalytic RNA —** RNA is not only the message, it is also the enzyme.
- **Translation —** decode the mRNA into polypeptide, fast and accurately enough.
- **The genetic code —** the mapping itself is degenerate, wobbly, and not quite universal.
- **The unifying idea:** every step is a decision point, and at almost every one the decision is made by an RNA structure rather than by a protein.

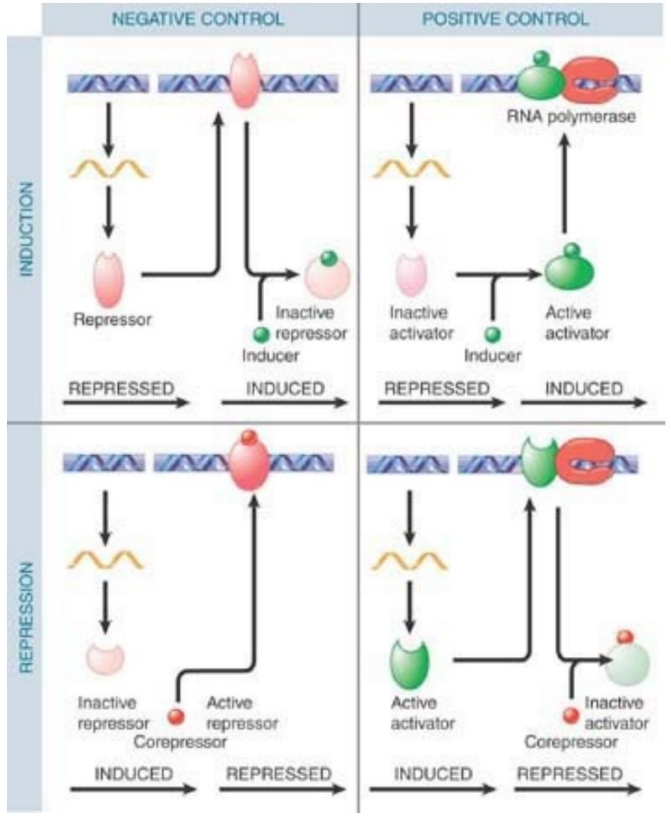
PART 1 · BACTERIAL REGULATION

Switching Genes On and Off

The Logic of Transcriptional Control

Lewin 24

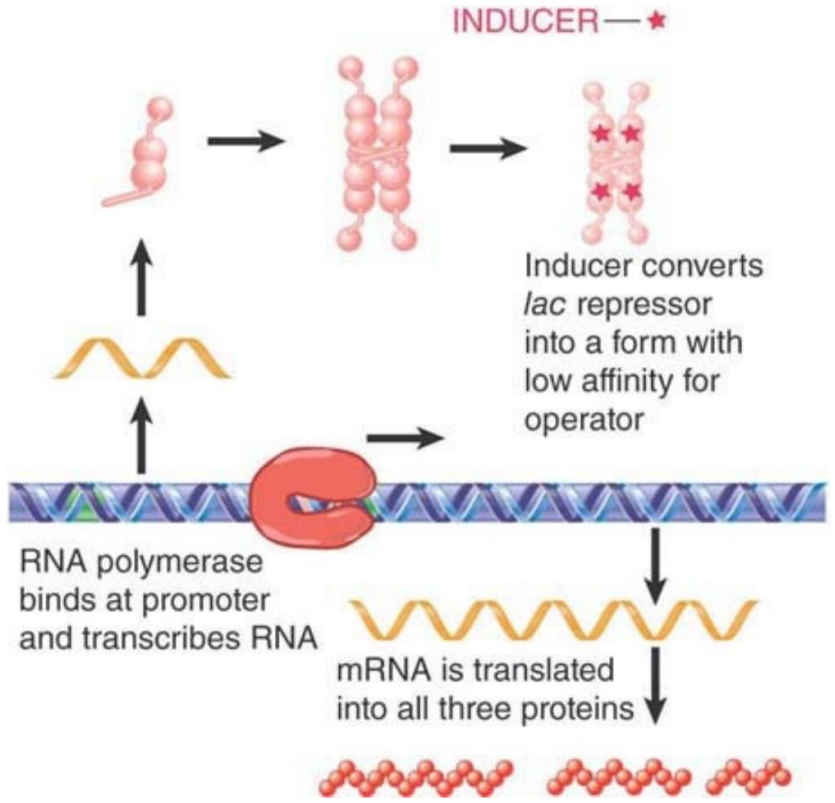
- **Regulator genes make trans-acting products** that recognise cis-acting sites — the same distinction you used for origins and promoters in Tema 3.
- **Negative control:** a repressor occupies an operator and blocks initiation; removing it turns the operon on.
- **Positive control:** an activator is required; without it the promoter is too weak to fire.
- **Inducible vs repressible:** a small-molecule effector either releases the block (inducer) or imposes it (corepressor).
- **Read any operon with two questions:** which protein binds the DNA, and what does the effector do to it? All four cells of the 2×2 are used in real genomes.



Lewin's Genes XII — Fig 24.4 — regulatory circuits are built from all four combinations of positive/negative control with inducible/repressible control.

lac: Negative Control and Induction

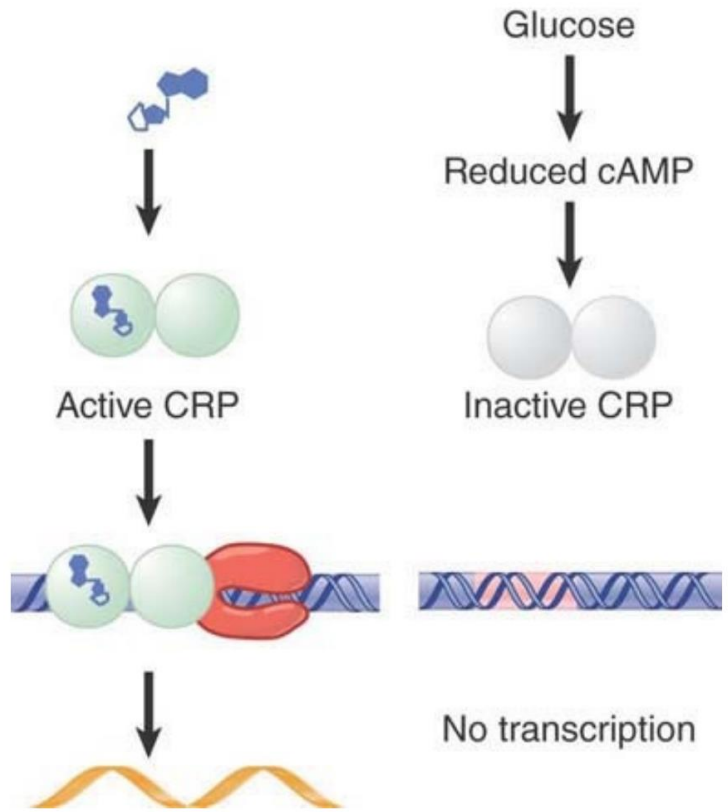
- **One transcription unit, three enzymes:** lacZ (β-galactosidase), lacY (permease), lacA (transacetylase) are induced coordinately, in that order along the message.
- **The LacI tetramer binds an operator that overlaps the start point** and sterically excludes polymerase; two auxiliary operators let one tetramer loop the DNA and raise repression ~100-fold.
- **Induction is allosteric:** allolactose binds the core domain, the DNA-binding headpieces lose affinity, and transcription resumes within seconds.
- **The genetics reads out the mechanism:** Oc is cis-dominant (a broken site), $lacI^-$ is trans-recessive (a diffusible product), $lacI^{-d}$ is dominant-negative (a poisoned subunit in the tetramer).
- **Repression is never absolute —** a basal trickle of permease and β-galactosidase must pre-exist, or the inducer could never get in and be made.



