

Bioinformatics

From next-generation sequencing reads to molecular interpretation

Dr Abu Suraih Sakhri

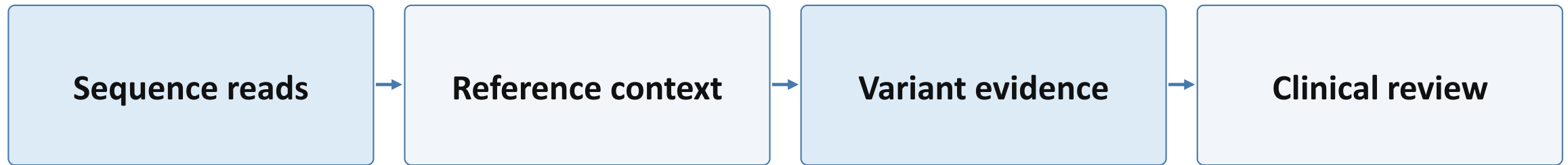
MBBS, MD (Pathology)

www.abusuraisakhri.com

Learning objectives

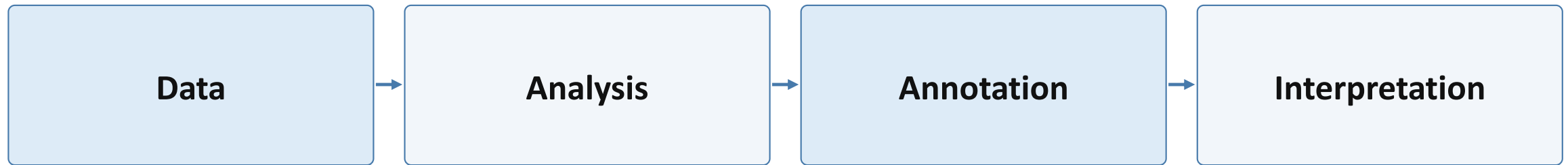
- Map NGS data from short reads to a clinical laboratory report
- Explain alignment, variant calling, sequencing depth, and VAF
- Distinguish annotations from clinical pathogenicity judgments
- Relate mutational signatures, TMB, and MSI to tumor assessment

Why sequencing requires bioinformatics



- NGS produces large collections of individual sequence reads
- A read has little clinical meaning without genomic context
- Computation organizes read-level evidence before laboratory review

Sequence data are not yet a diagnosis



- The instrument generates DNA sequence information
- Analytical pipelines transform raw observations into candidate changes
- Clinical significance is established after annotation and review

NGS measures many templates in parallel



- Individual DNA molecules are physically separated
- Local amplification makes each template detectable
- Parallel sequencing generates one read from each amplified template

How NGS differs from conventional Sanger sequencing

Feature	Sanger sequencing	Next-generation sequencing
Input	Usually homogeneous template	Heterogeneous DNA may be analyzed
Output	Composite region sequence	Many individual reads
Interpretation	Limited for mixed sequences	Computational read-level analysis

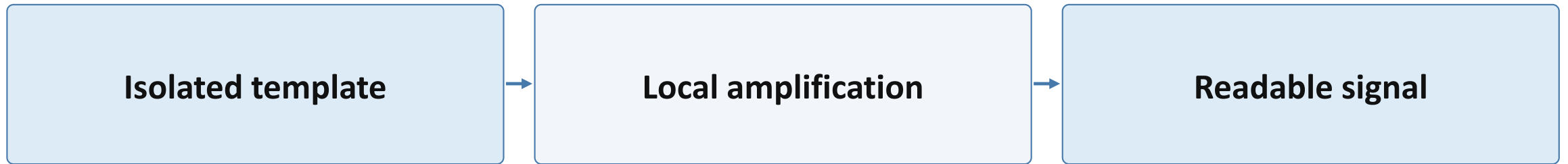
- Sanger usually requires a homogeneous DNA template
- Sequence mixtures can be difficult to interpret as averaged signals

Spatial separation preserves molecule identity



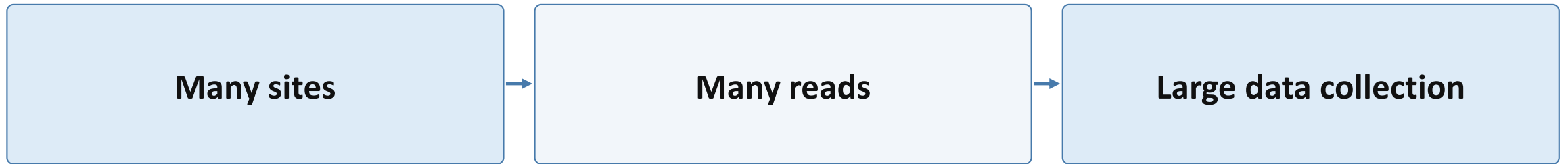
- The initial process isolates templates in space
- A separate site represents an individual starting molecule
- Independent sites can produce separately interpretable reads

Local amplification increases detectable signal



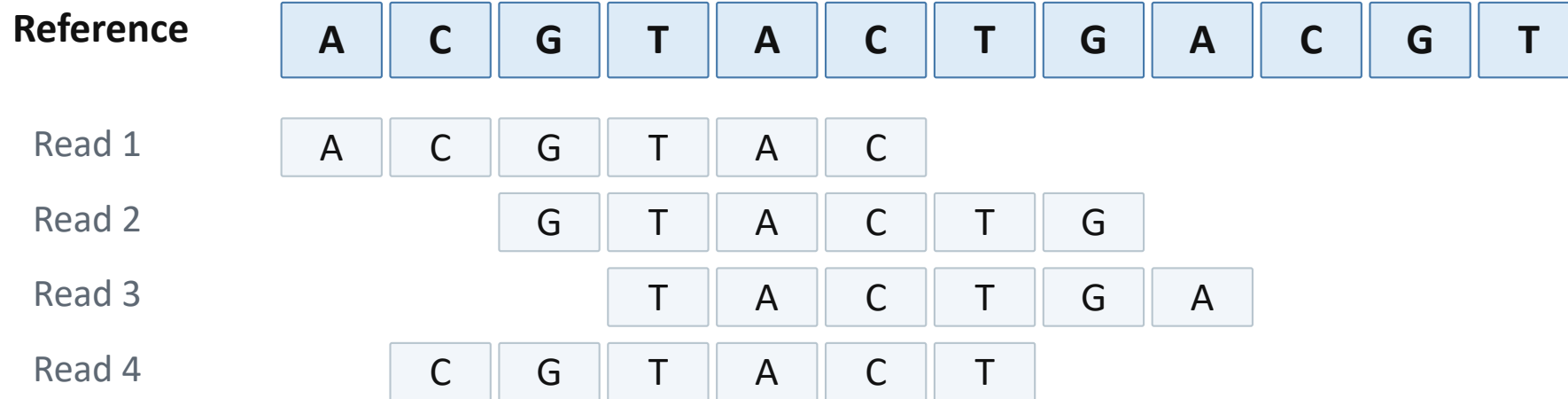
- Amplification occurs after spatial isolation
- Each original template is amplified at its own location
- Enough signal is required for accurate base detection

Parallel sequencing creates the data burden



- Millions of templates can be interrogated simultaneously
- Each amplified molecule contributes sequence information
- Interpretation therefore depends on scalable computational analysis

A read is a DNA fragment sequence



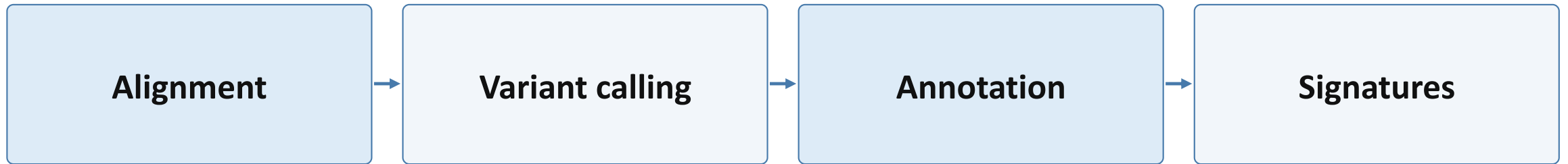
Original conceptual alignment schematic; Not patient sequence data.

