Transcriptomics in Crop Reseach

Introduction to RNAseq technology

Mary-Ann Blätke
JJ Szymanski

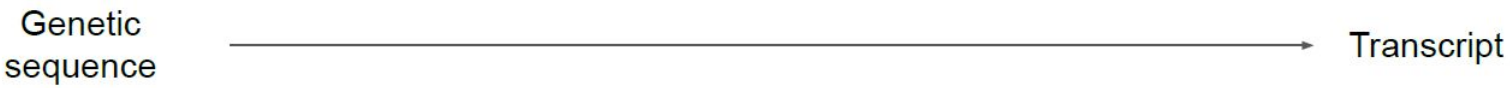
Mon 6th	Tue 7th	Wed 8th		Fri 17th
Gene to transcript Sequencing technologies Intro to bash & setup	Data formats Read mapping	Quality check Expression units Normalization		Catching up Q&A

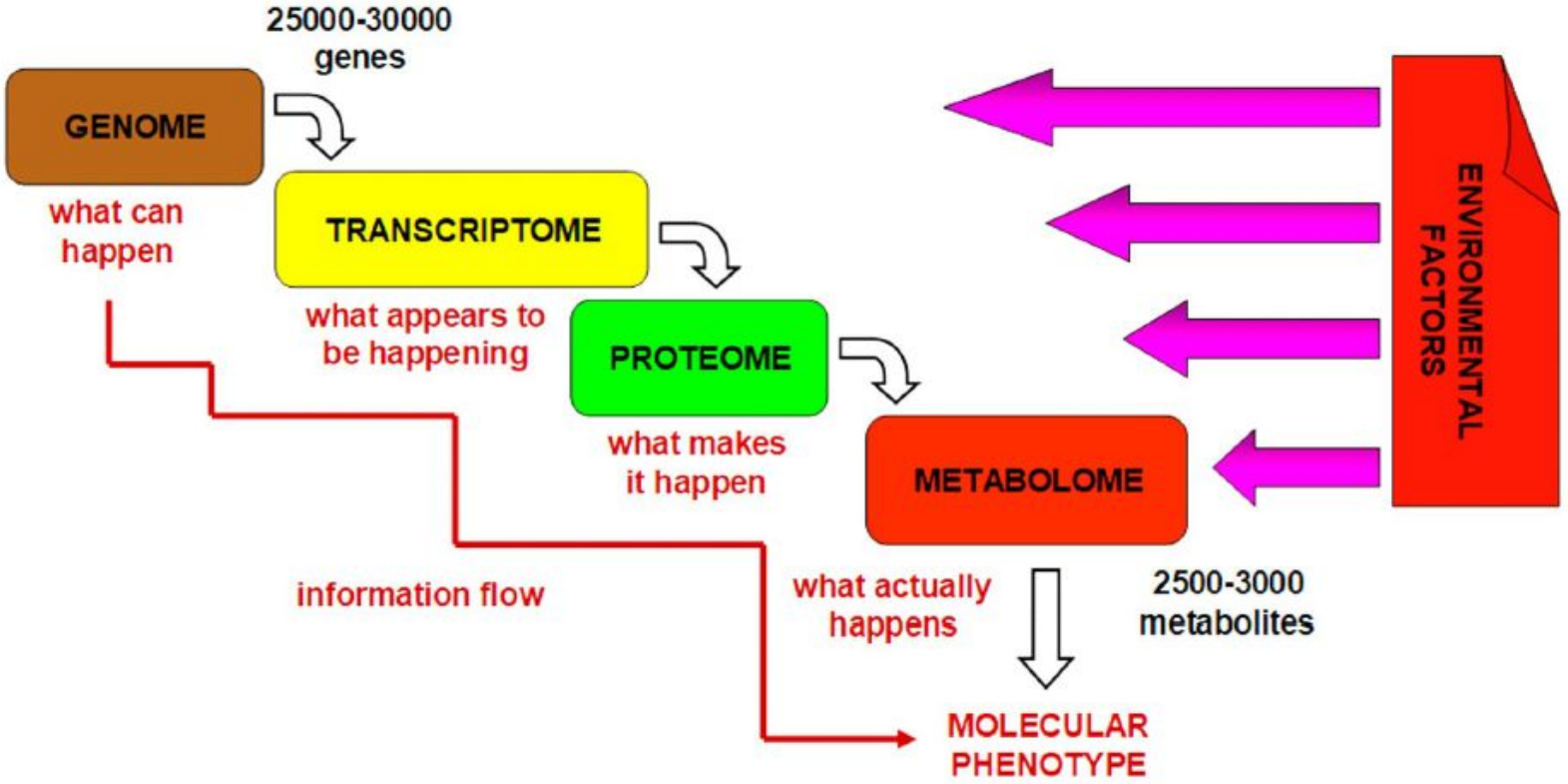
Gene to Transcript

Introduction to RNAseq technology

Mary-Ann Blätke
JJ Szymanski

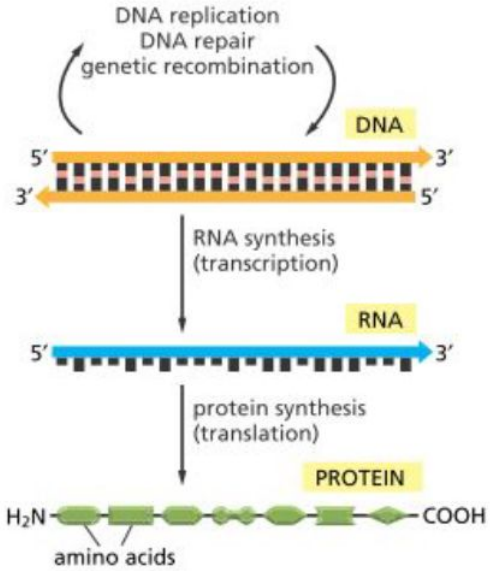

Genetic
sequence
Phenotype


Genetic
sequence
Transcript



J. Pers. Med. **2018**, *8*(3), 28; **https://doi.org/10.3390/jpm8030028**

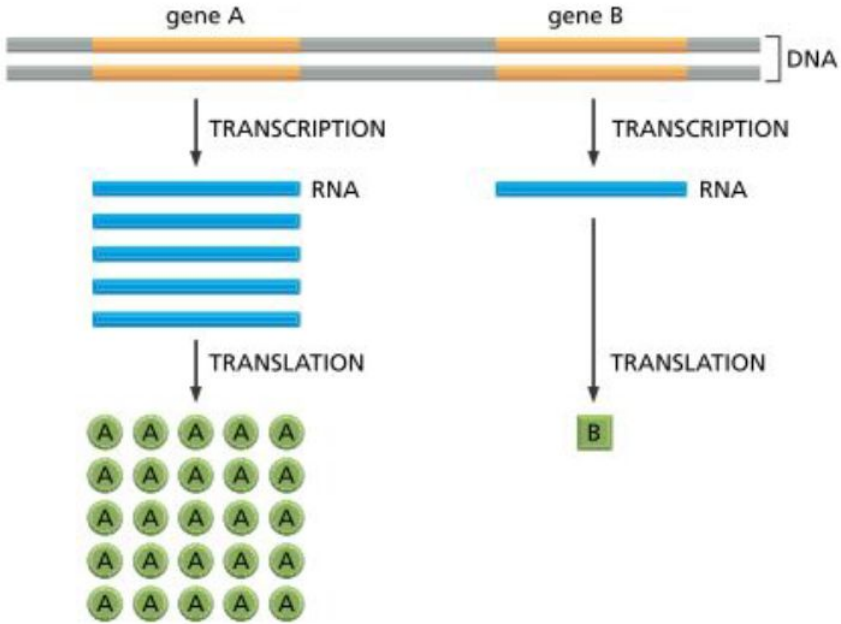
Just as a reminder



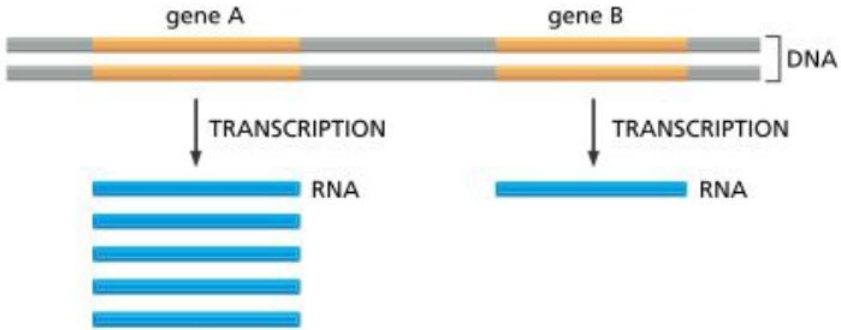
Alberts: Molecular Biology of the Cell

The amount of transcript being made is **not** the same for each gene.

Thus there a multiple levels of regulation affecting protein amount.



Here we will focus on the DNA → RNA story



RNA is made by RNA polymerases which are large multi-subunit enzymes in eukaryotes

Eukaryotes have **at least three RNA polymerases**:

- **RNA polymerase I (Pol I)** transcribes large ribosomal RNA (rRNA) genes
- **RNA polymerase II (Pol II)** transcribes messenger RNA (mRNA) genes
- **RNA polymerase III (Pol III)** transcribes a variety of RNAs including transfer RNA (tRNA) and 5S ribosomal RNA
- Plants have a fourth RNA polymerase that transcribes regulatory RNAs
- some plants have a fifth RNA polymerase

Bacteria and archaea have a **single RNA polymerase**

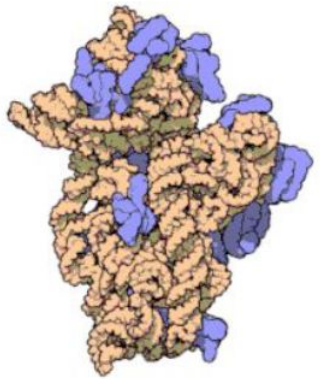
The cell contains many different types of RNA.

mRNA is studied for its role in gene regulation,

rRNA often makes up the **bulk amount of RNA in a cell ~80%**

Table 6–1 Principal Types of RNAs Produced in Cells

TYPE OF RNA	FUNCTION
mRNAs	messenger RNAs, code for proteins
rRNAs	ribosomal RNAs, form the basic structure of the ribosome and catalyze protein synthesis
tRNAs	transfer RNAs, central to protein synthesis as adaptors between mRNA and amino acids
