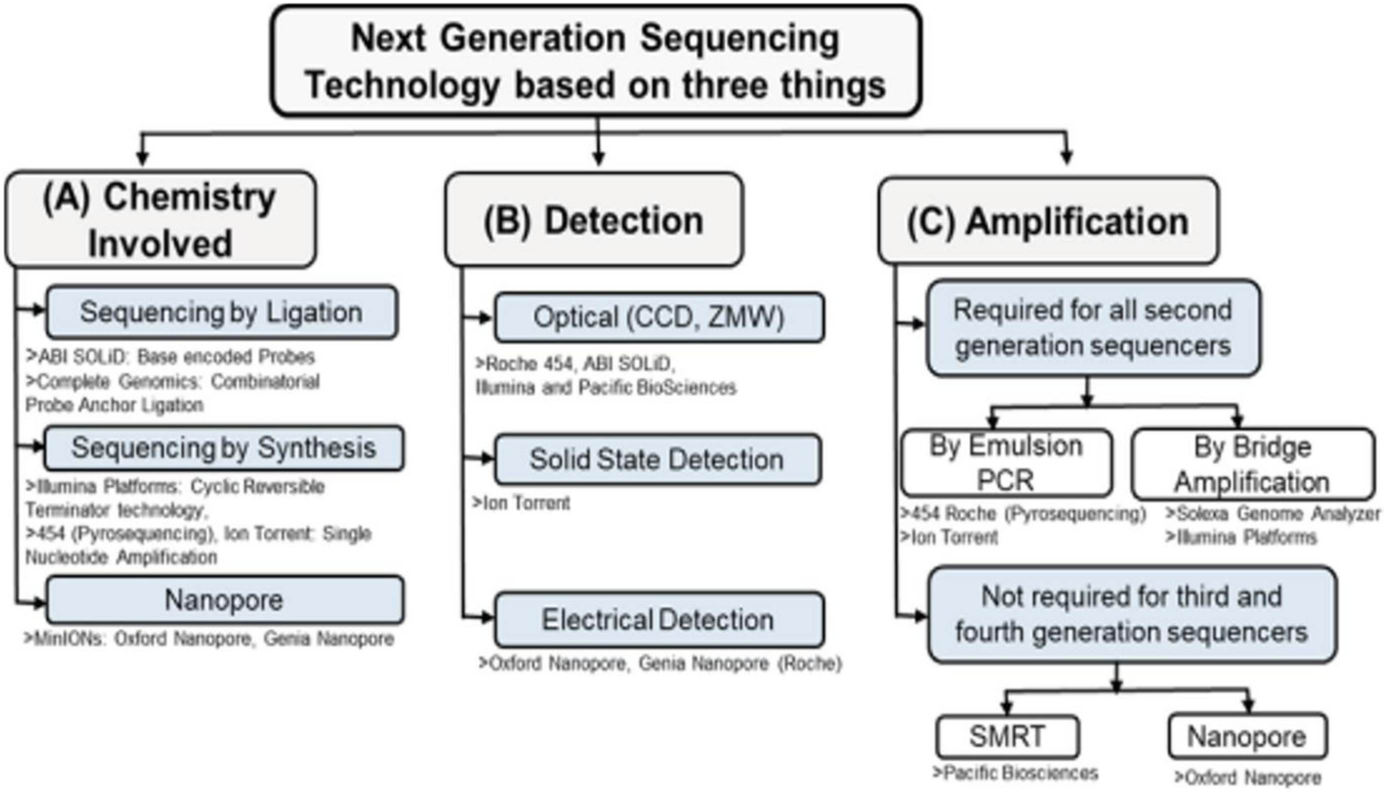
Genomics and Proteomics [22ZOOME31]

Next Generation Sequencing

Introduction

- Two major paradigms in next-generation sequencing (NGS) technology:
- short-read sequencing and long-read sequencing.
- Short-read sequencing approaches provide lower-cost, higher-accuracy data that are useful for population-level research and clinical variant discovery.
- By contrast, long-read approaches provide read lengths that are well suited for de novo genome assembly applications and full-length isoform sequencing.

- **Sequencing by Synthesis** (SBS) A sequencing approach that involves multiple parallel microsequencing addition events occurring on a surface, where data from each round is detected by imaging.
- SNP - Single-nucleotide polymorphism.



Next Generation Sequencing Technology based on three things

(A) Chemistry Involved

Sequencing by Ligation

>ABI SOLiD: Base encoded Probes
>Complete Genomics: Combinatorial Probe Anchor Ligation

Sequencing by Synthesis

>Illumina Platforms: Cyclic Reversible Terminator technology,
>454 (Pyrosequencing), Ion Torrent: Single Nucleotide Amplification

Nanopore

>MinIONs: Oxford Nanopore, Genia Nanopore

(B) Detection

Optical (CCD, ZMW)

>Roche 454, ABI SOLiD, Illumina and Pacific BioSciences

Solid State Detection

>Ion Torrent

Electrical Detection

>Oxford Nanopore, Genia Nanopore (Roche)

(C) Amplification

Required for all second generation sequencers

By Emulsion PCR

>454 Roche (Pyrosequencing)
>Ion Torrent

By Bridge Amplification

>Solexa Genome Analyzer
>Illumina Platforms

Not required for third and fourth generation sequencers

SMRT

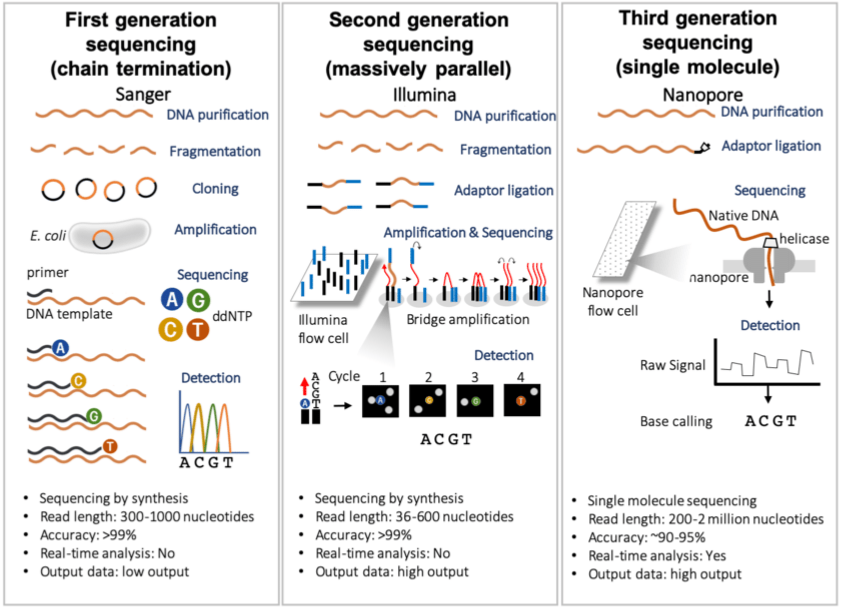
>Pacific Biosciences

Nanopore

>Oxford Nanopore

Genome sequencing

- Genome sequencing is an important step toward correlating genotypes with phenotypic characters.
- Sequencing technologies are important in many fields in the life sciences, including functional genomics, transcriptomics, oncology, evolutionary biology, forensic sciences, and many more.
- First generation sequencing: sequencing by synthesis (Sanger sequencing) &
  sequencing by cleavage (Maxam-Gilbert sequencing)
- Sanger sequencing - completion of various genome sequences (including human) and provided the foundation for development of other sequencing technologies
- Next generation sequencing" (NGS), and are further classified into second and third generation technologies.
- Although NGS methods have many advantages in terms of **speed, cost, and parallelism, the accuracy and read length of Sanger sequencing** is still superior and has confined the use of NGS mainly to resequencing genomes.



- Sequencing by synthesis
- Read length: 300-1000 nucleotides
- Accuracy: >99%
- Real-time analysis: No
- Output data: low output

- Sequencing by synthesis
- Read length: 36-600 nucleotides
- Accuracy: >99%
- Real-time analysis: No
- Output data: high output

- Single molecule sequencing
- Read length: 200-2 million nucleotides
- Accuracy: ~90-95%
- Real-time analysis: Yes
- Output data: high output

General Procedure for Genome Sequencing

1st step

- any genome sequencing project is generation of an enormous amount of sequencing data
- The raw data generated by sequencing is processed to convert the chromatograms into the quality values
- procedure is known as **“base calling” or fragment readout.**

Optional step

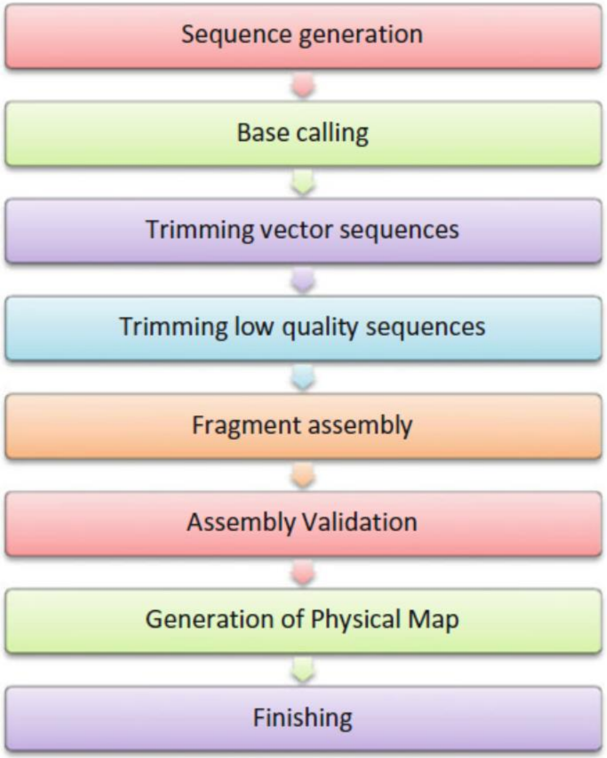
- The next step is optional as only those methods that involve library construction will **utilize trimming of vector sequences**.
- It is an important step as some parts of the vector are also sequenced by universal primers, along with the insert region.
- Further screening of good quality and bad quality regions is performed to ensure a more accurate sequence assembly.
- The processed data is then **assembled** by searching for fragment overlaps to form “contigs.”

Contigs and scaffolds

- because of repeat regions, the fragments may be wrongly aligned, leading to misassembly.
- Hence, assembly validation is done manually as well as by softwares to locate the correct position of reads.
- The contigs are then oriented one after the other with the help of mate-paired information to form ordered chains known as “scaffolds”

Finishing

- This step is followed by the final and most crucial step of genome sequencing:
- **Finishing.** It includes gap filling (also known as “minimum tiling path”) and reassessing the assembly



Sequence generation

Base calling

Trimming vector sequences

Trimming low quality sequences

Fragment assembly

Assembly Validation

Generation of Physical Map

Finishing

Next Generation Sequencing (NGS)

- Most commonly used platforms
- 454 GS Flx (Roche)
- HiSeq (Illumina)
- SoLID (Sequencing by Oligonucleotide Ligation and Detection) [Thermo Scientific] - commercially discontinued
- Ion Torrent (Thermo Scientific) – New version was launched -2015
- Pac Bio (Pacific Bioscience)

Features of sequencing platforms

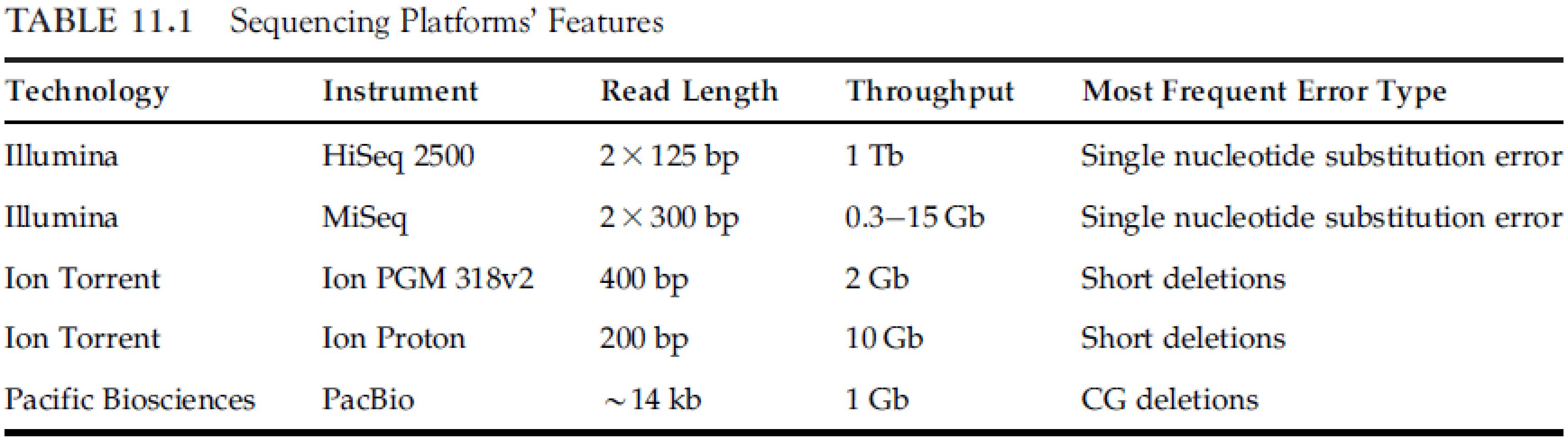


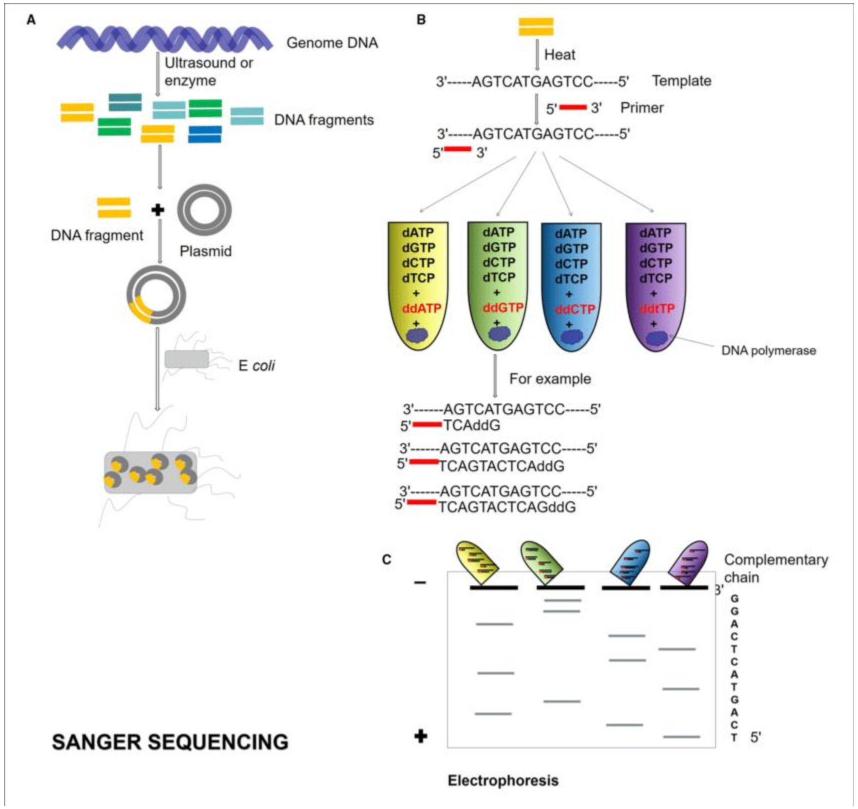
TABLE 11.1 Sequencing Platforms' Features

