Outline

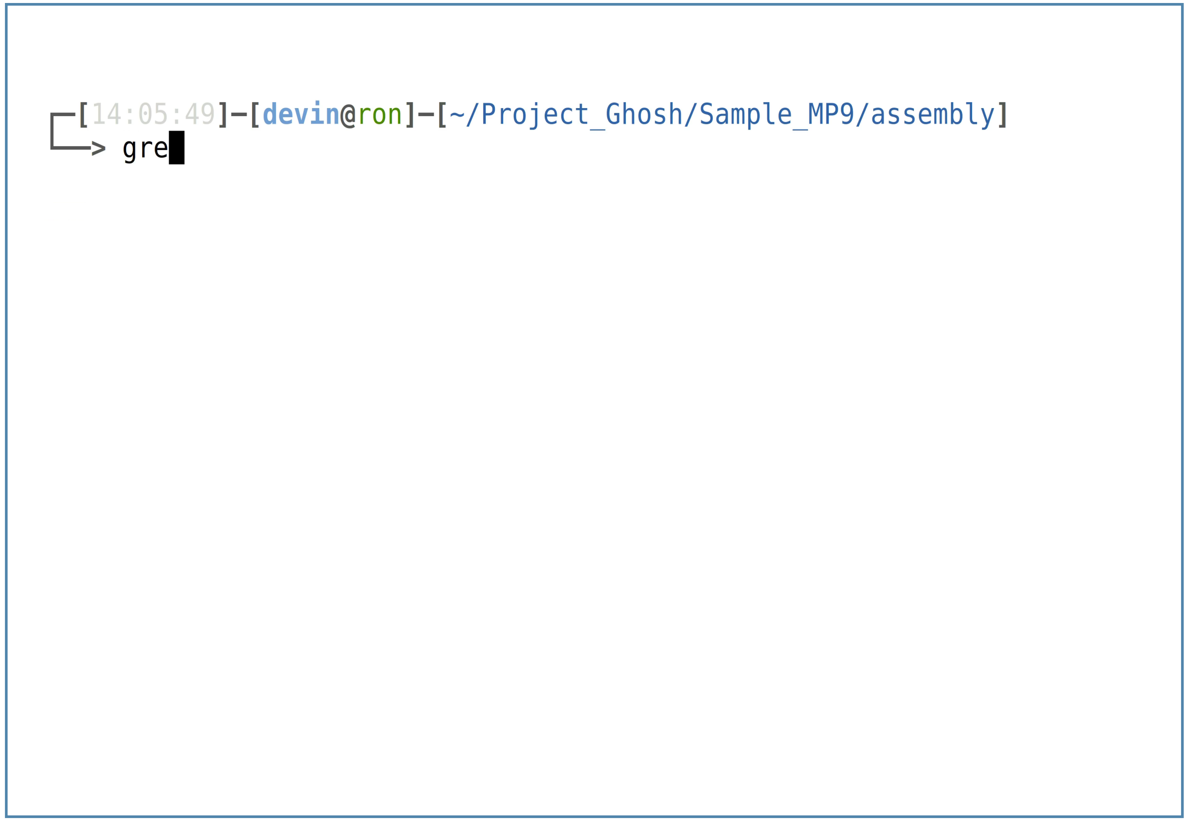
- Bioinformatics philosophy
  - We will be working in the bash or “command line” environment.
- Environmental genomics
  - We will focus on practice and principles critical to envirnomantsl genomics
- How is the data generated?
  - Critical to know limits and biases
- Data “structures”
  - Bioinformatics requires predictable data formats and conventions

Why are we pushing the Bourne Again SHell (BASH)?

- It is how real bioinformatics is done.
- Opens door to vast library of software.
- This is how people deal with all forms of “Big Data”.
- This is stable/durable knowledge.

Unfortunately:

- Most faculty are not experts.
- Hard to teach what you don’t know.
- This is where we provide the focused help.



```
┌─[14:05:49]─[devin@ron]─[~/Project_Ghosh/Sample_MP9/assembly]
└──> gre█
```

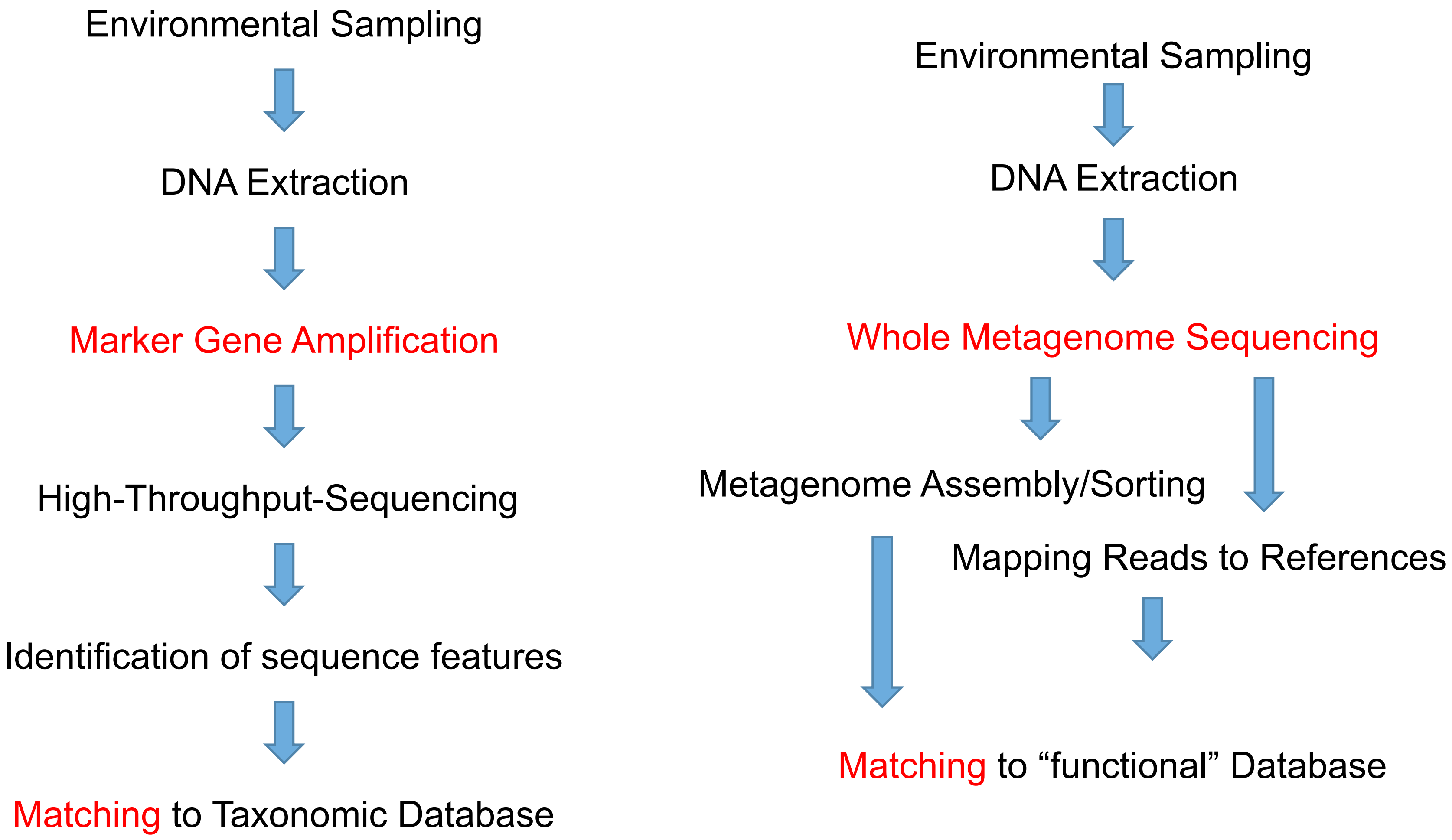
Concepts for environmental genomics

- Metagenomics: The study of genetic information gathered directly from the environment. **AKA: environmental genomics, ecogenomics, or community genomics**
  - Sometimes “Whole metagenome shotgun sequencing”
- Metabarcoding: uses “Universal” PCR primers to identify DNA from a mixture of organisms **AKA: Amplicon or Marker gene sequencing**.
- eDNA: DNA does not have to come directly from the organism.

Major Metagenomic Workflows

Metabarcoding

Whole Metagenome Shotgun



Environmental Sampling

DNA Extraction

Marker Gene Amplification

High-Throughput-Sequencing

Identification of sequence features

Matching to Taxonomic Database

Environmental Sampling

DNA Extraction

Whole Metagenome Sequencing

Metagenome Assembly/Sorting

Mapping Reads to References

Matching to “functional” Database

Every single step introduces biases

Major Workflows

Metabarcoding

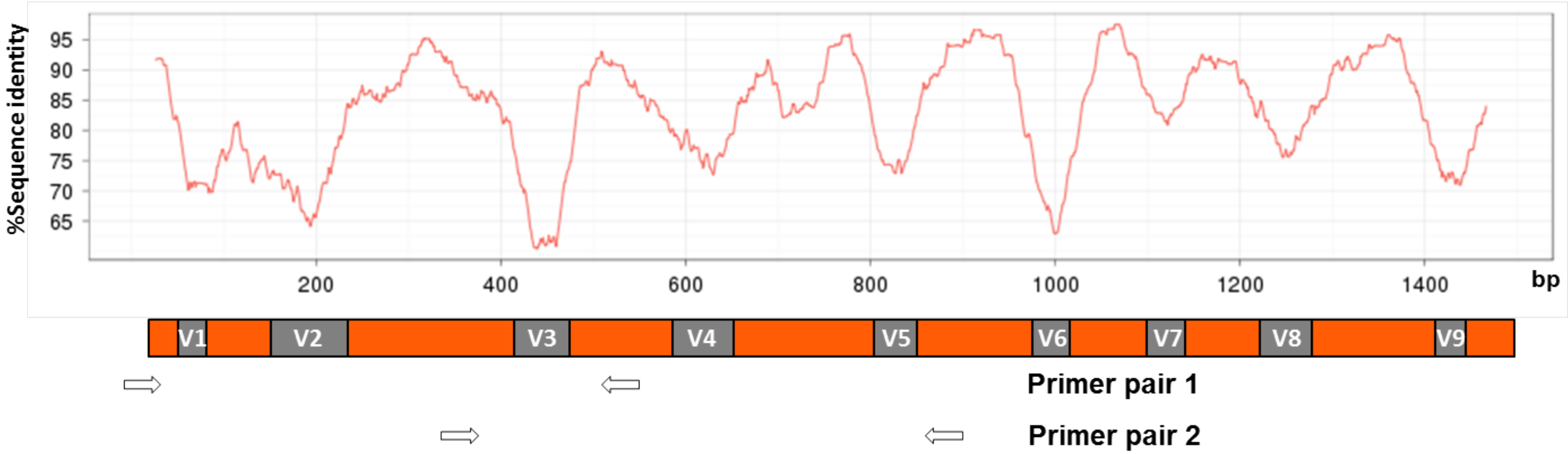
- Positive Aspects
  - Efficient
  - Possible to process 1000 of samples
- Negative Aspects
  - PCR Limitations
  - Limited by databases
  - Not directly related to ecosystem function

