Bioinformatics for High Throughput Sequencing

HSE-2025/2026

Materials

Course chat

Course materials

German Ashniev @SciKey

Game Rules

HW == Homework
(from 0 to 10)

Quiz == Exam
(from 0 to 10)

$$\textbf{GRADE} = \underbrace{\textbf{HW1} * 0.2 + \textbf{HW2} * 0.2 + \textbf{HW3} * 0.2}_{6} + \textbf{Quiz} * 0.4$$

```
if grade.count() < 3:
    IdontCare = True
    return("See You on retakes")
```

HW1 and HW2 deadline
1 week after assignment

HW3 deadline
2 weeks after assignment

*There are always "good" and "bad" options

Оценка по 10-балльной шкале	**Оценка по 5-балльной шкале за экзамен**	**Оценка в приложении к диплому НИУ ВШЭ**		**Оценка за зачет**
10	отлично (существенно превосходит ожидания)	A ++	Excellent	зачтено
9	отлично (превосходит ожидания)	A +	Very good	зачтено
8	отлично	A	Very good	зачтено
7	хорошо	B +	Good	зачтено
6	хорошо	B -	Good	зачтено
5	удовлетворительно	C +	Satisfactory	зачтено
4	удовлетворительно	C -	Satisfactory	зачтено
3	неудовлетворительно	F	Fail	не зачтено
2	неудовлетворительно	F	Fail	не зачтено
1	неудовлетворительно	F	Fail	не зачтено

Использование оценки "0"

Преподаватель имеет право поставить оценку "0" в следующих случаях:

- если студент не приступал к выполнению элемента контроля на занятии или в период сессии (например: сдал пустой лист, отказался от ответа, не явился на экзамен без уважительной причины);
- при обнаружении нарушений академических норм*, предусмотренных Правилами внутреннего распорядка обучающихся НИУ ВШЭ, таких как: списывание при выполнении письменной работы или при подготовке к ответу в устной форме, двойная сдача письменных работ, наличие плагиата в письменных работах, совершение подлога при выполнении письменных и устных работ, фабрикация данных и/или результатов работы, использование подсказок, применение технических средств для выполнения письменных или устных работ;
- незадекларированное использование генеративных моделей в соответствии с Регламентом организации проверки письменных учебных работ на наличие плагиата, использования генеративных моделей и размещения выпускных квалификационных работ обучающихся по программам бакалавриата, специалитета и магистратуры на корпоративном сайте (портале) НИУ ВШЭ;
- в иных случаях, установленных Положением об организации промежуточной аттестации и текущего контроля успеваемости студентов НИУ ВШЭ (например: использование материалов, запрещенных преподавателем, попытка общения с иными лицами, несанкционированные перемещения студентов и т.п.).

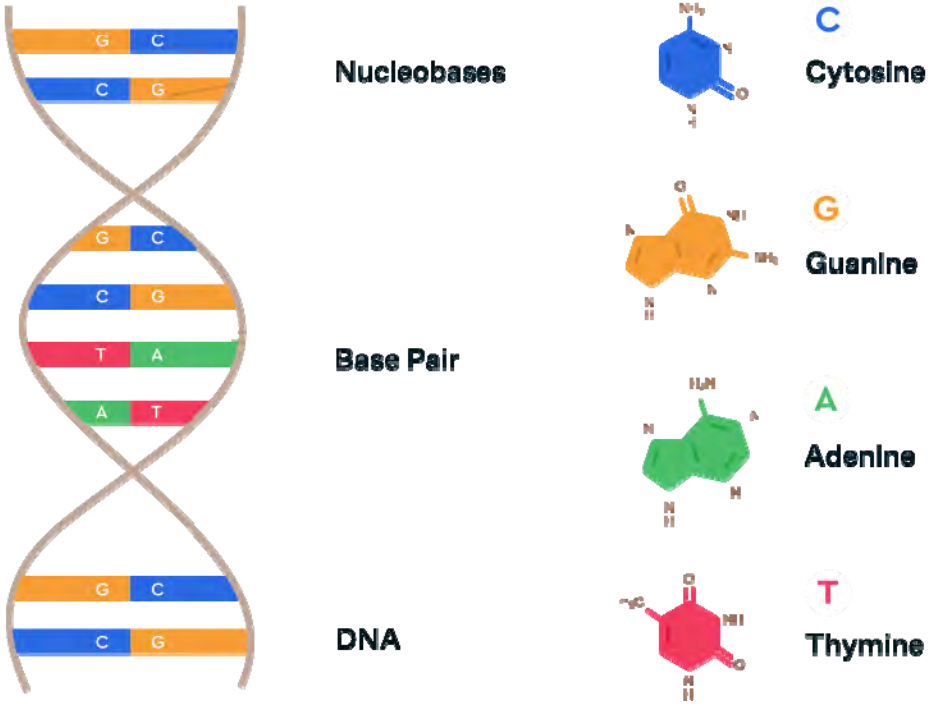
IV МОДУЛЬ 2025	Тема Занятия Подробно
04.04.2025	1) Overview lecture with introduction and course structure. 2) Rules. 3) Applications of Sequencing in various fields of knowledge.
11.04.2025	1) Evolution of DNA Sequencing Methods - Historical milestones - Comparative analysis: Throughput, accuracy, read length, and cost trends over time. - Emerging technologies: Single-molecule sequencing and epigenetic applications. 2) Sequencing Data File Formats - BCL - FASTQ - FAST5 3) Platform-Specific Data Structures 4) Sequencing Quality Assessment & Error Correction: - FastQC - MultiQC - Error Correction Tools 5) Read Alignment & Coverage Analysis - Short reads: BWA-MEM, Bowtie2. - Long reads: Minimap2, NGMLR. - samtools depth, mosdepth. 6) Reference Genomes and Model organisms 7) Mutation Databases - ClinVar, COSMIC, gnomAD, dbSNP 8) NGS Platform Errors & Comparative Analysis - Substitution errors. - PCR duplicates. - Homopolymers. - High indel rates.
18.04.2025	Introduction to Linux for Bioinformatics - Why Linux? - Core commands - Hands-on examples
25.04.2025	1) Major Steps in Sample Preparation for WGS and WES. 2) Required Equipment and Reagents. 3) DNA and RNA Extraction Methods. 4) Quality assessment. 5) Nucleic Acid Purification and Ribosomal RNA (rRNA) Depletion. 6) Library Preparation (Fragmentation and Tagmentation).

