



Analysis tools for resolving challenging medically relevant genes using nanopore whole-genome and digital panel data



Sergey Nurk², Alexey Zabelkin², Rebecca N. Palmer², Lynn Ly³, Rachel Martin², Ed Day², Heather Jeffery², Sissel Juul³, Philipp Rescheneder¹

¹Oxford Nanopore Technologies GmbH, Munich, Germany; ²Oxford Nanopore Technologies plc, Oxford, UK; ³Oxford Nanopore Technologies Inc., New York, USA

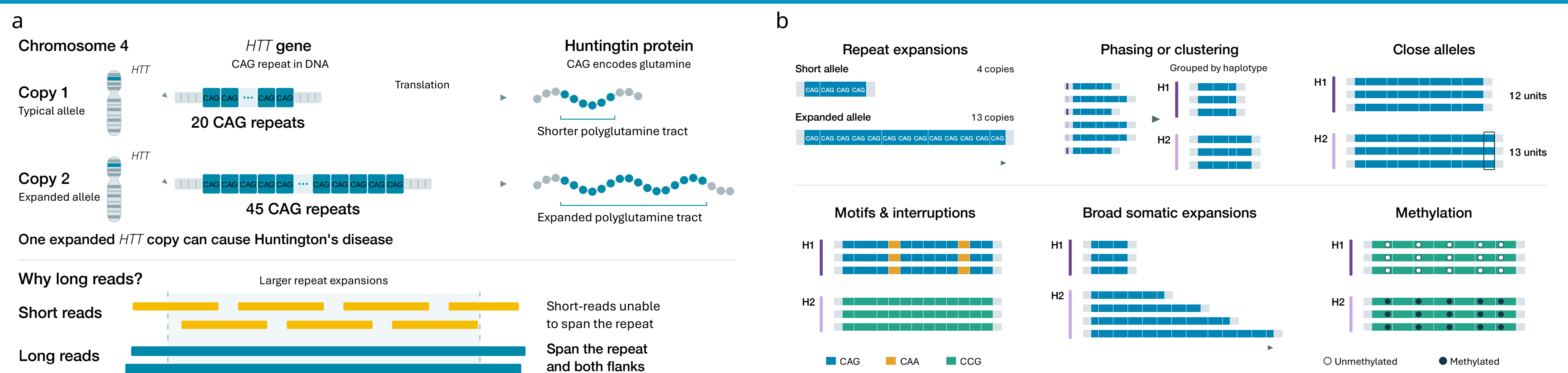
Abstract

Clinically important variation often occurs at loci that are not reliably resolved by conventional assays. Short-read sequencing and non-sequencing molecular assays can fail to resolve allele sequence, phase, motif composition, methylation, or copy number, particularly for short tandem repeat (STR) expansions and genes in segmental duplications or with pseudogenes. Nanopore whole-genome sequencing (WGS) can address these loci by spanning expanded repeats, informative variants, and structural breakpoints while preserving methylation information. Routine analysis requires streamlined, well-tested bioinformatic tools optimised for nanopore data and the characteristics of these medically relevant loci.

Mahi-mahi provides haplotype-resolved genotyping and annotation of clinically relevant STRs. It analyses spanning and partially overlapping reads, decomposes expected and unexpected motifs, clusters alleles, and reports repeat counts in repeat units (RU). It supports STRchive-derived and custom catalogues, compatibility with Stranger annotation, repeat-interruption analysis, and integration with methylation analysis. In 245 samples from the 1000 Genomes Project, sequenced with R10 chemistry, repeat lengths agreed with assembly-derived calls within one repeat unit in 98.6% of 10,927 filtered allele comparisons. Mahi-mahi detected all expected expansions in 14 Coriell cell lines, including complex motif configurations and length mosaicism at loci such as *FMR1*, *HTT*, *RFC1*, *C9orf72*, and *ATXN1*.

