VariantBench: An Agentic Benchmark for Genetic Variant Discovery and Interpretation

Aashka Bhowmick[*], Nasim Rahmatpour[*], Shivang Sharma[*], Qian Xu, Ananya Gupta, Asmita Lagwankar, Arjun Banerjee, Christopher Zou[*†]

LatchBio, San Francisco, CA, USA
[*]Equal contribution [†]Team lead

Correspondence: chris.zou@latch.bio

ABSTRACT

Analysis of data from genome sequencing and related assays is an increasingly important bottleneck in clinical and research settings. AI agents hold promise in automating and scaling such analysis. We present VariantBench, a verifiable benchmark of 118 evaluations that tests agents on their ability to conduct variant discovery and interpretation. Across 9,204 trajectories from 26 model-harness configurations, GPT-5.6 Sol / Codex and Claude Opus 4.8 / Pi led other configurations at a pass rate of 42.1% (149/354 attempts, 95% CI 33.8%–50.4%). Performance varies across problem domain, sequencing platforms, and read type, while a clinically validated case study suggests agents are not yet reliable in directing personalized medicine workflows. VariantBench provides a standard for measuring agentic performance in a domain seeing rapid AI adoption among researchers, clinicians, consumers, and patients alike.

Abbrevations in this Manuscript BCAA, branched-chain amino acid; CI, confidence interval; eQTL, expression quantitative trait locus; GWAS, genome-wide association study; HLA, human leukocyte antigen; LD, linkage disequilibrium; LINE, long interspersed nuclear element; LLM, large language model; ONT, Oxford Nanopore Technologies; QC, quality control; SNV, single-nucleotide variant; T2D, type 2 diabetes; VCF, Variant Call Format; WGS, whole-genome sequencing.

Introduction

The study of genetic variation is central to the study of every domain of life, from bacteria to pea plants [2]. In human biology, variant discovery and interpretation enables the development of new medicines, characterization of pathologies such as cancer and rare disease, and the study of genetic determinants of height, longevity, and other traits [3, 4, 5]. As in other omics-related domains, such work has historically been limited by sequencing costs, accuracy, and scalability. However, rapid improvements to sequencing and other assays have reduced this barrier [6]. The resulting data explosion has created exciting opportunities, such as population-scale biobanks and personalized medicine, while exposing new limitations in processing and analyzing biological data at scale [7, 8].

LLM-powered coding agents are ubiquitous in software engineering and offer a natural direction for automating and scaling biological data analysis [9, 10, 11, 12]. However, scientific data analysis requires more than writing software. Scientists explore messy datasets, iterate nonlinearly through tools and analysis steps, and draw careful conclusions. While doing so, they must balance domain-specific knowledge, software engineering concerns, and statistical rigor. To measure coding agent performance across these skills, we and others have published benchmarks spanning short, practical tasks in various domains (SpatialBench, SCBench, EpiBench), long-horizon tasks on real or rigorously simulated data (GeneBench-Pro, SCBench-Long, SpatialBench-Long, BiomniBench), tasks examining Internet use (BioMysteryBench), tasks requiring iterative molecular design (NucleoBench) and more [11, 12, 13, 14, 15, 16, 17, 18, 19]. To our knowledge no existing benchmark uses real-world data to examine agentic performance on variant discovery and interpretation.

Thus, we introduce VariantBench, an agentic benchmark of 118 evaluations across all stages of variant discovery and interpretation. Each VariantBench evaluation provides an agent with relevant files and a prompt that includes context and a gradeable field. Evaluations represent diverse input data types and effort levels across variant calling, clinical and personal genomics, and statistical and population genetics.

Across 9,204 trajectories from 26 model-harness pairs, the top-scoring models on their top-scoring harnesses are GPT-5.6 Sol and Claude Opus 4.8, with a 42.1% pass rate. The next best performing models are GPT-5.6 Terra, Claude Sonnet 5, and Grok 4.5. No model-harness pair exceeds a 50% trajectory pass rate. Manual examination of the failure modes suggests that while models are proficient at running routine pipelines, they struggle to inspect input data and adjust their analysis accordingly; ground conclusions in biological context; and integrate multiple data modalities.

Benchmark Design

We constructed benchmark evaluations based on independent reproduction of claims from peer-reviewed, publicly-available datasets. Each evaluation is built to be deterministically gradable, hardened against guessing and contamination, and robust across reasonable analysis choices. Numerical answers include tolerances based on multiple expert- or literature-validated workflows. Before inclusion in the benchmark, all evaluations underwent random expert review, screens for relevance and redundancy, and detailed trajectory scrutiny. See Methods for more details.

The final 118-problem suite covers three categories and 14 subcategories (Fig. 1a). Agents are called upon to perform QC and call variants across and within sequencing platforms, conduct and interpret statistical genetics analysis, and interpret genotypes in the context of disease, pedigrees, and supporting assays (Fig. 1b, c). We provide three representative evaluations in Table 1.

