Supplementary Material

1. Supplementary Results

Supplementary Table 1. The outcomes derived from the minimap2 tool encompassed the quantification of aligned reads based on MAPQ metrics exceeding thresholds of 10, 30, and 50. Across all experimental runs, the utilization of the modified index consistently yielded an augmentation in the ultimate MAPQ score as computed by the minimap2 tool.

k-mer and window sizes	MAPQ	% mapped reads with certain MAPQ	
		default index	modified index
k 15, w 8	10	92.128	92.154
	30	90.090	90.143
	50	87.901	87.981
k 21, w 11	10	92.065	92.152
	30	90.003	90.663
	50	87.790	88.912
k 27, w 14	10	91,671	91.716
	30	90.327	90.391
	50	88.597	88.687

Supplementary Table 2. The evaluation of the acquired VCF files against the reference was performed within "confident regions", utilizing HG002 raw reads and reference data sourced from PrecisionFDA Truth Challenge V2 aligned to the linear reference sequence GRCh38 (GCA_000001405.15).

k-mer and window sizes	Metric	SNVs		indels	
		default idx	modified idx	default idx	modified idx
k 15, w 8	TP	3330615	3332000	520316	520314
	FN	34512	33127	5153	5155
	FP	21526	21488	4311	4320
	Recall	0.98974	0.99016	0.99019	0.99019
	Precision	0.99358	0.99359	0.99179	0.99177
	F1-Score	0.99166	0.99187	0.99099	0.99098
k 21, w 11	TP	3329372	3335444	520264	520487
	FN	35755	29683	5205	4982
	FP	21208	20665	4318	4112

	Recall	0.98938	0.99118	0.99001	0.99052
	Precision	0.99367	0.99384	0.99178	0.99217
	F1-Score	0.99152	0.99251	0.99093	0.99134
k 27, w 14	TP	3334076	3336158	520452	520527
	FN	31051	28969	5017	4942
	FP	17957	17757	3705	3678
	Recall	0.99077	0.99139	0.99045	0.99060
	Precision	0.99464	0.99471	0.99293	0.99299
	F1-Score	0.99270	0.99304	0.99169	0.99179

Supplementary Table 3. The assessment of the generated VCF files against the reference was performed, utilizing HG002 raw reads and reference data sourced from PrecisionFDA Truth Challenge aligned to the linear reference sequence hs37d5.

k-mer and window sizes	Metric	SNVs		indels	
		default idx	modified idx	default idx	modified idx
k 15, w 8	TP	3052331	3052948	343329	343340
	FN	2316	1699	1237	1226
	FP	9772	9974	1589	1571
	Recall	0.99924	0.99944	0.99641	0.99644
	Precision	0.99681	0.99674	0.99539	0.99544
	F1-Score	0.99802	0.99809	0.99590	0.99594
k 21, w 11	TP	3053019	3053464	343301	343317
	FN	1628	1183	1265	1249
	FP	8127	8349	1541	1574
	Recall	0.99947	0.99961	0.99633	0.99638
	Precision	0.99735	0.99727	0.99553	0.99543
	F1-Score	0.99840	0.99844	0.99593	0.99591
k 27, w 14	TP	3053056	3053604	343318	343330
	FN	1591	1043	1248	1236
	FP	8012	8202	1517	1526
	Recall	0.99948	0.99966	0.99638	0.99641
	Precision	0.99738	0.99732	0.99556	0.99557

	F1-Score	0.99843	0.99849	0.99599	0.99599

Supplementary Table 4. Running time of the minimap2 alignment tool with HG002 raw reads from PrecisionFDA Truth Challenge aligned to the linear reference sequence hs37d5.

k-mer and window sizes	Time	
	default index	modified index
k 15, w 8	115m 11.703s	116m 15.529s
k 21, w 11	88m 27.425s	89m 54.367s
k 27, w 14	94m 22.050s	91m 35.477s

Supplementary Table 5. The assessment of the generated VCF files against the reference was performed, utilizing HG002 reference data from the PrecisionFDA Truth Challenge aligned to the linear reference sequence hs37d5. Synthetic reads generated by HG002 benchmark VCF file using ART toolset.