Whole Metagenome Shotgun

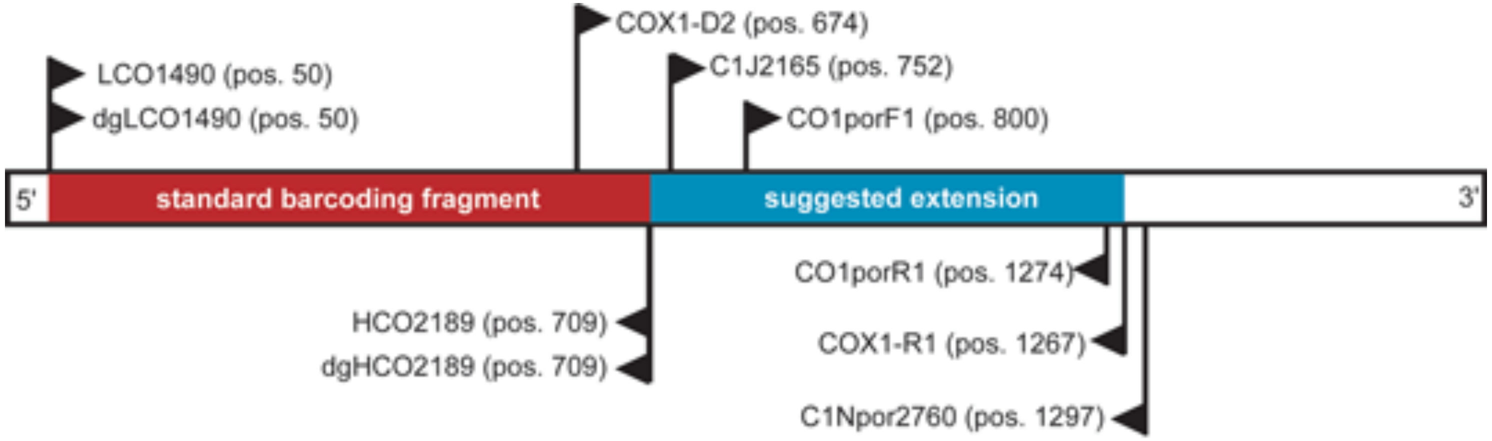
- Positive Aspects
  - Produces “functional” and taxonomic data
  - Not restricted by PCR All “organisms”
- Negative Aspects
  - Inefficient
  - Limited by databases

Marker Genes/Barcodes

Ribosomal RNA



Mitochondrial DNA



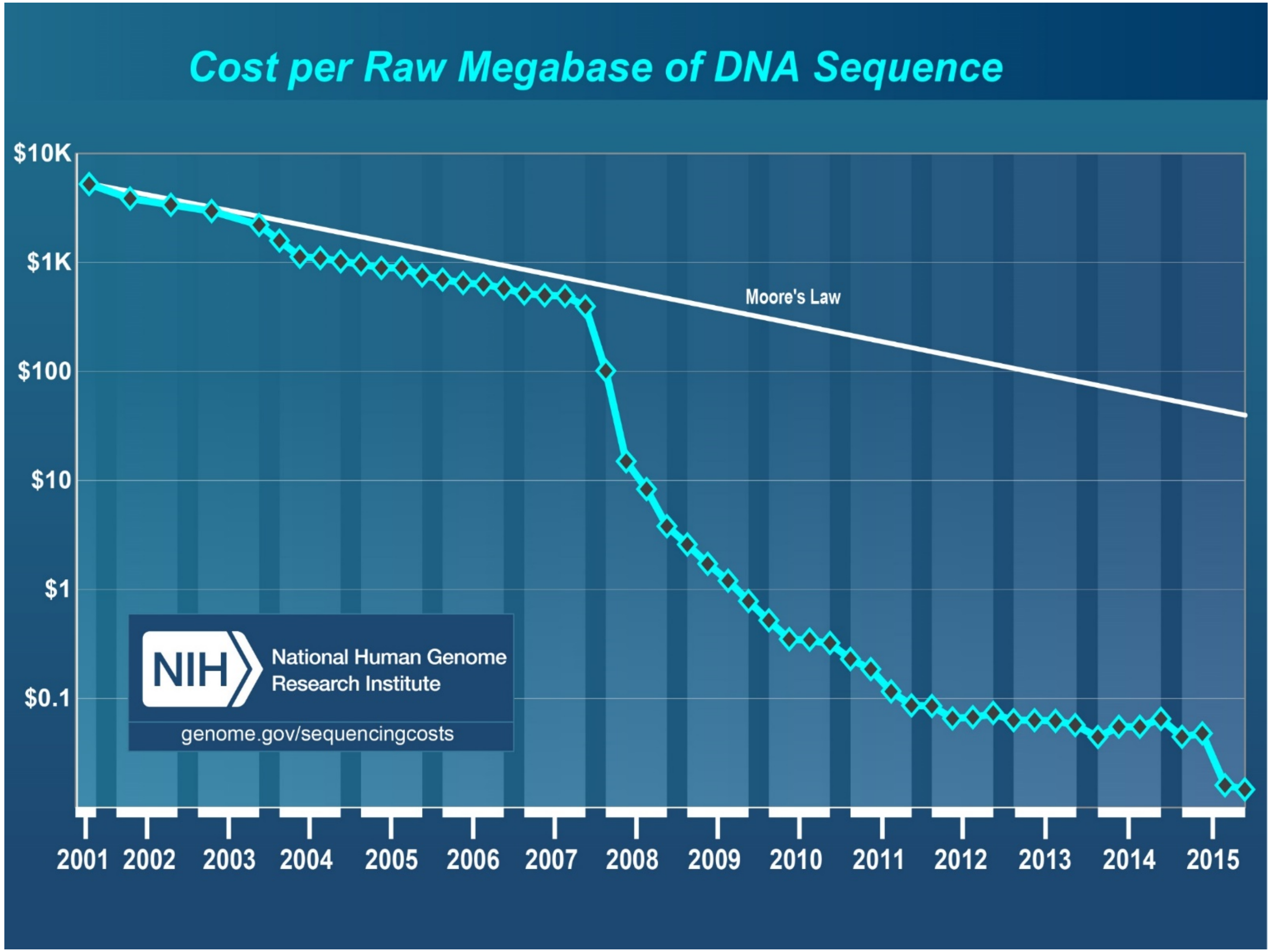
The generation of Genomic Data

Next Generation DNA sequencing technologies

A sea-change in genome-enabled biology

- The last several years have seen the development of fundamentally new sequencing technologies.
  - A process that continues…
- These technologies produce more data, and better data in many fewer steps.
- These changes make genomic analysis a core approach in diverse areas of biology

“Sequencing is now the cheapest part of sequencing” C. Titus Brown



Cost of Meta-barcoding

- Sampling $10 -$10,000 per sample
- DNA Isolation – $5-$10
- PCR $1-$5
- Sequencing $ 1 for 50,000 reads
- Analysis: $150/hour

Your experimental design should keep these relative costs in mind

Next Generation Sequencing Technologies

- The dominant sequencing approach today is Sequencing by Synthesis (SBS) from Illumina. This technology produces the vast majorities of datasets and will be the method used for your sequencing.
- Other interesting and useful approaches include the Single Molecule Real Time (SMRT) Sequencing approach by Pacific Biosciences and Nanopore Sequencing by Oxford Nanopore Technologies. Both of these technologies can produce much longer "reads" or continuous runs of sequence information.

Illumina Sequencing by synthesis

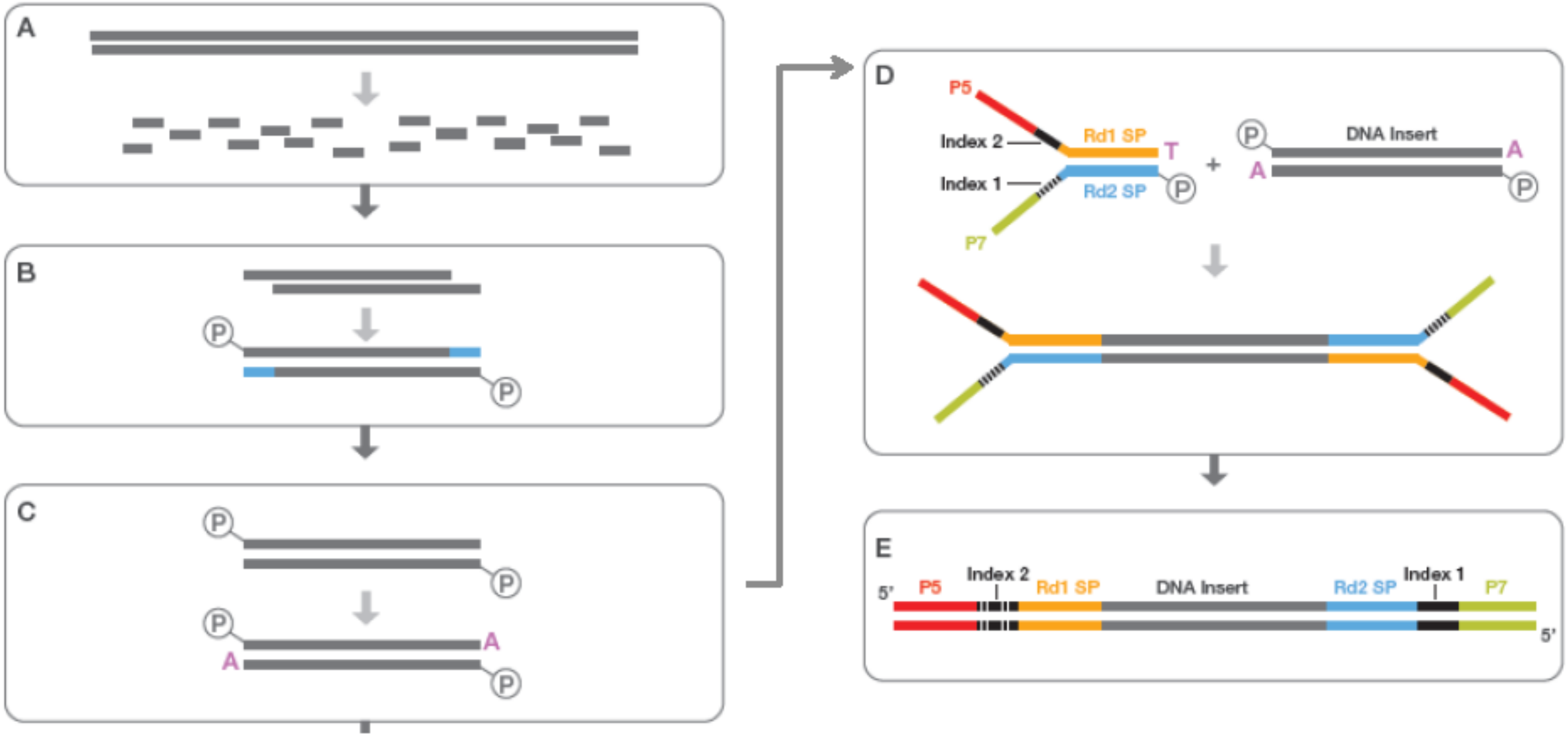
Watch a Video showing the basics of the technology

