Challenges in analysis and interpretation of clinical genetic data using different NGS Platforms and sequencing assays

Austrian Institute of Technology AIT
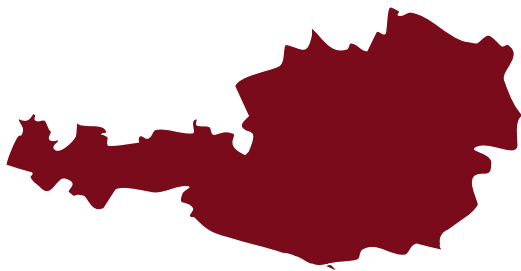
Stephan Pabinger

stephan.pabinger@ait.ac.at
@tadkeys

AIT - Austrian Institute of Technology

AUSTRIAN INSTITUTE
OF TECHNOLOGY

The largest applied research institute in **Austria**

Finance structure

- 30% contract research
- 30% co-financed research
- 40% Governmental funding

Owner structure

50.46%
Republic of Austria

49.54%
Federation of Austrian Industries

AIT Research Areas and Fields in Future Infrastructure Themes

Energy

Energy Infrastructure
- Smart Grids
- Smart Buildings
- Photovoltaics
- Thermal Energy Systems

Integrated Energy Systems
- Smart Cities and Regions
- Complex Energy Systems

Mobility

Transportation Infrastructure
- Environmentally-friendly transport infrastructure
- Cost-effective and resilient transport infrastructure
- Innovative road infrastructure safety strategies

Low-emission Transport
- High performance material
- Light-weight design of vehicle components
- Sustainable process

Multi-Modal Mobility Systems
- Human factors for personal mobility
- Integrated management of transport systems
- Real-time dynamic management of transportation systems

Safety & Security

Intelligent Vision Systems
- Multi- Camera Vision
- High-Speed Imaging

Future Networks and Services
- Advanced Applications in Sensor Networks
- Next-Generation Content Management Systems
- Secure Information Access in Distributed Systems

Highly Reliable Software and Systems
- Assessment and Testing of Autonomous and Safety-Critical Systems

Health & Environment

Biomedical & Biomolecular Health Solutions
- Preclinical and Clinical Diagnostics
- Molecular Diagnostics
- AAL Ambient Assisted Living
- Advanced Implant Solutions

Resource Exploitation and Management
- Exploitation of Biological Resources
- Microbial Detection
- Green Processes