snRNAs	small nuclear RNAs, function in a variety of nuclear processes, including the splicing of pre-mRNA
snoRNAs	small nucleolar RNAs, used to process and chemically modify rRNAs
scaRNAs	small cajal RNAs, used to modify snoRNAs and snRNAs
miRNAs	microRNAs, regulate gene expression typically by blocking translation of selective mRNAs
siRNAs	small interfering RNAs, turn off gene expression by directing degradation of selective mRNAs and the establishment of compact chromatin structures
Other noncoding RNAs	function in diverse cell processes, including telomere synthesis, X-chromosome inactivation, and the transport of proteins into the ER



Alberts: Molecular Biology of the Cell

Table 1. Summary of single-letter code recommendations

Symbol	Meaning	Origin of designation
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G or A	puRine
Y	T or C	pYrimidine
M	A or C	aMino
K	G or T	Ketone
S	G or C	Strong interaction (3 H bonds)
W	A or T	Weak interaction (2 H bonds)
H	A or C or T	not-G, H follows G in the alphabet
B	G or T or C	not-A, B follows A
V	G or C or A	not-T (not-U), V follows U
D	G or A or T	not-C, D follows C
N	G or A or T or C	aNy

5. DISCUSSION

The present nomenclature, summarised in Table 1, has been formulated to deal with incomplete specification of bases in nucleic acid sequences. In cases where two or more bases are permitted at a particular position the nomenclature permits the allocation of a single-letter symbol. The nomenclature may also be applied where uncertainty exists as to extent and/or identity. For double-stranded nucleic acids Table 2 permits the allocation of symbols to the complementary strand. Examples are given whereby the nomenclature is applied to sequences recognised by certain type II restriction endonucleases (Table 3) and to uncertainties in deriving a nucleic acid sequence from the corresponding amino acid sequence (Table 4).

Two applications fall outside the scope of the nomenclature and these are considered separately below.