Chromosome	Metric	SNVs		indels	
		default idx	modified idx	default idx	modified idx
1	TP	238882	239124	24846	24853
	FN	1082	840	2270	2263
	FP	230	234	2127	2115
	Recall	0.99549	0.99650	0.91629	0.91654
	Precision	0.99304	0.99902	0.92104	0.92147
	F1-Score	0.99726	0.99776	0.91865	0.91900
6	TP	186476	186758	20188	20189
	FN	1167	885	1652	1651
	FP	160	138	1515	1509
	Recall	0.99378	0.99528	0.92436	0.92440
	Precision	0.99914	0.99926	0.93015	0.93041
	F1-Score	0.99645	0.99727	0.92725	0.92740

Supplementary Table 6. The evaluation of the acquired VCF files against the reference was performed within "confident regions", utilizing HG002 raw reads and reference data sourced from PrecisionFDA Truth Challenge V2 aligned to the linear reference sequence GRCh38 (GCA_000001405.15). Index modified with the addition of population-specific mutations from gnomAD v3.1.2.

k-mer and window sizes	Added population	Metric	SNVs	indels
			modified idx	
k 27, w 14	Ashkenazi Jewish	TP	3334248	520451
		FN	30879	5018
		FP	18235	3729
		Recall	0.99082	0.99045
		Precision	0.99456	0.99289
		F1-Score	0.99269	0.99167
k 27, w 14	East Asian	TP	3334143	520441
		FN	30984	5028
		FP	18596	3765
		Recall	0.99079	0.99043
		Precision	0.99445	0.99282
		F1-Score	0.99262	0.99163
k 27, w 14	Admixed American	TP	3334087	520455
		FN	31040	5014
		FP	18244	3715
		Recall	0.99078	0.99046
		Precision	0.99456	0.99291
		F1-Score	0.99266	0.99169

Supplementary methods

1.Data

1.1 Sequencing data in FASTQ format

(HG002 35x)
https://precision.fda.gov/home/files/file-FpZG9Jj0xbJy0QXjB7yb6fpX-1
https://precision.fda.gov/home/files/file-FpZG9Jj0xbJfGbf01qBgKyg1-1

(HG002 50x)
https://precision.fda.gov/home/files/file-BvP0b3Q0Z8pJfQxYxbk8P2p0
https://precision.fda.gov/home/files/file-BvP0f600xKVyPQZj08ZjXPx4

1.2 Benchmark variant calls in VCF file

(HG002 GRCh38 v4.2.1)
https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NISTv4.2.1/GRCh38/HG002_GRCh38_1_22_v4.2.1_benchmark.vcf.gz

https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NISTv4.2.1/GRCh38/HG002_GRCh38_1_22_v4.2.1_benchmark.vcf.gz.tbi

(HG002 GRCh37 v3.2.2)
https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NISTv3.2.2/HG002_GIAB_highconf_IllFB-IllGATKHC-CG-Ion-Solid_CHROM1-22_v3.2.2_highconf.vcf.gz
https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NISTv3.2.2/HG002_GIAB_highconf_IllFB-IllGATKHC-CG-Ion-Solid_CHROM1-22_v3.2.2_highconf.vcf.gz.tbi

1.3 Confident regions in BED file

(HG002 GRCh38 v4.2.1)
https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NISTv4.2.1/GRCh38/HG002_GRCh38_1_22_v4.2.1_benchmark_noinconsistent.bed

(HG002 GRCh37 v3.2.2)
https://ftp-trace.ncbi.nlm.nih.gov/ReferenceSamples/giab/release/AshkenazimTrio/HG002_NA24385_son/NISTv3.2.2/HG002_GIAB_highconf_IllFB-IllGATKHC-CG-Ion-Solid_CHROM1-22_v3.2.2_highconf.bed

1.4 1000 Genomes Project phase 3 VCF files

(GRCh38)
Downloaded from:
ftp://hgdownload.soe.ucsc.edu/gbdb/hg38/1000Genomes/

(GRCh37)
Downloaded from:
http://hgdownload.cse.ucsc.edu/gbdb/hg19/1000Genomes/phase3/