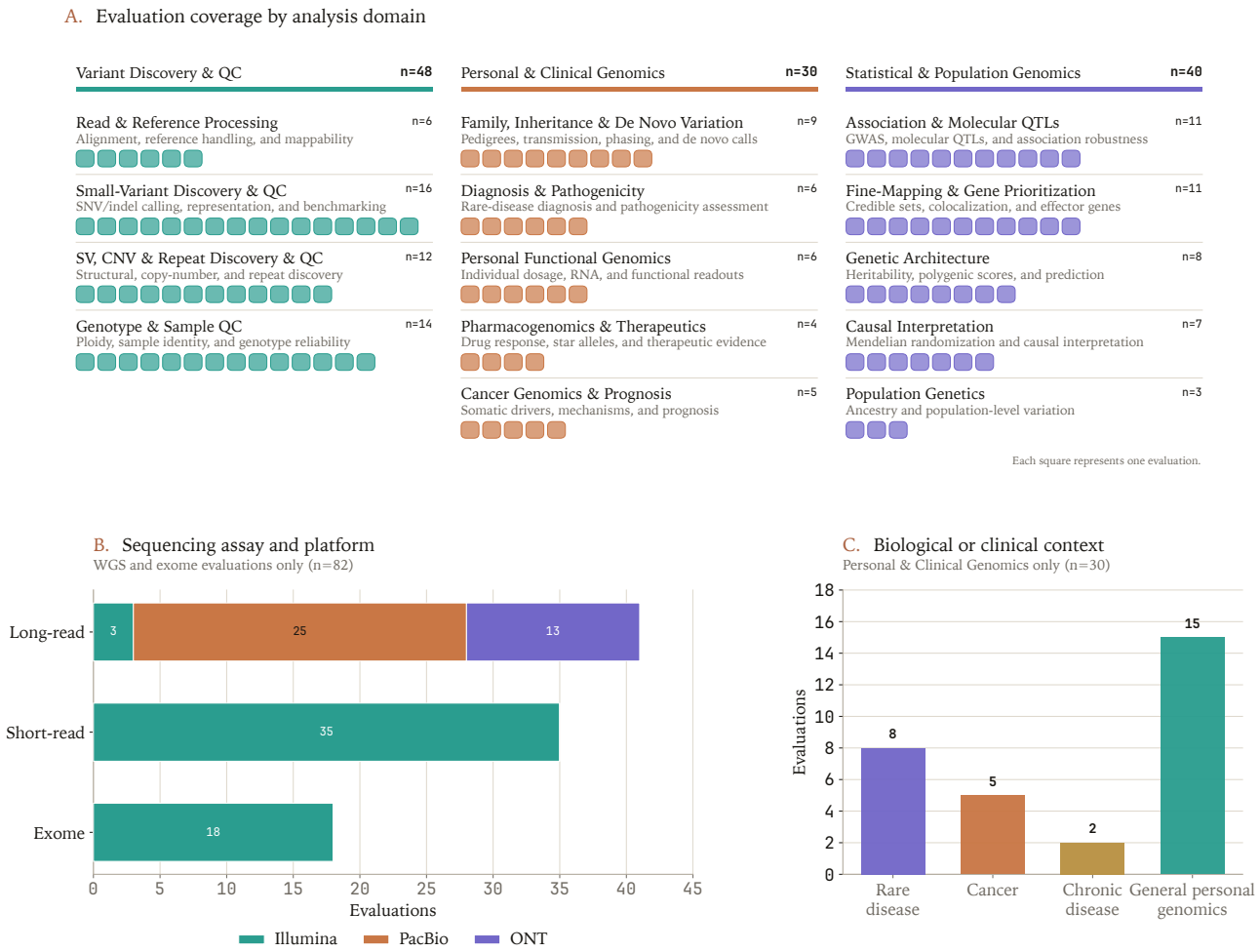

Figure 1: Composition of VariantBench. A, Overview of the 118 evaluations. **B,** Sequencing platform or data origin for relevant evaluations. Cross-platform WGS evaluations are counted once in each column where they have data; other origins include summary statistics, arrays, RNA sequencing, and derived tables or callsets without primary WGS or exome data. Five evaluations use mouse data, all from Illumina short-read WGS. **C,** Disease or clinical relevance among the 30 Personal & Clinical Genomics evaluations.

Example Evaluation	Provided Data	Ground Truth
Recognize phased neighboring variants that alter the same codon and adjust variant consequence annotations accordingly.	GATK UnifiedGenotyper Quartet joint VCF, aligned Illumina exome reads with phasing, reference genome, and associated BAM indexes [25].	Whether the agent recognizes a position-based caller's MNV limitations and performs haplotype-aware re-annotation on the full callset.
At an RFC1 repeat in an ONT genome, determine whether the expanded allele is pathogenic and report its G-base fraction.	ONT whole-genome read alignments over tandem-repeat regions, BAM index, and tandem-repeat region catalog [26].	Whether the agent identifies that there are no pathogenic alleles.
Use BCAA and T2D GWAS data to select *cis*-instrumented pathway proxies for BCKDK and determine whether there is genetic evidence for lowering BCAA levels reducing T2D risk.	European-ancestry BCAA and T2D GWAS summary statistics, an EUR LD reference, and locus/gene resources for instrument selection [27].	Whether the agent selects reasonable genes as instruments; whether it concludes that a small molecule drug that lowers BCAA does not necessarily lower T2D risk.