Description	Symbol	Bases represented					Complementary bases
		No.	A	C	G	T	
Adenine	**A**	1	A				T
Cytosine	**C**			C			G
Guanine	**G**				G		C
Thymine	**T**					T	A
Uracil	**U**					U	A
Weak	**W**	2	A			T	W
Strong	**S**			C	G		S
Amino	**M**		A	C			K
Ketone	**K**				G	T	M
Purine	**R**		A		G		Y
Pyrimidine	**Y**			C		T	R
Not A	**B**	3		C	G	T	V
Not C	**D**		A		G	T	H
Not G	**H**		A	C		T	D
Not T[a]	**V**		A	C	G		B
Any one base	**N**	4	A	C	G	T	N
Gap	**-**	0					-

a. ^ Not U for RNA

Cornish-Bowden (1985) 13(9): 3021–3030. wikipedia

CCCAGGGTCCGATGGGAAAGTGTAGCCTGCAGGCCCACACCTCCCCCTGTGAATCACGCCTGGCGGGACA

>gi|186704|Keratin Homo sapiens keratin
CCCAGGGTCCGATGGGAAAGTGTAGCCTGCAGGCCCACACCTCCCCCTGTGAATCACGCCTGGCGGGACA
AGAAAGCCCAAAACACTCCAAACAATGAGTTTCCAGTAAAATATGACAGACATGATGAGGCGGATGAGAG
GAGGGACCTGCCTGGGAGTTGGCGCTAGCCTGTGGGTGATGAAAGCCAAGGGGAATGGAAAGTGCCAGAC
CCGCCCCCTACCCATGAGTATAAAGCACTCGCATCCCTTTGCAATTTACCCGAGCACCTTCTCTTCACTC
AGCCTTCTGCTCGCTCGCTCACCTCCCTCCTCTGCACCATGACTACCTGCAGCCGCCAGTTCACCTCCTC
CAGCTCCATGAAGGGCTCCTGCGGCATCGGGGGCGGCATCGGGGCGGGCTCCAGCCGCATCTCCTCCGTC
CTGGCCGGAGGGTCCTGCCGCGCCCCCAACACCTACGGGGGCGGCCTGTCTGTCTCATCCTCCCGCTTCT
CCTCTGGGGGAGCCTATGGGTTGGGGGGCGGCTATGGCGGTGGCTTCAGCAGCAGCAGCAGCAGCTTTGG
TAGTGGCTTTGGGGGAGGATATGGTGGTGGCCTTGGTGCTGGCTTGGGTGGTGGCTTTGGTGGTGGCTTT
GCTGGTGGTGATGGGCTTCTGGTGGGCAGTGAGAAGGTGACCATGCAGAACCTCAATGACCGCCTGGCCT
CCTACCTGGACAAGGTGCGTGCTCTGGAGGAGGCCAACGCCGACCTGGAAGTGAAGATCCGTGACTGGTA
CCAGAGGCAGCGGCCTGCTGAGATCAAAGACTACAGTCCCTACTTCAAGACCATTGAGGACCTGAGGAAC
AAGGTGGGTGAATGGGCAGCAGAAGGCACCATTCCAGCTAGCTCCTTCTGGGAACAATTCATGCCCCAGG
CCGCTGAGACCTTAAGATTTCTCTATAGGACAGAGTCCACCCCAGATCCCTTCTTTCGAGGTCTTGGATG
CCCTAAGACTGATCAGTGAGAAGATGCTTTCCCTTCCCCAGGCCTCCTCATCCCCTTCTGATCTCAAATC

We will in almost all cases only write **one strand** of DNA in the FASTA format

FASTA format
One line with “>” then identifier
Multiple lines with sequence typically 80,120 etc characters per line

```
>gi|186704|Keratin Homo sapiens keratin
CCCAGGGTCCGATGGGAAAGTGTAGCCTGCAGGCCCACACCTCCCCCTGTGAATCACGCCTGGCGGGACA
AGAAAGCCCAAAACACTCCAAACAATGAGTTTCCAGTAAAATATGACAGACATGATGAGGCGGATGAGAG
GAGGGACCTGCCTGGGAGTTGGCGCTAGCCTGTGGGTGATGAAAGCCAAGGGGAATGGAAAGTGCCAGAC
CCGCCCCCTACCCATGAGTATAAAGCACTCGCATCCCTTTGCAATTTACCCGAGCACCTTCTCTTCACTC
AGCCTTCTGCTCGCTCGCTCACCTCCCTCCTCTGCACCATGACTACCTGCAGCCGCCAGTTCACCTCCTC
CAGCTCCATGAAGGGCTCCTGCGGCATCGGGGGCGGCATCGGGGCGGGCTCCAGCCGCATCTCCTCCGTC
CTGGCCGGAGGGTCCTGCCGCGCCCCCAACACCTACGGGGGCGGCCTGTCTGTCTCATCCTCCCGCTTCT
CCTCTGGGGGAGCCTATGGGTTGGGGGGCGGCTATGGCGGTGGCTTCAGCAGCAGCAGCAGCAGCTTTGG
TAGTGGCTTTGGGGGAGGATATGGTGGTGGCCTTGGTGCTGGCTTGGGTGGTGGCTTTGGTGGTGGCTTT
GCTGGTGGTGATGGGCTTCTGGTGGGCAGTGAGAAGGTGACCATGCAGAACCTCAATGACCGCCTGGCCT
CCTACCTGGACAAGGTGCGTGCTCTGGAGGAGGCCAACGCCGACCTGGAAGTGAAGATCCGTGACTGGTA
CCAGAGGCAGCGGCCTGCTGAGATCAAAGACTACAGTCCCTACTTCAAGACCATTGAGGACCTGAGGAAC
AAGGTGGGTGAATGGGCAGCAGAAGGCACCATTCCAGCTAGCTCCTTCTGGGAACAATTCATGCCCCAGG
CCGCTGAGACCTTAAGATTTCTCTATAGGACAGAGTCCACCCCAGATCCCTTCTTTCGAGGTCTTGGATG
CCCTAAGACTGATCAGTGAGAAGATGCTTTCCCTTCCCCAGGCCTCCTCATCCCCTTCTGATCTCAAATC
```

We will in almost all cases only write **one strand** of DNA in the FASTA format

FASTA format
One line with “>” then identifier
Multiple lines with sequence typically 80,120 etc characters per line

```
>gi|186704|Keratin Homo sapiens keratin
CCCAGGGTCCGATGGGAAAGTGTAGCCTGCAGGCCCACACCTCCCCCTGTGAATCACGCCTGGCGGGACA
AGAAAGCCCAAAACACTCCAAACAATGAGTTTCCAGTAAAATATGACAGACATGATGAGGCGGATGAGAG
GAGGGACCTGCCTGGGAGTTGGCGCTAGCCTGTGGGTGATGAAAGCCAAGGGGAATGGAAAGTGCCAGAC
```

