Oxford Nanopore Technologies

Resolving challenging medically relevant genes in the human genome with native Oxford Nanopore reads

Oxford Nanopore reads of unrestricted length allow copy number variation calling for genes with highly similar paralogs, along with tandem repeat expansion characterisation and motif identification

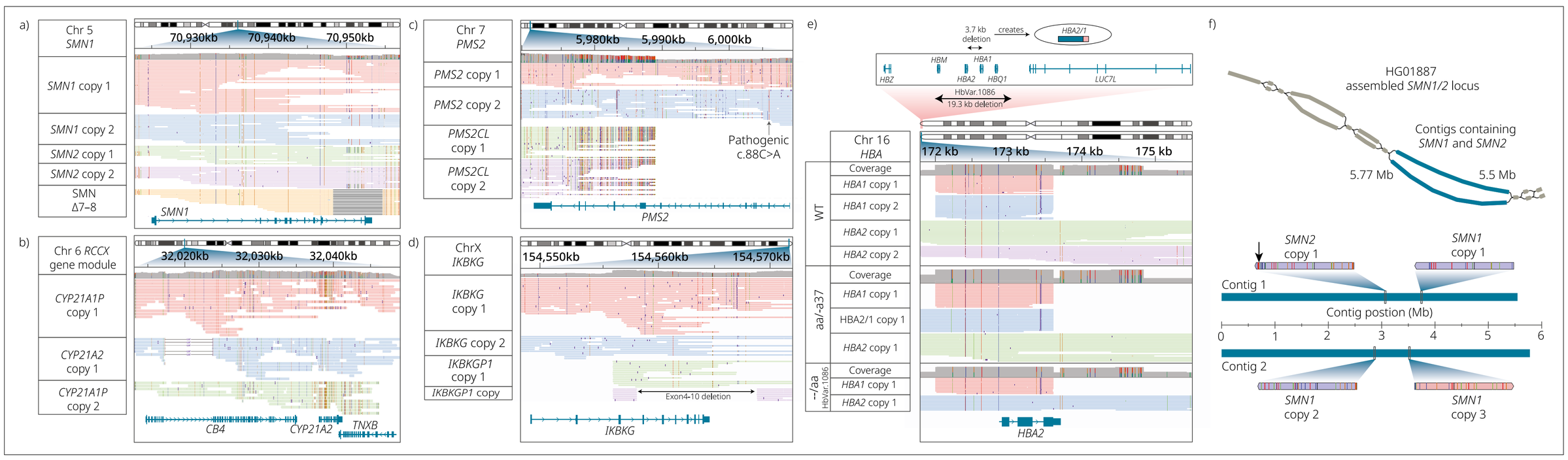

Fig. 1 a) Exon 7—8 deletion in *SMN1*, b) copy number loss of *CYP21A2*, c) pathogenic variant in *PMS2*, d) exon 4—10 deletion in *IKBKG*, e) deletions in *HBA1/2*, and f) assembly of SMA locus.

Distinguishing medically relevant genes from their paralogs and enumerating copy number changes through phasing or assembly-based methods

Several clinically relevant genes have highly similar paralogues or pseudogenes. As a result, determining copy number variants (CNVs) and identifying pathogenic variants pose significant technical challenges. For variant annotation, assignment to either the primary gene or pseudogene is required. Whole-genome Oxford Nanopore sequencing enabled phasing across these genes, thereby distinguishing homologous gene copies in four research cell lines. Loss of *SMN1*, for example by an exon 7—8 deletion in this carrier sample, is associated with spinal muscular atrophy (**Fig. 1a**), loss of *CYP21A2* at the RCCX locus, shown by total module deletion in this carrier sample, is associated with congenital adrenal hyperplasia (**Fig. 1b**), loss of functional *PMS2* is associated with Lynch syndrome (**Fig. 1c**) and loss of functional *IKBKG*, as shown in this female carrier, is associated with incontinentia pigmenti (**Fig. 1d**). Various deletions of *HBA1/2* are responsible for alpha thalassemia, resulting in either an *HBA2/1* fusion gene or complete loss of both genes, which can be annotated *de novo* using the phased haplotypes, compared to wild type (**Fig. 1e**). Assembly of the entire SMA locus was performed using Verkko with Herro-corrected ultra-long reads from HG01887. This resulted in haplotype-resolved contigs, greater than 5 Mb in length, that contained the *SMN1* and *SMN2* genes. Alignments of the *SMN1* reference gene to these contigs determined the correct overall ploidy, while identification of the discriminatory *SMN2* variant correctly enumerated the gene copy numbers (**Fig. 1f**).