Course plan part_1

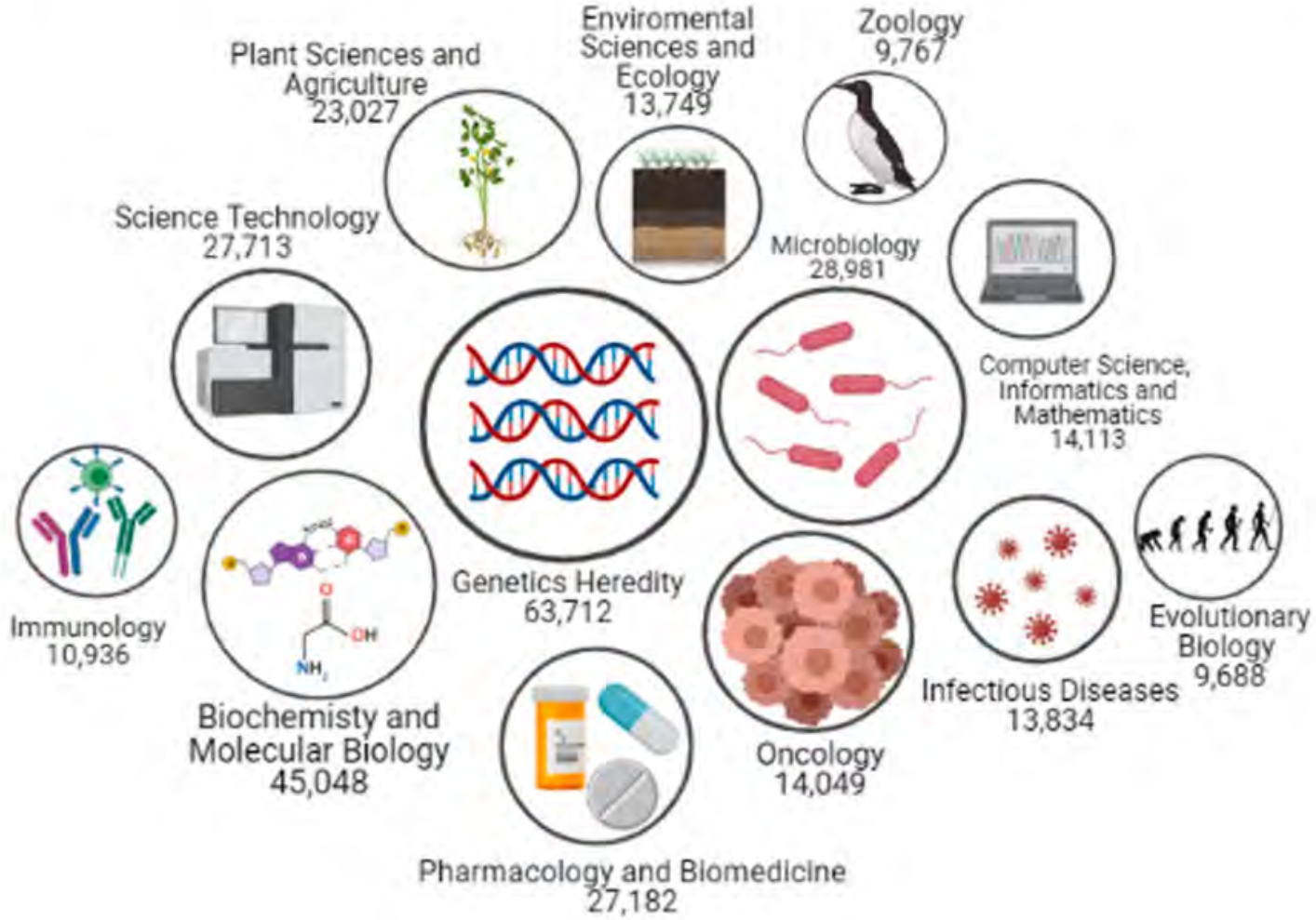
02.05.2025	
09.05.2025	1) Sanger Sequencing (First-Generation Sequencing) 2) Next-Generation Sequencing (NGS) Platforms: - Solexa (Pre-2006) Sequencing - Illumina (Post-2006) Sequencing - Ion Torrent (Thermo Fisher Scientific)
16.05.2025	- 454 Sequencing (Roche) - SOLiD Sequencing (Thermo Fisher Scientific) 3) Third-Generation Sequencing (Single-Molecule Sequencing)
23.05.2025	- PacBio SMRT Sequencing - Oxford Nanopore Sequencing 4) Emerging Sequencing Technologies - Helicos Biosciences (True Single-Molecule Sequencing)
30.05.2025	- BGI (DNBSEQ) - Quantum Sequencing (Quantapore) - Single-Molecule Fluorescence Sequencing (Genia, Stratos Genomics)
06.06.2025	Targeted sequencing methods and clinical applications (Exome, panels). Interesting cases from publications.
13.06.2025	**Экзамен**