Technology	Instrument	Read Length	Throughput	Most Frequent Error Type
Illumina	HiSeq 2500	2 × 125 bp	1 Tb	Single nucleotide substitution error
Illumina	MiSeq	2 × 300 bp	0.3–15 Gb	Single nucleotide substitution error
Ion Torrent	Ion PGM 318v2	400 bp	2 Gb	Short deletions
Ion Torrent	Ion Proton	200 bp	10 Gb	Short deletions
Pacific Biosciences	PacBio	~14 kb	1 Gb	CG deletions

differences regarding their chemistry, Read length, and throughput

Sanger capillary sequencing

- The ‘Sanger’ method relies on four separate polymerization reactions performed using **tritium radiolabeled primers**, where each reaction is supplied with small amounts of one chain-terminating 2,3-dideoxynucleoside triphosphate (ddNTP) to produce fragments of different lengths
- When the DNA polymerase incorporates a ddNTP at the 3′-end of the growing DNA strand, it lacks a 3′-hydroxyl group and chain elongation is terminated

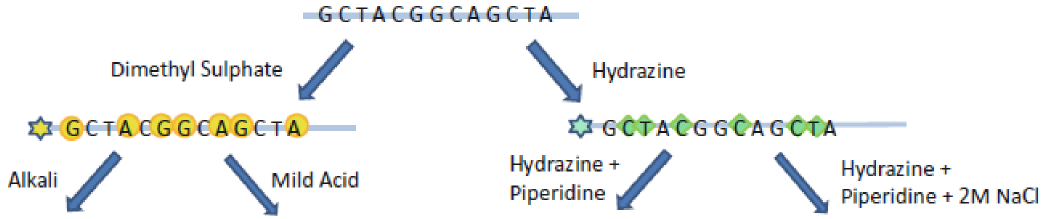

A
Genome DNA
Ultrasound or enzyme
DNA fragments
DNA fragment
+
Plasmid
E coli
SANGER SEQUENCING
B
Heat
3'-----AGTCATGAGTCC-----5'
Template
5' 3'
Primer
3'-----AGTCATGAGTCC-----5'
5' 3'
dATP dGTP dCTP dTCP + ddATP +
dATP dGTP dCTP dTCP + ddGTP +
dATP dGTP dCTP dTCP + ddCTP +
dATP dGTP dCTP dTCP + ddtTP +
DNA polymerase
For example
3'------AGTCATGAGTCC-----5'
5' TCAddG
3'------AGTCATGAGTCC-----5'
5' TCAGTACTCAddG
3'------AGTCATGAGTCC-----5'
5' TCAGTACTCAGddG
C
–
+
Complementary chain
3'
G G A C T C A T G A C T
5'
Electrophoresis

Maxam and Gilbert

- The same year, Maxam and Gilbert proposed an alternative, purely chemical, approach
- While extremely popular for years, with the improvement of the chain-termination method, this approach fell out of favor due to its technical complexity, use of hazardous chemicals, and difficulties to be scaled up.

First Generation Technologies

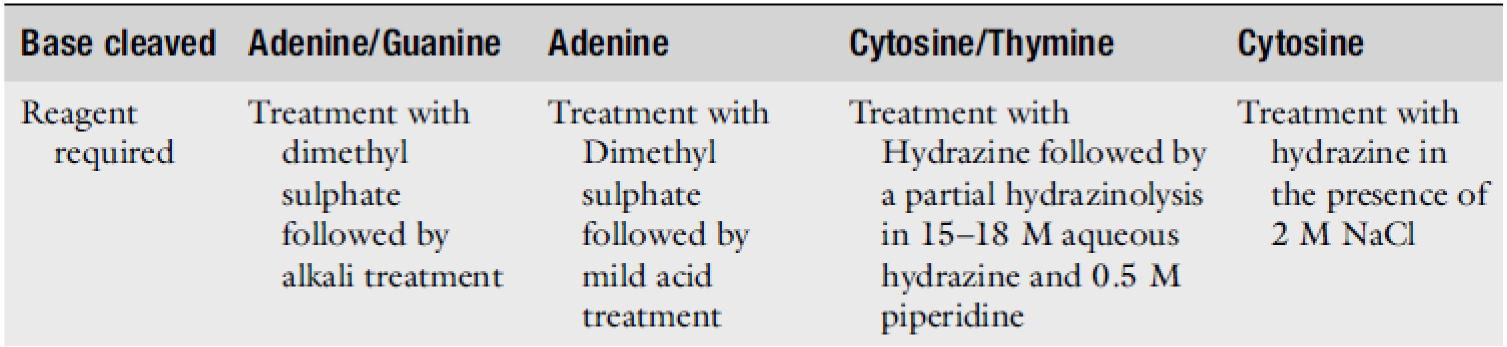
- The initial, first generation sequencing technologies were the sequencing-by-cleavage method established by Maxam-Gilbert and sequencing-by-synthesis developed by Sanger
- This technique first appeared in 1977 and is also known as the "chemical-degradation" method.
- The chemical reagents act on specific bases of existing DNA molecules and subsequent cleavage occurs
- In this technique, dsDNA is labeled with radiolabeled phosphorus at the 5′ end or 3′ end.
- The next step is to obtain ssDNA. This can be done by restriction digestion leading to sticky ends or denaturation at 90 C in the presence of DMSO, resulting in ssDNA.



✡			
		✡ G	
		✡ GC	✡ GC
✡ GCT	✡ GCT		
		✡ GCTA	✡ GCTA
✡ GCTAC			
✡ GCTACG			
		✡ GCTACGG	✡ GCTACGG
✡ GCTACGGC	✡ GCTACGGC		
✡ GCTACGGCA		✡ GCTACGGCAG	
		✡ GCTACGGCAGC	✡ GCTACGGCAGC
✡ GCTACGGCAGCT	✡ GCTACGGCAGCT		

- Only one strand is purified and divided into four samples.
- Each sample is treated with a different chemical reagent
- all four reaction products when run separately on 20 % polyacrylamide gel containing 7 M urea, reveal the point of breakage and the pattern of bands can be read directly to determine the sequence.
- Use of hazardous chemicals and incomplete reactions make this method unsuitable for large-scale DNA sequencing.

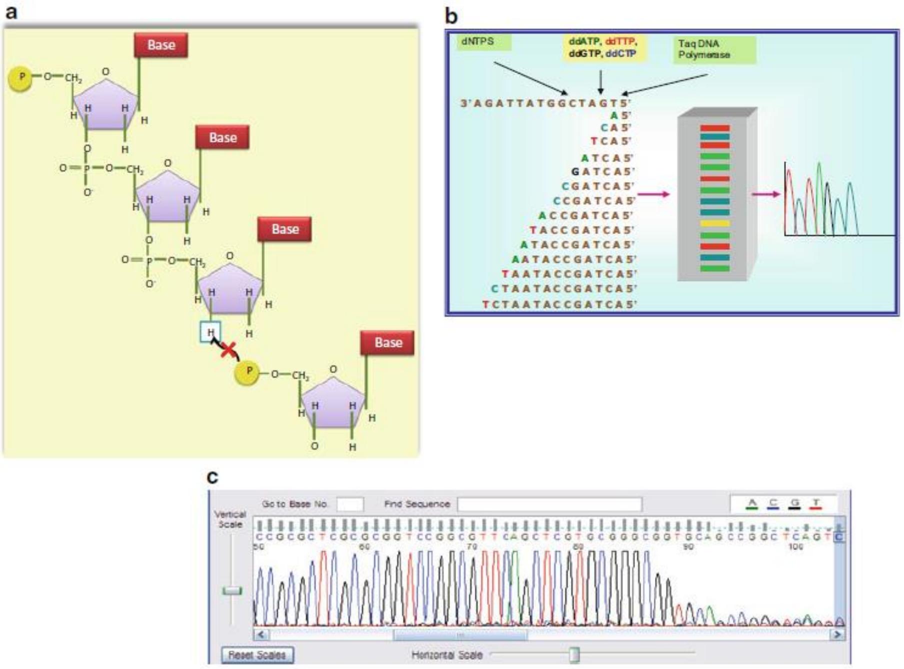
Reagents used in Maxam-Gilbert method for sequencing



Base cleaved	Adenine/Guanine	Adenine	Cytosine/Thymine	Cytosine
Reagent required	Treatment with dimethyl sulphate followed by alkali treatment	Treatment with Dimethyl sulphate followed by mild acid treatment	Treatment with Hydrazine followed by a partial hydrazinolysis in 15–18 M aqueous hydrazine and 0.5 M piperidine	Treatment with hydrazine in the presence of 2 M NaCl

Chain Termination Sequencing or “dideoxy sequencing

- This technique is also known as **chain termination sequencing or “dideoxy sequencing.”**
- Sanger sequencing has played a crucial role in understanding the **genetic landscape of the human genome**.
- It was developed by Frederick Sanger in 1975, but was commercialized in 1977.
- The technique is based on the principal usage of dideoxy ribonucleoside triphosphates that lack 3’hydroxyl group



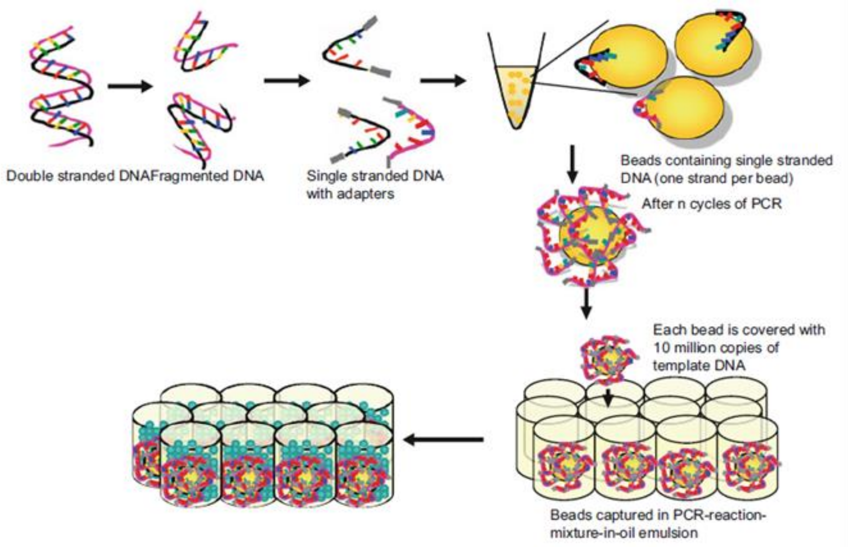
Dideoxy chain termination method.

(a) DNA synthesis by addition of dNTP (that has free 3′-OH) in the synthesizing strand. But once a ddNTP is incorporated, the strand synthesis stops as no 3′-OH is available to form a phosphodiester bond with the next dNTP.

(b) formation of a series of chains by incorporation of ddNTPs which can by separated by capillary electrophoresis.