Serious
https://www.youtube.com/watch?v=womKfikWlxM

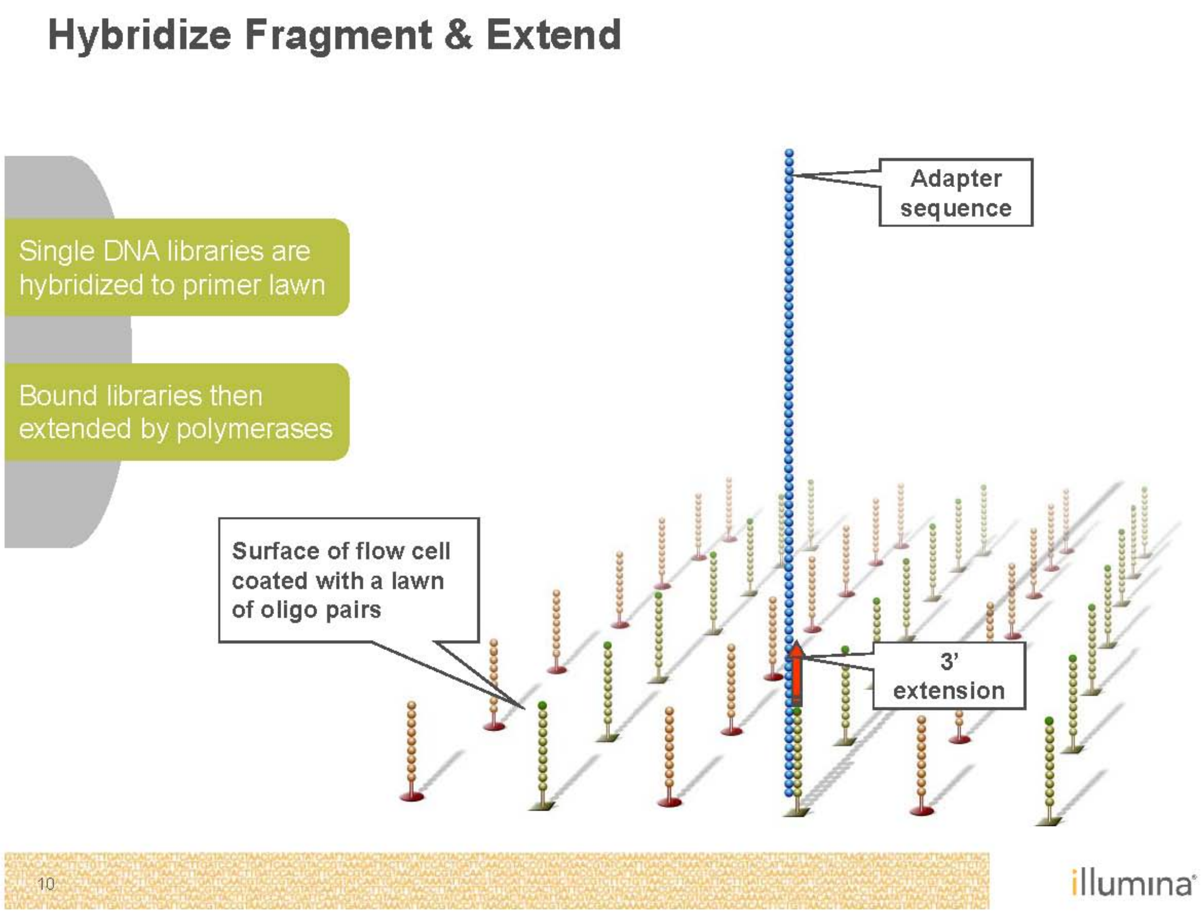
Not so serious
https://www.youtube.com/watch?v=-7GK1HXwCtE

Basic Process

DNA



Sequencing on the inside surface of a flowcell



Hybridize Fragment & Extend

Single DNA libraries are hybridized to primer lawn

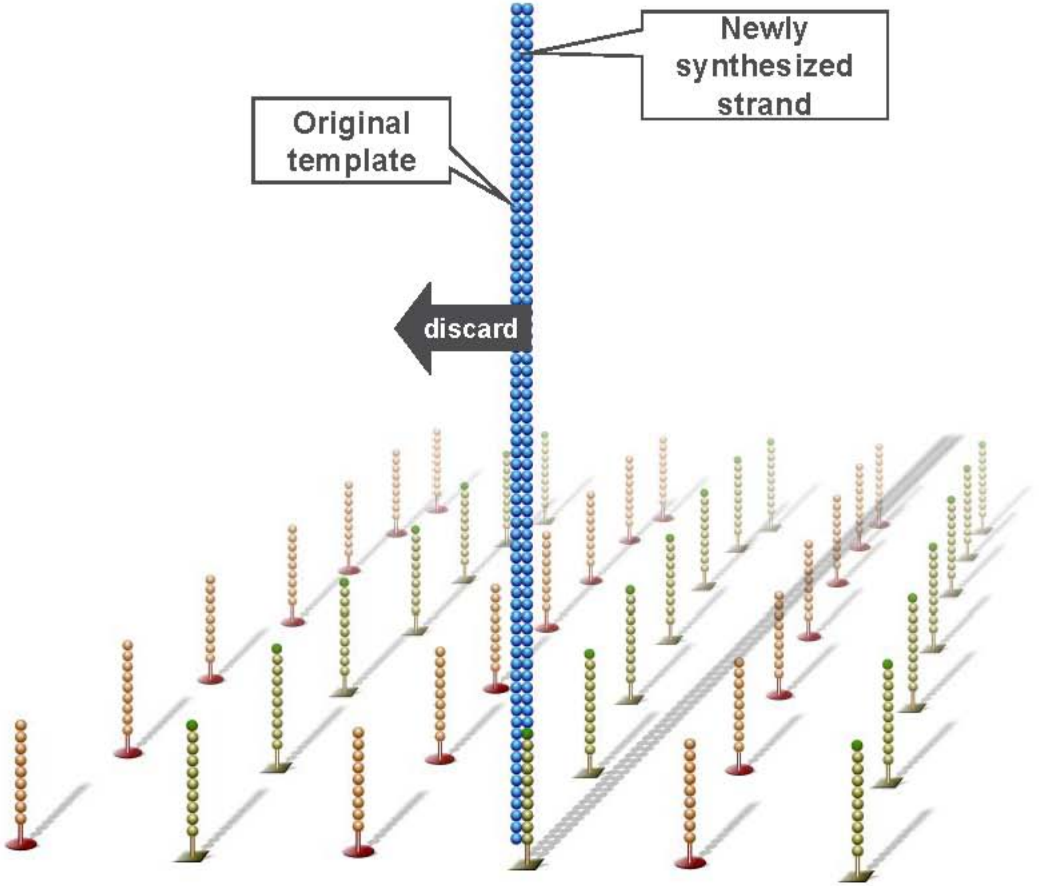
Bound libraries then extended by polymerases