Innovation Systems

Foresight & Governance
- New R&I Processes and Systems
- Anticipatory Governance

Technology Experience
- Contextual Experience
- Experience Foundations
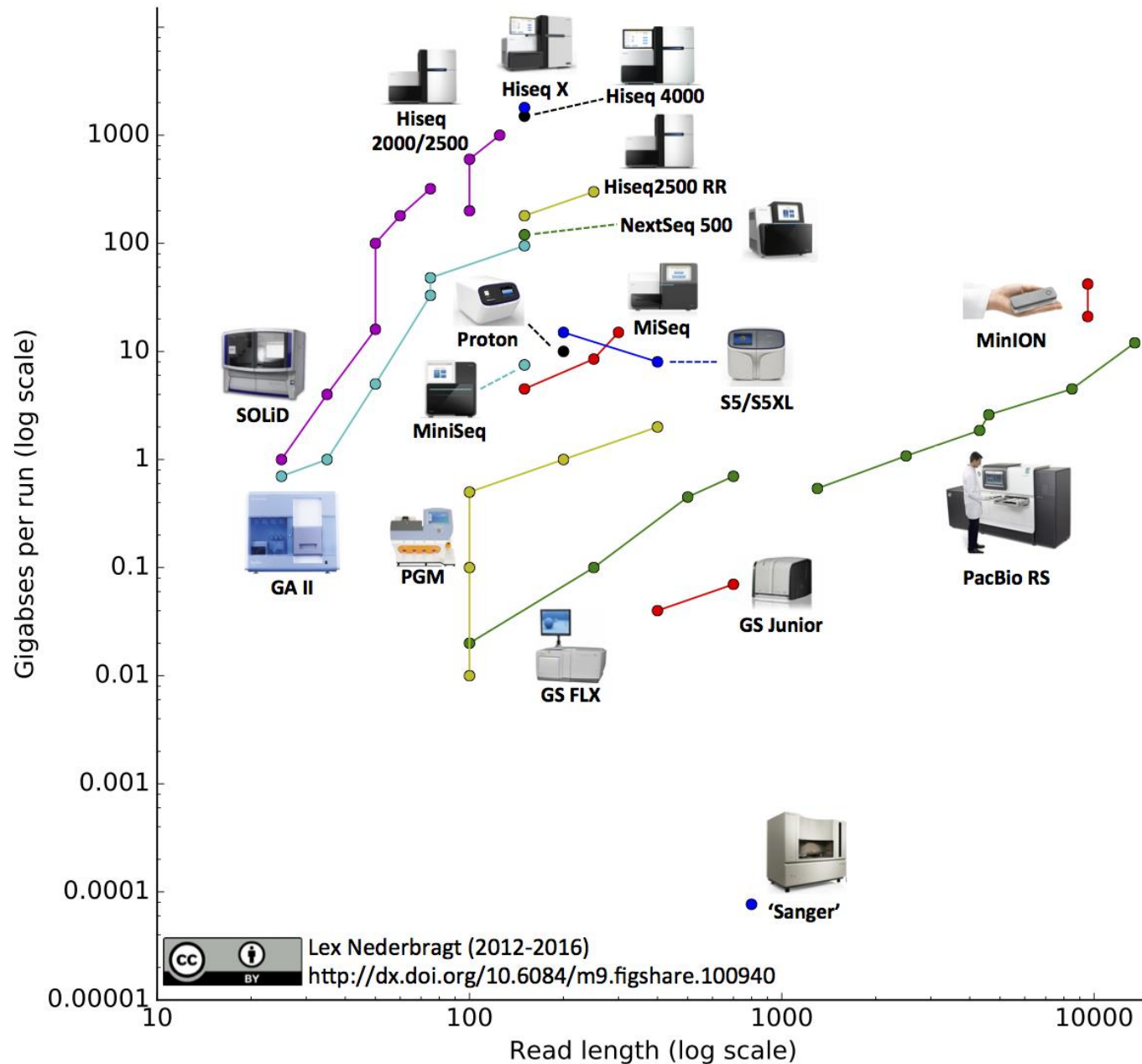
Identify effective ways for early diagnosis of diseases

Saliva diagnostics

Department

Research Area and Research Field

Challenges in analysis and interpretation of clinical genetic data using different NGS Platforms and sequencing assays