1.5 gnomAD v3.1.2 VCF files (GRCh38)

Downloaded from:
https://storage.googleapis.com/gcp-public-data--gnomad/release/3.1.2/vcf/genomes/

1.6 Additional files for Base Quality Score Recalibrator

(GRCh38)
ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/dbsnp_138.hg38.vcf.gz
ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/dbsnp_138.hg38.vcf.gz.tbi
ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/Mills_and_1000G_gold_standard.indels.hg38.vcf.gz
ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/Mills_and_1000G_gold_standard.indels.hg38.vcf.gz.tbi

(GRCh37)
ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/b37/dbsnp_138.b37.vcf.gz
ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/b37/Mills_and_1000G_gold_standard.indels.b37.vcf.gz

2. Commands

2.1 Command for trimming using cutadapt v1.18

cutadapt -q 0,0 --minimum-length 60 -a AGATCGGAAGAGCACACGTCTGAACTCCAGTCA -A AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT --pair-filter=any -o $output_fastq_file1 -p $output_fastq_file2 $input_fastq_file1 $input_fastq_file2

2.2 Command for minimap2 index creation using minimap2_index_modifier (minimap2 v2.24)

minimap2 -d $output_index -k $k -w $w --vcf-file-with-variants $input_vcf [--parse-haplotype] $ref_fasta

2.3 Commands for read alignment (minimap2 v2.24, samtools v1.18)

minimap2 -a -Y -f $n,$m -x sr -R '@RG\tID:HG002\tSM:HG002\tPL:ILLUMINA' $modified_index $trimmed_fastq1 $trimmed_fastq2 > $output_intermediate_bam

samtools fixmate -m -O SAM $output_intermediate_bam - | samtools sort - | samtools markdup -d 100 - $output_bam
samtools index $output_bam $output_bam_index

2.4 Command for Base Quality Score Recalibration (GATK v4.2.6.1, samtools v1.18)

```
gatk BaseRecalibrator \
   -R $ref_fasta \
   -I $input_bam \
   --read-index $input_bam_index \
   --use-original-qualities \
   -O $report_file \
   --known-sites $dbsnp \
   --known-sites $mills_and_1000g

gatk ApplyBQSR \
   --add-output-sam-program-record \
   -R $ref_fasta \
   -I $input_bam \
   --read-index $input_bam_index \
   --use-original-qualities \
   -O $output_recalibrated_bam \
   -bqsr $report_file \
   --static-quantized-quals 10 \
   --static-quantized-quals 20 \
   --static-quantized-quals 30
```

samtools index $output_recalibrated_bam $output_bam_recalibrated_index

2.5 Command for variant calling (GATK v4.2.6.1)

```
gatk HaplotypeCaller \
   -R $ref_fasta \
   -I $input_recalibrated_bam \
   --read-index $input_recalibrated_bam_index \
   -O $output_vcf \
   -contamination 0 \
   --native-pair-hmm-threads 1
```

2.6 Command for VCF files comparison (hap.py v0.3.15)

hap.py $benchmark_vcf $ result_vcf -f $confident_regions_bed -r $ref_fasta \
-V --engine=vcfeval --no-leftshift --no-decompose --gender=none

2.7 Commands for synthetic read generation (bcftools v1.19, ART v2.5.8)

(example for chromosome 1, same commands for chromosome 6)

```
bcftools consensus -H R -f $input_fasta_chr1 $input_vcf_chr1 > $out_fasta1
bcftools consensus -H A -f $input_fasta_chr1 $input_vcf_chr1 > $out_fasta2
art_illumina -ss HS25 -sam -i  $out_fasta1 -p -l 150 -f 19 -m 200 -s 10 -qs 30 -qs2 30 -o ${prefix1}
art_illumina -ss HS25 -sam -i  $out_fasta2 -p -l 150 -f 19 -m 200 -s 10 -qs 30 -qs2 30 -o ${prefix2}
cat ${prefix1}1.fq ${prefix2}1.fq > $output_fq1
cat ${prefix1}2.fq ${prefix2}2.fq > $output_fq2
```