- A read contains the base sequence of one measured fragment
- Short reads must be related to the reference genome
- The biological question concerns the source locus, not the read alone

High throughput creates analytical scale

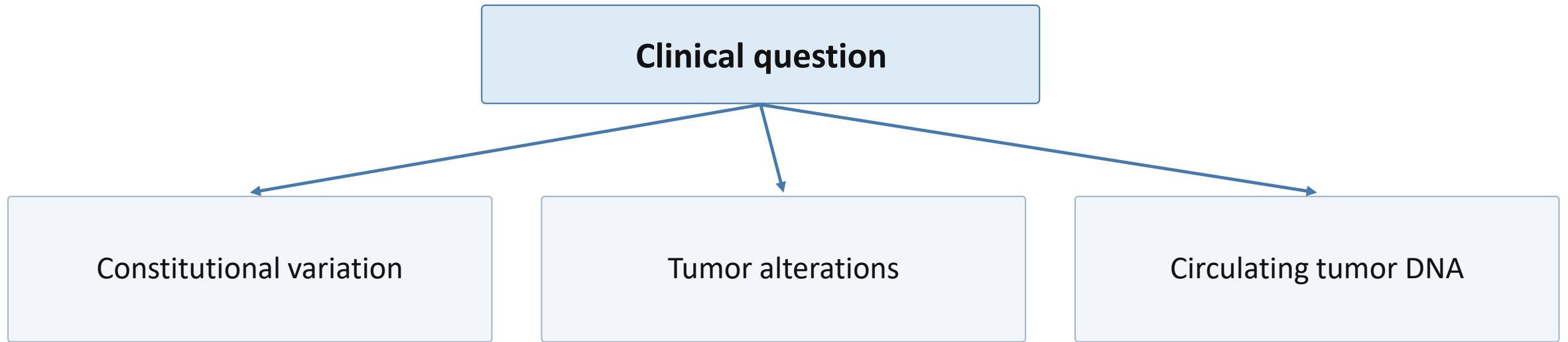
- High-throughput instruments can produce billions of reads
- Data volumes reaching terabytes per day
- Storage and interpretation are therefore inseparable from sequencing

A pipeline is an ordered analytical workflow



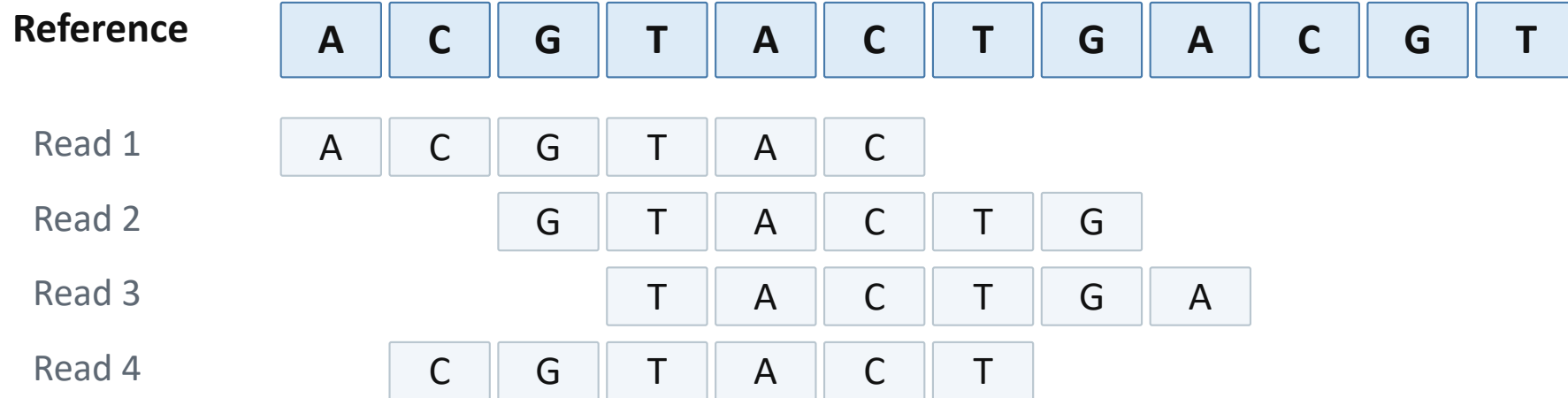
- Analysis methods vary by application and specimen type
- The core tasks are alignment, variant calling, and interpretation
- Additional analyses can characterize aggregate mutation patterns

Specimen and clinical question shape the analysis



- Constitutional disease and cancer testing ask different questions
- DNA can be heterogeneous within a tissue or blood specimen
- Pipeline selection and interpretation must reflect the application

Read alignment provides genomic context



Original conceptual alignment schematic; Not patient sequence data.

- Alignment maps sequencing reads to a reference genome
- A genomic coordinate makes a read interpretable
- The mapped collection becomes the substrate for subsequent comparison

The reference genome is a coordinate system

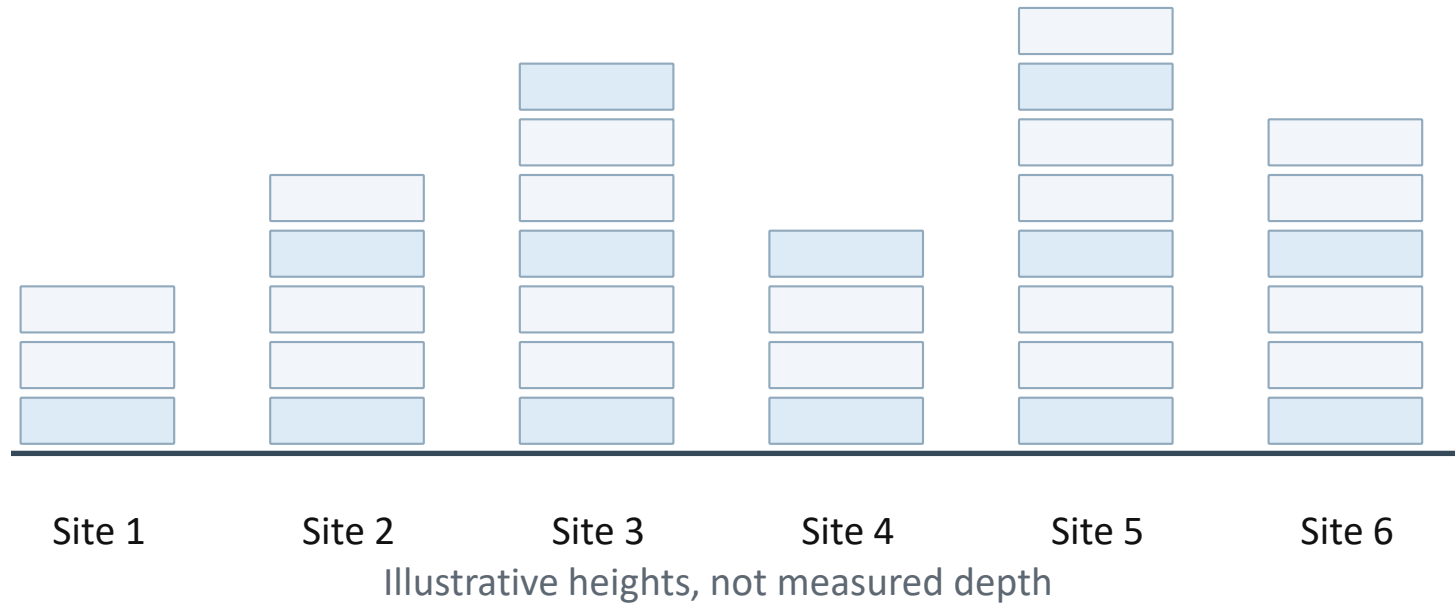
Reference	A	C	G	T	A	C	T	G	A	C	G	T
Read 1	A	C	G	T	A	T						
Read 2			G	T	A	T	T	G				
Read 3				T	A	T	T	G	A			
Read 4		C	G	T	A	C	T					

Red = hypothetical alternate base; Example is not patient data.

- Mapped reads can be viewed at corresponding reference positions
- A reference base is the comparator for detecting variation
- Alignment alone does not determine clinical significance

Many reads may cover the same locus

Illustrative read coverage across neighboring positions



- Multiple molecules can yield reads overlapping one site
- Overlapping observations supply repeated evidence
- Their accumulation provides sequencing depth at that site

Alignment and variant calling are distinct

Where?