Course plan part_2

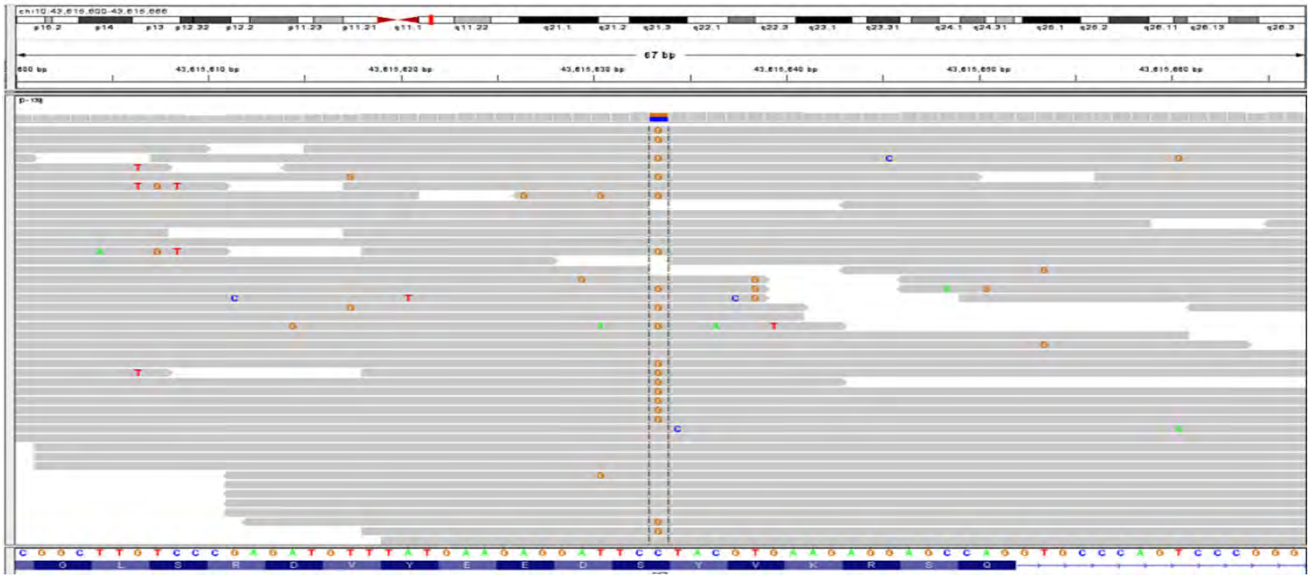
DNA sequencing - a laboratory technique in order determine the exact sequence of nucleotides, or bases, in a DNA molecule.



NGS Research Fields and Publication Distribution



Sequencing Application Examples- Inherited Conditions



Discovery tool: Single gene disorders
i.e. AD – Kabuki syndrome (MLL)

Causative mutations for multigenic diseases –superior to "one by one" approach of traditional sequencing

Diagnostic advancements for diseases with overlapping symptoms, multiple possible syndromes/genes

Inherited Conditions- Challenges and Opportunities

Challenges

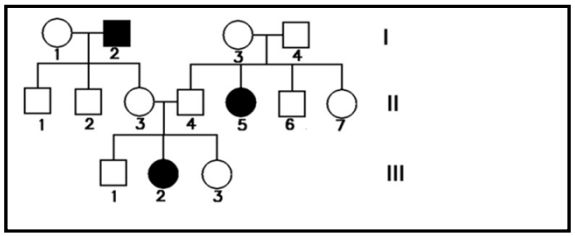
Example:
Monogenic disorders

- Novel missense mutations
- Germ line mosaicism
- Structural aberrations
- Imprinting effects
- Epigenetic factors

Opportunities

Example:
Multifactorial disease

- Risk loci more often in non-coding or inter-gene regions
- Pathogenicity of variants often unclear- less testing vs. monogenic disease
- Reference human genome cataloguing of variants = more test offerings



Sequencing Application Examples- Neoplastic Conditions

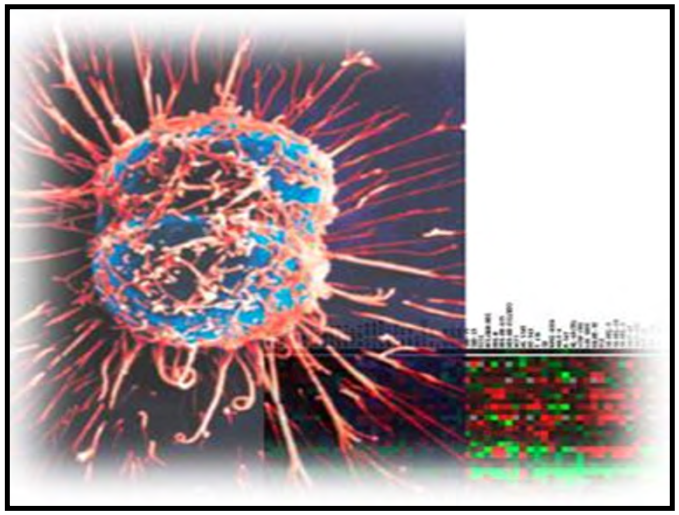
Cancer susceptibility genes

Risk assessment
Risk management

Somatic/driver mutations

Micro-RNAs

Methylation
Epigenetic changes



Alterations in gene expression

Molecular profiling

Tumor sub-typing

Patient stratification

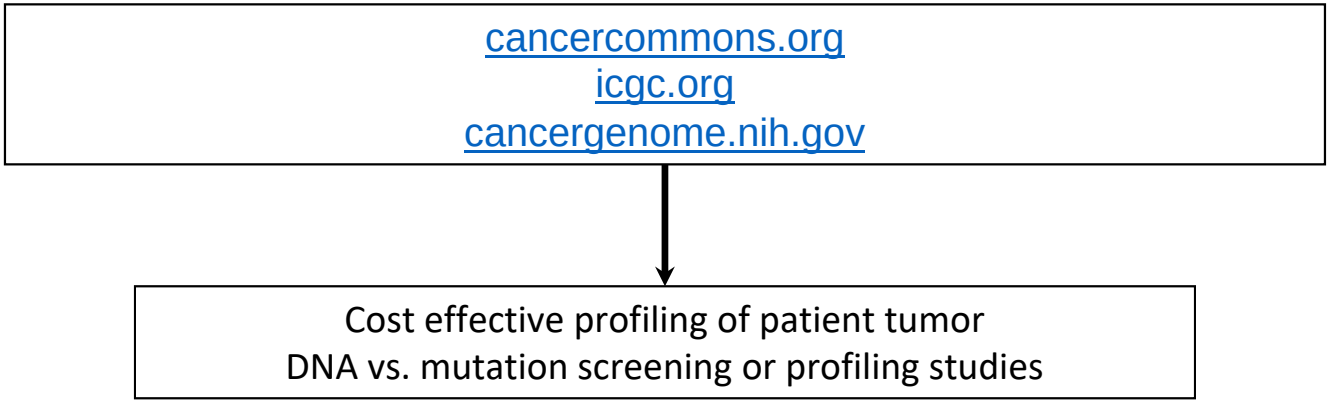
Predictions of therapeutic response during personalized treatment

