Correcting Modification-Mediated Errors in Nanopore Sequencing by Nucleotide Demodification and in silico Correction

Chien-Shun Chiou[1], Bo-Han Chen[1], You-Wun Wang[1], Nang-Ting Kuo[2], Chih-Hsiang Chang[2], Yao-Ting Huang[2*]

[1] Centers for Disease Control, Taichung, Taiwan.

[2] Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan.

*Corresponding email: ythuang@cs.ccu.edu.tw

Abstract

The accuracy of Oxford Nanopore Technology (ONT) sequencing has significantly improved thanks to new flowcells, sequencing kits, and basecalling algorithms. However, novel modifications untrained in the basecalling models can seriously reduce the quality. This paper reports a set of ONT-sequenced genomes with unexpected low quality (~Q30) due to extensive new modifications. Demodification by whole-genome amplification (WGA) significantly improved the quality of all genomes (~Q50-60) while losing the epigenome. We developed a computational method, Modpolish, for correcting modification-mediated errors without WGA. Modpolish produced high-quality genomes and uncovered the underlying modification motifs without loss of epigenome. Our results suggested that novel modifications are prone to ONT errors, which are correctable by WGA or Modpolish without additional short-read sequencing.

Keywords: DNA modifications, Nanopore Sequencing, Whole-Genome Amplification.

Background

The Oxford Nanopore Technology (ONT) is a popular long-read sequencing platform that enables real-time sequencing for point-of-care medical applications, such as the diagnosis of infectious and newborn diseases within hospitals [1, 2]. Despite its great potential and popularity, the accuracy of ONT was inferior to those of other platforms (e.g., Illumina and PacBio HiFi). Recently, the quality of ONT sequencing has significantly improved thanks to new flowcells (e.g., R10.4), sequencing kits (e.g., Kit 14), and basecalling algorithms (e.g., Bonito). For example, by using the R10.4 flowcells, near-perfect microbial genomes from isolates or metagenomes can be reconstructed by ONT-only sequencing without short-read polishing [3].

However, because the throughput of R10.4 is much lower due to slower sequencing speed, most projects rely on the R9.4 flowcells for higher yield. Although the upcoming sequencing kit will further enhance the accuracy (e.g., Kit 14), postassembly genome polishing is still compulsory for removing ONT systematic errors regardless of the flowcell or kit versions. Systematic errors are recurrent basecalling errors at the same locus, which are not correctable by the consensus of read pileups (e.g., Racon) [4]. Homopolymer errors (i.e., indels) were the primary source of ONT systematic errors. Thanks to several machine-learning algorithms, these errors

have been significantly reduced by read-based (e.g., Medaka) or reference-based (e.g., Homopolish) polishing methods [5]. These algorithmic advances have produced high-quality ONT genomes sufficient for downstream analysis (e.g., >Q50) [3, 6].

Unfortunately, the ONT signals are ultra-sensitive to various modifications (e.g., 5mC, 6mA). More than 17 and 160 modification types have been found in DNA and RNA, respectively, and the number is still growing (e.g., DNA adducts, N4-acetyldeoxycytosine) [7, 8]. These modifications disturb the electrical current and result in unfixable systematic errors [9]. Note that these modification-mediated errors cannot be eliminated by new flowcells and sequencing kits (e.g., R10.4 and Kit 14) which aim to reduce homopolymer errors. Furthermore, existing basecalling and polishing algorithms (e.g., Guppy and Medaka) were trained for capturing only a few modifications (e.g., 5mC, 5hmc, 6mA). Consequently, the quality of ONT sequencing is unreliable when novel modifications extensively edit the genome.

This paper presents a set of unexpected low-quality genomes due to extensive novel modifications. We show that the removal of modifications by whole-genome amplification (WGA) significantly improves the quality of all genomes. A novel computational method is developed for correcting these modification errors without WGA.

Results

Unusual low-quality of ONT genomes due to extensive modifications

We sequenced 12 microbial strains of *Listeria monocytogenes* using Illumina and ONT (~200-990Mbp) (Figure 1(a), Supplementary Tables S1 and S2). The ONT reads were assembled into genomes with sequencing errors further polished by the-state-of-the-art tools (Supplementary Table S3, see Methods). The Illumina and ONT reads were hybrid assembled for evaluation purposes (Supplementary Table S4). When compared with the Illumina/ONT hybrid assemblies (Figure 1(b)), seven ONT-only genomes exhibited high quality (HQ) ranging from Q47 to Q60 (e.g., R19-2905 and R20-0088). However, five isolates (R20-0026, R20-0030, R20-0127, R20-0148, and R20-0150) showed unexpectedly low quality (LQ) varying from Q27 to Q34. The accuracy of these five LQ genomes remained unimproved after replicated ONT sequencing (data not shown). Further investigation of the five LQ genomes revealed excessive amounts of mismatch errors (1,228-5,780) compared with the seven HQ ones (3-36 mismatches) (Figure 1(c)). Homopolymer errors (i.e., indels) were not the source of inferior quality (7-306, Supplementary Table S5).

Manual inspection revealed that these mismatches were ONT basecalling errors uncorrected after genome polishing (Figure 1(d) and Supplementary Figure S1). As

mismatch errors in ONT are mainly due to epigenetic modifications, we computed the frequency of well-known methylation in these isolates (see Method and Supplementary Table S6). In terms of 5-methylcytosine (5mC), the numbers of modified loci in the five LQ genomes (~240-340k) were not significantly higher than those in the HQ ones (210-345k, P=0.89, Figure 1(e)). Similarly, the numbers of N^6-methyladenine (6mA) modifications also showed no significant difference between the LQ and HQ groups (98-218k v.s. 126-223k, P=0.34, Figure 1(f)). Because the numbers of mismatch errors in LQ genomes are significantly higher than those of HQ ones (P=0.005, Figure 1(g)), we suspected ONT basecalling algorithms failed to distinguish the novel modifications in the LQ isolates.

