A Quick Guide to the

Analytics Behind Genomic Testing

Elaine Gee, PhD
Director, Bioinformatics
ARUP Laboratories

Learning Objectives

- Catalogue various types of bioinformatics analyses that support clinical genomic testing
- Enumerate types of variant classes
- Describe algorithmic methods for variant detection by NGS
- Compare and contrast germline and somatic clinical bioinformatics pipeline methodologies
- Discuss the infrastructure complexity required to support analytics for NGS testing at scale in the cloud
- Explain validation strategies for bringing best-in-class pipelines into clinical production

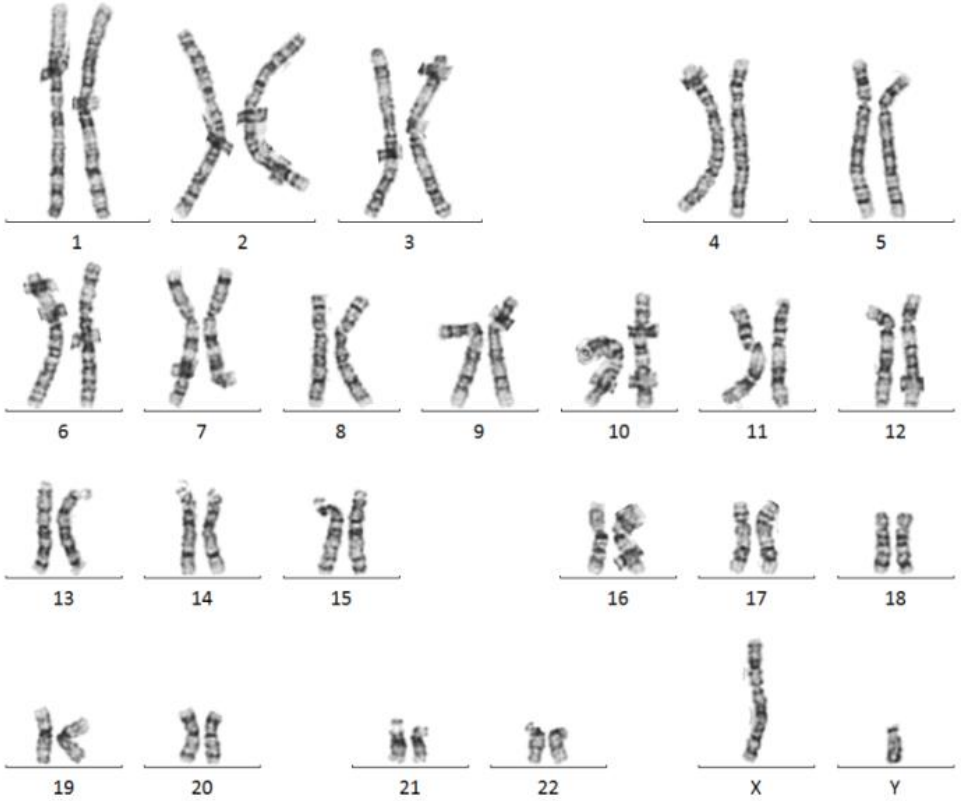
The Human Reference Genome

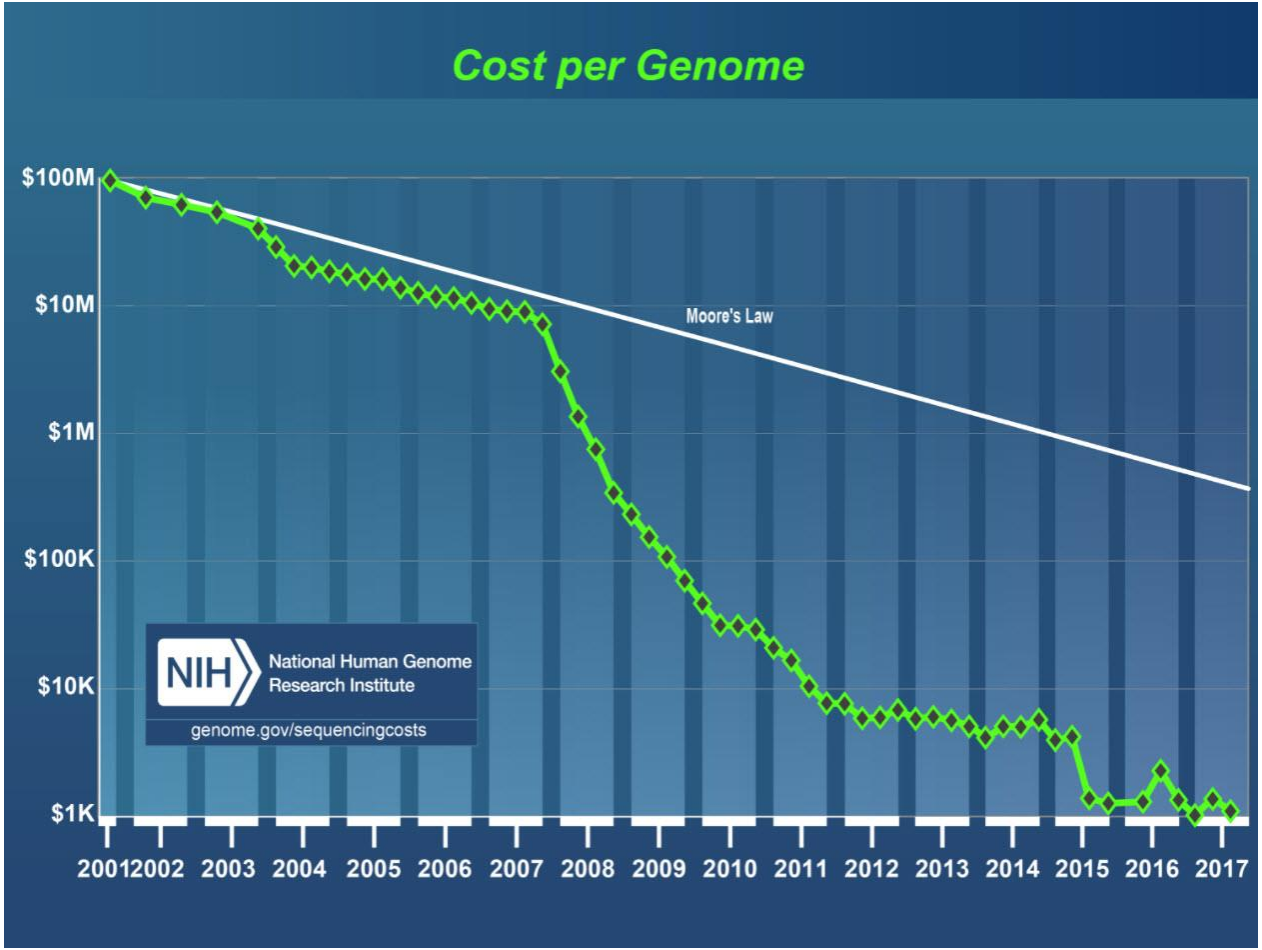
~3B base pairs

structured into

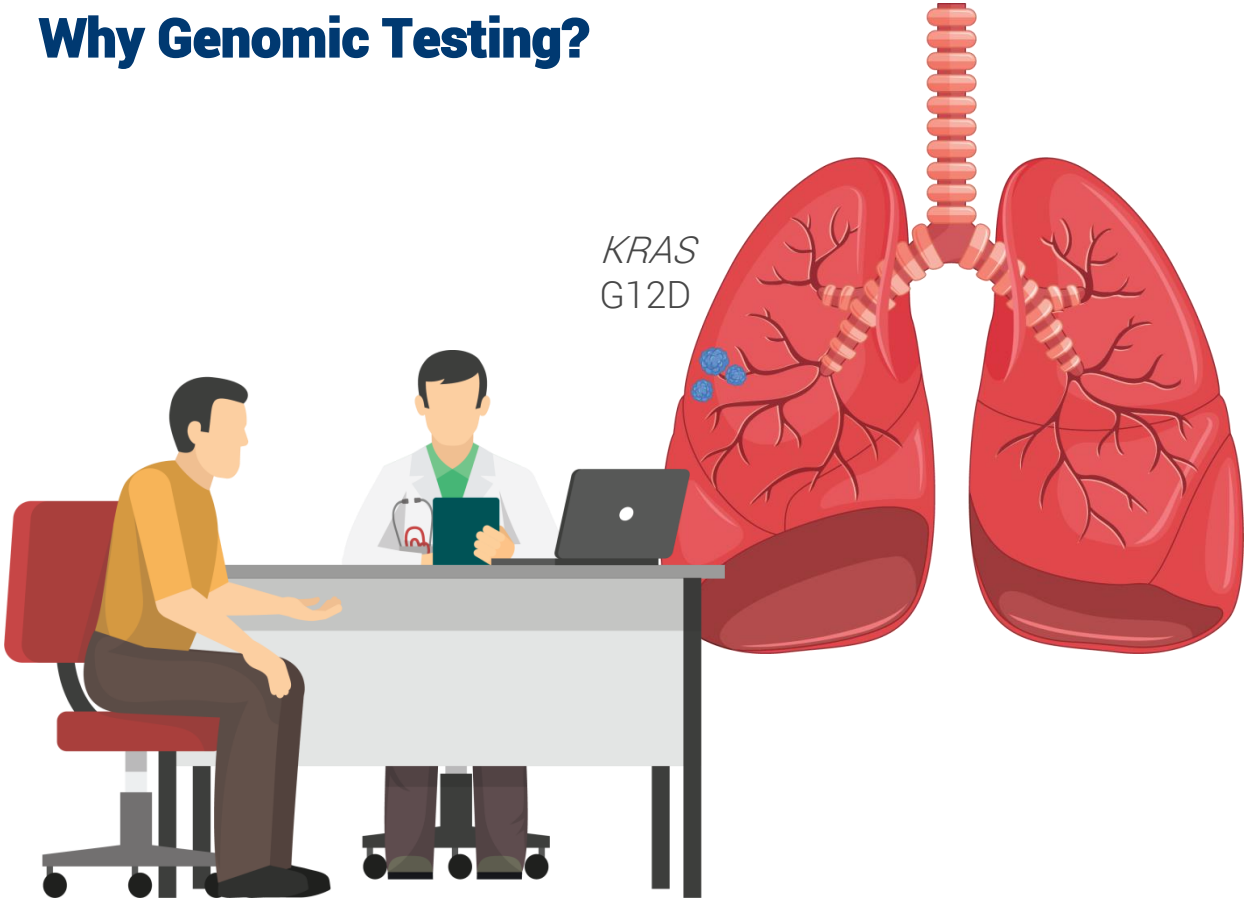
23 chromosome pairs

3,098,825,702	base pairs
20,805	coding genes
14,181	pseudogenes
196,501	gene transcripts




Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017

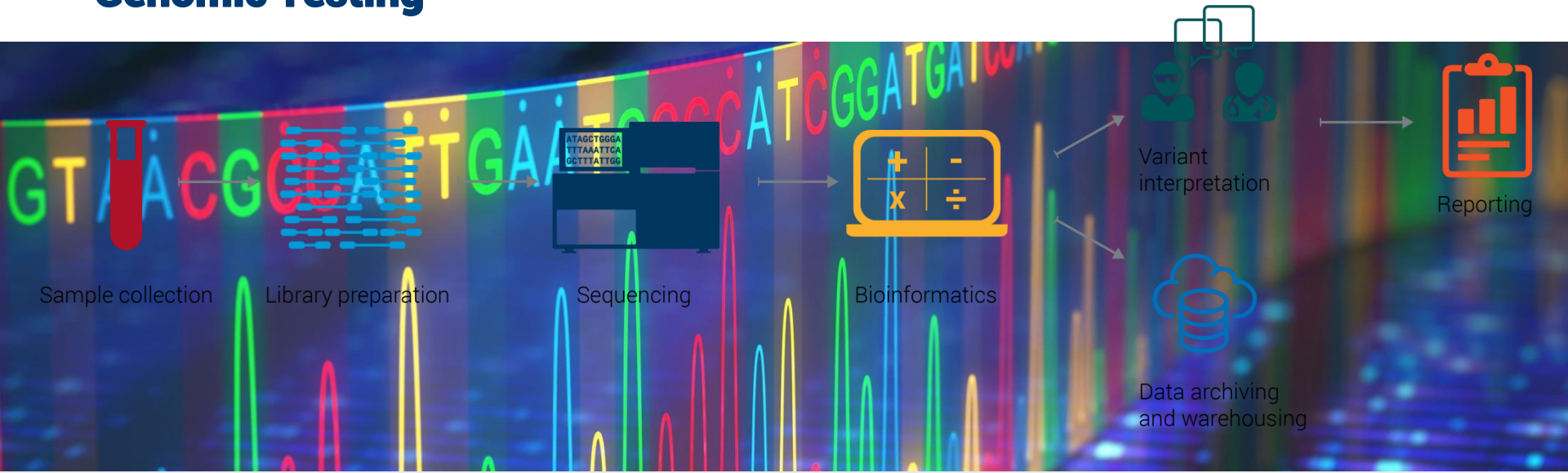
Why Genomic Testing?