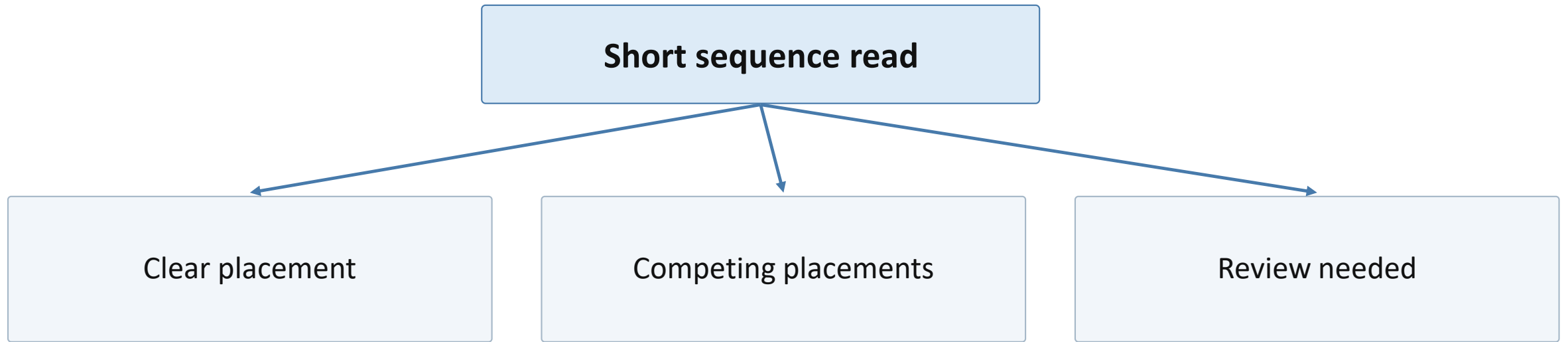
What differs?



What does it mean?

- Alignment asks where each sequence fits in the reference
- Variant calling asks how the sample differs at those positions
- Clinical interpretation asks why a particular difference matters

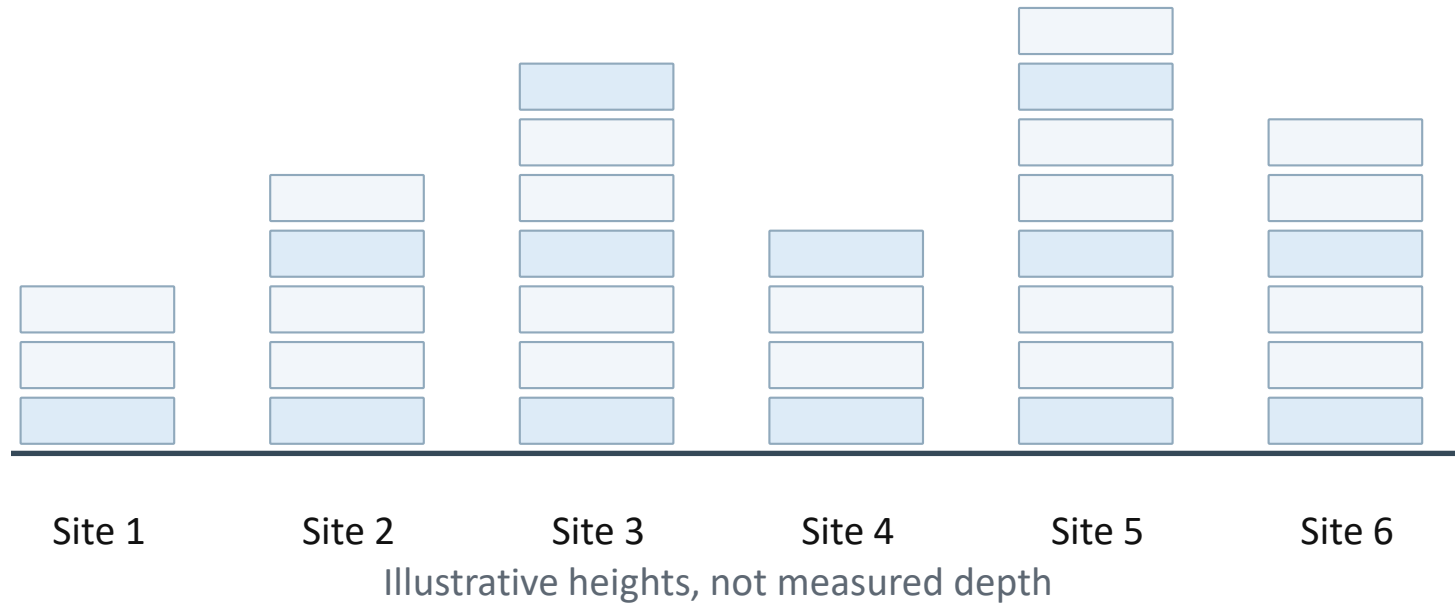
Ambiguous mapping illustrates a limitation



- A short sequence may not uniquely identify one genomic location
- Reference context and analytical review influence confidence
- Not every read supports an unambiguous variant conclusion

Depth is the number of reads covering a position

Illustrative read coverage across neighboring positions



- Sequencing depth describes local read coverage
- Depth can vary across the regions being interrogated
- A reported variant must be interpreted relative to its evidence

Breadth and depth answer different questions

Feature	Breadth	Depth
Question	Which regions are covered?	How many observations at a site?
Primary relevance	Scope of detection	Sensitivity to low-frequency alleles
Design influence	Target selection	Allocation of sequencing reads

- Breadth describes how much genomic territory is interrogated
- Depth describes how many reads cover a position

Low-frequency alleles require sufficient depth

Reference and alternate molecules in a sample



R = reference support A = alternate support

A minority alternate signal can coexist with abundant reference DNA

Conceptual schematic — not patient data or an assay validation plot.

- An altered allele may be present in only a minority of molecules
- More overlapping reads increase the chance of observing it
- Low-VOF detection therefore depends strongly on local depth

Variant allele frequency is a read proportion

Reads overlapping a single locus



R = reference support A = alternate support

$$\text{VAF} = \text{alternate-supporting reads} / \text{all evaluated reads at that site}$$

Conceptual schematic — not patient data or an assay validation plot.

- VAF is the fraction of reads at a locus supporting an alternate allele
- VAF describes observed evidence in the sequenced specimen
- It is not, by itself, a clinical classification or diagnosis

VAF: Distinguish support from total coverage

Reads overlapping a single locus



R = reference support A = alternate support

VAF = alternate-supporting reads / all evaluated reads at that site

Conceptual schematic — not patient data or an assay validation plot.

- The numerator is reads supporting the specified variant
- The denominator is reads considered at that position
- Interpret the proportion together with depth and sample context

A low VAF is not the same as a small effect

Feature	Observed allele fraction	Clinical effect
Meaning	Relative read support	Potential disease relevance
Evidence	Sequencing observations	Annotation and clinical evaluation
Do not infer	Pathogenicity from frequency	Variant abundance from pathogenicity

- VAF quantifies how often a change is observed in sample reads
- Pathogenicity concerns biological or clinical significance

Mixed cell populations complicate interpretation

Reference and alternate molecules in a sample



R = reference support A = alternate support

A minority alternate signal can coexist with abundant reference DNA

Conceptual schematic — not patient data or an assay validation plot.

- NGS can analyze specimens containing different DNA molecules
- Tumor-containing samples may mix altered and unaltered DNA
- Observed VAF reflects the analyzed material, not a pure cell population

Variant calling systematically compares sequences

Reference	A	C	G	T	A	C	T	G	A	C	G	T
Read 1	A	C	G	T	A	T						
Read 2			G	T	A	T	T	G				
Read 3				T	A	T	T	G	A			
Read 4		C	G	T	A	C	T					

Red = hypothetical alternate base; Example is not patient data.

- The called result expresses a difference from a reference sequence
- Calling depends on the aligned sequence observations
- This step precedes biological annotation and clinical interpretation

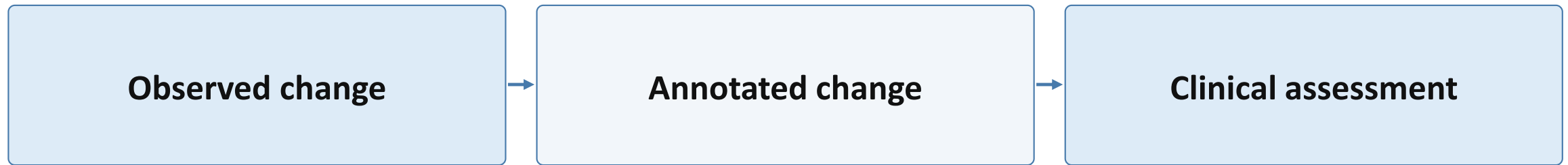
The alternate base must be supported by reads

Reference	A	C	G	T	A	C	T	G	A	C	G	T
Read 1	A	C	G	T	A	T						
Read 2			G	T	A	T	T	G				
Read 3				T	A	T	T	G	A			
Read 4		C	G	T	A	C	T					

Red = hypothetical alternate base; Example is not patient data.