High-quality ONT genomes by WGA demodification

We removed the modifications in all microbial samples by WGA (Figure 2(a)), which randomly amplifies the genome fragments without retaining any epigenetic modification (see Methods). The WGA-demodified samples were sequenced by ONT, assembled into chromosomes, and compared with the Illumina/ONT hybrid genomes (Figure 2(a), Supplementary Tables S7 andS8). The five LQ genomes after WGA exhibited significantly higher quality than those without demodifications (e.g., Q27 to Q53 in R20-0026) (Figure 2(b), Supplementary Table S9). In particular, the amounts

of mismatch errors significantly reduced after demodification (e.g., 5,780 to 16 in R20-0026) (Figure 2(c)). Consequently, the unexpected low quality of ONT was due to excessive novel modifications untrained in their basecalling model. The demodification by WGA can produce high-quality ONT genomes without the need for Illumina short reads.

However, while WGA successfully erased these modifications, the sequencing cost increased by two factors. First, WGA required a higher sequencing depth (~100x) for assembling a complete genome when compared with ordinary ONT sequencing (~30x) (Figure 2(d) and Supplementary Figures S2-3). It was due to the uneven amplification of WGA, which led to non-uniform sequencing depth and a fragmented assembly at moderate coverage. Second, the WGA-demodified samples may reduce the ONT yields. We observed the numbers of available/active pores could sometimes decrease quickly (e.g., less than 100 pores after 12h) (Figure 2(e)), which was possibly owing to the hyperbranched structure unresolved after WGA. Consequently, the sequencing cost of WGA-demodified samples using ONT is much higher than ordinary sequencing.

in silico **correction of modification-mediated errors by Modpolish**

We developed a novel computational method (called Modpolish) for correcting these modification-mediated errors without WGA and prior knowledge of the modifications. Modpolish identifies and corrects the modification-mediated errors by investigating basecalling quality, basecalling consistency, and evolutionary conservation (Figure 3(a), see Method). Briefly, because the ONT signals are disturbed by modifications, the basecalling quality is usually low, and the basecalled nucleotides are often inconsistent at the modified loci. In conjunction with the conservation degree measured by closely-related genomes, only the modified loci with ultra-high conservation will be corrected by Modpolish, avoiding false corrections of strain variations.

We assessed the accuracy of Modpolish by comparing the quality of the ONT-only genomes (polished by Medaka/Homopolish) with those further polished by Modpolish. The results indicated that Modpolish significantly improved the genome quality of all LQ genomes (Figure 3(b), Supplementary Table S10). For instance, the quality of R20-0030 improved from Q34 to Q60, and the number of mismatches decreased from 1,228 to 33 (Figure (3(c)). We observed that the number of mismatches in R20-0026 reduced dramatically (i.e., from 5,780 to 143). However, the quality improvement (i.e., from Q27 to Q45) was slightly inferior to the others due to the 143

uncorrected mismatches. Note that no false corrections were made on the seven HQ genomes, implying the correction specificity of Modpolish is high.

The multilocus sequencing typing (MLST) indicated that R20-0026 belonged to the sequence type ST1081 and the remaining four LQ strains (i.e., R20-0030, R20-0127, R20-0148, R20-150) were ST87. Hence, we investigated whether an identical modification system extensively edited the genomes of these two lineages. Sequence analysis of the modified loci revealed that the modifications of ST1081 were on the GCTGG motif (Figure 3(d)). On the other hand, the modification sites of all ST87 strains centered on the GCAGC motif (Figure 3(e)). Therefore, two modification systems seem specific to each of the two lineages. In addition, while both motifs are not palindromic, their reverse complements (i.e., CCAGC, GCTGC) were also hotspots of modifications (Supplementary Figure S4). Because the mismatches frequently appeared on both strands at the same loci (Supplementary Figure S5), the unknown modification may symmetrically edit both strands. Although their underlying mechanisms remained unclear, the two systems extensively modified the genomes at specific motifs with high conservation within each lineage, leading to excessive amounts of basecalling yet correctable errors.

Comparison of phylogeny reliability under extensive modifications

Because sequencing errors alter the genetic distances between strains, we assessed the reliability of phylogeny using ONT with or without modification-error removal. We reconstructed the core-genome MLST (cgMLST) phylogeny of the five LQ strains sequenced and assembled by four methods: ONT-only sequencing, WGA-demodified ONT, ONT with Modpolish, and hybrid ONT/Illumina sequencing (Figure 4(a)). The WGA-demodified genomes perfectly clustered with the ONT/Illumina hybrid for each strain in both clades (ST87 and ST1081). The ONT genomes corrected by Modpolish clustered with the hybrid and WGA-demodified genomes in both clades. But the genetic distance slightly deviated from them, especially in the ST1081 clade. The ONT-only genomes were phylogenetic distant from the others due to excessive amounts of modification-mediated errors.

When comparing each method in the seven HQ isolates, ONT with WGA was slightly worse than the original ONT and Modpolish in six strains (e.g., 47 v.s. 6 mismatches in R20-0088) (Figure 4(b)), except for the R19-2905 isolate (i.e., 12 v.s. 36 mismatches). These mismatches slightly increased the genetic distance to the others (Supplementary Figure S6). Nevertheless, phylogenetic analysis indicated that the genomes of all methods were perfectly clustered for each HQ strain (Supplementary Figure S7), implying the number of mismatches is less than that of strain variations. Consequently, all methods can produce reliable phylogeny when free of novel

158 modifications. But when new modifications extensively edit the genome, only ONT
159 with WGA or Modpolish can provide sufficient typing accuracy without additional
160 Illumina sequencing.

161 Discussion

162 This paper presented a set of unexpected low-quality ONT genomes due to extensive
163 modifications untrained in the basecalling models. Demodification by WGA
164 successfully improved the genome quality while losing the epigenome. The *in silico*
165 method, Modpolish, removed these modification-mediated errors without prior
166 knowledge of modifications and uncovered the modified motifs while retaining the
167 epigenome. When unknown modifications extensively shaped the genome, ONT with
168 WGA or Modpolish produced nearly identical cgMLST profiles as hybrid
169 ONT/Illumina did. On the other hand, the phylogeny of ONT-only genomes was
170 disturbed by modification-mediated errors. Therefore, ONT with WGA or Modpolish
171 is robust to modification-mediated errors without the need for additional Illumina
172 sequencing.

173 Quality reduction of ONT on novel modifications

174 Existing ONT basecalling algorithms only capture a few methylations (e.g., 5mC,
175 5hmc, 6mA) and ignore the vast amount of other modifications. Theoretically, species-

176 specific modifications can be distinguished by training bespoke models for one
177 organism (e.g., Taiyaki). But practically, it is infeasible to train models for hundreds of
178 modifications in the biosphere. Especially in metagenomic sequencing, the usage of
179 any particular model is biased against other modifications. For instance, a meta-
180 epigenomic sequencing uncovered 22 methylation systems in a single microbial
181 community [10]. Hence, if WGA is not an option, modification-mediated errors are
182 better removed at the postassembly stage as each assembled contig can be polished
183 independently.

