The biology of big data

a primer on the biological foundations of functional genomics

Wendy Valencia-Montoya (TF 2020), adapted from Irina Shlosman (TF 2019), Kate Shulgina (TF 2018), and Laura Bagamery (TF 2016)

Preamble

> We are drowning in a sea of data and starving for knowledge.

Sydney Brenner, Nobel lecture, 2002

> It's a biology that I like to call low-input, high-throughput, no-output biology.

Sydney Brenner, "The next 100 years in biology", 2006

Over the past decade, technological advances have shattered previous limitations on the scope and scale on which biological data can be collected. A single experiment can now generate gigabytes or even terabytes of data, capturing the totality of an organism's genetic information; its entire brain, in three dimensions, with single-neuron resolution; or the individual activity levels of each of tens of thousands of its genes.

The increasing focus on generating high-volume data sets has also been criticized by prominent scientists within the field. Detractors argue that such mass data production prioritizes data quantity over quality and promotes less intellectually rigorous techniques of merely scouring large source material for interesting correlations, rather than thoughtfully formulating concrete hypotheses and designing controlled experiments to test them directly.

This is simply not true. The analysis of high-volume data is but another scientific investigation that requires the integration of careful hypothesis formulation, experimental design, and validation (historically well-emphasized in biology) with the computational expertise to extract information from massive data sets (still a relatively young skill within the field, but one that is now as acutely necessary for a young scientist).

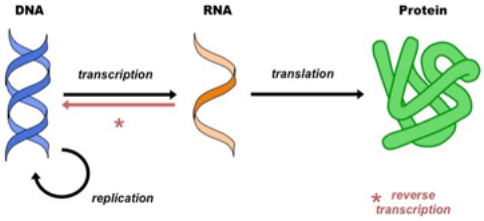
This course will explore computational methods for analyzing high-throughput data in great detail. This primer is intended to provide background exploring where, conceptually and technically, such data come from and how these factors dictate the biological questions that they are capable of answering.

The central dogma

All living organisms descent from a common ancestor and use the same molecules to store genetic information. However, every individual possesses a unique set of traits that distinguish it from members of other species and, in subtler ways, from other members of its own species. These traits, which are passed down from the organism's ancestors, are complex manifestations of units of heritable information known as **genes**.

How does an organism store, transmit, and interpret this genetic information that specifies its identity?

The **central dogma** of molecular biology posits that cells possess chemical codes detailing the assembly instructions for macromolecular machinery that promotes reactions underlying the fundamental physiological processes and states characteristic of the cell. Stored genetic material, in the form of DNA, is decoded to produce RNA intermediates, which are in turn decoded to build instruments of biological activity, proteins:

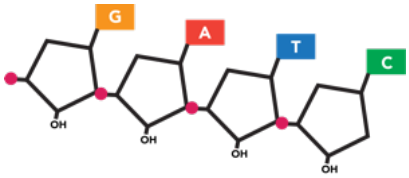


RNAs and proteins that are created in this way are referred to as **gene products**, and their synthesis is known as **gene expression**.

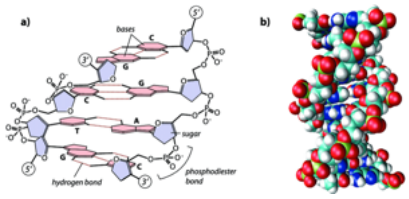
Genetic information at the molecular scale

On a molecular level, a gene is a sequence of **deoxyribonucleic acid**, or **DNA**. A molecule of DNA consists of two chains coiled around each other to form a double helix.

Each strand is built of **nucleotide** subunits, which are made of three parts: a five-carbon sugar (deoxyribose), a phosphate group, and a nitrogenous **base**. The four bases found in DNA nucleotides are known as **adenine (A)**, **guanine (G)**, **thymine (T)**, and **cytosine (C)**, and the sequence of these bases in DNA encodes the genetic information.



Schematic representation of a string of DNA nucleotides--linkages between the phosphate groups (pink) and five-carbon sugars (pentagons) of successive nucleotides comprise the strand backbone while the bases specify information content (here, GATC).



a. Molecular structure of nucleotides. b. 3D representation of the double-helical structure of DNA.

The double-stranded nature of DNA adds redundancy, as particular **base pairs (bp)** co-occur at the same position across sister strands and share weak bonds. These specifically pair A with T and G with C. The sequences of the two strands are thus **complementary**, with each reflecting the content of the other in a set of corresponding bases. This repetitive nature allows the fidelity of the sequence to be preserved in the event that one strand is damaged.

The information contained on each strand is also directional, as linkages within a DNA

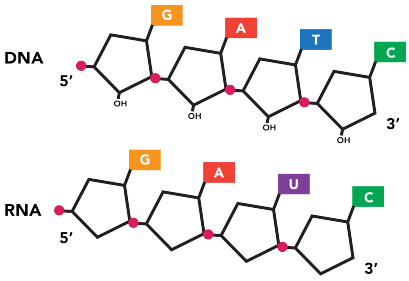
molecule imbue the structure with chemical polarity. Within the chain, a nucleotide is bound to the nucleotide that precedes it through its own phosphate group and to the nucleotide that follows it at its deoxyribose group. Under a particular convention for numbering the carbons within the five-carbon sugar, these linkages occur at the fifth carbon (which bears the phosphate group) and third carbon, respectively. Additional nucleotides can only be added to the end of the strand at which a nucleotide possesses an accessible third carbon. All reactions relying on the synthesis of new nucleic acid polymers--a category which includes both DNA replication and DNA sequence decoding (*to be explained*)-- must thus proceed along the polymer in a single direction known as **5' to 3'** *("five-prime to three-prime")*. This is the sole direction in which a strand "runs" and in which its genetic information can be read.