Some older programs can only parse few characters in the identifier or expect certain line lengths

```
CTGGCCGGAGGGTCCTGCCGCGCCCCCAACACCTACGGGGGCGGCCTGTCTGTCTCATCCTCCCGCTTCT
```

Many programs have issues with line endings
This is /r/n in windows CR LF
/n in Linux LF
And
/r in some old Macs CR

```
CCGCTGAGACCTTAAGATTTCTCTATAGGACAGAGTCCACCCCAGATCCCTTCTTTCGAGGTCTTGGATG
CCCTAAGACTGATCAGTGAGAAGATGCTTTCCCTTCCCCAGGCCTCCTCATCCCCTTCTGATCTCAAATC
```

NCBI identifiers [edit]

The NCBI defined a standard for the unique identifier used for the sequence (SeqID) in the header line. This allows a sequence that was obtained frc database to be labelled with a reference to its database record. The database identifier format is understood by the NCBI tools like `makeblastdb` a `table2asn` . The following list describes the NCBI FASTA defined format for sequence identifiers.[5]

Type	Format(s)	Example(s)
local (i.e. no database reference)	`lcl|integer` `lcl|string`	`lcl|123` `lcl|hmm271`
GenInfo backbone seqid	`bbs|integer`	`bbs|123`
GenInfo backbone moltype	`bbm|integer`	`bbm|123`
GenInfo import ID	`gim|integer`	`gim|123`
GenBank	`gb|accession|locus`	`gb|M73307|AGMA13GT`
EMBL	`emb|accession|locus`	`emb|CAM43271.1|`
PIR	`pir|accession|name`	`pir||G36364`
SWISS-PROT	`sp|accession|name`	`sp|P01013|OVAX_CHICK`
patent	`pat|country|patent|sequence-number`	`pat|US|RE33188|1`
pre-grant patent	`pgp|country|application-number|sequence-number`	`pgp|EP|0238993|7`
RefSeq	`ref|accession|name`	`ref|NM_010450.1|`
general database reference (a reference to a database that's not in this list)	`gnl|database|integer` `gnl|database|string`	`gnl|taxon|9606` `gnl|PID|e1632`
GenInfo integrated database	`gi|integer`	`gi|21434723`
DDBJ	`dbj|accession|locus`	`dbj|BAC85684.1|`
PRF	`prf|accession|name`	`prf||0806162C`
PDB	`pdb|entry|chain`	`pdb|1I4L|D`
third-party GenBank	`tpg|accession|name`	`tpg|BK003456|`
third-party EMBL	`tpe|accession|name`	`tpe|BN000123|`
third-party DDBJ	`tpd|accession|name`	`tpd|FAA00017|`
TrEMBL	`tr|accession|name`	`tr|Q90RT2|Q90RT2_9HIV1`

wikipedia

NCBI identifiers [edit]

The NCBI defined a standard for the unique identifier used for the sequence (SeqID) in the header line. This allows a sequence that was obtained frc database to be labelled with a reference to its database record. The database identifier format is understood by the NCBI tools like `makeblastdb` a `table2asn`. The following list describes the NCBI FASTA defined format for sequence identifiers.[5]

Type	Format(s)	Example(s)
local (i.e. no database reference)	`lcl|integer` `lcl|string`	`lcl|123` `lcl|hmm271`
GenInfo backbone seqid	`bbs|integer`	`bbs|123`
GenInfo backbone moltype	`bbm|integer`	`bbm|123`
GenInfo import ID	`gim|integer`	`gim|123`
GenBank	`gb|accession|locus`	`gb|M73307|AGMA13GT`
EMBL	`emb|accession|locus`	`emb|CAM43271.1|`
PIR	`pir|accession|name`	`pir||G36364`
SWISS-PROT	`sp|accession|name`	`sp|P01013|OVAX_CHICK`
patent	`pat|country|patent|sequence-number`	`pat|US|RE33188|1`
pre-grant patent	`pgp|country|application-number|sequence-number`	`pgp|EP|0238993|7`
RefSeq	`ref|accession|name`	`ref|NM_010450.1|`
general database reference (a reference to a database that's not in this list)	`gnl|database|integer` `gnl|database|string`	`gnl|taxon|9606` `gnl|PID|e1632`
GenInfo integrated database	`gi|integer`	`gi|21434723`
DDBJ	`dbj|accession|locus`	`dbj|BAC85684.1|`
PRF	`prf|accession|name`	`prf||0806162C`
PDB	`pdb|entry|chain`	`pdb|1I4L|D`
third-party GenBank	`tpg|accession|name`	`tpg|BK003456|`
third-party EMBL	`tpe|accession|name`	`tpe|BN000123|`
third-party DDBJ	`tpd|accession|name`	`tpd|FAA00017|`
TrEMBL	`tr|accession|name`	`tr|Q90RT2|Q90RT2_9HIV1`

WHY COMPLEXITY?

wikipedia

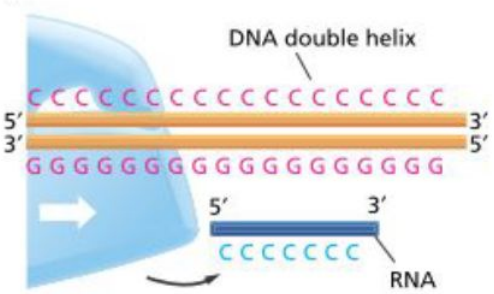
We will in almost all cases only write **one strand** of DNA in the FASTA format

```
>gi|186704|Keratin Homo sapiens keratin
CCCAGGGTCCGATGGGAAAGTGTAGCCTGCAGGCCCACACCTCCCCCTGTGAATCACGCCTGGCGGGACA
AGAAAGCCCAAAACACTCCAAACAATGAGTTTCCAGTAAAATATGACAGACATGATGAGGCGGATGAGAG
GAGGGACCTGCCTGGGAGTTGGCGCTAGCCTGTGGGTGATGAAAGCCAAGGGGAATGGAAAGTGCCAGAC
CCGCCCCCTACCCATGAGTATAAAGCACTCGCATCCCTTTGCAATTTACCCGAGCACCTTCTCTTCACTC
AGCCTTCTGCTCGCTCGCTCACCTC...
```

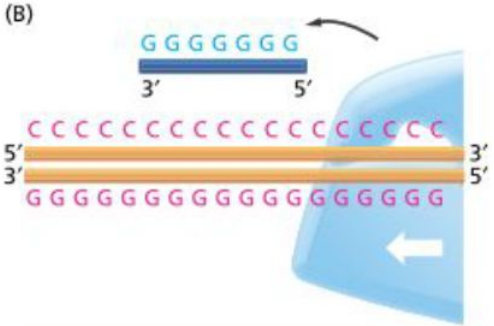
We will in almost all cases only write **one strand** of DNA in the FASTA format