Table 1: Example VariantBench evaluations.

Results

Frontier Model Harness-Pairs Cluster at a 40% Pass Rate

We evaluated VariantBench across 26 agentic configurations. Each configuration features one model and one agentic harness (Claude Code, Pi, or Codex) that provides tools and prompt structure [20, 21, 22]. We evaluated configurations in a sandbox with approximately 170 top-line R, Python, and external software libraries, including both general-purpose scientific libraries and domain-specific tools.

No model-harness pair reached a 50% pass rate across attempts (Fig. 2a). GPT-5.6 Sol / Codex and Claude Opus 4.8 Max / Pi, the best-performing pairs, achieved pass rates of 42.1% (149/354 attempts, 95% CI 33.8–50.4). The best configurations featuring models from XAI and Google Deepmind, Grok 4.5 / mini-swe-agent and Gemini 3.5 Flash / Pi, passed 38.1% and 35.3% of attempts respectively. All configurations completely fail at least 44% of VariantBench evaluations, while Claude Opus 4.8 / mini-swe-agent and Claude Sonnet 5 / mini-swe-agent have the highest proportion of partial completions at 28.8% of evaluations (Fig. 2c).

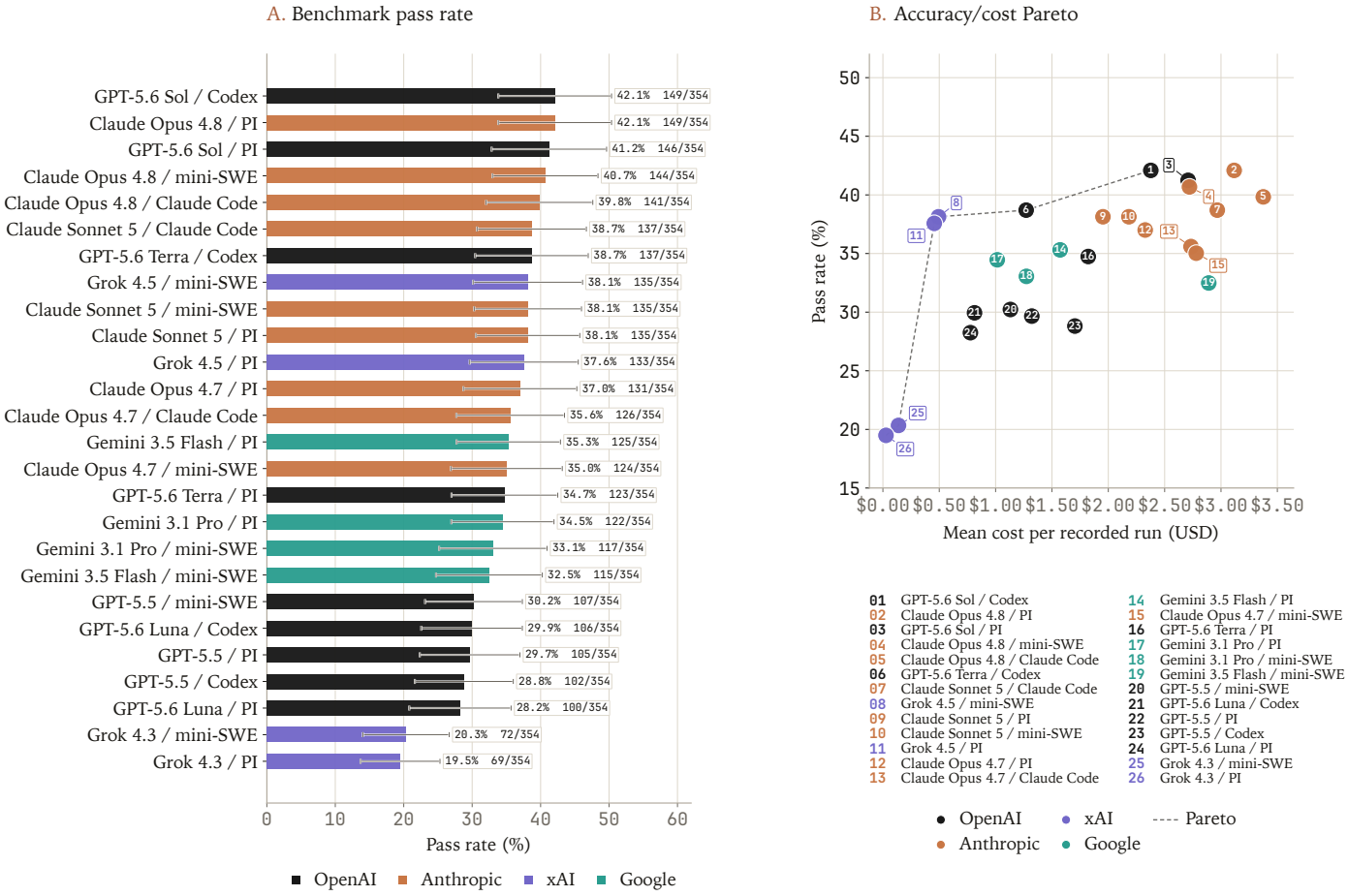

Figure 2: Benchmark performance across model and harness configurations. A, Overall pass rate for each of 26 model-harness configurations. **B,** Pass rate versus mean cost per recorded run; numbered dots map to the complete configuration labels below the plot.

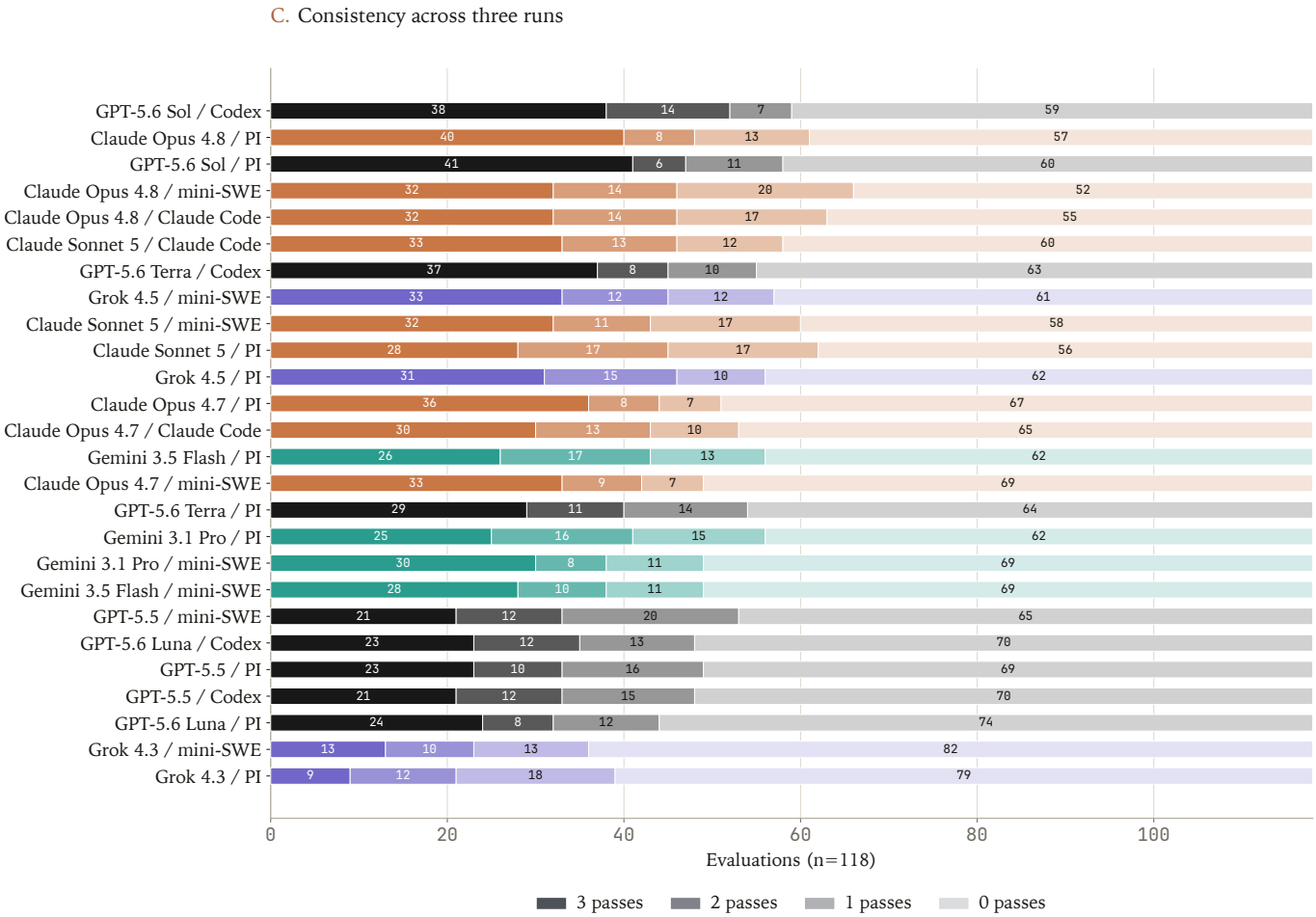

Figure 2: Benchmark performance across model and harness configurations (continued). C, Evaluations with three, two, one, or zero successful runs among three attempts.