(c) Electropherogram of a DNA sequence

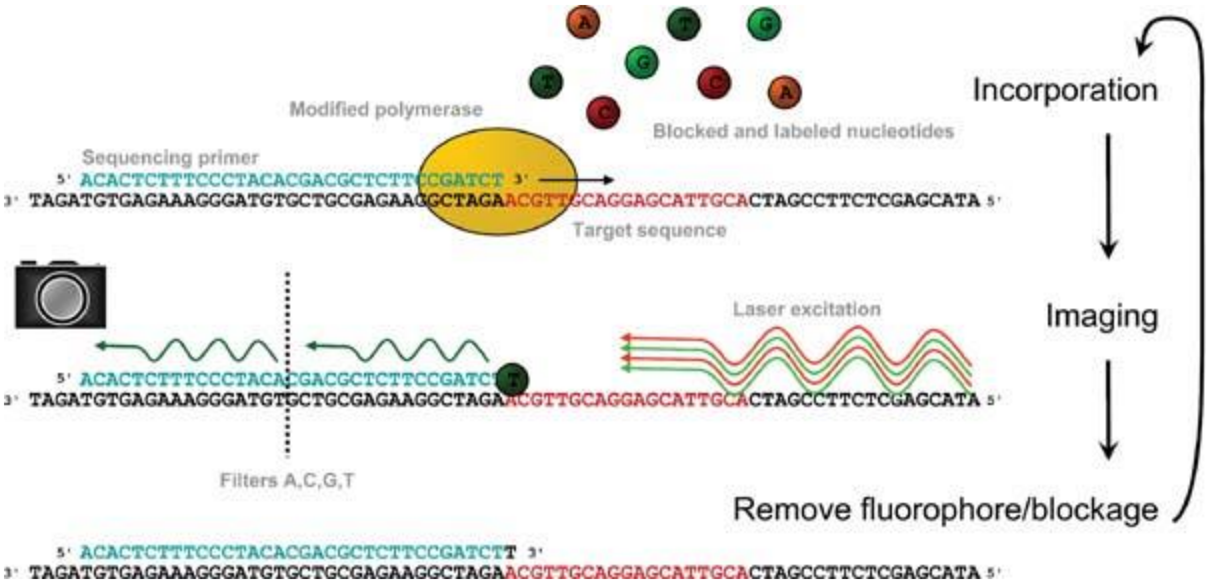
Roche 454 (pyrosequencing)


Double stranded DNAFragmented DNA
Single stranded DNA with adapters
Beads containing single stranded DNA (one strand per bead)
After n cycles of PCR
Each bead is covered with 10 million copies of template DNA
Beads captured in PCR-reaction-mixture-in-oil emulsion

Pyrosequencing. In this method, the double stranded DNA is fragmented into small pieces and then denatured to get single strands, which are ligated with adaptors at both the ends.
Each DNA fragment gets
bound to microbead and is amplified by emulsion PCR, generating many copies of a single molecule

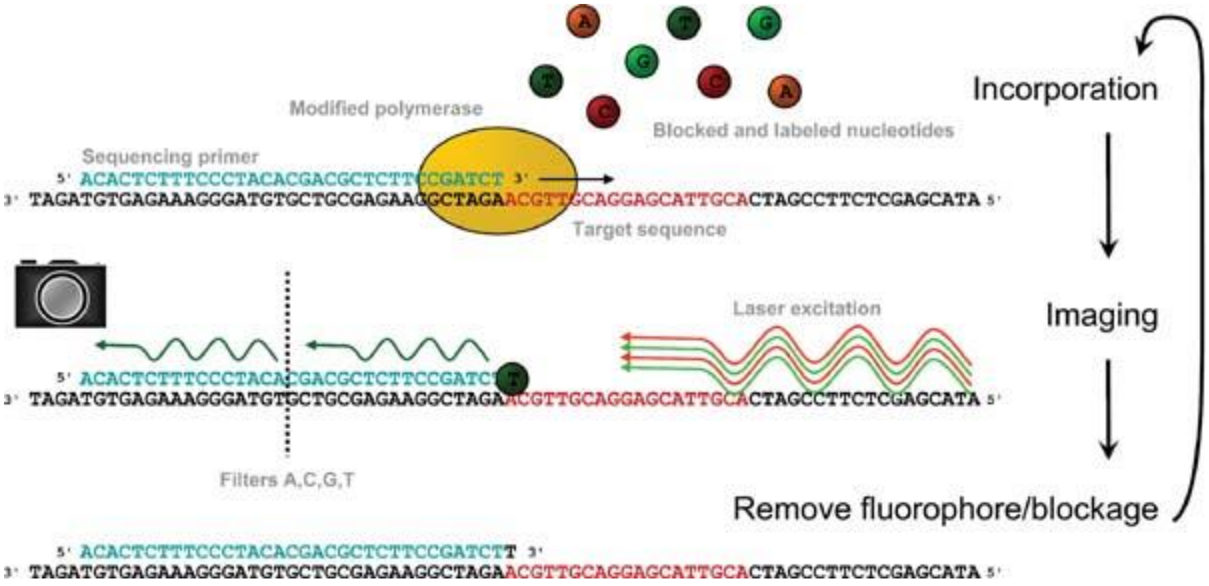
Illumina Platform

- Performs clonal amplification and sequencing via synthesis using reversible terminator chemistry with DNA polymerase and fluorophore-labeled terminator nucleotides.


A
T
G
G
T
C
C
A
Incorporation
Modified polymerase
Blocked and labeled nucleotides
Sequencing primer
5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT 3'
3' TAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAACGTTGCAGGAGCATTGCACTAGCCTTCTCGAGCATA 5'
Target sequence
Laser excitation
Imaging
5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT
T
3' TAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAACGTTGCAGGAGCATTGCACTAGCCTTCTCGAGCATA 5'
Filters A,C,G,T
Remove fluorophore/blockage
5' ACACTCTTTCCCTACACGACGCTCTTCCGATCTT 3'
3' TAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAACGTTGCAGGAGCATTGCACTAGCCTTCTCGAGCATA 5'

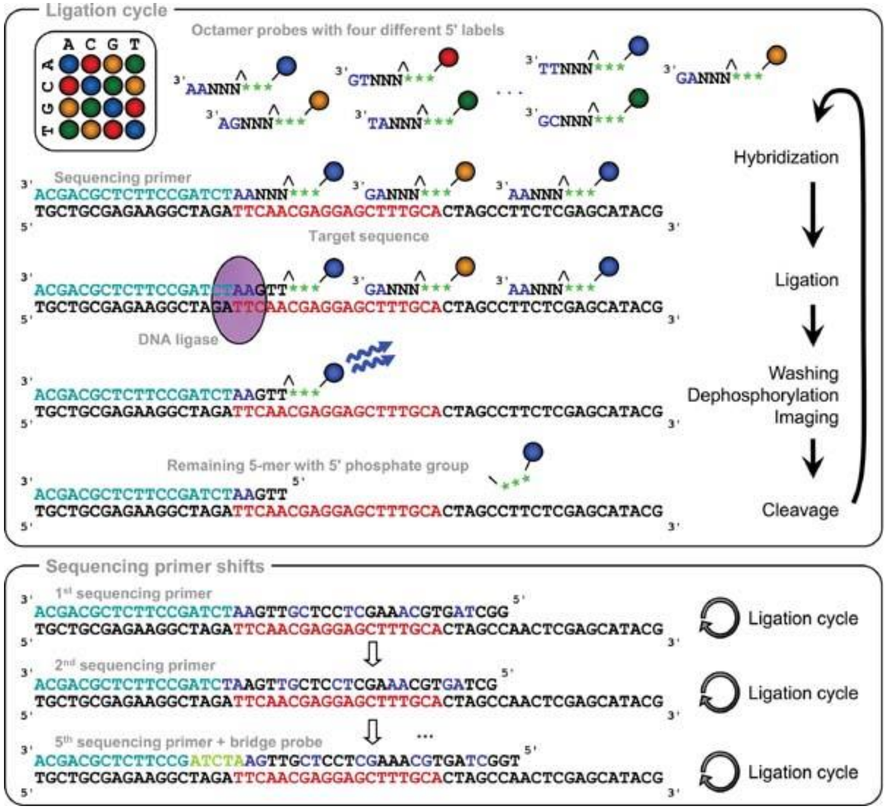
How is the reversible terminator chemistry applied?

- Sequencing primers are annealed to the adapters of the sequences to be determined.
- Polymerases are used to extend the sequencing primers by incorporation of fluorescently labeled and terminated nucleotides.


A
T
G
G
T
C
C
A
Modified polymerase
Blocked and labeled nucleotides
Incorporation
Sequencing primer
5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT 3'
3' TAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAACGTTGCAGGAGCATTGCACTAGCCTTCTCGAGCATA 5'
Target sequence
Laser excitation
Imaging
5' ACACTCTTTCCCTACACGACGCTCTTCCGATCT
T
3' TAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAACGTTGCAGGAGCATTGCACTAGCCTTCTCGAGCATA 5'
Filters A,C,G,T
Remove fluorophore/blockage
5' ACACTCTTTCCCTACACGACGCTCTTCCGATCTT 3'
3' TAGATGTGAGAAAGGGATGTGCTGCGAGAAGGCTAGAACGTTGCAGGAGCATTGCACTAGCCTTCTCGAGCATA 5'

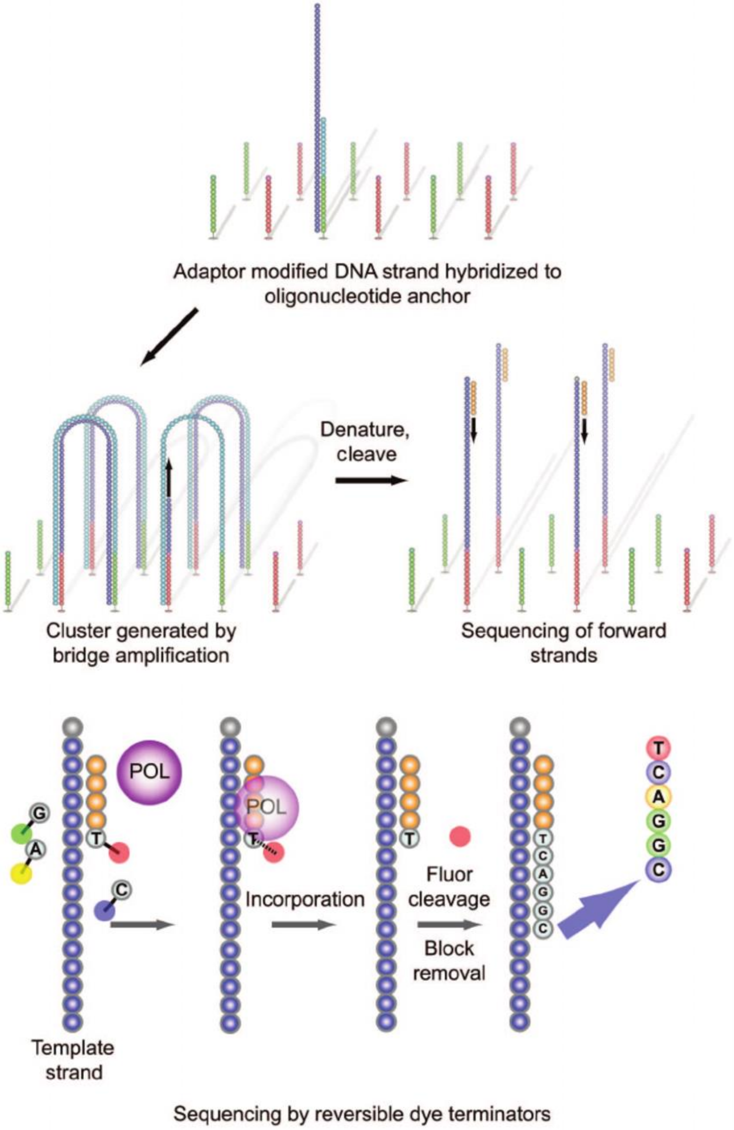
Principle of sequencing by synthesis

- Consists of using solid-phase PCR, in which single-stranded DNA fragments ligated to adapters are attached to a flow cell via hybridization


Ligation cycle
Octamer probes with four different 5' labels
A C G T
A C G T
3' AANNN***
3' AGNNN***
3' GTNNN***
3' TANNN***
...
3' TTNNN***
3' GCNNN***
3' GANNN***
Sequencing primer
3' ACGACGCTCTTCCGATCTAANNN*** 3' GANNN*** 3' AANNN***
TGCTGCGAGAAGGCTAGATTCAACGAGGAGCTTTGCACTAGCCTTCTCGAGCATACG 3'
5'
Target sequence
3' ACGACGCTCTTCCGATCTAAGTT*** 3' GANNN*** 3' AANNN***
5' TGCTGCGAGAAGGCTAGATTCAACGAGGAGCTTTGCACTAGCCTTCTCGAGCATACG 3'
DNA ligase
3' ACGACGCTCTTCCGATCTAAGTT***
5' TGCTGCGAGAAGGCTAGATTCAACGAGGAGCTTTGCACTAGCCTTCTCGAGCATACG 3'
Remaining 5-mer with 5' phosphate group
3' ACGACGCTCTTCCGATCTAAGTT 5'
5' TGCTGCGAGAAGGCTAGATTCAACGAGGAGCTTTGCACTAGCCTTCTCGAGCATACG 3'
Hybridization
Ligation
Washing
Dephosphorylation
Imaging
Cleavage
Sequencing primer shifts
1st sequencing primer
3' ACGACGCTCTTCCGATCTAAGTTGCTCCTCGAAACGTGATCGG 5'
5' TGCTGCGAGAAGGCTAGATTCAACGAGGAGCTTTGCACTAGCCAACTCGAGCATACG 3'
Ligation cycle
2nd sequencing primer
3' ACGACGCTCTTCCGATCTAAGTTGCTCCTCGAAACGTGATCG 5'
5' TGCTGCGAGAAGGCTAGATTCAACGAGGAGCTTTGCACTAGCCAACTCGAGCATACG 3'
Ligation cycle
...
5th sequencing primer + bridge probe
3' ACGACGCTCTTCCGATCTAAGTTGCTCCTCGAAACGTGATCGGT 5'
5' TGCTGCGAGAAGGCTAGATTCAACGAGGAGCTTTGCACTAGCCAACTCGAGCATACG 3'
Ligation cycle

Formation of bridge

- Amplification of DNA begins with the adapter of the free 3' end of the molecule binding to a complementary oligonucleotide, attached to the flow cell surface, forming a bridged structure.
- Thus, the PCR occurs through the addition of unlabeled nucleotides and other necessary reagents.

