Specialty Journal of Pharmacognosy, Phytochemistry, and Biotechnology
ISSN: 3062-441X
2024, Volume 4, Page No: 171-182

Available online at: www.galaxypub.co/page/journals

Galaxy Publication

High-Throughput Pharmacogene Haplotyping with Oxford Nanopore Adaptive Sampling in 1,036 Clinically Relevant Genes

Aung Kyaw[1], Htet Lin[1]*, Min Thura[1]

[1]Department of Pharmacognosy, Faculty of Pharmacy, University of Yangon, Yangon, Myanmar.

***E-mail** ✉ htet.lin.pg@gmail.com

Received: 19 May 2024; **Revised:** 08 August 2024; **Accepted:** 12 August 2024

ABSTRACT

Pharmacogenomics (PGx) examines how genetic differences between individuals influence medication outcomes, creating the possibility of customizing drug dosing for each person. Existing focused PGx testing approaches primarily rely on microarrays, PCR-based systems, or short-read sequencing technologies. While these methods are effective for detecting single-nucleotide variants (SNVs) and insertion/deletion events (INDELs), they are not suitable for identifying large structural alterations or for delivering clear haplotype phasing needed for accurate star-allele classification. In this study, we applied adaptive sampling from Oxford Nanopore Technologies to selectively enrich a set of 1,036 pharmacogenomically relevant genes curated from the PharmGKB resource. By comparing the results to established reference datasets, we verified reliable variant detection and star-allele determination across five Genome in a Bottle samples. We also demonstrate that up to three samples may be multiplexed on a single PromethION flow cell without compromising performance, achieving recall and precision rates of 99.35% and 99.84% for the targeted loci. These findings promote the integration of nanopore sequencing into clinical PGx workflows.

Keywords: Pharmacogenomics, Nanopore sequencing, Targeted enrichment, Haplotype resolution, Star-allele identification

How to Cite This Article: Kyaw A, Lin H, Thura M. High-Throughput Pharmacogene Haplotyping with Oxford Nanopore Adaptive Sampling in 1,036 Clinically Relevant Genes. Spec J Pharmacogn Phytochem Biotechnol. 2024;4:171-82. https://doi.org/10.51847/xutu52cul2

Introduction

Optimizing medication strategies through pharmacogenomics (PGx) aims to enhance treatment efficacy while reducing harmful reactions. PGx focuses on how genomic variability influences drug metabolism and activity, thereby enabling dose adjustments tailored to each patient. More than 95% of people possess at least one actionable variant in a pharmacogene, underscoring the broad clinical potential of PGx [1–3]. Recently, the PREPARE clinical trial—the largest prospective evaluation of pre-emptive genotyping to date—offered strong evidence supporting panel-based PGx testing [2, 3]. Additionally, PGx-related recommendations appear in the labeling of over 360 approved medications[1]. To standardize the interpretation of genetic diversity, the Pharmacogene Variation Consortium (PharmVar) employs the star (*) allele system [4, 5]. Clinical guidance based on this nomenclature has already been published by CPIC and the Dutch Pharmacogenetics Working Group, specifying how PGx information can refine prescribing choices. Because germline results remain relevant throughout life—especially when obtained in early adulthood—targeted, preventive PGx screening is increasingly proposed for integration into standard medical care.

A major transition has taken place in recent years from PCR and array-based PGx genotyping platforms toward massively parallel sequencing (MPS) strategies [6]. The adoption of MPS is intended to overcome several limitations of earlier methods by enabling analysis of rare or less-characterized alleles and by supporting the detection of copy number variations (CNVs). Previous assessments showed that even specialized PGx microarrays failed to capture roughly 25% of clinically annotated variants listed in PharmGKB [7]. Likewise, the PREPARE project interrogated only 50 germline variants across 12 genes with the PCR-driven LGC SNPline system. Due to

economic considerations, most clinical laboratories rely on targeted MPS panels or whole-exome sequencing (WES). However, WES excludes noncoding and regulatory sequences [8]. Whole-genome sequencing (WGS), which avoids these gaps, has been proposed as an alternative. As short-read sequencing continues to decrease in cost, WGS is expected to supersede WES in many contexts [9]. Nonetheless, WGS produces very large datasets requiring substantial computational resources, and it raises ethical issues surrounding incidental findings beyond the PGx scope [10].

Short-read platforms also have inherent constraints that affect PGx interpretation. With read lengths typically capped at around 600 bp, traditional MPS cannot consistently resolve haplotypes, and larger structural variants (SVs) may remain undetected. SVs encompass longer genomic segments and therefore influence more base pairs per person than SNVs. Furthermore, genes with high sequence similarity—such as CYP2D6 in proximity to the related CYP2D7 and CYP2D8 pseudogenes—pose mapping challenges for short-read sequencing [9, 11].

It is widely recognized that capturing every form of genomic alteration is essential for accurately predicting resulting traits. Long-read sequencing (LRS) has previously been applied in PGx research to obtain single-nucleotide changes, structural rearrangements, and phase information across difficult genomic regions. Multiple studies have already shown its suitability for early clinical implementation [12–15]. Both major LRS platforms have undergone major accuracy upgrades in recent years—ONT through its Q20+ chemistry, and PacBio through HiFi sequencing. Despite this, the majority of targeted workflows still depend on long-range PCR amplification prior to sequencing [16], a procedure known to introduce false chimeric molecules that subsequently distort haplotype interpretation [17, 18]. A PCR-independent alternative, nanopore Cas9-targeted sequencing (nCATS), has been used to resolve the CYP2D6–CYP2D7 region [19, 20], yet this approach still requires labor-intensive guide design for every locus in a PGx panel.

To circumvent these limitations, we employed ONT's adaptive sampling system to enrich pharmacogenetically relevant genes curated from PharmGKB [21]. Because ONT instruments evaluate reads as they are produced, fragments can be either retained or ejected based on a rapid preliminary alignment to a reference sequence. This strategy boosts coverage of desired regions without modifying the library workflow and ensures the sequencer devotes nearly all pore activity to target molecules. Importantly, the gene list can be updated instantly whenever new actionable pharmacogenes emerge, without optimization cycles. We anticipate that long, PCR-free reads will allow more reliable haplotype definitions and yield improved predictions of functional outcomes at a lower overall cost.

