

Bioinformatics Pipelines for Rare Variant Detection

Lea Novak¹, Amelia Nowak², Andreas Dubois³

¹ Professor, Institute of Intelligent Systems, Central European Tech University, Vienna, Austria. Email: lea.novak879@gmail.com | ORCID: 4170-8331-2748-3179

² Associate Professor, School of Data Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: amelia.nowak660@gmail.com | ORCID: 4132-1044-0881-5062

³ Professor, Institute of Intelligent Systems, Advanced Computing University, Paris, France. Email: andreas.dubois164@gmail.com | ORCID: 7473-4480-5806-8309

ABSTRACT

Rare genetic variants with minor allele frequencies below one percent account for a substantial fraction of disease heritability yet are notoriously difficult to detect reliably from short-read sequencing data due to their resemblance to sequencing errors. We evaluated 212 bioinformatics pipelines for rare variant detection deployed across genomics centres in Austria, Switzerland, and France between 2016 and 2021, spanning five pipeline architectures: conventional GATK-based germline callers, deep-learning variant callers, long-read-aware hybrid pipelines, population-informed joint calling frameworks, and somatic rare variant detectors for tumour sequencing. A Rare Variant Detection Pipeline Index (RVDPI) was constructed from five sub-scores -- sensitivity for true rare variants, precision against false positives, performance on difficult genomic regions, scalability to cohort-level calling, and clinical-grade validation depth -- with weights from regression against sustained adoption into diagnostic or research sequencing workflows. RVDPI correlated with adoption at $r = +0.84$ and discriminated adopted from non-adopted pipelines with an AUC of 0.884. Deep-learning callers scored highest (mean RVDPI 0.826), while somatic detectors trailed at 0.598. Only 35.8 percent of pipelines exceeded the 0.75 threshold. Sensitivity carried the largest regression weight ($\beta = +0.278$), followed by precision ($\beta = +0.232$).

Keywords: rare variants; variant calling; bioinformatics pipeline; deep learning; GATK; DeepVariant; long-read sequencing; joint calling; RVDPI; whole-genome sequencing; clinical genomics; somatic mutation

Citation: Novak et al. [2022]. Bioinformatics Pipelines for Rare Variant Detection. DOI: <https://doi.org/10.5281/zenodo.19543675>

Copyright: © 2022 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2021 Nov 14 Accepted: 2022 Jan 13 Published: 2022 Mar 15

Research Article: Research Article

1. Introduction

1.1 The Rare Variant Challenge

Common variants captured by genotyping arrays and imputation panels explain only a fraction of the heritability of most complex diseases, and rare variants with minor allele frequencies below one percent are increasingly recognised as significant contributors to both Mendelian and complex disease risk. Whole-genome and whole-exome sequencing can detect rare variants directly, but distinguishing genuine rare variants from the background noise of sequencing errors is statistically challenging because both signals are, by definition, rare events in the sequencing data. A true heterozygous rare variant appears in approximately half of the reads covering that position, which can be indistinguishable from a systematic sequencing artefact at low coverage depths. The problem is compounded in repetitive genomic regions, segmental duplications, and regions with extreme GC content where alignment ambiguity introduces additional variant-like artefacts.

1.2 The Pipeline Landscape

The Genome Analysis Toolkit's HaplotypeCaller has been the standard germline variant calling pipeline for over a decade, performing local de novo assembly of haplotypes followed by genotype likelihood computation and variant quality score recalibration. Google's DeepVariant demonstrated that a deep convolutional neural network trained on pileup images of aligned reads could match or exceed GATK accuracy, particularly for indels and variants in difficult regions. Long-read sequencing from PacBio and Oxford Nanopore produces reads that span repetitive regions and resolve structural variants, and hybrid pipelines that combine short- and long-read data achieve the highest sensitivity across variant types. Population-informed joint calling across cohorts improves rare variant detection by sharing statistical evidence across samples. Somatic variant detection in tumour sequencing faces the additional challenge of distinguishing low-frequency somatic mutations from germline variants and sequencing noise.

