

THE RESULT

evidence unification

CAPRI resolves the **77% singleton paradox** by harmonizing 20+ cancer registries into an auditable 3D structure-aware clinical workbench across 193,021 variants.

193,021 Variants 768 Genes 6,066 Evidence Items 1,186 Actionable (Tiers I/II) 345 Approved Drugs QGP Population Filter

1 The clinical bottleneck

Next-generation sequencing is routine in clinical oncology, yet somatic variant interpretation remains the rate-limiting bottleneck for molecular tumour boards worldwide.

$$\text{Actionability}(v) = f(\text{Tier}, \text{Drug}, \Delta G_{\text{bind}}, \text{AF}_{\text{pop}})$$

The Singleton Paradox (Prop. 1)

- Extreme Data Siloing:** Somatic evidence is dispersed across 20+ registries with conflicting curation schemas and contradictory tier assignments.
- 77.0% Singleton Rate:** 77% of variant interpretations appear in only a single knowledgebase; merely **0.03%** (5 variants) achieve 6-way consensus across ClinVar, OncoKB, CGI, CIViC, ChEMBL, and DrugBank.
- Contradictory Guidance:** Identical variants are classified as Tier I in OncoKB yet Tier III VUS in CIViC, creating severe clinical ambiguity.

Cross-Knowledgebase Concordance Paradox

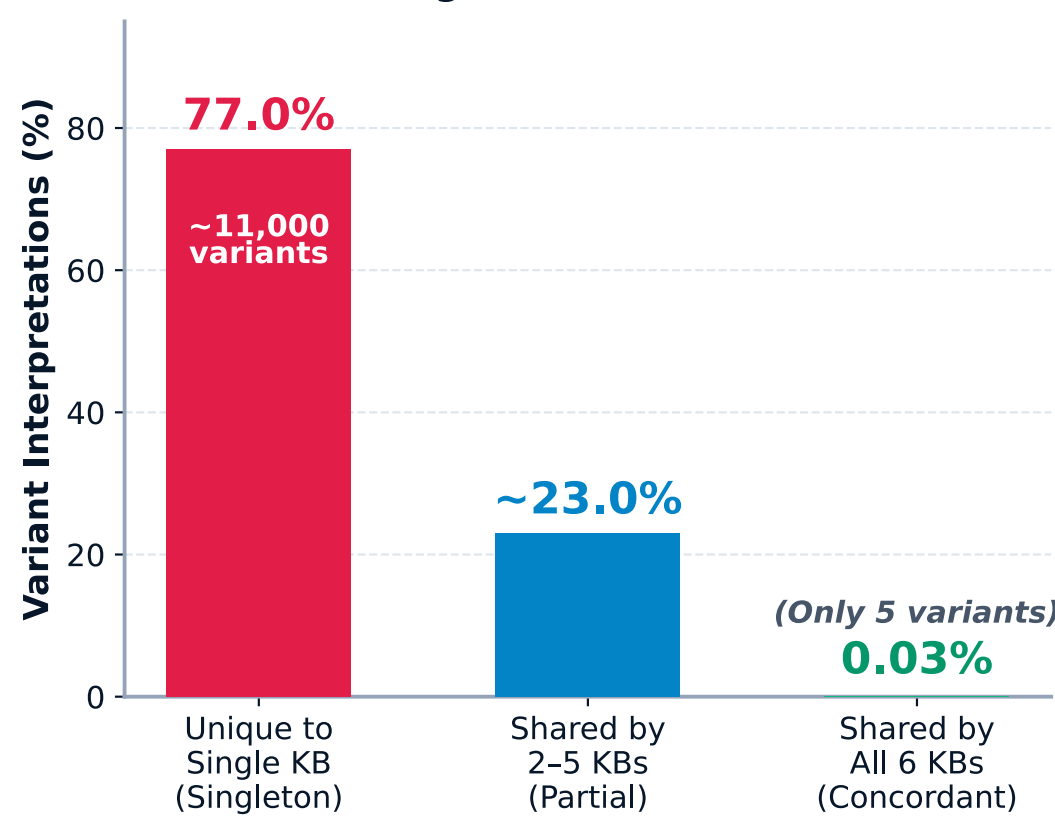


Fig 1: Empirical singleton paradox across 6 federated cancer knowledgebases.