Denature Double-stranded DNA

Double-stranded molecule is denatured

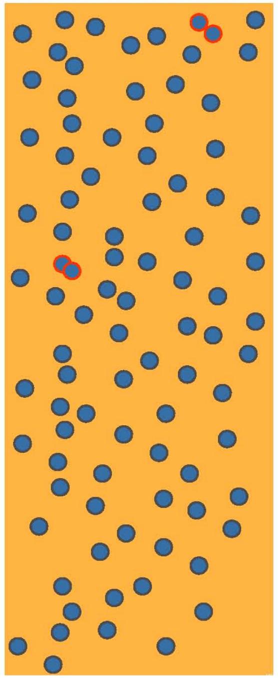
Original template washed away

Newly synthesized strand is covalently attached to flow cell surface

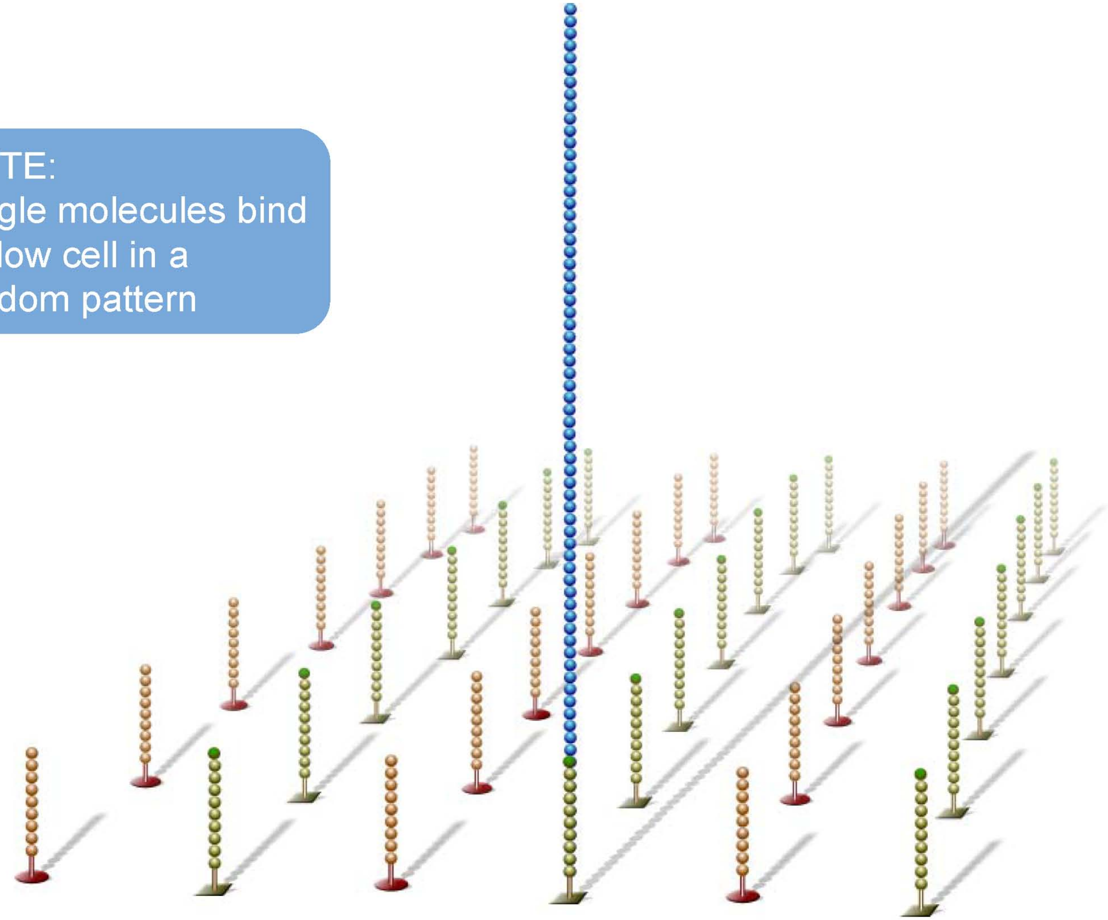


illumina

Hybridize Fragment & Extend

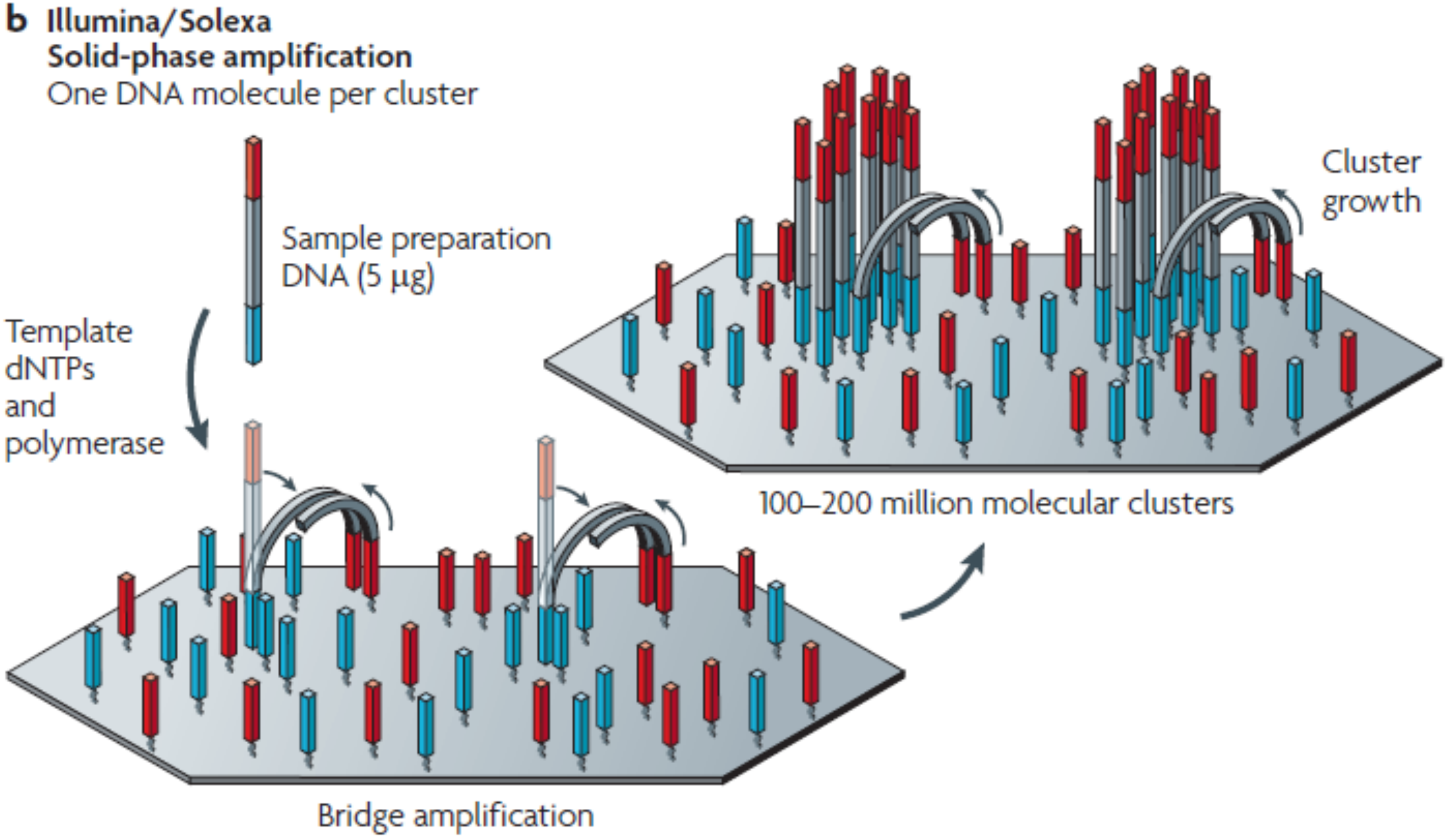


NOTE:
Single molecules bind to flow cell in a random pattern

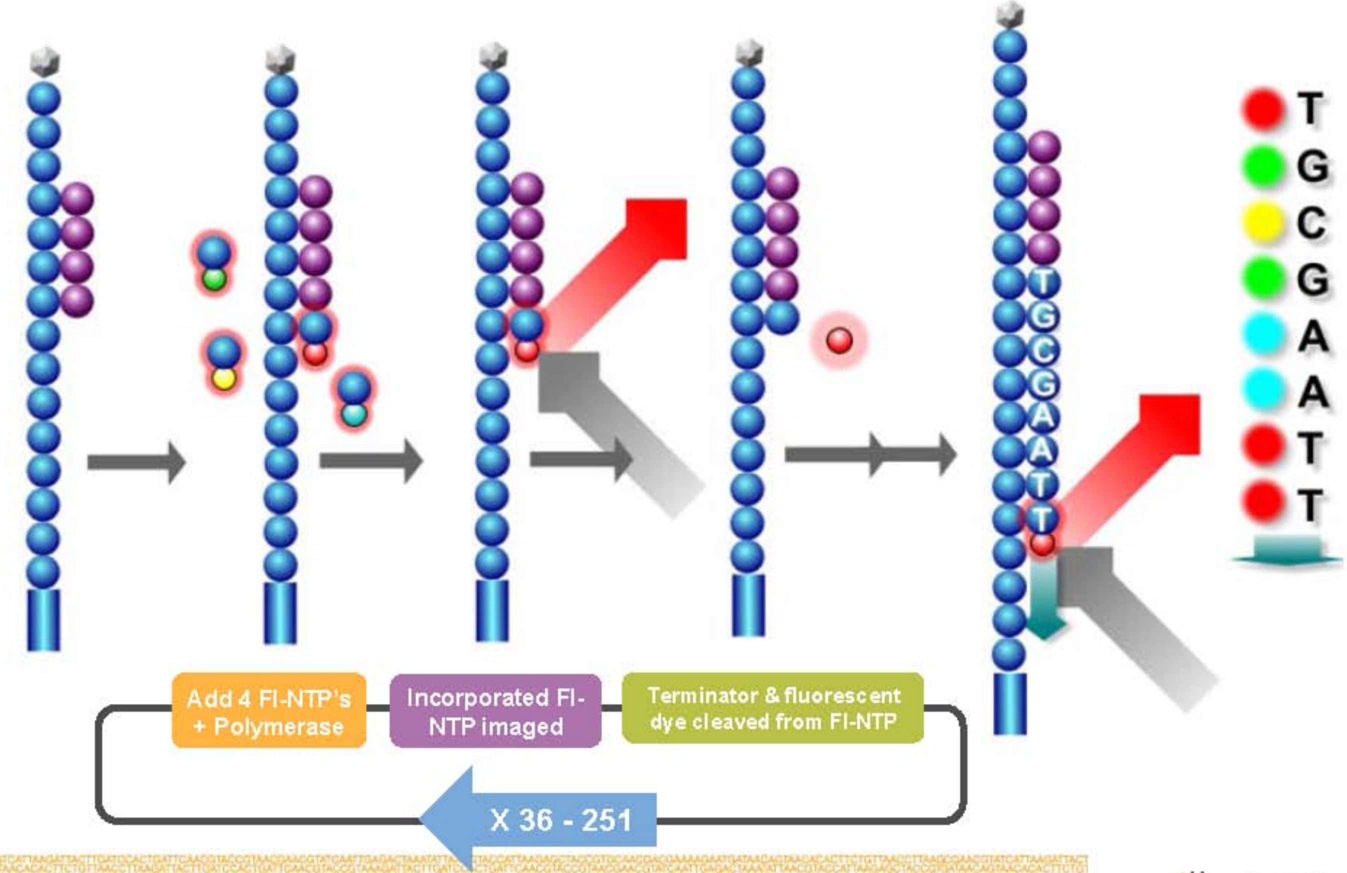


Making sequencing massively parallel

- Solid Phase


b Illumina/Solexa
Solid-phase amplification
One DNA molecule per cluster
Sample preparation DNA (5 μg)
Template dNTPs and polymerase
Bridge amplification
Cluster growth
100–200 million molecular clusters

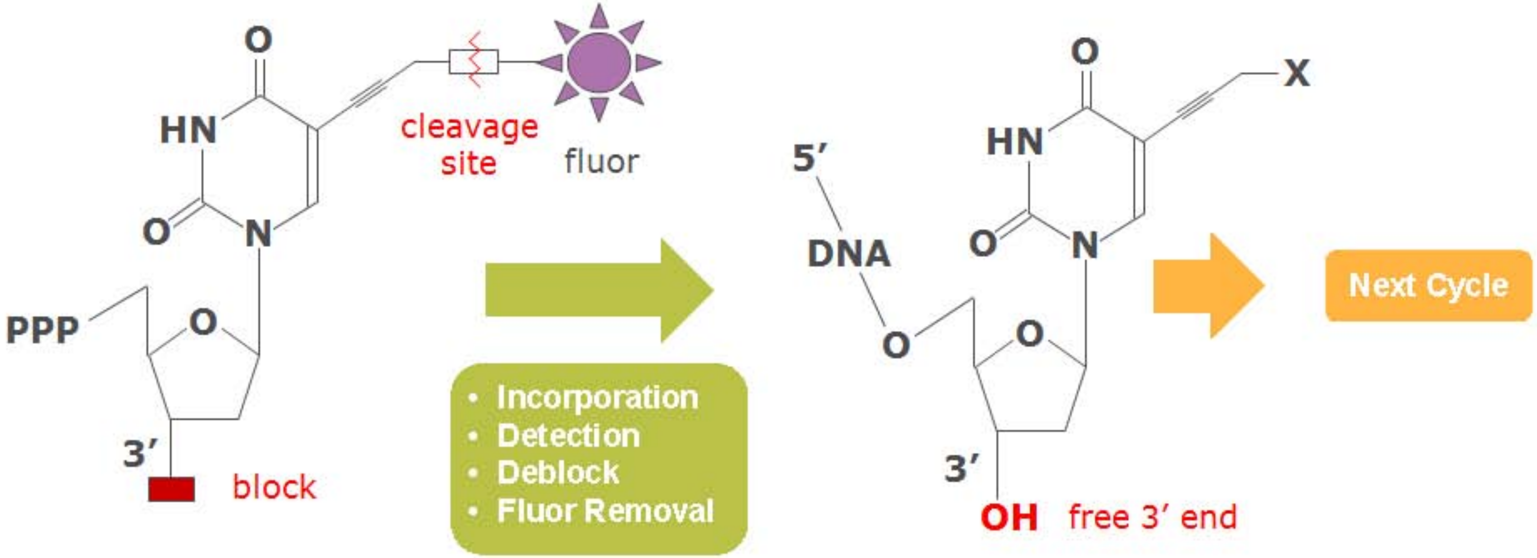
Sequencing By Synthesis


T
G
C
G
A
A
T
T
T
G
C
G
A
A
T
T
Add 4 Fl-NTP's + Polymerase
Incorporated Fl-NTP imaged
Terminator & fluorescent dye cleaved from Fl-NTP
X 36 - 251

illumina

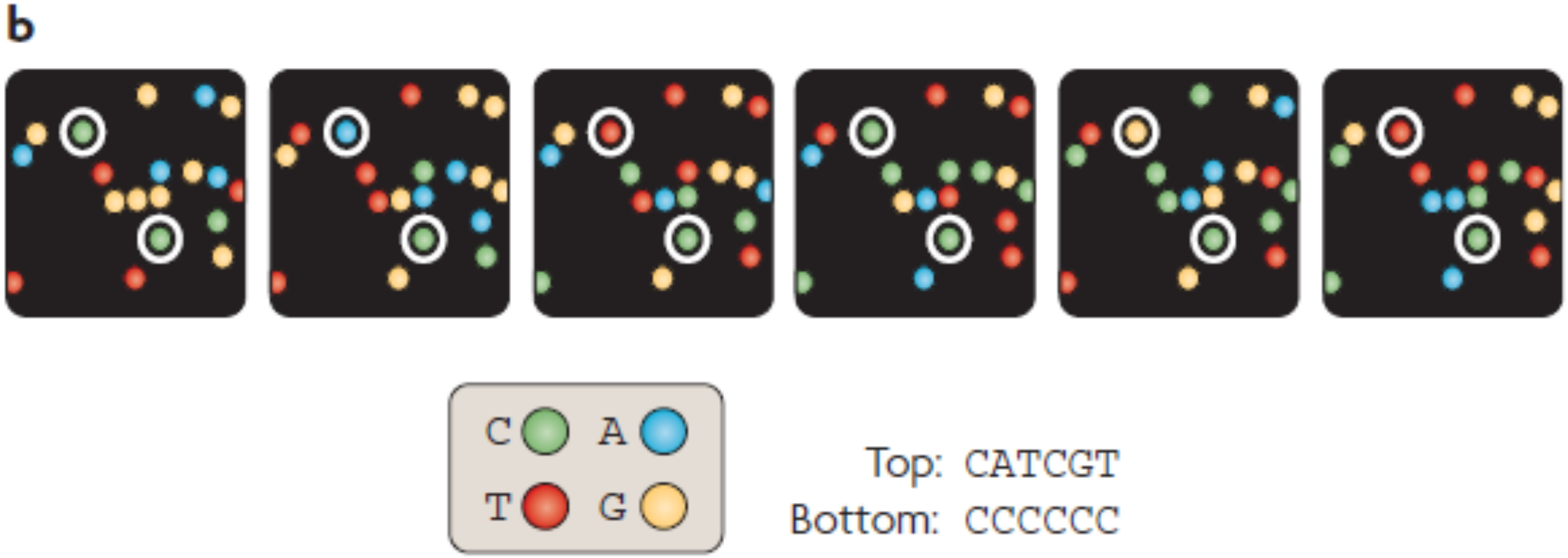
Reversible Terminator Chemistry

- All 4 labeled nucleotides in 1 reaction
- Higher accuracy
- No problems with homopolymer repeats