- A reference position provides the expected base
- Aligned sequence reads may show the reference or an alternate base
- The evidence needs assessment before a change is reported

Variant identity and clinical meaning are separate



- A called sequence change is an analytical observation
- Some changes are common polymorphisms
- Clinical assessment determines whether a change is relevant to disease

Germline polymorphisms illustrate baseline variation

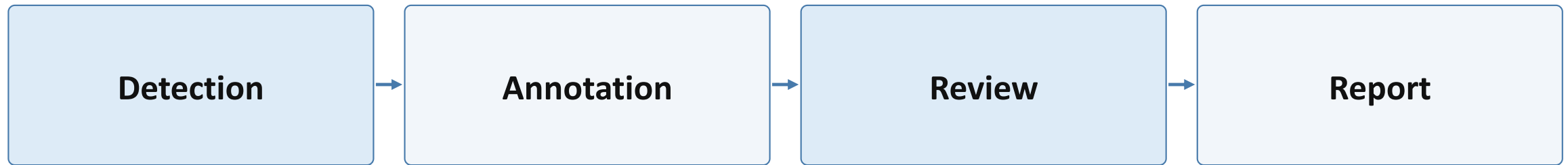
- Individuals naturally differ from the reference genome
- Adequate sequencing is intended to identify variants within target intervals
- A mismatch with the reference does not automatically imply disease

Coverage constrains a negative interpretation

Feature	Positive finding	No variant detected
Evidence	Alternate reads observed	No qualifying alternate evidence
Interpretive issue	Validity and biological effect	Was detection adequately supported?
Context	Source specimen	Scope and depth of the assay

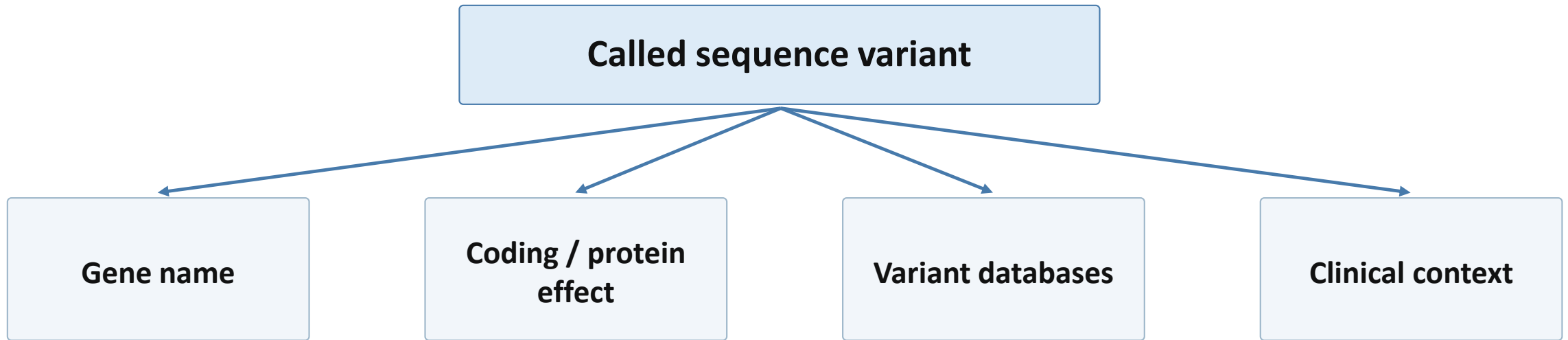
- Failure to observe a variant depends on what was examined
- Insufficient depth weakens inference about rare alleles

Analytical detection is not pathogenicity



- Variant calling identifies candidate sequence alterations
- Annotation adds molecular and contextual information
- Laboratory interpretation assigns clinical relevance where supported

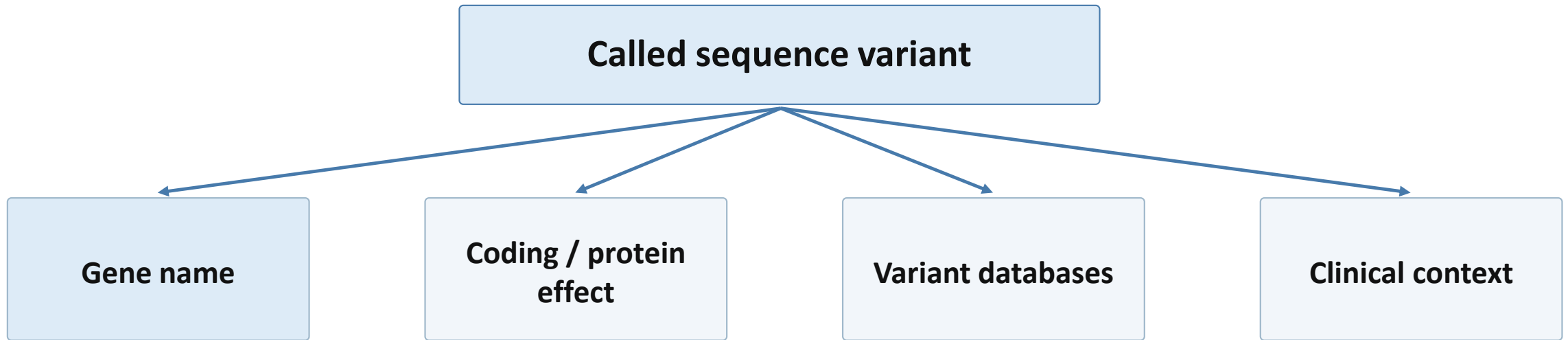
Annotation turns coordinates into descriptions



Annotation assembles evidence; Laboratory review determines what is reportable.

- A genomic change can be linked to a gene
- Coding changes and predicted protein effects may be added
- Population, pathogenicity, and clinical databases offer context

Gene identification is a first annotation layer



Annotation assembles evidence; Laboratory review determines what is reportable.

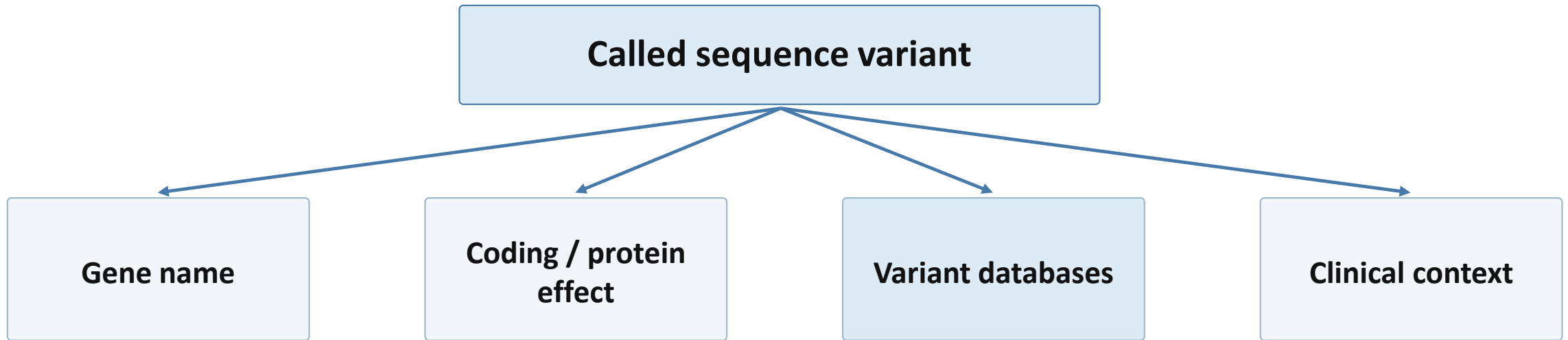
- The genomic location may be assigned to a named gene
- Gene identity provides functional context
- The gene name alone cannot establish a disease-causing effect

Sequence change and protein effect differ

Feature	Coding annotation	Protein consequence
Describes	Nucleotide-level change	Possible amino-acid or functional effect
Evidence type	Sequence mapping	Computed consequence / prediction
Pitfall	Gene name is not diagnosis	Prediction is not proof

- The coding-level description identifies an alteration in sequence
- Protein-effect predictions describe possible consequences