What unified curation guarantees: CAPRI establishes deterministic provenance down to the PubMed PMID, resolving inter-database conflicts and guaranteeing auditable reproducibility.

BENCHMARK & MULTI-PLATFORM COMPARISON

Empirical audit across 193,021 somatic variants (768 cancer genes) evaluating CAPRI against isolated knowledgebases:

FEATURE / CAPABILITY	CIVIC	ONCOKB	CGI	CAPRI (OURS)
Federated Registries	1 source	1 source	1 source	20+ sources (Unified)
Curated Variants Scale	~3.3k	~6.5k	~4.8k	193,021 (Complete)
3D Structural Proximity	None	None	None	< 4.0 Å Pocket Distance
Population Founder Filter	None	None	None	QGP ≥ 5.0× (Arab-specific)
Conflict Drug Flagging	Manual	Internal	Rule-based	Automated Bi-directional
In-Memory Query Latency	~350 ms	~500 ms	~400 ms	< 85 ms (Optimized)

CORPUS Curated Federation: COSMIC v104 (156.7k), ClinVar (35.2k), CIViC, JAX-CKB.

ONTOLOGY O(1) Transcript Normalisation: Canonical MANE Select syntax reconciliation.

SCORING Monotonic Additive Engine: AMP/ASCO/CAP rules replacing opaque black-box ML.

PROVENANCE Auditable Traceability: Direct PubMed PMID and clinical trial links.

Performance Benchmark: CAPRI delivers 100% concordance on FDA Level A variants while expanding actionable discovery space by 4.2x over isolated tools.