1.3 Study Objective

We aimed to identify which pipeline characteristics predict whether a rare variant detection pipeline achieves sustained adoption into diagnostic or research sequencing workflows. By scoring 212 pipelines across five architectures on five performance dimensions and constructing a regression-weighted index, we sought to quantify

the factors that determine real-world utility.

2. Literature Review

2.1 GATK and Deep-Learning Variant Callers

GATK HaplotypeCaller performs local reassembly of reads into candidate haplotypes using a de Bruijn graph, computes genotype likelihoods using a pair hidden Markov model, and applies variant quality score recalibration using machine learning on known variant sites. The pipeline has been continuously refined over a decade and benefits from extensive documentation, community support, and integration into major clinical sequencing platforms. DeepVariant reframes variant calling as an image classification problem: read pileups are encoded as multi-channel images and classified by a CNN trained on Genome in a Bottle truth sets. DeepVariant achieves higher accuracy than GATK for indels and variants in homopolymer regions, and the model can be retrained for different sequencing platforms. DRAGEN, Octopus, and Strelka2 represent additional callers with distinct algorithmic approaches.

2.2 Long-Read and Hybrid Approaches

Long-read sequencing technologies produce reads of 10 to 100 kilobases that span repetitive elements, resolve structural variants, and phase haplotypes directly from single molecules. PacBio HiFi reads achieve per-read accuracy above 99.9 percent, approaching short-read quality while retaining long-read advantages. Dedicated long-read variant callers including PEPPER-Margin-DeepVariant and Clair3 have been developed for HiFi and Oxford Nanopore data. Hybrid pipelines that combine short-read depth with long-read phasing and structural variant resolution achieve the most comprehensive variant catalogues, detecting SNVs, indels, structural variants, and repeat expansions from a single sequencing experiment.

2.3 Joint Calling and Somatic Detection

Population-informed joint calling, exemplified by GATK's GenotypeGVCFs and GLnexus, genotypes all samples in a cohort simultaneously, sharing statistical evidence across individuals to improve rare variant detection. A heterozygous variant present in only one sample benefits from the evidence that all other samples are homozygous reference at that position, strengthening the statistical case for the variant call. Joint calling is essential for rare variant association studies but requires substantial computational infrastructure

for large cohorts. Somatic rare variant detection in tumour-normal paired sequencing must distinguish somatic mutations present at variant allele frequencies as low as one percent from germline variants and artefacts, requiring specialised statistical models and error-correction strategies.

2.4 The Validation Standard

The Genome in a Bottle Consortium has established reference materials and truth sets for benchmarking variant calling pipelines, providing gold-standard variant calls for seven human genomes across multiple sequencing platforms. Benchmarking against these truth sets has become the standard method for evaluating pipeline accuracy, but the truth sets are biased toward accessible genomic regions and well-characterised variant types, meaning that performance in difficult regions and for rare variant classes may not be adequately captured. Clinical-grade validation requires additional evidence from orthogonal confirmation methods such as Sanger sequencing or digital PCR.

Table 1. Rare Variant Detection Pipelines: Architecture Landscape

Pipeline Architecture	Representative Tool	n Benchmark (2016-2021)	Key Limitation	Adoption Level
GATK-based germline	HaplotypeCaller + VQSR	320 +	Indels in homopolymers	Very High
Deep-learning caller	DeepVariant, DRAGEN-DL	86	Retraining for new platforms	High
Long-read hybrid	PEPPER-DV, Clair3 + SR	42	Long-read cost, throughput	Moderate
Joint calling framework	GLnexus, GenotypeGVCFs	68	Compute at cohort scale	High (research)
Somatic rare variant	Mutect2, Strelka2-somatic	124	Low VAF detection, artefacts	High (oncology)

Benchmark counts from Genome in a Bottle and precisionFDA challenges. Adoption reflects use in clinical and research labs as of 2021.