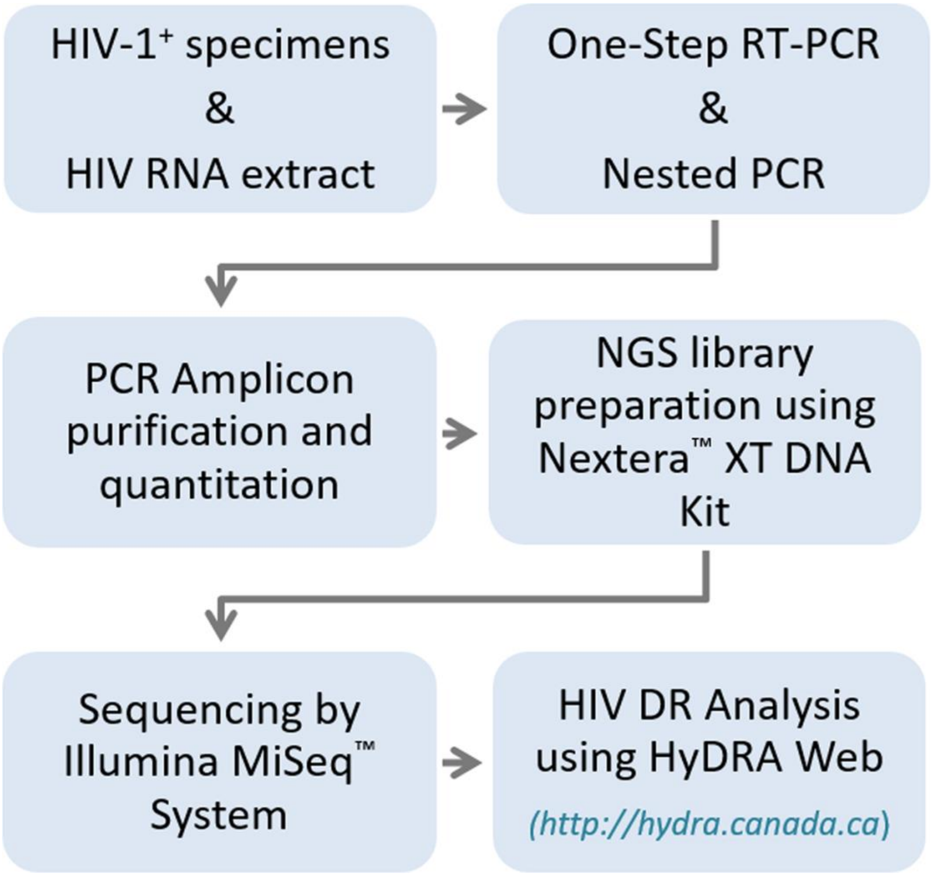

Adaptor modified DNA strand hybridized to oligonucleotide anchor
Denature, cleave
Cluster generated by bridge amplification
Sequencing of forward strands
POL
G
A
T
C
Incorporation
Fluor cleavage
Block removal
T
C
A
G
G
C
Template strand
Sequencing by reversible dye terminators

Incorporation of fluorophore-labeled terminator nucleotides

- Clusters of identical molecules are obtained at the end of this stage, and sequencing reactions occur within each cluster, with steps for the incorporation of fluorophore-labeled terminator nucleotides (chemically blocked 3’OH ends), excitation and reading;
- These steps are repeated for each nucleotide of the sequence.
- Base reading is performed in each sequencing cycle through sequential analysis of images captured by charge-coupled device cameras

MiSeq platform

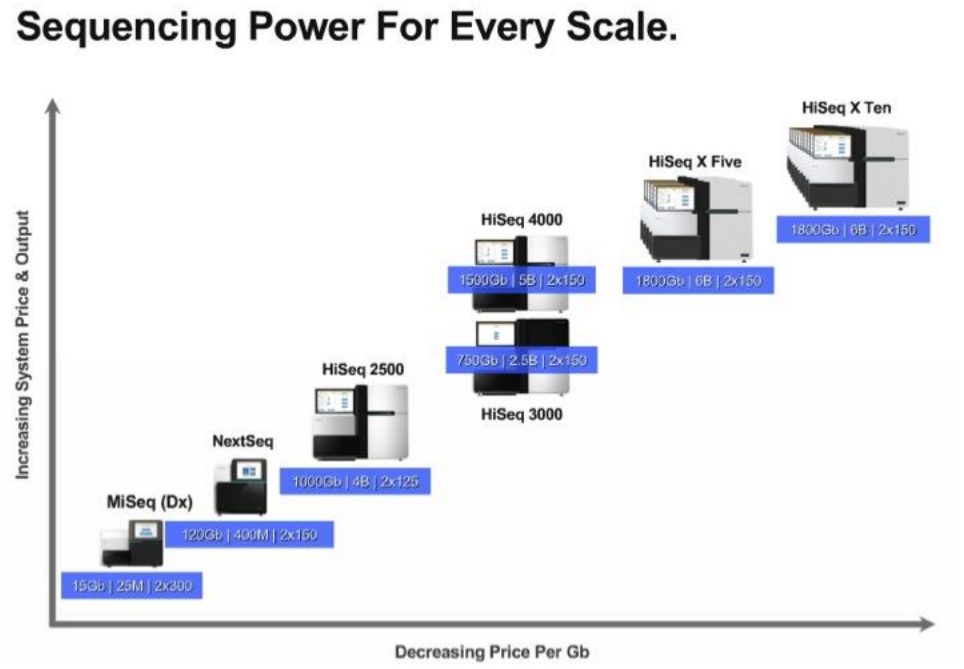
- The MiSeq platform is suitable for **the sequencing of small genomes such as those of bacteria**, as its throughput ranges from **0.3 to 15 Gb**, with reading lengths of up to **300 bp, using paired libraries**.



HIV-1+ specimens & HIV RNA extract
One-Step RT-PCR & Nested PCR
PCR Amplicon purification and quantitation
NGS library preparation using Nextera™ XT DNA Kit
Sequencing by Illumina MiSeq™ System
HIV DR Analysis using HyDRA Web (http://hydra.canada.ca)

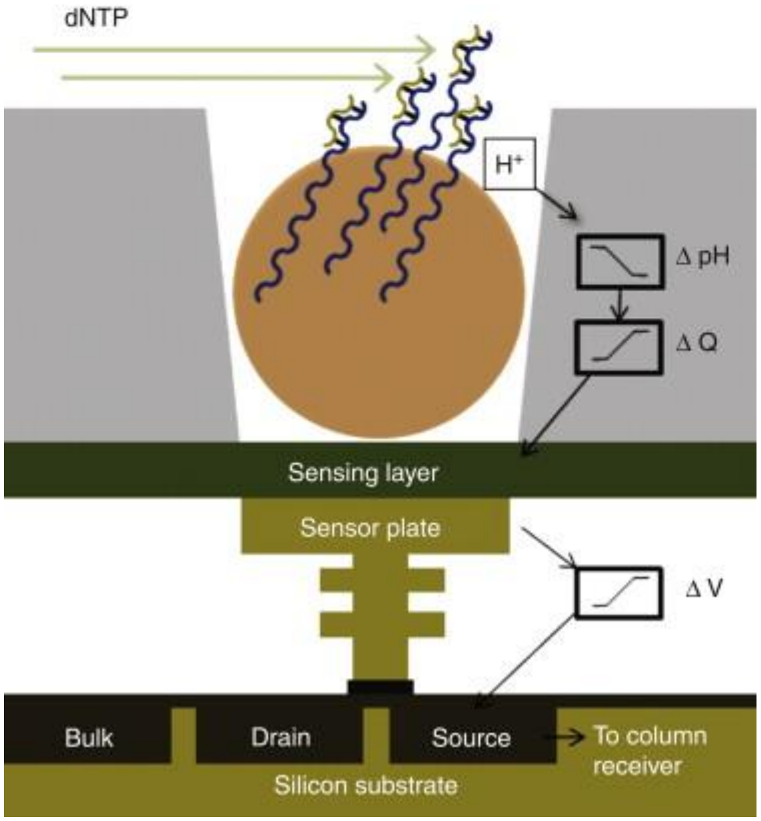
Other versions of illumina

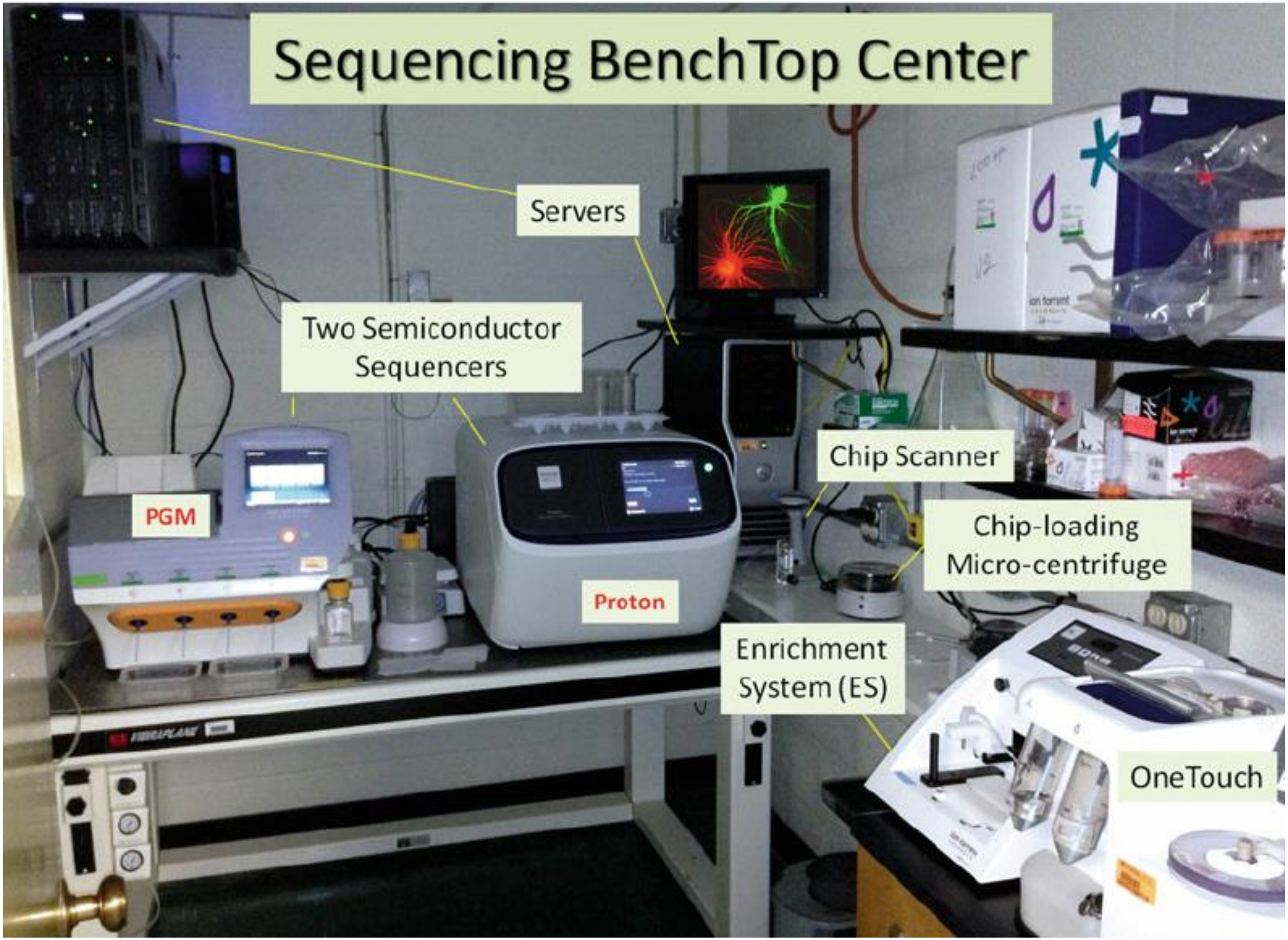
- HiSeq 3000 and 4000, which exhibit an approximate throughput of 125-750 Gb and 125-1500 Gb, respectively, with the **same maximum read length for both platforms** (150 bp), and allow the use of paired libraries



Ion Torrent Platform

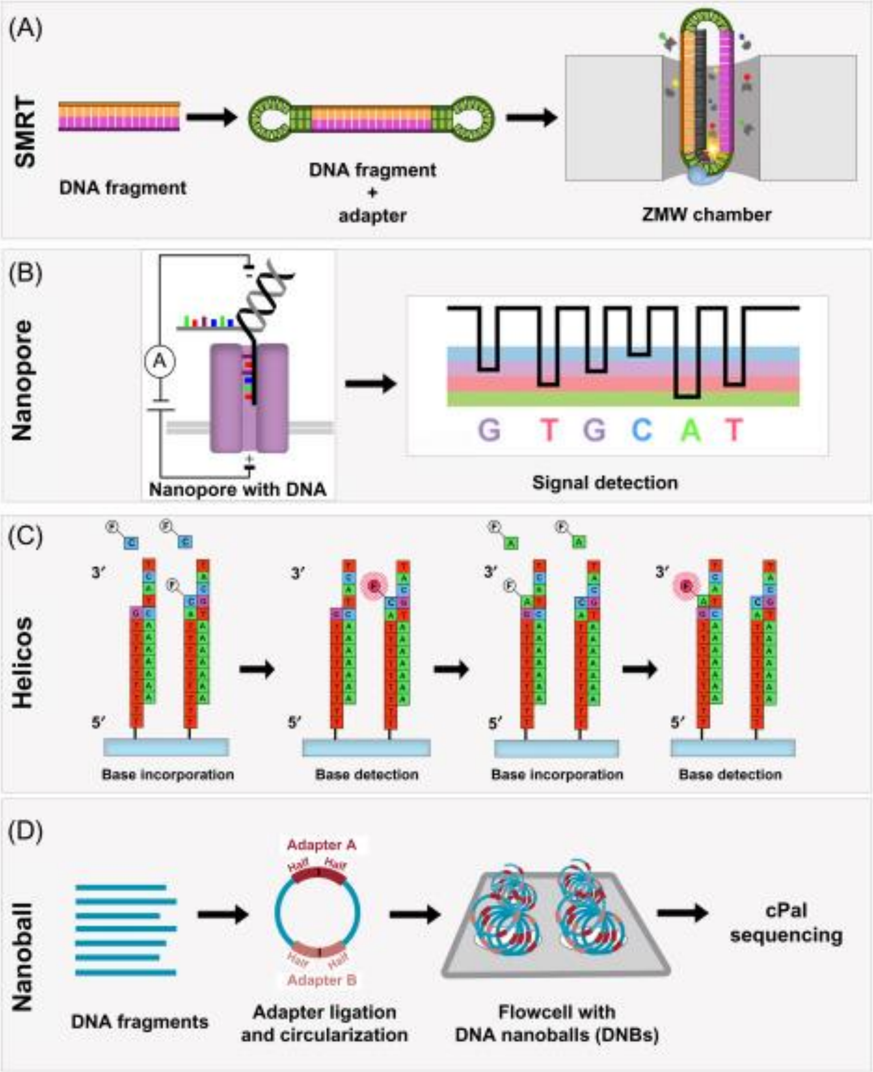
- The Ion Torrent platform Personal Genome Machine (PGM) uses the sequencing-by-synthesis (SBS) method, in which one H+ ion is released upon the incorporation of each nucleotide in a DNA molecule by the polymerase enzyme, changing the pH and allowing the identification of the nucleotide that is added during the synthesis of the DNA strand.