Sillago characterises clinically relevant genes located in segmental duplications using nanopore WGS and digital panel data. It phases reads using informative variants and breakpoints, estimates copy number, builds haplotype consensus sequences, calls variants against primary genes, and detects gene-pseudogene fusions. Locus-specific annotations highlight known clinically relevant variation, including *F8* inversions and *SMN1/SMN2* deletions. We evaluated Sillago using 35 annotated reference cell lines and samples from the 1000 Genomes Project across loci including *SMN1/SMN2*, *GBA*, *HBA1/HBA2*, *F8*, *RCCX/CYP21A2*, and *PMS2*. In the reference panel, it matched *SMN1* copy number in 35/35 samples and *SMN2* copy number in 34/35. Together, Mahi-mahi and Sillago enable streamlined analysis of challenging medically relevant regions beyond standard variant calling for rare disease research, carrier screening, and population genomics.

Genetic and epigenetic features of short tandem repeats



Short tandem repeats (STRs) consist of adjacent copies of a short DNA motif. More than 70 disease-associated tandem repeat loci are catalogued in STRchive¹, with locus-specific interpretation supported by annotation tools such as Stranger². Changes in repeat number can alter gene function and cause inherited disease. In *HTT* (a), expansion of the CAG repeat adds extra glutamine building blocks to the huntingtin protein; the schematic contrasts illustrative alleles with 20 and 45 repeat copies.

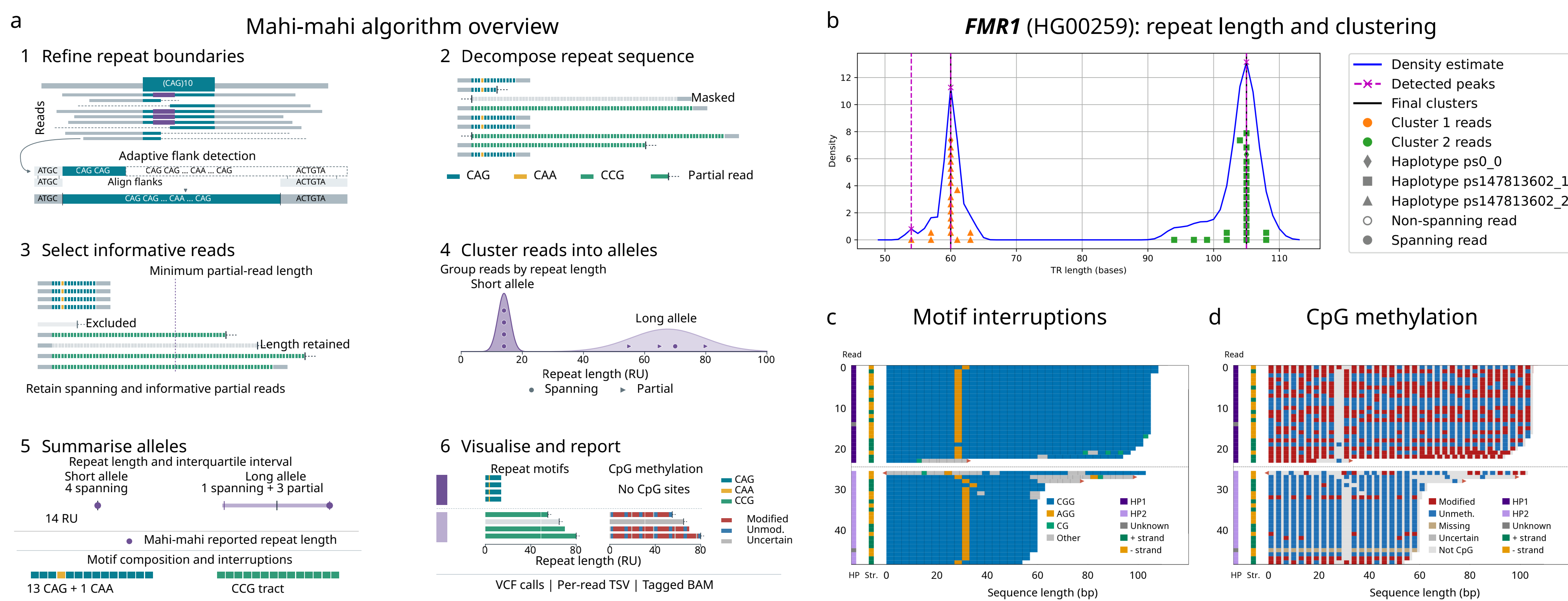
A repeat expansion may extend beyond the span of short reads. Long nanopore reads can span the full repeat and its flanking sequence, retaining the context needed to locate and characterise individual alleles. Repeat length is only one part of the full picture (b). Alleles may differ in motif composition or interruptions, while somatic variation can produce a broad distribution of lengths within a sample. Methylation provides an additional layer of information at relevant loci. Analysing individual molecules preserves the relationship between these features and supports phasing or clustering into alleles. Reliable interpretation therefore requires repeat counts, sequence structure, and read-level variation to be considered together, with locus-specific thresholds and appropriate reference evidence.