Materials and Methods

Target gene selection and .bed file construction

The PGx gene list was assembled from the clinical_annotations.tsv file downloaded from PharmGKB [21]. All clinically annotated variants appearing in this file were included. Target intervals were assigned as follows: variants located inside exons or introns triggered inclusion of the full gene body; variants situated outside gene boundaries were incorporated when located within 100 kb of the transcription start or end site. In these cases, the interval extended from the gene boundary to the variant position. This procedure produced 3,347 total variants—3,262 SNVs and 85 INDELs—spanning 1,036 genes. An additional 20 kb was appended to both sides of each interval to accommodate long DNA fragments whose initial bases may fall outside the actual target. Any overlapping intervals were merged. The final .bed file covered 5.68% of the genome, which fits within ONT's recommended 1%–10% enrichment range [22].

DNA sample collection and QC

Reference materials NA12878 (HG001), HG01190, NA19785, NA24385 (HG002), and NA24631 (HG005) were obtained from the Coriell Institute (Camden, NJ, United States). Concentrations were confirmed using the PicoGreen fluorescence assay (ThermoFisher, Waltham, MA, United States). DNA fragment size distributions were examined on the Femto Pulse platform using the Agilent 165 kb Genomic DNA kit (Agilent Technologies, Santa Clara, CA, United States).

Library preparation

For HG001 sequenced on a PromethION R9.4.1 flow cell, libraries were prepared using ONT's SQK-LSK110 ligation-based protocol. A total of 1.5 µg of intact DNA was subjected to repair and end-processing. After ligation

of adapters and cleanup, 70.5 fmol of library was diluted in elution buffer (ONT, Oxford, United Kingdom) and divided into three equal fractions. The first fraction (23.5 fmol) was loaded onto the flow cell. After 22 h and 40 h of runtime, the flow cell was washed following the Flow Cell Wash Kit protocol, and the second and third aliquots were loaded.

For HG002 and HG005 sequenced on R10.4.1 flow cells, the Native Barcoding Kit 24 V14 (SQK-NBD114.24) was used. Each sample began with 4 µg of unfragmented DNA, and reagent volumes were doubled to yield a larger library. After end-repair and purification, concentrations were measured using PicoGreen. Equimolar amounts—based on the lowest-concentration sample—were included in the barcoding reaction to ensure uniform read distribution. Pooled, barcoded libraries underwent adapter ligation and cleanup, producing 79 fmol of final library. Of this, 50 fmol was loaded. After 36 h, the library was recovered according to ONT's recovery protocol, the flow cell was washed, and the recovered material, plus the remaining 29 fmol, was reloaded.

For HG001, HG01190, and NA19785 processed on R10.4.1 flow cells, the approach was similar. Each sample started with 3 µg of high-molecular-weight DNA and produced 180 fmol of library. An initial 60 fmol was loaded to maximize pore utilization. After 20 h, the flow cell was washed and supplemented with 70 fmol. At 42 h, the flow cell was washed again, and the remaining 50 fmol was added. Because unused pore capacity remained at 72 h, library recovery was performed with an added 10 µL of sequencing buffer, and sequencing resumed.

Data analysis pipeline

Raw .fast5 files were re-basecalled using ONT's super-accurate (SUP) Guppy models: version 6.4.2-gpu for the R9.4.1 and R10.4.1 HG002/HG005 runs, and 6.5.7-gpu for the R10.4.1 HG001/HG01190/NA19785 run. Reads scoring below Q10 were excluded. For downstream steps, we retained only reads actively accepted by the adaptive sampling (AS) algorithm, preventing the inclusion of short, rejected fragments that often fail to map uniquely. Basecalled reads were aligned to GRCh38 using minimap2 v2.18-r1015 [23]. Variant detection was then carried out with Clair3 v0.1-r10, followed by phasing using WhatsHap v2.0 with --include-homozygous --indels --distrust-genotypes.

Benchmarking was performed with hap.py. Recall was computed as truth_tp / (truth_tp + truth_fn) and precision as query_tp / (query_tp + query_fp). Although whole-genome benchmarking typically excludes true negatives because the overwhelming number of reference-matching sites skews interpretation [24], in our case, only the predefined panel variants are assessed. Thus, concordant reference calls are meaningful for PGx applications, where confirming the absence of an ALT allele is as important as identifying it. Accuracy was therefore calculated as (truth_tp + truth_tn) / total_variants_in_truth_set.

For HG001, the hybrid Genome in a Bottle–Platinum Genomes benchmark described by Krusche *et al*. was used. For HG002 and HG005, the GIAB v4.2.1 truth sets served as reference [25].

For 24 of the 1,036 genes in the target panel—designated as VIP genes by PharmGKB—star-allele calls were generated using Aldy v4.4 (https://github.com/0xTCG/aldy) [26, 27]. For HG001, HG01190, and NA19785, assignments were compared against the Genetic Testing Reference Material Coordination (GeT-RM) consensus. No validated star-allele sets exist for HG002 or HG005, so results are provided as a reference only.

Results and Discussion

Sequencing summary statistics and evaluation of variant calling performance

To establish baseline performance of adaptive sampling (AS) on the PGx panel, HG001 was processed using the SQK-LSK110 ligation workflow and sequenced on a PromethION R9.4.1 flow cell. This experiment achieved an average 47× coverage over targets, with >99.9% of bases reaching 20× depth **(Table 1)**. Among the 3,347 PGx variants in our design, 3,262 appear in the HG001 truth set, and 99.69% were accurately identified **(Table 2)**.

Table 1. Summary statistics for the R9.4.1 and R10.4.1 multiplex sequencing runs.

Flow-cell chemistry	**R9.4.1**	**R9.4.1**	**R9.4.1**	**R10.4.1**	**R10.4.1**	**R10.4.1**
Sample (GIAB reference)	HG001	HG002	HG005	HG001	HG01190	NA19785
Total sequenced bases (Gb)	49.9	55.9	73.3	20	32	24
Mean sequencing depth on-target	47×	40×	23×	20×	32×	24×
Percentage of target regions ≥30×	98.8%	91.2%	11.1%	5.1%	64.6%	17.8%

Percentage of target regions ≥20×	99.9%	99.4%	71.2%	57.7%	97.6%	81.1%
Percentage of target regions ≥15×	99.96%	99.8%	93.9%	89.1%	99.6%	97.0%

Table 2. Variant calling metrics for HG001 using the R9.4.1 flow cell relative to the Krusche *et al*. reference.