Lewin's Genes XII — Fig 24.9 — inducer converts the repressor to a low-affinity form, freeing the operator so RNA polymerase can initiate.

lac: Positive Control and Catabolite Repression

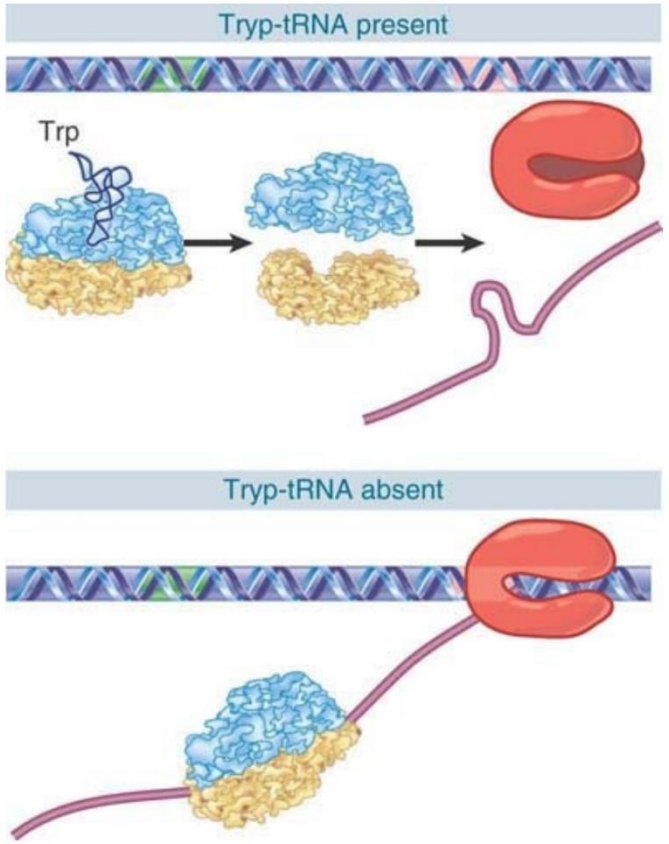
- **Relieving repression is necessary but not sufficient:** the lac promoter is intrinsically weak and needs help to initiate efficiently.
- **CRP (CAP) plus cAMP** binds upstream, bends the DNA by more than 90°, and recruits polymerase through a direct contact with the α subunit C-terminal domain.
- **Glucose lowers cAMP,** so CRP stays inactive — catabolite repression is positive control being withdrawn, not an extra repressor being added.
- **The operon is an AND gate:** lactose present AND glucose absent. This is combinatorial control in its minimal form.
- **One global signal, many local decisions —** the same CRP–cAMP serves dozens of catabolic operons, which is how the cell imposes a hierarchy of carbon sources.



Lewin's Genes XII — Fig 24.26 — by lowering cyclic AMP, glucose inhibits transcription of every operon that depends on CRP.

trp: Attenuation Couples Transcription to Translation

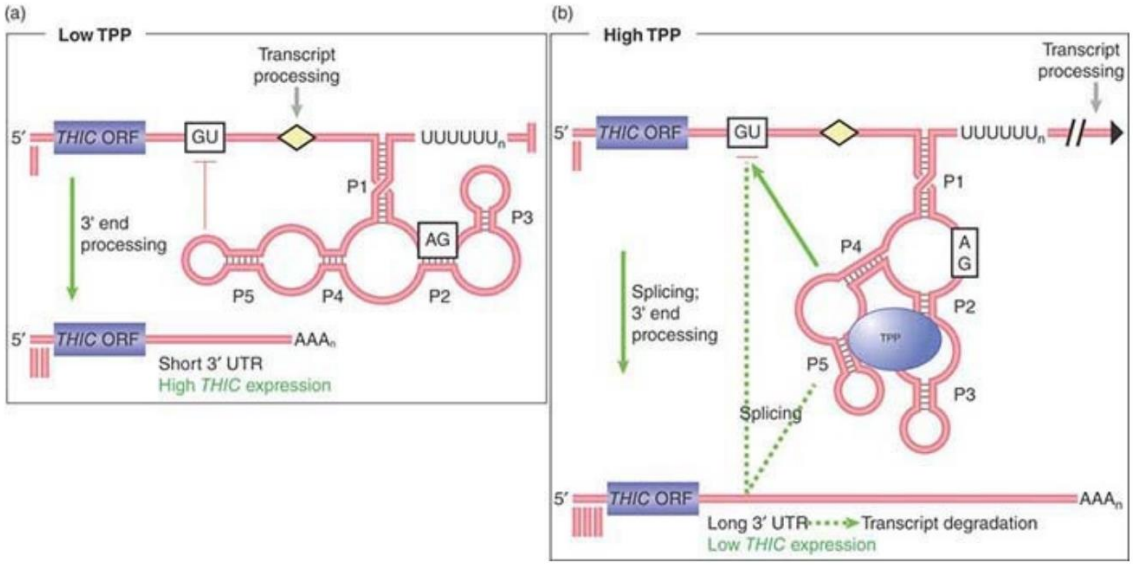
- **Two controls in series:** a repressor gives coarse control in response to free tryptophan, and the attenuator gives fine control in response to charged tRNA^Trp — together ~700-fold.
- **The sensor is a translated peptide:** the leader ORF contains two consecutive Trp codons, so ribosome progress reports on tRNA charging.
- **Trp plentiful** → the ribosome runs through region 1–2, the 3:4 terminator hairpin forms, and polymerase terminates before the structural genes.
- **Trp scarce** → the ribosome stalls on the Trp codons, region 2 pairs with 3 instead, the terminator cannot form, and polymerase reads on.
- **This only works because bacteria transcribe and translate the same molecule simultaneously** — a genuinely prokaryotic solution, impossible once a nuclear envelope intervenes.



Lewin's Genes XII — Fig 24.37 — with tryptophan tRNA available the ribosome translates the leader and releases, the terminator hairpin forms and polymerase stops; without it the ribosome stalls, the hairpin cannot form and polymerase continues.

Riboswitches: The mRNA Senses the Metabolite

- **No protein at all:** an aptamer in the 5′ UTR binds a small molecule directly, and ligand binding changes the RNA fold.
- **The fold change flips an expression platform —** forming or destroying a terminator hairpin, or occluding the ribosome-binding site.
- **Ligands are the metabolites of ancient pathways:** TPP, SAM, FMN, lysine, glycine, guanine, cobalamin — consistent with a very old regulatory strategy.
- **Eukaryotes use them too:** the TPP riboswitch is the one class conserved into plants and fungi, where it acts by changing splice-site choice rather than termination.
- **Some riboswitches are ribozymes:** glmS cleaves its own mRNA when glucosamine-6-phosphate binds — sensor and effector in one molecule, and the bridge into Part 4.



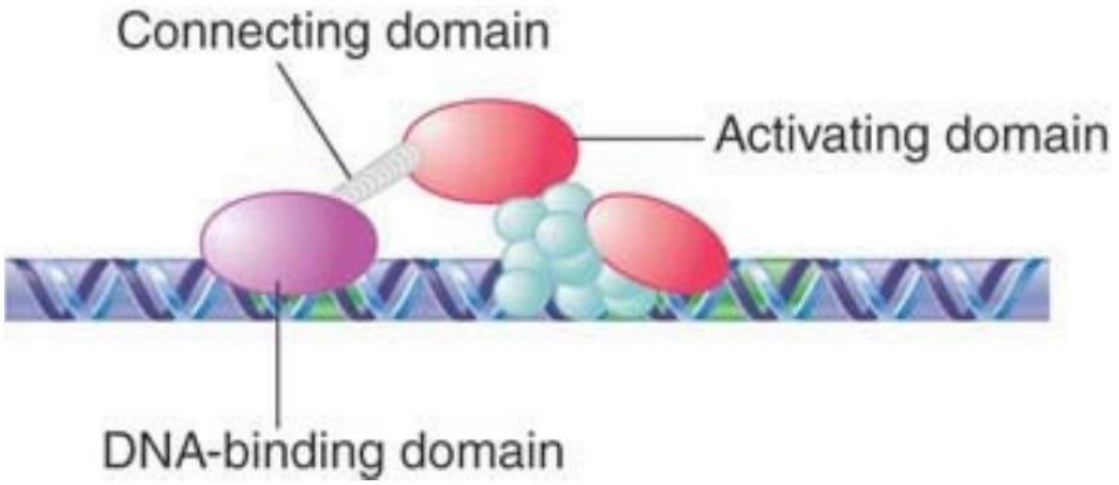
Lewin's Genes XII — Fig 29.2 — a TPP-binding riboswitch controls NMT1 expression by switching alternative splicing of the pre-mRNA.