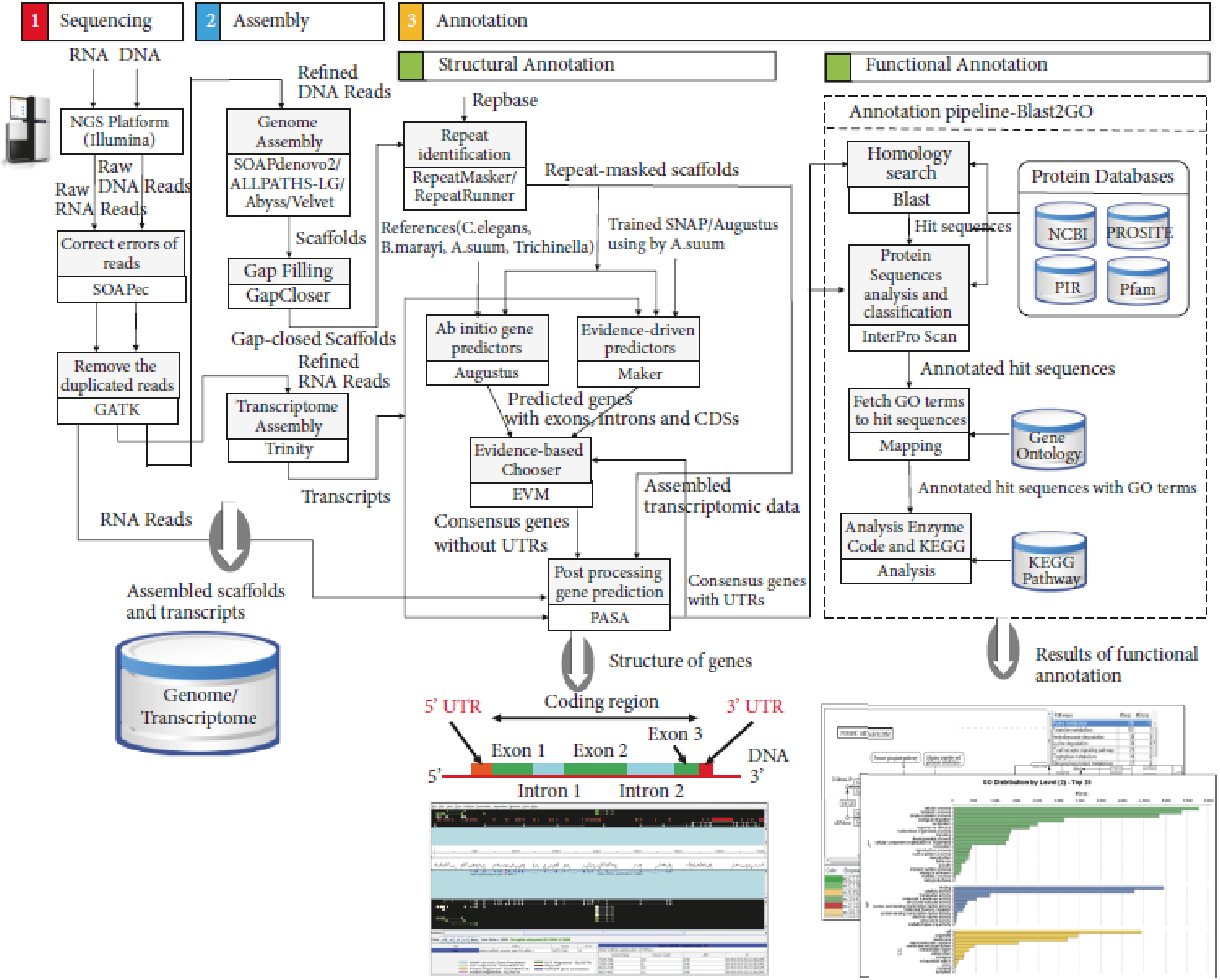

Sequencing BenchTop Center
Servers
Two Semiconductor Sequencers
PGM
Proton
Chip Scanner
Chip-loading Micro-centrifuge
Enrichment System (ES)
OneTouch

PacBio Platform

- PacBio technology from Pacific Biosciences is based on Single-Molecule Real-Time (SMRT) sequencing and is considered a next-generation technology.
- PacBio sequencing identifies the sequence information of the target DNA molecule during the replication process.
- A closed single-stranded circular DNA model referred to as SMRTbell is initially produced from adapters ligated at both ends of the double-stranded target DNA in hairpin format



	Throughput	Length	Quality	Costs	Applications	Main sources of errors
Sanger	6Mb/day	800nt	10^{-4}-10^{-5}	~500$/Mb	Small sample sizes, genomes/scaffolds, InDels/SNPs, long haplotypes, low complexity regions, ...	Polymerase/amplification, low intensities/missing termination variants, contaminant sequences
454/Roche	750Mb/day	400nt	10^{-3}-10^{-4}	~20$/Mb	Complex genomes, SNPs, structural variation, indexed samples, smallRNA^{+}, mRNAs^{+}, ...	Amplification, mixed beads, intensity thresholding, homopolymers, phasing, neighbor interference
Illumina	5000Mb/day	100nt	10^{-2}-10^{-3}	~0.50$/Mb	Complex genomes, counting (SAGE, CNV ChIP, small RNA), mRNAs, InDels/ homopolymers, structural variation, bisulphite data, indexing, SNPs^{+}, ...	Amplification, mixed clusters/neighbor interference, phasing, base labeling
SOLiD	5000Mb/day	50nt	10^{-2}-10^{-3}	~0.50$/Mb	Complex small genomes, counting (SAGE, ChIP, small RNA, CNV), SNPs, mRNAs, structural variation, indexing, ...	Amplification, mixed beads, phasing, signal decline, neighbor interference
Helicos	5000Mb/day	32nt	10^{-2}	<0.50$/Mb	Non-amplifiable samples, counting (SAGE, ChIP, small RNA), ...	Polymerase, low intensities / thresholding, molecule loss/termination


1 Sequencing
2 Assembly
3 Annotation
Structural Annotation
Functional Annotation
RNA DNA
NGS Platform (Illumina)
Raw RNA Reads
Raw DNA Reads
Correct errors of reads
SOAPec
Remove the duplicated reads
GATK
Refined DNA Reads
Genome Assembly
SOAPdenovo2/ ALLPATHS-LG/ Abyss/Velvet
Scaffolds
Gap Filling
GapCloser
Gap-closed Scaffolds
Refined RNA Reads
Transcriptome Assembly
Trinity
Transcripts
RNA Reads
Assembled scaffolds and transcripts
Genome/ Transcriptome
Repbase
Repeat identification
RepeatMasker/ RepeatRunner
Repeat-masked scaffolds
References(C.elegans, B.marayi, A.suum, Trichinella)
Trained SNAP/Augustus using by A.suum
Ab initio gene predictors
Augustus
Evidence-driven predictors
Maker
Predicted genes with exons, introns and CDSs
Evidence-based Chooser
EVM
Consensus genes without UTRs
Assembled transcriptomic data
Post processing gene prediction
PASA
Consensus genes with UTRs
Structure of genes
5' UTR
Coding region
3' UTR
Exon 1
Exon 2
Exon 3
DNA
5'
3'
Intron 1
Intron 2
Annotation pipeline-Blast2GO
Homology search
Blast
Hit sequences
Protein Databases
NCBI
PROSITE
PIR
Pfam
Protein Sequences analysis and classification
InterPro Scan
Annotated hit sequences
Fetch GO terms to hit sequences
Mapping
Gene Ontology
Annotated hit sequences with GO terms
Analysis Enzyme Code and KEGG
Analysis
KEGG Pathway
Results of functional annotation
GO Distribution by Level (2) - Top 20

Genomics and Proteomics [22ZOOME31]

Structural Genomics and Genome Annotation

Introduction

- Genomes represent the **starting point of genetic studies**
- **A genome** is the complete genetic information of an organism or a cell.
- Single or double stranded nucleic acids store this information in a **linear or in a circular sequence**.
- Genome sequence - an organism’s blueprint: the set of instructions dictating its biological traits.
- Central dogma of protein synthesis
- majority of RNA sequences originate from protein coding genes; that is, they are processed into messenger RNAs (mRNAs) which, after their export to the cytosol, are translated into proteins
- proteins - main functional and structural players in the cell.
- The delineation of the complete set of protein-coding genes and their alternative splice forms - essential to the task of translating the information in the sequence of the genome into biologically relevant knowledge
- drafts of the human genome
- To precisely determine this sequence, progressively more efficient technologies characterized by increased accuracy, throughput and sequencing speed have been developed.
- drafts of the human genome

STRUCTURAL GENOMICS

- De Novo Assembly
- Reference Assembly
- Genome Annotation

De Novo Assembly

- The de novo genome assembly process consists of grouping reads obtained through sequencing via base pairing to represent the **complete genome without the aid of reference sequences**.
- Problems in this process are **repetitions in the genome,** especially in eukaryotic organisms, regions with **low sequencing coverage due to biases such as GC bias,** and sequencing artifacts peculiar to each platform
- Thus, the development of assembly algorithms to evaluate issues such as these is quite complex.
- Therefore, different computational approaches are used for genome assembly, such as **greedy algorithms, overlap_layout_consensus (OLC), and De Bruijn graphs**

Greedy algorithms

- The greedy algorithm approach is characterized by an **extensive computational effort** because it is based on the **analysis of all possible sequence alignments.**
- The application of this method in conjunction with De Bruijn graphs and OLC is common
- examples of software that utilize in this approach.
  - SSAKE
  - SHARCGS and
  - VCAKE

Overlap layout consensus (OLC)

- Divided into three steps:
- First consists of **identifying the possible overlap in the set of reads.**
- Second is followed by the **construction of a graph based on the identified overlaps**, and
- the consensus sequence is then finally **generated through an algorithm** that tours the graph, visiting each node exactly once.
- Examples of software that utilize in this approach.
  - Newbler
  - Mira and
  - Edena

De Bruijn graphs

- consists of the fragmentation of reads into k-mer sizes, in which k represents the read length.
- Subsequently, they are evaluated to find overlaps of k-1-mers among the k-mers initially generated, considering the identity of the bases
- Velvet - used software platforms for the assembly of prokaryotic genomes using short reads
- SPADES is employed for data obtained through the Ion Torrent platform, whose reads exhibit random lengths.
- SOAPdenovo and ALL-PATHS-LG software are examples of applications of De Bruijn graphs
- widely used for the assembly of eukaryotic genomes, particularly because of their ability to optimize memory consumption during the process.
- These genomes require a **high computational effort due to their complexity, with the presence of the major repetitive regions and various chromosomes, and their large size, generally on the order of gigabases,** in addition to the small number of complete eukaryotic genome projects
- A single consensus sequence representing a genome, or various sequences representing portions of a genome that need to be oriented and arranged in a process known as scaffolding, can be obtained after the assembly process using each one of the aforementioned approaches
- The scaffolds exhibit regions that were not represented during the assembly process for reasons such as the stringency of the adopted parameters and coverage biases, producing regions referred to gaps, and represented by the letter “N”

- The use of different sequencing strategies can be useful in resolving these regions, which often originate from repetitions in the genome.
- **Paired genomic libraries**, such as **paired-end and mate-pair libraries**, are used to represent these areas because they employ long fragments of several kilobases and only sequence the ends, such that the distance between the pairs is known
- Some software is applied **after the assembly process to generate scaffolds and to resolve gap regions.**
- Thus, the overlap between the ends of contiguous sequences (contigs) is analyzed and can ensure the extension of the sequence - SSPACE software
- GAPFILLER tools that use paired libraries to resolve gaps
- Evaluation of the assembly process is performed using metrics such as N50, which **evaluates the length of the sequences** produced, in addition to the **average length and number of contigs as well as largest and smallest contigs.**
- The assembled contigs can be evaluated through the **mapping of paired reads,** when available, to confirm the results
- The core eukaryotic genes can be sought in the results of the assembly for eukaryotic genomes using the **Genome Assembly Gold-standard Evaluation (GAGE)** tool

Reference Assembly

- The mapping of sequences produced by NGS platforms is one of the most common applications of genomic data.
- However, this process represents a computational challenge due to the characteristics of the sequences originating from these platforms, **mostly short reads.**
- Several obstacles must be overcome to perform this task, such as accurate mapping of the reads to reference sequences, distinguishing between errors that occurred during sequencing and true genetic variations, and the difficulty of handling the large amount of data produced by these platforms
-

- BWA which uses the **Burrows_Wheeler transformation algorithm** to increase mapping speed
- SHRiMP - compatible with data in **letter space and color space format produced by the SOLiD platform**
- SOAP2 which was developed to **conduct single nucleotide polymorphism studies**, whose current version features significant improvements in **memory management and alignment speed**
- TopHat2 which is suitable **for data from the Ion Torrent platform** and
- mrsFAST which **examines all possibilities for mapping to the reference genome**, making it useful for variance detection studies

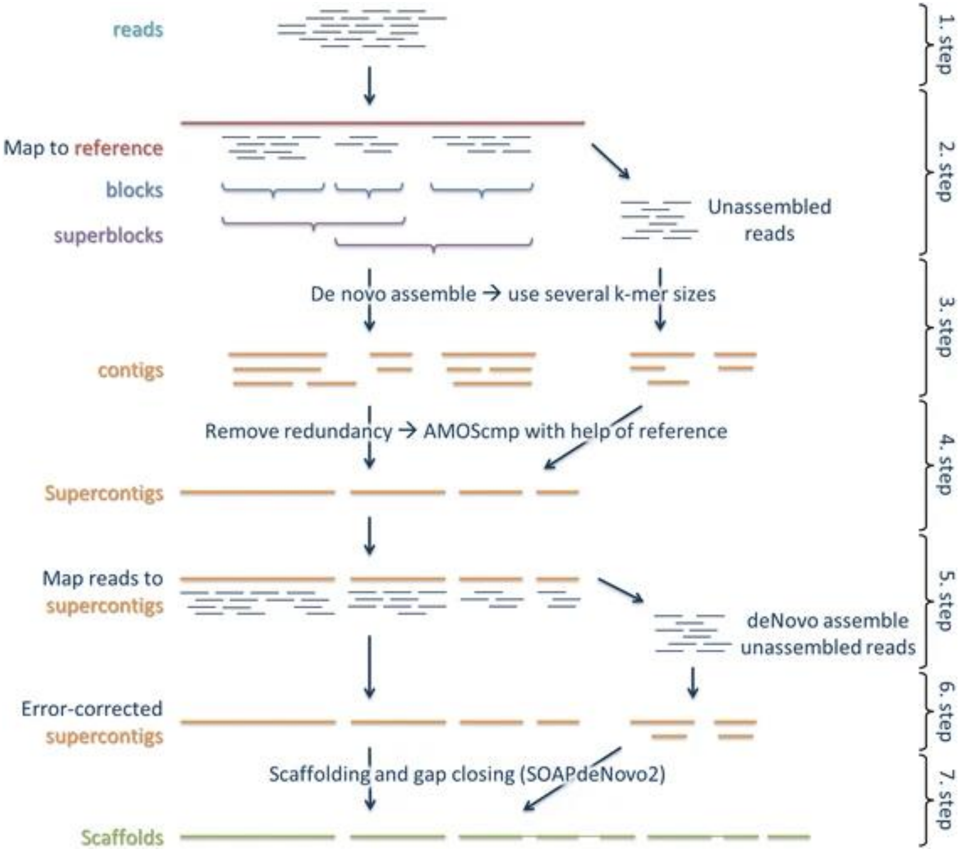
Genome Annotation

- Genome annotation consists of describing the function of the **product of a predicted gene (through an in silico approach).**
- This can be achieved using bioinformatics software with specific features, including (1) signal sensors (e.g., for TATA box, start and stop codon, or poly-A signal detection),
- (2) content sensors (e.g., for G1C content, codon usage, or dicodon frequency detection), and
- (3) similarity detection (e.g., between proteins from closely related organisms, mRNA from the same organism, or reference genomes)
- The method for predicting gene and genome structures (e.g., tRNAs, rRNAs, promoter regions) is associated with the applied assembly strategies and sequencing platforms