The two strands within a double helix are oriented such that the 5' end of one strand is paired with the 3' end of its sister. As the two strands run in opposite, or **antiparallel** directions, each sequence is, relative to its sister, a series of opposite bases running backwards, or the **reverse complement**.

Double-stranded DNA. The 5' and 3' ends of each strand are marked and occur at opposite sides of the molecule for each chain to produce antiparallel alignment. Complementary base pairing occurs between A-T and C-G couples across strands.

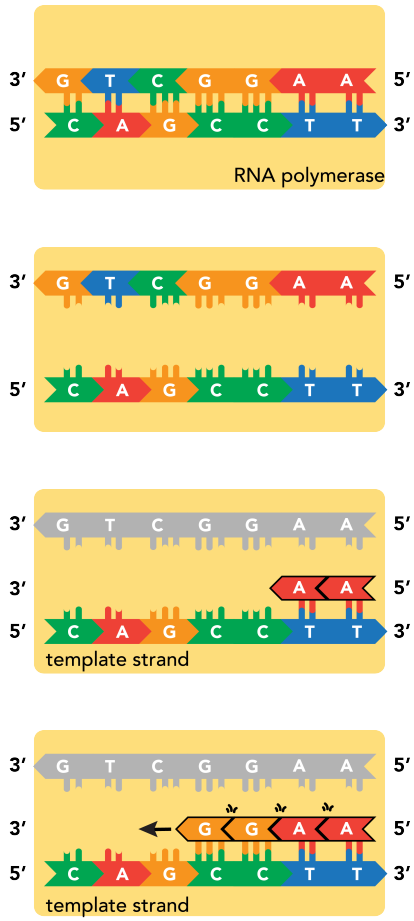
Decoding DNA: transcription

Genetic information is extracted from DNA by **transcription**, in which the double helix unfurls and one strand serves as the template for the production of a molecule of **ribonucleic acid**, or **RNA** from the gene source. The structure of RNA is similar to that of DNA, but its phosphate-sugar backbone incorporates a slightly different sugar (ribose) and the base **uracil (U)** is substituted for thymine. Like thymine, uracil associates with adenine bases, although RNA is typically single-stranded and exhibits base pairing primarily in complex structures in which a single RNA strand partially folds against itself.



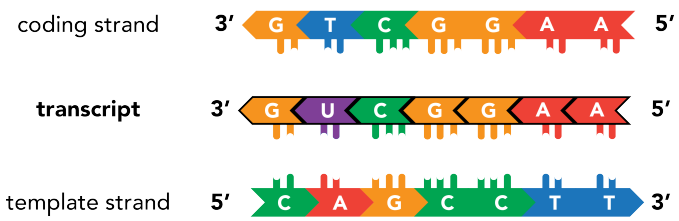
Equivalent nucleotide sequences expressed in DNA (top) and RNA (bottom). RNA contains the base uracil in lieu of thymine, and the sugars within its backbone lack a hydroxyl group found in deoxyribose. The 5' and 3' ends of each sequence are labeled.

Transcription is capable of generating RNA products, known as **transcripts**, using either strand within a double helix as a template. The protein that catalyzes transcription, **RNA polymerase** unwinds a small portion of the double helix, which allows free RNA nucleotides to base pair to the template. The polymerase then links nucleotides into a continuous polymer by spurring the formation of **phosphodiester bonds** between the ribose sugars and phosphate groups of successive nucleotides. RNA, like DNA, can be synthesized solely from its 5' end along its 3' end, and this is the direction in which RNA polymerase moves down a nascent transcript.



Mechanism of transcription. RNA polymerase (depicted as a yellowish blob surrounding the DNA) binds a section of DNA and causes local dissociation of the two DNA strands. The strand from which RNA will be produced is the template. RNA nucleotides (depicted with black outlines) form complementary base pairs with the template. RNA polymerase moves processively along DNA to link RNA nucleotides and synthesize a transcript 5' to 3'.

As the RNA lies antiparallel to its DNA source material, the sequence of a transcript will always be the reverse complement of its template. The template's sister strand, also the reverse complement of the template, features the sequence that will actually match the transcript (with U-T substitutions). This strand is accordingly known as the **coding strand**. Which strand is coding and which is template can vary among the many genes present within a stretch of DNA.

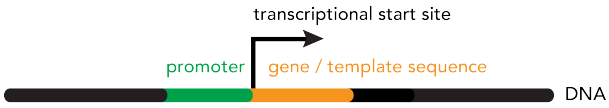


An RNA transcript (center) is the reverse complement of its template DNA (bottom) but contains an analogous sequence to the other, "coding" strand (top).

Genes--sequences that encode transcripts in this way--constitute only a fraction of an organism's total DNA content. How, then, is a particular DNA sequence recognized as a gene, and how is it targeted for transcription?

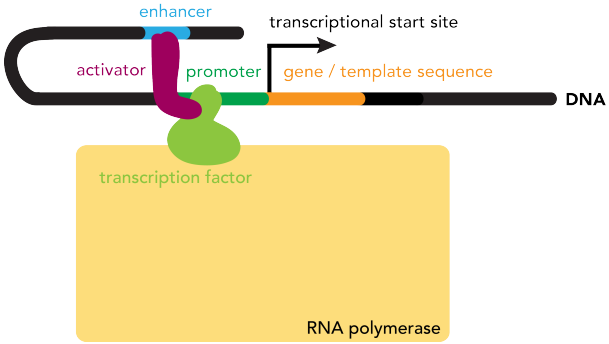
A sequence that serves as template for an RNA molecule is prefaced by a **promoter**

region, a stretch of DNA that defines the gene's start site. The promoter serves at the binding site at which RNA polymerase will be loaded onto the sequence prior to transcription.