Failure Mode Analysis

Averaged across configurations, 12 evaluation subcategories scored between 20% and 40% (Fig. 3A–C). Genotype & Sample QC scored lower (Fig. 3A), while Personal Functional Genomics scored higher, reaching 100% performance in some configurations (Fig. 3B, D). Pass rates additionally varied by sequencing read type and by sequencing platform, with evaluations involving long-read platforms and data proving more difficult (Fig. 3E, F).

Manual review of failure modes suggests that agentic systems have capability gaps in applying tacit knowledge, exploring data with the right priors, and exercising scientific judgement. Only Grok 4.5 and GPT-5.6 Sol recorded passes against an evaluation concerning difficulties with G-rich repetitive regions in long-read sequencing, despite the fact that trajectory examinations reveal models are aware of the concept. Across three evaluations requiring models to make decisions around variant calls based on LINE, Alu, or structural variant elements, the best configuration scored 44.4% (0/3 LINE, 3/3 Alu, 1/3 structural variants); across configurations, agents scored 1.3%, 17.9%, and 12.8% on attempts for LINE, Alu, and structural variant inspection respectively. An evaluation testing judgement in weighting *cis*-eQTL signals, *trans*-regulated genes, and credible-set variants to nominate a causal effector gene recorded an overall pass rate of 64.1% across all attempts (GPT-5.6 Sol / Codex: 2/3 attempts).

Interestingly, we find our results are highly robust to the sandbox agents run in (Fig. 3G). When the tool-rich sandbox environment was replaced by a minimal one containing only Python, R, and Mamba, we observed no significant difference in overall Claude Opus 4.8 / Pi and GPT-5.5 / Pi pass rates (42.1% vs.

43.8% and 29.7% vs. 29.9%, respectively). Combining results across the two configurations, we find that 73 evaluations are unchanged in results between sandbox choices. Of those that change, we find a minimal environment trades failures in Variant Discovery & QC for gains in Personal & Clinical Genomics and Statistical & Population Genomics.