- Genome annotation can be divided into three basic categories.
- 1. **Nucleotide level annotation**, which seeks to identify the physical location of DNA sequences to determine where components such as genes, RNAs, and repetitive elements are located.
- Sequencing and/or assembly errors at this stage can result in false pseudogenes through ins/dels.
- 2. **Protein-level annotation,** which seeks to determine the possible functions of genes, identifying which one a given organism does or does not have.
- 3. **Process-level annotation,** which aims to identify the pathways and processes in which different genes interact, assembling an efficient functional annotation.
- The last two levels, sequencing and/or assembly errors may compromise the inference of the true gene function because of reduced similarity

The complete genome sequence - deduced from the overlaps of these shorter fragments, a process defined as de novo genome assembly

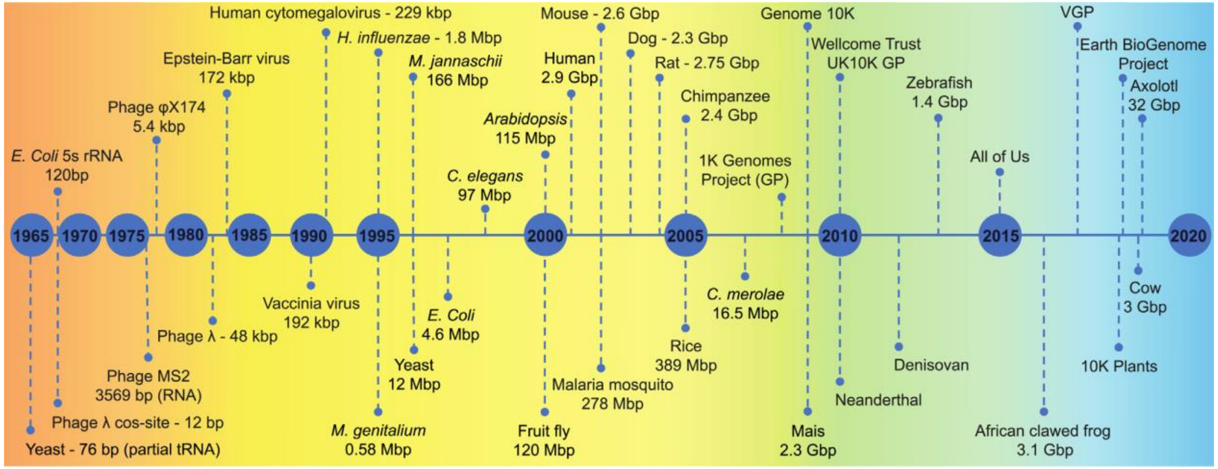
1. **Raw reads get quality trimmed (1. step) and**
2. **mapped against a reference (2. step).**
3. **Reference mapped reads are grouped into blocks with continuous read coverage. These blocks are then combined into superblocks until a total length of at least 12 kb is reached. Superblocks are overlapping by at least one block. Each superblock and all unmapped reads are separately de novo assembled (3. step).**
4. **Resulting contigs are merged into non-redundant supercontigs (4. step).**
5. **reads are mapped back to the supercontigs and unmapped reads are de novo assembled to get additional supercontigs.**
6. **All supercontigs are error corrected with back mapped reads (6. step) and**
7. **afterwards used for scaffolding and gap closing (7. step)**


reads
1. step
Map to reference
blocks
superblocks
Unassembled reads
2. step
De novo assemble → use several k-mer sizes
contigs
3. step
Remove redundancy → AMOScmp with help of reference
Supercontigs
4. step
Map reads to supercontigs
deNovo assemble unassembled reads
5. step
Error-corrected supercontigs
6. step
Scaffolding and gap closing (SOAPdeNovo2)
Scaffolds
7. step

- due to time and cost constraints, only an individual per species was addressed, and its sequence generally represents the **‘reference’ genome for the species**.
- Reference genomes can guide re-sequencing efforts in the same species, acting as a **template for read mapping.**
- They can be annotated to understand gene function or used **to design gene manipulation experiments.**
- Sequences from different species can be aligned and compared **to study molecular evolution**.
- Due to the impact of reference genomes in all these downstream applications, it is paramount that their sequence is as much complete and error-free as possible.

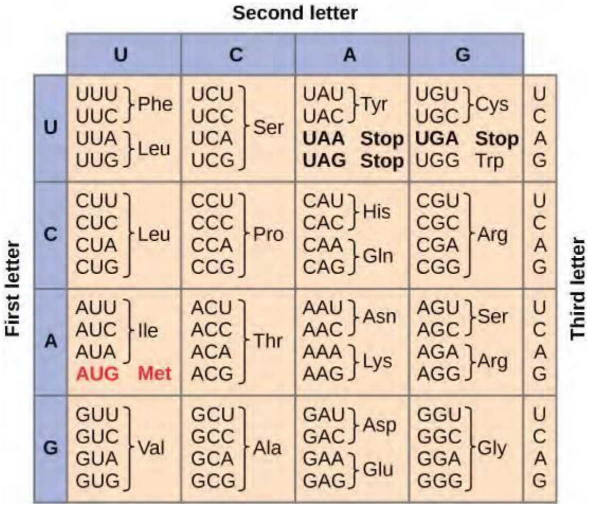
Sequencing nucleic acids in the XX century

- 1953 - Watson and Crick - published their seminal paper unraveling the double helix structure of DNA
- “so far as is known the sequence of bases along the chain is irregular” and “the sequence of bases on a single chain does not appear restricted in any way”, two features that entail a role DNA in the storage of genetic information, and highlight the importance of determining the exact sequence of bases along the chain
- 1953 - marked the first sequencing of a biological molecule.
- Thanks to a refined partition chromatography method, Sanger was able to sequence the two chains of insulin protein



Why is it importance?

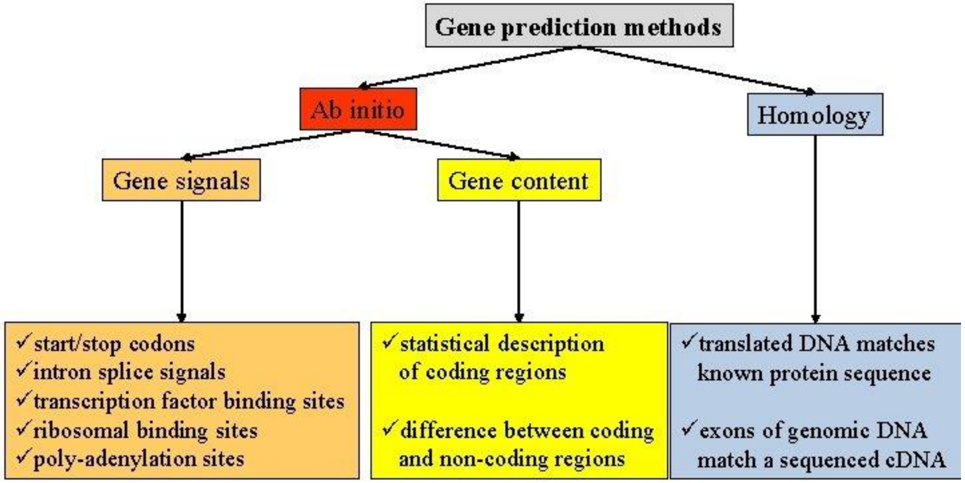
- **Gene finding** is the process of identifying genome sequence regions representing stretches of DNA that encode **biologically active products, such as proteins or functional noncoding RNAs**
- The coding regions of a genome contain the instructions to **build functional proteins.**
- If transcript or protein sequences are already known for an organism, these can be used to **determine the location of the corresponding genes in the genomic sequence**.
- Cost-effective next generation sequencing (NGS) techniques have become widely available, which can produce tens to hundreds of million sequence reads per run and can be used **to sequence complementary DNA (cDNA) generated from RNA transcripts.**
- **Natural selection** acting on the encoded protein product restricts the rate of mutation in coding sequences (CDSs) compared to non-functional genomic DNA.
- Protein-coding genes in particular can be identified by their characteristic **three periodic pattern of conservation.**
- Due to the degeneracy of the genetic code, **different nucleotides at the third codon** position often encode the same amino acid, thereby allowing for alterations at this codon position **without changing the encoded protein**.



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

- Approaches which **use information about expressed sequences or sequence conservation** are called **“extrinsic,”** as they require additional knowledge besides the genomic sequence of the organism being analysed.
- The “intrinsic” approach to **gene identification**, which is based on the evaluation of **characteristic differences between coding and noncoding genomic sequence regions.**
- characteristic ***differences in the distribution of short DNA oligomers between coding and noncoding regions*** (often called **codon bias**).

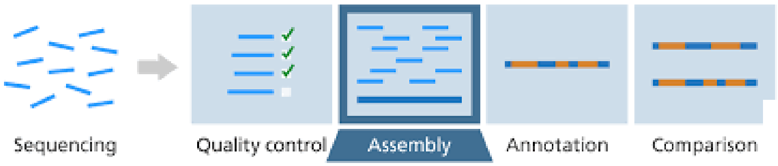
Categories of gene prediction programs

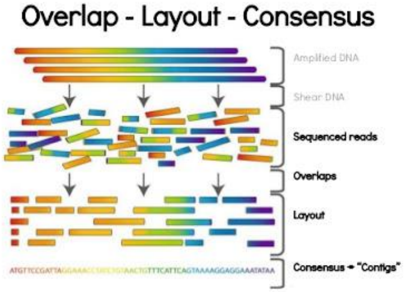

Gene prediction methods
Ab initio
Homology
Gene signals
Gene content
✓start/stop codons
✓intron splice signals
✓transcription factor binding sites
✓ribosomal binding sites
✓poly-adenylation sites
✓statistical description of coding regions
✓difference between coding and non-coding regions
✓translated DNA matches known protein sequence
✓exons of genomic DNA match a sequenced cDNA

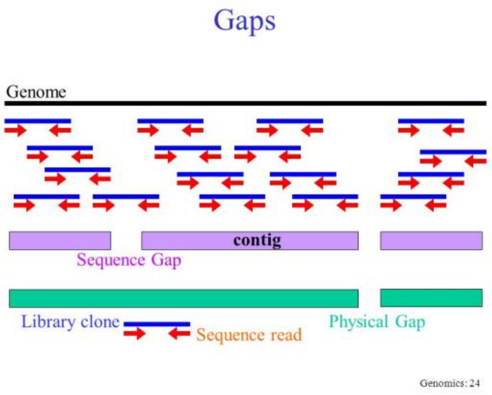
Intrinsic methods: without reference to known sequences
Extrinsic methods: with reference to known sequences

Lecture 3. Gene Finding and Sequence Annotation

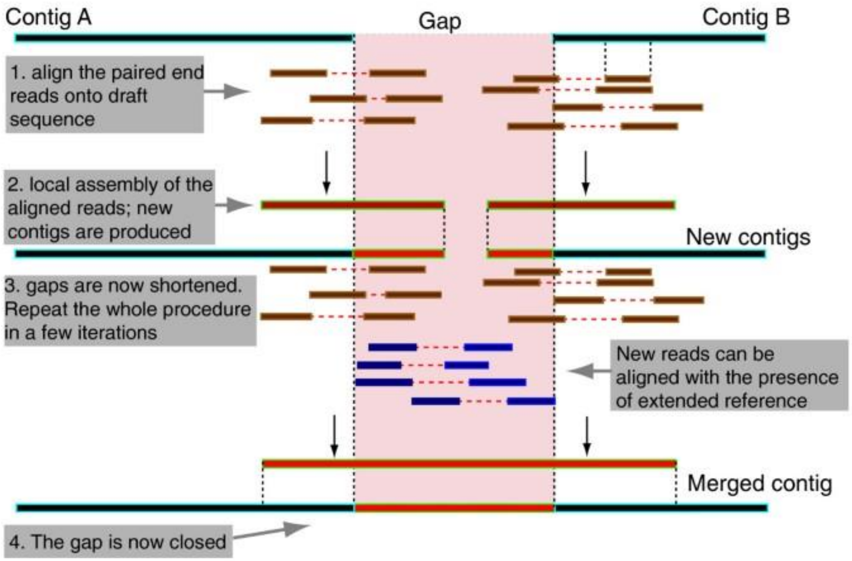
- **Assembler:** A computer program that pieces together overlapping reads to reconstruct the original sequence.
- **gap** Unsequenced area between two contigs.
- **uncaptured (physical gap)** Unsequenced area between two contigs with no subclones spanning it.
- **Captured gap (a sequence gap)** Unsequenced area between two contigs spanned by at least one subclone.
- **Contig** ‘A set of gel readings that are related to one another by overlap of their sequences.
- All gel readings belong to **one and only one contig**, and each contig contains at least one gel reading.
- The gel readings in a contig can be summed to form a contiguous consensus sequence and the length of this sequence is the length of the contig’.