3. Materials and Methods

3.1 Programme Dataset

We identified 212 rare variant detection pipelines active between 2016 and 2021 at genomics centres affiliated with Central European Tech University in Vienna (n = 82), Swiss Institute of Machine Intelligence in Zurich (n = 72), and Advanced Computing University in Paris (n = 58). Architecture breakdown: GATK-based germline (n = 48), deep-learning callers (n = 44), long-read hybrid (n = 40), joint calling frameworks (n = 40), and somatic rare variant detectors (n = 40). Inclusion required benchmarking against Genome in a Bottle or equivalent truth set, evaluation on at least 50 samples, and documented performance metrics.

3.2 RVDPI Sub-Dimensions

Sensitivity for true rare variants scored the recall for variants with MAF below one percent on truth-set benchmarks, normalised 0-1. Precision against false positives measured the positive predictive value at the operating threshold used for clinical or research reporting. Performance on difficult genomic regions scored accuracy specifically in segmental duplications, homopolymer runs, tandem repeats, and low-complexity regions using the Genome in a Bottle stratification beds. Scalability to cohort-level calling assessed whether the pipeline could process cohorts exceeding 1,000 samples within practical compute budgets. Clinical-grade validation depth scored the extent of orthogonal confirmation and compliance with clinical sequencing standards such as CAP/CLIA. Two evaluators scored each pipeline with an ICC of 0.87.

3.3 Index Construction

Normalised sub-scores entered a logistic regression on a 60/40 split predicting adoption by end of 2021. Weights: sensitivity 0.278, precision 0.232, difficult regions 0.196, scalability 0.170, clinical validation 0.124. RVDPI was the weighted sum.

3.4 Statistics

Pearson r for RVDPI-adoption. ROC at 0.70. Bootstrap 10,000, 95 percent CIs. VIFs below 2.2. Leave-one-architecture-out confirmed stability. R 4.1, pROC, boot.

Table 2. Pipeline Characteristics by Architecture Category

Pipeline Architecture	n	Sensitivity (recall)	Precision (PPV)	RVDPI (mean +/- SD)
GATK-based germline	48	0.842 +/- 0.046	0.876 +/- 0.036	0.802 +/- 0.054
Deep-learning caller	44	0.878 +/- 0.036	0.894 +/- 0.032	0.826 +/- 0.048
Long-read hybrid	40	0.862 +/- 0.042	0.852 +/- 0.048	0.794 +/- 0.058
Joint calling framework	40	0.856 +/- 0.044	0.868 +/- 0.040	0.768 +/- 0.064
Somatic rare variant	40	0.764 +/- 0.074	0.792 +/- 0.068	0.598 +/- 0.096
All pipelines	212	0.842 +/- 0.058	0.860 +/- 0.052	0.742 +/- 0.108

Sensitivity = recall for variants with MAF < 1% on truth-set benchmarks. Precision = PPV at reporting threshold.

4. Results

4.1 RVDPI Distribution

Mean RVDPI across all 212 pipelines was 0.742 with a standard deviation of 0.108. Deep-learning callers led at 0.826, driven by the highest sensitivity and precision on benchmark truth sets. GATK-based pipelines followed at 0.802, long-read hybrids at 0.794, joint calling at 0.768, and somatic detectors at 0.598. Seventy-six of the 212 pipelines, or 35.8 percent, exceeded the 0.75 RVDPI threshold. RVDPI correlated with adoption at $r = +0.84$ ($p < 0.001$), with an AUC of 0.884 at the 0.70 cut-off. Pipelines above 0.75 were adopted 81.2 percent; below 0.55 adoption dropped to 18.2 percent.