184 **Limitations of ONT with WGA**

185 The cost of WGA ONT is higher than ordinary sequencing due to several side
186 effects of the amplification protocol. First, the amplified DNA may still contain a
187 hyperbranched structure after Flap endonuclease (e.g., T7) cleavage. The
188 hyperbranched DNA may block the pores during ONT sequencing and reduce the
189 available pores and yields. In addition, the usage of endonuclease cleavage also
190 decreased the read lengths. In conjunction with the uneven amplification, WGA
191 requires higher coverage (~100x) for reconstructing a complete genome than ordinary
192 ONT sequencing (~30x). Notably, the usage of WGA discards the entire methylome.
193 The loss of modifications would prohibit any epigenetic study using ONT.

194 **Limitations of ONT with Modpolish**

While Modpolish eliminated most modification-mediated errors, the correction power was lower in the ST1081 isolate. The lack of ST1081 genomes in NCBI RefSeq decreased the sensitivity of Modpolish. As the algorithm only corrects the loci of high evolutionary conservation, a sufficient number of closely-related genomes is necessary. Therefore, Modpolish is more suitable for common instead of rare lineages.

Nevertheless, Modpolish retains all modifications after ONT sequencing while WGA loses the epigenome. Epigenetic methylation has been thought to contribute to the rapid adaptation of resistance [11]. For instance, phase-variable adenine DNA methyltransferases (e.g., ModA11 and ModA12) increase susceptibility to cloxacillin and ciprofloxacin in *Neisseria meningitidis* [12]. The resistance due to overexpression of efflux pumps (e.g., sugE) has been linked to the lack of the Dcm-mediated 5mC silencing [13]. Therefore, Modpolish should be used when the epigenome is the focus of the study.

Functional implications of the two modification systems

We discovered two pentanucleotide motifs, GCTGG (CCAGC) and GCAGC (GCTGC), specific to each of the two lineages (ST1081 and ST87). In ST1081, the GCTGG (CCAGC) motif is part of *chi* sites, hotspots of homologous recombination mediated by the RecBC enzyme [14, 15]. As phages cut by restriction enzymes are

213 further degraded by RecBC [16], modifications on the GCTGG motif may be part of
214 the defending system of ST1081, which protect itself against the RecBC cleavage.

215 In ST87 strains, the GCAGC/GCTGC (i.e., GCWGC) motif was the known target
216 of the orphan methyltransferase M.BatI [17]. M.BatI produced fully-methylation on 5′-
217 GCWGC-3′ and hemimethylation on 5′-GCSGC-3′. Reinvestigation of the modified
218 sites in ST87 showed the existence of both GCWGC and GCSGC (Supplementary
219 Figure S4). Interestingly, M.BatI increased toxicity when expressed in *E coli* in their
220 study, which was concordant with the elevated virulence of ST87 strains.

221 Hence, the two lineages possessed two distinct modification systems for defensive
222 purposes and increasing virulence. Although further investigations are required to
223 assess their biological function, modifications that have acquired regulatory effects in
224 bacteria are usually conservative within a clade [18]. Consequently, our *in silico*
225 algorithm successfully utilize the conservation for correcting modification errors.

226 Conclusion

227 This paper reported a set of unexpectedly low-quality genomes due to novel
228 modifications untrained in the ONT basecalling model. The increasing number of new
229 modifications found by single-molecular sequencing or high-resolution mass
230 spectrometry will unavoidably reduce the ONT accuracy. New ONT flowcells,

231 sequencing kits, and basecalling algorithms aim to resolve the homopolymer issue but
232 not modification-mediated errors. Our study showed that these modification-mediated
233 errors can be effectively corrected by preassembly amplification or postassembly
234 polishing without additional short-read sequencing, producing high-quality genomes
235 reliable for downstream analysis.

236 Materials and Methods

237 **Bacterial isolates.** Twelve *Listeria monocytogenes* isolates used in this study were
238 obtained from hospitals recovered from listeriosis patients in Taiwan between 2019 and
239 2020. The isolates were submitted to the Taiwan Centers for Disease Control for further
240 identification and genotyping. The isolates belonged to serogroups IIa (5 isolates), IIb
241 (6 isolates), and IVb (1 isolate) and sequence type (ST) 1, ST5 (2 isolates), ST87 (4
242 isolates), ST101, ST155, ST378, ST1081, and ST1532.

243 **Whole genome sequencing.** WGS of bacterial isolates was conducted in the Central
244 Region laboratory of Taiwan CDC using the Illumina MiSeq sequencing platform
245 (Illumina Co., USA) and the Nanopore sequencing platform (Oxford Nanopore
246 Technologies, Inc., UK). DNA of bacterial isolates was extracted using the Qiagen
247 DNeasy blood and tissue kit (Qiagen Co., Germany). Illumina DNA library
248 construction was performed using the Illumina DNA Prep, (M) Tagmentation system

(Illumina Co.), and sequencing was run with the MiSeq reagent kit version 3 (2X 300 bp), manipulated according to the manufacturer's instructions. Nanopore DNA library construction was performed using the Rapid Barcoding Kit and sequencing was run using the MinION device and R9.4 chemistry.

Removal of modifications of nucleotides using whole-genome amplification. DNA Bacterial Genomic DNA was amplified using the REPLI-g Advanced DNA Single Cell Kit (Qiagen, Hilden, Germany), manipulated according to the manufacturer's instructions. The amplified DNA was purified using the KAPA HyperPure Beads (Roche, Basel, Switzerland) before subjecting to Nanopore sequencing.