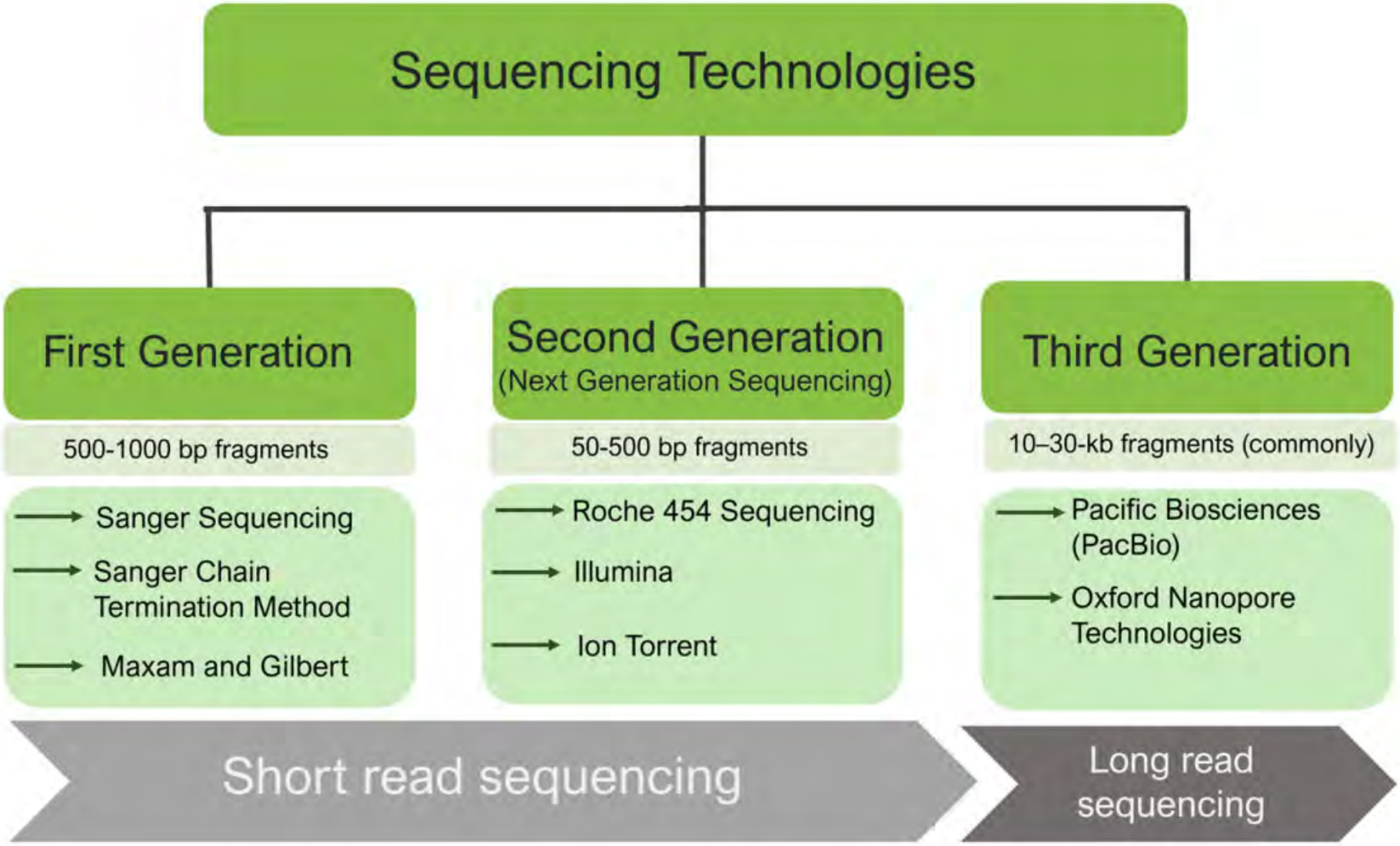
Therapeutic monitoring

Prognosis

Sequencing Application Examples- Neoplastic Conditions

- Mutation panel screening
- Exome and transcriptome screening
- Genome sequencing-comparison to normal tissue/reference sample




Sequencing Technologies
First Generation
500-1000 bp fragments
Sanger Sequencing
Sanger Chain Termination Method
Maxam and Gilbert
Second Generation
(Next Generation Sequencing)
50-500 bp fragments
Roche 454 Sequencing
Illumina
Ion Torrent
Third Generation
10–30-kb fragments (commonly)
Pacific Biosciences (PacBio)
Oxford Nanopore Technologies
Short read sequencing
Long read sequencing