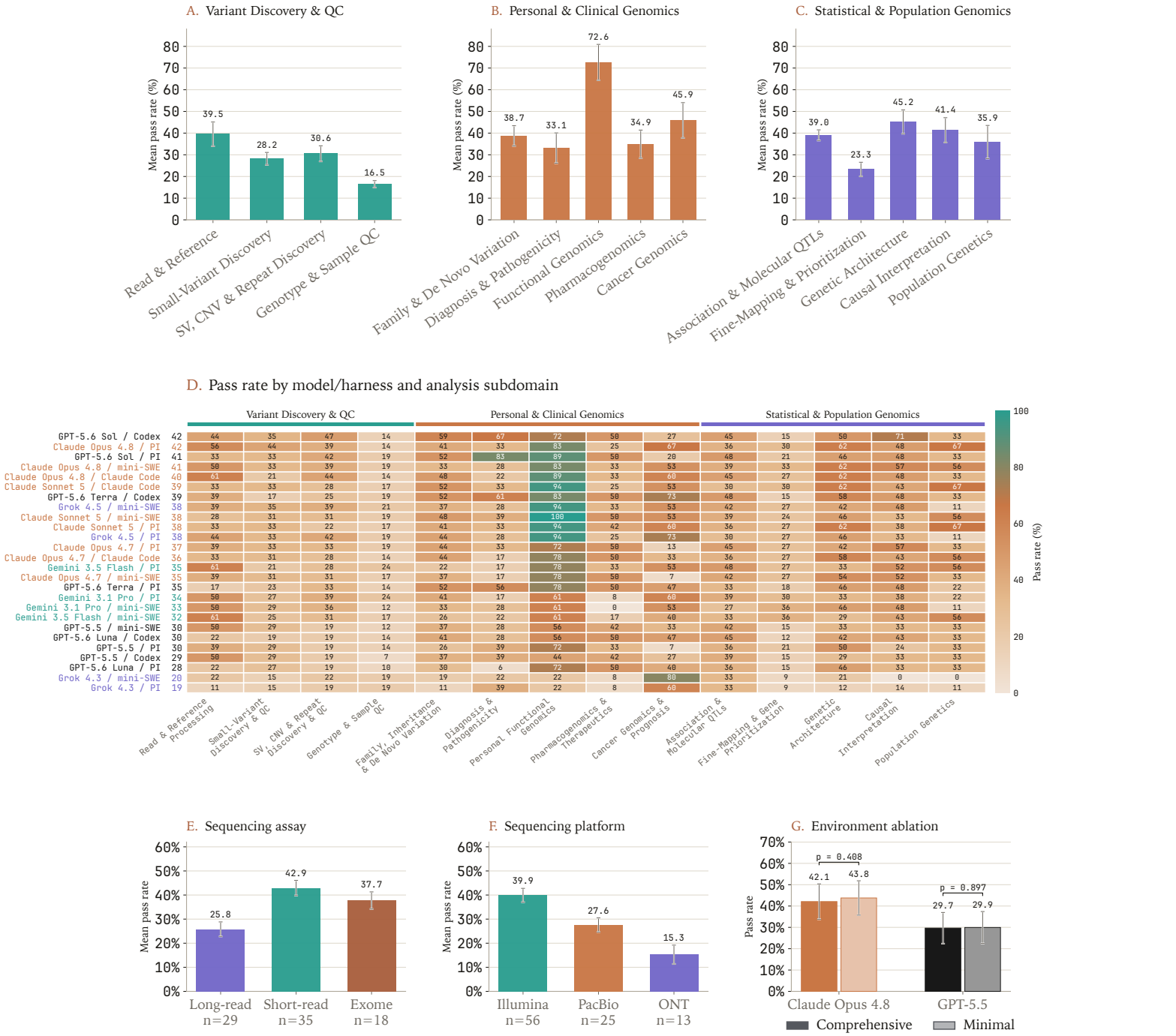

Figure 3: Performance across benchmark subsets and environments. A–C, Mean configuration-level pass rate for each analysis subcategory, grouped by broad category: (A) Variant Discovery & QC, (B) Personal & Clinical Genomics, and (C) Statistical & Population Genomics. **D,** Pass rate for each model–harness configuration across the 14 evaluation subcategories, with rows ordered by overall pass rate. **E,** Mean configuration-level pass rate for short-read WGS, long-read WGS, and exome evaluations. **F,** Mean configuration-level pass rate for evaluations using Illumina, PacBio, or ONT data. **G,** Pass rates with the comprehensive and minimal environments for Claude Opus 4.8 and GPT-5.5 using the Pi harness.

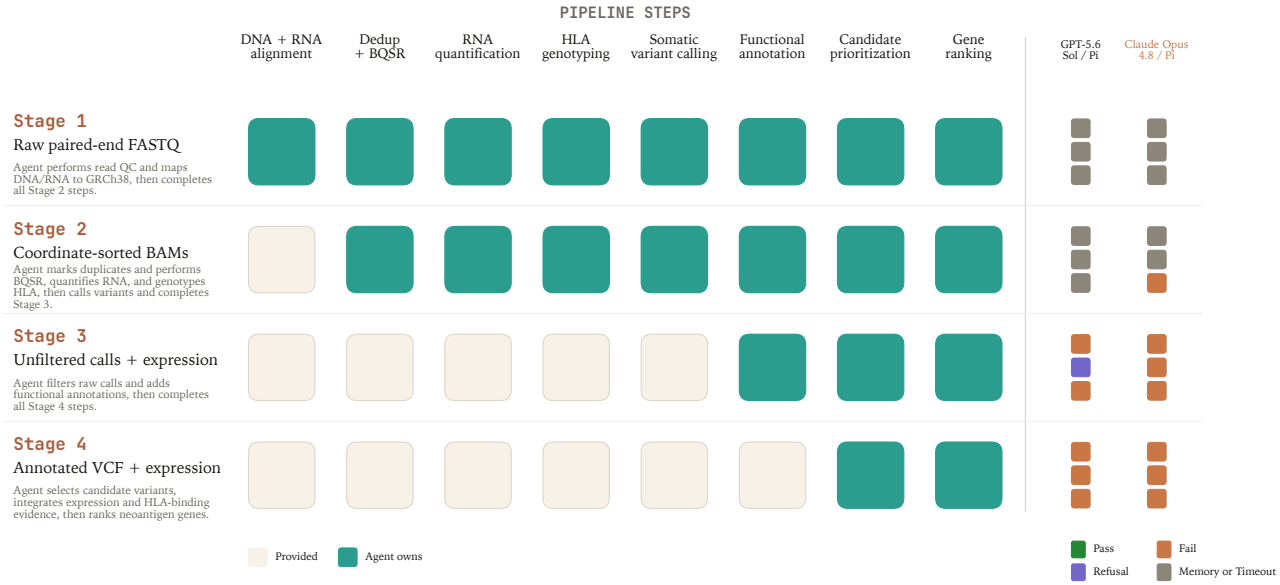

Figure 4: Performance on neoantigen design evaluations. A, Schematic of neoantigen design pipeline and evaluations. **B,** Performance on three attempts with GPT-5.6 Sol / Pi and Claude Opus 4.8 / Pi.

Agents Have Mixed Effectiveness on Real-World Neoantigen Design

Recent case studies have shown that treatments based on personalized omics measurements can dramatically improve outcomes compared to standard of care. In one notable example with publicly available data, Patient S was diagnosed with osteosarcoma and achieved remission by designing a cancer vaccine against neoantigens found in their cancer [23, 24].

We obtained consent from Patient S's care management team and designed four evaluations testing whether agents are capable of identifying target genes for neoantigen design based on a patient's tumor sequencing data. Patient S's vaccine design involved roughly eight pipeline steps (Fig. 4a). Our evaluations test agents' abilities to independently design and progress through this pipeline.