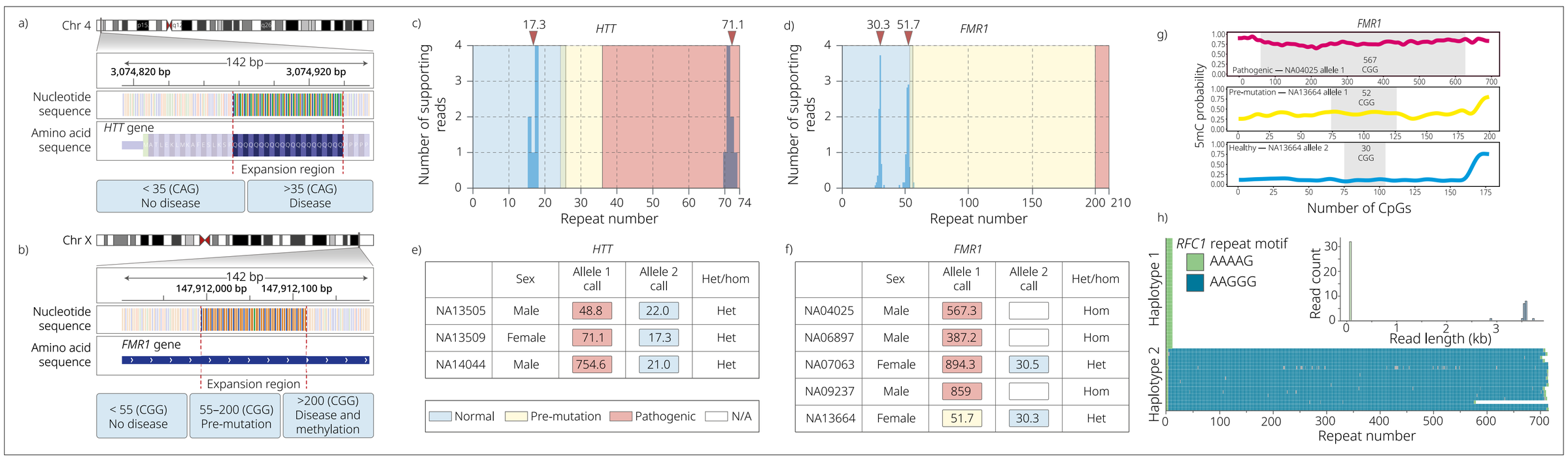

Fig. 2 a) *HTT* gene, b) *FMR1* gene, c) *HTT* repeat expansion, d) *FMR1* pre-mutation, e) *HTT* samples, f) *FMR1* samples, g) 5mC methylation in *FMR1*, and h) *RFC1* motif expansion.

Quantifying tandem repeat expansions in *FMR1* and *HTT* plus methylation in *FMR1*, and characterising motif switching and expansion in the AluSx3 element of *RFC1*

Short tandem repeats (STRs) are repetitive DNA sequences consisting of 2–6 bp repeat units. Changes in repeat lengths are associated with numerous genetic disorders. For example, Huntington's disease is characterised by a CAG repeat in the *HTT* gene (**Fig. 2a**). Fragile X syndrome (FXS) has an expansion of a CGG repeat in the 5′ UTR of *FMR1* (**Fig. 2b**). Eight cell line research samples derived from subjects with either Huntington's disease or FXS were sequenced. Sample NA13509 contains one *HTT* haplotype with an expansion into the full mutation range (**Fig. 2c**). Sample NA13664 was derived from a woman with developmental delays but not full FXS; however, one *FMR1* allele is on the pre-mutation borderline (**Fig. 2d**). **Tables 2e and 2f** show correct classification of known expanded alleles into full or pre-mutation status for all samples. The *FMR1* CGG repeat within the promoter region has a distinct correlation between its 5mC methylation and CGG repeat length, with hypomethylation found in healthy repeat lengths, and increasing methylation levels with increasing repeat lengths (**Fig. 2g**). Biallelic repeat expansions in *RFC1* are associated with cerebellar ataxia, neuropathy, and vestibular areflexia syndrome (CANVAS). This expansion occurs within the poly(A) tail of an AluSx3 element in the form of an AAGGG motif. In an additional sample with an expanded *RFC1* allele, sequence clustering and motif detection were used to identify reads from the expanded allele and determine the motif structure (**Fig. 2h**).

PO_1246(EN)_V3_19May2026