When studying genomes it is important to keep in mind that genes can be **encoded on both strands of the DNA**.

```
>gi|186704|Keratin Homo sapiens keratin
CCCAGGGTCCGATGGGAAAGTGTAGCCTGCAGGCCCACACCTCCCCCTGTGAATCACGCCTGGCGGGACA
AGAAAGCCCAAAACACTCCAAACAATGAGTTTCCAGTAAAATATGACAGACATGATGAGGCGGATGAGAG
GAGGGACCTGCCTGGGAGTTGGCGCTAGCCTGTGGGTGATGAAAGCCAAGGGGAATGGAAAGTGCCAGAC
CCGCCCCCTACCCATGAGTATAAAGCACTCGCATCCCTTTGCAATTTACCCGAGCACCTTCTCTTCACTC
AGCCTTCTGCTCGCTCGCTCACCTC...
```

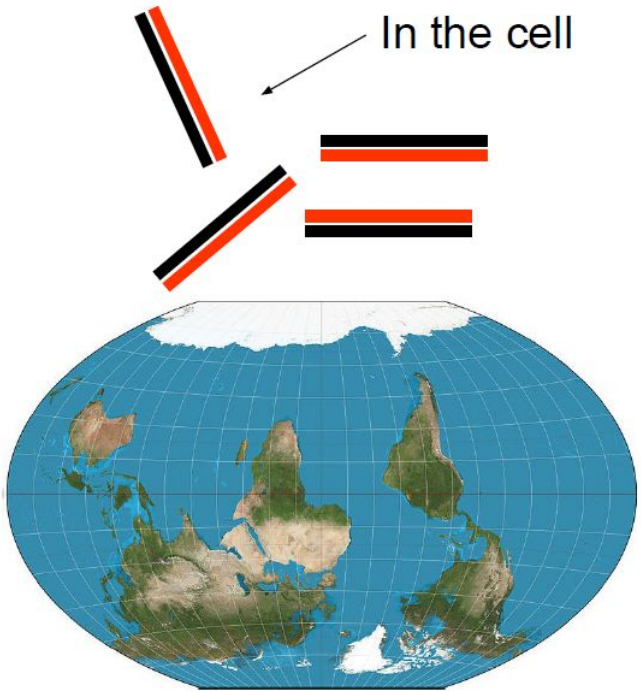


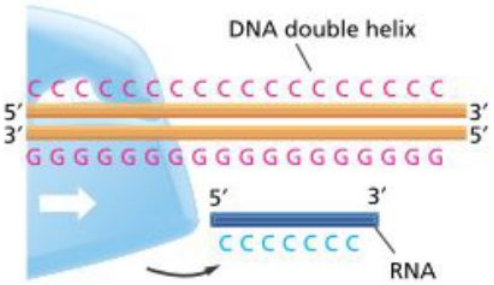
an RNA polymerase that moves from left to right makes RNA by using the bottom strand as a template



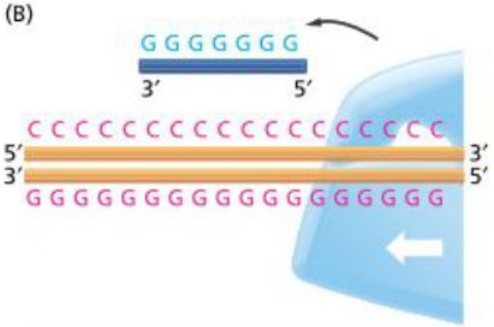
an RNA polymerase that moves from right to left makes RNA by using the top strand as a template

Alberts: Molecular Biology of the Cell





an RNA polymerase that moves from left to right makes RNA by using the bottom strand as a template

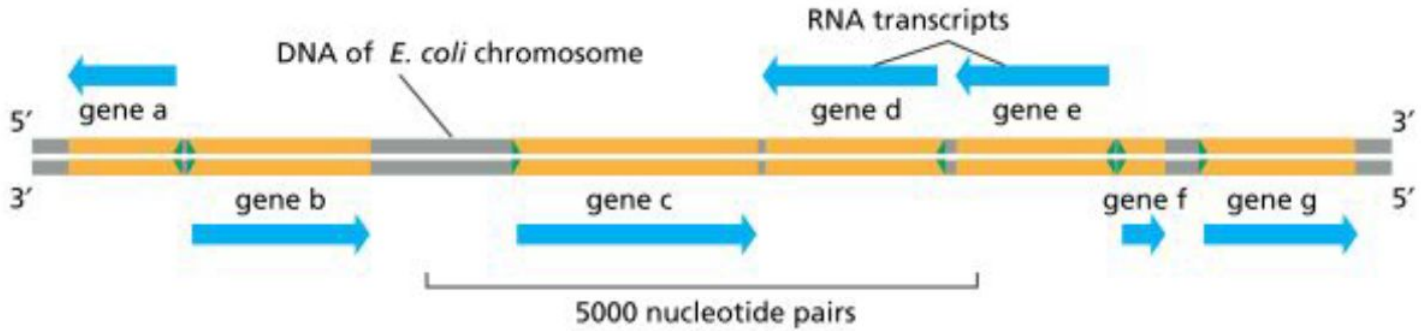


an RNA polymerase that moves from right to left makes RNA by using the top strand as a template

Alberts: Molecular Biology of the Cell

But how does an RNA polymerase know which strand to read from and where to start transcription?

Promoter elements direct the RNA Polymerase. These regions on the DNA often consist of short **DNA stretches with conserved sequence** to which the some auxiliary factors bind.