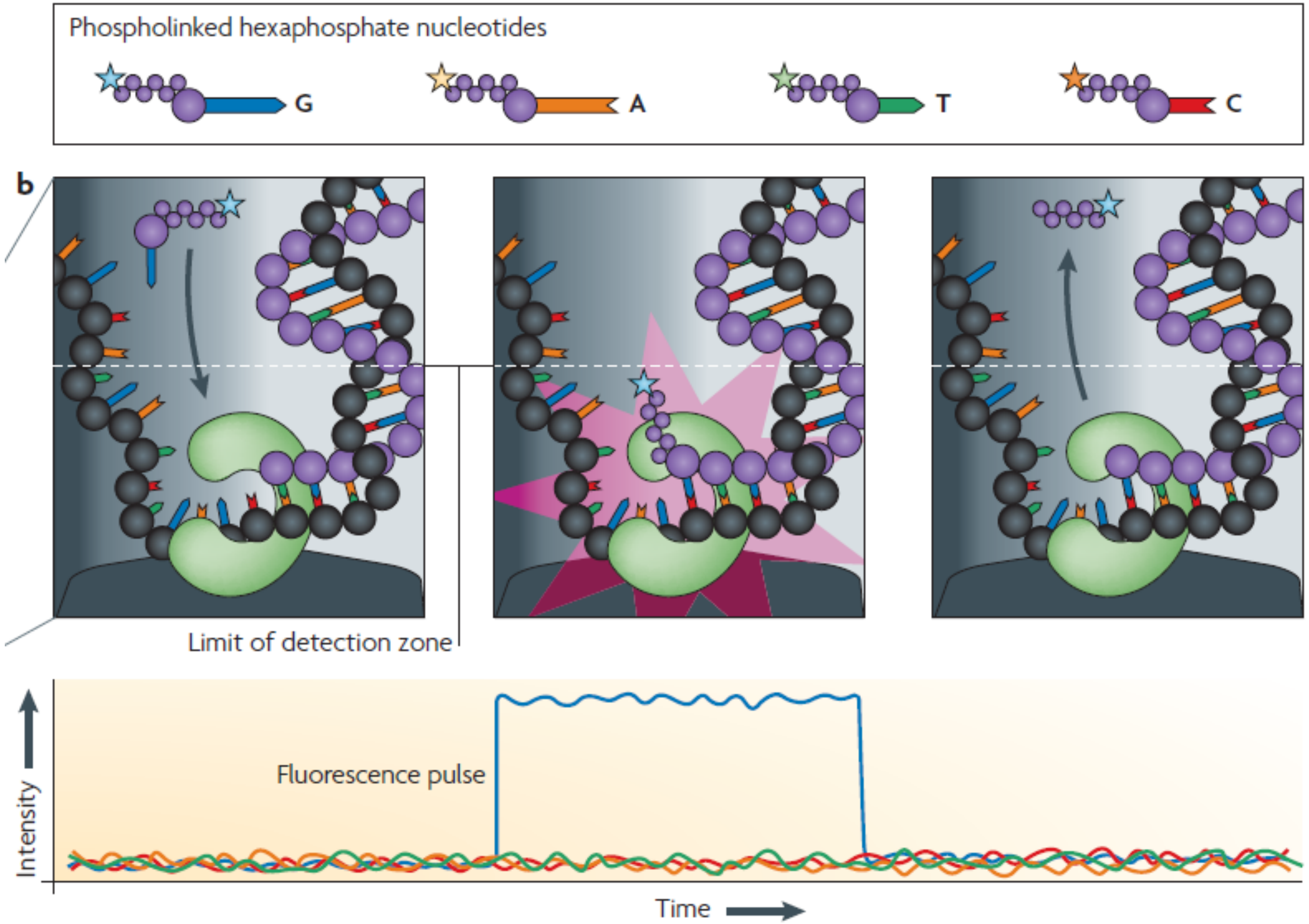


- Incorporation
- Detection
- Deblock
- Fluor Removal

Illumina Stargazing



Pac Bio

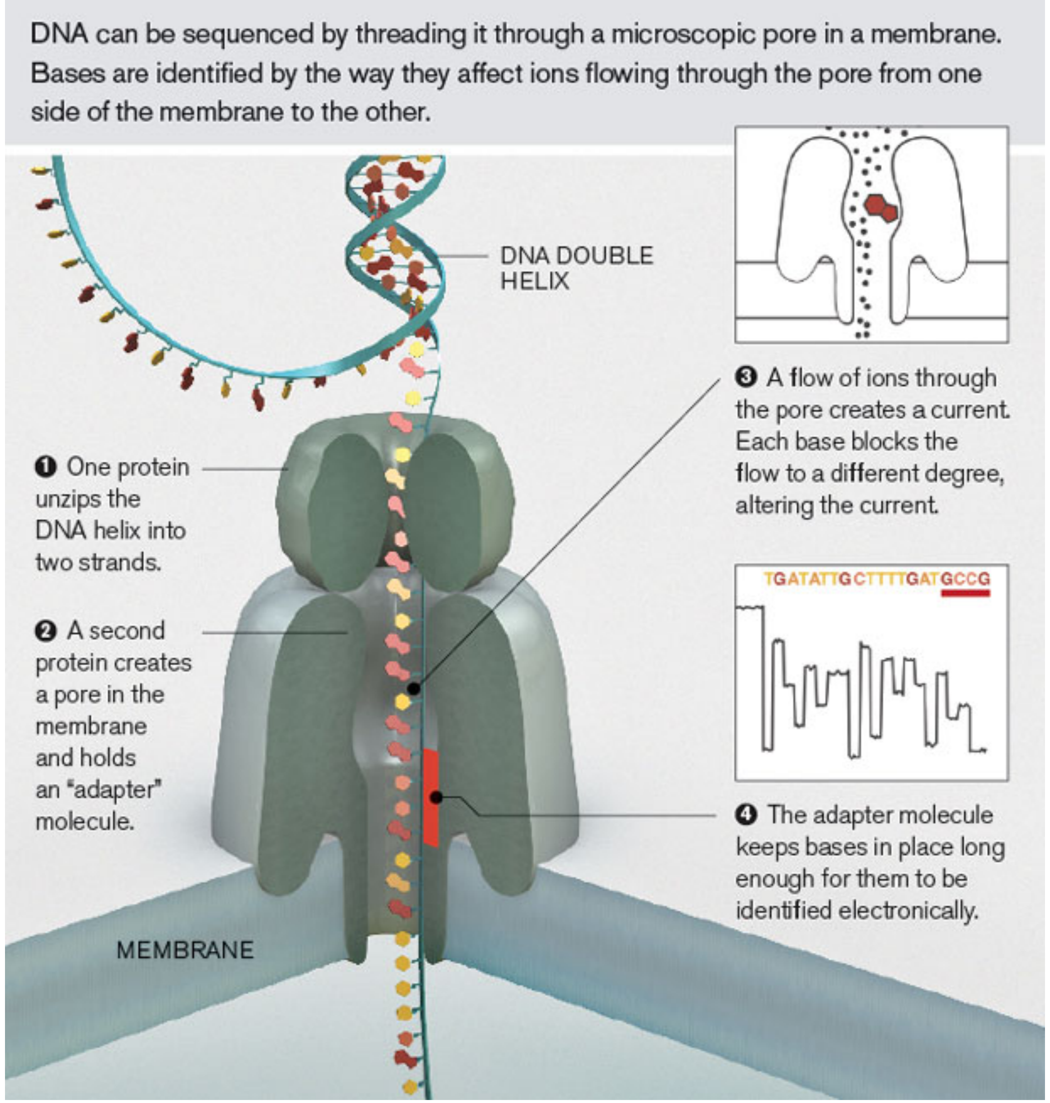


Watch a Video showing the basics of the technology

https://www.youtube.com/watch?v=hBr0TJg-N6U

Nanopore Sequencing

https://www.youtube.com/watch?v=3UHw22hBpAk



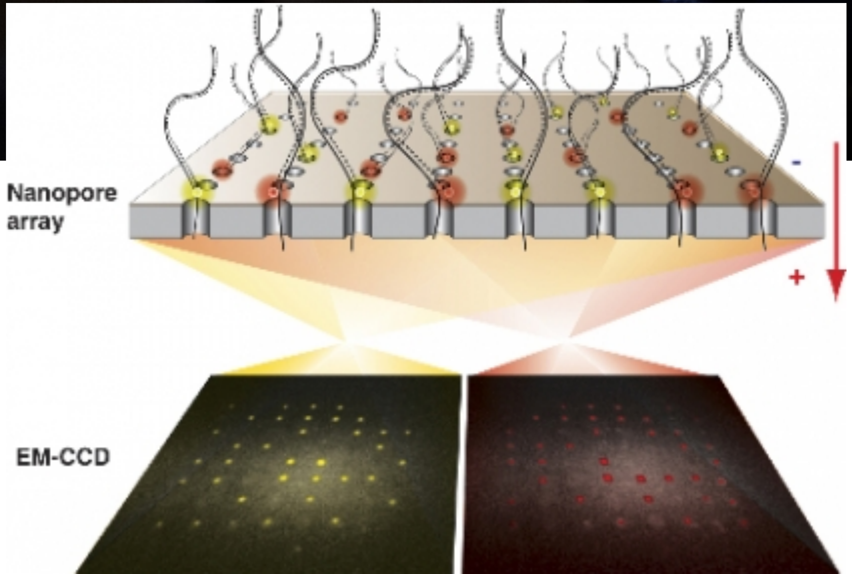
DNA can be sequenced by threading it through a microscopic pore in a membrane. Bases are identified by the way they affect ions flowing through the pore from one side of the membrane to the other.

❶ One protein unzips the DNA helix into two strands.

❷ A second protein creates a pore in the membrane and holds an "adapter" molecule.

❸ A flow of ions through the pore creates a current. Each base blocks the flow to a different degree, altering the current.

❹ The adapter molecule keeps bases in place long enough for them to be identified electronically.