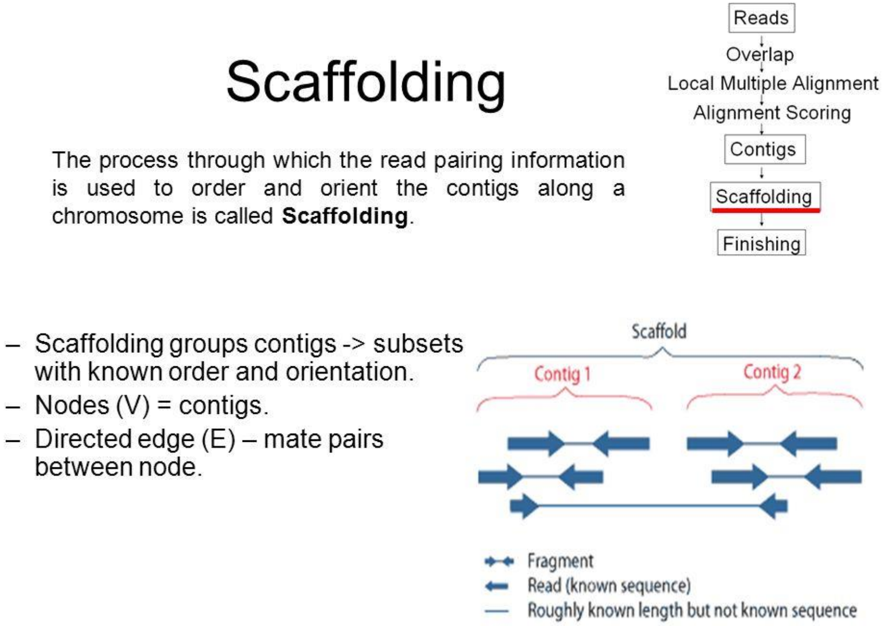




- Finishing The process of improving a draft assembly composed of **shotgun sequencing reads, resolving misassembled regions**, closing sequence gaps, and validating low-quality regions to produce a highly accurate finished DNA sequence (<1 error in 10 000 bp).
- paired (sister) reads Sequences generated from both ends of a DNA fragment. Such reads are oriented toward each other and the distance between them is equal to the template length.


Contig A
Gap
Contig B
1. align the paired end reads onto draft sequence
2. local assembly of the aligned reads; new contigs are produced
New contigs
3. gaps are now shortened. Repeat the whole procedure in a few iterations
New reads can be aligned with the presence of extended reference
Merged contig
4. The gap is now closed

- quality score The probability of a wrong **basecall.**
- A Phrap quality score of X corresponds to an error probability of approximately 10–X/10 (a score of 30 means that the error probability is 1/1000 or 99.9% accuracy for a base in the assembled sequence).
- **read** Gel reading generated during the DNA sequence process.
- **repeats** Sequences of varying lengths found in multiple copies in the genome.
- **scaffold** A group of ordered and oriented contigs.
- For large genomes, even long reads fail to generate end-to-end chromosome sequences, requiring linkage information to orient and order the contigs, a process known as **scaffolding.**



Scaffolding

The process through which the read pairing information is used to order and orient the contigs along a chromosome is called **Scaffolding**.

Reads
Overlap
Local Multiple Alignment
Alignment Scoring
Contigs
Scaffolding
Finishing

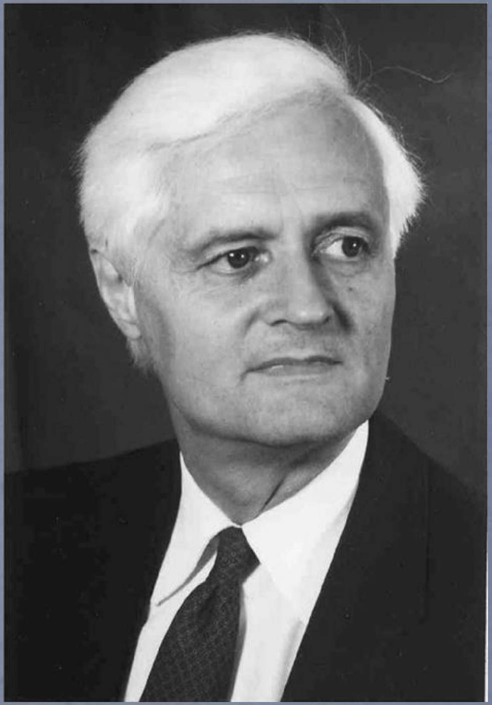
- Scaffolding groups contigs -> subsets with known order and orientation.
- Nodes (V) = contigs.
- Directed edge (E) – mate pairs between node.

Genomics and Proteomics
[22ZOOME31]

Pharmacogenetics

Pharmacogenetics

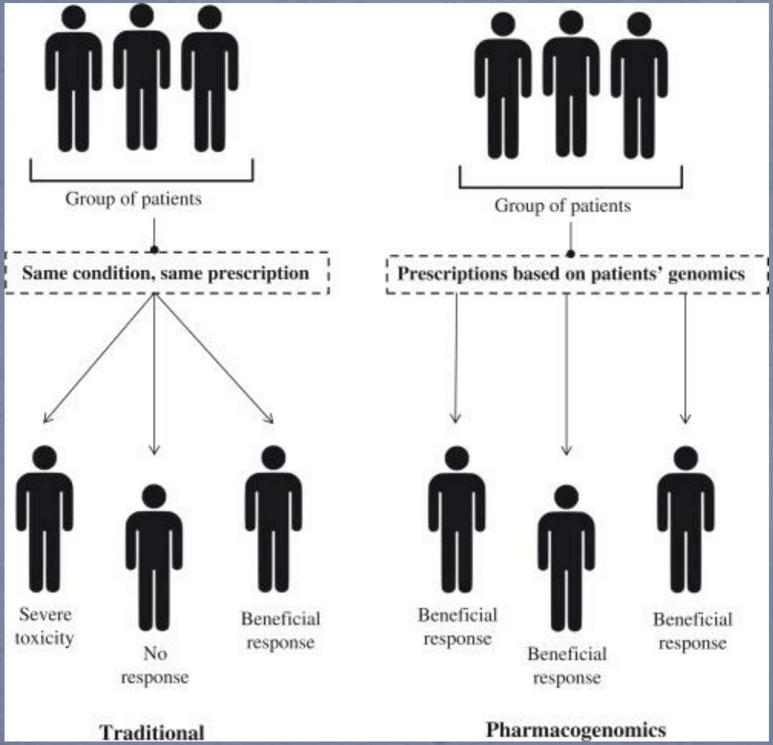
- *Individualized drug therapy* represents a major portion of ***personalized medicine***.
- **Friedrich O. Vogel**, coined the term *pharmacogenetics* and defined it as "the study of heritable variability in drug response", or, simply, "gene–drug interactions."



1925 -2006

Vision towards Pharmacogenetics

- *Drug responses* will depend upon **each patient's genetic make-up**
- Arno G. Motulsky



Founder of medical genetics

Our genetic make-up determines our *individual drug response*.

Each individual's drug response is ***holistic*** in that it actually encompasses five contributing influences:

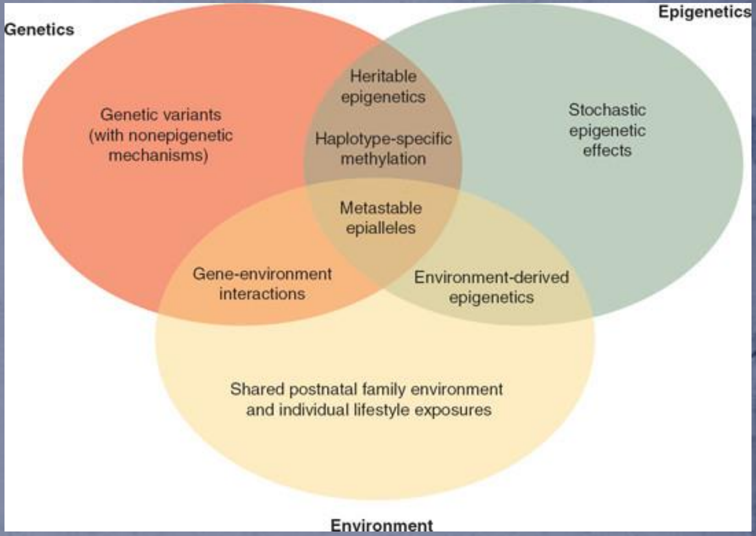
(1) ***genotype*** (DNA single-nucleotide variants, insertions, deletions, duplications, and inversions);

(2) ***epigenetic effects*** (DNA methylation, RNA interference, histone modifications, and chromatin remodeling);

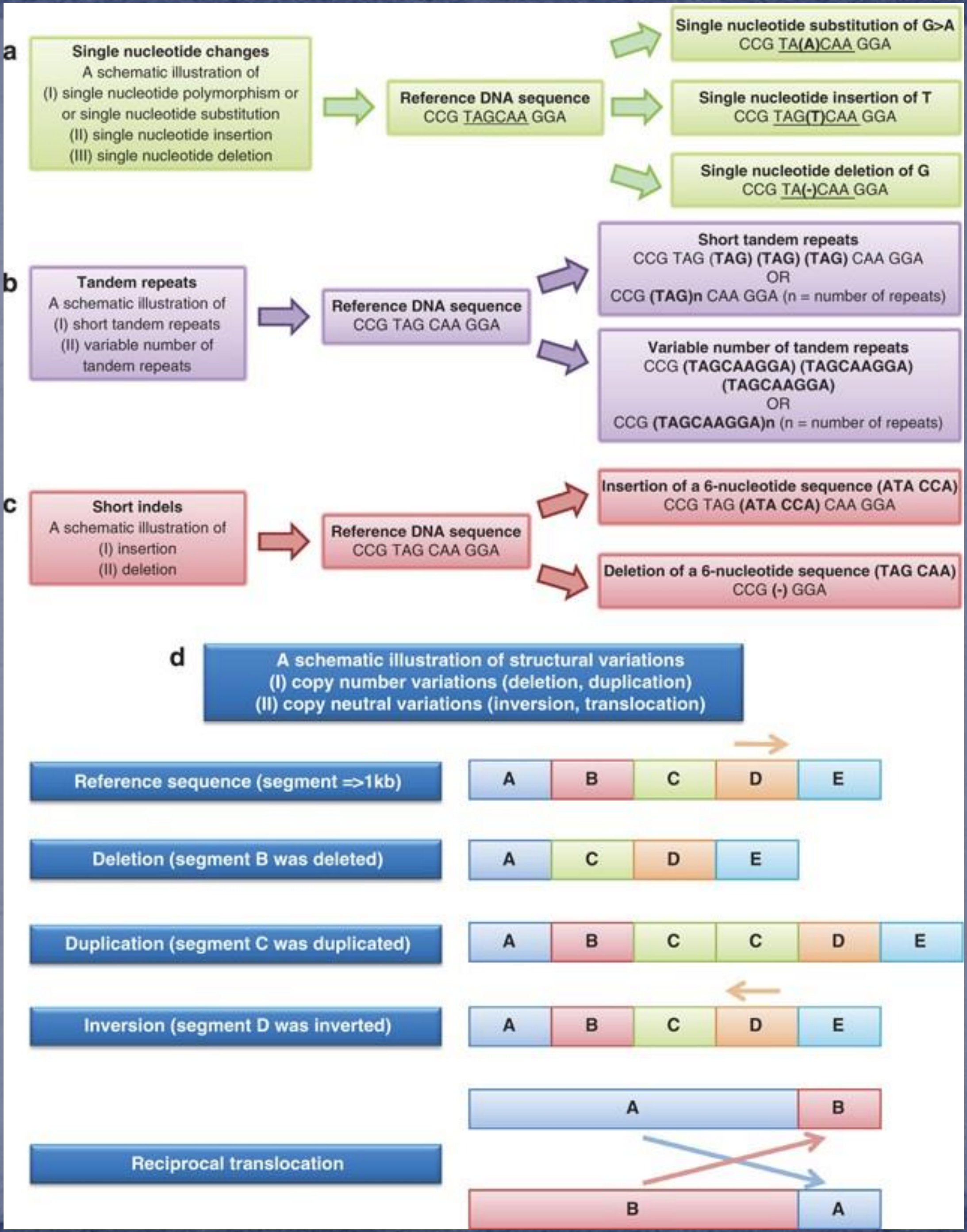
(3) **endogenous influences** (age, gender, ethnicity, exercise, various disease states, and functional status of kidneys and other organs);

(4) **environmental factors** (diet, cigarette smoking, lifestyle, drug–drug interactions, and significant exposure to occupational chemicals and other environmental pollutants); and

(5) **microbiome** differences specific to each person


Genetics
Epigenetics
Environment
Genetic variants (with nonepigenetic mechanisms)
Heritable epigenetics
Haplotype-specific methylation
Stochastic epigenetic effects
Metastable epialleles
Gene-environment interactions
Environment-derived epigenetics
Shared postnatal family environment and individual lifestyle exposures

Genotype (DNA single-nucleotide variants, insertions, deletions, duplications, and inversions);



a

Single nucleotide changes
A schematic illustration of
(I) single nucleotide polymorphism or
or single nucleotide substitution
(II) single nucleotide insertion
(III) single nucleotide deletion

Reference DNA sequence
CCG TAGCAA GGA

Single nucleotide substitution of G>A
CCG TA**(A)**CAA GGA

Single nucleotide insertion of T
CCG TAG**(T)**CAA GGA

Single nucleotide deletion of G
CCG TA**(-)**CAA GGA

b

Tandem repeats
A schematic illustration of
(I) short tandem repeats
(II) variable number of
tandem repeats

Reference DNA sequence
CCG TAG CAA GGA

Short tandem repeats
CCG TAG **(TAG) (TAG) (TAG)** CAA GGA
OR
CCG **(TAG)n** CAA GGA (n = number of repeats)