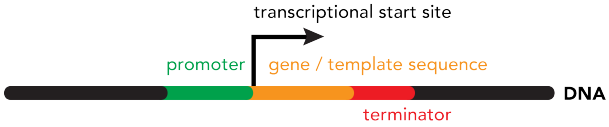
*A promoter region precedes, or appears **upstream**, of a gene. The site of transcriptional initiation is traditionally indicated with an arrow.*

Many factors influence the frequency with which RNA polymerase will bind to such a promoter region. These include the strength of the promoter--its intrinsic binding affinity for the polymerase--as well as the presence of absence of various accessory proteins, known as **transcription factors**, that influence the interaction between RNA polymerase and template DNA. Transcription factors may assist in either the recruitment or repulsion of RNA polymerase from a DNA sequence. **Activators** positively regulate RNA polymerase binding and thus transcription rates, while **repressors** negatively regulate these processes. Protein factors that alter the rate of transcription may also do so by binding DNA at regulatory regions other than the promoter, typically beyond the 5' end, or **upstream** of the gene but also past the 3' end (**downstream**), which can be relatively close to or far from the genes that they target. A complex interplay of all of these factors regulates, on the level of transcription, the extent to which a particular gene is expressed.



A more complicated view of the interactions that regulate transcription. Transcription factors can bind to the promoter region and alter the binding propensity of RNA polymerase. Additionally, other protein factors may bind at additional sites, near or distant, to affect transcription. Here, an upstream regulatory region known as an enhancer is bound by an activator. Though the enhancer region is far upstream, looping of the DNA back over itself allows the activator bound at this distant site to interact with other gene regulatory machinery to increase RNA polymerase binding and thereby the transcription rate.

In contrast, the cues that mark the end of a gene are more straightforward. Directly downstream of the transcribed region, a **terminator** sequence triggers the release of RNA polymerase from DNA.

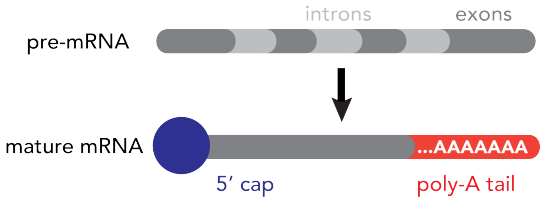


A terminator region directly downstream of the gene. Movement of RNA polymerase across the terminator sequence will stimulate the undbinding of the enzyme from the DNA sequence and thereby terminate transcription.

Decoding RNA: translation

Only a subset of RNA molecules provide template from which proteins are produced. These are **messenger RNAs** (**mRNAs**). Such RNAs undergo a series of post-transcriptional modifications that mark them as information sources for future proteins.

In **eukaryotes** (cells that contain a defined **nucleus** in which genetic information is stored, among other membrane-delineated compartments that segregate biochemical reactions), a **pre-mRNA**, or **primary transcript**, the native RNA form that is the direct product of transcription, is converted into a **mature mRNA** that is ready to be decoded into protein through three principal processing steps: **5' capping**, **polyadenylation**, and **RNA splicing**.

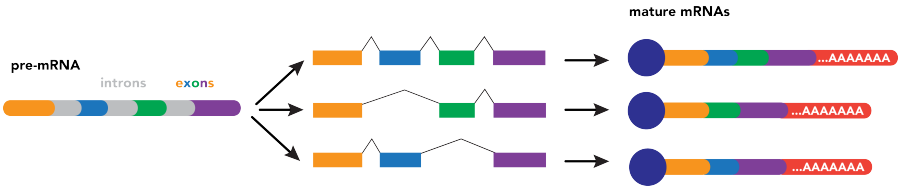


An overview of RNA processing. The 5' end of the transcript is outfitted with a cap, the 3' end is extended with an A-rich tail, and some internal sequences are stripped from the pre-mRNA while the remainder are spliced into a single, continous, mature transcript.

First, even as the RNA is still being transcribed, the 5' end of the nascent transcript is fitted with a **5' cap**, a single modified guanine nucleotide. This cap both identifies the transcript as an mRNA (for purposes of nuclear export and protein production) and is protective against degradation.

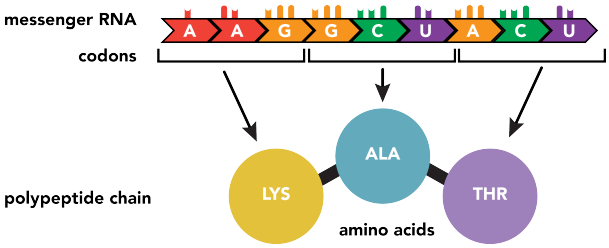
After transcription is complete, the 3' end of the transcript is extended with the addition of roughly 200 nucleotides, almost all of which are adenine bases. This **poly-A tail** is not specified with the DNA content of the gene and is synthesized without a template. Like the 5' cap, this modification, **polyadenylation**, marks the RNA as protein-coding and increases the stability of the molecule.

Not all regions within a coding transcript encode for protein sequence. In the third stage of RNA processing, noncoding **intron** regions are excised from mRNA precursors, leaving only coding regions, or **exons**, comprising mature mRNA. This editing step, **RNA splicing**, occurs by a complicated mechanism in which a large set of proteins and one class of RNAs that are not protein-coding in themselves, **small nuclear RNAs** (**snRNAs**) constitute a **spliceosome** complex that catalyzes the formation of loops within intron sequences followed by their complete cleavage from their adjacent exons and the connection of free exon ends. In some transcripts with numerous exons, splicing reactions may combine subsets of exons to produce distinct combinations and thereby distinct protein-coding sequences from a single primary transcript sequence. This flexibility in RNA maturation is known as **alternative splicing**. It is important to note that alternative splicing adds another layer of complexity to the analysis of RNAseq data. So far, we have made an implicit assumption that the output of an RNAseq experiment, e.g. ci that map to a particular sequence, can be unambiguously assigned to a single transcript i. For many genes this assumption does not hold, since they have multiple **isoforms**, and a given read could correspond to a number of transcripts. In these cases, a more sophisticated model is necessary to assign reads to transripts - an exercise that you will go through in one of the upcoming homeworks!