Data structures and conventions

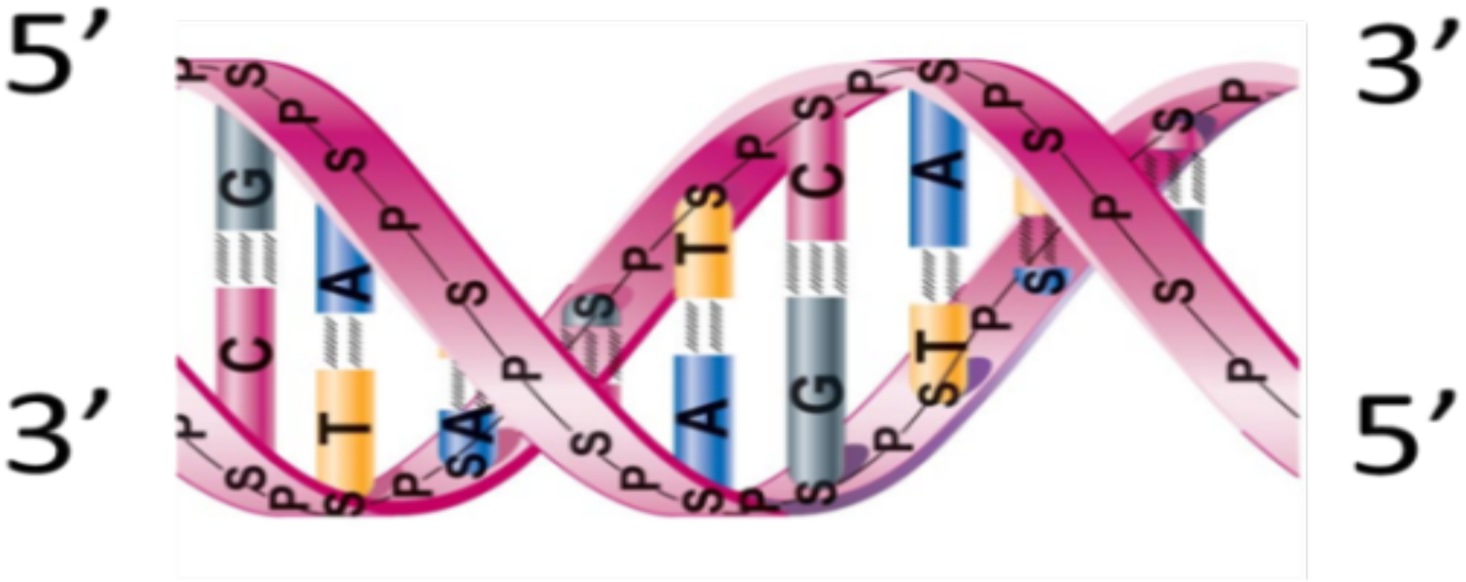
- The interaction of computers and data requires that we follow strict conventions for how we communicate genomic data.
- Bioinformatics relies on the use of common file formats.

Some simple concepts about DNA sequence data

In bioinformatics, DNA or RNA is depicted in a single string from 5’ to 3’ and only one strand shown (saves space). Similarly, protein sequences are always written from the N-terminus to the C- terminus.

“It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material.”

In J.D. Watson and F.H.C. Crick, 'A Structure for Deoxyribose Nucleic Acid,' Letter in *Nature* (25 Apr 1953)



The FASTA File Format

Protein FASTA

```
>gi|129295|sp|P01013|OVAX_CHICK GENE X PROTEIN (OVALBUMIN-RELATED)
QIKDLLVSSSTDLDTTLVLVNAIYFKGMWKTAFNAEDTREMPFHVTKQESKPVQMMCMNNSFNVATLPAE
KMKILELPFASGDLSMLVLLPDEVSDLERIEKTINFEKLTEWTNPNTMEKRRVKVYLPQMKIEEKYNLTS
VLMALGMTDLFIPSANLTGISSAESLKISQAVHGAFMELSEDGIEMAGSTGVIEDIKHSPESEQFRADHP
FLFLIKHNPTNTIVYFGRYWSP
```

Nucleotide FASTA

```
>SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
```

Quality Files with PHRED quality scores are often created in parallel

```
>SRR014849.1 EIXKN4201CFU84 length=93
18 10 5 3 2 1 1 1 1 1 1 1 1 1 1 1 22 37 31 22 16 11 6 1 26 34 30 11 33 26 30 21
33 26 25 36 32 16 36 32 16 36 32 20 6 24 33 25 30 25 2 24 36 32 15 35 31 17
36 32 20 6 25 29 20 30 25 4 32 26 32 2332 26 30 24 33 26 35 31 14 28 27 30 22
28 24 27 17 32 23 28 28
```

Extension	Meaning
.fna	fasta nucleic acid
.ffn	FASTA nucleotide of gene regions
.fasta (.fas)	generic fasta
.faa	fasta amino acid
.fq (.fastq)	FASTQ file

Quality scores and meaning. Quality scores = $-\log_{10}$(probability of error)

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

FASTQ file format

The general format is similar in style to FASTA but has quality score attached in same file and given as ASCII characters to save space. This is the common format for raw sequence data.

@title and optional description
sequence line(s)
+optional repeat of title line
quality line(s)

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGCTTTTTTTGTTTGGAACCGAAAGG
GTTTTGAATTTCAAACCCTTTTCGGTTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#"""""""""""7F@71,’";C?,B;?6B;:EA1EA
1EA5’9B:?:#9EA0D@2EA5’:>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

Sequences are expected to be represented in the standard IUB/IUPAC amino acid and nucleic acid codes.

The nucleic acid codes are:

A	adenosine	C	cytidine	G	guanine
T	thymidine	N	A/G/C/T (any)	U	uridine
K	*G/T (keto)*	*S*	*G/C (strong)*	*Y*	*T/C (pyrimidine)*
M	*A/C (amino)*	*W*	*A/T (weak)*	*R*	*G/A (purine)*
B	*G/T/C*	*D*	*G/A/T*	*H*	*A/C/T*
V	*G/C/A*	*-*	*gap of indeterminate length*		

The amino acid codes are:

A	alanine	P	proline
B	aspartate/asparagine	Q	glutamine
C	cystine	R	arginine
D	aspartate	S	serine
E	glutamate	T	threonine
F	phenylalanine	*U*	*selenocysteine*
G	glycine	V	valine
H	histidine	W	tryptophan
I	isoleucine	Y	tyrosine
K	lysine	Z	glutamate/glutamine
L	leucine	X	any
M	methionine	*	translation stop
N	asparagine	*-*	*gap of indeterminate length*

GFF files describe the position of genes in a FASTA file.

The **general feature format** (**gene-finding format**, or **generic feature format**, **GFF**) is a file format used for describing genes and other features of DNA, RNA and protein sequences. The filename extension is .GFF.

In a GFF file there are 9 columns of information for each feature in a DNA sequence (FASTA file).

Position	Name	
1	sequence	
2	source	
3	feature	
4	start	
5	end	
6	score	
7	strand	
8	frame	
9	attributes.	

Sequence Alignment Map (SAM) Files or BAM for binary version

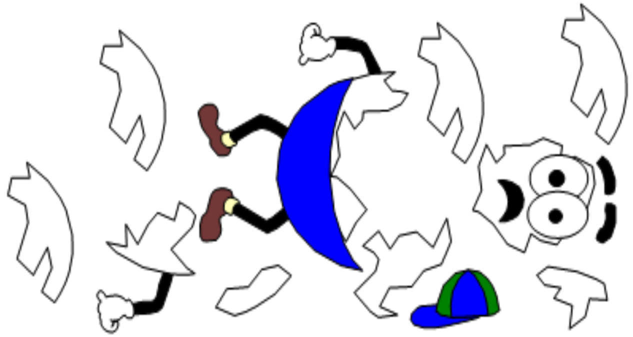
11 mandatory fields in each row: separated by spaces in a text file

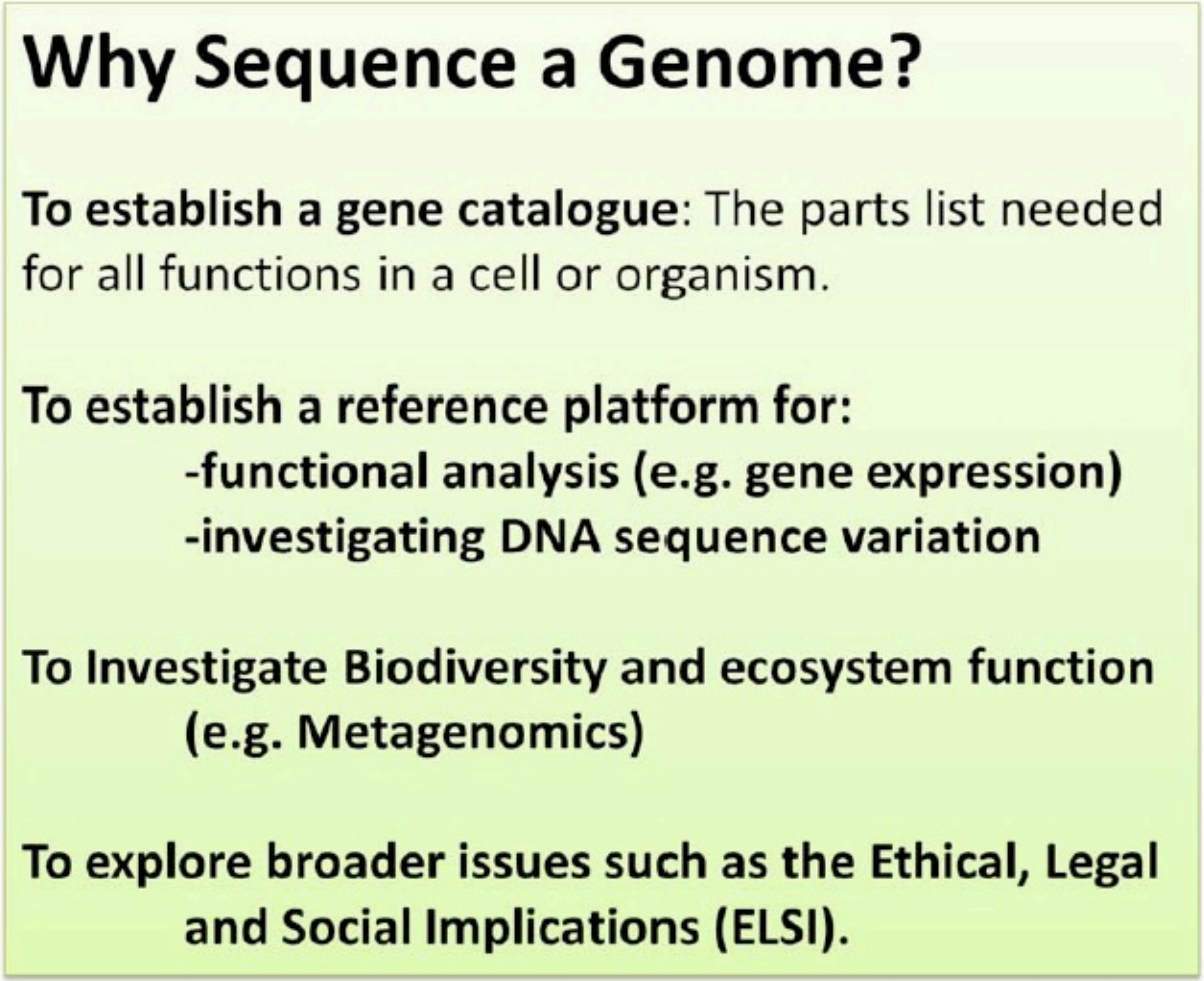
1=Read Name, 2=bitwise flag, 3=Reference Sequence, 4=start position of read in reference, 5= Map Quality, 6= CIGAR string, 7= paired end read reference, 8=position of mate, 9=distance between reads, 10=Read sequence, 11=read quality.

```
1:497:R:-272+13M17D24M 113 chr1 497 37 37M chr15 100338662 0

CGGGTCTGACCTGAGGAGAACTGTGCTCCGCCTTCAG

0;==-==9;>>>>>=>>>>>>>>>>>=>>>>>>>>>>
```

Sequencing and assembling a genome

The process of sequencing genomes typically involves breaking the genome up and then attempting to put Humpty-Dumpty back together again.





Why Sequence a Genome?

To establish a gene catalogue: The parts list needed for all functions in a cell or organism.