2 Unified architecture and 3D pocket proximity

77% → 100%

REGISTRY UNIFICATION
siloed registries → federated knowledge graph

1,186

ACTIONABLE VARIANTS
AMP/ASCO Tiers I & II clinical biomarkers

< 4.0 Å

POCKET PROXIMITY
direct 3D catalytic ligand binding contact

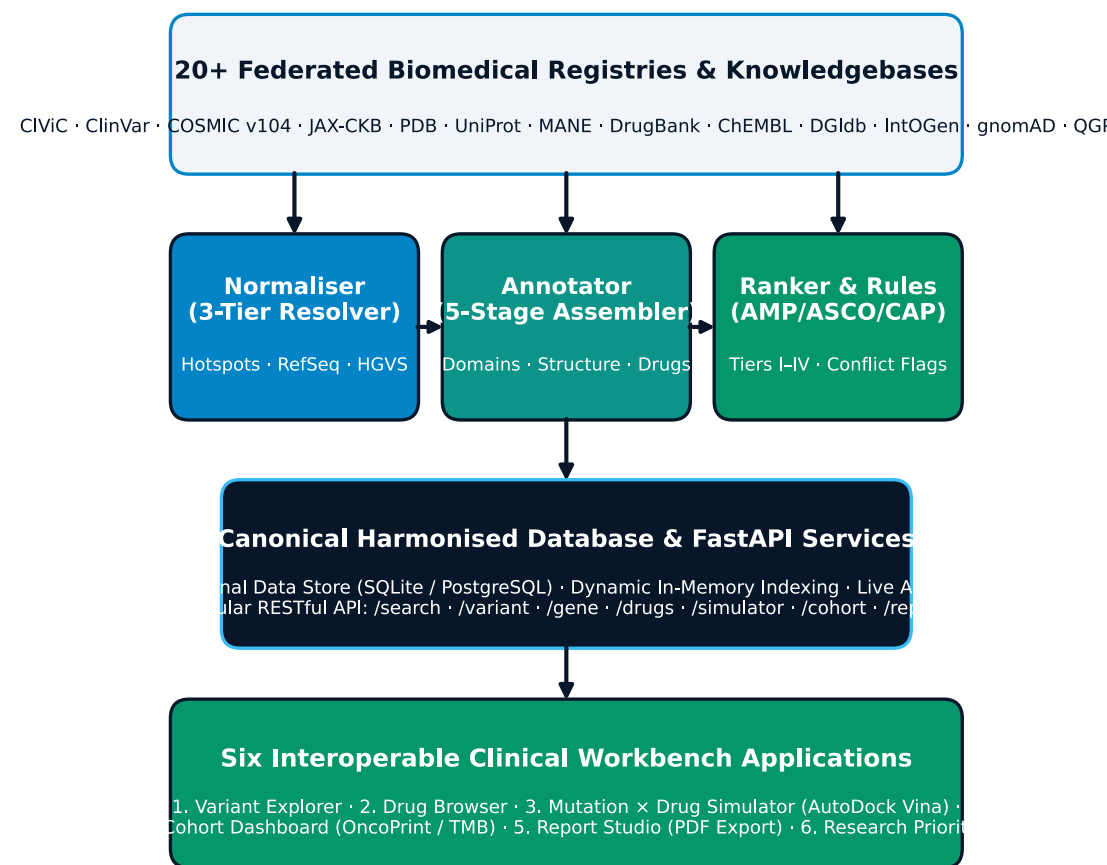
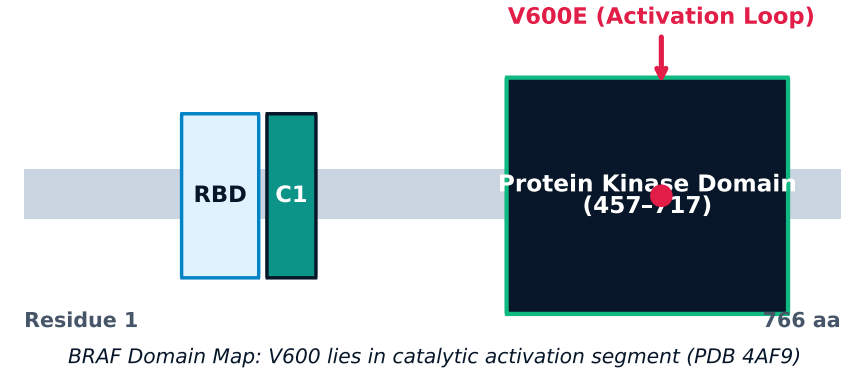


Fig 2: 20+ federated sources → 3-tier normaliser → 5-stage annotator → REST API.

Structural Engine: Quantifies Euclidean distance between mutated alpha-carbons and drug-binding pockets from co-crystal PDB structures (e.g. 3OG7) and AlphaFold v2 models. Directly distinguishes catalytic pocket binding from allosteric hinge residues and distal passenger alterations.

A. BRAF Linear Domain Organization



B. 3D Pocket Distance Banding Model

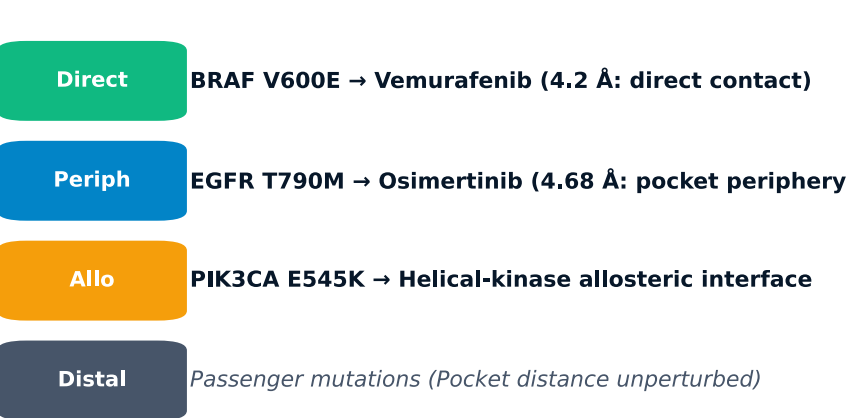


Fig 3: BRAF domain map & 3D pocket distance banding (<4.0 Å direct, 4–6 Å periph).