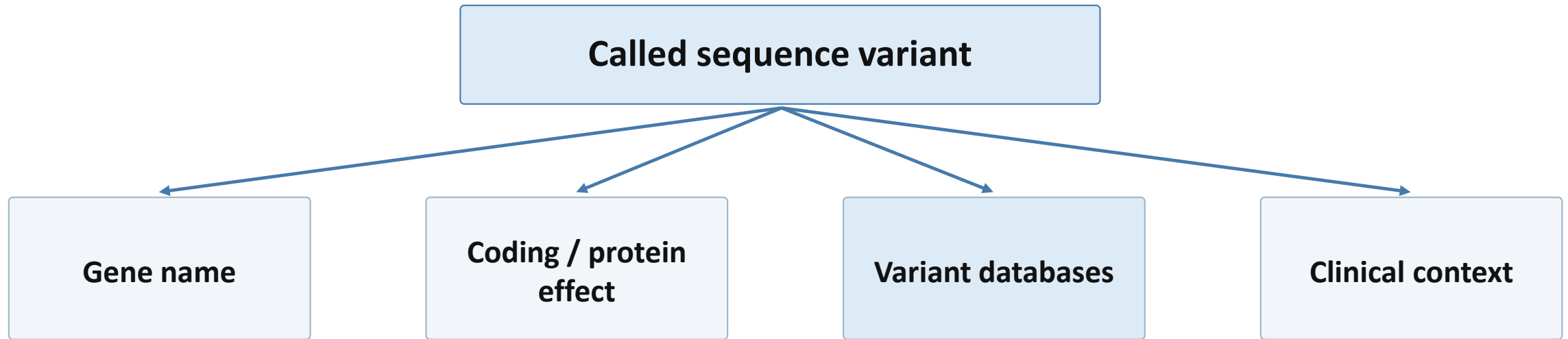
Population databases provide frequency context



Annotation assembles evidence; Laboratory review determines what is reportable.

- Commonly occurring variants can be recognized through databases
- Frequency information helps place a finding in perspective
- Commonness and clinical significance must not be conflated

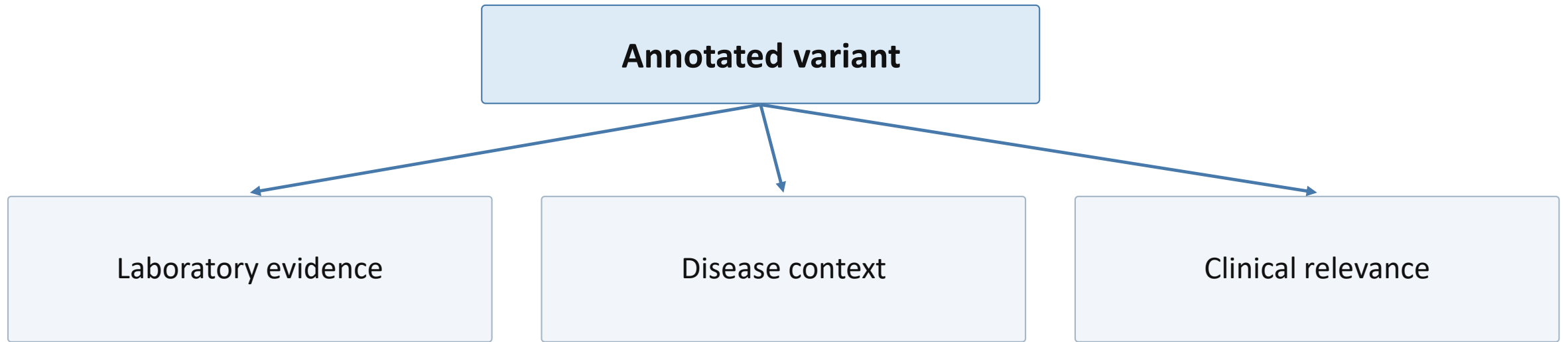
Pathogenic variant resources add disease context



Annotation assembles evidence; Laboratory review determines what is reportable.

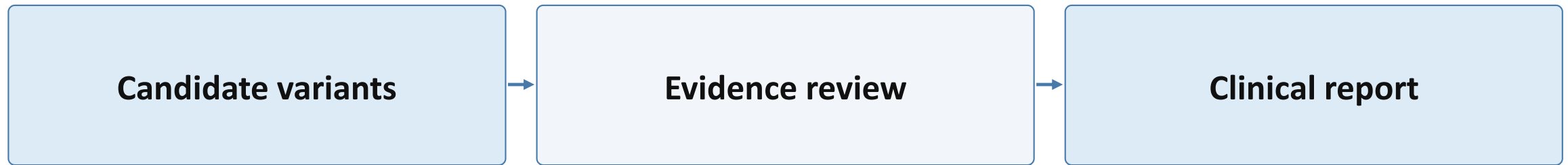
- Databases may catalog variants recognized as pathogenic
- Reported pathogenicity information requires expert assessment
- The clinical laboratory remains responsible for reviewing evidence

Clinical information completes the annotation



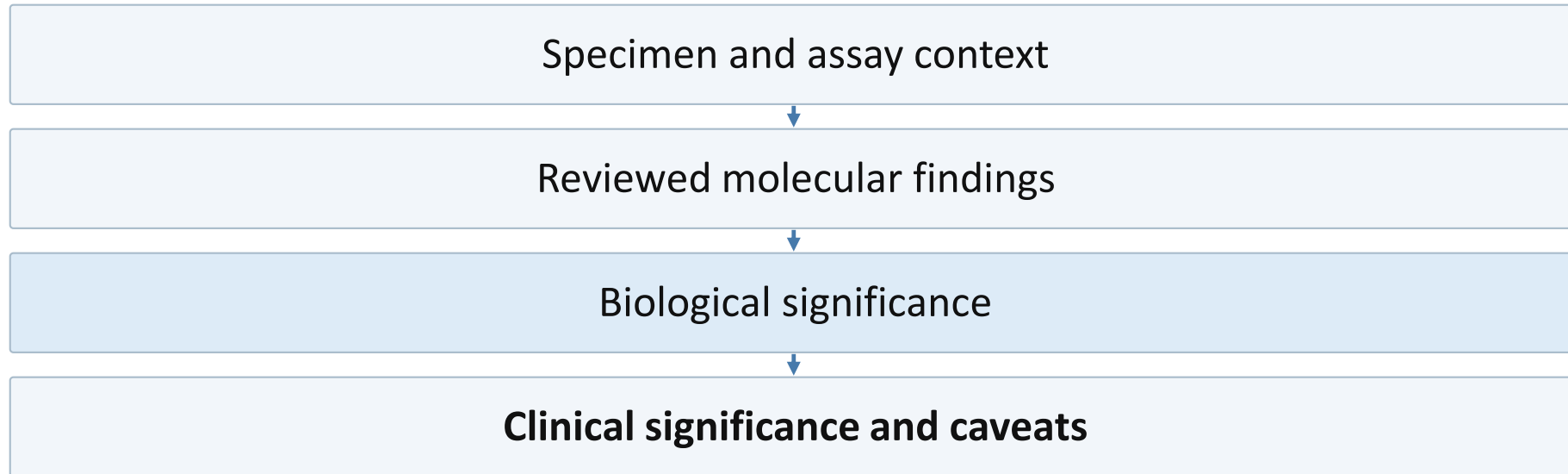
- Clinical history can affect the relevance of a molecular finding
- The same observed alteration may answer different diagnostic questions
- Interpretation connects the result to the patient context

The laboratory review is a necessary gate



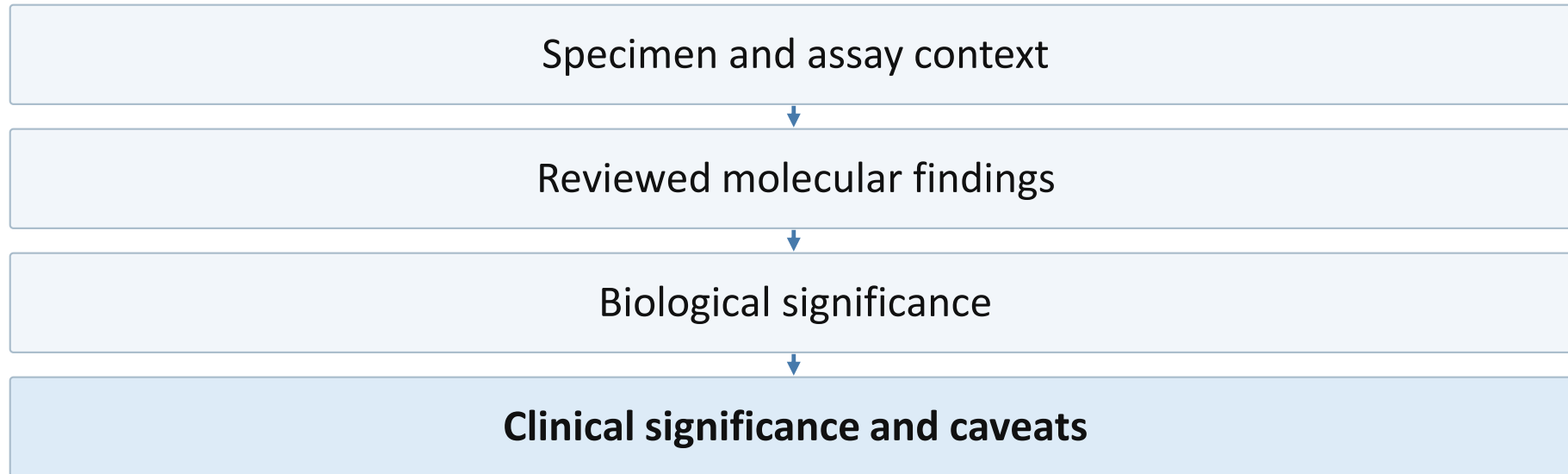
- Annotated output is not automatically a final clinical report
- The clinical laboratory reviews identified changes
- Only reviewed findings are incorporated into reporting