1 in 4
cancer deaths are from lung cancer.

~222,500
new cases of lung cancer in the U.S. in 2017.

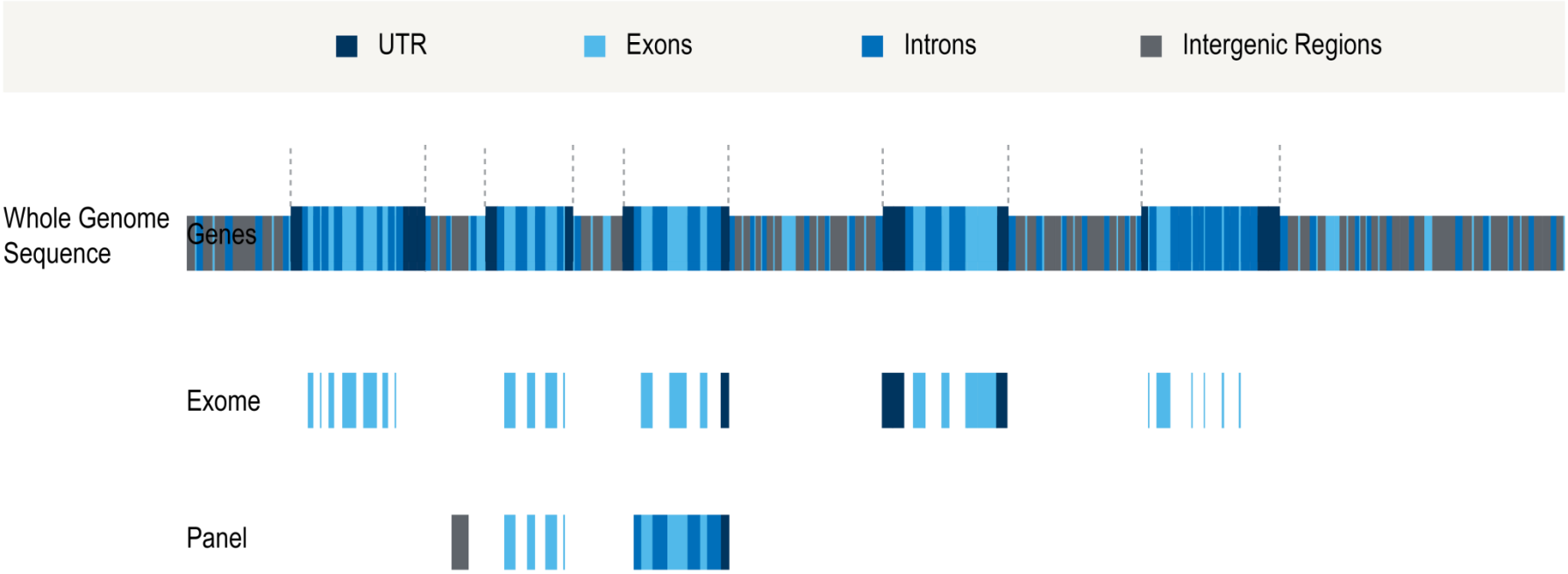
Genomic Testing



Short-Read Sequencers
Illumina
Ion-Torrent

Long-Read Sequencers
PacBio
NanoPore
10X
Nanostring

Types of NGS Testing—Somatic & Germline

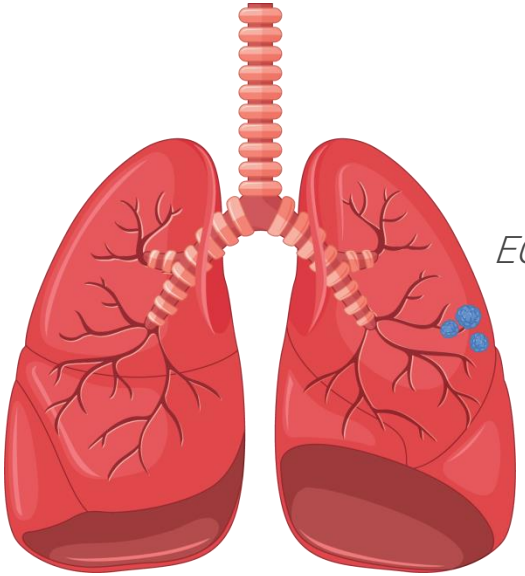


Types of NGS Testing—cfDNA and ctDNA

Non-Invasive Prenatal Testing (NIPT)

Trisomy 21
(Down Syndrome)

Liquid Biopsy



EGFR

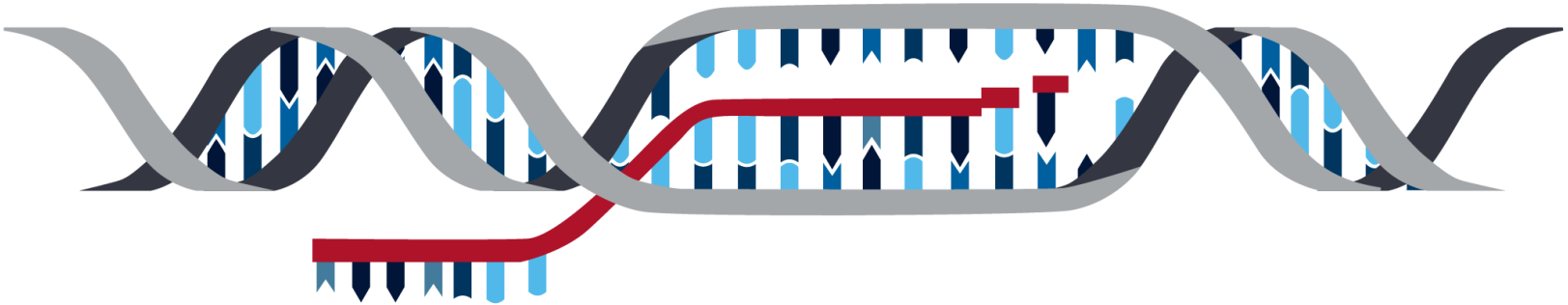
Non-small cell lung cancer

Types of NGS Testing—Infectious Disease

- Virus ✓
- Bacteria
- Fungus
- Protozoa

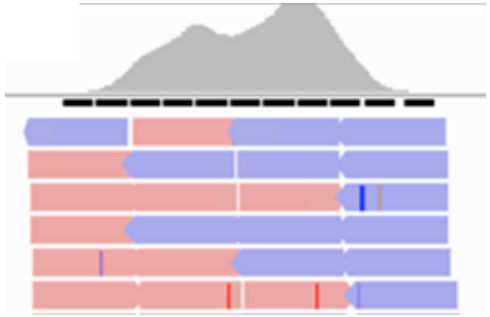
Types of NGS Testing—RNA-Seq

- Alternate transcripts
- Novel gene isoforms
- Gene fusions



Role of Clinical Bioinformatics

Build pipelines



Provide **supplemental information** for clinical interpretation and **quality control**