3 Algorithms

Algorithm 3.1: Minimizer search for the reference genetic sequences

Algorithm 3.1 Minimizer search for the reference genetic sequences

Input: w – window size, k – kmer length, s – sequence ($|s| \geq w + k - 1$)
Output: (w, k) – minimizers and their positions

function GetMinimizers(s, k, w)
    $M \leftarrow \varnothing$
    for $i = 1$ to $|s| - w - k - 1$ **do** ▷ Find minimum value
        $m \leftarrow \infty$
        for $j = 0$ to $w - 1$ **do**
            $u \leftarrow s_i + j_k$
            $v \leftarrow \overline{s}_i + j_k$ ▷ $\overline{s}$ – reversed strand
            if $u \neq v$ **then**
                $m = min(m, min(u, v))$
            end if
        end for
        for $j = 0$ to $w - 1$ **do** ▷ Collect minimizers
            $u \leftarrow s_i + j_k$
            $v \leftarrow \overline{s}_i + j_k$
            if $u < v$ and $u = m$ **then**
                $M \leftarrow M \cup (m, i + j, 0)$
            end if
            if $u > v$ and $v = m$ **then**
                $M \leftarrow M \cup (m, i + j, 1)$
            end if
        end for
    end for
    return M
end function

Algorithm 3.2: Additional minimizer search for SNV

Algorithm 3.2 Additional minimizer search for SNV

Input: w – window size, k – kmer length, s – sequence ($|s| \geq 2 \cdot (w+k) - 1$), (var, pos) – variant and its position
Output: (w, k) – minimizers and their positions

function GetAdditionalMinimizers(s, k, w, var, pos)
  $subs \leftarrow s[pos - k - w + 1] \dots s[pos + k + w - 1]$
  $subs[k + w - 1] \leftarrow var$ ▷ Modify sequence with SNV
  $M \leftarrow \varnothing$
  for $i = 0$ to $2 \cdot (w + k - 1)$ **do** ▷ Find minimum value
    $m \leftarrow \infty$
    for $j = 0$ to $w - 1$ **do**
      $u \leftarrow subs_i + j_k$
      $v \leftarrow \overline{subs}_i + j_k$ ▷ $\overline{subs}$ – reversed strand
      if $u \neq v$ **then**
        $m = min(m, min(u, v))$
      end if
    end for
    for $j = 0$ to $w - 1$ **do** ▷ Collect minimizers
      $u \leftarrow subs_i + j_k$
      $v \leftarrow \overline{subs}_i + j_k$
      if $u < v$ and $u = m$ **then**
        $M \leftarrow M \cup (m, i + j, 0)$
      end if
      if $u > v$ and $v = m$ **then**
        $M \leftarrow M \cup (m, i + j, 1)$
      end if
    end for
  end for
  return M
end function

Algorithm 3.3: Searching valid combinations of SNVs in a window of length k

Algorithm 3.3 Searching valid combinations of SNVs in a window of length k

Input: $genotypes$ – list of SNVs' genotypes, $begin_{ptr}, end_{ptr}$ – pointers to SNVs in current window
Output: $Combs$ – array of SNVs combinations in a window of length k

function GetCombs(s, k, w)
  $Combs \leftarrow \varnothing$
  $Size \leftarrow size(genotypes)$
  for $i = 0$ to $Size - 1$ **do**
    $w_{ptr} \leftarrow begin_{ptr}$
    $CombsWithPos \leftarrow \varnothing$
    while $w_{ptr} \neq end_{ptr}$ **do**
      if $w_{ptr} \rightarrow genotypes[i] = 0$ **then**
        $CombsWithPos \cup (w_{ptr} \rightarrow allele_{ref}, w_{ptr} \rightarrow pos)$
      end if
      if $w_{ptr} \rightarrow genotypes[i] = 1$ **then**
        $CombsWithPos \cup (w_{ptr} \rightarrow allele_{alt}, w_{ptr} \rightarrow pos)$
      end if
      $w_{ptr} \leftarrow w_{ptr} \rightarrow next$
    end while
    $Combs \cup CombsWithPos$
  end for
  return $Combs$
end function