Category	Total PGx variants in HG001 reference	Alternate (ALT) variants in truth set	Alternate (ALT) variants called in R9.4.1 data	Recall (%)	Precision (%)	Accuracy (%)
All variants	3,262	1,229	1,224	99.59	99.59	99.69
Single nucleotide variants (SNVs)	3,186	1,218	1,214	99.67	99.75	99.79
Insertions/deletions (INDELs)	76	11	10	90.91	83.33	96.05
Number of genes evaluated	1,023	—	—	—	—	—

Next, to reduce sequencing cost per sample, HG002 and HG005 were barcoded and processed together using the Q20+ Kit14 chemistry on early-access R10.4.1 flow cells. Mean depths of 40× (HG002) and 23× (HG005) were obtained on a single flow cell **(Table 1)**. Although this represents approximately half the coverage of the initial run, both samples retained >99% recall **(Table 3)**.

Table 3. Variant calling results for barcoded HG002 and HG005 compared with their GIAB truth sets on a single R10.4.1 flow cell.

Sample (GIAB)	Variant category	Total PGx variants in reference	Alternate (ALT) variants in truth set	Alternate (ALT) variants called in R10.4.1 data	Recall (%)	Precision (%)	Accuracy (%)
HG002	All variants	3,229	1,048	1,042	99.43	99.81	99.75
	Single nucleotide variants (SNVs)	3,155	1,039	1,035	99.62	100.00	99.87
	Insertions/deletions (INDELs)	74	9	7	77.78	77.78	94.59
	Number of genes evaluated	1,008	—	—	—	—	—
HG005	All variants	2,813	1,039	1,030	99.13	100.00	99.68
	Single nucleotide variants (SNVs)	2,762	1,032	1,025	99.32	100.00	99.75
	Insertions/deletions (INDELs)	51	7	5	71.43	100.00	96.08
	Number of genes evaluated	991	—	—	—	—	—

A second multiplex run combined HG001, HG01190, and NA19785 on another R10.4.1 flow cell. These samples were selected because comprehensive star-allele definitions exist. As expected, increased multiplexing was anticipated to moderately reduce precision and recall, but coverage remained sufficient to identify larger structural events. **Table 4** lists SNV and INDEL recall/precision for HG001. Equivalent comparisons for HG01190 and NA19785 could not be generated due to the absence of validated truth datasets.

Table 4. Variant calling metrics for HG001 using the R10.4.1 flow cell containing pooled HG001/HG01190/NA19785 libraries. Truth sets for the latter two samples are unavailable.

Sample (GIAB)	Variant category	Total PGx variants in reference genome	Alternate (ALT) variants in truth set	Alternate (ALT) variants detected in R10.4.1 data	Recall (%)	Precision (%)	Accuracy (%)
HG001	All variants	3,262	1,229	1,224	99.35	99.84	99.70

Single nucleotide variants (SNVs)	3,186	1,218	1,214	99.43	99.92	99.75
Insertions/deletions (INDELs)	76	11	10	90.91	90.91	97.65
Number of genes evaluated	1,023	—	—	—	—	—

Structural variant calling and haplotype phasing

To resolve haplotypes for detected SNVs, indels, and more complex alterations, we employed WhatsHap, which leverages the extended span of ONT reads. Since ONT read length is dictated by the size of the DNA fragments entering the flow cell, individual reads can traverse numerous loci, including sizable or intricate structural rearrangements.

Because pharmacogenomic variation is traditionally expressed using *-allele nomenclature, we next processed the AS-derived sequence data with Aldy v4.4 to assign star alleles automatically. The library preparation strategy avoided PCR, so formation of synthetic chimeras was not expected. Outputs are presented in **Table 5**.

Table 5. Concordance between GeT-RM PGx star-allele designations and those inferred by Aldy for the VIP genes.

Gene	HG001 (GIAB)	HG001 ALDY R9.4.1	HG001 ALDY R10.4.1	HG01190 GET-RM	HG01190 ALDY R10.4.1	NA19785 GET-RM	NA19785 ALDY R10.4.1
CFTR	– (*WT/*WT)Δ	*WT/*WT	*WT/*WT	–	*WT/*WT	–	*WT/*WT
COMT	–	*Met/*ValA	*Met/*ValA	–	*Met/*ValA	–	*Met/*ValB
CYP1A2	*1F/*1F	*1M/*1M	*1M/*1M	*1A/*1A	*1B/*1B	*1L/*1L	*1L/*1L
CYP2A13	*1A/*1A≠	*1/*1	*1/*1	*1A/*1A≠	*1/*1	–	*1/*1
CYP2A6	*1/*1	*1+*1/*12	*1+*1/*12	*1/*1	*1/*1	*1/*1≠	*1/*1
CYP2B6	*1/*1	*1/*1	*1/*1	*1(*5)/*1(*27)	*1/*5	*1/*1	*1/*5
CYP2C19	*1/*2	*1/*2	*1/*2	*1/*2	*1/*2	*1/*1	*1/*1
CYP2C8	*1/*3	*3/*5	*1/*3	*1/*3	*1/*3	*1/*1	*1/*1
CYP2C9	*1/*2	*1/*2	*1/*2	*2/*61	*1/*61	*1/*1	*1/*1
CYP2D6	*3/*4+*68	*3 + *82/*4 +*132	*4N.ALDY/ *10+*82	*4/*5	*4/*4	*1/*2+*13	*2/*13
CYP2E1	no consensus (*5)/*7	*1/*5A_7A _1B	*1/*5A_7A_ 1B	*1/*7	*1/*7	*7/*7≠	*4/*5
CYP2J2	*1/*1≠	*1/*1	*1/*1	*1/*7≠	*1/*7	–	*1/*1
CYP3A4	*1/*1	*1/*1	*1/*1	*1/*1B	*36/*36	*1/*1	*1/*36
CYP3A5	*3/*3	*3/*3	*3/*3	*1/*1	*1/*1	*1/*3	*1/*3
CYP4F2	*1/*1	*1/*1	*1/*1	*1/*3	*1/*3	*3/*3≠	*3/*3
DPYD	*1/(*4)	*4/*5	*4/*5	*1/*9	*1/*9	*1/*1	*1/*1
G6PD	NEG≠	*B/*B	*B/*B	NEG≠	*B/*B	NEG≠	*B/*B
GSTP1	*A/*C; *B/*D	*A/*C	*A/*C	*A/*B	*A/*B	*A/*B≠	*A/*B
NAT2	*4/*5	*4/*5	*4/*5	*4/*4	*4/*4	no consensus	*7/*7 (curated: *7/*12)
NUDT15	– (*1/*1)‡	*1/*1	*1/*1	– (*1/*1)‡	*1/*1	– (*1/*1)‡	*1/*1
SLCO1B1	*1/*15	*1/*15	*1/*15	*1/*1	*1/*1	*1/*1	*1/*37
TPMT	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1	*1/*1

UGT1A1	*1/*28	*60_80/*112	*1/*28_60_80_93	no consensus (*37)/*60	*1/*60_80 (curated: *1/*37_60_80)	no consensus	*1/*28_60_80_93
VKORC1	H1/(H9)	*H1/*H8	*H1/*H9	*H7/*H7	*H7/*H7	*H1/*H1	*H1/*H1

Light blue, dark blue, and orange denote exact matches, expanded interpretations, or discordant calls, respectively. White boxes indicate missing reference designations.