To establish a reference platform for:

-functional analysis (e.g. gene expression)

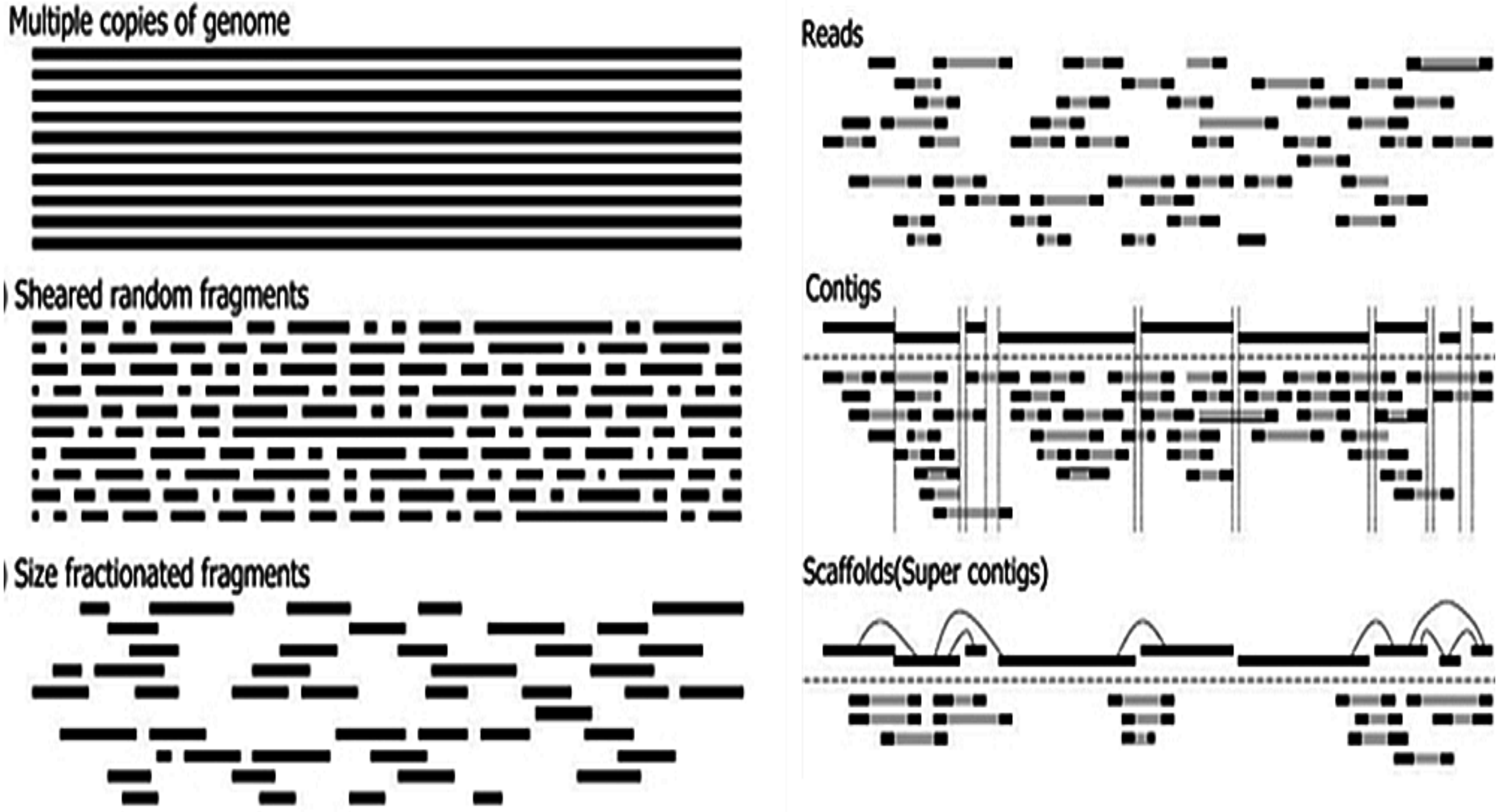
-investigating DNA sequence variation

To Investigate Biodiversity and ecosystem function (e.g. Metagenomics)

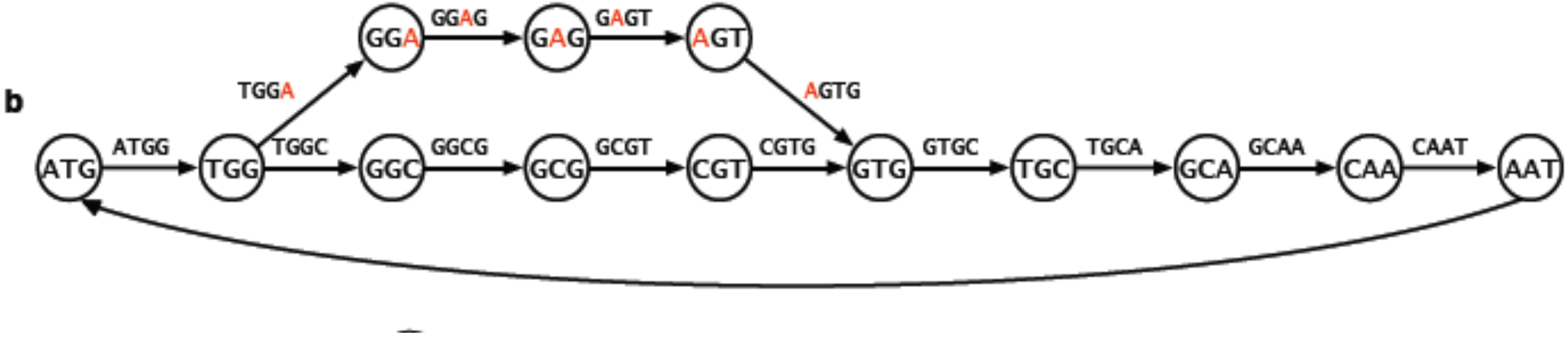
To explore broader issues such as the Ethical, Legal and Social Implications (ELSI).

There are many diverse reasons for sequencing genomes.

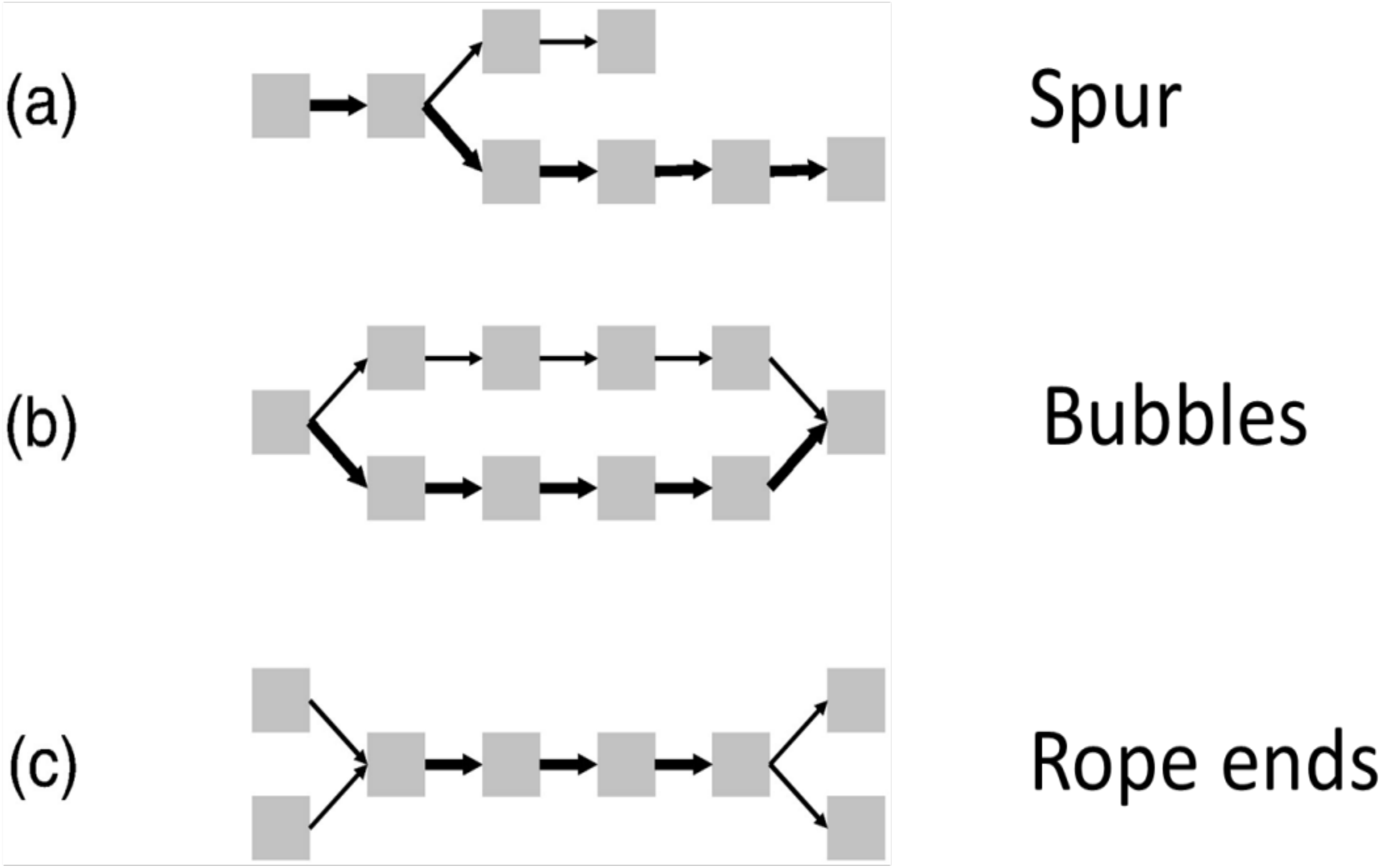
Sequencing a genome


Multiple copies of genome
) Sheared random fragments
) Size fractionated fragments
Reads
Contigs
Scaffolds(Super contigs)

In real data, there will be sequencing errors and polymorphisms. In the figure below, a single base difference results in two paths that diverge and then converge. This could be caused by a sequencing error in the middle of a read or polymorphisms. If this represents heterozygosity, the paths may have equal representation.



In the diagram below, the path complexities include spurs that will result from a sequencing error at the end of a read, bubbles as shown above and “rope ends”. Rope ends depict two different paths that share a common set of k-mers. These are the result of repeats that are greater than the length of a k-mer.



Assembly output and assessing the quality of an assembly:

De novo and assembly produces two main outputs.

- **Contig file** (FASTA or multi FASTA)
- **SAM file:** SAM (Sequence Alignment/Map) format is a generic format for storing large nucleotide sequence alignments.

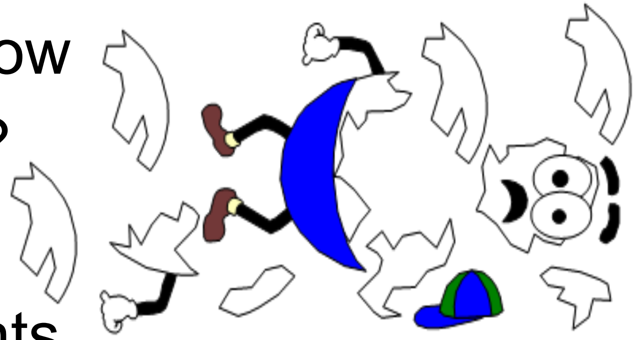
How do we assess the quality of an assembly? There are three basic measures of assembly quality:

1. **N50:** A measure of average contig size. Specifically, ½ of the genome is assembled in contigs of this size or greater.
2. **Depth of coverage:** A measure of how much information is available for each base call.
3. **Completeness of the gene catalogue:** What percentage of the genes are assembled into contigs?