PacBio®, 2021

Sequencing Application Examples- Other Considerations

Different NGS platforms have different capabilities

RNA and DNA sequence changes

DNA copy number variations

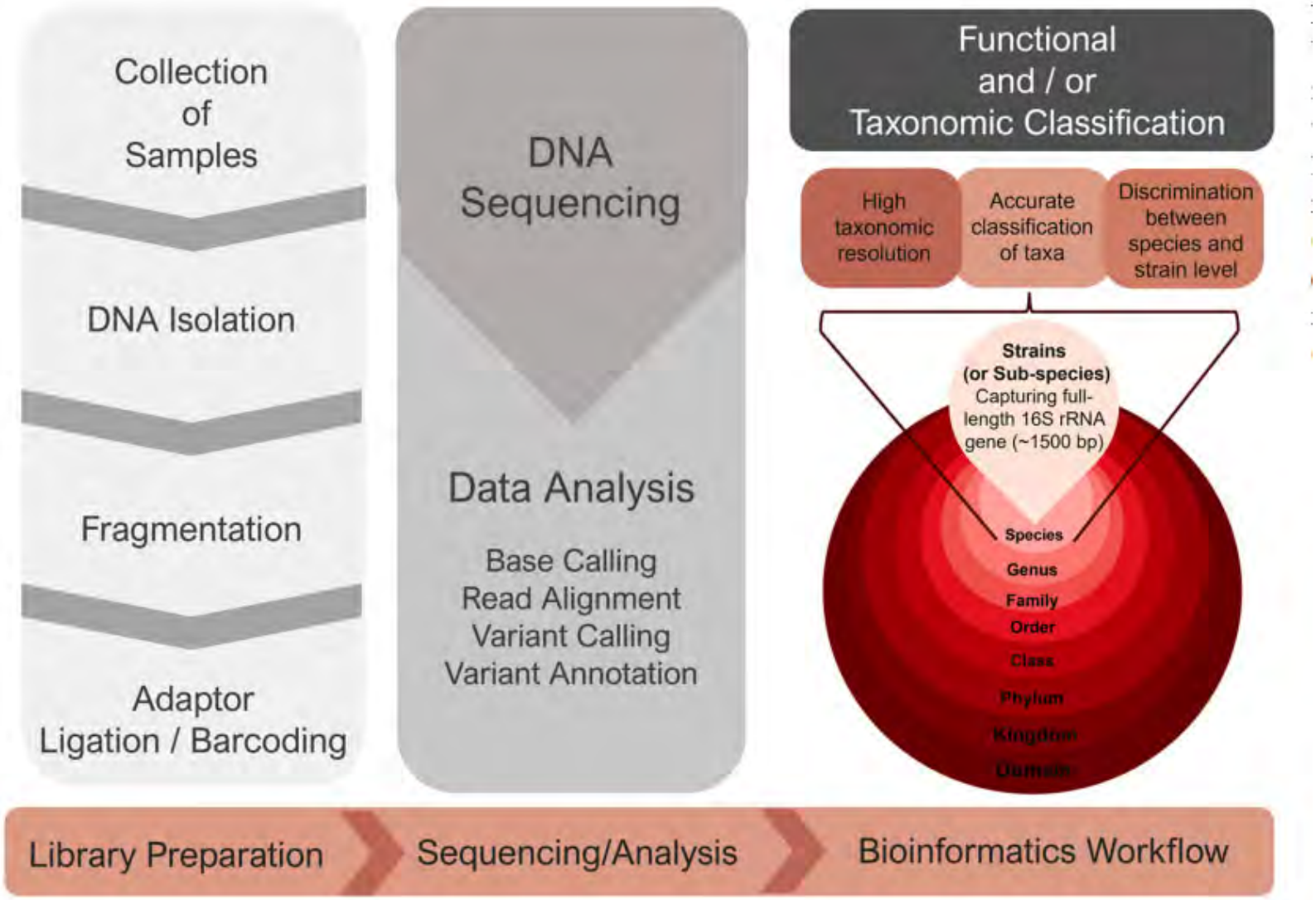
DNA rearrangements

RNA expression profiles

Methylation

↓

A single method usually provides only part of this variety of information - cost , specimen type, and application considerations important


Collection of Samples
DNA Isolation
Fragmentation
Adaptor Ligation / Barcoding
DNA Sequencing
Data Analysis
Base Calling
Read Alignment
Variant Calling
Variant Annotation
Functional and / or Taxonomic Classification
High taxonomic resolution
Accurate classification of taxa
Discrimination between species and strain level
Strains (or Sub-species)
Capturing full-length 16S rRNA gene (~1500 bp)
Species
Genus
Family
Order
Class
Phylum
Kingdom
Domain
Library Preparation
Sequencing/Analysis
Bioinformatics Workflow

NGS Application Examples- Other Considerations

NGS- significant false positive rate → **Mutation confirmation Usually by Sanger sequencing-will platform evolution eliminate?**

Variable % tumor cells and variable % tumor cells with (presumably) secondary mutation → **May overlap with NGS false positive rate**

Low level mutations- not easily confirmed by Sanger sequencing (higher detection threshold ≈ 15-20%) without more sensitive mutation screening - DGGE, dHPLC, pyrosequencing or mutation enrichment- i.e. COLD PCR

Numerous heterogeneous aberrations- i.e. oncologic applications need algorithm development

Clinical Utility

- Balance of net health benefits vs. harm
- NGS –transformative for personalized treatment of disease
- Clinical indication - includes test rationale, patient population and clinical scenarios
- Principles of comparative effectiveness- requires individualized evidence-based approach for each patient