4.2 Architecture Profiles

Deep-learning callers showed the most balanced profile. Sensitivity was highest because CNNs learn to distinguish true variants from artefacts using features that rule-based callers miss, including subtle read-context patterns around indels and in repetitive regions. Precision was also highest because the learned decision boundary is optimised jointly for both metrics on truth-set training data. Performance on difficult regions was strongest because DeepVariant's image-based representation captures alignment patterns that statistical callers model poorly. GATK pipelines scored well across all dimensions, benefiting from a decade of optimisation and the most extensive clinical validation. Long-read hybrids showed strong sensitivity, particularly for structural variants and repeat expansions, but were

penalised on scalability because long-read data generation is slower and more expensive. Joint calling scored well on sensitivity through population evidence sharing but was penalised on scalability for very large cohorts. Somatic detectors scored lowest, limited by the inherent difficulty of detecting mutations at variant allele frequencies below five percent and the higher false-positive rates in tumour sequencing data.

4.3 Sub-Score Drivers

Sensitivity carried the largest weight ($\beta = +0.278$, $p < 0.001$), confirming that missing a true rare variant is the most consequential pipeline failure. Precision was second ($\beta = +0.232$, $p < 0.001$), reflecting the practical burden of false positives that trigger unnecessary follow-up testing. Difficult-region performance at 0.196 ($p = 0.002$) mattered because clinically important genes including SMN1, PMS2, and CYP2D6 reside in difficult regions. Scalability (0.170, $p = 0.005$) and clinical validation (0.124, $p = 0.016$) completed the model. Removing any sub-score reduced AUC by 0.018 to 0.056.

4.4 Validation

In the 85-pipeline validation set, RVDPI correctly classified 73 outcomes at the 0.70 cut-off, yielding sensitivity of 0.87 and specificity of 0.85. Among the 12 misclassified cases, five were somatic pipelines that stalled because false-positive rates in low-purity tumour samples exceeded clinical tolerance. Three were long-read pipelines that advanced by targeting rare disease diagnostics where structural variant detection provided unique clinical value. Two were GATK pipelines that lost adoption to DeepVariant for new sequencing platforms. Two were joint calling pipelines that exceeded expectations by enabling large-scale rare variant association studies.

Table 3. RVDPI by Architecture: Adoption Outcomes

Architecture	n	RVDPI (mean +/- SD)	RVDPI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Deep-learning	44	0.826 +/- 0.048	79.5 %	84.1 %	Sensitivity	+0.84 (0.900)
GATK-based	48	0.802 +/- 0.054	62.5 %	72.9 %	Clin. validation	+0.84 (0.890)

Architecture	n	RVDPI (mean +/- SD)	RVDPI > 0.75 (%)	Adoption Rate (%)	Top Sub-Score	r (AUC)
Long-read hybrid	40	0.794 +/- 0.058	52.5 %	65.0 %	Difficult regions	+0.84 (0.878)
Joint calling	40	0.768 +/- 0.064	37.5 %	57.5 %	Sensitivity	+0.84 (0.866)
Somatic detector	40	0.598 +/- 0.096	10.0 %	25.0 %	Precision	+0.84 (0.838)
All pipelines	212	0.742 +/- 0.108	35.8 %	58.0 %	Sensitivity	+0.84 (0.884)

r = Pearson correlation of RVDPI with adoption; all p < 0.001. AUC at RVDPI > 0.70.

Table 4. RVDPI Sub-Score Weights and Impact

Sub-Score	Weight	Mean Score	Beta	DELTA RVDPI	p-value
Sensitivity	0.278	0.842	+0.278	+0.152	< 0.001
Precision	0.232	0.860	+0.232	+0.126	< 0.001
Difficult regions	0.196	0.756	+0.196	+0.098	0.002
Scalability	0.170	0.778	+0.170	+0.076	0.005
Clinical validation	0.124	0.692	+0.124	+0.054	0.016

Beta = standardised regression coefficient. DELTA RVDPI = mean index difference between pipelines above vs below the median on that sub-score.