Assembly of sequence reads. Illumina sequence reads for each isolate were assembled using the SPAdes assembler version 3.12.0 (http://cab.spbu .ru/software/spades/) [19]; both Illumina sequence reads and Nanopore sequence reads for each isolates together were assembled to complete the full genomic sequences using the Unicycler Assembler [20]. The Nanopore reads for each isolate (in FAST5 file) were subjected to basecalling using the Guppy basecaller (https://nanoporetech.com/). In the ONT-only assembly, the sequences (in FASTQ file) were assembled using Flye (https://github.com/fenderglass/Flye)[21], then polished using the Racon (https://github.com/lbcb-sci/racon) [4], the Medaka

(https://github.com/nanoporetech/medaka), and the Homopolish (https://github.com/ythuang0522/homopolish) [5]. Methylations (i.e., 5mC, 6mA) in the ONT-only genomes were called by Megalodon (https://github.com/nanoporetech/megalodon). The Integrative Genome Viewer (IGV) was used for visualizing the ONT modification errors [22]. The genome quality was assessed by fastmer (https://github.com/jts/assembly_accuracy).

cgMLST analysis. Assembled Illumina contigs, assembled and polished Nanopore contigs, and assembled complete genomic sequence (obtained from assembling Illumina sequences and Nanopore sequences) for each isolate were used to generate core-gene multilocus sequence typing (cgMLST) profiles (based on 2,172 core genes) using an in-house-developed cgMLST profiling tool available on the cgMLST@Taiwan website (http://rdvd.cdc.gov.tw/cgMLST). Phylogenetic trees were constructed with cgMLST profiles using the minimum spanning tree algorithm and the tool provided on the cgMLST@Taiwan website.

Overview of Modpolish. The proposed computational method, Modpolish, aims to remove modification-mediated errors by investigating the inconsistency of basecalled nucleotides, qualities of basecalled alleles, and evolutionary conservation at the modified loci. Modpolish is an extension from Homopolish, a polishing algorithm

285 designed for correcting ONT homopolymer errors [5]. Figure 5 depicts the workflow
286 of Modpolish. The closely-related genomes are first identified by screening against a
287 compressed representation of microbial genomes. The genome sequences are then
288 retrieved on the fly and compared with the draft genome. We only retain closely-related
289 genomes of high nucleotide and structural similarity. Given the alignment matrix of
290 reads, qualities, and homologs, Modpolish identifies potential-modified loci of
291 inconsistent basecalling and low quality and only corrects the mismatch errors highly
292 conserved in homologs. The details are described in the following sections.

293 **Collection of homologs by nucleotide and structural similarity.** The draft genome
294 (to be polished) is scanned against the virus, bacteria, or fungus genomes compressed
295 by Mash as (MinHash) sketches, which is a reduced representation of all microbial
296 genomes in NCBI RefSeq [23]. Subsequently, top *t* (default 20) closely related genomes
297 will be retrieved on the fly. Mash estimated the Jaccard similarity between the draft and
298 related genomes over a subset of *k*-mers. Though very fast, this method has low
299 resolution at distinguishing closely-related genomes because the small subset of *k*-mers
300 may not capture the few strain variations. Consequently, the genome similarity has to
301 be re-estimated using more sensitive approaches.

302 Subsequently, each downloaded genome is compared against the draft genome
303 using FastANI for computing the average nucleotide identity (ANI) at a higher

304 resolution than Mash [24]. FastANI chops the two genomes into pieces and aligns them
305 against each other for speedup. However, it only considers the aligned segments for
306 ANI estimation and ignores the unaligned portions (Supplementary Figure S8(a)). The
307 unaligned segments imply these two genomes differ by structural variations (i.e.,
308 vertically-/horizontally-transferred genes). As small and large variants are both genetic
309 footprints of strain variations during evolution, Modpolish also computes the structural
310 similarity (average-structural identity, ASI), defined as the percentage of aligned
311 segments. We only retain the related genomes with sufficient ANI (>99%) and ASI
312 (>90%) for subsequent error correction. These emprical cutoffs were determined by
313 investigating the distributions of ANI and ASI in real microbial genomes.

314 **Correction of modification-mediated errors by reads and homologs.** These closely-
315 related genomes with sufficient ANI and ASI are aligned against the draft genome via
316 minimap2 (with asm5 option) [25]. The raw ONT reads are also mapped against the
317 draft genome by minimap2 (with map-ont option). We extract the basecalled
318 nucleotides, basecalling qualities, and homologous alleles from the alignments. The
319 aligned homologs, reads, and qualities are converted into a table of several summary
320 statistics (Supplementary Figure S8(b)).

321 The summary statistics include the allele counts of A, T, C, and G separately for
322 homologs and ONT reads, ignoring the insertion and deletion gaps. We identify the

potentially modified sites according to the allele discordancy and average quality (see also Supplementary Figure S8(b)). The allele discordancy is the frequency of alternative alleles (i.e., non-major ones) at one locus. The average quality was computed by averaging the qscores from all read bases at the same locus. A potentially-modified locus is defined as the allele discordancy greater than 5% and the average quality score below 15, which were empirically observed from the modification-mediated errors.

For each potentially-modified locus, if all the homologous alleles are 100% conserved, we will correct the erroneous nucleotide into the alternative allele concordant with the homologs. These stringent criteria aimed for specificity instead of sensitivity, ensuring little or no false corrections would be made. We also implemented a motif-aware mode when the modification system is known in advance. If the user specifies a known modification motif (e.g., CCGAC), the program will additionally correct loci according to the provided pattern by lowering the homologous conservation ratio from 100% to 80%.

Data and software availability

The genomes sequenced and assembled by Illumina, ONT, and WGA ONT are deposited in the NCBI with BioProject (xxxxxx). Modpolish was implemented as a

subcommand in the Homopolish package, which is freely available at (https://github.com/ythuang0522/homopolish/).

Conflict of interests

The authors declare no conflict of interests.

Reference

1. Gorzynski JE, Goenka SD, Shafin K, Jensen TD, Fisk DG, Grove ME, Spiteri E, Pesout T, Monlong J, Baid G *et al*: **Ultrarapid Nanopore Genome Sequencing in a Critical Care Setting**. *New England Journal of Medicine* 2022, **386**(7):700-702.
2. Gu W, Deng X, Lee M, Sucu YD, Arevalo S, Stryke D, Federman S, Gopez A, Reyes K, Zorn K *et al*: **Rapid pathogen detection by metagenomic next-generation sequencing of infected body fluids**. *Nat Med* 2021, **27**(1):115-124.
3. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, Albertsen M: **Oxford Nanopore R10.4 long-read sequencing enables near-perfect bacterial genomes from pure cultures and metagenomes without short-read or reference polishing**. *bioRxiv* 2021:2021.2010.2027.466057.
4. Vaser R, Sović I, Nagarajan N, Šikić M: **Fast and accurate de novo genome assembly from long uncorrected reads**. *Genome Res* 2017, **27**(5):737-746.
5. Huang Y-T, Liu P-Y, Shih P-W: **Homopolish: a method for the removal of systematic errors in nanopore sequencing by homologous polishing**. *Genome biology* 2021, **22**(1):1-17.
6. Zhang P, Jiang D, Wang Y, Yao X, Luo Y, Yang Z: **Comparison of De Novo Assembly Strategies for Bacterial Genomes**. *International Journal of Molecular Sciences* 2021, **22**(14):7668.