Alternative splicing. Any of multiple exons may be joined in multiple combinations to produce alternative mature mRNAs from the same starting transcript. Skipped sequences are traditionally depicted with carats spanning between the included exons. The 5' cap and poly-A tail are depicted in royal blue and red, respectively.

Once mRNAs have been processed, genetic information can then be extracted from RNA by **translation**, in which an RNA transcript serves as the template for the production of a protein, which consists of one or more **polypeptide chains**. A polypeptide chain is a linear sequence of **amino acid** subunits. Twenty different amino acids can appear in polypeptides, yet they are specified from a nucleotide code comprised of just four distinct base types. This level of variation is possible by specifying amino acids with groups of three consecutive nucleotides. The set correspondence between RNA triplets, or **codons**, and their resulting amino acids is the **genetic code**.



The genetic code. Triplets of RNA nucleotides encode specified amino acids.

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

The entire genetic code. The codons specify the amino acids (listed with their standard three-letter abbreviations) or regulatory signals listed.

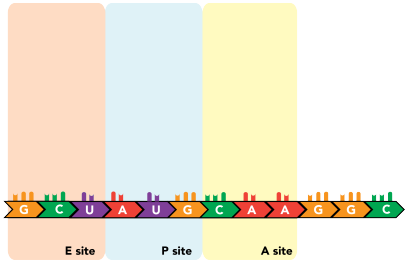
Of the sixty-four possible nucleotide triplets, the majority specify particular amino acids with redundancy in codons, but one encodes both an amino acid (methionine) and a general signal to initiate polypeptide synthesis if it has not already begun. This start site is not necessarily the first codon within the mature mRNA, and sequence upstream of it, the **5' untranslated region** (**UTR**), will not be integrated into the polypeptide chain. Similarly, translation ceases with one of the three **stop codons** which specify not amino acids, but the cessation of translation itself. Sequence further downstream, the **3' UTR** is also omitted from the protein sequence. These sequences perform regulatory functions in recruiting translational machinery and otherwise regulating translation rate, mRNA localization and stability, and similar properties.

An mRNA, with start and stop codons labeled. The sequences internal to these codons is translated into protein; the flanking regions are the 5' and 3' untranslated regions.

The act of translation involves three distinct classes of RNA molecules: the mRNAs that serve as coding templates; **transfer RNAs** (**tRNAs**), which act as adaptor molecules that bind particular amino acids and deliver them to the site of the appropriate codon; and **ribosomal RNAs** (**rRNAs**), which function as critical subunits of **ribosomes**, the sprawling

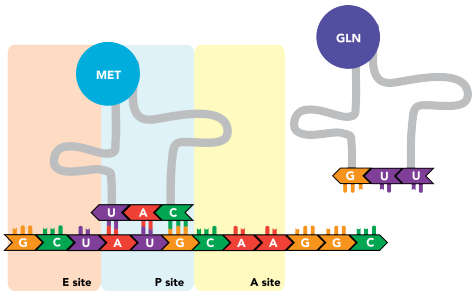
cellular machines that perform the synthetic activity of attaching an amino acid to a nascent polypeptide chain.

The ribosome contains three sites that bind mRNAs and tRNAs. The **A site** binds aminoacyl-tRNAs, or **charged** tRNAs, bound to their amino acid cargo. The **P site** binds **peptidyl-tRNA**, an RNA molecule that carries not only its own target amino acid but the entire nascent peptide chain, to which it is bound through the last added amino acid **residue** (the tRNA's own target). The **E site** is where tRNAs with cargo that has been successfully integrated into the polypeptide exit the ribosome.

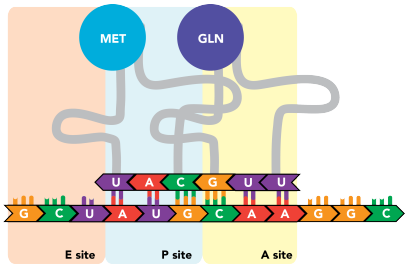


The ribosome and its sites, bound to an mRNA.

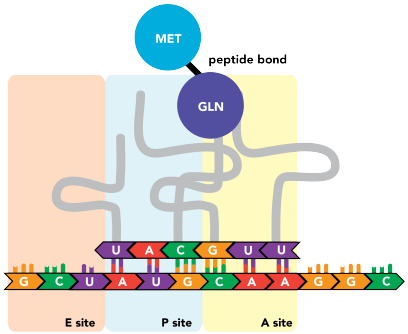
Translation is initiated with a start codon (AUG) in the P site of a partially assembled ribosome. An appropriately charged tRNA is folded into a confirmation that permits binding of an amino acid (here, methionine) at a single-stranded stretch at its 3' end and leaves an **anticodon**, the reverse complement of the mRNA codon specifying the same amino acid, exposed along another face of the molecule. For the first translated peptide, this tRNA is an **initiator tRNA**, and its binding to the start codon will promote full ribosome assembly.



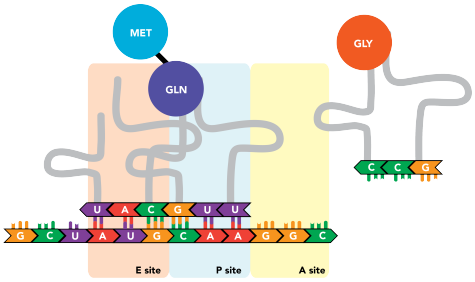
Once this occurs, a tRNA complementary to the next codon, which resides in the A site, enters the ribosome, and the codon and anticodon engage in base pairing. There are now amino acids in two ribosomal compartments: one attached to the initiator tRNA in the P site, and the next successive amino acid attached to the tRNA in the A site.