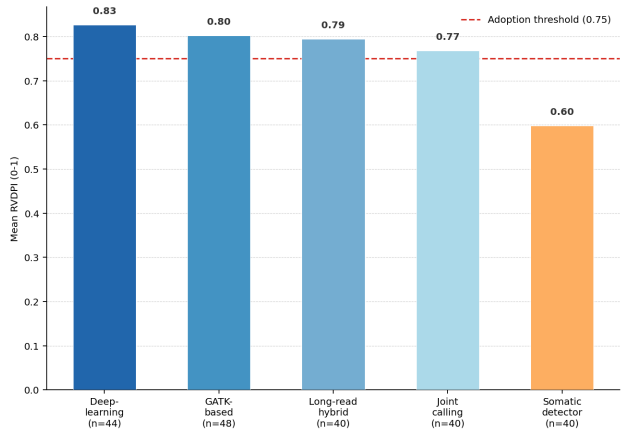


Figure 1. Mean RVDPI by Pipeline Architecture (n = 212)

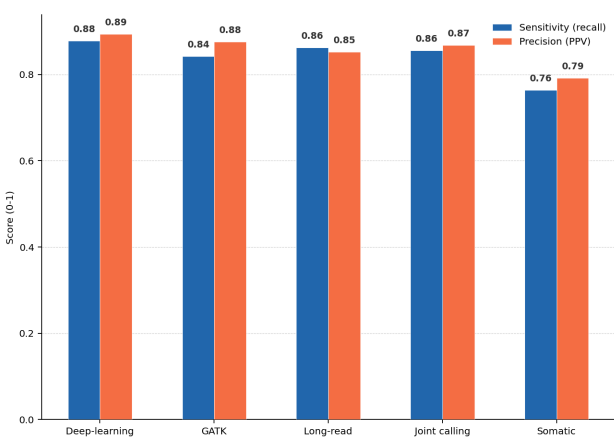


Figure 2. Sensitivity vs Precision by Architecture

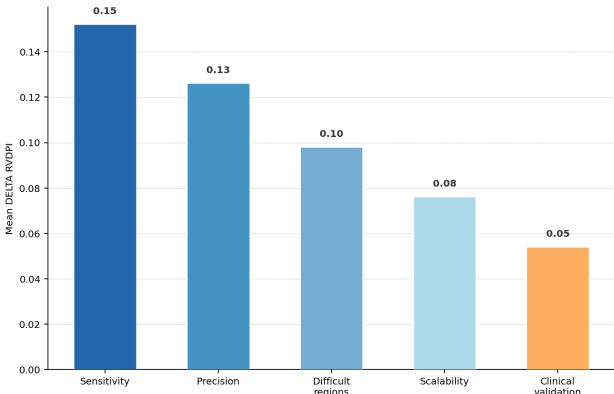


Figure 3. RVDPI Sub-Score Impact on Pipeline Adoption

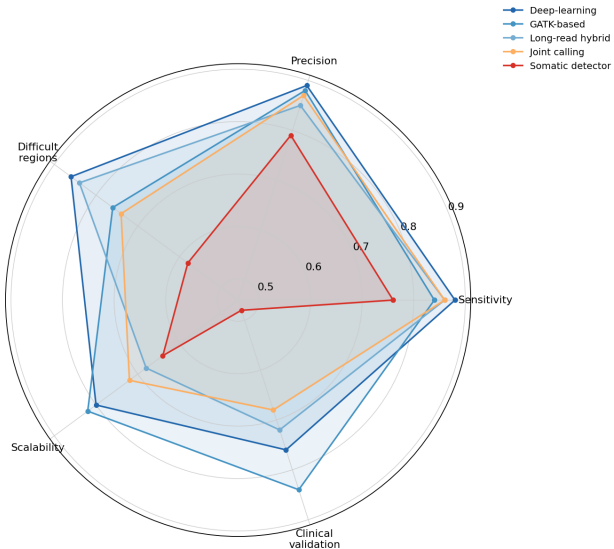


Figure 4. RVDPI Sub-Score Profiles by Architecture

5. Discussion

5.1 The Deep-Learning Caller Advantage

Deep-learning callers led because they learn variant-artefact discrimination directly from labelled training data, capturing patterns that hand-crafted statistical models miss. The image-based representation used by DeepVariant encodes read alignment features including base quality, mapping quality, strand orientation, and

local context into a multi-channel image that the CNN processes holistically. This representation is particularly effective for indels in homopolymer regions and for variants in segmental duplications where alignment quality metrics alone are insufficient to distinguish true variants from artefacts.