PART 2 · EUKARYOTIC REGULATION

Combinatorial Control on a Chromatin Template

Activators Are Modular

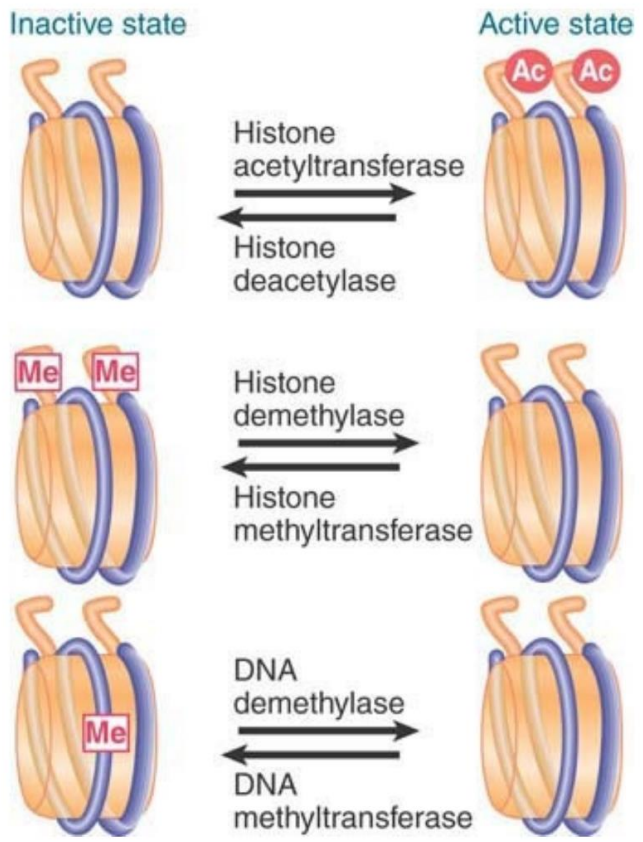
- **A typical activator has two separable parts:** a DNA-binding domain (zinc finger, homeodomain, bHLH, bZIP) and an activation domain, connected flexibly.
- **Domain swaps prove independence:** Gal4's DNA-binding domain fused to a foreign activation domain still activates — which is exactly why the two-hybrid assay works.
- **Flexibility explains enhancers:** because the connector does not care about distance or orientation, the DNA simply loops to bring the activation domain to the promoter.
- **Activation is usually indirect:** activation domains recruit coactivators — Mediator, TFIID/TAFs, histone acetyltransferases — rather than touching polymerase themselves.
- **Combinatorial arithmetic:** a few hundred factors, used in overlapping combinations at composite enhancers, specify tens of thousands of distinct expression patterns.



Lewin's Genes XII — Fig 26.6 — the DNA-binding and activating functions of a transcription factor are independent domains joined by a flexible connector.

Chromatin Is the Regulatory Layer

- **The eukaryotic default is OFF:** nucleosomes occlude promoters, so activation means clearing or repositioning them, not merely removing a repressor.
- **ATP-dependent remodellers** (SWI/SNF, RSC, ISWI, SWR1) slide, evict, or exchange nucleosomes wherever a sequence-specific activator points them.
- **Covalent marks set the state:** acetylation (HATs) opens; H3K9me and H3K27me close; DNA methylation at CpG locks it in.
- **Marks are read, not just written:** bromodomains bind acetyl-lysine, chromodomains bind methyl-lysine — so one modification recruits the next machine.
- **This is the layer that makes states heritable** through cell division without any change in sequence — the operational definition of epigenetics, and the subject of Tema 09.



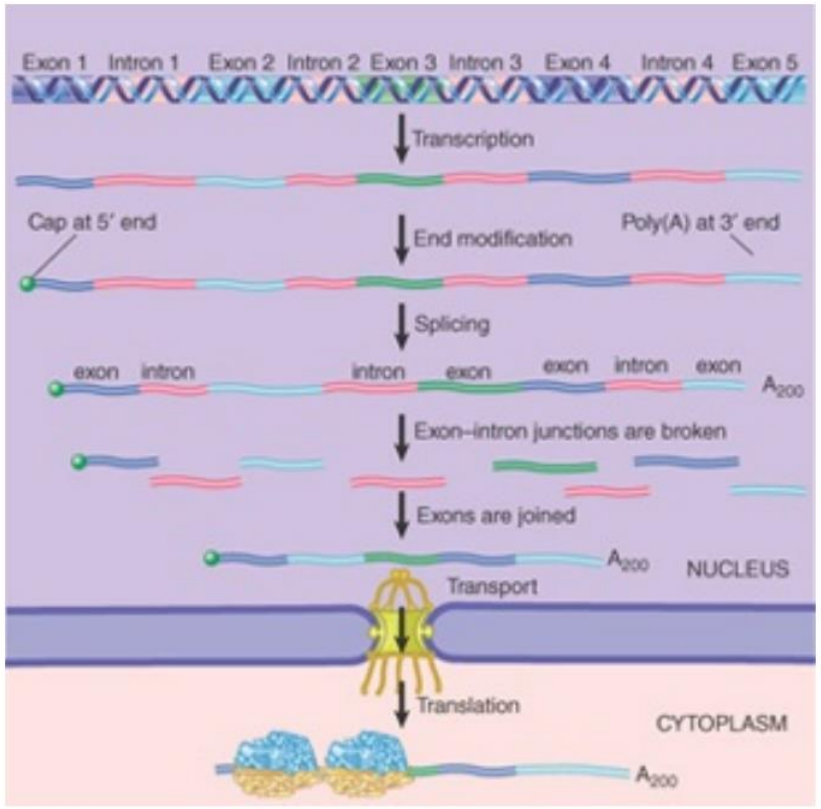
Lewin's Genes XII — Fig 26.27 — histone acetylation activates chromatin; DNA methylation and methylation at specific histone sites inactivate it.