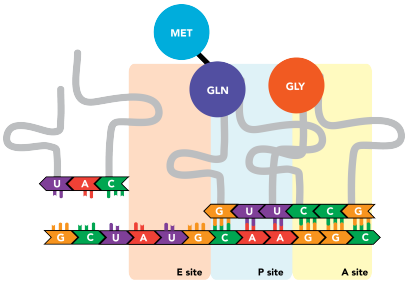
The peptide in the P site is transferred from its tRNA onto the amino acid in the A site, and the two are linked by a **polypeptide bond**.



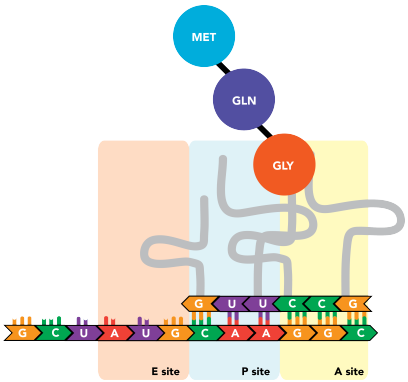
The ribosome then shifts once again.



The initiator tRNA, now in the E site and relieved of its cargo, is released from the ribosome and the transcript. Meanwhile new tRNA binds to the next successive codon in the A site.



The nascent polypeptide is transferred to the end of this next amino acid.



This process repeats itself as the ribosome moves processively along the mRNA molecule, triplet by triplet, from the 5' to 3' ends of the transcript, until a stop codon appears in the A site. There, **release factors** bind the codon and prompt the ribosome to eject its completed polypeptide chain and dissociate from the mRNA.

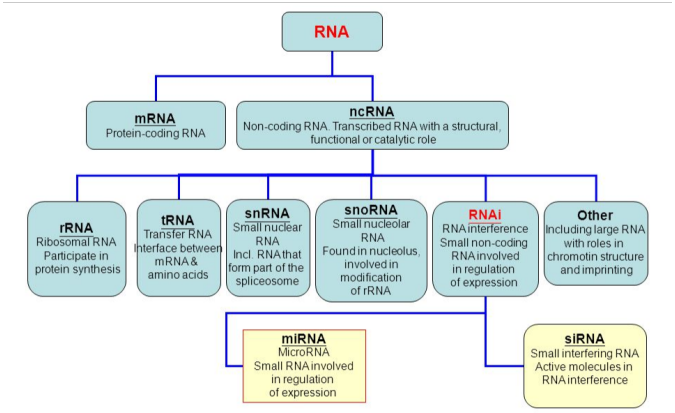
Finally, one or more linear chains of polypeptides fold into three-dimensional structures to form proteins. These proteins perform critical functions in virtually all biochemical processes within a cell--including, among many others, the transcriptional and translational activities described above.

The central dogma, revisited

We have now examined the means by which information flows from DNA sequences through RNA transcripts and into protein structure, as posited by the central dogma. However, it is crucial to recognize that all heritable sequence information does not homogeneously travel down this pathway in its totality. The discovery of retroviruses, for instance, has revealed that RNAs can act as templates for the synthesis of DNA via *reverse* transcription. In some iterations, the central dogma merely stipulates that the flow of genetic information is a path that ends with proteins, which are not capable of transmitting genetic information backwards into nucleic acids or into other proteins, allowing for more flexible interactions between nucleic acids themselves.

And even under more conventional biological paradigms, information can be lost or altered at various stages in this transmission pipeline. Not all DNA encodes genes; some sequences, including promoters, terminators, and enhancers, are regulatory regions that affect the transcription rate of nearby genes without generating gene products themselves. Other so-called "junk" DNA performs unknown functions, or possibly even none at all.

Similarly, not all transcripts encode mRNAs; other, **noncoding** RNAs, including tRNAs and rRNAs, are functional molecules in their own right without being subjected to translation. Not all regions of a coding transcript encode peptide sequence; introns are removed from mRNA precursors before translation, and in organisms that practice alternative splicing, so too are some exons. Not even all sequences within a mature mRNA code for protein, as there are untranslated regions that precede and follow the start and stop codons, respectively.

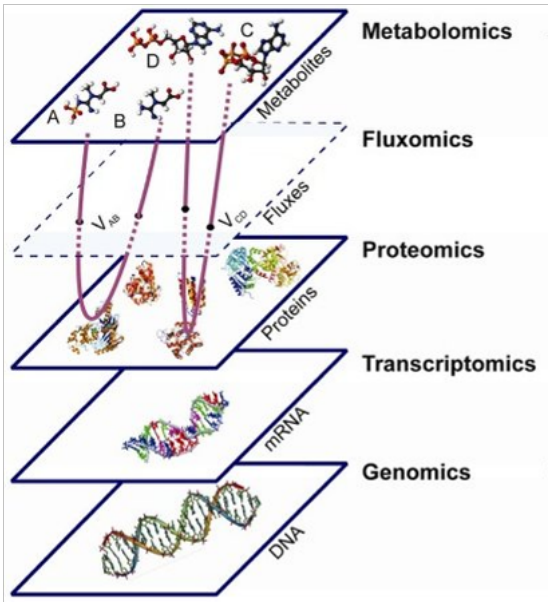


Types of RNA and functions.