References

- Hiatt et al. STRchive: a dynamic resource detailing population-level and locus-specific insights at tandem repeat disease loci. *Genome Med.* 17, 29 (2025). doi:10.1186/s13073-025-01454-4.
- Clinical Genomics, Stranger: annotation of tandem repeat pathogenicity. <https://github.com/Clinical-Genomics/stranger> (accessed 18 October 2026).
- Park, J., Kaufman, E., Valdimanis, P. N., and Bafna, V. TRViz: A Python library for decomposing and visualizing tandem repeat sequences. *Bioinform. Adv.* 3, vbad058 (2023). doi:10.1093/bioadv/vbad058.
- Gustafson et al. High-coverage nanopore sequencing of samples from the 1000 Genomes Project to build a comprehensive catalog of human genetic variation. *Genome Res.* 34, 2061-2073 (2024). doi:10.1101/gr.792723.124.
- Ren, Gu, and Chaisson, vamous: variable-number tandem repeats annotation using efficient motif sets. *Genome Biol.* 24, 175 (2023). doi:10.1186/s13059-023-03010-y.
- Dolzhenko et al. Characterization and visualization of tandem repeats at genome scale. *Nat. Biotechnol.* 42, 1606-1614 (2024). doi:10.1038/s41587-023-02057-3.
- <https://www.pacb.com/technology/puretarget/>
- Chen et al. Genome-wide profiling of highly similar paralogous genes using HiFi sequencing. *Nat. Commun.* 16, 2340 (2025). doi:10.1038/s41467-025-57505-2.
- Shaw, J., Boucher, C., Yu, Y. W., Noyes, N., and Li, H. Long-read reconstruction of many diverse haplotypes with devider. *Genome Res.* 35, 2637-2649 (2025). doi:10.1101/gr.280510.125.
- Coriell Institute for Medical Research, GM03814. https://catalog.coriell.org/0/Sections/Search/Sample_Detail.aspx?Product=CC&Ref=GM03814 (accessed 18 October 2026).
- Prior et al. Characterization of reference materials for spinal muscular atrophy genetic testing: a GeT-RM collaborative project. *J. Mol. Diagn.* 23, 103-110 (2021). doi:10.1016/j.jmoldx.2020.10.011.