HGMD®

gnomAD

COSMIC
Catalogue Of Somatic Mutations In Cancer

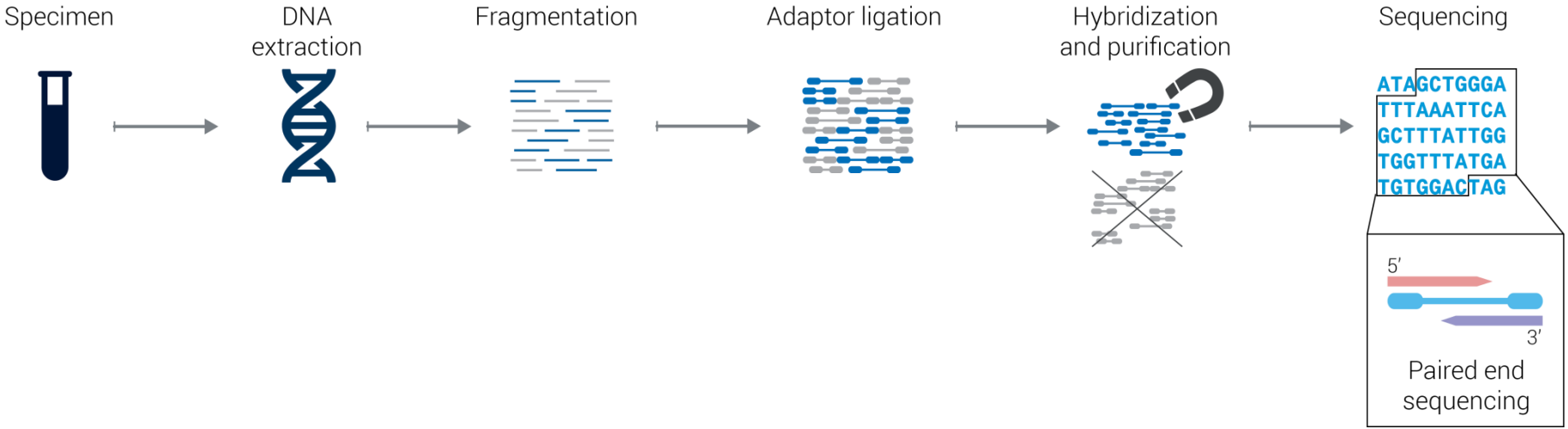
Other computationally heavy analytics are involved in **evaluating**:

Design of new panels

Identification of genetic patterns in patient cohorts

Discovery of gene pathways

Understanding **bioinformatics** requires understanding **the laboratory process.**

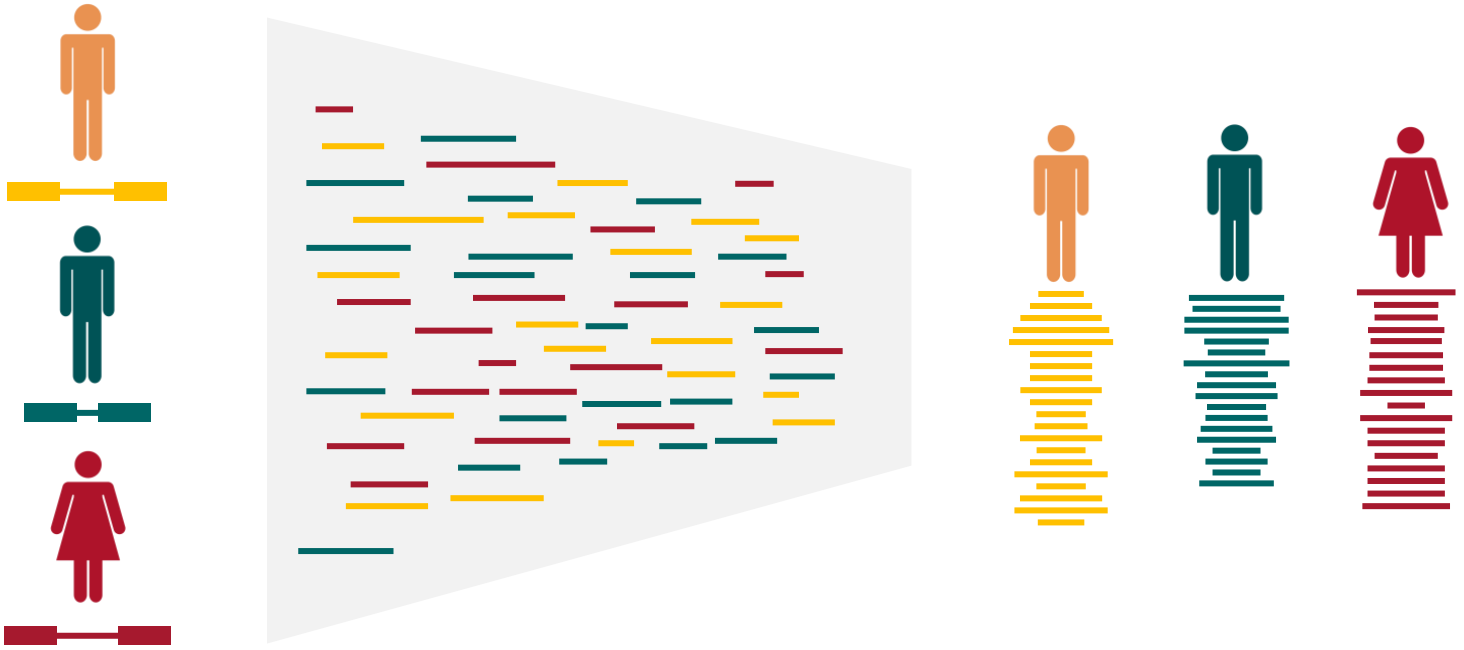


Variant Calling Pipeline

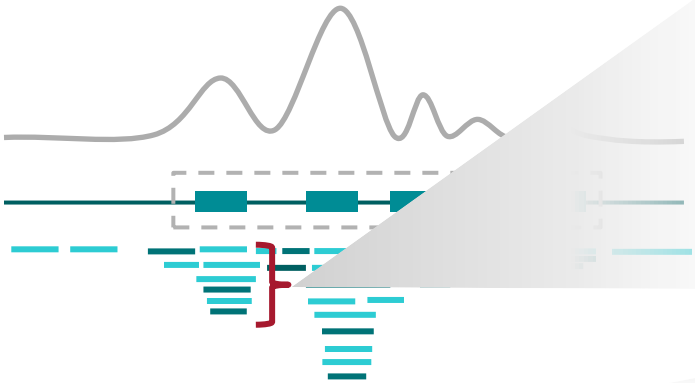
Steps in a bioinformatics pipeline:

1. Sample demultiplexing
2. Read alignment
3. BAM polishing steps
4. Variant calling
5. Variant annotations
6. QC calculations

Step 1: Sample Demultiplexing



Step 2: Read Alignment



Read Alignment

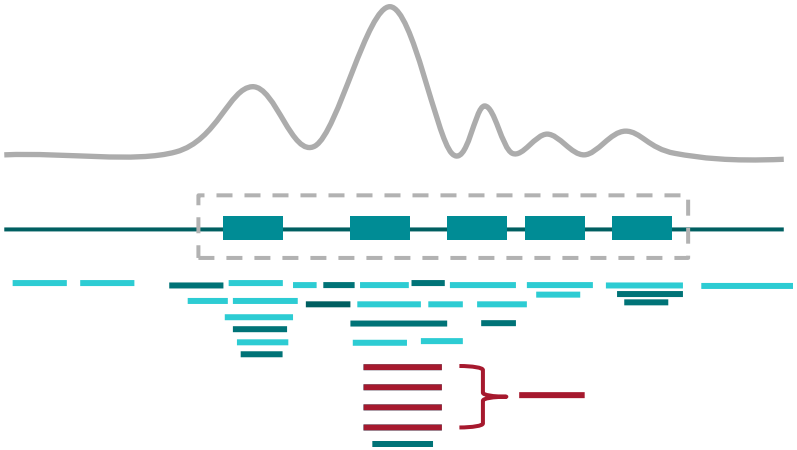
```
Coor     12345678901234  5678901234567890123456789012345
ref      AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1        TTAGATAAAGGATA*CTG
+r002           aaaAGATAA*GGATA
+r003        gcctaAGCTAA
+r004                        ATAGCT.............TCAGC
-r003                              ttagctTAGGC
-r001/2                                        CAGCGGCAT
```