Most importantly of all, the mere existence of coding sequences does not guarantee that their gene products will be expressed at particular levels, or at all. Consider the variety of tissues types present within the human body. Muscle, skin, and liver cells display wide variation in biochemical identity and physiological state despite possession of identical DNA sequences. This is due to differential gene expression across lineages (a key reason that high-throughput RNA sequencing often focuses on distinct tissue types). The ability to regulate the production of protein from given genes is an essential means by which cells can adopt distinctive biochemical traits in response to internal and external cues. An organism's genetic information is a script that supplies with the possibility to read from many possible roles and speaking lines. The set of gene products that it expresses can vary, and often do so dynamically and in response to specific environmental conditions.

A global view of genetic information and its manifestations

It is now possible to analyze the totality of an organism's DNA content, its **genome**; its complete set of mRNAs, its **transcriptome**; and every protein that it contains, its **proteome**.



Omics refers to the collective technologies used to characterise and quantify pools of biological molecules and explore their roles and relationships. These data sets--different varieties of 'omes--reflect different biological properties of an organism. The genome represents the entirety of the genetic material from which an organism can derive its protein and RNA content. Genomic data is particularly valuable for characterizing species, identifying new species, or identifying the ways in which the genes and DNA content of one organism differ from those of another. Among two organisms, one may be **mutant**, possessing heritable alterations within its DNA relative to "normal" or **wildtype** members of its species (note, however, that while wildtype quite literally refers to the form of the organism found in the wild, laboratory strains of most model organisms (bacteria, fish, sand mice) have been extensively inbred relative to their peers found in nature). (Also, more jargon: variants of a gene with different sequence identities are **alleles**, and an organism's alleles constitute its genetic identity, or **genotype**. A genotype is distinct from the physical traits that manifest within an organism, its **phenotype**, which is additionally dependent upon environmental variables).

As noted, the gene set from which an organism *can* express its gene products will vary from the set of genes that are actually being expressed for many reasons. By regulating which gene products it produces at what times, an organism can control its physiology. Analysis of the transcriptome is useful to identify changes in gene expression, which can occur dynamically, as cells proceed through various developmental, cell cycle, or other time-delimited processes, or respond to an environmental perturbation; genetically, as a function of the distinct genes that wildtype and mutant organisms possess as possible sources of transcripts, and as potential gene products that might regulate gene expression directly or indirectly in their own right; or epigenetically, as in different tissue types within a multicellular organism.

As with the distinctions between the genome and transcriptome, the content of the proteome cannot be inferred from the transcriptome. As with transcripts, the proteins that a cell contains can vary with time and context, but transcripts are short-lived, and many proteins are not. The proteome includes not just recently translated proteins, but many other proteins that a cell has produced and retained over the course of its lifetime. Protein-level information is thus a more direct and informative readout of the actual cellular machinery with which a particular cell, at a particular time, can perform biochemical reactions. Mass characterization of the entire proteome remains a relatively difficult and laborious process. In contrast, high-throughput sequencing of nucleotide content has become an increasingly accessible experimental technique for delineating a genome or transcriptome.

DNA sequencing

To understand DNA sequencing, it is necessary to understand the mechanics of DNA replication. As in transcription, DNA replication begins with the unwinding of the double helix (the **denaturation** of the double-stranded macromolecule) to produce single strands that function as template. Free nucleotides are added sequentially to the growing polymer as they bind to their complementary base pairs on the template. This synthesis occurs exclusively at the 3' end of the growing chain. The result is a DNA strand that, as in the double helix, is the reverse complement of its template, with the two strands in antiparallel orientation. During replication of the genome, both parental strands serve as

template simultaneously.

Methods for identifying the order of nucleotides within a sequence typically do so by exploiting the processive nature of replication, using the query DNA as template and detecting the addition of particular nucleotides with base-specific inhibitors and/or fluorescent dyes. The oldest (and still quite common) technique is **Sanger sequencing**. In it, source DNA is divided among four reaction tubes. Each tube contains the same short DNA sequence that is the reverse complement of a particular site on one of the source strands, where it will bind and seed polymerization of a new strand, directing replication and thus sequencing data toward this starting point. This sequence is known as a **oligonucleotide**, or **primer**. Each reaction also contains the protein that incorporates nucleotides into DNA (**DNA polymerase**) and DNA nucleotides (**deoxynucleotides**, or **dNTPs**---this is the generalized term for nucleotides of all four bases: **dATP**, **dGTP**, **dTTP**, and **dCTP**. The reactions differ in that each contains a **dideoxynucleotide** (**ddNTP**) with one of the four bases. This nucleotide is incorporated into nascent DNA normally, but because it lacks the 3' hydroxyl group that allows for the addition of additional nucleotides at the 3' end of a growing strand, it will terminate the sequencing reaction and serve as the 3' terminal end of the chain. The four ddNTPs are present at low concentrations relative to the dNTPs, so that when replication requires the relevant base, the terminating ddNTP is incorporated with low probability, and synthesis continues with the proper dNTP in the majority of occurrences.

When multiple copies of the template are replicated under these conditions, the result is a series of DNA fragments of different sizes, all of which incorporated a ddNTP at a different instance of the appropriate base and terminated afterwards. The lengths of these fragments can be used to infer the location in the sequence at which the base is present, and with four reactions--one for each base--the identity of the base at every position in the source DNA can be determined. In the original Sanger method, ddNTPs are labeled radioactively. The DNA fragments are injected into a meshwork (an **agarose gel**) and an electric field is applied, causing the negatively-charged phosphate backbones of DNA molecules to propel them down the gel. With the lattice structure of the gel impeding larger molecules, the rate of migration through the gel is dependent on the size, so that fragments are separated based on length. Autoradiographic detection techniques on the sorted gel can then detect the exact lengths of the hot fragments, revealing precisely where in the sequence a ddNTP was incorporated as an indicator of nucleotide identity at that site. In the modern adaptation, each ddNTP is labeled with a fluorescent dye of a different color, allowing the incorporation of each ddNTP to be distinguished within a single reaction instead of four.