Reports should explain biological significance



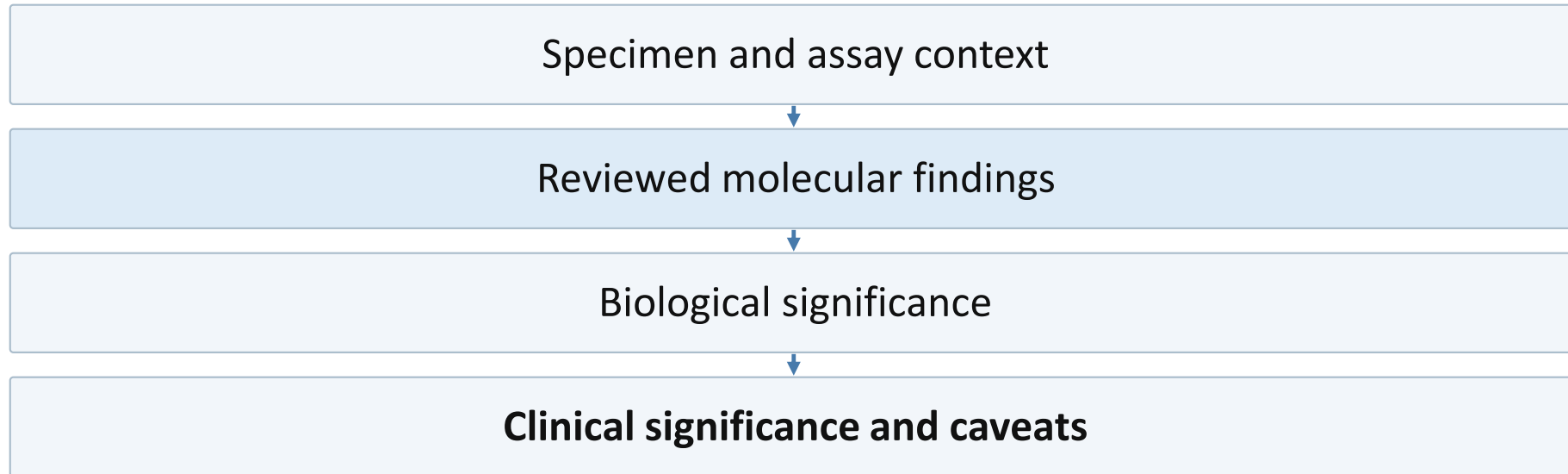
- Clinicians need more than a variant name or coordinate
- A concise description should address biological meaning
- The report should distinguish clinically important findings

Reports should explain clinical significance



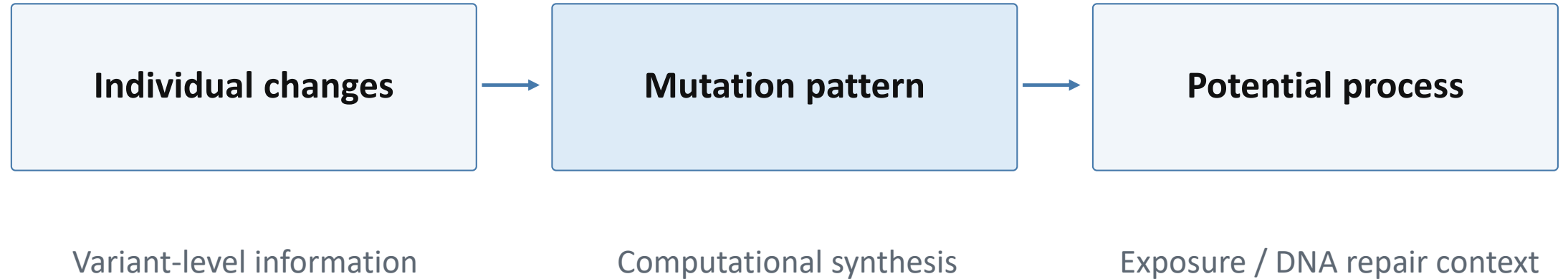
- Molecular knowledge evolves faster than routine clinical familiarity
- A report translates genomic interpretation into patient-relevant terms
- Evidence and remaining uncertainty should be communicated

Tumor reports may contain multiple alterations



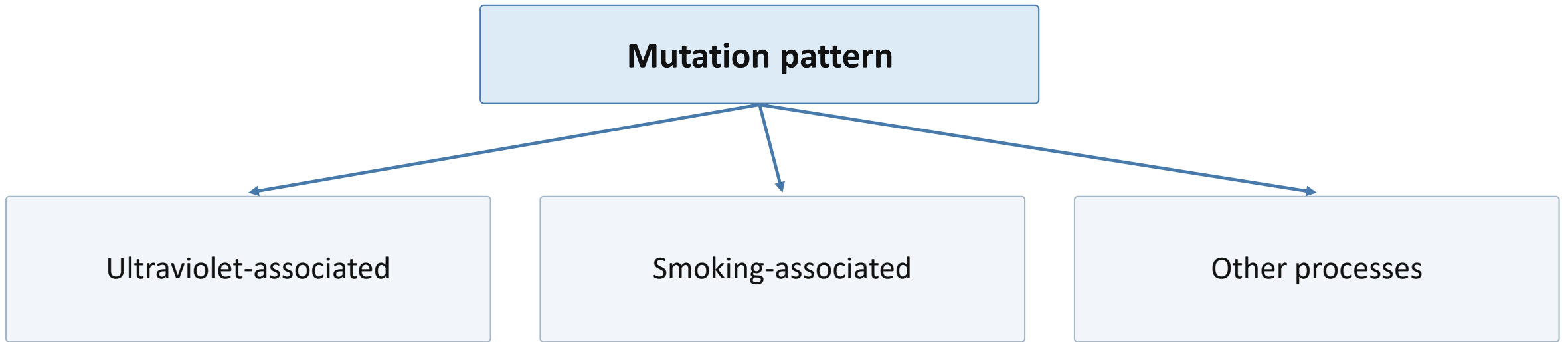
- Cancers can harbor numerous detected changes
- The clinically meaningful subset requires interpretation
- Variant lists alone are an insufficient summary of disease relevance

Mutational signatures aggregate patterns



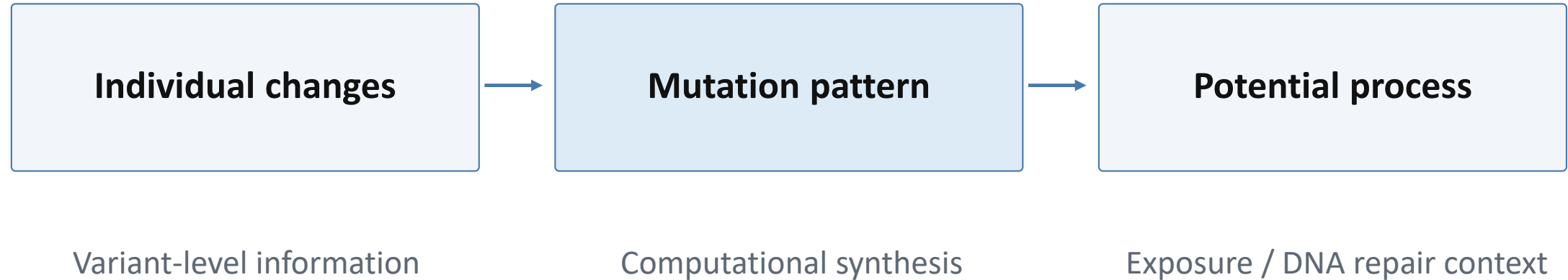
- Single variants describe discrete sequence changes
- Signatures examine patterns across multiple mutations
- Patterns can provide clues about underlying mechanisms

Some patterns suggest environmental exposure



- Ultraviolet light as one signature-associated exposure
- Smoking is another cited source of mutation patterns
- A signature is an etiologic clue, not direct proof of an exposure