SAM Format

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA     *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA        * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC        *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC              * SA:Z:ref,9,+,5S6M,30,1;
r001  147 ref 37 30 9M         =  7 -39 CAGCGGCAT          * NM:i:1
```

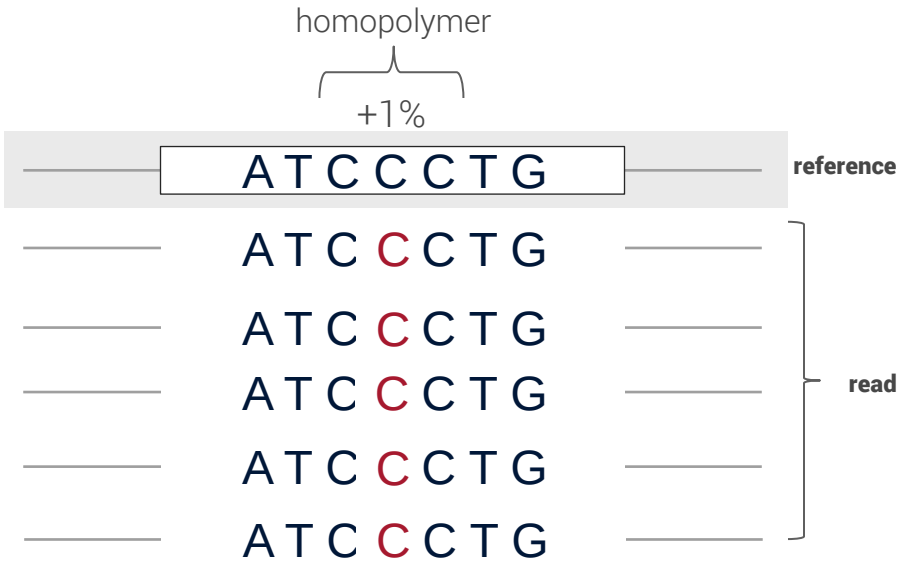
Step 3: BAM Polishing Steps

PCR Duplicate Removal

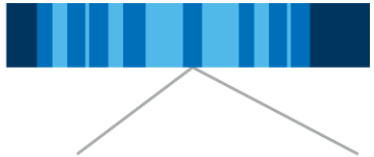


Base Quality Score Recalibration

Q30 Phred base quality score → 99.9% → 1/1000

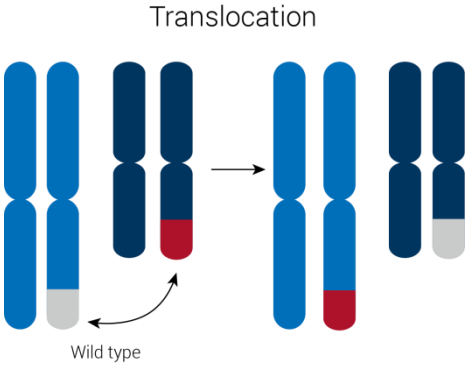


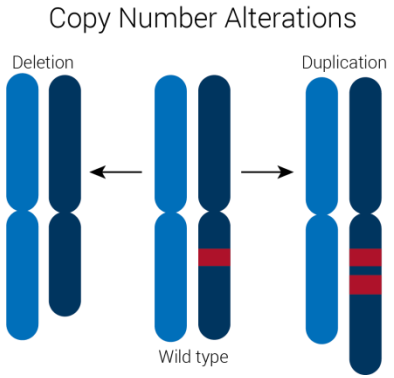
Step 4: Variant Calling by Class



Reference Sequence	**AGGTCCACTGAC**
Single Nucleotide Variant (SNV)	AGCTCCACTGAC
Multi Nucleotide Variant (MNV)	AGCAGCACTGAC
Insertion	AGCGTCCACTGAC
Deletion	AG...TCCACTGAC
Indel	AGCTA...TCCACTGAC
Inversion	AGGTCCAAGTCC
Duplication	AGGTCCACACTGAC

Structural Variants





Example Variant Calling Algorithms

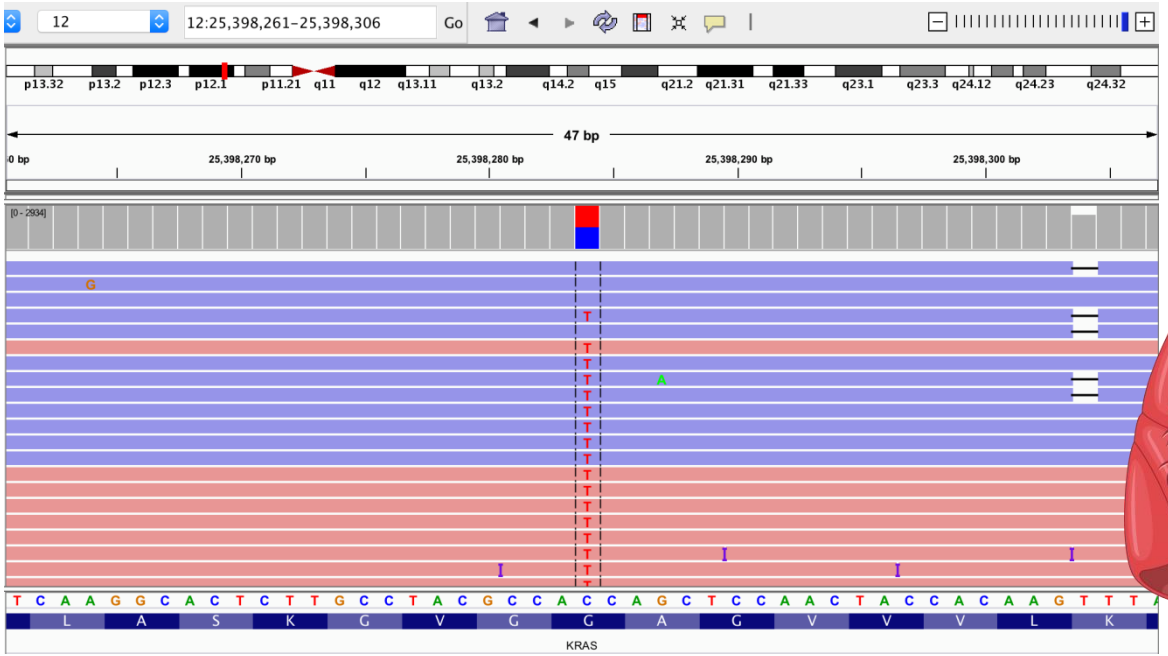
SNV/Insertion/Deletion

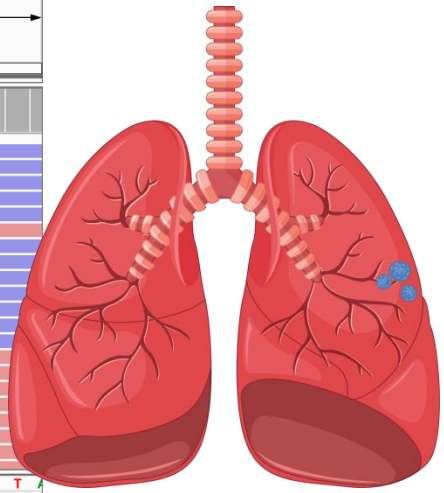
- Position based callers (GATK Unified Genotyper, LoFreq)
- Local de-novo assembly of haplotypes (GATK Haplotype Caller)
- Graph based variant callers (Graph Genome)
- Neural networks (Deep Variant)

Duplications/Structural variants

- Pattern growth approach (Pindel)
- Split reads, discordant paired-end reads (Manta, DELLY, CREST)
- kmer + de-novo assembly (BreaKmer)
- Unmapped or partially mapped reads (ITD Assembler)
- Depth of coverage + background error correction + principal component analysis (XHMM)
- Tumor/normal
- B allele frequency

Example *KRAS* G12D Variant Cell