Across Stage 3 and Stage 4 evaluations, GPT-5.6 Sol / Pi and Claude Opus 4.8 / Pi failed across three trails to recall several of the neoantigens Patient S's care team progressed to experimental validaiton and eventually clinical use (Fig. 4b). Examining their trajectories, we found both agents set strict variant-allele-frequency cutoffs, excluded a MAP2 frameshift mutation by filtering on upstream variant calling tags, and excluded BTD, a higly expressed gene, based on its low-expression affected isoform. Thus, while both agent configurations produced plausible neoantigen workflows, a neoantigen design campaign based on this set of agent trajectories would have missed several candidates that were clinically useful.

Discussion

VariantBench measures whether LLM-powered agents can conduct practical work in variant discovery and interpretation across 118 evaluations. The strongest model-harness pairs pass 42.1% of evaluation attempts, suggesting that agents are useful but not yet reliable for such work. Examination of failure modes suggests that agents are capable of implementing and computing canonical analysis steps but struggle to make sound judgments when exploring data, choosing workflows, and interpreting results. These failure modes are conspicuous when considering the strongest model-harness pairs fail to rediscover clinically actionable neoantigens.

VariantBench has several limitations. First, we constructed VariantBench based on publicly available datasets. While this makes our benchmark easier to reproduce, it caused difficulties in acquiring genotyping data. VariantBench is thus missing canonical evaluations in some subcategories, such as calculating genome-wide association statistics or conducting in-sample fine-mapping. Moreover, we cannot exclude the possibility of training contamination in our results. To mitigate this possibility, we stripped identifying information from file names, headers, and obfuscated identifying study details where appropriate. These limitations point to promise in future work involving synthetic or consented real-world datasets.

Second, we used GPT-5.5 extensively during benchmark construction. While each benchmark evaluation represents a relevant analysis skill regardless of model-specific failures, the distribution of results suggests that we may have inadvertently identified weaknesses specific to GPT-5.5 that were resolved in subsequent OpenAI releases (Fig. 2a). We are actively investigating what these differences may be.

Most importantly, VariantBench limits grading for each evaluation to deterministic endpoint results. These are useful for measurement and interpretation, but we and others have observed that grading endpoints misses trajectory information and requires numeric tolerances to accommodate multiple correct workflows. Future work will involve benchmarks capable of measuring both longer-horizon and more open-ended signals.

Methods

Benchmark Composition and Data

Candidate evaluations were derived from published variant discovery and interpretation workflows with publicly available datasets. Each evaluation is defined as a single JSON file, including an evaluation prompt, a grader type, metadata, structured notes, and pointers to input data. Input data was defined such that each evaluation would test one major analysis concept. Identifying metadata and headers were removed from input data. Where possible, agents were provided with a superset of the data required to prevent the presence or absence of certain data from hinting at workflows.

During review, we retained candidate evaluations when their conclusions could be derived from scratch from the input data, were robust to tooling and workflow changes as long as the underlying scientific idea was unchanged, and were expressed well through a gradable endpoint answer. Candidates were revised or removed when their prompts over-specified the method, when their trajectories showed evidence of contamination, when multiple reasonable analysis choices produced incompatible answers or could not reproduce the answer, or when graders could not distinguish between plausible and implausible answers. After review, candidates were screened for broad relevance to variant discovery and interpretation and redundancy against existing problems.

Evaluation Format and Deterministic Grading

Evaluations are scored via a deterministic endpoint grading of the final structured answer. Each evaluation specifies a target answer schema. Agents return their answers as JSON, which is fed into a grading script. VariantBench features many types of graders, all built from typed field checks: numeric fields with tolerances, set-valued fields measured by Jaccard similarity; categorical fields with exact match; multiple choice; dictionary comparison; and more. When an evaluation specifies multiple graders, it receives a passing score only when all fields pass. Runs with unparseable fields were not counted as failures and instead rerun.

Agent Runs and Execution

We ran each of the 118 VariantBench evaluations three times across 26 model-harness configurations, for a total of 9,204 runs. A configuration is a model paired with Claude Code, Pi, or Codex, which gives a model tools, filesystem access, prompts, and turn structure. Each run executed in an identical containerized sandbox preloaded with Python, R, Mamba, and a suite of scientific and bioinformatics tools, except when running with a minimal sandbox, which excluded the suite of tools. Each sandbox provided six CPU cores, 34 GiB of memory, 128 GiB of disk, and a six-hour wall-clock limit. For our neoantigen pipeline evaluations, we provided sandboxes with 470 GiB of memory, 512 GiB of disk, and an eight-hour wall-clock limit to allow for memory-intensive alignment workflows. The sandboxes also provided internet access, so agents are were to manage their dependencies live. Each evaluation's data files were staged into the sandbox workspace with their file names only.