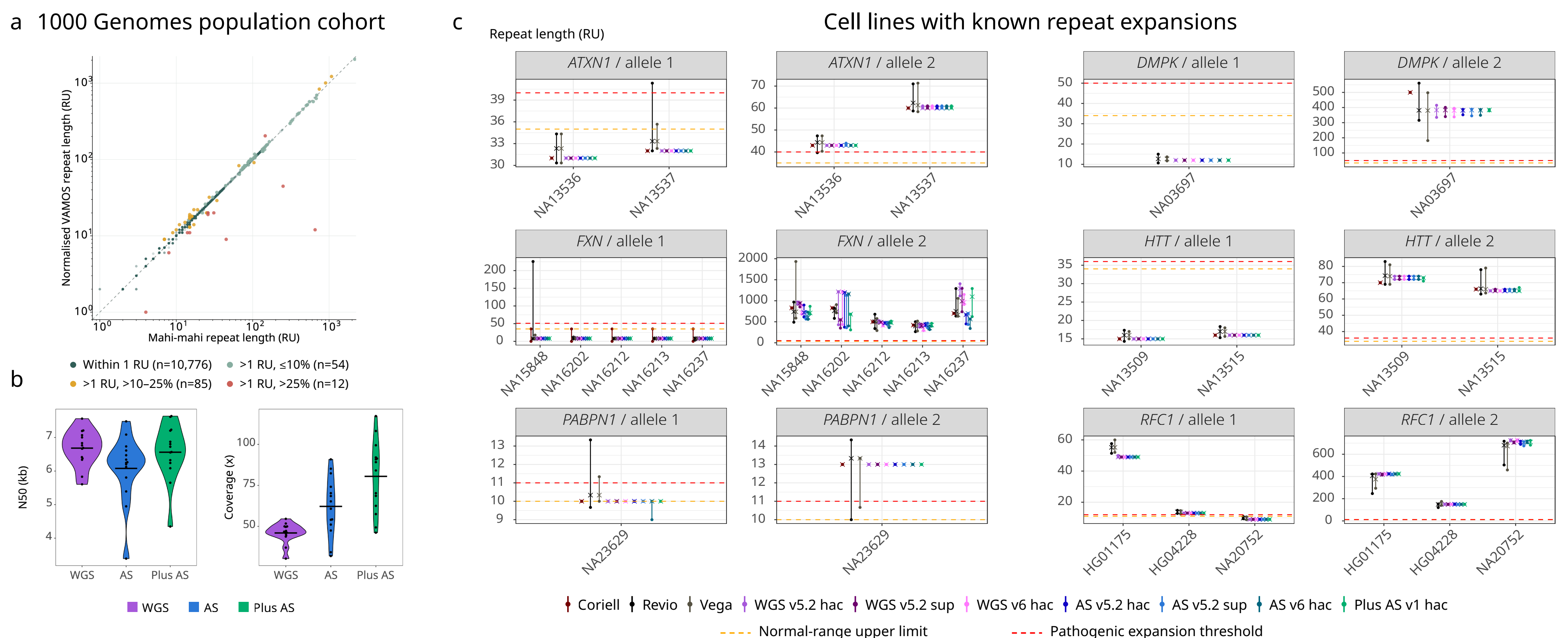
Oxford Nanopore Technologies, the Wheel icon, MinION are registered trademark of Oxford Nanopore Technologies plc in various countries. All other brands and names contained are the property of their respective owners. © 2026 Oxford Nanopore Technologies plc. All rights reserved. Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition.