Clinical Utility-Challenges

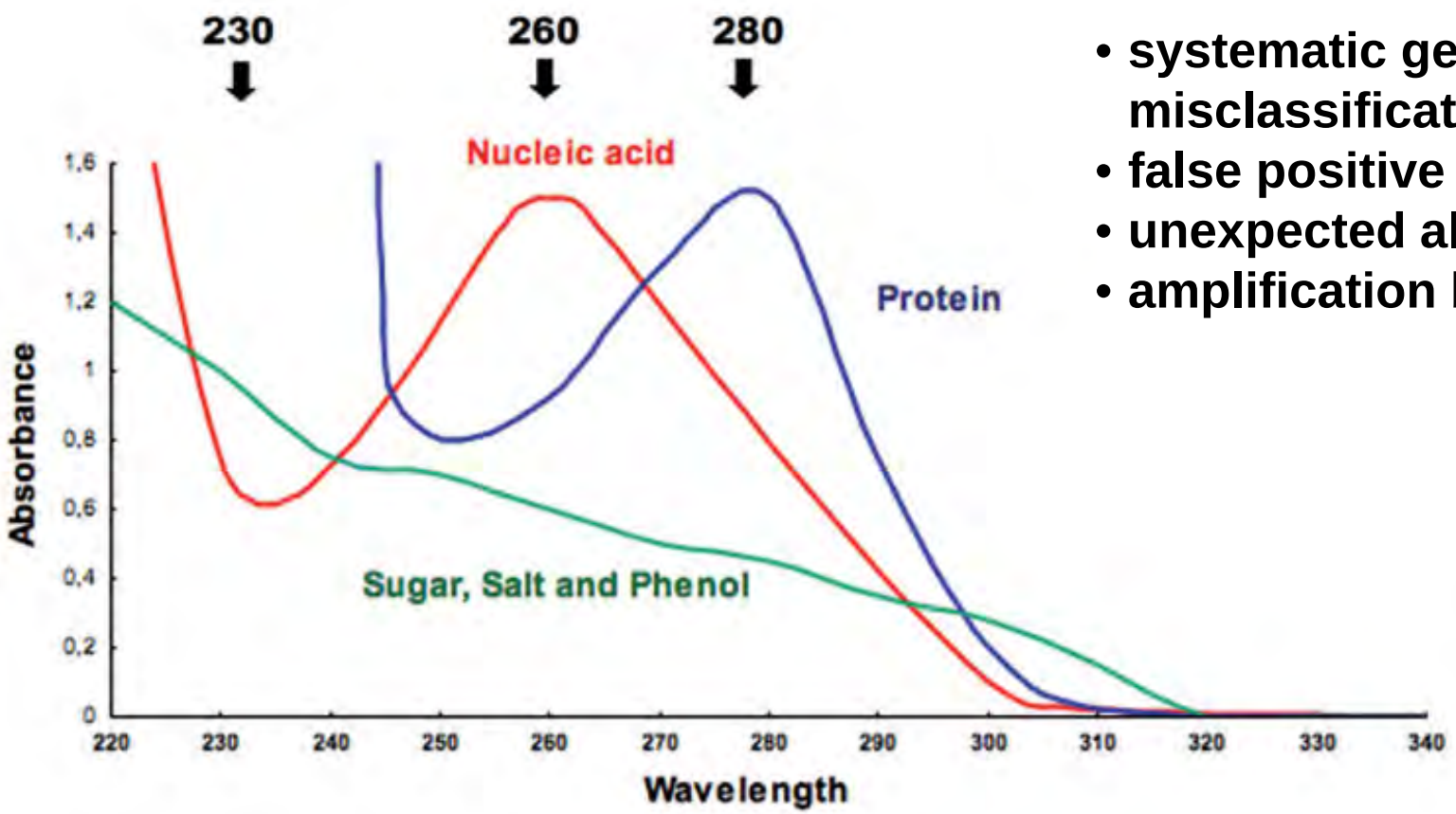
NGS data density = frequently encountered variants of unknown significance

- Which variants are clinically actionable?
- Development of evidence-based scientific standards to evaluate utility in in different patient populations for accurate risk estimation
- Careful selection of patients for genome sequencing and genetic counseling-crucial