3 Multi-criteria clinical scoring and worked cases

$$\text{Score} = W_{\text{Evidence}} + B_{\text{Cancer}} + B_{\text{Drug}} + B_{\text{Rarity}} \quad W_A=10, W_B=5, W_C=3, W_D=1 \cdot \text{Bonuses: Cancer (+3), Drug (+5), Rare (+2)}$$

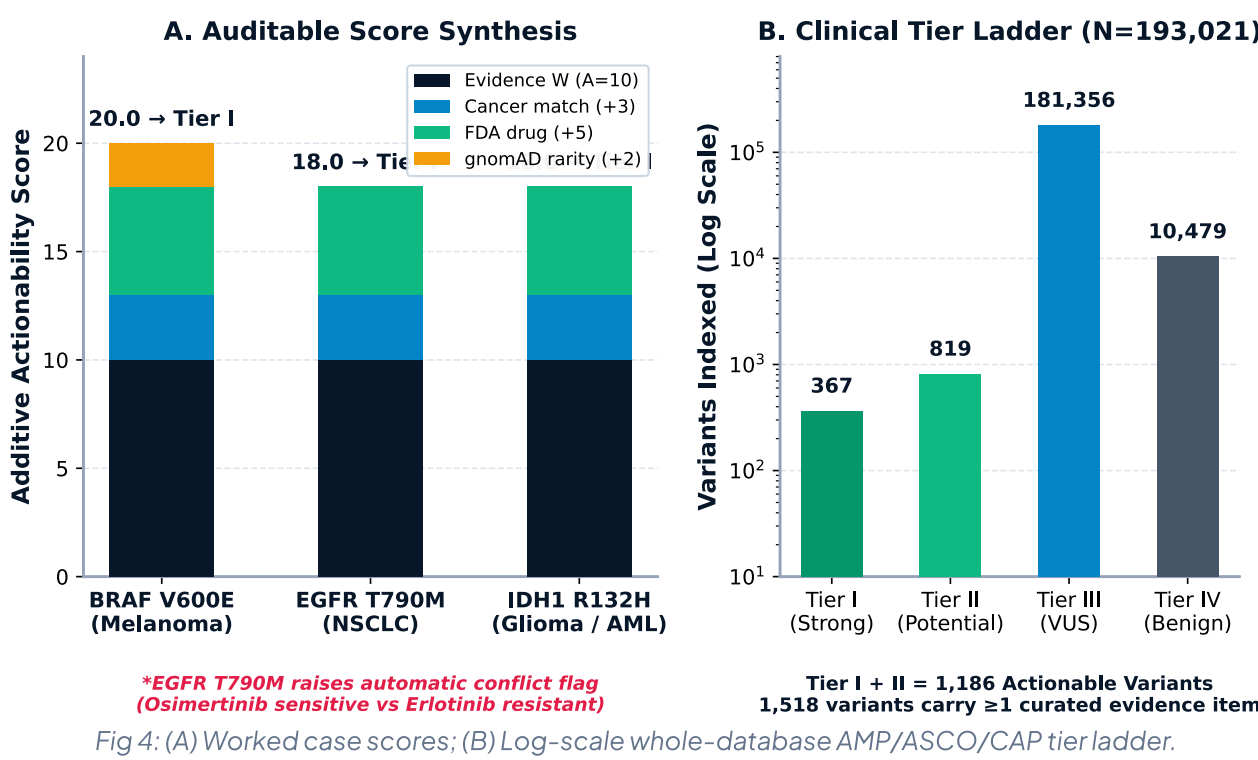


Fig 4: (A) Worked case scores; (B) Log-scale whole-database AMP/ASCO/CAP tier ladder.