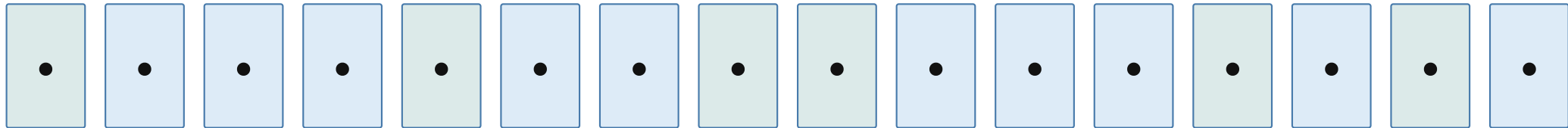
DNA repair defects can shape mutational patterns



- Failure of DNA repair alters the processes generating mutations
- Informatics can detect patterns compatible with repair defects
- The interpretation concerns a process, not merely a variant count

Tumor mutational burden summarizes burden

Observed sequence alterations

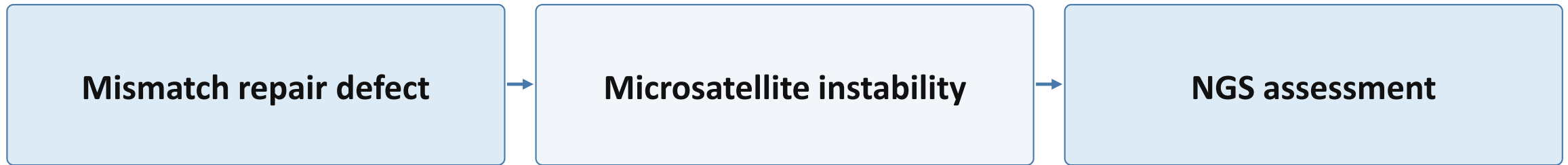


Aggregate mutation burden in the assessed territory

Conceptual grouping only: No mutation count, cutoff, or specimen shown.

- TMB quantifies mutation burden rather than naming one driver
- Informatics may estimate this feature from NGS data
- The clinical role discussed by Robbins concerns immunotherapy response

Microsatellite instability signals repair dysfunction



- Defects in mismatch repair can produce microsatellite instability
- MSI is a distinct feature from the total mutation burden
- NGS-derived data can be used to assess MSI status

TMB and MSI are distinct readouts

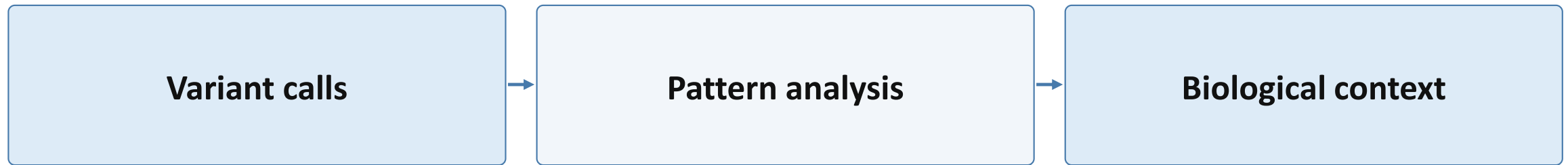
Feature	TMB	MSI
Main construct	Mutation burden	Microsatellite instability
Associated process	Aggregate mutation accumulation	Mismatch repair dysfunction
Interpretive caution	Not a specific mutation	Not identical to high TMB

- TMB expresses overall tumor mutation burden
- MSI characterizes instability linked to mismatch repair dysfunction

NGS biomarkers inform immunotherapy discussions

- High TMB and MSI with checkpoint-inhibitor responsiveness
- This is a clinical association, not a guarantee for an individual patient
- Interpretation must remain linked to the specific clinical setting

Signature analysis adds a process-level layer



- Variant calling lists changes at individual sites
- Signature analysis considers relationships across many events
- Combining layers gives a more complete tumor profile

NGS scope determines what can be interpreted

Whole genome: Coding and noncoding regions

Whole exome: Coding exons

Selected gene panel

Conceptual nested scopes — area is not proportional to sequence size.

- Clinical testing can target selected genes, exons, or the whole genome
- Broader territory generally introduces more interpretation work
- Bioinformatics must match the assay's intended genomic scope

Targeted panels prioritize relevant regions

Whole genome: Coding and noncoding regions

Whole exome: Coding exons

Selected gene panel

Conceptual nested scopes — area is not proportional to sequence size.

- A panel sequences selected genes or exons
- Concentrating reads can maximize depth and reduce cost
- Smaller reportable territory can limit interpretation burden

Whole-exome sequencing broadens coding coverage

Whole genome: Coding and noncoding regions

Whole exome: Coding exons

Selected gene panel

Conceptual nested scopes — area is not proportional to sequence size.

- WES enriches protein-encoding exons
- Coding exons as roughly 1.5% of the genome
- The broader scope is useful in unresolved genetic disorders

Whole-genome sequencing extends beyond exons

Whole genome: Coding and noncoding regions

Whole exome: Coding exons

Selected gene panel

Conceptual nested scopes — area is not proportional to sequence size.

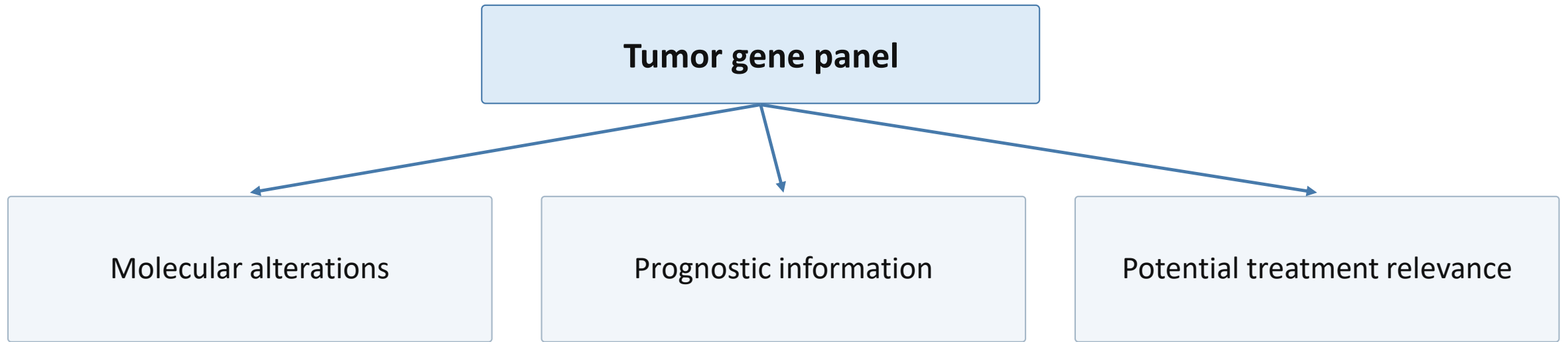
- WGS examines coding and noncoding genome territory
- It can detect structural rearrangements when sequencing is sufficient
- High costs and informatics challenges constrain clinical implementation

The assay choice changes analytical workload

Assay	Analytical emphasis	Scope of interrogation
Gene panel	Focused depth and selected variants	Clinical gene set
WES	Broad coding-region changes	Exons
WGS	Genome-wide and structural findings	Coding + noncoding regions

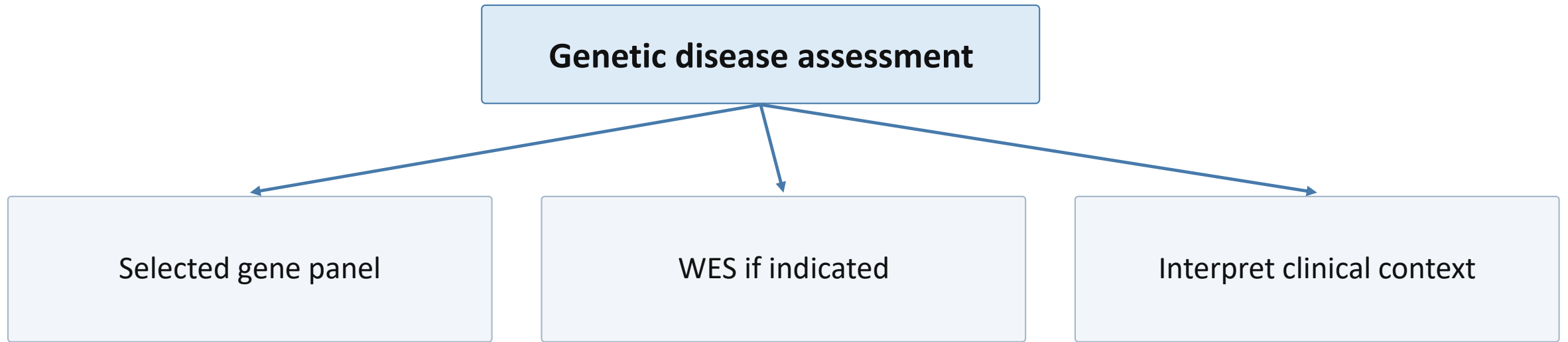
- Panels restrict interrogation to deliberately selected genes
- WES increases coding-region interpretation