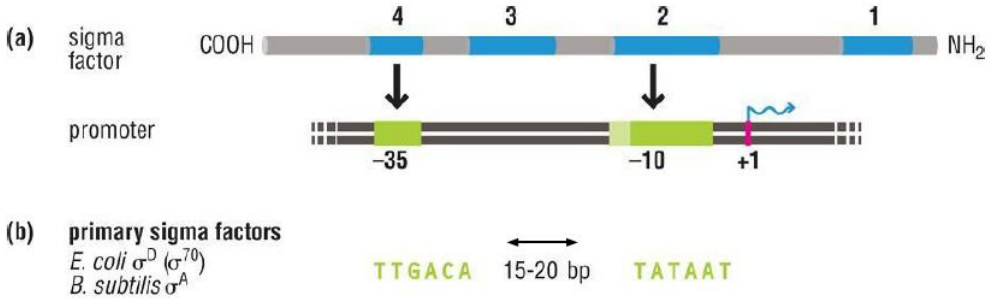
▲ Promoter elements

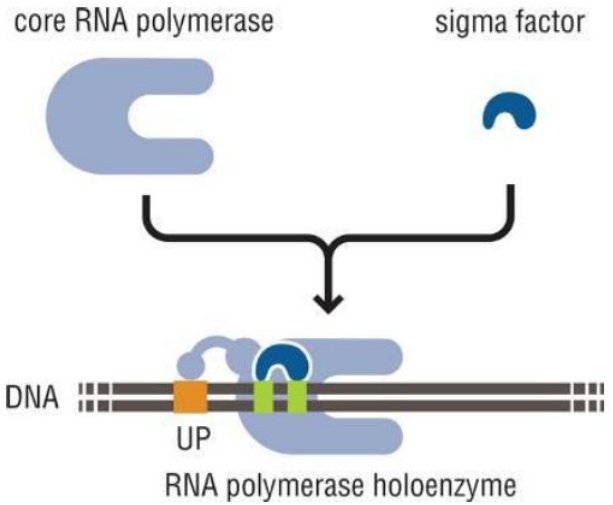
Alberts: Molecular Biology of the Cell

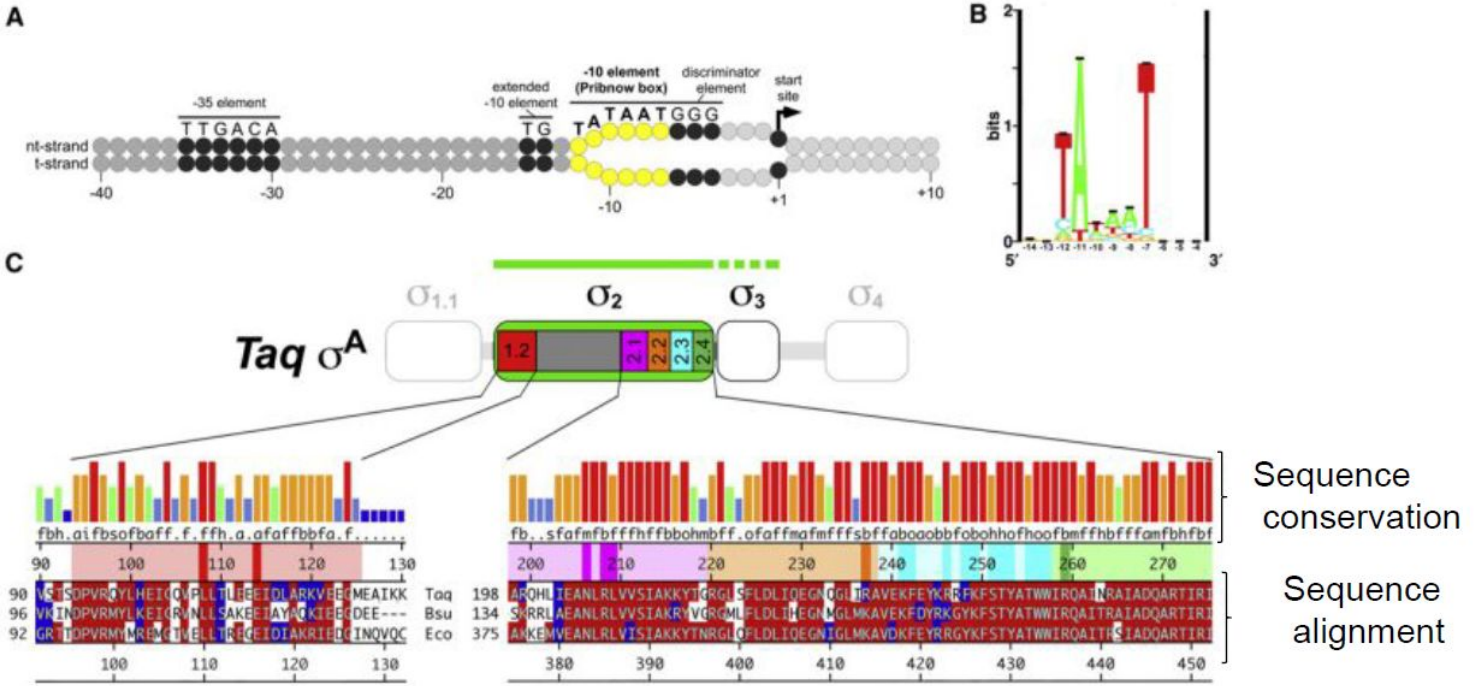
Basal bacterial promoters generally have two elements: a -35 element and a -10 element. These are roughly 35 and 10 bases upstream of the transcription start site

Sigma factors bind sequences that define the bacterial promoters and each sigma factor has sequences it prefers to bind to, and has a preferred spacing between -35 and -10

Some promoters might have additional elements, e.g. very active ones have an AT rich sequence the ***UP element*** which is contacted by the C-terminal domain of RNA Polymerase α subunit







Promoter motifs recognized by primary bacterial RNAP factors. Circles are bases and black and yellow circles are recognized.

The resulting sequence logo (another better? Way to sequence frequency histograms)

Cell 147 , 1257–1269

Eukaryotes and transcriptional start points

Eukaryotic RNA polymerases need the TATA binding protein (TBP) to initiate transcription (this is part of TFIID) TBP binds to the TATA box if this is present (~25-30 bp upstream of TS start site)

It was presumed that not all genes have a TATA box but only about 1/3 of genes.