7. Wang S, Xie H, Mao F, Wang H, Wang S, Chen Z, Zhang Y, Xu Z, Xing J, Cui Z *et al*: **N4-acetyldeoxycytosine DNA modification marks euchromatin regions in Arabidopsis thaliana**. *Genome Biology* 2022, **23**(1):5.
8. Xu L, Seki M: **Recent advances in the detection of base modifications using the Nanopore sequencer**. *Journal of Human Genetics* 2020, **65**(1):25-33.
9. Schatz MC: **Nanopore sequencing meets epigenetics**. *Nature Methods* 2017, **14**(4):347-348.
10. Hiraoka S, Okazaki Y, Anda M, Toyoda A, Nakano S-i, Iwasaki W: **Metaepigenomic analysis reveals the unexplored diversity of DNA methylation in an environmental prokaryotic community**. *Nature Communications* 2019, **10**(1):159.
11. Ghosh D, Veeraraghavan B, Elangovan R, Vivekanandan P: **Antibiotic Resistance and Epigenetics: More to It than Meets the Eye**. *Antimicrobial Agents and Chemotherapy* 2020, **64**(2):e02225-02219.
12. Jen FE-C, Seib KL, Jennings MP: **Phasevarions Mediate Epigenetic Regulation of Antimicrobial Susceptibility in Neisseria meningitidis**. *Antimicrobial Agents and Chemotherapy* 2014, **58**(7):4219-4221.
13. He G-X, Zhang C, Crow RR, Thorpe C, Chen H, Kumar S, Tsuchiya T, Varela MF: **SugE, a New Member of the SMR Family of Transporters, Contributes to Antimicrobial Resistance in Enterobacter cloacae**. *Antimicrobial Agents and Chemotherapy* 2011, **55**(8):3954-3957.
14. Aoki H, Kajino K, Arakawa Y, Hino O: **Molecular cloning of a rat chromosome putative recombinogenic sequence homologous to the hepatitis B virus encapsidation signal**. *Proceedings of the National Academy of Sciences* 1996, **93**(14):7300-7304.
15. Ponticelli AS, Schultz DW, Taylor AF, Smith GR: **Chi-dependent DNA strand cleavage by RecBC enzyme**. *Cell* 1985, **41**(1):145-151.
16. Simmon VF, Lederberg S: **Degradation of bacteriophage lambda deoxyribonucleic acid after restriction by Escherichia coli K-12**. *J Bacteriol* 1972, **112**(1):161-169.
17. Furuta Y, Miura F, Ichise T, Nakayama SMM, Ikenaka Y, Zorigt T, Tsujinouchi M, Ishizuka M, Ito T, Higashi H: **A GCDGC-specific DNA (cytosine-5) methyltransferase that methylates the GCWGC sequence on both strands and the GCSGC sequence on one strand**. *PLOS ONE* 2022, **17**(3):e0265225.

18. Seshasayee ASN, Singh P, Krishna S: **Context-dependent conservation of DNA methyltransferases in bacteria**. *Nucleic acids research* 2012, **40**(15):7066-7073.
19. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD *et al*: **SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing**. *J Comput Biol* 2012, **19**(5):455-477.
20. Wick RR, Judd LM, Gorrie CL, Holt KE: **Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads**. *PLoS Comput Biol* 2017, **13**(6):e1005595.
21. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL *et al*: **metaFlye: scalable long-read metagenome assembly using repeat graphs**. *Nature Methods* 2020, **17**(11):1103-1110.
22. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer**. *Nature biotechnology* 2011, **29**(1):24-26.
23. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM: **Mash: fast genome and metagenome distance estimation using MinHash**. *Genome Biology* 2016, **17**(1):132.
24. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S: **High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries**. *Nature Communications* 2018, **9**(1):5114.
25. Li H: **Minimap2: pairwise alignment for nucleotide sequences**. *Bioinformatics* 2018, **34**(18):3094-3100.

Figure Legends

Figure 1. Quality comparison of 12 microbial strains using ONT-only and ONT/Illumina hybrid sequencing. (a) Workflow of ONT-only and ONT/Illumina hybrid assembly; (b) Q scores; (c) number of mismatches; (d) comparison of ONT and Illumina reads by IGV; (e) numbers of 5mC, 6mA, and mismatches between HQ/LQ strains.

Figure 2. Quality improvement of ONT by WGA demodification. (a) Worflow of WGA-demodified ONT; (b) Q scores of the WGA-demodified and ONT-only genomes; (c) numbers of mismatches of the WGA-demodified and ONT-only genomes; (d) WGA genome quality with respect to sequencing depth; (e) numbers of active/available pores during WGA-demodified and ordinary ONT sequencing.

Figure 3. Correction of modification-mediated errors by Modpolish. (a) Workflow of Modpolish; (b) Q scores before and after Modpolish; (c) numbers of mismatchs before and after Modpolish; (d) the sequence motif of modification on ST1081; (e) the sequence motif of modifications on ST87.

Figure 4. Comparison of phylogeny reliability of four methods. (a) The cgMLST phylogeny of the five LQ strains sequenced and assembled by four methods: ONT-only sequencing (ONT), WGA-demodified ONT (ONT_WGA), ONT with Modpolish

(ONT_Modpolish), and hybrid ONT/Illumina sequencing (Hybrid_WGS); (b) the cgMLST distances of ONT, ONT_WGA, and ONT_Modpolish to the Hybrid_WGS assembled genomes.

Figure 5. Illustration of Modpolish workflow. A set of closely-related genomes are first retrieved by screening the compressed sketches of RefSeq genomes. We retain the genomes with sufficient nucleotide and structural similarity. The selected genomes and ONT reads are aligned onto the draft genome, generating a pileup matrix of homologs, reads, and qualites. Modpolish only corrects modification-mediated errors with inconsistent read alleles, low quality, and high conservation in homologs.