PART 3 · RNA SPLICING

One Gene, Many Messages

Pre-mRNA Processing: Three Modifications

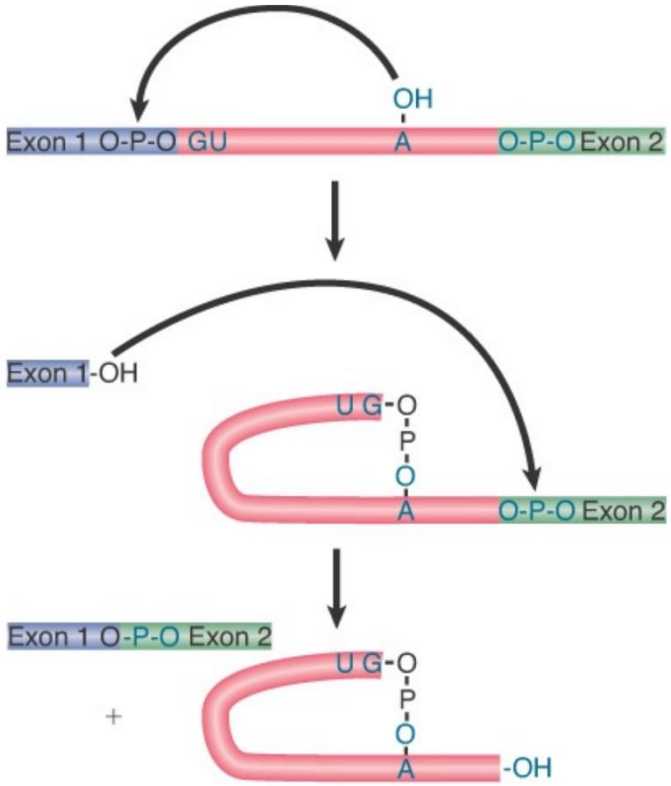
- **5′ cap:** m7G joined by an unusual 5′–5′ triphosphate bond, added within ~30 nt of initiation; it protects the end and is the eIF4E landing pad for translation.
- **Splicing:** exon–intron junctions are broken and exon ends joined, with no net change in the number of phosphodiester bonds.
- **3′ end:** made by cleavage ~10–30 nt after AAUAAA followed by poly(A) addition — the end is manufactured, not defined by where polymerase stops.
- **All three are co-transcriptional** and coupled to the phosphorylated Pol II CTD, which recruits the capping, splicing and cleavage machineries in turn.
- **Processing licenses export:** the exon junction complex deposited by splicing marks the mRNA as correctly made, and later drives nonsense-mediated decay if a stop codon lies upstream of it.



Lewin's Genes XII — Fig 19.1 — nuclear RNA is modified at both ends and spliced to remove introns before export to the cytoplasm.

The Chemistry: Two Transesterifications

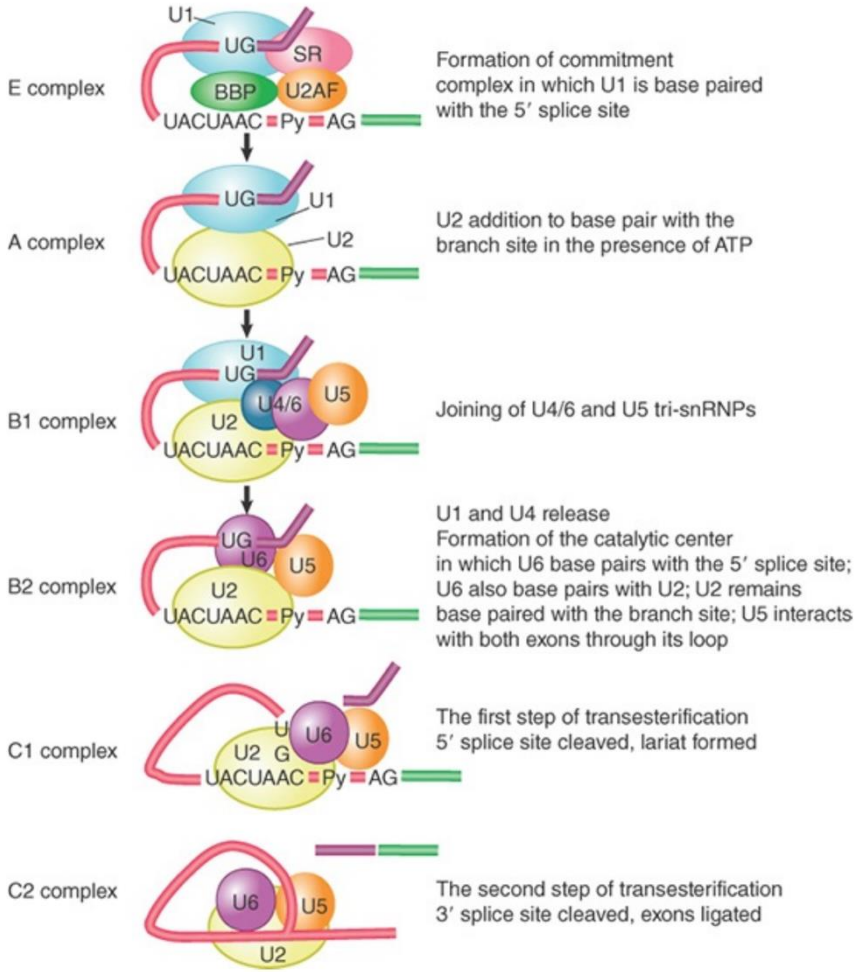
- **The ends are marked by the GU–AG rule,** with a branch-point A some 20–40 nt upstream of the 3′ site and a polypyrimidine tract between them.
- **Step 1:** the 2′-OH of the branch-point adenosine attacks the 5′ splice site, freeing exon 1 and creating a lariat through an unusual 2′–5′ bond.
- **Step 2:** the newly exposed 3′-OH of exon 1 attacks the 3′ splice site, ligating the exons and releasing the intron lariat.
- **Bond count is conserved,** so the chemistry itself costs no ATP — ATP is spent on spliceosome assembly and rearrangement, not on bond formation.
- **Hold on to this mechanism:** it is exactly what a group II intron does with no protein present at all, which is the argument of Part 4.



Lewin's Genes XII — Fig 19.6 — nuclear splicing proceeds by two transesterifications in which an –OH group attacks a phosphodiester bond, releasing the intron as a lariat.

The Spliceosome Cycle

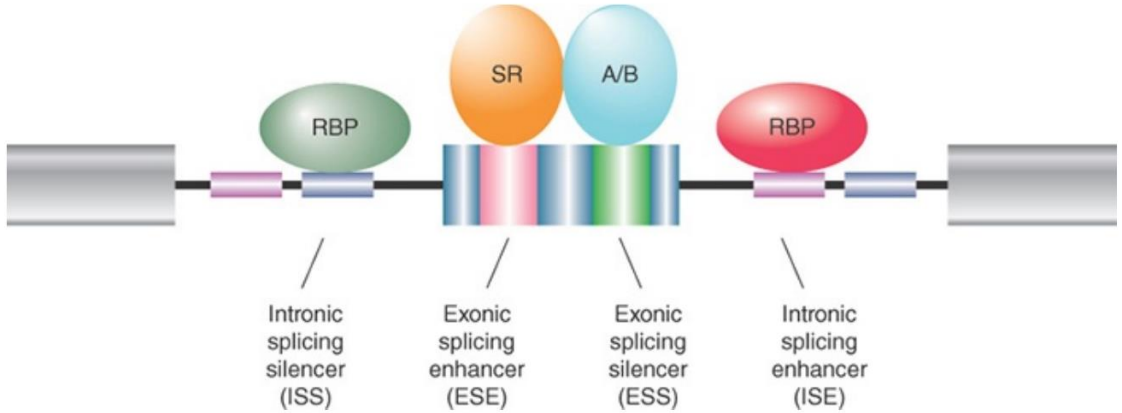
- **E (commitment) complex:** U1 base pairs with the 5′ splice site, SF1/BBP binds the branch point, U2AF binds the polypyrimidine tract and 3′ site. ATP-independent.
- **A complex:** U2 replaces SF1 and base pairs with the branch site, bulging the branch adenosine out so its 2′-OH is available. ATP-dependent.
- **B complex:** the U4/U6·U5 tri-snRNP joins, then U1 and U4 are released — the activation step that commits the substrate.
- **C complex:** U6 pairs with the 5′ splice site and with U2 to build the catalytic centre, while U5 tethers both exon ends through its loop; the two transesterifications follow.
- **What is catalytic here is RNA:** U6 and U2 snRNA position the two Mg^{2+} ions. A 12 MDa machine with ~170 proteins, and the chemistry is still done by the RNA.
- **Finding the sites:** vertebrates use exon definition (small exons in huge introns), yeast uses intron definition — which is why vertebrate splicing is so much more regulatable.



