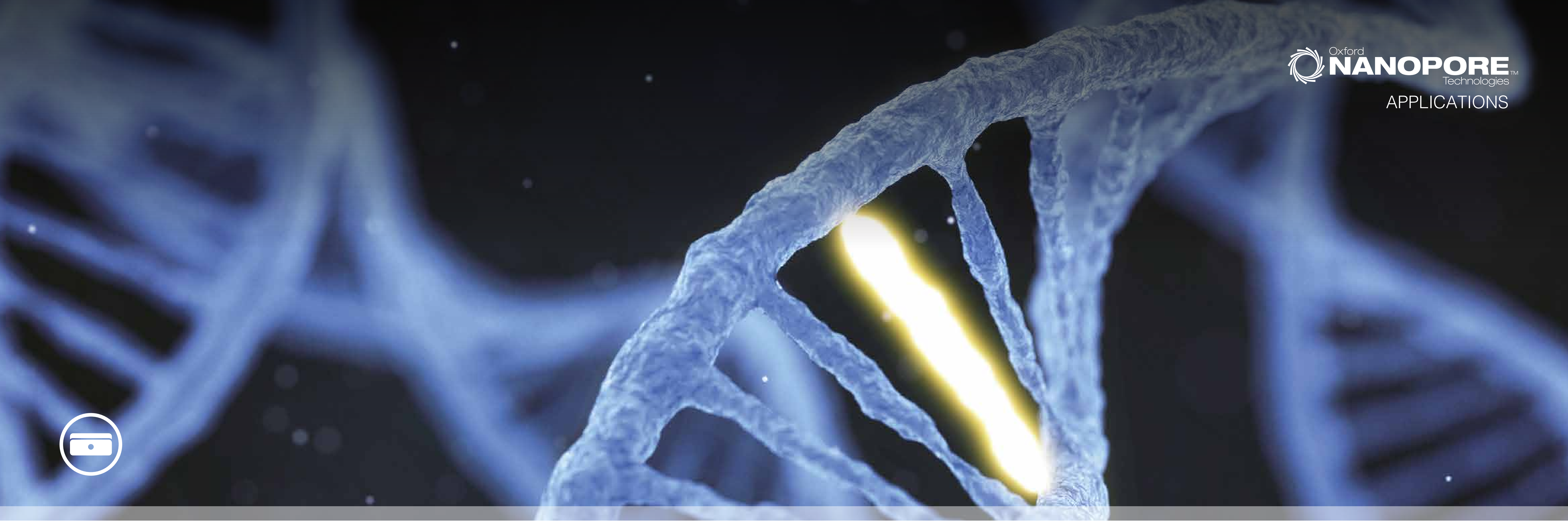

Oxford NANOPORE Technologies
APPLICATIONS

Performance benchmarking for calling and phasing of single-nucleotide polymorphisms and structural variants

The length, accuracy and low bias of nanopore reads makes them ideally suited to the characterisation and phasing of structural variants and single-nucleotide polymorphisms across the entire genome

Contact: **apps_posters@nanoporetech.com** More information at: **www.nanoporetech.com** and **publications.nanoporetech.com**

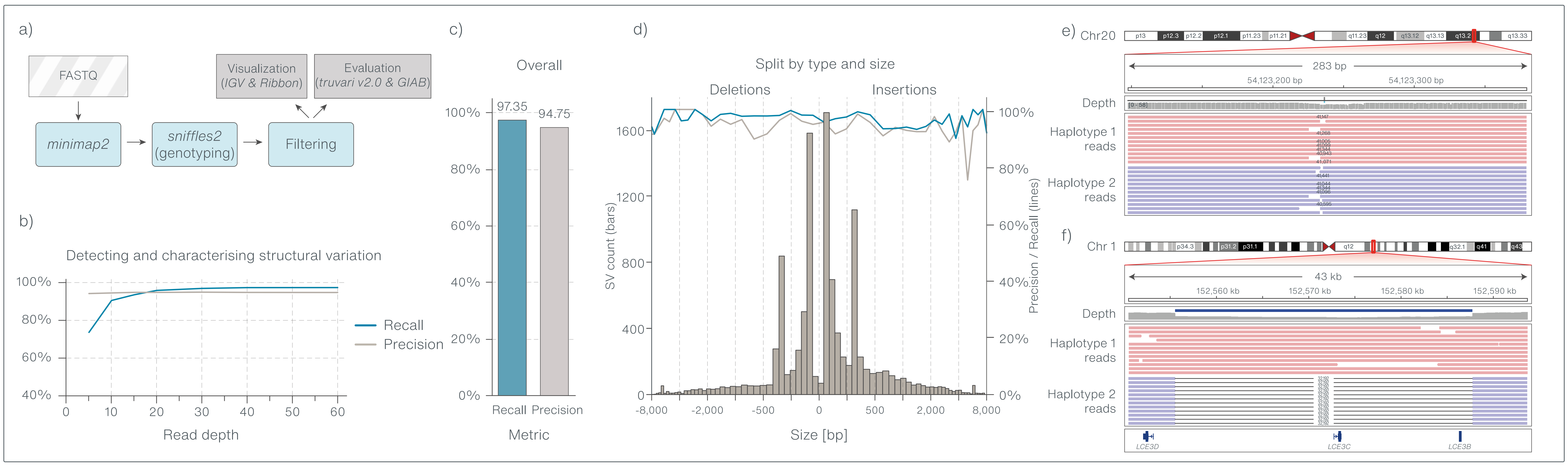

Fig. 1 Calling structural variants (SV) with nanopore reads a) bioinformatics workflow b) and c) precision and recall d) calling different SV classes e) calling very long insertions and deletions