Supplementary Information

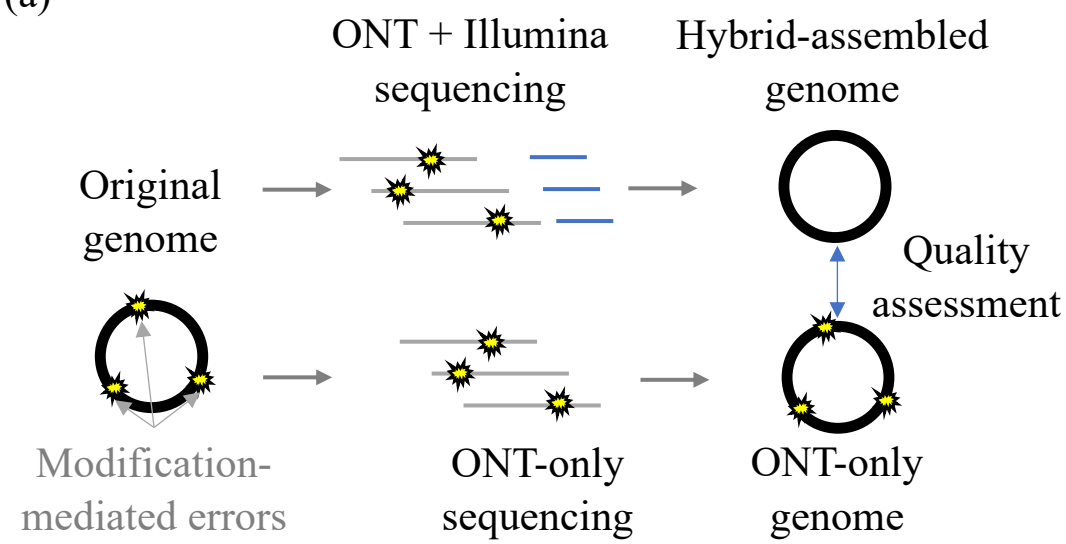
Additional file 1

Additional file 1 includes Supplementary Figures S1-8.

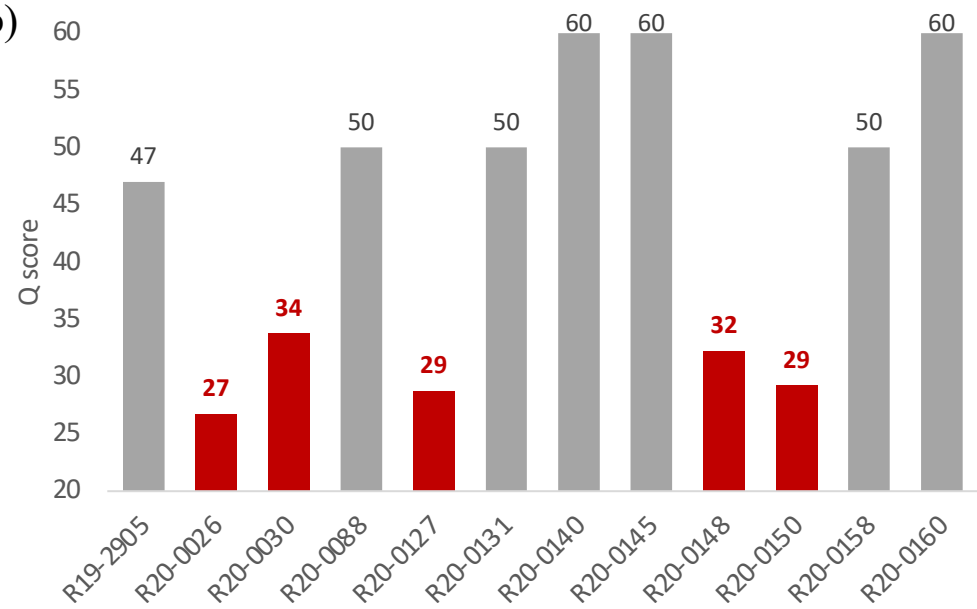
Additional file 2

Additional file 2 includes Supplementary Tables S1-10.

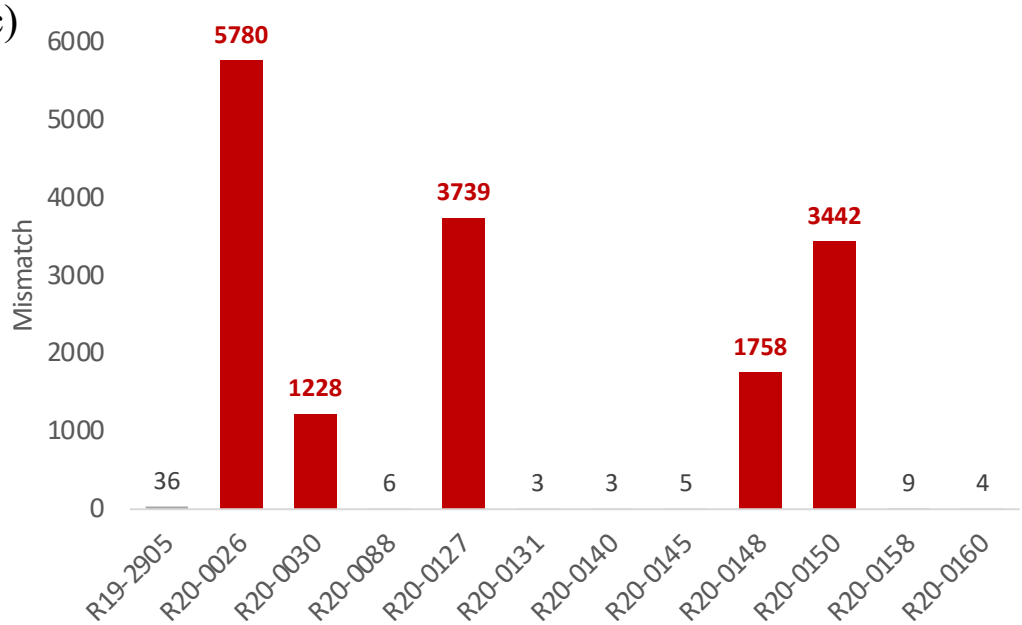
(a)