Δ No diplotype was listed in GeT-RM; based on Pranesh *et al*. (2019), *WT/*WT was selected. ≠ Diplotype originates from GeT-RM non-consensus material (single-assay evidence). For G6PD, "NEG" reflects the absence of the A+/A– allele in the Tech Open Array; Aldy's *B/*B corresponds to wild-type. ‡ No GeT-RM diplotype existed; following Liu *et al*. (2023), a call was assigned [28].

We used PromethION's adaptive sampling to digitally enrich a panel of 1,036 pharmacogenes, evaluating its suitability for individualized medication optimization. Several major clinical centers—such as Erasmus MC in the Netherlands—already employ PGx testing, often based on limited TaqMan-style assays available on request. Yet each gene in these small panels typically includes only a narrow set of variants. In contrast, some institutions utilize broader sequencing-based diagnostics. Notably, the ACMG recently revised its recommendations for clinical PGx analysis [29], covering targeted testing, WES/WGS, and CNV analysis. Approximately 34.49% (269/780) of PGx tests cataloged in the NIH GTR database involve full-coding-region sequencing [30]. Nonetheless, intronic and regulatory variants—such as those known to impact CYP2C19 and CYP3A4—are frequently absent [31, 32]. Our adaptive-sampling workflow simultaneously retrieves sequence data for all targeted loci and inherently retains CNV information, representing an advance over existing genotyping technologies.

The selected panel covers established PharmGKB pharmacogenes and also includes loci with weaker current evidence but potential future clinical relevance. For each locus, the entire genomic region is examined, incorporating introns and up to 20 kb flanking both ends. Because the target file is flexible, newly validated PGx genes can be incorporated without re-designing laboratory protocols.

Recall and precision assessment for HG001, HG002, and HG005

Using the GIAB HG001 standard on an R9.4.1 flow cell, we obtained 99.59% recall and 99.59% precision for PharmGKB-listed variants. These metrics parallel Illumina NovaSeq WGS results at ~30X coverage [33], which reported 99.53% recall and 99.57% precision using GATK HaplotypeCaller.

In the following experiment, HG002 and HG005 were jointly sequenced on a single R10.4.1 flow cell using ONT's latest chemistry. We anticipated that the improved raw read accuracy would enable reliable genotyping of two samples despite reduced coverage. Prior modeling indicates that ~6X and 8X coverage allows recovery of 98% of homozygous and 90% of heterozygous variants, respectively [34]. As shown in **Table 3**, our results exhibit comparable or even superior precision relative to HG001.

Wagner *et al*. recently introduced a benchmark for 273 autosomal CMRG genes in HG002, many of which remain unresolved in GIAB v4.2.1 due to sequencing or variant-level complexity [35]. Of the 1,036 genes in our panel, 32 overlap with the CMRG set. Across these genes, all PharmGKB SNVs and indels achieved perfect recall and precision against the CMRG reference.

Final experiment: three-sample multiplexing on R10.4.1

In the last sequencing run, we evaluated whether HG001, HG01190, and NA19785 could be processed together on a single R10.4.1 flow cell. These GIAB samples were selected both to compare HG001 performance with the earlier non-multiplexed R9.4.1 run and because curated *-allele reference calls exist for all three. The recall and precision outcomes are presented in **Table 4**. Although the mean depth decreased from 47X to 20X, total recall for PharmGKB-listed variants declined by only 0.24%, while precision rose from 99.59% to 99.84%. These data confirm that the shift from R9.4.1 to R10.4.1 provides noticeable improvements for PGx-relevant variant detection. As ONT now ships R10-series flow cells as the standard, we infer that multiplexing three samples on a single PromethION R10.4.1 flow cell—capturing 5.68% of the genome—is suitable for PGx workflows.

Running three barcoded samples together reduces the estimated per-sample expenditure to €320. For comparison, Twist Bioscience has released a long-read capture panel for PacBio platforms, but it targets only 49 genes and

would require redesign of enrichment probes whenever new loci must be added; additionally, the hybridization and reagent steps add extra cost.[3]

Long-read phasing success rate

The long-read datasets were also used to phase both haplotypes, an essential step for accurate *-allele resolution. Short-read or imputation-assisted tools do not reliably capture subject-specific PGx phasing, whereas long reads directly link alleles without depending on population references or trio data. Using the phased truth sets for HG001, HG002, and HG005, we computed phasing success across 47 PharmGKB VIP genes for which reference phasing was available **(Figure 1)**. Substantial differences among samples were seen for genes such as CFTR, CYP2A6, CYP2C19, NRAS, and NUDT15. For HG001, R9- versus R10-based discrepancies likely reflect depth and chemistry, while much larger sample-to-sample differences are best explained by variation in the number of heterozygous sites available for phasing.

For example, CFTR contains 39 phased reference variants in HG001 but 280 in HG002. A higher abundance of heterozygous positions naturally produces longer haplotype blocks, improving the fraction of variants that can be phased. The effect of unphased calls is also magnified when the reference set for a gene is small. As seen for CYP2A6, HG001 had 1 correctly phased variant out of 3 reference positions, whereas HG002 had 1 out of 1.