Key challenges for genome assembly:

Intrinsic Challenges:

1. ***Heterozygosity:*** The alleles of a gene are not the same, yet we typically force them into a single consensus sequence.
2. ***Paralogy vs. Alleleism:*** Genes come from other genes by a process of duplication. This results in two or more similar genes in an organism. There are two alleles in a diploid organism that are very similar. How do you tell a duplicated gene from alleles of a gene?
3. ***Sequence complexity:*** Simple sequence repeats (SSR), large-scale repeats like transposable elements (TEs).



Extrinsic Challenges:

1. ***Quality of DNA sequences (sequencing errors):*** Each sequencing technology has specific patterns of error. For example, pyrosequencing typically has high error rates associated with runs of a single nucleotide.
2. ***Length of DNA sequence reads:*** Shorter reads are less likely to be unique or to include many unique K-mers (see below).
3. ***Coverage:*** Depth of Coverage is a random process at best. Consequently some regions of the genome will have low levels of coverage.
4. ***Memory intensive:*** Inherently requires large amounts of RAM for assembly and storage for input and output.
5. ***Software:*** Need for approaches that are flexible, user friendly and powerful.

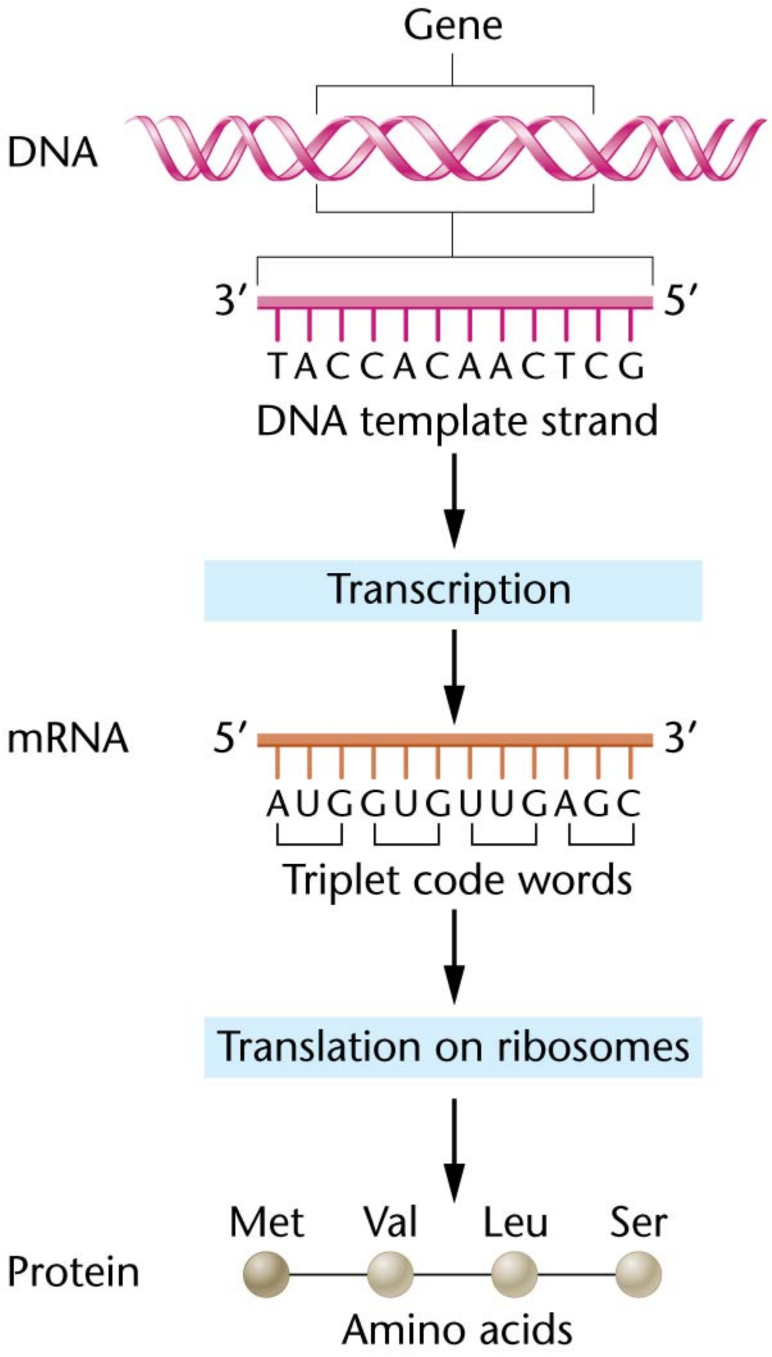
Genome annotation and inferring function

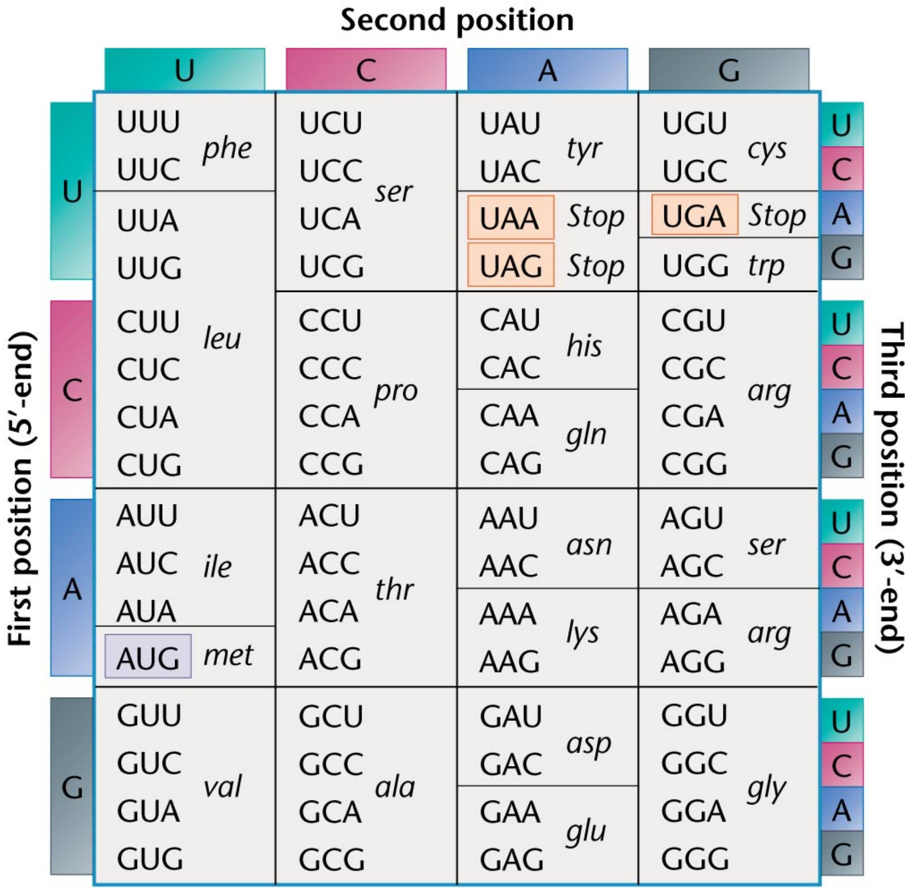
Once we have assembled a genome into one or more large "contigs" how do we "read" the DNA sequences and predict the genes and their functions

Inferring Function from a DNA Sequence

- We use our understanding of cellular processes and evolution to predict the existence and function of genes in DNA sequences.
  - The near universal nature of the genetic code makes it possible to predict what protein sequences can be encoded by any DNA molecule.
  - Evolution allows us to compare genes and their proteins from one species to another.
  - When we have demonstrated the function of a gene in a model organism we often assume it will serve the same or similar role in other species

Most concepts in Bioinformatics rely on core knowledge of Genetics



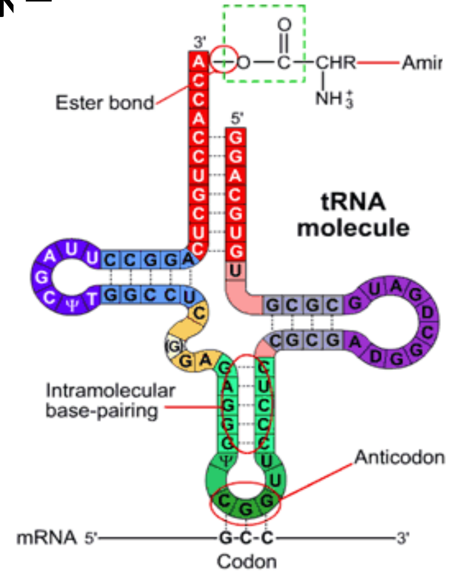


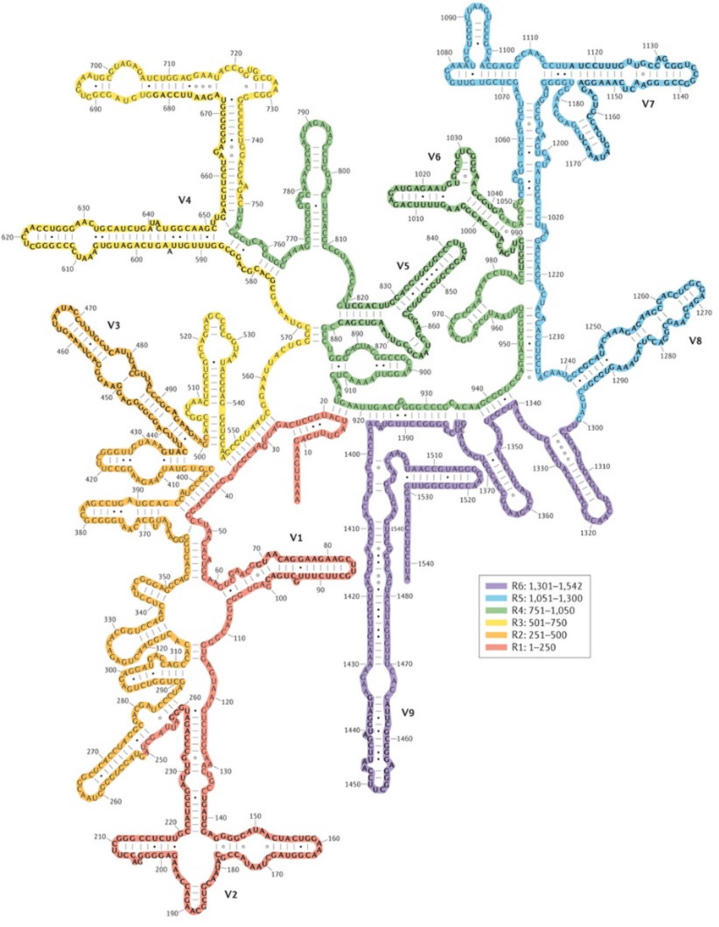
First position (5′-end)	Second position				Third position (3′-end)
	U	C	A	G	
U	UUU *phe*	UCU *ser*	UAU *tyr*	UGU *cys*	U
	UUC	UCC	UAC	UGC	C
	UUA	UCA	UAA *Stop*	UGA *Stop*	A
	UUG	UCG	UAG *Stop*	UGG *trp*	G
C	CUU *leu*	CCU *pro*	CAU *his*	CGU *arg*	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA *gln*	CGA	A
	CUG	CCG	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

▨ Initiation ▨ Termination

Beyond protein coding genes

- Not all genes encode proteins
- How would you find the genes for transfer RNAs and ribosomal RNAs in a DNA molecule?
    - You can look for similar sequences identified in related organisms
    - You can consider their special features lik~ secondary structures





Nature Reviews | Microbiology