Principle

sequence library

Sequencer

Short sequence reads

Reference

Alignment

Variants
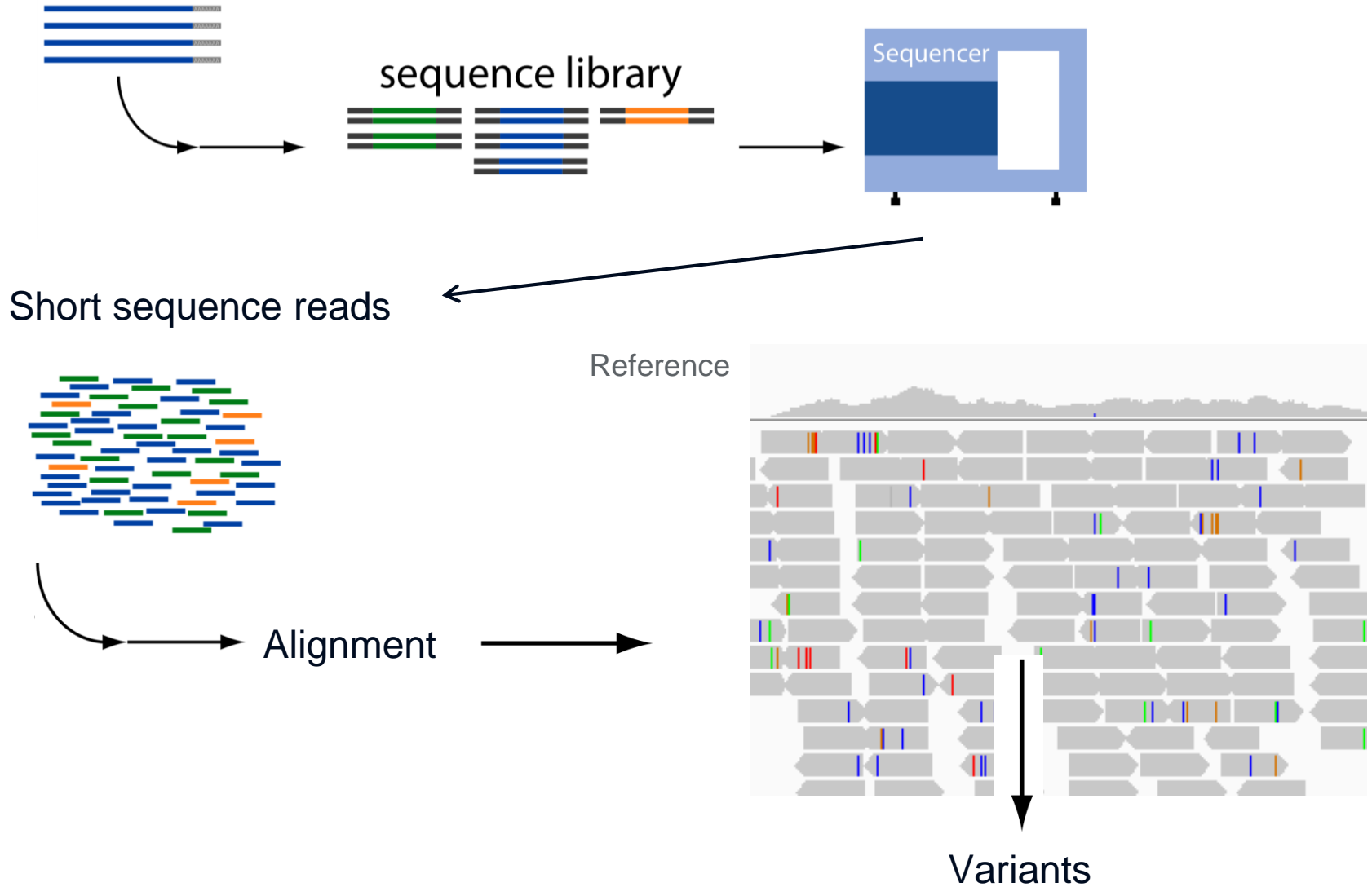
Properties of different technologies

Instrument	Amplification	Run time	Millions of Reads/run	Bases / read	Gbp/run
Illumina MiSeq	BridgePCR	5-55h	1-22	50-600	0.3-13.2
Illumina NextSeq 500	BridgePCR	11-30h	130-400	75-300	19.5-120
Illumina HiSeq 2500	BridgePCR	10h - 11days	300-2000	50-300	15-500
Ion Torrent - PGM	emPCR	2-7h	0.475-4.75	200-400	0.095-1.9
Ion Torrent - Proton	emPCR	4-6h	70-500	175	12.25-87.5
Pacific Biosciences RS II	None	2 hrs.	0,03	3000	0,09
Oxford Nanopore MinION (forecast)	None	≤6 hrs.	0,1	9000	0,9

http://www.molecularecologist.com/next-gen-fieldguide-2016/

Error Rates

Instrument	Primary Errors	Single-pass Error Rate (%)	Final Error Rate (%)
Illumina	Substitutions	~0.1	~0.1
Ion Torrent	INDELs	~1	~1
Oxford Nanopore	Deletions	≥4	4
PacBio RS	INDELs	~13	≤1