The Sanger method remains a simple and reliable way to perform sequencing with high fidelity, although the sequence quality eventually deteriorates, limiting the interval of DNA that can be sequenced in a single reaction. This is typically a sequence fragment, or **read**, of several hundred base pairs. When determining the sequence of longer stretches of DNA, multiple Sanger reactions must be performed with primers spaced across the region to cover its sequence collectively.

Sanger sequencing represents the first generation of sequencing technology; the high-throughput permutations that have arrived since are known as **next-generation sequencing**. These are capable of performing sequencing from multiple sites in parallel to sequence full genomes with efficiency. One such technology is **pyrosequencing**, also known as **454 sequencing** after its corporation of origin. In pyrosequencing, multiple single-stranded copies of template are bound to beads that are partitioned into separate wells on a large plate. Sequencing occurs in multiple cycles---first a primer, DNA polymerase, several additional enzymes, and just one of the four possible dNTPs flow over the well. DNA replication proceeds as far as it can with this single dNTP; if this dNTP appears next in the sequence, it is incorporated. A dNTP contains three phosphate groups, but a nucleotide in a DNA chain contains just one, and integration of the nucleotide into a growing chain prompts the release of the other two phosphates in the form of **pyrophosphate**. The other enzymes present in the reaction mixture react with pyrophosphate in several steps to, ultimately, emit visible light, and a detector measures the light output of each well in response. The intensity of light reflects the number of dNTPs that are incorporated. After a time, the initial reaction mixture is washed out and replaced with another containing a different dNTP. This continues for many, many cycles.

The multiplexed nature of pyrosequencing makes it rapid and low-cost. It is also relatively accurate, but when the same base appears many times in succession, it becomes difficult to distinguish their precise numbers. Thus, depending on the sequence identity, the accuracy of pyrosequencing can vary, with repetitive sequences subject to higher error rates. The reads produced by pyrosequencing are much shorter than those generated by Sanger sequencing.

The vast majority of high-throughput sequencing now relies on **Illumina sequencing**

(again, corporate branding). Illumina sequencing begins with the random fragmentation of larger DNA sequences into short fragments. Adaptor sequences that are serve as unique barcodes for each fragment are **ligated** onto both fragment ends. Such a fragment collection is known as a **library**. The contents of the library are denatured and flown over a plate covered with small sequences that are complementary to the adaptors. Many copies of the same adaptor complements (for both ends of a fragment) appear together in well-delineated clusters on the plate, and the fragments bind to their appropriate adaptors on the plate at both ends in a bridge conformation. When a primer, DNA polymerase, and dNTPs are added, DNA replication occurs. The products are then denatured, but both strands remain attached to the plate through their adaptor complements. After several cycles, each cluster on the plate contains many copies of both strands of the original template. Sequencing both of two DNA strands, **paired-end sequencing** permits greater accuracy.

Then Illumina sequencing actually begins. Similar to Sanger sequencing, the plate is exposed to reaction mixture containing primer, DNA polymerase, and four distinctly-labeled fluorescent nucleotides, but in Illumina sequencing, *all* of the nucleotides added inhibit DNA replication once they are incorporated--but reversibly so, as each consists of a nucleotide with a removable inhibitor attached its 3' end. In the presence of this reaction mixture, a single nucleotide is added, and replication is inhibited thereafter. The identity of this nucleotide is determined from its fluorescence. Then the reaction mixture is washed out the inhibitors are cleaved from the nascent DNA chains, and fresh reaction mixture is added. This continues for many, many cycles.

Like pyrosequencing, Illumina sequencing collects sequence information quickly and on massively parallelized scales, and the presence of both complements of the template DNA within clusters generates even more sequence data. The sequential detection process also means that Illumina does not suffer from the same issues with repetitive sequences as pyrosequencing. However, Illumina reads are extremely short, at under 100 bp.

The short read lengths characteristic of all high-throughput sequencing techniques mean that sequence data is heavily fragmented and reads must be compiled into longer, more coherent forms in post-processing. (This complex, randomly fragmented output is also why such sequencing techniques are sometimes referred to as **shotgun sequencing**.) To reconstruct a genome, low-quality and extremely fragments are purged and the remaining fragments are clustered based on matching, overlapping sequence information, forming **contigs** (continguous sequences). When the genome of the organism has previously been characterized, contigs can be **mapped**, or **aligned**, directly to this **reference genome** by scanning the genome for a matching sequence. When there is no prior sequence information available, the fragments must be **assembled** *de novo*. In this case, contigs are stitched together into a final assembly without the benefit of a reference template. Genomic data can then be **annotated**, to mark features of interest, such as genes, transcriptional binding sites (and many more) within particular base pair regions.

RNA sequencing

Like the genome, the transcriptome can be characterized by nucleotide sequencing, known as **RNA-Seq**. However, RNA is prone to degradation and existing sequencing technologies are designed to sequence DNA using a DNA polymerase. To deal with these issues, RNA is isolated and subjected to reverse transcription, producing **complementary DNA** (**cDNA**) to the transcript sequence. Sequencing then proceeds with cDNA as it does genomic DNA--adaptors are attached, a library is prepared, and so on.

Like DNA polymerase, the reverse transcriptase used to make the cDNA library needs a primer to provide the 3' end to add new nucleotides. Some protocols use oligo-dT primers, which are complementary to the A-rich poly-A tail selectively added to the 3' end of mRNAs in many organisms. Collections of random primers can also be used, which can bind to other RNA species in the sample that are not mature mRNAs.