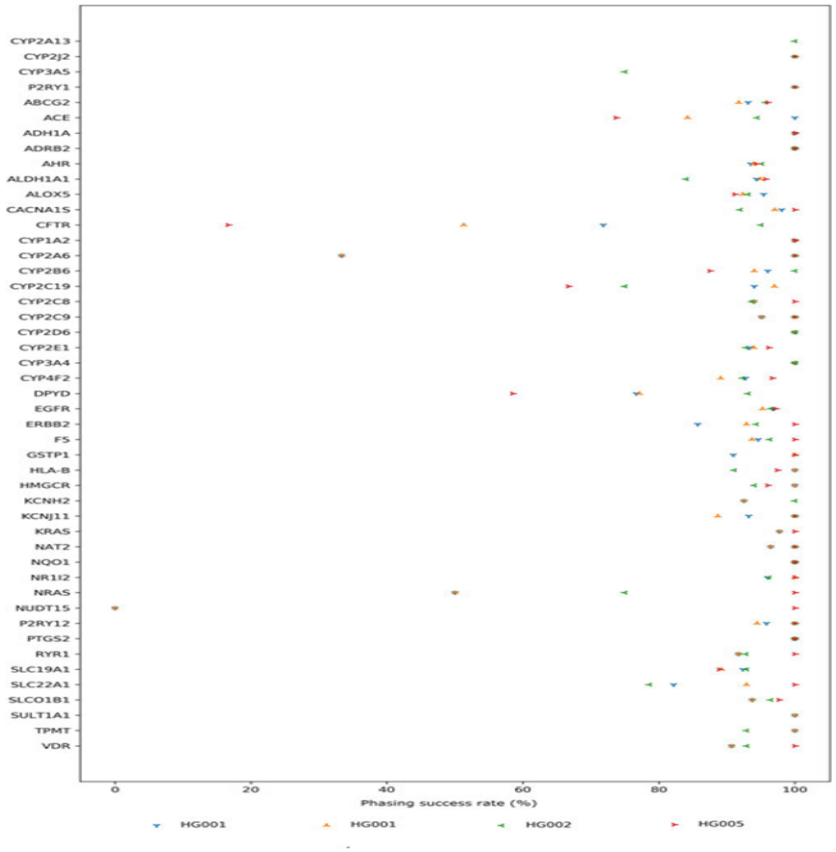

Figure 1. Proportion of correctly phased variants relative to truth sets. Only the 47 VIP genes with available reference phasing information are displayed.

Genes with very large genomic spans predictably had more unphased positions, as even long reads could not extend across entire loci. When comparing genes with ≤90% versus ≥90% phasing success, a clear separation in genomic span is observed. **Figure 2** summarizes the relationship between gene size and phasing success across samples. These patterns emphasize the need for high-molecular-weight DNA input: longer molecules enable reads to encompass more variants, producing longer phasing blocks. Standard extraction methods designed for short-read sequencing or microarray analysis may not consistently provide DNA of sufficient length for long-read PGx applications.

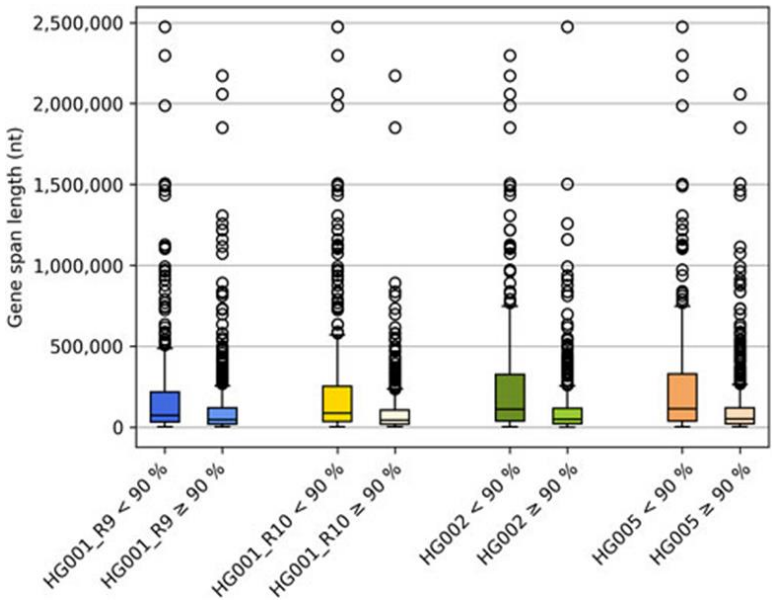

Figure 2. Boxplots showing gene span distributions for genes with <90% versus ≥90% phasing accuracy across reference datasets.

Star-allele annotation using Aldy 4

In the final stage of analysis, we applied Aldy 4 to assign *-alleles across the PGx genes included in our panel [26, 27]. Extracting maximal pharmacogenomic value from our dataset requires a tool that can jointly interpret SNVs, INDELs, CNVs, and the phasing information produced by long-read sequencing. Many commonly used platforms—such as PharmCAT—rely solely on .vcf input files for haplotype determination and consequently overlook gene copy alterations or fusion events [36]. In contrast, Aldy 4 accepts entire WGS .bam files. PyPGx includes a workflow tailored for long-read data but processes only .vcf files and lacks structural variant support [37]. Cyrius also accepts WGS .bam input, though it is specialized for short-read sequencing and restricted to the CYP2D6 locus [38]. Benchmark studies from Graansma *et al.* and Shugg *et al.* support Aldy as the most comprehensive solution for this application [39, 40]. The *-allele results for the 24 Aldy-supported VIP genes are listed in **Table 5**.

For evaluation, we used the GeT-RM star-allele assignments for HG001, HG01190, and NA19785 and supplemented these with additional validated resources where GeT-RM lacked coverage [41–43]. Aldy reports both major and minor haplotypes, but only major *-alleles were compared since GeT-RM contains no minor-allele annotations.

Using R9.4.1 data, Aldy produced 14 matching, 8 non-matching, and 2 unreferenced calls. For three of the discordant loci—CYP1A2, DPYD, and GSTP1—our results actually refined or corrected the listed reference haplotypes. In CYP1A2, GeT-RM does not include the *1M allele; although *1M shares rs762551 with *1F, our dataset clearly demonstrated the presence of rs2472304 on both alleles. In DPYD, *4 was identified only by several assays (Affymetrix DMET; LifeTech Taqman LDT; Agena iPLEX ADME PGx Pro), and none assessed *5. For GSTP1, prior assays could not differentiate between *A/*C and *B/*D, including an NGS panel [44]. Aldy incorrectly assigned the CYP2A6 diplotype by inferring the *12 structural variant. Discordances for CYP2C8, UGT1A1, and VKORC1 correspond to known issues with homopolymer stretches on ONT R9.4.1 flow cells [45]. The misidentification of the CYP2D6 *4+*68 haplotype was anticipated, as Aldy also missed this in its original publication [27].

For the R10.4.1 runs, Aldy produced correct calls for CYP2C8, VKORC1, and UGT1A1, even with reduced depth. This aligns with improvements from ONT's dual-reader pore design, which enhances performance in homopolymers.