```
#CHROM  POS       ID  REF  ALT  QUAL  FILTER  INFO                                    FORMAT      Sample
12      25398284  .   C    T    1409  PASS    AF=0.498;DP=1843;DP4=440,469,397,521;SB=7  GT:AF:DP4   0/1:0.498:440,469,397,521
```

Step 5: Variant Annotations

VCF variant

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	Sample
12	25398284	.	C	T	1409	PASS	AF=0.498;DP=1843;DP4=440,469,397,521;SB=7	GT:AF:DP4	0/1:0.498:440,469,397,521

Annotated variant

Gene	Location	Nuc. Change	Protein Change	Variant Type	Depth	Allele Freq
KRAS	chr12: 25398284	c.35G>A	p.Gly12Asp	Nonsynonymous	2869	49.8

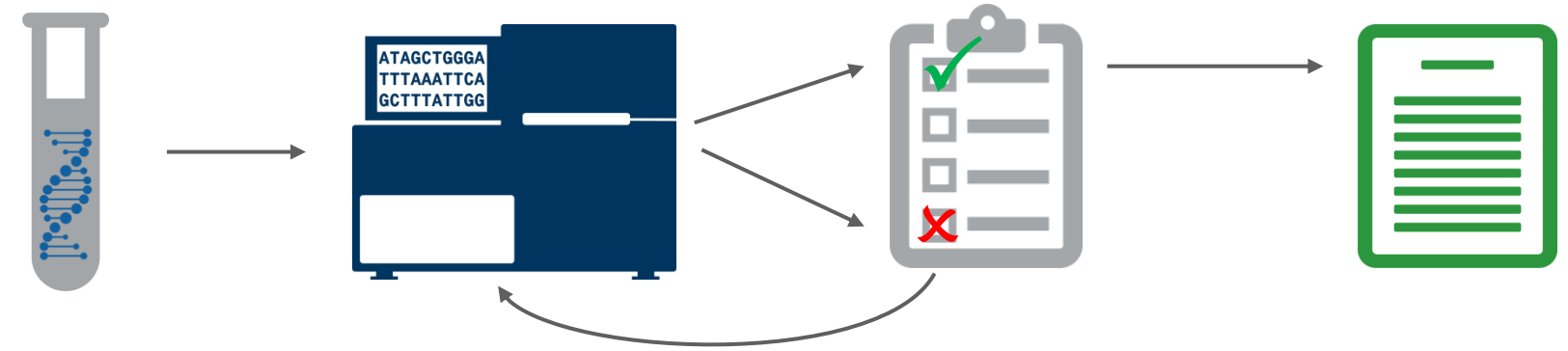
The VCF variant includes:

- chromosome
- position
- ID
- reference base
- alternate base
- variant quality
- meta-information
  - information and individual format fields
  - filter flags

The annotated variant includes:

- Gene
- Gene Transcript
- Nucleotide change (cdot)
- Protein change (pdot)
- Variant Type
  - Polymorphism
  - Synonymous
  - Non-synonymous
    - Nonsense
    - Missense
  - Frame shift

Step 6: QC Calculations

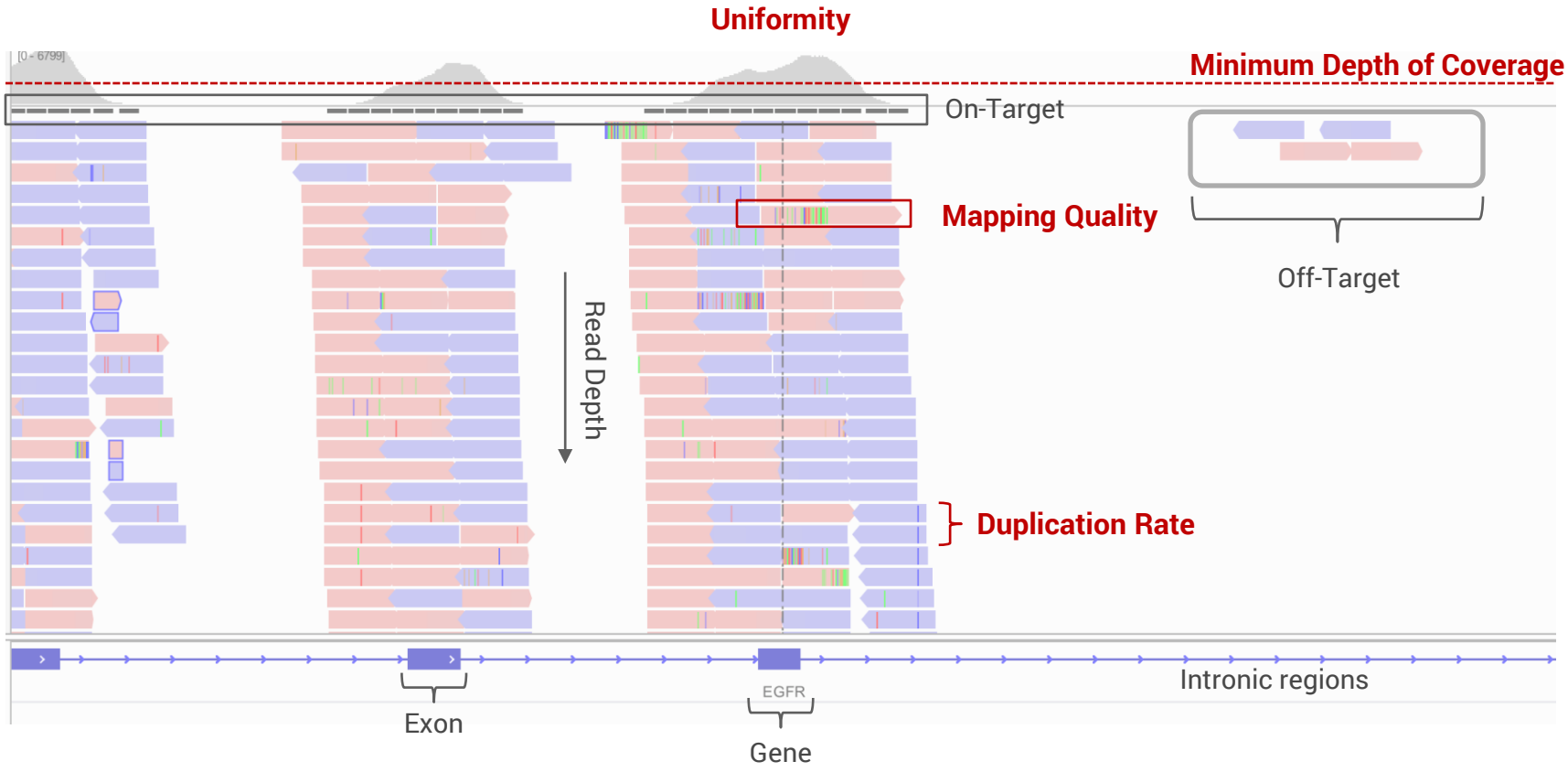


Run Level
Cluster density
Base call quality score
Fragment size

Sample Level
Depth coverage
Uniformity
Mapping quality
Duplication rate