Lewin's Genes XII — Fig 19.12 — splicing proceeds through discrete complexes (E, A, B1, B2, C1, C2) as snRNPs recognise the consensus sequences and rearrange.

Alternative Splicing and Its Regulation

- **More than 90% of human multi-exon genes are alternatively spliced —** this, not gene number, is where proteome complexity comes from (~20,000 genes, far more proteins).
- **Cis elements tune site strength:** exonic and intronic splicing enhancers and silencers (ESE, ESS, ISE, ISS) sit near the alternative site.
- **Trans factors read them:** SR proteins generally promote inclusion, hnRNP A/B generally repress, and the local SR:hnRNP ratio is the tissue-specific signal.
- **Same machinery, several outcomes:** cassette exons, mutually exclusive exons, alternative 5′ or 3′ sites, and intron retention.
- **The classic developmental case:** the Drosophila Sxl → tra → dsx cascade sets sex entirely by splice-site choice, each step regulating the next.
- **Now a drug target:** risdiplam (FDA-approved 2020) is a small molecule that forces SMN2 exon 7 inclusion, treating spinal muscular atrophy by editing a splicing decision.



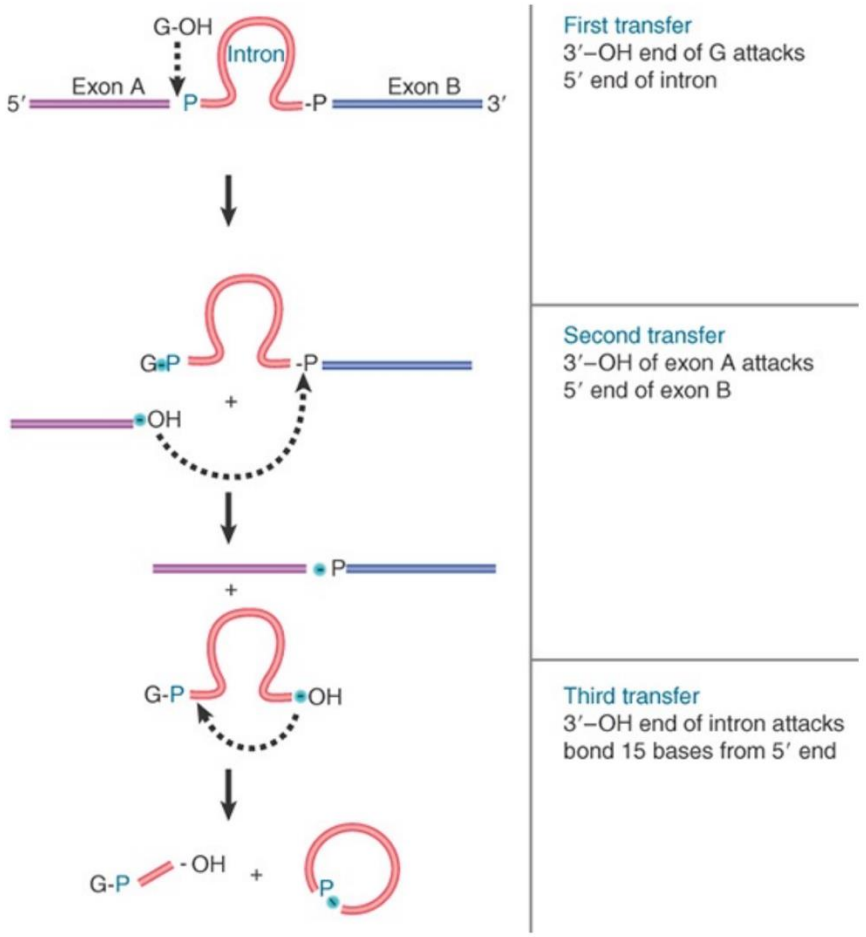
Lewin's Genes XII — Fig 19.24 — exonic and intronic enhancers and silencers modulate splice-site selection; SR proteins generally bind enhancers, hnRNP proteins generally bind silencers.

PART 4 · CATALYTIC RNA

When the Enzyme Is the RNA

Self-Splicing: The First Ribozyme

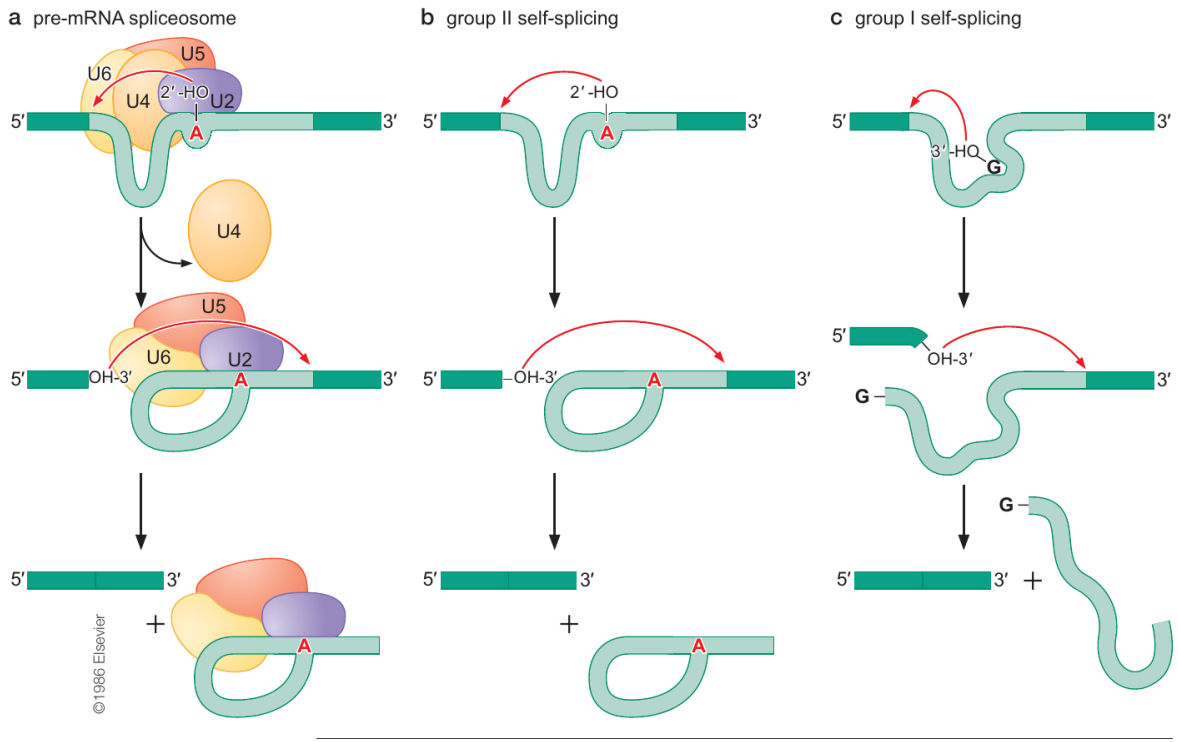
- **Cech, 1982:** the Tetrahymena rRNA intron excises itself in vitro with no protein present. Altman, in parallel: the catalytic subunit of RNase P is its RNA. Nobel Prize 1989.
- **Group I chemistry:** a free guanosine 3′-OH attacks the 5′ splice site; the released exon then attacks the 3′ site; the intron leaves linear and later circularises.
- **Same logic as the spliceosome, different nucleophile —** an exogenous G here, an internal branch-point A there. Both are transesterifications, both conserve bond number.
- **Catalysis comes from folding, not from side chains:** a conserved P4–P7 core, an internal guide sequence that base-pairs the 5′ exon into position, and specifically bound Mg^{2+}.
- **It is a real enzyme:** rate enhancements of 10^{10}–10^{11} over the uncatalysed reaction, with genuine substrate binding — chemically poorer than protein, but unambiguously catalytic.