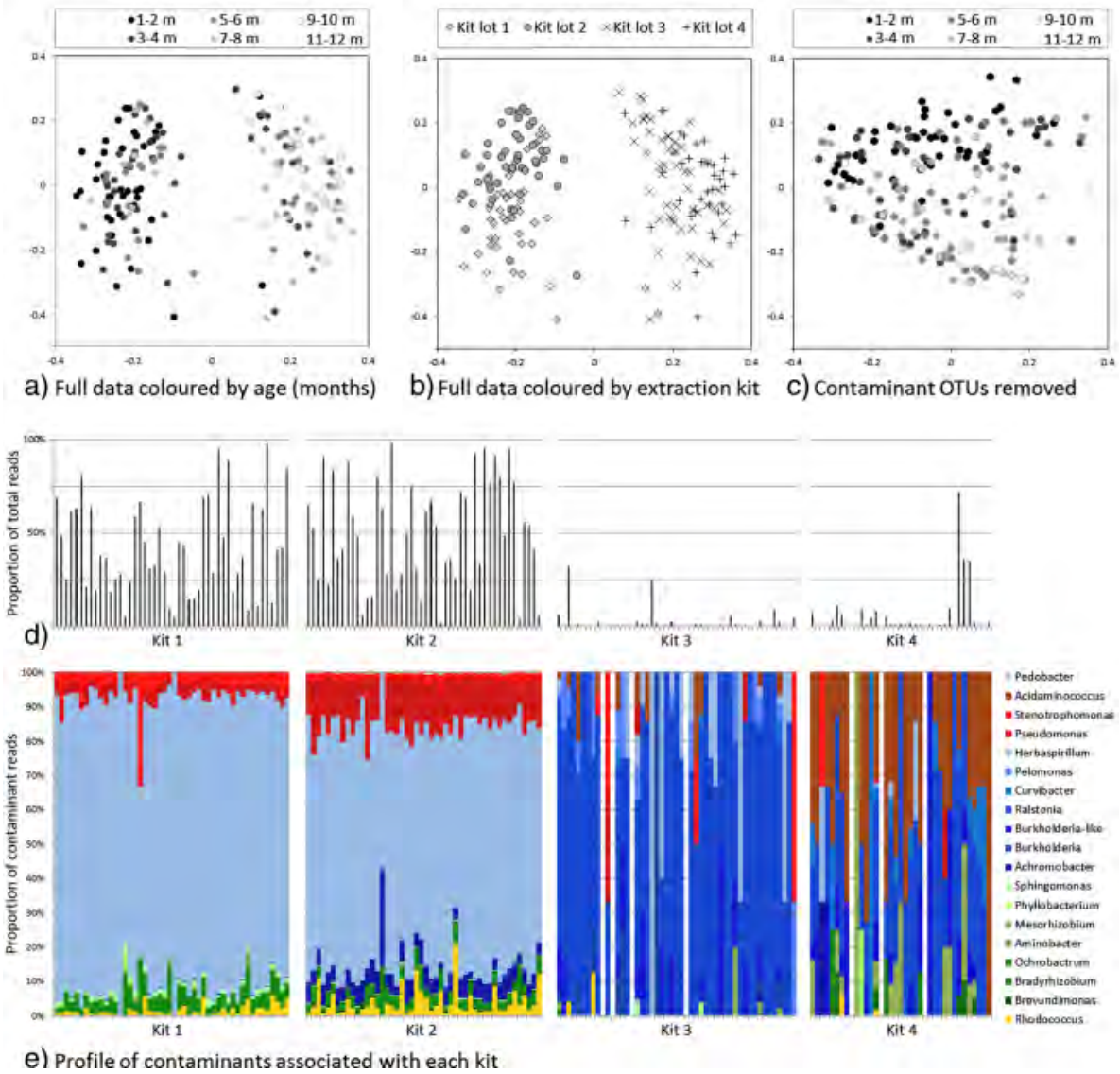
Sequencing. Main problems

- **Sample Contamination** (salts, proteins, and other chemicals)
- **Impure Template DNA** (contaminants from reagents or environmental sources)
- **Failed Sequencing Reactions** (universal primers may not work with certain plasmids)
- **Errors in Sequencing Data** (positive and negative errors)
- **Chimeric Fragments**
- **Computational Complexity** (NP-hard)
- **Inhibitors in Reagents** (inhibition of DNA polymerase)
- **Quality Control Issues** (labor-intensive)

Sample Contamination



- systematic genotype misclassification
- false positive associations
- unexpected allele ratios
- amplification bias



*Nasopharyngeal samples from Thailand.

Impure Template DNA

a) The PCoA plot appears to show age-related clustering; however,

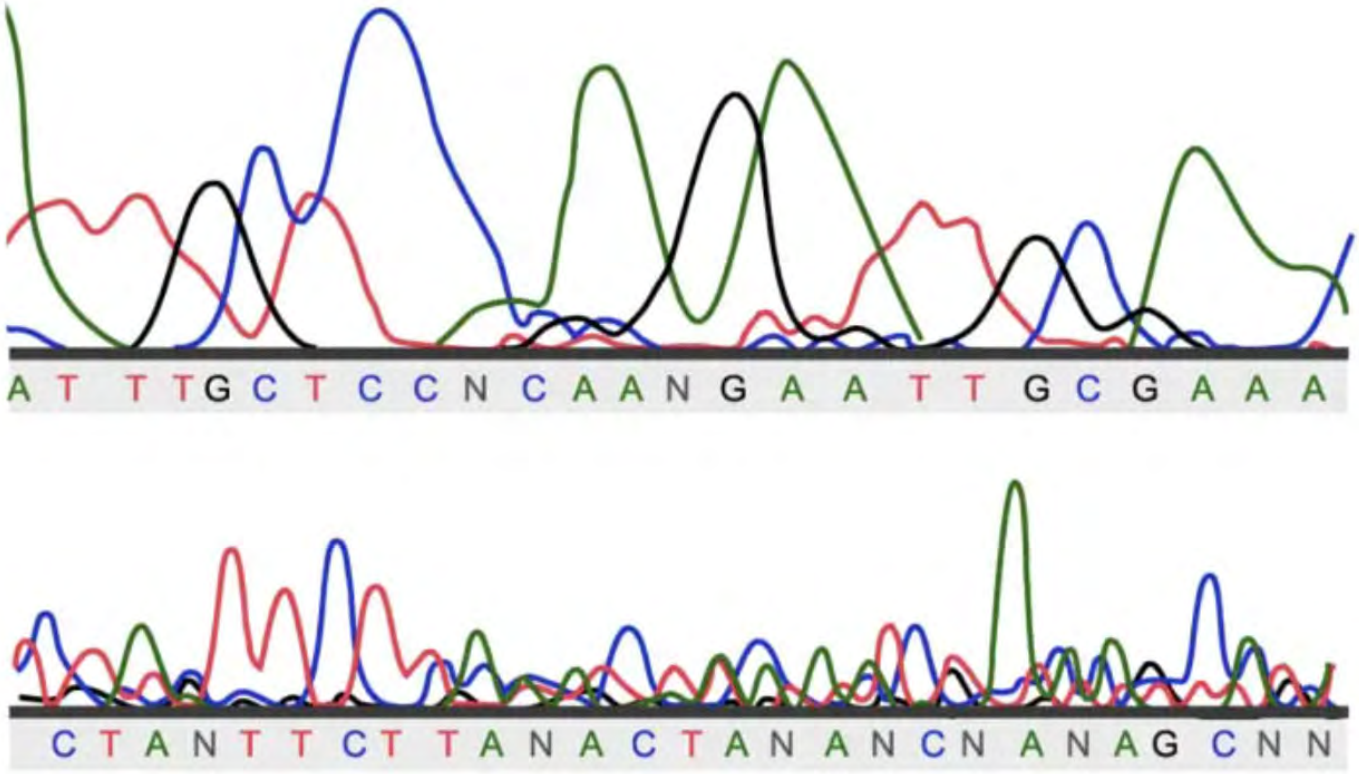
b) extraction kit lot explains the pattern better.

c) When coloured by age, the plot shows the loss of the initial clustering pattern after excluding contaminant OTUs from ordination.

d) The proportion of reads attributed to contaminant OTUs for each sample, demonstrating that the first two kits were the most heavily contaminated.

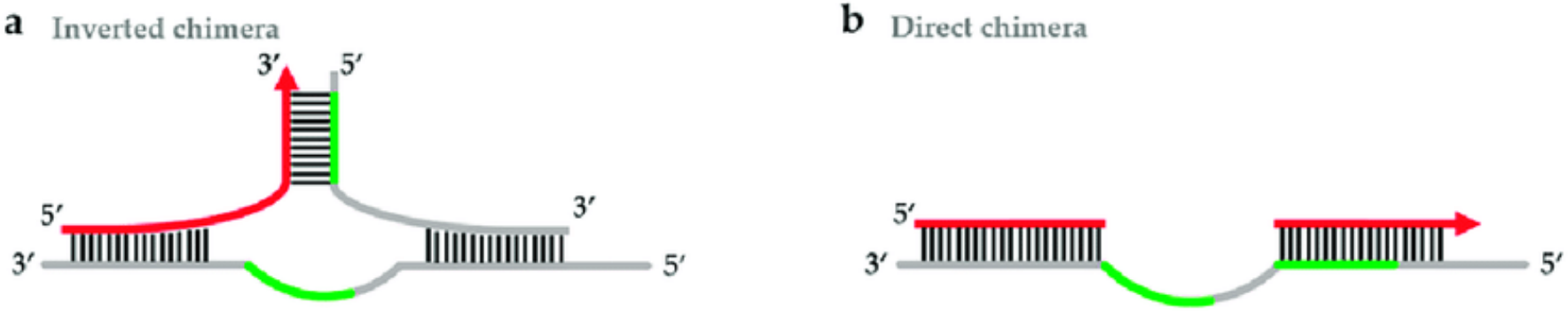
e) Genus-level profile of contaminant OTUs for each kit used.