Sequencing Techniques

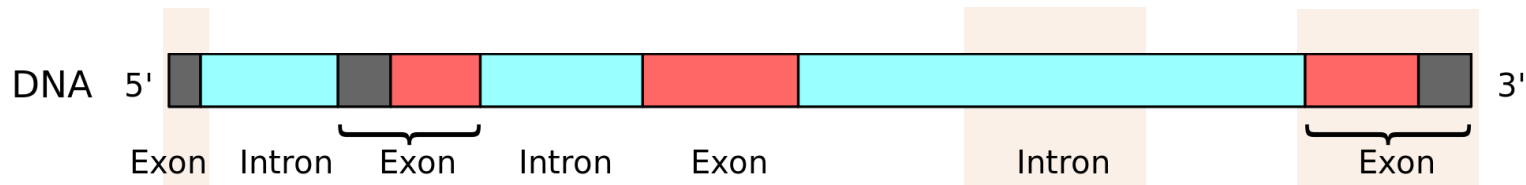DNA 5' Exon Intron Exon Intron Exon Intron Exon 3'
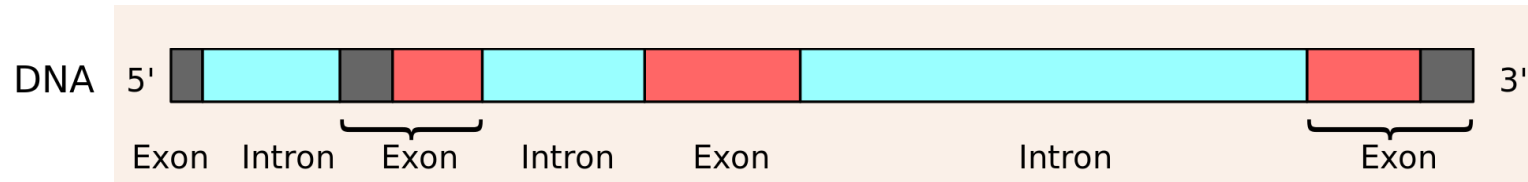
- Whole genome sequencing

Sequencing Techniques
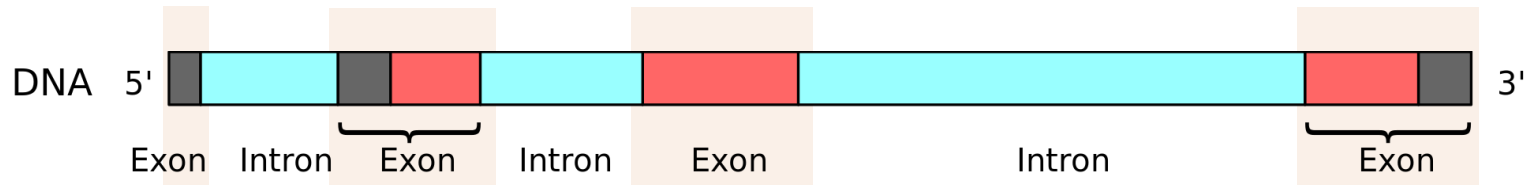
- Whole genome sequencing
- Whole exome sequencing
- Custom capture

Sequencing Techniques

- Whole genome sequencing
- Whole exome sequencing
- Custom capture
- Amplicon sequencing

What is the best technology for my use-case?

- Clinical question?
- Number of samples?
- Cost?
- Future strategies?

Challenges of current technologies

Amplification errors

- All polymerases have an inherent error rate (10^{-6} - 10^{-7})

GC bias

- PCR bias against GC rich sequences
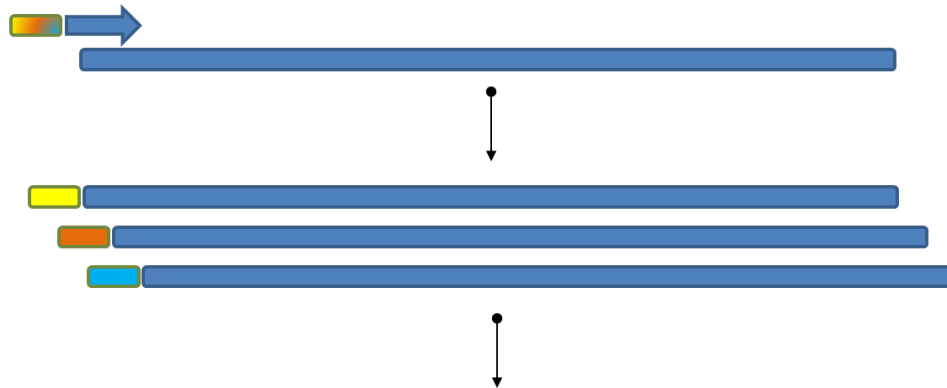- Exome capture bias against GC rich sequences

Trouble detecting small insertions and deletions

- Capture baits may not hybridize well
- Capture cannot be used to reliably detect large CNVs