Tumor profiling links alterations to care



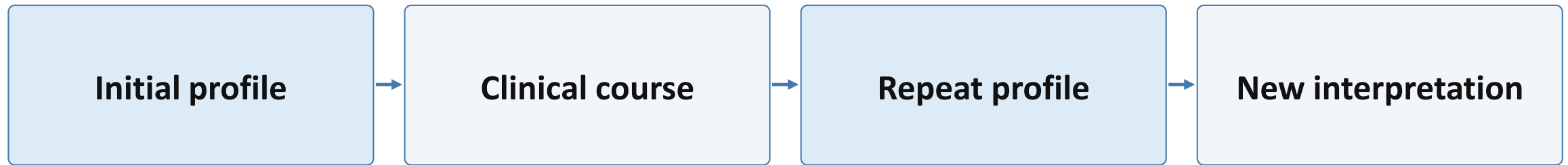
- Tumors contain distinct sets of somatic alterations
- Gene panels seek clinically informative molecular changes
- Reporting must distinguish prognostic from potentially treatable findings

Constitutional disease testing seeks a genetic cause



- NGS can investigate inherited or developmental disorders
- Panels are used in disorders such as cardiomyopathy or deafness
- WES can help identify molecular diagnoses after prior inconclusive testing

Repeat profiling can assess acquired resistance



- A recurrent cancer may differ molecularly from the original tumor
- Repeat NGS testing can identify relevant resistance-associated changes
- Interpret the new findings in relation to prior treatment and specimen

Liquid biopsy makes allele sensitivity important

Reference and alternate molecules in a sample



R = reference support A = alternate support

A minority alternate signal can coexist with abundant reference DNA

Conceptual schematic — not patient data or an assay validation plot.

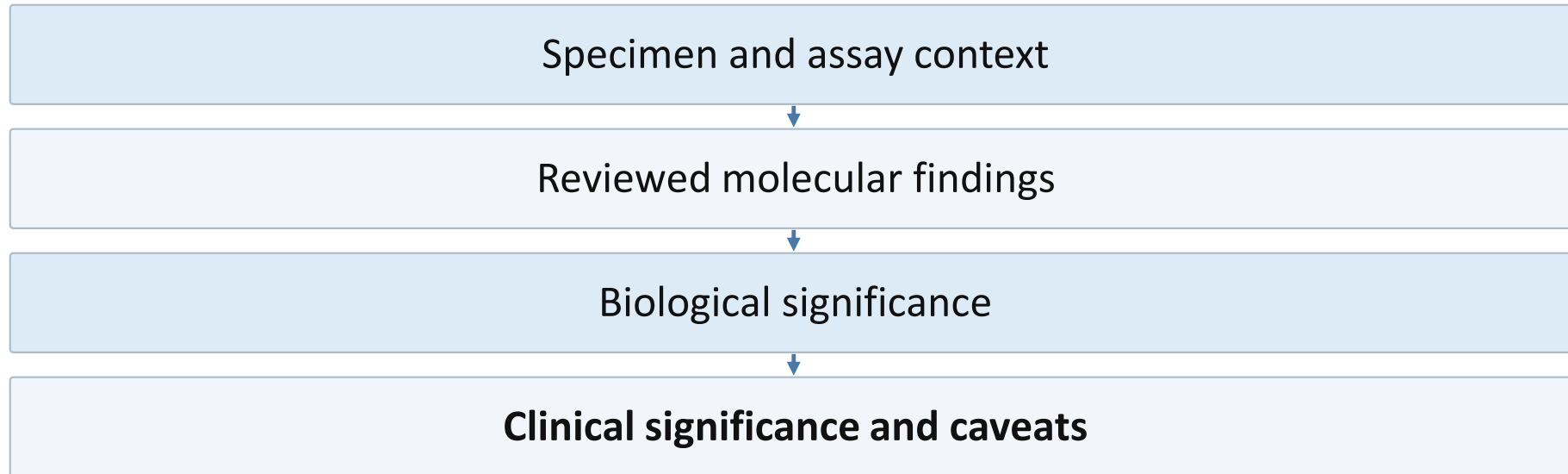
- Tumor-derived DNA may circulate in peripheral blood
- Tumor-related alterations can be infrequent in mixed DNA
- Low-VAF analysis and assay-specific interpretation become important

Circulating DNA requires specimen-aware interpretation

Feature	Tumor tissue analysis	Circulating tumor DNA analysis
Material	Direct tumor-containing specimen	DNA recovered from blood
Evidence	Tumor sequence profile	Tumor-associated signal in circulation
Caution	Tumor heterogeneity	Low-frequency signal and sample context

- Circulating tumor cells and cell-free tumor DNA are different analytes
- NGS can detect tumor-associated sequence changes in blood

A defensible report answers three questions



- What sequence change was detected, and with what supporting evidence?
- What is its relevant molecular or biological annotation?
- What clinical interpretation is justified for this specimen?

Case: A suspected inherited disorder

Which step turns a called variant into a useful diagnostic result?

Interpretation integrates annotation and the clinical phenotype; A reference mismatch alone is insufficient.

Hypothetical teaching case; No patient-derived results are shown.

- A child with a suspected genetic disorder undergoes sequencing
- Many sequence differences are identified relative to the reference
- Prioritize laboratory review, annotation, and phenotype concordance

Case: Heterogeneous cancer sequencing

Does a low observed VAF make the alteration clinically unimportant?

No. VAF describes the observed fraction of reads, not the variant's biological or clinical consequence.

Hypothetical teaching case; No patient-derived results are shown.

- A tumor specimen contains altered and unaltered DNA molecules
- A low-frequency alteration is seen among overlapping reads
- Depth, specimen composition, and reportable clinical meaning must be assessed

Case: Tumor pattern information

Why does a mutational signature add information beyond a variant list?

It summarizes a pattern that may indicate an exposure or DNA repair mechanism.

Hypothetical teaching case; No patient-derived results are shown.

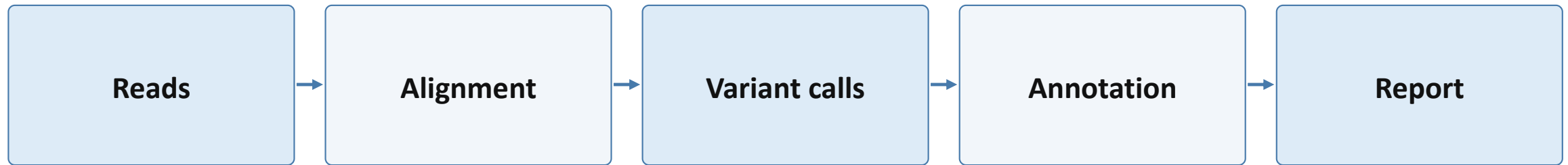
- A cancer profile shows numerous sequence alterations
- Aggregate analysis identifies a pattern relevant to a repair process
- Individual variants and process-level features provide complementary evidence

The principal pitfalls are category errors

Feature	Analytical output	Question answered
Alignment	Where does a read map?	Coordinates
Variant call / VAF	What differs / in what fraction?	Local evidence
Annotation / review	What does the finding mean?	Biology + clinical context

- Alignment is not variant calling; Variant calling is not diagnosis
- Sequencing depth is not the same as genomic breadth

An integrated molecular-diagnostic workflow



- The reference genome anchors the aligned reads
- Read evidence becomes candidate variants and annotated findings
- Clinical laboratory review converts selected findings into a report

Knowledge check: Interpret the analytical layers

- A low-VOF finding needs assessment of read support and context
- An annotated coding change still requires clinical interpretation
- A tumor mutation pattern can provide mechanism-level information

References and source boundary

- Robbins, Cotran & Kumar: Pathologic Basis of Disease, 11th ed. (2026)
- Chapter 5, “Genes and Human Diseases,” Bioinformatics, pp. 164–165
- Adjacent NGS principles and clinical applications provide teaching context
- Schematics are original conceptual illustrations, not patient observations