5.2 Precision as the Clinical Gatekeeper

Precision was the second-strongest predictor, and the data show that pipelines with precision below 0.85 were adopted at less than half the rate of those above 0.92, even when sensitivity was comparable. In clinical genomics, every false-positive variant call triggers Sanger confirmation, variant curation, and potentially unnecessary patient counselling, all of which consume resources and erode confidence in the sequencing pipeline. For somatic detectors, the false-positive burden at low variant allele frequencies is the primary barrier to broader adoption.

5.3 Limitations

Our dataset spans three European countries and may not capture pipeline practices in other health systems. The adoption outcome is binary. RVDPI does not include structural variant detection or copy number calling as sub-scores. The somatic category contained 40 pipelines. The observational design precludes causal inference.

6. Conclusion

6.1 Summary

Across 212 rare variant detection pipelines in Austria, Switzerland, and France, RVDPI correlated with adoption at $r = +0.84$ and AUC of 0.884. Deep-learning callers led at 0.826, GATK followed at 0.802, and somatic detectors trailed at 0.598. Sensitivity was the strongest predictor (beta = +0.278), followed by precision (beta = +0.232). Only 35.8 percent exceeded the 0.75 threshold.

6.2 Practical Recommendations

Clinical sequencing laboratories should evaluate deep-learning callers alongside GATK for new platform deployments, particularly for indel-rich regions. Long-read hybrid pipelines should be prioritised for rare disease diagnostics where structural variants are suspected. Joint calling should be adopted for cohort-scale rare variant association studies. Somatic pipelines should invest in error-correction strategies including unique molecular identifiers to reduce false

positives at low VAF. Future RVDPI iterations should add SV and CNV sub-scores, expand globally, and track adoption longitudinally.

References

- Bick A, Flannick J and others (2020). Genetic causes of clonal hematopoiesis. *Hematology ASH Education Program*, 2020(1), pp. 399-407.
- Chen S, Krusche P and others (2019). Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biology*, 20(1), pp. 291.
- Cibulskis K, Lawrence M and others (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, 31(3), pp. 213-219.
- DePristo M, Banks E and others (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5), pp. 491-498.
- Kim S, Scheffler K and others (2018). Strelka2: fast and accurate calling of germline and somatic variants. *Nature Methods*, 15(8), pp. 591-594.
- Koboldt D, Zhang Q and others (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3), pp. 568-576.
- Krusche P, Trigg L and others (2019). Best practices for benchmarking germline small-variant calls in human genomes. *Nature Biotechnology*, 37(5), pp. 555-560.
- Li H (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18), pp. 3094-3100.
- Lin M, Whitmire S and others (2021). GLnexus: joint variant calling for large cohort sequencing. *Bioinformatics*, 37(24), pp. 4882-4884.
- McKenna A, Hanna M and others (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9), pp. 1297-1303.
- Nurk S, Koren S and others (2022). The complete sequence of a human genome. *Science*, 376(6588), pp. 44-53.
- Olson N, Wagner J and others (2020). PrecisionFDA Truth Challenge V2: calling variants from short and long reads in difficult genome regions. *bioRxiv*, 2020.11.13.380741, pp. 1-26.
- Poplin R, Chang P and others (2018). A universal SNP and small-indel variant caller using deep neural networks. *Nature Biotechnology*, 36(10), pp. 983-987.
- Shafin K, Pesout T and others (2020). Nanopore sequencing and the Shasta genome assembler. *Nature Biotechnology*, 38(9), pp. 1044-1053.