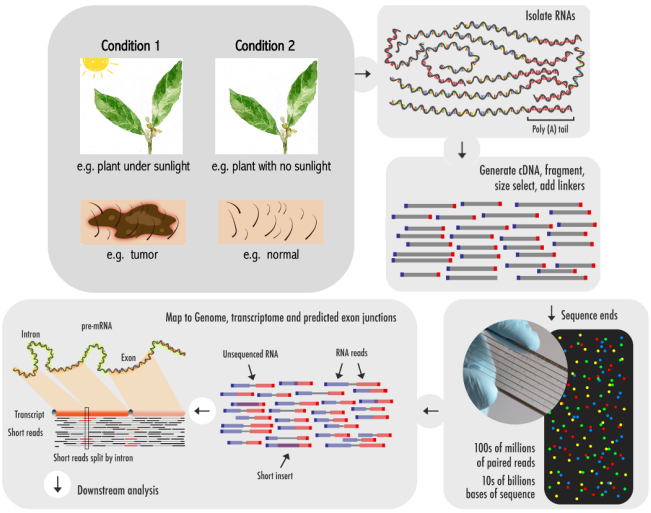
RNA-Seq can be performed on total RNA content or on mRNAs. If the focus is on sequencing mRNAs, mature, processed mRNAs possessing a poly-A tail can be selectively purified due to their affinity for complementary oligonucleotides of Ts, often expressed on easily-extracted beads. The resulting sequence information can be aligned to a reference genome, to a reference transcriptome, or assembled *de novo*.

Though the procedures for sequencing genomic and transcriptomic content are quite similar, the nature of the output is quite different. The genome is a stable, long-term storage form for genetic material, and genomic sequencing typically focuses on merely determining sequence identity (although there are applications that attempt to determine the copy number of genes or chromosomes). In contrast, the transcriptome is in constant

flux, and identifying its contents is insufficient for its characterization--it is critical to know the both the identity and quantity of each transcript.

For this reason, high **coverage** or **sequencing depth**, the number of reads that represent a particular sequence, must be extremely high for RNA-Seq, even above genomic sequencing. Further, transcript abundance is a relative value, dependent upon both the number of reads of an individual transcript and the total sample. Low coverage does not allow for accurate representation of lower-level transcripts. While desired coverage is a relatively straightforward calculation for a characterized genome, the size of the transcriptome can vary across conditions.

Another complication within RNA-Seq is the identification and resolution of ambiguous sequence that can map to multiple sites, or can correspond to multiple potential RNAs due to the prevalence of RNA splicing as a means of generating distinct RNAs. Use of sequencing techniques that generate long reads can assist with some mapping difficulties, but ultimately alignments require statistical assumptions regarding the likelihood of mapping to a particular site, inferring the probability of a particular alignment from other coverage characteristics of the region.



A workflow for a RNA-seq experiment.

The output of RNA-Seq is thus not simply a sequence but a catalog of transcripts with a measure of transcript abundance for all transcripts detected. Transcript abundance is typically calculated by normalizing number of reads for a given transcript to its length, then normalizing for sequencing depth--normalizing to the length-normalized value for all samples, divided by one million. The resulting units are **TPM**, or **transcripts per million**. These normalization steps are intended to account for both bias in coverage for longer transcripts, as well as variability in yield across experiments--TPM values should represent the proportion of reads that map to a transcript in a way that is comparable across experiments.

TPM values obtained in an RNAseq experiment are typically a *relative*, rather than an *absolute* measure of true transcript abundance in the sample, collected under very specific -- if not unique -- conditions. Therefore, beware of overinterpreting RNA-Seq data and remember that *apparent* fluctuations in TPM values for a particular gene of interest can arise either from true changes in the level of expression of this gene *or* from changes in the abundances of some other transcripts in the population. However, clever controls, particularly the use of known concentrations of particular RNAs included within samples, can allow for the validation of measurements and can even yield absolute RNA quantities. The final complication that can skew RNA-Seq results has to do with the fact that mRNA stability can vary quite wildly from transcript to transcript, with typical half-lives on the order of several hours - i.e. the same order of magnitude as the time that it takes to carry out an RNA-Seq experiment. This means that between t=0, when the sample is harvested, and t=final, when transcript abundance is quantified, the mRNA profile of the sample might have changed if a subset of less stable transcripts has been degraded. Keep this in mind when you consider the experimental set up of an RNA-Seq experiment, and think about how this phenomenon can explain the mystery of Moriarti's experiments on this week's homework!

Despite all the above-mentioned biases, with careful and informed analysis, it is possible to infer a great deal from RNA-Seq data.

Slides used in section September 11, 2020 can be found here file.

Image credits/links

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2Fa-The-organization-of-a-double-stranded-DNA-molecule-the-two-strands-running-in_fig3_35389831&psig=AOvVaw2FWG8vtXof0HCdIgdQ6gxZ&ust=1599574407851000&source=images&cd=vfe&ved=0CA0QjhxqFwoTCNj3uZm sCFQAAAAAdAAAAABAw

https://www.google.com/url?sa=i&url=http%3A%2F%2Fib.bioninja.com.au%2Fstandard-level%2Ftopic-2-molecular-biology%2F27-dna-replication-transcri%2Fcentral-dogma.html&psig=AOvVaw2eY6Lml5LMoZOFa7eNRn9V&ust=1599579172771000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCMConfau1-sCFQAAAAAdAAAAABAD

https://www.google.com/url?sa=i&url=https%3A%2F%2Fslideplayer.com%2Fslide%2F7959310%2F&psig=AOvVaw3hIrLwMvLH5cd1z62LP6Y0&ust=1599578746436000&sour sCFQAAAAAdAAAAABAD

Modified from: https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.technologynetworks.com%2Fgenomics%2Farticles%2Frna-seq-basics-applications-and-protocol-299461&psig=AOvVaw3O1S9rSBCG_xnadOOy7iPd&ust=1599574012534000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCICst-au1-sCFQAAAAAdAAAAABAD