For HG01190 (R10.4.1), Aldy yielded improved calls for CYP1A2, CYP2B6, CYP2C9, and CYP3A4, whereas it failed for CYP2D6 and UGT1A1. Since *1B of CYP1A2 is not part of GeT-RM, we manually confirmed its presence. For CYP2B6, some assays reported *1/*1 without evaluating *5, while Affymetrix identified *5 but incorrectly assumed rs36079186, leading to a *27 interpretation. The *1/*61 vs. *2/*62 discrepancy for CYP2C9 was also described by Liu *et al.* (2023). Regarding CYP3A4, the *36 allele had not been tested by GeT-RM;

Aldy's *36/*36 call is technically correct, but this haplotype was removed from PharmVar after v5.2.17, so the current notation is *1/*1.

For NA19785, improved *-allele calls for CYP2B6 and SLCOB1 agree with Liu *et al*. (2023). CYP2E1 alleles *4 and *5 were not included in Agena's assay but were validated using IGV. Aldy correctly called *1/*36 for CYP3A4, though—as above—*36 has since been retired and should be listed as *1. For NAT2, Aldy reported *7/*7, but manual review showed the correct diplotype is *7/*12. Low read depth led Aldy to misattribute rs1799931 to the *12 allele. As with the other reference samples, Aldy did not correctly resolve CYP2D6.

Across all evaluated samples, CYP2D6 remained problematic due to its highly complex genomic architecture. In HG01190, one allele was mistaken for *4 instead of the *5 deletion, likely caused by CYP2D7-derived reads misaligning to CYP2D6. Both HG001 and NA19785 harbor intricate CYP2D6–CYP2D7 hybrid configurations; Aldy reconstructed only portions of the final diplotypes. Since Aldy's fusion-breakpoint identification relies primarily on SNV-based literature patterns—which may not distinguish all hybrid forms—more sophisticated approaches, such as pangenome graph assemblies, may be required to resolve CYP2D6 structures [46].

**"In addition, diplotyping the UGT1A1 locus remains notably difficult. The UGT Nomenclature Committee currently recognizes 113 haplotypes.[4] The GeT-RM datasets do not offer unified genotype calls for HG01190 or NA19785, even though each was analyzed on four and three platforms, respectively. For HG001, the consensus assignment is *1/*28, yet the Affymetrix assay additionally detected *60 and *93, and Agena classified it as *1/60. In our R9 dataset, the UGT1A128 allele was not identified, but appeared in the R10 dataset, likely benefiting from the improved homopolymer handling. Aldy also flagged *112 (C>A variant) across all samples; however, since A is now the reference base, this was manually revised to *1. In HG01190, phased read inspection revealed the *37 allele (9 TA repeats), which Aldy missed, likely due to insufficient depth. Moreover, Aldy consistently reported *60 and *80 for all samples, and manual review confirmed that both variants occur on the same chromosomal copy. Overall, updated allele definitions and studies on larger populations are needed to clarify true UGT1A1 haplotype composition. Our findings suggest that ONT long-read sequencing can help refine UGT1A1 characterization.

For HG002 and HG005, GeT-RM does not provide *-allele references.

Future perspectives

Alternative adaptive-sampling strategies could be examined to expand throughput for this PGx panel. At present, MinKNOW evaluates the start of each molecule, aligns it, and determines whether it aligns to the target .bed regions. Readfish, an open-source replacement, offers more flexibility [47], notably yielding faster rejection decisions. In our hands, an average of 830–880 bp (~2 s of sequencing) was produced before a read was rejected. Readfish has reported shorter N50s (~500 bp) for rejected reads, though its efficiency on PromethION is still lower than on MinION/GridION due to slower unblocking. ReadBouncer, using k-mer classification, performs quicker accept/reject decisions but cannot handle very large references such as full human genomes [48]. A more recent framework, BOSS-RUNS, dynamically modifies the regions of interest based on accumulated sequence data and relays this information to Readfish, enabling prioritization of under-covered loci. However, its computational demands make it impractical for human-scale targets [49].

Another challenge involves equimolar pooling of high-molecular-weight DNA for a single PromethION flow cell, especially when samples differ in size distribution. Mechanical shearing could equalize fragment lengths, but shorter input molecules reduce AS efficiency and shrink haplotype blocks, which is detrimental for PGx phasing. Recently, computational solutions have emerged. Just as AS enriches target genes, it can be adapted to enrich specific barcodes. While MinKNOW cannot currently combine gene-level AS with barcode selection, Readfish enables barcode-aware AS [50]. Even so, a fully automated system that balances barcode representation and harmonizes coverage across samples and targets would be useful. SwordFish offers such balancing for SARS-CoV-2 amplicons but has not been shown to scale to large genomes [51].

The PGx field itself continues to evolve, and new strategies for linking diplotypes to functional phenotypes are emerging. Instead of relying on categorical *-allele nomenclature, continuous prediction scales have been proposed. Machine-learning models trained on entire gene sequences can infer phenotypes, even for variants not assigned to any *-allele. Improved performance has already been demonstrated for CYP2D6 in tamoxifen- and venlafaxine-treated groups [13, 52]. Our phased long-read, PCR-free datasets may offer superior input for such approaches. As a limitation, we emphasize that wider cohort-level validation will be needed before considering clinical deployment.

Conclusion

Our study demonstrates that applying adaptive sampling on the PromethION platform yields extensive PGx information spanning 1,036 genes, including SNVs, INDELs, structural variants, phasing results, and -allele calls. The long-read diplotyping framework we propose is comprehensive and well-positioned to integrate future clinical insights. We show that up to three samples can be multiplexed on one PromethION flow cell with strong performance, achieving recall and precision rates of 99.35% and 99.84% for targeted variants. At present, the main restriction on accurate -allele assignment comes from the available bioinformatic tools. Enhancements to AS could further increase sample multiplexing and improve accuracy by raising enrichment performance and balancing coverage. Ultimately, we conclude that targeted long-read sequencing represents a powerful approach for advancing personalized medicine in PGx."