Failed Sequencing Reactions
Errors in Sequencing Data



Severely low signal intensity due to hardware failure or a failed reaction

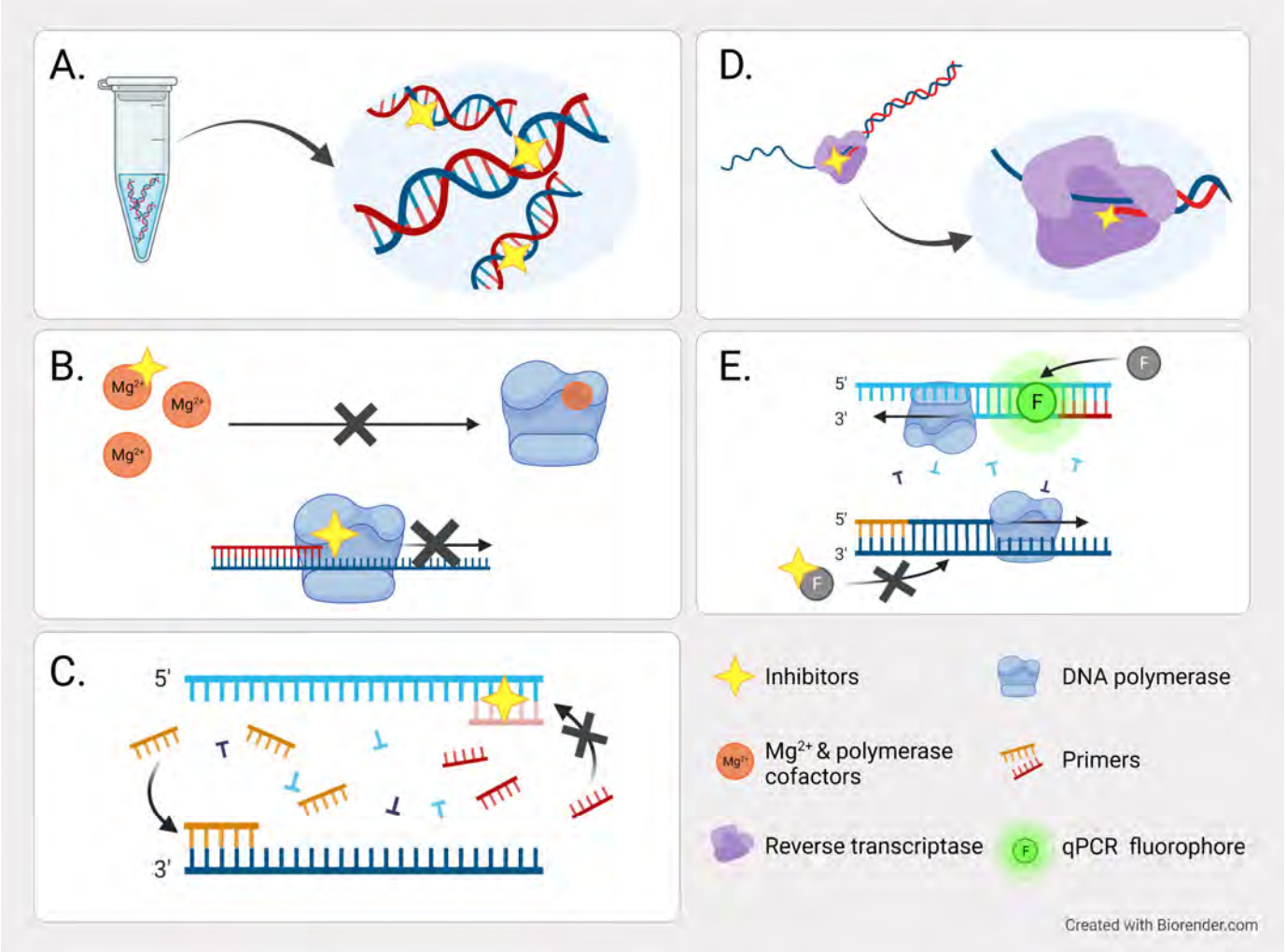
Chimeric Fragments



There was branch migration reaction during the MDA processing, because of there exists same sequence (green line in a,b)

- In **(a)**, the 3' end of a displaced strand attaches to another template. Here, the 3' end (marked by a red arrow) pairs with a sequence on the 5' strand (green line), leading to the formation of an **inverted chimera**.
- In **(b)**, the displaced 3' end pairs with a sequence on the 3' strand (green line), resulting in a **direct chimera**.

Inhibition of DNA polymerase



Informed Consent and Ethical Considerations

- Create patient awareness of benefits and harms
- No specific guidance exists- institutional policies vary
- Potential for anxiety and uncertainty exist especially for variants of unknown significance
- Discovery of incidental findings unrelated to the disease in question

Analytical Considerations-Regulation, Assay Validation, and Reference Materials

- Sequences are not truly complete – gaps in reads, GC rich regions, bioinformatics limitations with indel variant calling
- "gold standard" comparison- Sanger sequencing, however the technical capabilities are dwarfed by NGS
- Regardless - all NGS steps must be evaluated, and quality control metrics must be in place- is sequencing portions of a reference genome(s) sufficient?
- Development of reference materials (RMs) for meaningful validation is key