Molecular barcoding

Library generation

STEP 1 - Barcoding

STEP 2 - Amplification

Consensus sequence generation

BC1 ATCGATCAGTCACGTAGGGTACCCGATTACCTTACAG**A**ATCCGATCCATTCGAAATCGGGA
BC1 ATCGA**C**CAGTCACGTAGGGTACCCGATTACCTTACAGGATCCGATCCATTCGAAATCGGGA
BC1 ATCGATCAGTCACGTAGGGTAC**G**CGATTACCTTACAGGATCCGATCCA**A**TCGAAATCGGGA
BC1 ATCGATCAGTCACGTAGGGTACCCGATTACCTTACAGGATCCGATCCATTCGAAATCG**C**GA

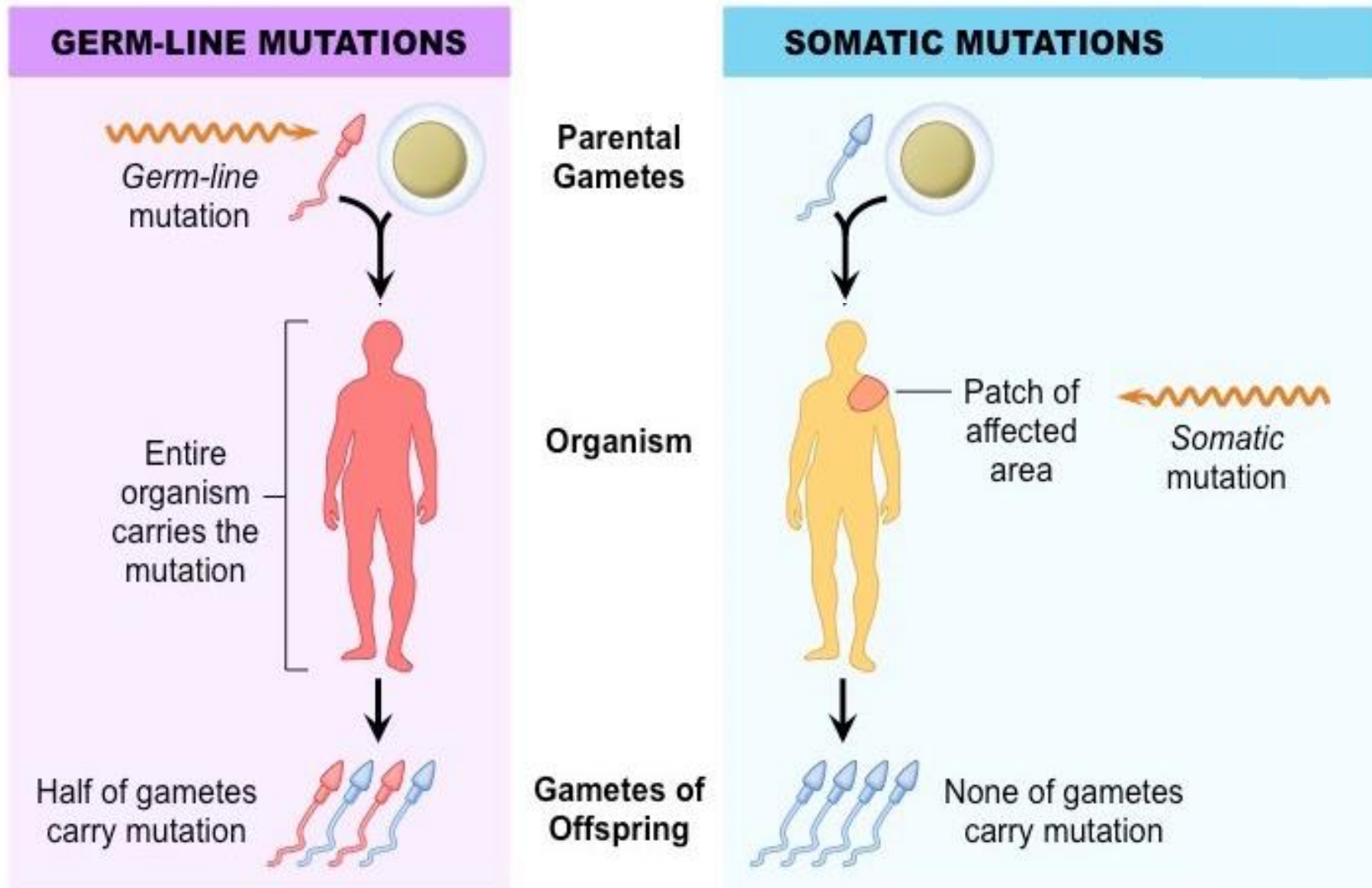ATCGATCAGTCACGTAGGGTACCCGATTACCTTACAGGATCCGATCCATTCGAAATCGGGA

random barcode mix   unique barcodes   sequencing adaptors