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Ji Y, Skierka JM, Blommel JH, Moore BE, VanCuyk DL, Bruflat JK, et al. Preemptive pharmacogenomic testing for precision medicine: a comprehensive analysis of five actionable pharmacogenomic genes using next-generation DNA sequencing and a customized CYP2D6 genotyping cascade. J Mol Diagn. 2016;18(3):438–45. doi:10.1016/j.jmoldx.2016.01.003
2. van der Wouden CH, Böhringer S, Cecchin E, Cheung KC, Dávila-Fajardo CL, Deneer VH, et al. Generating evidence for precision medicine: considerations made by the Ubiquitous Pharmacogenomics Consortium when designing the PREPARE study. Pharmacogenet Genomics. 2020;30(6):131–44. doi:10.1097/FPC.0000000000000405
3. Swen JJ, van der Wouden CH, Manson LEN, Abdullah-Koolmees H, Blagec K, Blagus T, et al. A 12-gene pharmacogenetic panel to prevent adverse drug reactions: an open-label, multicentre, controlled, cluster-randomised crossover implementation study. Lancet. 2023;401(10374):347–56. doi:10.1016/S0140-6736(22)01841-4
4. Gaedigk A, Ingelman-Sundberg M, Miller NA, Leeder JS, Whirl-Carrillo M, Klein TE, et al. The Pharmacogene Variation (PharmVar) Consortium: incorporation of the human cytochrome P450 allele nomenclature database. Clin Pharmacol Ther. 2018;103(3):399–401. doi:10.1002/cpt.910
5. Gaedigk A, Casey ST, Whirl-Carrillo M, Miller NA, Klein TE. Pharmacogene Variation Consortium: a global resource and repository for pharmacogene variation. Clin Pharmacol Ther. 2021;110(3):542–5. doi:10.1002/cpt.2321
6. Tafazoli A, Guchelaar HJ, Miltyk W, Kretowski AJ, Swen JJ. Applying next-generation sequencing platforms for pharmacogenomic testing in clinical practice. Front Pharmacol. 2021;12:693453. doi:10.3389/fphar.2021.693453
7. Tilleman L, Weymaere J, Heindryckx B, Deforce D, Van Nieuwerburgh F. Contemporary pharmacogenetic assays in view of the PharmGKB database. Pharmacogenomics. 2019;20(4):261–72. doi:10.2217/pgs-2018-0167
8. Tilleman L, Heindryckx B, Deforce D, Van Nieuwerburgh F. Pan-cancer pharmacogenetics: targeted sequencing panels or exome sequencing? Pharmacogenomics. 2020;21(15):1073–84. doi:10.2217/pgs-2020-0035
9. Tafazoli A, Guchelaar HJ, Miltyk W, Kretowski AJ, Swen JJ. Applying next-generation sequencing platforms for pharmacogenomic testing in clinical practice. Front Pharmacol. 2021;12:693453. doi:10.3389/fphar.2021.693453

10. Johnson SB, Slade I, Giubilini A, Graham M. Rethinking the ethical principles of genomic medicine services. Eur J Hum Genet. 2020;28(2):147–54. doi:10.1038/s41431-019-0507-1
11. Nofziger C, Paulmichl M. Accurately genotyping CYP2D6: not for the faint of heart. Pharmacogenomics. 2018;19(13):999–1002. doi:10.2217/pgs-2018-0105
12. Fukunaga K, Hishinuma E, Hiratsuka M, Kato K, Okusaka T, Saito T, et al. Determination of a novel CYP2D6 haplotype by targeted sequencing and long-read sequencing in the Japanese population. J Hum Genet. 2021;66(2):139–49. doi:10.1038/s10038-020-0815-x
13. van der Lee M, Allard WG, Vossen RH, Baak-Pablo RF, Menafra R, Deiman BA, et al. Toward predicting CYP2D6-mediated variable drug response from gene sequencing data. Sci Transl Med. 2021;13(603):eabf3637. doi:10.1126/scitranslmed.abf3637
14. Scott ER, Yang Y, Botton MR, Seki Y, Hoshitsuki K, Harting J, et al. Long-read HiFi sequencing of NUDT15. Hum Mutat. 2022;43(11):1557–66. doi:10.1002/humu.24457
15. van der Lee M, Rowell WJ, Menafra R, Guchelaar HJ, Swen JJ, Anvar SY. Application of long-read sequencing to elucidate complex pharmacogenomic regions. Pharmacogenomics J. 2022;22(1):75–81. doi:10.1038/s41397-021-00259-z
16. Liau Y, Maggo S, Miller AL, Pearson JF, Kennedy MA, Cree SL. Nanopore sequencing of CYP2D6 allows haplotyping and detection of duplications. Pharmacogenomics. 2019;20(14):1033–47. doi:10.2217/pgs-2019-0080
17. Ammar R, Paton TA, Torti D, Shlien A, Bader GD. Long-read nanopore sequencing for detection of HLA and CYP2D6 variants. F1000Res. 2015;4:17. doi:10.12688/f1000research.6037.1
18. Laver TW, Caswell RC, Moore KA, Poschmann J, Johnson MB, Owens MM, et al. Pitfalls of haplotype phasing from amplicon-based long-read sequencing. Sci Rep. 2016;6(1):21746. doi:10.1038/srep21746
19. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat Biotechnol. 2020;38(4):433–8. doi:10.1038/s41587-020-0407-5
20. Rubben K, Tilleman L, Deserranno K, Tytgat O, Deforce D, Van Nieuwerburgh F. Cas9-targeted nanopore sequencing improves CYP2D6-CYP2D7 hybrid allele genotyping. PLoS Genet. 2022;18(9):e1010176. doi:10.1371/journal.pgen.1010176
21. Whirl-Carrillo M, Huddart R, Gong L, Sangkuhl K, Thorn CF, Whaley R, et al. An evidence-based framework for evaluating pharmacogenomics knowledge. Clin Pharmacol Ther. 2021;110(3):563–72. doi:10.1002/cpt.2350
22. Oxford Nanopore Technologies (ONT). Nanopore community: adaptive sampling [Internet]. 2020. Available from: https://community.nanoporetech.com/docs/plan/best_practice/adaptive-sampling/v/ads_s1016_v1_revi_12nov2020
23. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094–100. doi:10.1093/bioinformatics/bty191
24. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, et al. Best practices for benchmarking germline small-variant calls. Nat Biotechnol. 2019;37(5):555–60. doi:10.1038/s41587-019-0054-x
25. Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, et al. Benchmarking challenging small variants with linked and long reads. Cell Genomics. 2022;2(5):100128. doi:10.1016/j.xgen.2022.100128
26. Numanagić I, Malikić S, Ford M, Qin X, Toji L, Radovich M, et al. Allelic decomposition and exact genotyping of highly polymorphic genes. Nat Commun. 2018;9(1):828. doi:10.1038/s41467-018-03273-1
27. Hari A, Zhou Q, Gonzaludo N, Harting J, Scott SA, Qin X, et al. An efficient genotyper and star-allele caller for pharmacogenomics. Genome Res. 2023;33(1):61–70. doi:10.1101/gr.277075.122
28. Liu Y, Lin Z, Chen Q, Chen Q, Sang L, Wang Y, et al. PAnno: a pharmacogenomics annotation tool for clinical genomic testing. Front Pharmacol. 2023;14:1008330. doi:10.3389/fphar.2023.1008330
29. Tayeh MK, Gaedigk A, Goetz MP, Klein TE, Lyon E, McMillin GA, et al. Clinical pharmacogenomic testing and reporting: a technical standard of the ACMG. Genet Med. 2022;24(4):759–68. doi:10.1016/j.gim.2021.12.009
30. Rubinstein WS, Maglott DR, Lee JM, Kattman BL, Malheiro AJ, Ovetsky M, et al. The NIH genetic testing registry. Nucleic Acids Res. 2013;41(D1):D925–35. doi:10.1093/nar/gks1173