ONT + Illumina sequencing
Hybrid-assembled genome
Original genome
Quality assessment
Modification-mediated errors
ONT-only sequencing
ONT-only genome

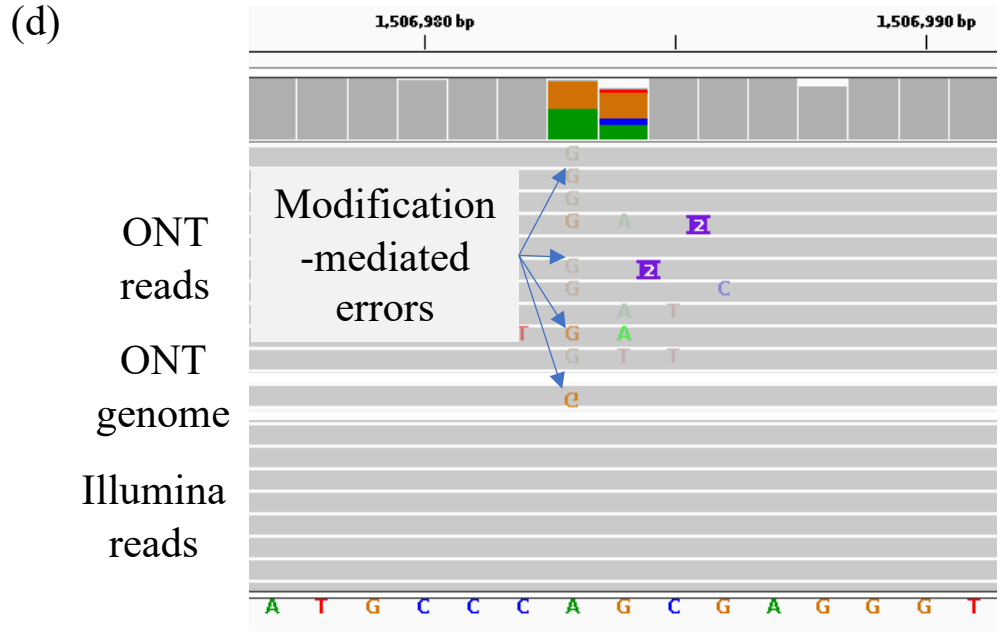
(b)


Q score
60
55
50
45
40
35
30
25
20
47
27
34
50
29
50
60
60
32
29
50
60
R19-2905
R20-0026
R20-0030
R20-0088
R20-0127
R20-0131
R20-0140
R20-0145
R20-0148
R20-0150
R20-0158
R20-0160

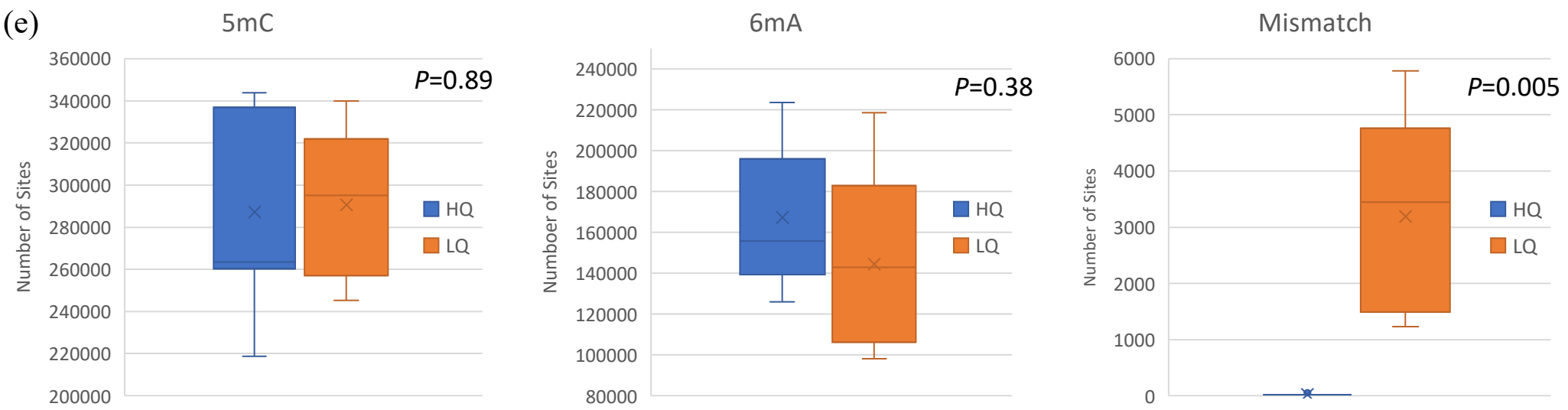
(c)