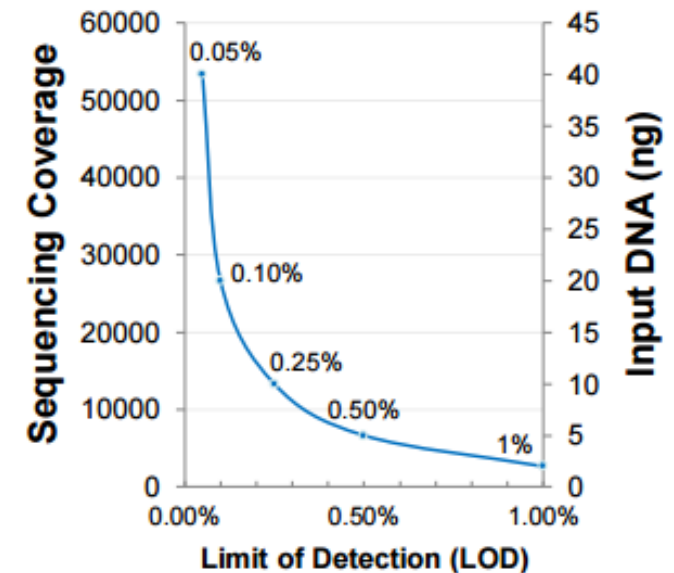
Types of variants

Somatic mutations – coverage considerations

Theoretical coverage

Number of cells	DNA (ng) Amount	Max Coverage	Sensitivity (4x Cov)
166.667	1000	333.333	0,001%
16.667	100	33.333	0,012%
6.667	40	13.333	0,03%
3.333	20	6.667	0,06%
1.667	10	3.333	0,12%
167	1	333	1,2%
17	0,1	33	12%

cell free DNA (Ion Torrent)

https://cofactorgenomics.com/heterogenous-dna-sequencing-lower-limits-minor-allele-frequency-sensitivity/
https://www.thermofisher.com/content/dam/LifeTech/Documents/PDFs/EACR-2016-Detection-of-somatic-mutations.pdf