Lewin's Genes XII — Fig 21.2 — group I self-splicing proceeds by transesterifications in which bonds are exchanged directly; a free guanosine initiates the first transfer.

One Chemistry, Three Machines

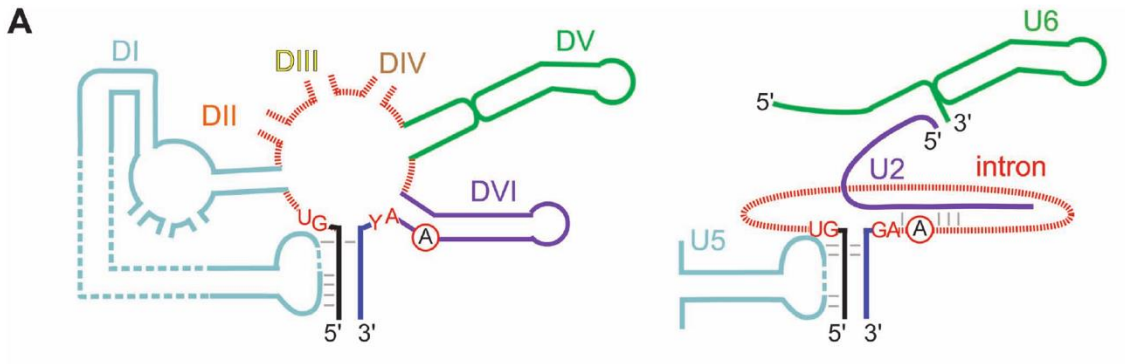
- **Group I:** an exogenous guanosine attacks; the intron departs linear.
- **Group II:** an internal branch-point adenosine attacks; the intron departs as a lariat; entirely self-catalysed.
- **Spliceosome:** an internal branch-point adenosine attacks; the intron departs as a lariat; catalysis is supplied in trans by snRNAs.
- **So the spliceosome and a group II intron differ mainly in delivery:** one carries its catalytic RNA as separate, reusable pieces, the other has it built into the intron.
- **The wider ribozyme set:** RNase P, hammerhead and hairpin ribozymes, the glmS riboswitch, the ribosome's peptidyl transferase — plus the whole of RNA interference.



Watson, Molecular Biology of the Gene 7e — Fig 14-8 — the spliceosome-mediated reaction compared with group II and group I self-splicing; the chemistry of group II is essentially that of the spliceosome.

Group II Introns Became the Spliceosome

- **A group II intron is two things at once:** a ribozyme (six domains, DI–DVI) and a retroelement, carrying an intron-encoded protein that is reverse transcriptase + maturase + endonuclease.
- **The correspondence is domain-for-domain:** DV ↔ the U6 internal stem-loop, DVI ↔ U2 with the branch point, DI's exon-binding site ↔ U5 — same catalytic triad, same two-Mg^{2+} active site.
- **They are mobile:** retrohoming reverse-splices the intron RNA directly into DNA, then target-primed reverse transcription copies it — the same TPRT mechanism you saw for LINEs in Tema 3.
- **The evolutionary scenario:** bacterial group II introns arriving with the mitochondrial endosymbiont, proliferating in the proto-eukaryotic genome, then fragmenting into the snRNAs.
- **Even the protein kept its ancestry:** Prp8, the spliceosome's largest and most conserved protein, retains a reverse-transcriptase-like fold derived from an intron-encoded RT.
- **One element, four legacies:** spliceosomal introns, the snRNAs, the spliceosome itself, and the non-LTR retrotransposon/telomerase RT lineage.



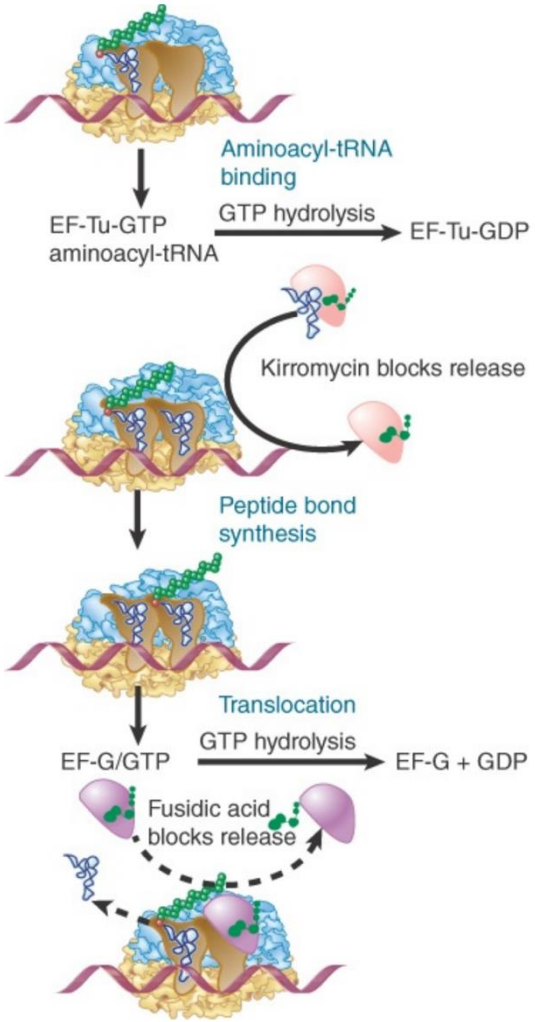
Mobile DNA III (ASM Press, 2015) — Fig 5A (Lambowitz & Belfort) — the group II intron (left) and the spliceosome (right) drawn with corresponding elements in matching colours: DV/U6, DVI/U2, DI-exon-binding/U5.

PART 5 · TRANSLATION & THE GENETIC CODE

Decoding the Message

The Elongation Cycle

- **Three tRNA sites:** A for the incoming aminoacyl-tRNA, P for peptidyl-tRNA, E for the exiting deacylated tRNA; mRNA and tRNAs move together, three nucleotides at a time.
- **EF-Tu·GTP delivers aminoacyl-tRNA** and hydrolyses GTP only after correct codon–anticodon pairing is sensed — the hydrolysis step is the fidelity checkpoint, not the binding step.
- **Peptidyl transfer** moves the growing chain from the P-site tRNA onto the A-site amino acid; the tRNAs then occupy hybrid A/P and P/E states.
- **EF-G·GTP drives translocation,** and its structure mimics the shape of the EF-Tu–tRNA complex — molecular mimicry that lets both use the same site.
- **Alternation is enforced by that shared factor-binding site,** which is exactly why kirromycin freezes EF-Tu release and fusidic acid freezes EF-G release.
- **Speed and accuracy:** ~15–20 amino acids per second in bacteria at ~10^{-4} errors per codon — a million-fold worse than replication, but proteins are disposable and DNA is not.