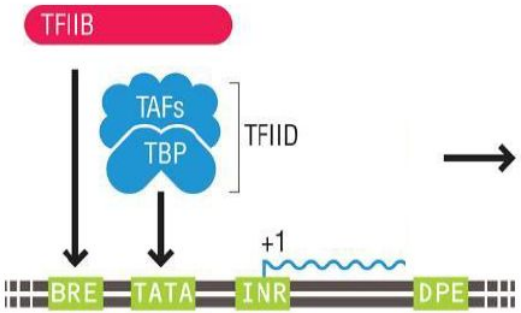
• The first step in assembling one transcription initiation complex is often binding of TFIID to the TATA box

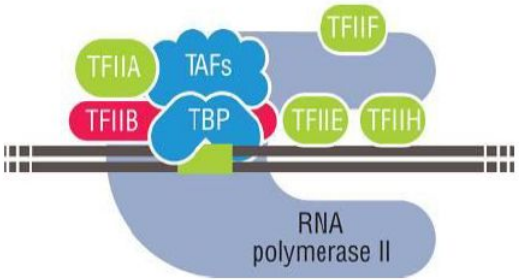
• TFIID binds to the TATA box via TBP, which binds to the minor groove of DNA, inducing strong distortions in the DNA and thus local DNA unwinding

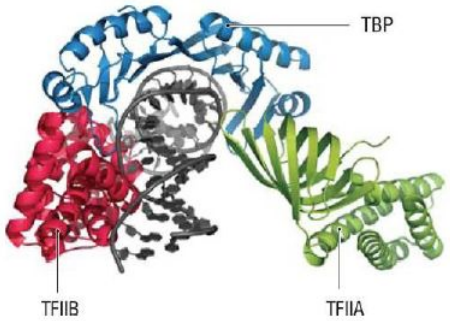
•Other components of TFIID, called TBP-associated factors (TAFs), mediate recognition of other promoter elements like INR and DPE

• After TFIID has associated with DNA, TFIIB is recruited. This recognizes the BRE promoter element and binds asymmetrically, helping to determine the transcription direction. TFIIB has some similarities to bacterial sigma factor

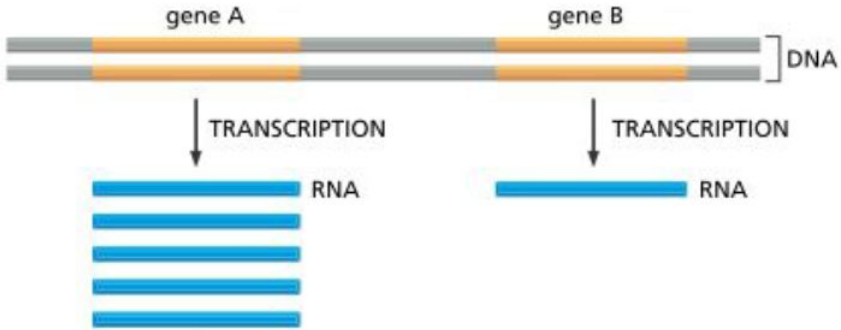
• After TFIID and TFIIB have bound, TFIIA binds, and stabilizes the TBP-DNA interactions, then TFIIE and TFIIH (TFIIH catalyzes ATP-powered DNA unwinding)







Here we will focus on the DNA → RNA story



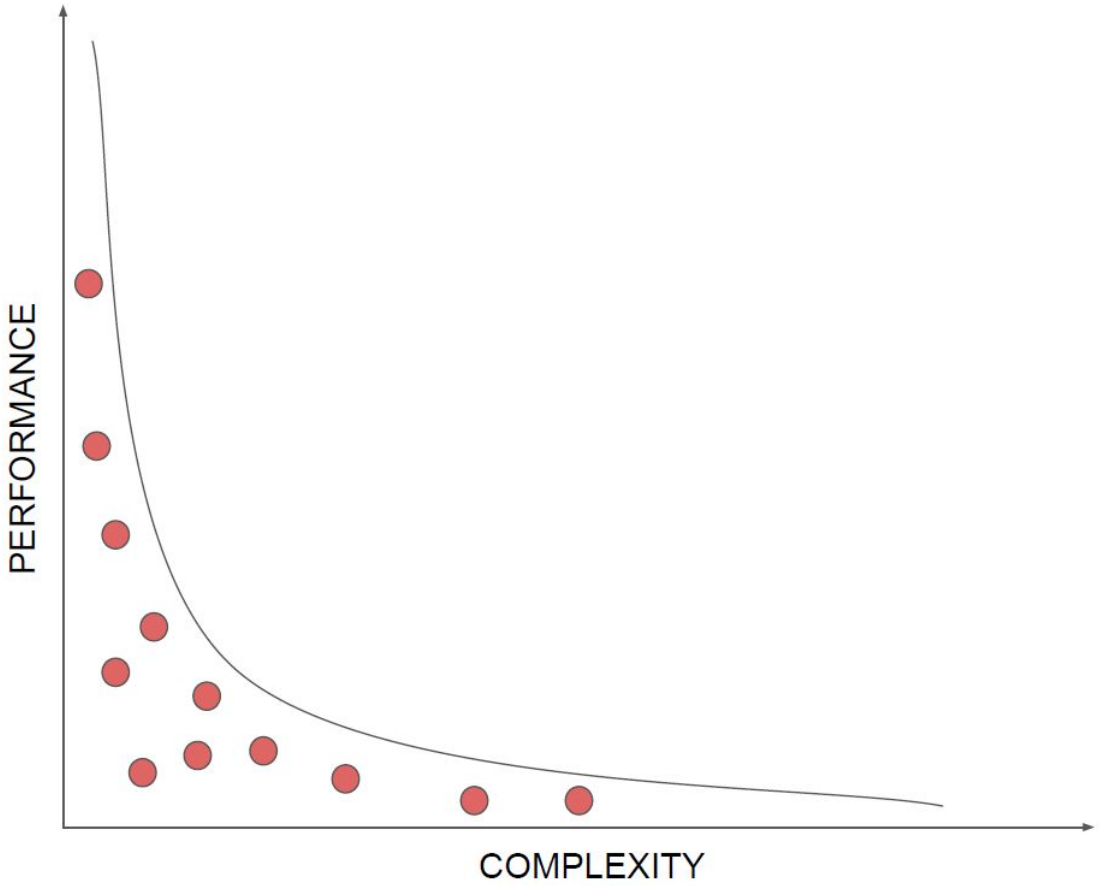
RNA is made by RNA polymerases which are large multi-subunit enzymes in eukaryotes