Analysis challenges

QC

- Quality trimming / filtering – what cutoff?
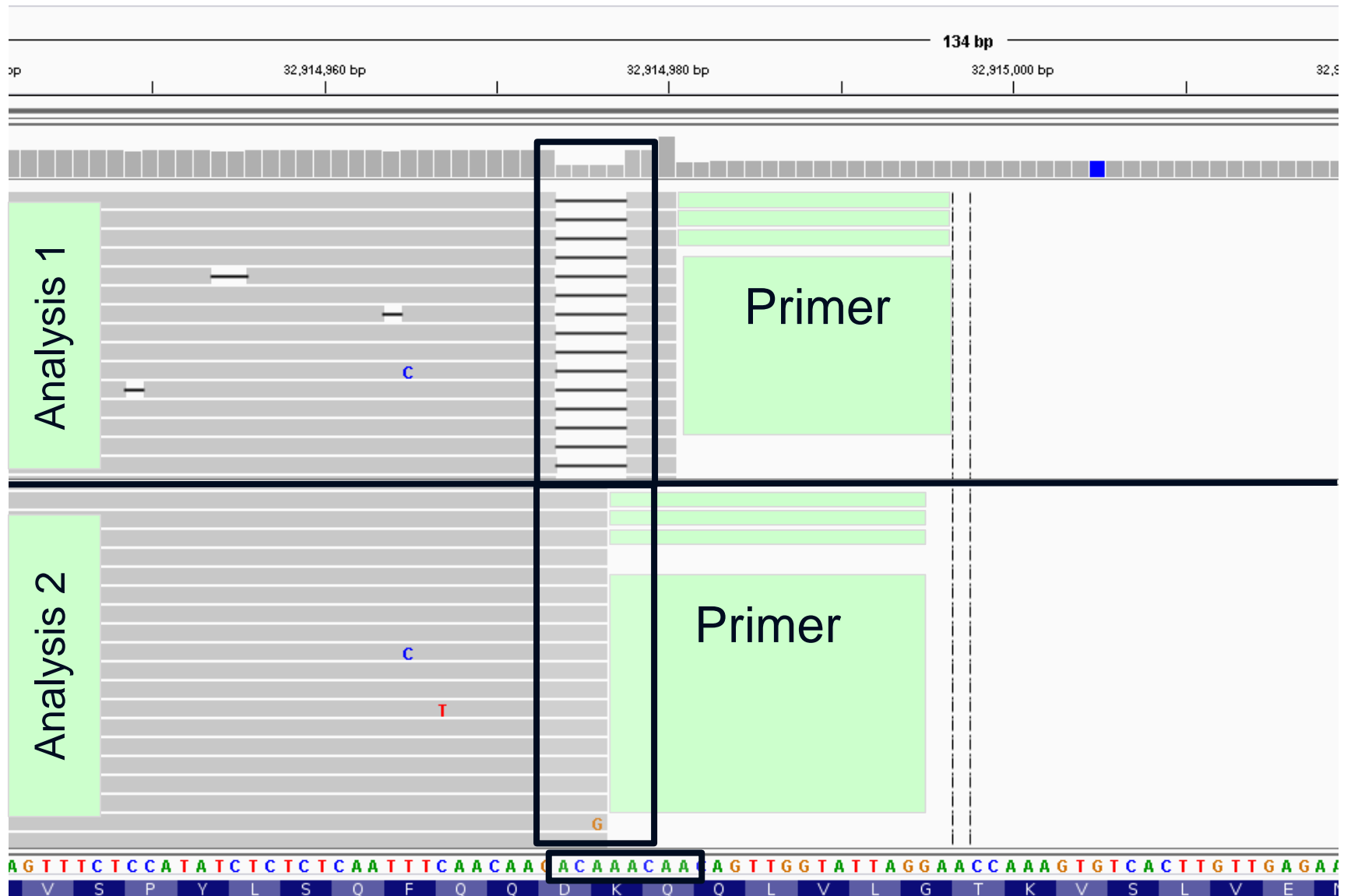- Correct primer sequences

Mapping

- Correct tool
- Choice of reference

Variant calling

- Which tool? Combine several?
- Germline, somatic
- Structural variations
- Parameters

Primer trimming

Primer trimming – before mapping

```
REF:                    AACAAGACAAACAACAGTTGGTATTAGGAA

Reads:                       AACAAGACAACAGTTGGTATTAGGAA
                             AACAAGACAACAGTTGGTATTAGGAA


Primer:                              ACAGTTGGTATTAGGAA


Read trimmed:           AACAAGACA


Alignment:  ref: AACAAGACAAACAACAGTTGGTATTAGGAA
                 AACAAGACA
```

→ **NO INDEL**

Primer trimming – after mapping

```
REF:              AACAAGACAAACAACAGTTGGTATTAGGAA

Reads:                 AACAAGACAACAGTTGGTATTAGGAA
                       AACAAGACAACAGTTGGTATTAGGAA

Alignment:  ref: AACAAGACAAACAACAGTTGGTATTAGGAA
          reads: AACAAG    ACAACAGTTGGTATTAGGAA

Primer:                        ACAGTTGGTATTAGGAA

Alignment:  ref: AACAAGACAAACAACAGTTGGTATTAGGAA
Trimmed reads  : AACAAG    ACA
```

→ **INDEL**

Interpretation of variants (technical)

Variants

- Check strand-bias
- Check coverage
- Homopolymer region

Analysis system

- Be careful with stringent default filtering settings
- Know your analysis system (avoid black-boxes)
- Ability to use own databases

Sources of error

- Contaminations through barcodes
- PCR amplification
- FP through sampling (e.g.: skin tissue when taking blood)

-> Clinical interpretation

Recommendations

- Choose sequencing system according to your needs

- Use transparent analysis systems

- Optimize analysis settings to use-case

- Check technical properties of variants (coverage, strand, qualities, …)

- Look at variants in genome browser