Outcomes and Statistical Analysis

The result set contains three attempts for every problem and every model-harness configuration. Several of the runs recorded timeouts at the six-hour wall-clock limit: we counted these as non-passing runs because our independent reproduction of every evaluation required far less time. Outside of our neoantigen pipeline evaluations, we observed no out of memory errors or unparseable answers during our runs. We calculate the endpoint pass rate as the number of correct runs over the sum of correct and incorrect runs. Uncertainty intervals are 95% Student-t intervals computed across per-evaluation mean pass rates, with the evaluation as the sampling unit.

Data Availability

Results files, selected evaluations, and representative trajectories will be made available at `https://github.com/latchbio/variantbench`.

References

[1] A full list of datasets, concepts, and other references used in VariantBench construction will be made available at `https://github.com/latchbio/variantbench`.

[2] Mendel, G. Experiments on Plant Hybridization. *Proceedings of the Natural History Society of Brünn* (1866).

[3] Rusina, P. V., Falaguera, M. J., Romero, J. M. R., McDonagh, E. M., Dunham, I. & Ochoa, D. Genetic support for FDA-approved drugs over the past decade. *Nat. Rev. Drug Discov.* **22**, 864 (2023). [Link].

[4] Miki, Y., Swensen, J., Shattuck-Eidens, D. et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**, 66–71 (1994). [Link].

[5] Cipriani, V., Vestito, L., Magavern, E. F. et al. Rare disease gene association discovery in the 100,000 Genomes Project. *Nature* (2025). [Link].

[6] Berger, B. & Yu, Y. W. Navigating bottlenecks and trade-offs in genomic data analysis. *Nat. Rev. Genet.* **24**, 235–250 (2023). [Link].

[7] All of Us Research Program Genomics Investigators. Genomic data in the All of Us Research Program. *Nature* **627**, 340–346 (2024). [Link].

[8] LatchBio. Manifesto. [Link].

[9] Huang, K., Zhang, S., Wang, H. et al. Biomni: A general-purpose biomedical AI agent. *Science* (2026). [Link].

[10] Mitchener, L., Laurent, J. M., Andonian, A., Tenmann, B., Narayanan, S., Wellawatte, G. P., White, A., Sani, L. & Rodriques, S. G. BixBench: A comprehensive benchmark for LLM-based agents in computational biology. *arXiv* arXiv:2503.00096 (2025).

[11] Qu, Y., Lu, Y., Tu, X., Zhang, S., She, T., Shaw, A. G., Shih, J.-H., Zhao, B. et al. BiomniBench: Process-level evaluation of LLM agents for real-world biomedical research. *bioRxiv* 2026.05.12.724604 (2026).

[12] Anthropic. Evaluating Claude for bioinformatics with BioMysteryBench. [Link].

[13] Workman, K., Yang, Z., Muralidharan, H. & Le, H. SpatialBench: Can agents analyze real-world spatial biology data? *arXiv* arXiv:2512.21907 (2025).

[14] Workman, K., Yang, Z., Muralidharan, H., Abdulali, A. & Le, H. scBench: Evaluating AI agents on single-cell RNA-seq analysis. *arXiv* arXiv:2602.09063 (2026).

[15] Muralidharan, H., Baskar, R., Lee, S. H., Proctor, T. & Workman, K. EpiBench: Verifiable evaluation of AI agents on epigenomics analysis. *arXiv* arXiv:2606.13602 (2026).

[16] Li, J. H. & Ho, A. J. GeneBench-Pro: Evaluating multistage statistical reasoning in genomics, quantitative biology, and translational biomedicine. *bioRxiv* 2026.06.29.735386 (2026).

[17] Diks, I., Yang, Z., Banerjee, A., Proctor, T. & Workman, K. scBench-Long: Verifiable benchmarking of long-horizon single-cell biology. *arXiv* arXiv:2606.26563 (2026).

[18] Diks, I., Muralidharan, H., Proctor, T. & Workman, K. Verifiable benchmarking of long-horizon spatial biology (SpatialBench-Long). *arXiv* arXiv:2605.28065 (2026).

[19] Shor, J., Strand, E. & McLean, C. Y. NucleoBench: A large-scale benchmark of neural nucleic acid design algorithms. *bioRxiv* 2025.06.20.660785 (2025).

[20] Anthropic. Claude Code. [Link].

[21] OpenAI. Introducing Codex. [Link].

[22] Zechner, M. Pi: A customizable coding-agent harness. [Link].

[23] Sijbrandij, S. Osteosarcoma data. [Link].

[24] Houston Methodist. Houston Methodist develops first personalized mRNA cancer vaccine for osteosarcoma (2026). [Link].

[25] Corpas, M., Valdivia-Granda, W., Torres, N. et al. Crowdsourced direct-to-consumer genomic analysis of a family quartet. *BMC Genomics* **16**, 910 (2015). [Link].

[26] Cortese, A., Simone, R., Sullivan, R. et al. Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. *Nature Genetics* **51**, 649–658 (2019). [Link].

[27] Tambets, R., Jesse, M., Kronberg, J. et al. Genetic analysis of circulating metabolic traits in 619,372 individuals. *Nature* (2026). [Link].