Mahi-mahi: haplotype-resolved tandem repeat analysis



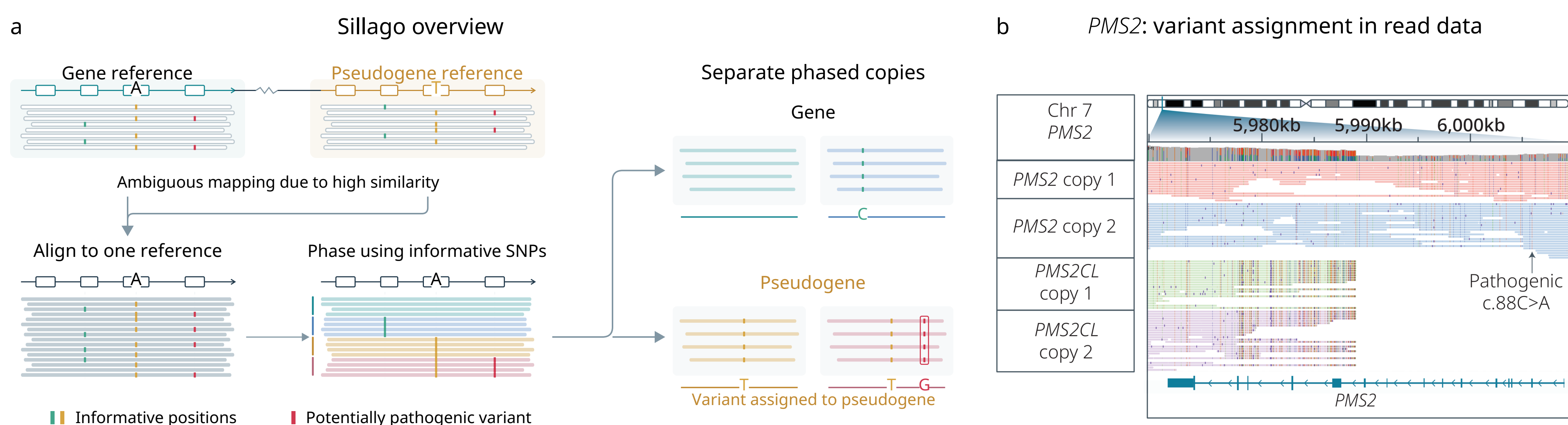
Mahi-mahi retrieves reads overlapping each repeat and realigns reference flanks to refine its boundaries, including reads with clipped or misaligned sequence (a). Repeat sequences are decomposed into motifs with TRviz³, preserving their order and counts. Spanning reads and sufficiently informative non-spanning reads are clustered by repeat length; motif patterns and existing haplotags help resolve closely spaced or overlapping alleles. Reads with strand-specific evidence of basecalling error are excluded from motif and modification summaries while retaining their length evidence. Allele summaries report repeat length and read-level variation, motif composition, and CpG modifications. Output includes a VCF, a per-read TSV, a tagged BAMfile, and visualisations supporting downstream interpretation. Built-in loci derived from STRchive and Stranger catalogues, custom loci, and Stranger-compatible annotation are supported. HG00259 *FMR1* outputs illustrate repeat-length clustering (b), motif interruptions (c), and CpG methylation (d) in Mahi-mahi.

Mahi-mahi: repeat-length concordance with reference data



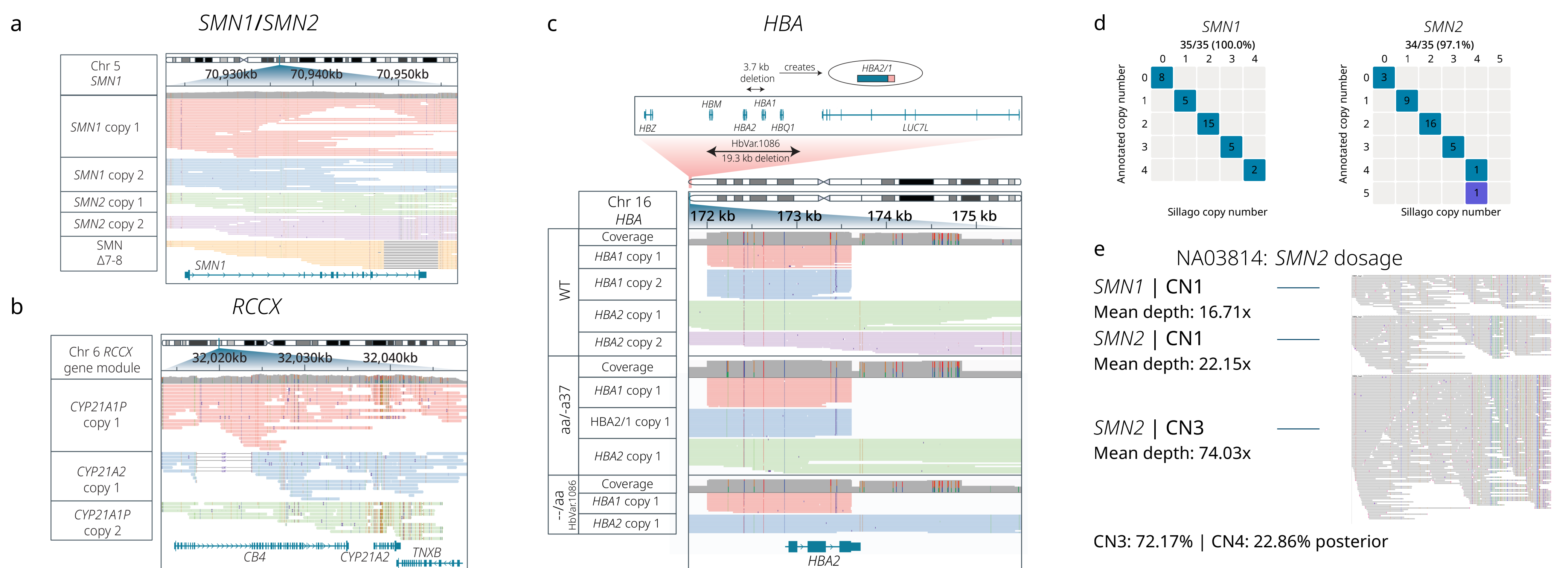
In 245 samples from the 1000 Genomes Project⁴, Mahi-mahi repeat lengths were compared with vamous⁵ calls from haplotype-resolved assemblies to evaluate performance in a population cohort (a). Across 10,927 allele comparisons at 57 loci, 94.2% matched exactly and 98.6% agreed within one repeat unit. Repeat-length definitions were harmonised before comparison, and missing, low-depth, or uninterpretable calls were excluded. Next, 14 cell lines with known repeat expansions were sequenced using WGS and Adaptive Sampling (AS) for targeted enrichment during sequencing. Read length N50 and depth are shown in (b). Results were evaluated against their expected repeat lengths (not reported for *RFC1*) and TRGT⁶ calls from public PacBio PureTarget Revio and Vega datasets⁷ (c). Mahi-mahi detected all 14 expected Coriell expansions.