31. Morales-Rosado JA, Goel K, Zhang L, Åkerblom A, Baheti S, Black JL, et al. Next-generation sequencing of CYP2C19 in stent thrombosis. Cardiovasc Drugs Ther. 2021;35(3):549–59. doi:10.1007/s10557-020-06988-w
32. Zhou Y, Tremmel R, Schaeffeler E, Schwab M, Lauschke VM. Rare-variant pharmacogenomics: challenges and opportunities. Trends Pharmacol Sci. 2022;43(10):852–65. doi:10.1016/j.tips.2022.07.002
33. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, et al. High-coverage whole-genome sequencing of the expanded 1000 Genomes cohort. Cell. 2022;185(18):3426–3440.e19. doi:10.1016/j.cell.2022.08.004
34. Mahmoud M, Huang Y, Garimella K, Audano PA, Wan W, Prasad N, et al. Utility of long-read sequencing for All of Us. bioRxiv. 2023. doi: (preprint)
35. Wagner J, Olson ND, Harris L, McDaniel J, Cheng H, Fungtammasan A, et al. Curated variation benchmarks for challenging medically relevant autosomal genes. Nat Biotechnol. 2022;40(5):672–80. doi:10.1038/s41587-021-01158-1
36. Sangkuhl K, Whirl-Carrillo M, Whaley RM, Woon M, Lavertu A, Altman RB, et al. PharmCAT: a pharmacogenomics clinical annotation tool. Clin Pharmacol Ther. 2020;107(1):203–10. doi:10.1002/cpt.1568
37. Lee SB, Shin JY, Kwon NJ, Kim C, Seo JS. ClinPharmSeq: a targeted sequencing panel. PLoS One. 2022;17(7):e0272129. doi:10.1371/journal.pone.0272129
38. Chen X, Shen F, Gonzaludo N, Malhotra A, Rogert C, Taft RJ, et al. Cyrius: accurate CYP2D6 genotyping using whole-genome sequencing. Pharmacogenomics J. 2021;21(2):251–61. doi:10.1038/s41397-020-00205-5
39. Graansma LJ, Zhai Q, Busscher L, Menafra R, van den Berg RR, Kloet SL, et al. From gene to dose: refining CYP2C19 phenotype predictions. Front Pharmacol. 2023;14:1076574. doi:10.3389/fphar.2023.1076574
40. Shugg T, Ly RC, Osei W, Rowe EJ, Granfield CA, Lynnes TC, et al. Computational pharmacogenotype extraction from clinical NGS. Front Oncol. 2023;13:1199741. doi:10.3389/fonc.2023.1199741
41. Pratt VM, Everts RE, Aggarwal P, Beyer BN, Broeckel U, Epstein-Baak R, et al. Characterization of 137 genomic DNA reference materials. J Mol Diagn. 2016;18(1):109–23. doi:10.1016/j.jmoldx.2015.08.005
42. Gaedigk A, Turner A, Everts RE, Scott SA, Aggarwal P, Broeckel U, et al. Reference materials for CYP2D6 genetic testing. J Mol Diagn. 2019;21(6):1034–52. doi:10.1016/j.jmoldx.2019.06.007
43. Gaedigk A, Boone EC, Scherer SE, Lee SB, Numanagić I, Sahinalp C, et al. CYP2C8, CYP2C9, and CYP2C19 characterization. J Mol Diagn. 2022;24(4):337–50. doi:10.1016/j.jmoldx.2021.12.011
44. Ramudo-Cela L, López-Martí JM, Colmeiro-Echeberría D, De-Uña-Iglesias D, Santomé-Collazo JL, Monserrat-Iglesias L. Development of an NGS panel for clinical pharmacogenetics. Farm Hosp. 2020;44(6):243–53. doi:10.7399/fh.11353
45. Stevanovski I, Chintalaphani SR, Gamaarachchi H, Ferguson JM, Pineda SS, Scriba CK, et al. Comprehensive genetic diagnosis of tandem repeat expansion disorders. Sci Adv. 2022;8(9):eabm5386. doi:10.1126/sciadv.abm5386
46. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, et al. A draft human pangenome reference. Nature. 2023;617(7960):312–24. doi:10.1038/s41586-023-05896-x
47. Payne A, Holmes N, Clarke T, Munro R, Debebe BJ, Loose M. Readfish enables targeted nanopore sequencing. Nat Biotechnol. 2021;39(4):442–50. doi:10.1038/s41587-020-00746-x
48. Ulrich JU, Lutfi A, Rutzen K, Renard BY. ReadBouncer: precise adaptive sampling. Bioinformatics. 2022;38(Suppl_1):i153–60. doi:10.1093/bioinformatics/btac223
49. Weilguny L, De Maio N, Munro R, Manser C, Birney E, Loose M, et al. Dynamic adaptive sampling using Bayesian experimental design. Nat Biotechnol. 2023;41:1018–25. doi:10.1038/s41587-022-01580-z
50. Alexander P, Rory M, Nadine H, Christopher M, Matt C, Matthew L. Barcode aware adaptive sampling for Oxford nanopore sequencers. bioRxiv. 2022.
51. Munro R, Holmes N, Moore C, Carlile M, Payne A, Tyson JR, et al. Real-time monitoring and adaptive sampling of viral nanopore sequencing. Front Genet. 2023;14:1138582. doi:10.3389/fgene.2023.1138582
52. McInnes G, Dalton R, Sangkuhl K, Whirl-Carrillo M, Lee SB, Tsao PS, et al. Transfer learning enables prediction of CYP2D6 haplotype function. PLoS Comput Biol. 2020;16(11):e1008399. doi:10.1371/journal.pcbi.1008399