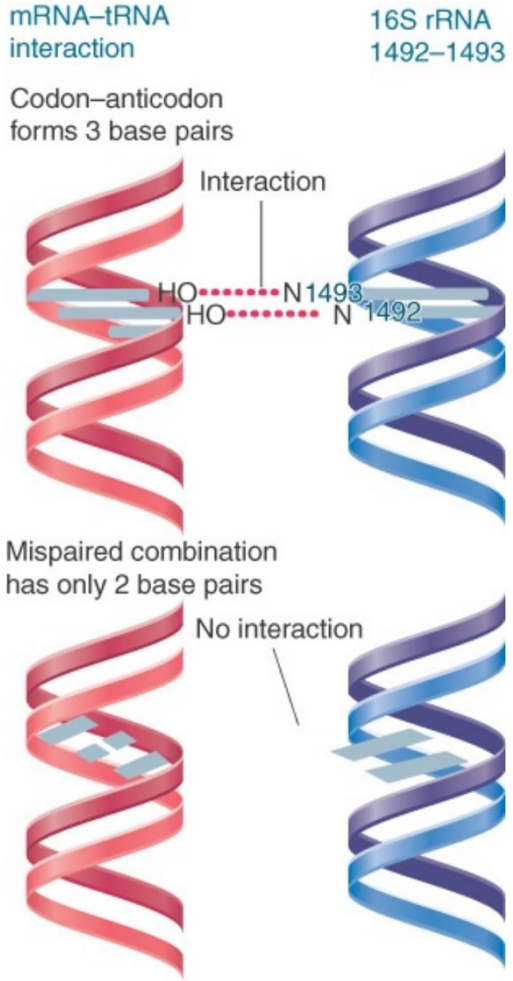


Lewin's Genes XII — Fig 22.30 — EF-Tu and EF-G bind alternately as the ribosome accepts aminoacyl-tRNA, forms the peptide bond, and translocates.

The Ribosome Is a Ribozyme

Lewin 22

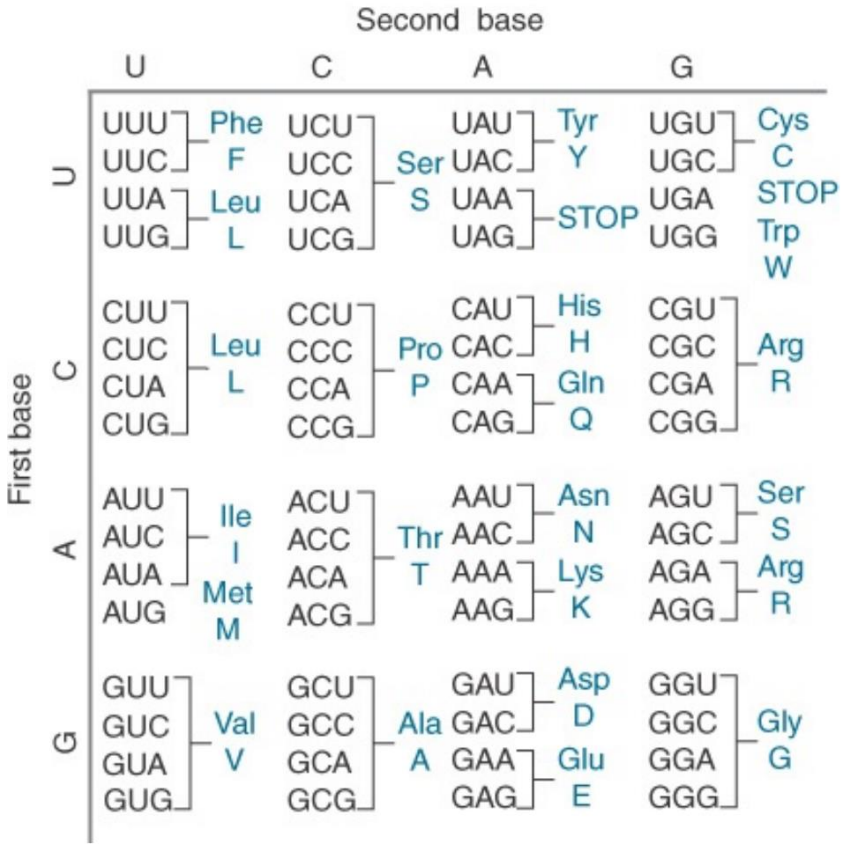
- **The peptidyl transferase centre is pure RNA:** crystal structures show no protein within ~18 Å of the reaction — 23S rRNA does the chemistry.
- **Catalysis is largely entropic:** the rRNA positions the substrates and the P-site tRNA's own A76 2′-OH participates in shuttling the proton.
- **Decoding is rRNA's job too:** 16S bases A1492 and A1493 flip out of their helix and read the minor-groove geometry of the codon–anticodon duplex.
- **Geometry, not stability, is the test:** only Watson–Crick shape at positions 1 and 2 triggers the domain closure that licenses EF-Tu to hydrolyse GTP.
- **Fidelity is a two-stage filter:** initial selection before hydrolysis, then proofreading after it, so a near-cognate tRNA gets two chances to be rejected.
- **Both facts point the same way as Part 4:** the catalytic core of the translation machine is RNA, and must predate the proteins now wrapped around it.



Lewin's Genes XII — Fig 22.47 — correct codon–anticodon pairing supports interaction with adenines 1492 and 1493 of 16S rRNA; a mispaired tRNA–mRNA duplex cannot.

The Genetic Code

- **64 triplets:** 61 sense codons plus 3 stops. The code is degenerate — most amino acids have several codons — but unambiguous: no codon ever means two things.
- **Degeneracy is structured, not random:** the third base matters least (third-base degeneracy), and chemically similar amino acids tend to have similar codons.
- **So the code buffers mutation:** most third-position changes are synonymous, and many first-position changes are conservative substitutions.
- **This is what makes Ka/Ks work** (Tema 2): synonymous sites approximate neutral evolution, so the ratio of nonsynonymous to synonymous change measures selection.
- **Codon usage is biased and species-specific,** tracking tRNA abundance — which is why heterologous expression so often needs codon optimisation.

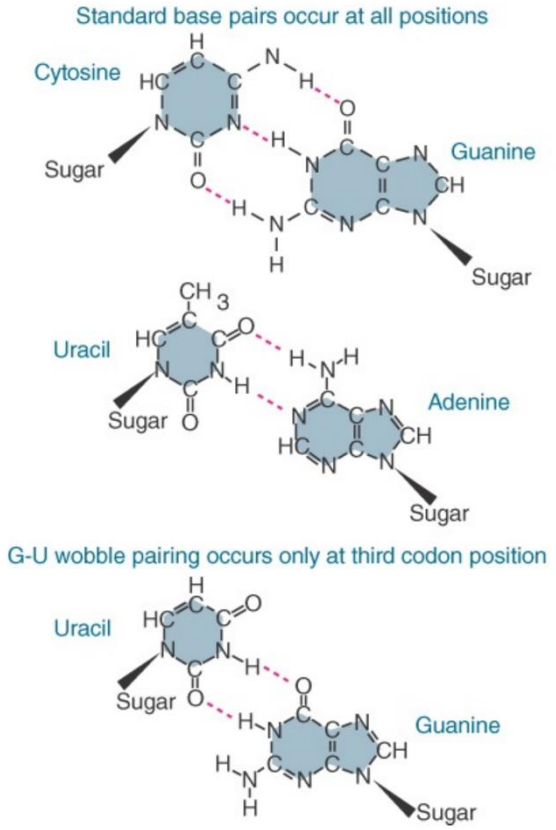


Lewin's Genes XII — Fig 23.1 — all 64 triplets have meaning: 61 specify amino acids and 3 cause termination.