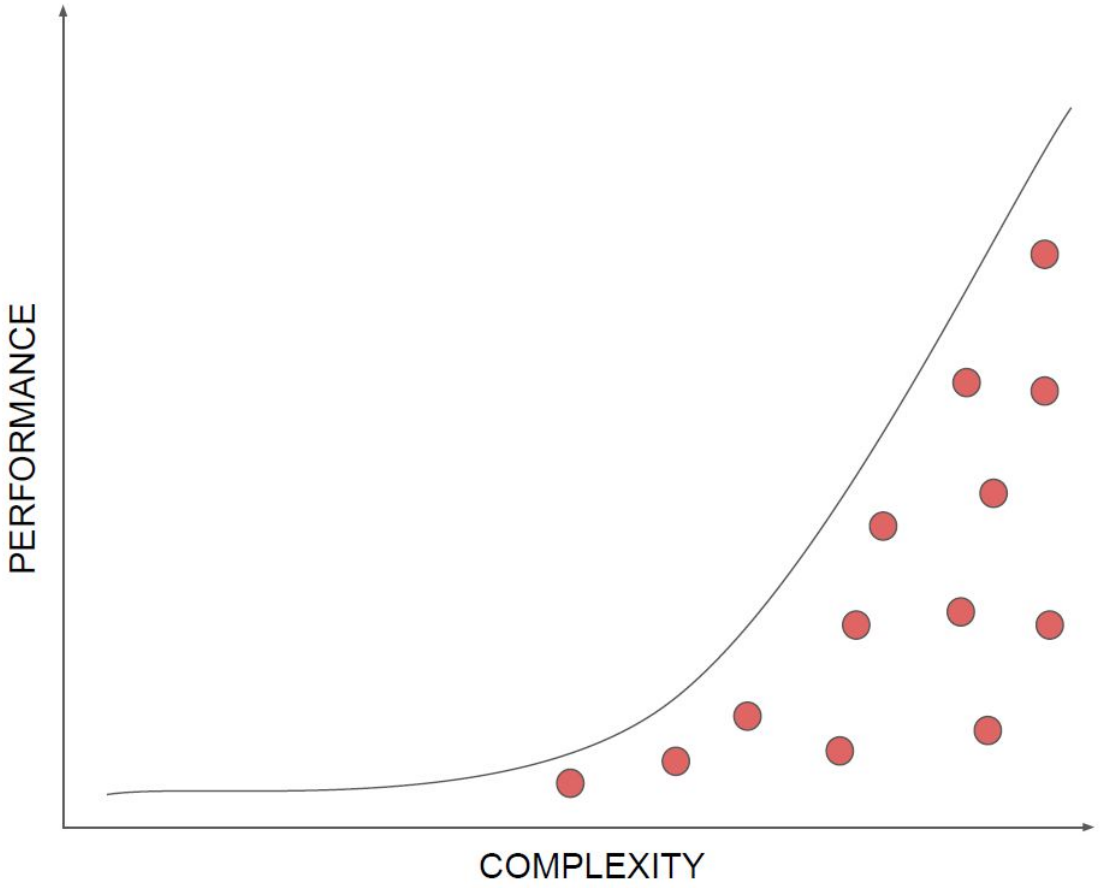
Eukaryotes have **at least three RNA polymerases**:

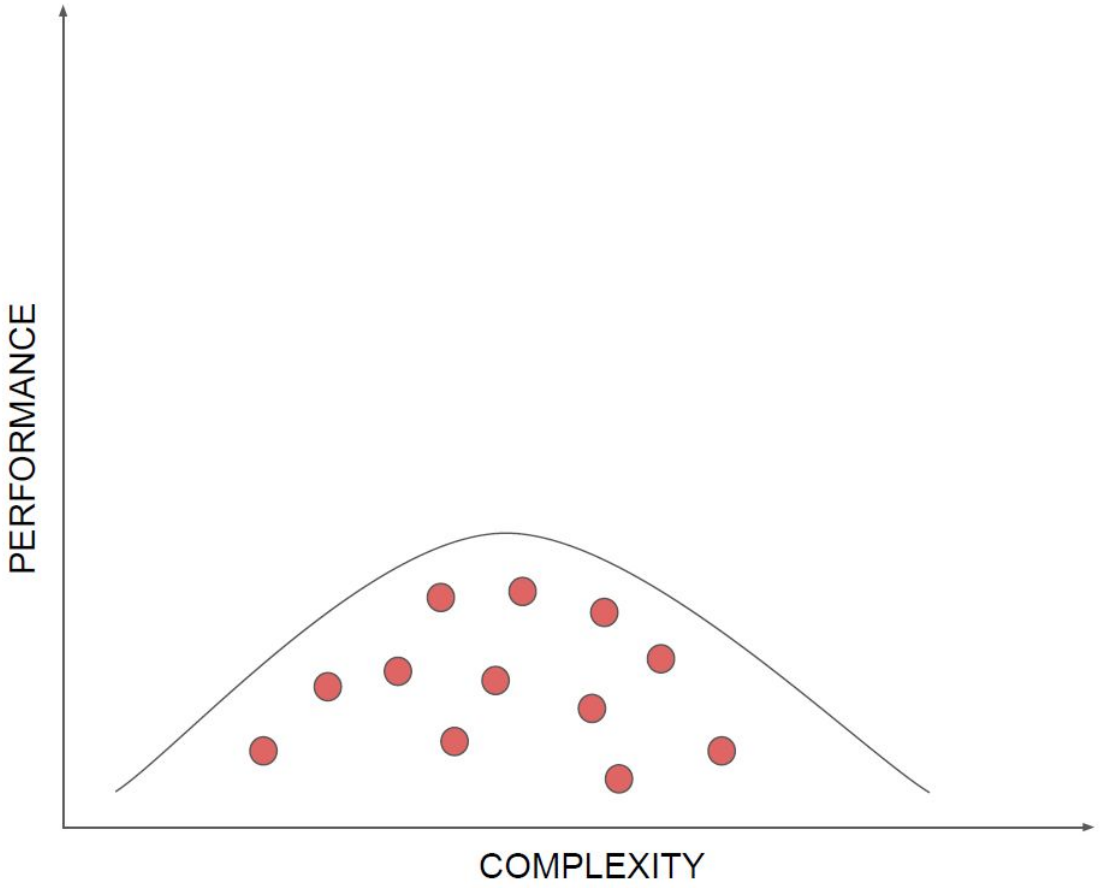
- **RNA polymerase I (Pol I)** transcribes large ribosomal RNA (rRNA) genes
- **RNA polymerase II (Pol II)** transcribes messenger RNA (mRNA) genes
- **RNA polymerase III (Pol III)** transcribes a variety of RNAs including transfer RNA (tRNA) and 5S ribosomal RNA
- Plants have a fourth RNA polymerase that transcribes regulatory RNAs
- some plants have a fifth RNA polymerase

WHY COMPLEXITY?

Bacteria and archaea have a **single RNA polymerase**


PERFORMANCE
COMPLEXITY


PERFORMANCE
COMPLEXITY


PERFORMANCE
COMPLEXITY