Sillago: phasing in homologous genomic regions



In addition to tandem repeats, medically relevant variation occurs in genes with highly similar copies or pseudogenes. Their sequence similarity makes read placement ambiguous, obscuring whether a variant affects a functional gene or its pseudogene. Long reads link variants to informative differences between copies, enabling them to be phased and assigned to the relevant gene⁸. Sillago (a) remaps homologous reads to a common reference and phases them using a positional de Bruijn graph⁹. Read-supported graph simplification combines informative variants and structural breakpoints. Shared variant patterns link reads into haplotypes even when individual reads cannot be mapped uniquely to their original copy. Read depth estimates copy number; haplotype consensus sequences support variant calling (VCF) and locus-specific annotation (YAML). Tagged reads (BAM) enable visual inspection. The *PMS2* example (b) assigns the labelled c.88C>A variant to a copy of *PMS2*.

Sillago: structural variant analysis and copy-number concordance



Alongside *PMS2/PMS2CL*, Sillago supports *SMN1/SMN2*, *RCCX*, *GBA*, and *HBA*. SMN reads (a) separate intact *SMN1* and *SMN2* copies from an exon 7-8 deletion. *RCCX* (b) distinguishes *CYP21A2* and *CYP21A1P* read groups. *HBA* examples (c) show a 3.7 kb deletion with an *HBA2/1* hybrid and a 19.3 kb deletion. In 35 cell lines with known SMN status, Sillago matched *SMN1* copy number in 35/35 and *SMN2* in 34/35 (d). In NA03814 (e), Sillago called *SMN2* CN4: one CN1 cluster plus a cluster favouring CN3 (72.2% posterior probability) over CN4 (22.9%). Coriell notes cytogenetic instability and lists five *SMN2* copies¹⁰. GeT-RM assays reported four, four-plus, or five copies¹¹ but retained CN5 as consensus truth.

Conclusions

Mahi-mahi and Sillago extend nanopore analysis to challenging medically relevant loci. In 245 samples from the 1000 Genomes Project, Mahi-mahi agreed with assembly-derived repeat lengths within one repeat unit in 98.6% of 10,927 filtered allele comparisons. It detected all expected expansions in 14 Coriell cell lines. In 35 annotated reference cell lines, Sillago matched *SMN1* and *SMN2* copy number in 35/35 and 34/35 samples. These results show how both tools streamline nanopore analysis for rare disease research, carrier screening, and population genomics.