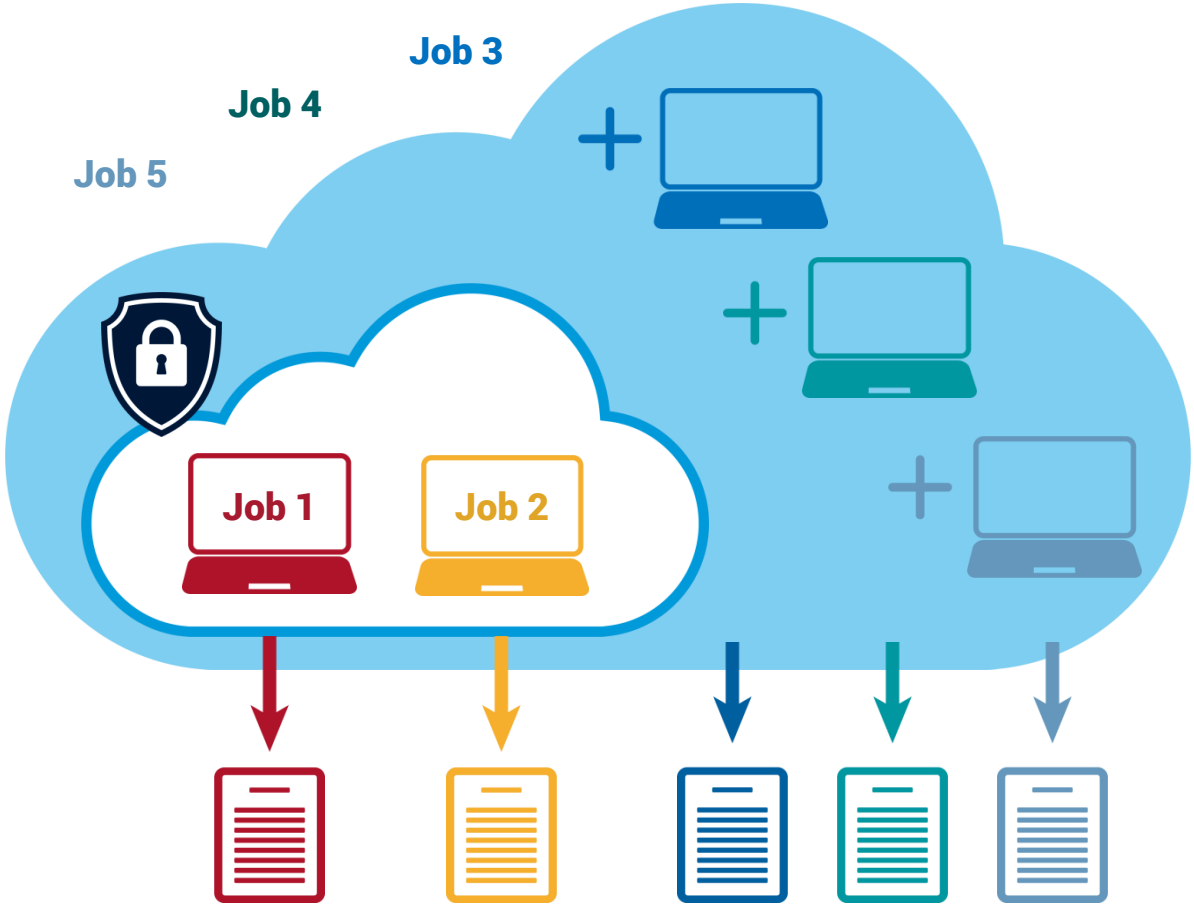
Variant Level
Novel variants
Known variants
Transition-to-transversion ratio

Sample-Level QC Metrics for Targeted Capture

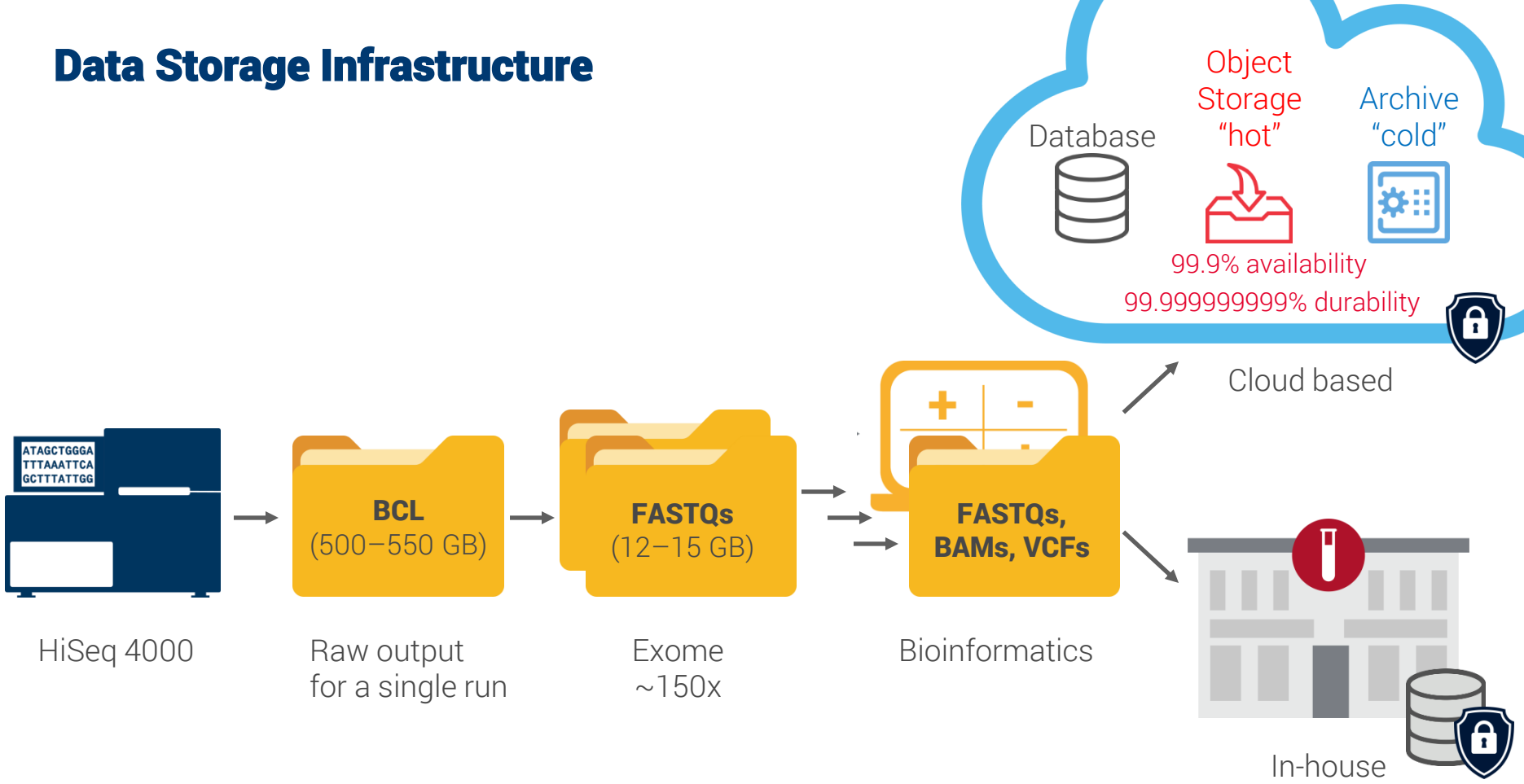


Compute Infrastructure for Data Processing

How does a bioinformatics job get executed in clinical production?



Data Storage Infrastructure



Bioinformatics Pipeline Validation

- Recommendations from CAP/AMP
  - 17 recommendation statements
  - 59 variants tested in each variant class
- Example Statistics
  - Positive percentage agreement (PPA)
  - Positive predictive value (PPV)
  - Reproducibility
  - Allelic fraction lower limit of detection
- Validation required prior to use in clinical production

Summary

Catalogue various types of bioinformatics analyses that support clinical genomic testing

Enumerate types of variant classes

Describe algorithmic methods for variant detection by NGS

Compare and contrast germline and somatic clinical bioinformatics pipeline methodologies

Discuss the infrastructure complexity required to support analytics for NGS testing at scale in the cloud

Explain validation strategies for bringing best-in-class pipelines into clinical production

Questions?

Elaine Gee, PhD
Director of Bioinformatics
ARUP Laboratories
elaine.gee@aruplab.com