Wobble: Fewer tRNAs Than Codons

Lewin 23

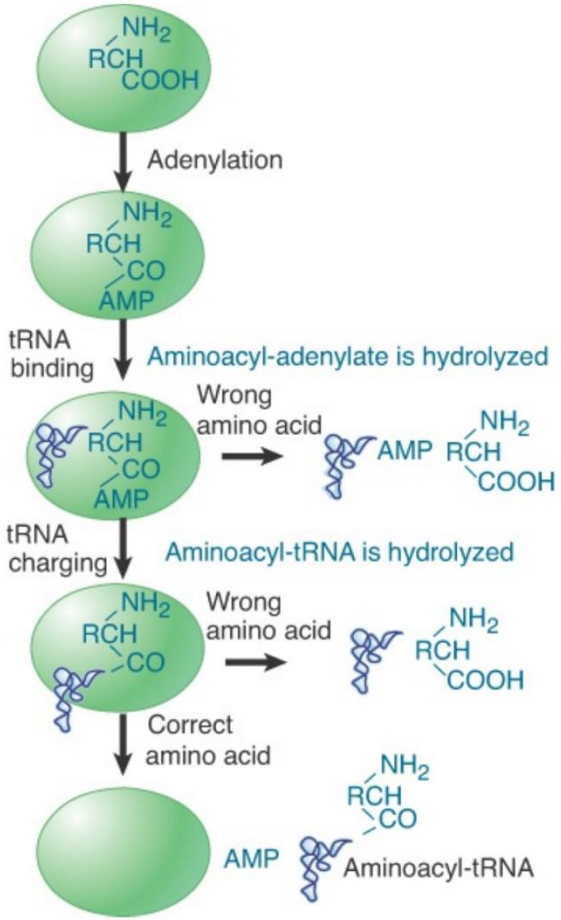
- **The problem:** 61 sense codons but far fewer tRNA species — around 40 in most cells, and only 22 in human mitochondria.
- **Crick's wobble hypothesis:** pairing at codon position 3 / anticodon position 1 is geometrically loose, so G·U pairs are tolerated there and nowhere else.
- **Modified bases extend the range:** inosine (deaminated A) at the wobble position reads U, C or A — the widest reach of any anticodon base.
- **And modifications can also restrict it:** 2-thiouridine can only form one hydrogen bond with G, so it pairs with A alone — sharpening rather than broadening specificity.
- **The consequence:** unique codon assignments are only possible when the third base is G or U, which constrains how the code could have expanded.



Lewin's Genes XII — Fig 23.4 — wobble allows G·U pairs between the third base of the codon and the first base of the anticodon; standard pairs occur at all positions.

Charging Fidelity: The Synthetases Decide

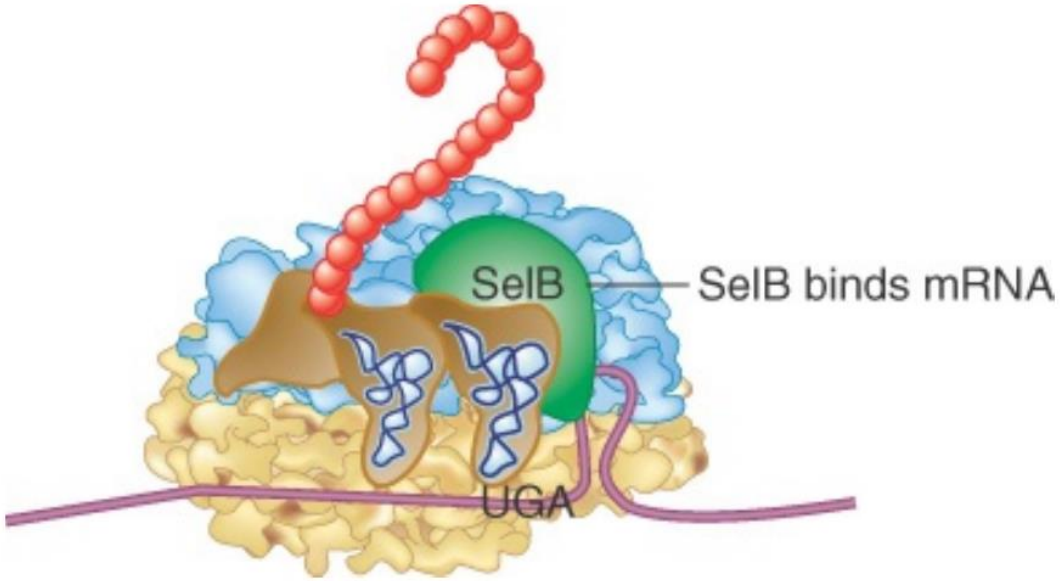
- **The real specificity step is charging, not decoding:** the ribosome checks the anticodon, so whatever amino acid the synthetase attached is what gets incorporated.
- **Two chemical steps:** the amino acid is activated as an aminoacyl-adenylate using ATP, then transferred to the 2′- or 3′-OH of the tRNA's terminal A.
- **Two unrelated classes:** class I and class II synthetases have different folds and approach opposite faces of the tRNA — the code was, in effect, implemented twice.
- **Editing is built in:** a second active site hydrolyses wrong products, either before transfer or after — the double-sieve. IleRS excludes valine to about 1 in 3,000.
- **Classic proof that the ribosome does not check:** Cys-tRNA^Cys chemically reduced to Ala-tRNA^Cys inserts alanine at every cysteine codon. The tRNA is the adaptor; the synthetase writes the code.



Lewin's Genes XII — Fig 23.16 — proofreading by aminoacyl-tRNA synthetases: the noncognate aminoacyl-adenylate can be hydrolysed before transfer, or the mischarged tRNA after it.

Recoding: The Code Is Not Frozen

- **Selenocysteine, the 21st amino acid:** a UGA stop is read as Sec when a downstream SECIS hairpin recruits the dedicated elongation factor SelB — context changes the meaning of a codon.
- **Pyrrolysine, the 22nd:** UAG read as Pyl in some archaea and bacteria, using its own synthetase and tRNA.
- **Programmed frameshifting:** a slippery sequence plus a downstream pseudoknot makes the ribosome shift −1 or +1 at a defined efficiency — used by retroviruses to set the Gag:Gag-Pol ratio.
- **Suppressor tRNAs:** an anticodon mutation lets a tRNA read a stop codon, rescuing a nonsense mutation — the phenomenon that first proved the code is read by tRNA.
- **The code is not quite universal:** mitochondrial genomes reassign codons repeatedly, and some ciliates read UAA/UAG as glutamine. Small, isolated genomes can afford to change.
- **Now engineered on purpose:** reassigning codons to incorporate non-canonical amino acids is standard synthetic-biology practice.



Lewin's Genes XII — Fig 23.11 — SelB is a dedicated elongation factor that delivers tRNA^Sec to a UGA codon followed by a stem-loop in the mRNA.

Where the Field Has Moved

Seminars · 7 ago

- **Splicing is not only co-transcriptional.** Up to ~40% of mammalian introns are removed after transcription terminates, making timing itself a regulatory variable (Choquet et al., Nat Rev Genet 2025 — seminar paper 10).
- **Where a gene sits in the nucleus changes how well it is spliced.** Proximity to nuclear speckles, which concentrate splicing factors, predicts splicing efficiency — 3D genome organisation feeding into RNA processing (Bhat et al., Nature 2024 — paper 11, lectura complementaria).
- **The epitranscriptome is a regulatory layer in its own right.** m⁶A, m⁵C and pseudouridine affect every stage of an mRNA's life, written and erased by dedicated enzymes and read by YTH-domain proteins (Gilbert et al., Annu Rev Biochem 2023 — seminar paper 12).
- **And it is druggable:** STM2457, a selective METTL3 inhibitor, reduces m⁶A on leukaemogenic mRNAs and impairs AML growth in vivo (Yankova et al., Nature 2021).
- **The methodological warning for your own work:** RNA modifications and isoform-level regulation both distort standard RNA-seq quantification. Gene-level counts can be right while the biology is happening one level down.

Five Decisions, One Recurring Answer

- **Which genes are read** — repressors, activators, and increasingly the RNA itself (attenuators, riboswitches).
- **Which parts of the transcript survive** — the spliceosome, whose catalytic core is snRNA, tuned by SR/hnRNP ratios.
- **Who does the chemistry** — RNA, in group I and II introns, RNase P, the spliceosome and the peptidyl transferase centre.
- **How faithfully it is decoded** — synthetase double-sieve editing, then rRNA-based geometric proofreading at the A site.
- **What the codons mean** — a degenerate, error-buffering, wobble-read code that is neither frozen nor quite universal.
- **The thread:** every layer that looks like it should need a protein turns out to have an RNA solution underneath it — which is the strongest argument that RNA came first.