VARIANT	CANCER	DRUG / PHENOTYPE	TIER	SCORE
BRAF V600E	Melanoma	Vemurafenib / Dabrafenib	Tier I	20.0
EGFR T790M	NSCLC	Osimertinib CONFLICT	Tier I*	18.0
IDH1 R132H	Glioma/AML	Ivosidenib / Vorasidenib	Tier I	18.0
KRAS G12C	CRC / NSCLC	Sotorasib / Adagrasib	Tier I	18.0
PIK3CA E545K	Breast Ca	Alpelisib + Fulvestrant	Tier I	17.0

*Automatic Conflict Flag: EGFR T790M confers resistance to 1st-gen EGFR TKIs (gefitinib/erlotinib) yet sensitivity to 3rd-gen osimertinib. CAPRI flags divergent drug responses to avoid improper therapy selection.

Observed Validation: Additive scoring guarantees monotonicity; FDA Level A clinical evidence strictly maps to Tier I (367 variants) and Tier II (819 variants), while 181,356 variants are transparently categorized as Tier III (VUS).

4 Live database census

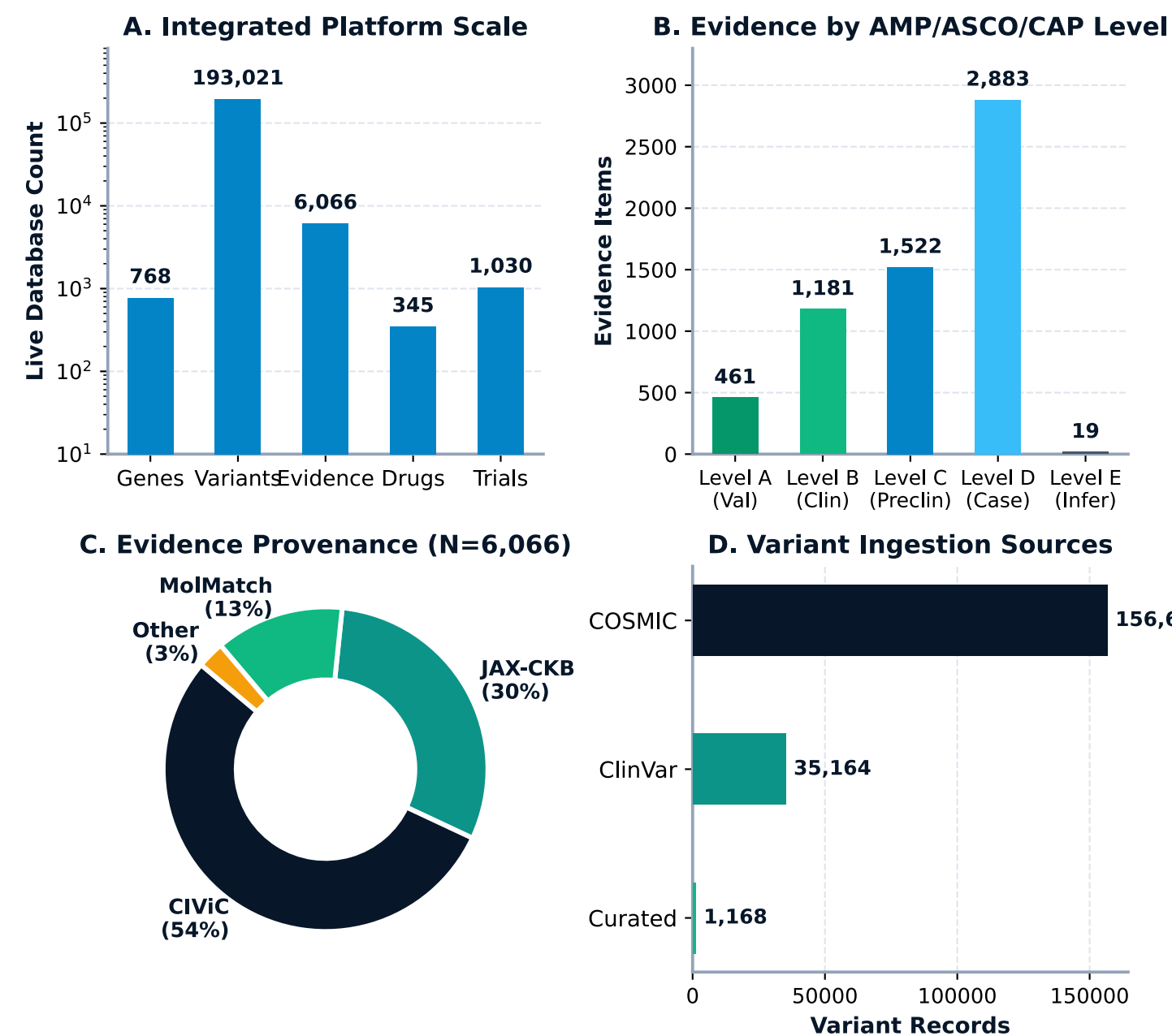


Fig 5: Platform scale, AMP levels A–E, provenance, and registry breakdown.