The read lengths enabled by nanopore sequencing make it feasible to call long structural variants with high sensitivity and specificity, even in difficult regions of the human genome

Our newest version of the SV analysis pipeline (Fig. 1a) benefits from *minimap2* v.2.24's improved performance in aligning long reads in areas of structural variation and repeats. Furthermore, *sniffles2*'s new tandem-repeat-aware functionality leads to more consistent placement of variant calls within reference tandem repeat sequence, increasing standard validation-set concordance. To validate this pipeline we called SVs with up to 60x of HG002 data and compared this to the most recent GIAB truth-set. We found high overall performance with >95% precision and >97% recall (Figs. 1b and 1c). The size distribution of the SVs clearly shows the expected ALU peak at around 300 bp and a LINE1 peak above 5,000 bp. Low-complexity short tandem repeats are known to be under-represented in the current human reference and this is seen as a higher number of insertions than deletions in our data. We found that high precision is independent of read depth. Recall is more directly influenced by read depth, with 20x being required for identifying >95% of all variants. However, even with as little as 10x, >90% are detected. Performance is largely independent of SV size (Fig. 1d); even very large SVs are called with base-pair precision (Figs. 1e and f). These examples show SV-calling combined with read phasing: Fig. 1e) shows a homozygous 40 kb insertion, captured in single reads from each haplotype; Fig. 1f) shows a 30 kb heterozygous deletion that covers *LCE3B* and *LCE3C*. Deletion of these two genes results in susceptibility to psoriasis.

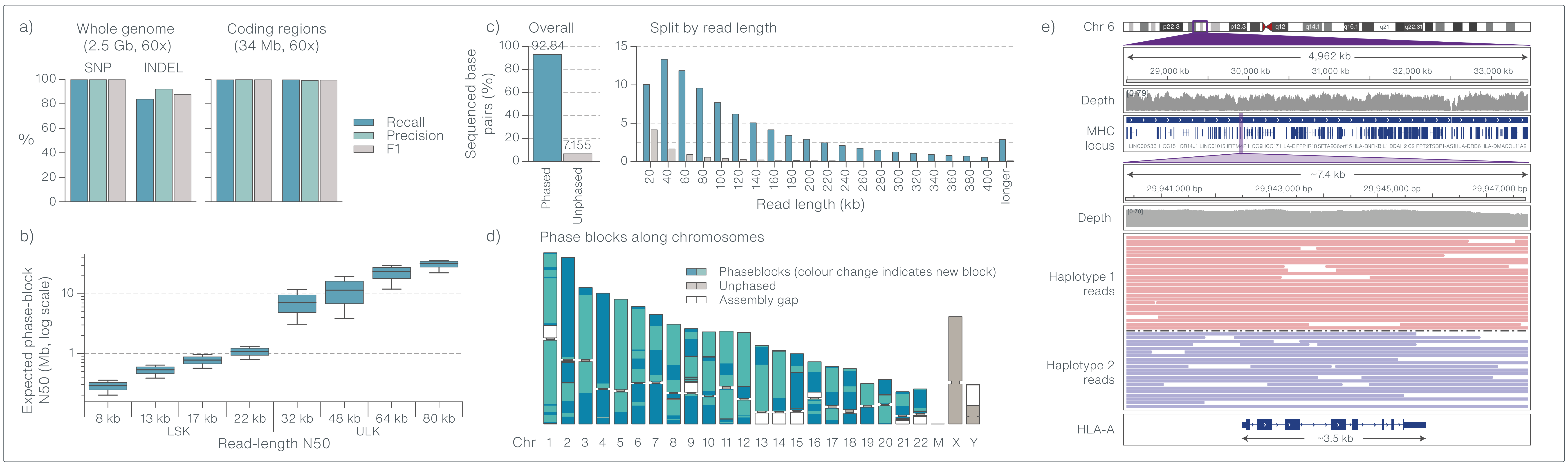

Fig. 2 SNP-calling a) recall and precision using publicly available datasets b)-d) phasing e) phasing genes in the major incompatibility complex

The combination of long reads with highly sensitive and specific calling of single-nucleotide polymorphisms simplifies phasing and results in exceptionally long phase-blocks

To test SNP- and indel-calling performance we used *Clair3* on a 60x nanopore HG002 dataset. Results show high precision (>99.8%) and recall (>99.8%) for substitutions across the whole genome (Fig. 2a left). In coding regions, which contain >96% of ClinVar pathogenic indels, results also show high precision (>99.7%) and recall (>99.2%) for indel calls (Fig. 2a right). The length of nanopore reads can be leveraged to assign SNP, SVs, and methylation calls to their homologous chromosome of origin. By using information from overlapping long reads the resulting phase blocks can span tens of megabases, typically only interrupted by long stretches of homozygosity. To show the influence of read length on phase-block length, we used *whatshap* on LSK and ultra-long (ULK) HG002 data with increasing read lengths (Fig. 2b). For 10-15 kb reads we see a phase-block N50 of between 100-800 kb, sufficient to fully phase roughly 60% of all gene bodies. Increasing read length increases phase-block contiguity up to a hundred-fold, with N50 exceeding 30 Mb for the longest ULK datasets, allowing full phasing of >94% of all genes. A high percentage of phased reads allows phasing of small variants, SVs and methylation calls. We used *whatshap haplotag* to scan raw reads for phased variants. Using a ULK dataset we could assign >92% of all sequenced nucleotides to a haplotype (Fig. 2c) and cover >98% of the entire genome (excluding centromeres and telomeres) with phased reads. Four chromosomes had > 95% of sequence covered by just two phase blocks (Fig. 2d). To demonstrate phasing performance on medically relevant loci, we choose the MHC region, where we could cover 99.6% of the 6 Mb locus with 2 phase blocks and fully phase all genes (Fig. 2e).