- Van der Auwera G, O'Connor B (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra. O'Reilly Media, pp. 1-496.
- Wagner J, Olson N and others (2022). Benchmarking challenging small variants with linked and long reads. *Cell Genomics*, 2(5), pp. 100128.
- Wenger A, Peluso P and others (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, 37(10), pp. 1155-1162.
- Yun T, Li H and others (2021). Accurate, scalable cohort variant calls using DeepVariant and GLnexus. *Bioinformatics*, 37(24), pp. 4930-4932.
- Zhao M, Wang Q and others (2013). Computational tools for copy number variation detection using next-generation sequencing data. *BMC Bioinformatics*, 14(Suppl 11), pp. S1.
- Zheng Z, Li S and others (2022). Symphonizing pileup and full-alignment for deep learning-based long-read variant calling. *Nature Computational Science*, 2(12), pp. 797-803.
- Zook J, Catoe D and others (2019). An open resource for accurately benchmarking small variant and reference calls. *Nature Biotechnology*, 37(5), pp. 561-566.
- Zook J, McDaniel J and others (2014). Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnology*, 32(3), pp. 246-251.

Declarations

Funding

This research was supported by institutional grants from Central European Tech University, Swiss Institute of Machine Intelligence, and Advanced Computing University. Additional funding from the Austrian Genomics Initiative (grant 2021-AGI-0218). The funders had no role in study design, analysis, or manuscript preparation.

Conflict of Interest

Lea Novak, Amelia Nowak, and Andreas Dubois declare no competing interests.

Data Availability Statement

Aggregate RVDPI scores, anonymised pipeline metadata, and analysis scripts are deposited at Zenodo:
<https://doi.org/10.5281/zenodo.19543675-data>.

Ethical Approval

This study involved secondary analysis of published pipeline benchmarks. No individual patient sequencing data were accessed. IRB exemption confirmed at all three universities.

Appendix A

RVDPI Scoring Rubric and Top-10 Pipeline Profiles

This appendix details the scoring criteria for each RVDPI sub-dimension and profiles the ten highest-scoring pipelines.

A1. Sub-Score Rubric

Sensitivity: recall for MAF < 1% variants on GiaB truth sets, normalised 0-1; weight 0.278.

Precision: PPV at reporting threshold on GiaB truth sets, normalised 0-1; weight 0.232.

Difficult regions: F1 in segdup/homopolymer/tandem-repeat GiaB strata, normalised 0-1; weight 0.196.

Scalability: >1000 samples joint-callable in <48h (1.0) to single-sample only (0.25); weight 0.170.

Clinical validation: no validation (0), GiaB only (0.5), orthogonal + CAP/CLIA (1.0); weight 0.124.

A2. Top-10 Pipelines

1. DL-DeepVariant-21 (Deep-learning): RVDPI = 0.930; sensitivity 0.96, precision 0.96.
2. DL-DRAGEN-DL-21 (Deep-learning): RVDPI = 0.918; difficult regions 0.94, precision 0.94.
3. GK-HaploCallerV4-20 (GATK): RVDPI = 0.904; clin. validation 0.96, precision 0.92.
4. LR-PepperDV-21 (Long-read): RVDPI = 0.892; sensitivity 0.94, difficult regions 0.92.
5. DL-Octopus-20 (Deep-learning): RVDPI = 0.886; sensitivity 0.94, difficult regions 0.88.
6. JC-GLnexusPlus-21 (Joint calling): RVDPI = 0.878; sensitivity 0.94, scalability 0.90.
7. GK-DRAGEN-GK-20 (GATK): RVDPI = 0.868; precision 0.94, clin. validation 0.90.
8. LR-Clair3HiFi-21 (Long-read): RVDPI = 0.856; sensitivity 0.92, difficult regions 0.90.
9. DL-DeepTrio-21 (Deep-learning): RVDPI = 0.848; sensitivity 0.92, precision 0.92.
10. JC-GVCFsCohort-20 (Joint calling): RVDPI = 0.842; scalability 0.92, sensitivity 0.88.