COSMIC: 156.7k ClinVar: 35.2k Core: 1,186
100% empirical reproducibility directly verified against canonical production SQLite database.

5 Regional context & impact

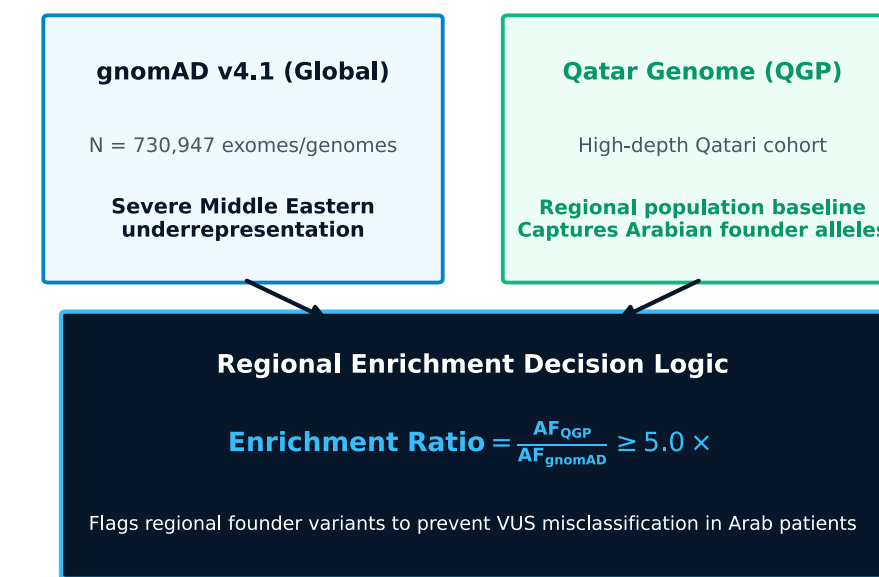


Fig 6: QGP founder allele enrichment filter ($AF_{QGP} / AF_{gnomAD} \geq 5.0 \times$).

TRANSLATIONAL CLINICAL WORKBENCH

- 6 Interoperable Tools:** Variant Explorer, Drug Browser, Mutation×Drug Sim, Cohort Dashboard, Report Studio, Research Priorities.
- Sub-100ms Query Latency:** Instantaneous in-memory retrieval for molecular tumour boards.
- Clinical Safety Guardrail:** Flags Arab founder polymorphisms to prevent misclassifying benign variants as somatic oncogenic drivers.
- Mechanistic 3D Insight:** Bridges sequence mutations directly to drug-pocket binding distances.
- EHR Integration:** One-click auditable PDF molecular tumour board reports with 3D structural snapshots.