Variable number of tandem repeats
CCG **(TAGCAAGGA) (TAGCAAGGA)**
(TAGCAAGGA)
OR
CCG **(TAGCAAGGA)n** (n = number of repeats)

c

Short indels
A schematic illustration of
(I) insertion
(II) deletion

Reference DNA sequence
CCG TAG CAA GGA

Insertion of a 6-nucleotide sequence (ATA CCA)
CCG TAG **(ATA CCA)** CAA GGA

Deletion of a 6-nucleotide sequence (TAG CAA)
CCG (-) GGA

d

A schematic illustration of structural variations
(I) copy number variations (deletion, duplication)
(II) copy neutral variations (inversion, translocation)

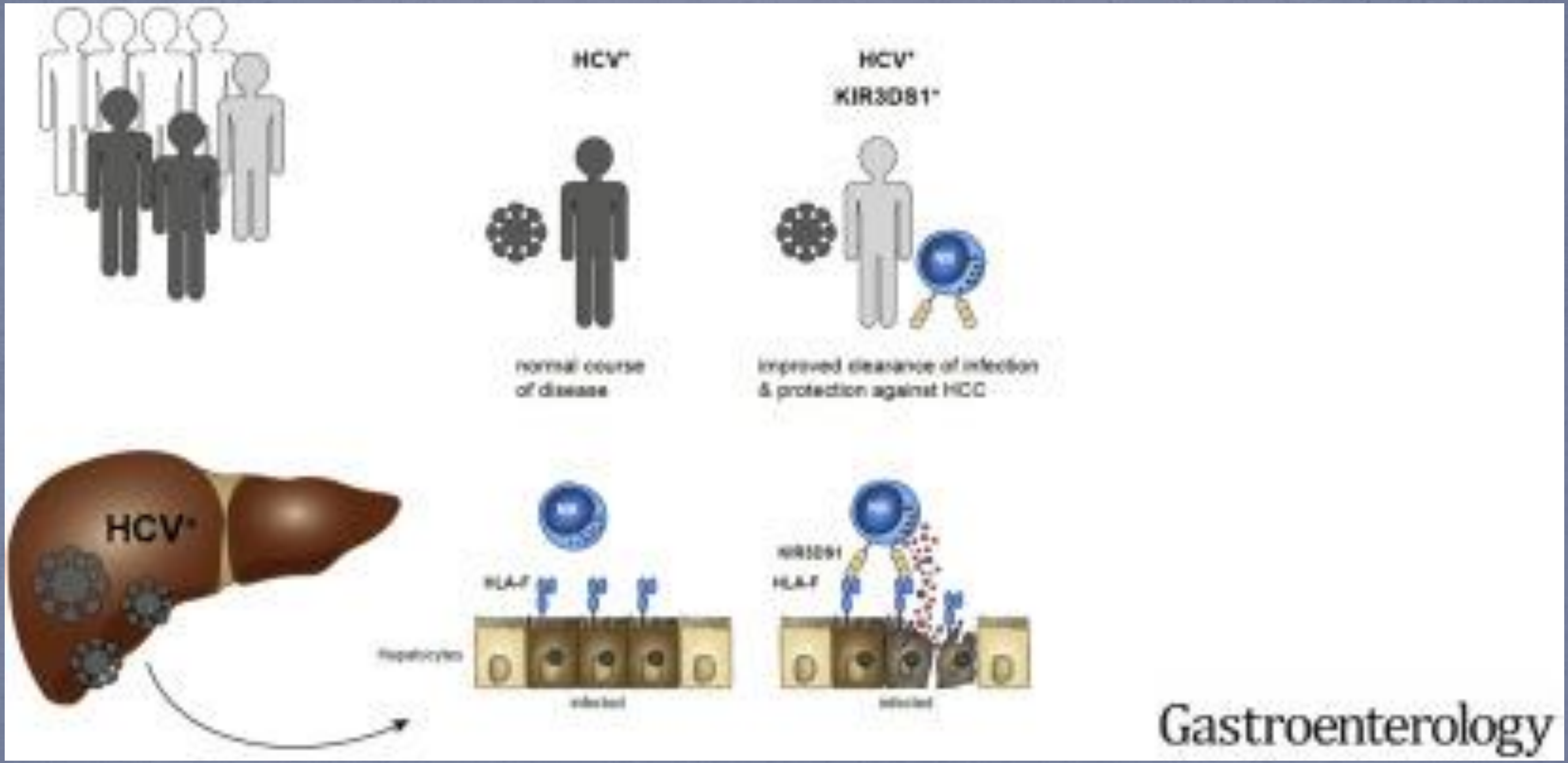
Reference sequence (segment =>1kb)	A	B	C	D	E	
Deletion (segment B was deleted)	A	C	D	E		
Duplication (segment C was duplicated)	A	B	C	C	D	E
Inversion (segment D was inverted)	A	B	C	D	E	

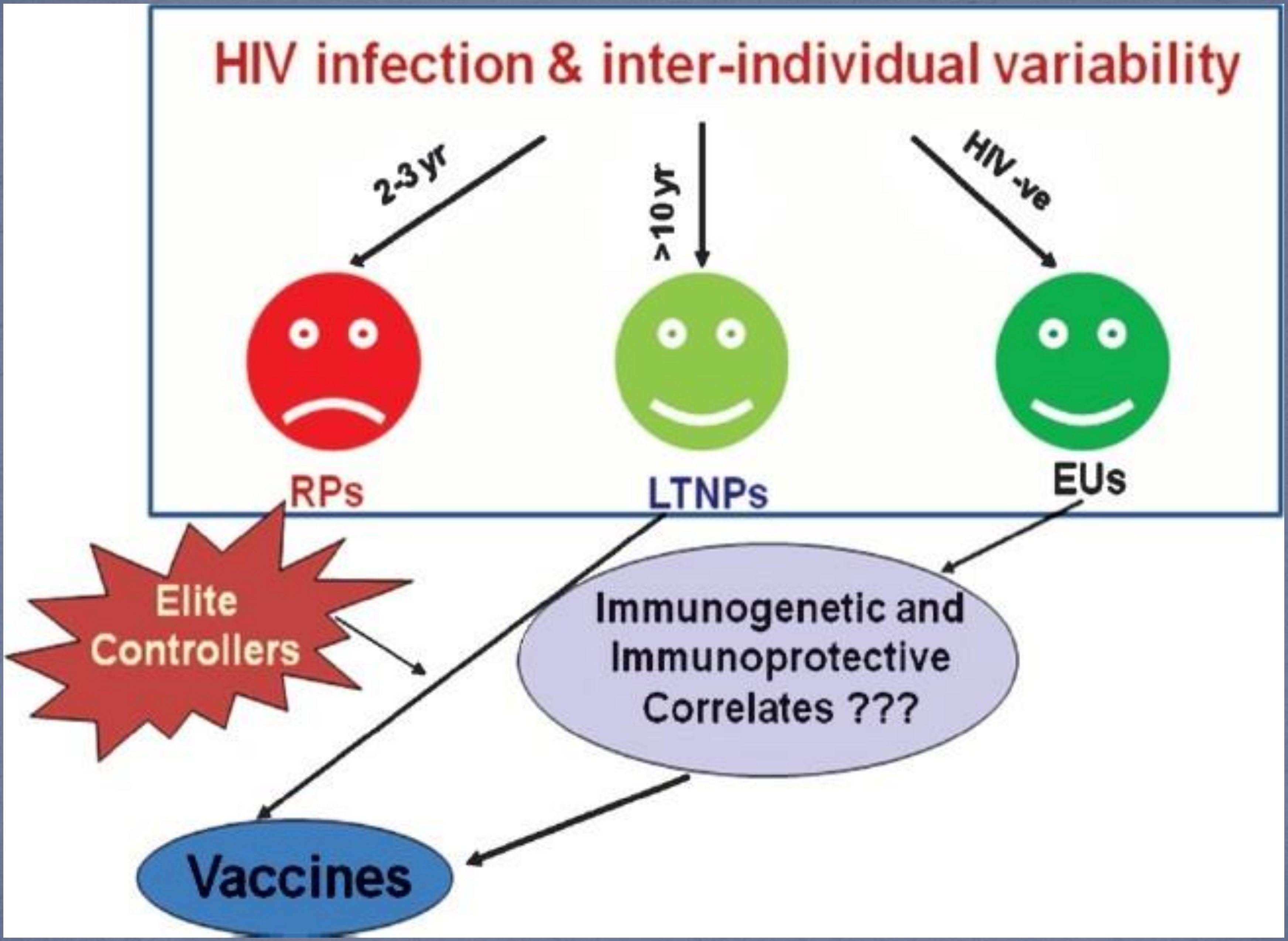
Reciprocal translocation

A B

B A

Role of KIR3DS1 ?





HIV infection & inter-individual variability

2-3 yr

>10 yr

HIV -ve

RPs

LTNPs

EUs

Elite Controllers

Immunogenetic and Immunoprotective Correlates ???

Vaccines

Types of Drug Responses

- *Inter individual variability in drug response* is defined as an “effect of varying intensity occurring in different individuals receiving a specified drug dose,” or “requirement of a range of doses (concentrations) in order to produce an effect of specified intensity in each patient”

Classifications of drug response

- (1) the desired beneficial effect (*efficacy*);
- (2) *adverse effect*;
- (3) no effect (*therapeutic failure*); and
- (4) *toxic effect*.
  - The latter two effects are particularly **dependent on drug dosage.**
  - Dose-*in*dependent ADRs can happen at any dosage and are generally unpredictable.

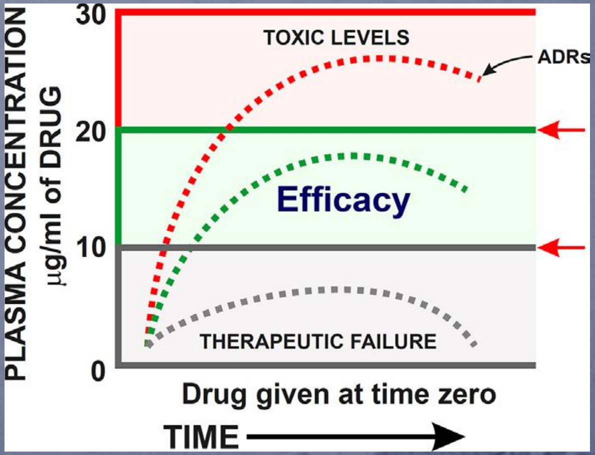
Classifications of drug response

- (1) the desired beneficial effect (*efficacy*);
- (2) *adverse effect*;
- (3) no effect (*therapeutic failure*); and
- (4) *toxic effect*.
  - The latter two effects are particularly **dependent on drug dosage.**
  - Dose-*in*dependent ADRs can happen at any dosage and are generally unpredictable.

- if the *parent drug* is the component responsible for efficacy as well as toxicity.
- If a *metabolite* is the *active principle* (i.e., that which causes the efficacy) and it is also the toxicant, then one can replace the word "drug" with "metabolite" and describe the same events leading to *efficacy* versus *therapeutic failure* versus *toxic levels*
- A classic example in which the metabolite is more biologically active than the parent drug is the parent drug codeine, which becomes activated by the CYP2D6 enzyme into the more biologically active metabolite, morphine

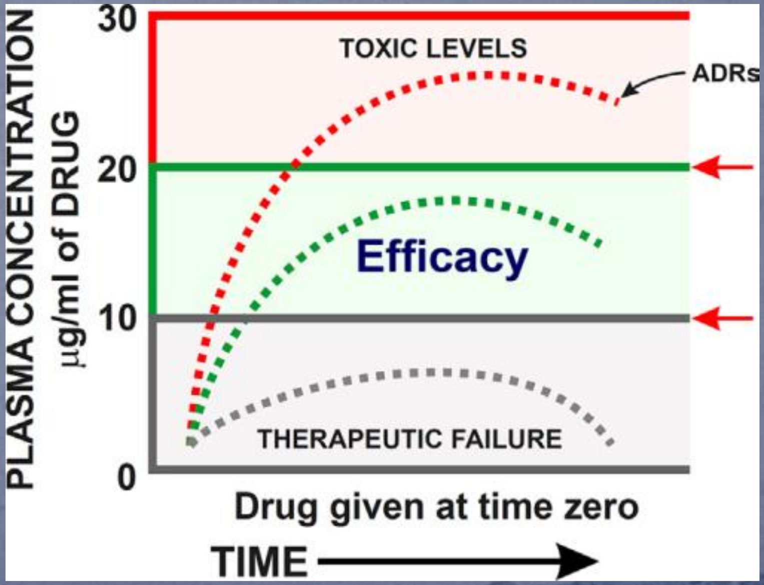
Therapeutic failure

- Treating a patient with a drug, is to maintain optimal plasma levels of the *active principle* (drug) in the *therapeutic range*.
- If the dose of drug is **too small**, the interval of administration inadequate, bioavailability of the active drug too low, or the active drug too rapidly etabolized and thus quickly cleared—then the active drug level in the blood might never reach its effective oncentration.
- This would lead to absence of the expected response (*therapeutic failure*).



Toxic effect

- If the dose of drug is **too large**, intervals of administration too short, or the active drug poorly metabolized and thus too slowly excreted, this accumulation can lead to ***toxic concentrations***.
- Levels that cause toxicity will lead to ADRs; such a drug response might occur because of drug accumulation at toxic levels in blood, and accumulation might occur in one or more critical target organs, or in both blood and target organs.



Adverse Drug Reactions (ADRs)

- (1) dose-dependent;
- (2) dose-independent;
- (3) dose- and time-dependent (cumulative); and
- (4) time-related withdrawal reactions +
- Dose-independent ADRs comprise **idiosyncratic drug** reactions and allergic reactions; +
- A better understanding of PGx should help reduce morbidity and mortality caused by ADRs, but should especially help prevent ADRs caused by *idiosyncratic dose-**in**dependent drug reactions*.

- Certain patients develop harmful adverse drug reactions (ADRs) after taking a medication
- Undesired reactions to a drug or its metabolite(s) can potentially be serious and even life threatening.
- According to the FDA Adverse Event Reporting System (FAERS), over one million cases of ADRs were observed in 2014 and this number is steadily increasing as reporting systems become more accessible to physicians
- Two main classifications of ADRs:
- ***Predictable ADRs*** occur due to the pharmacological activity of a drug or its metabolites
- ***idiosyncratic ADRs*** are primarily observed as an immune system response
- The major biological pathway capable of triggering such idiosyncratic ADRs is activated by a drug's direct binding with human leukocyte antigen (HLA) protein variants