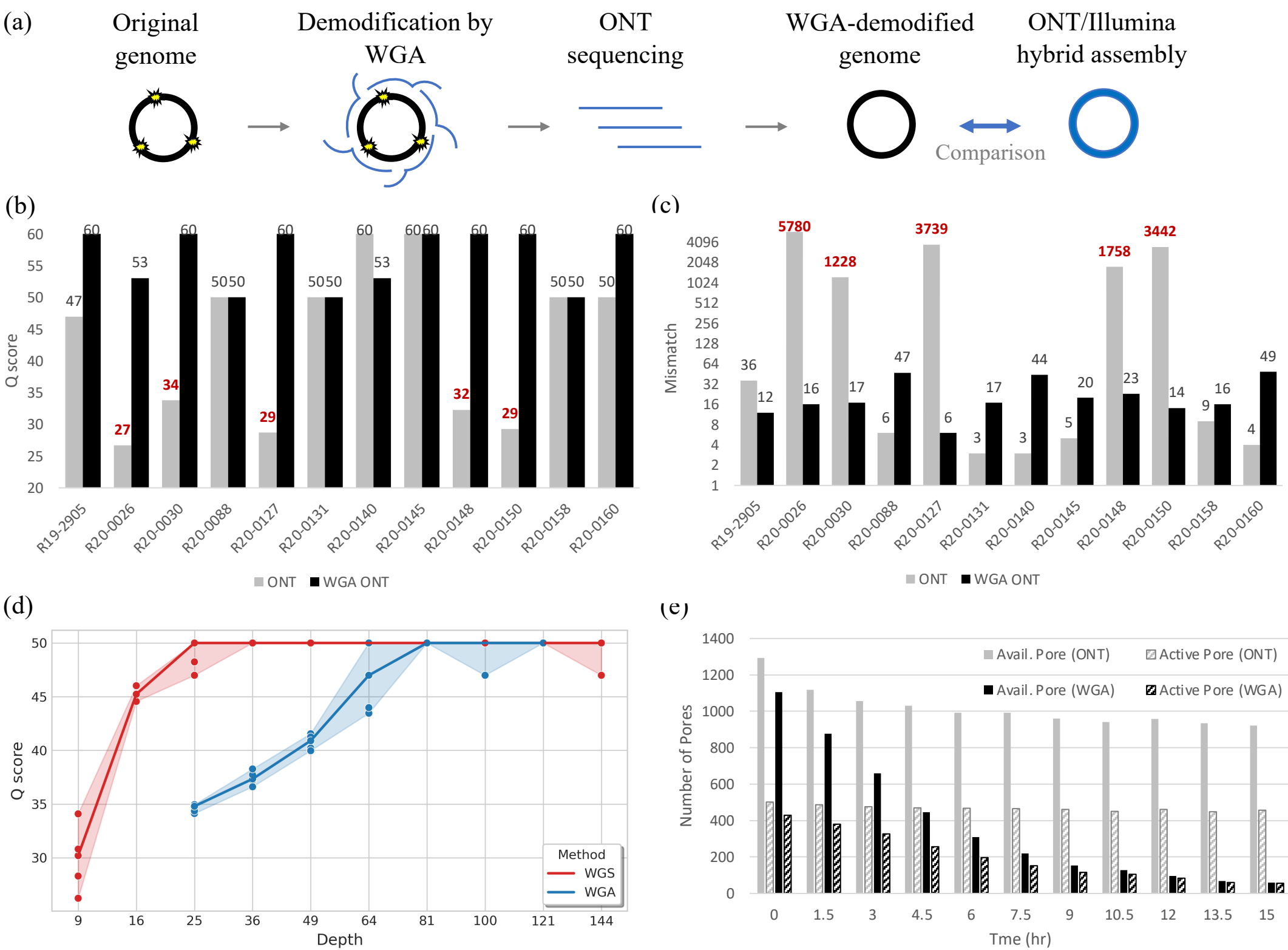

Mismatch
6000
5000
4000
3000
2000
1000
0
36
5780
1228
6
3739
3
3
5
1758
3442
9
4
R19-2905
R20-0026
R20-0030
R20-0088
R20-0127
R20-0131
R20-0140
R20-0145
R20-0148
R20-0150
R20-0158
R20-0160

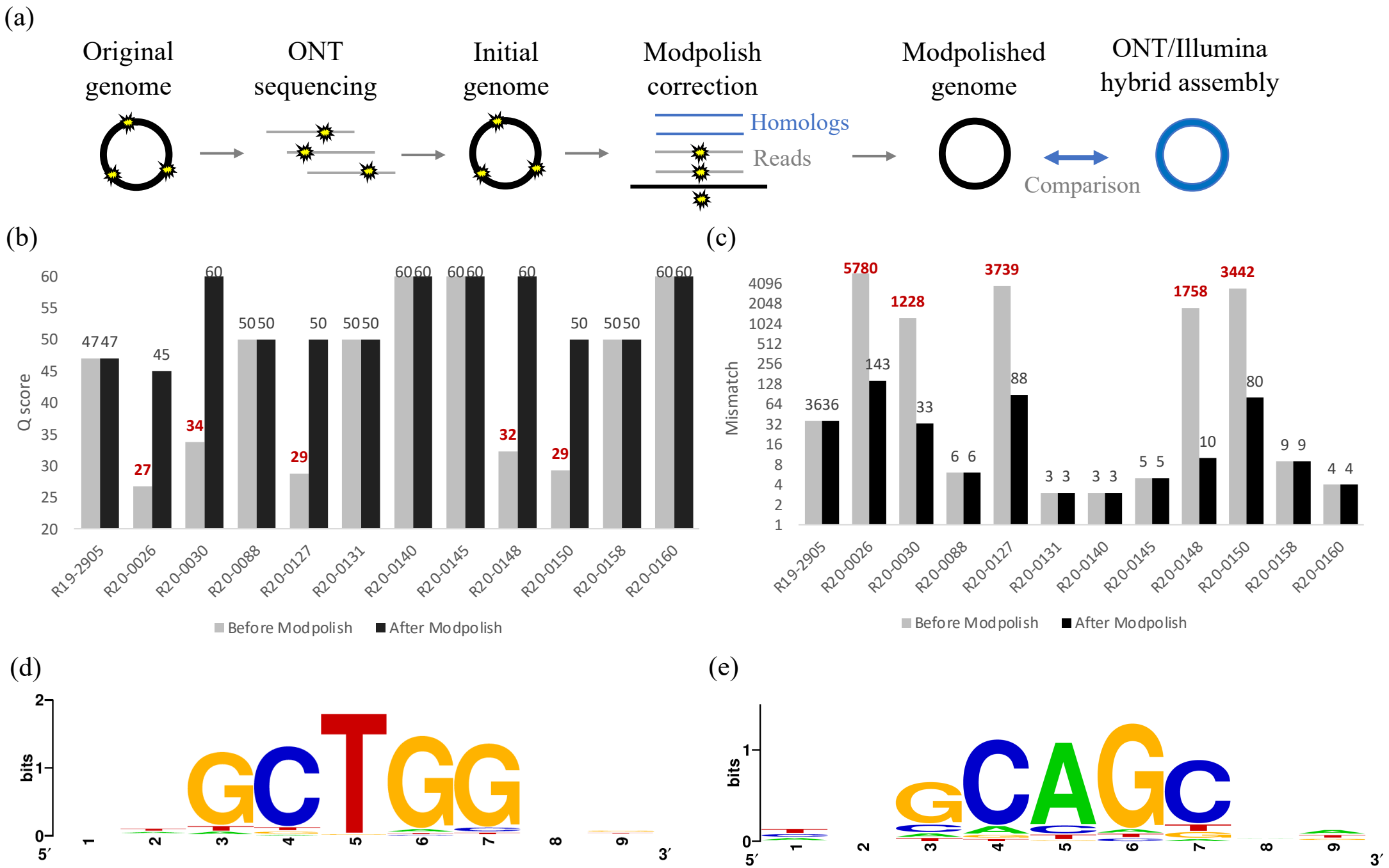
(d)

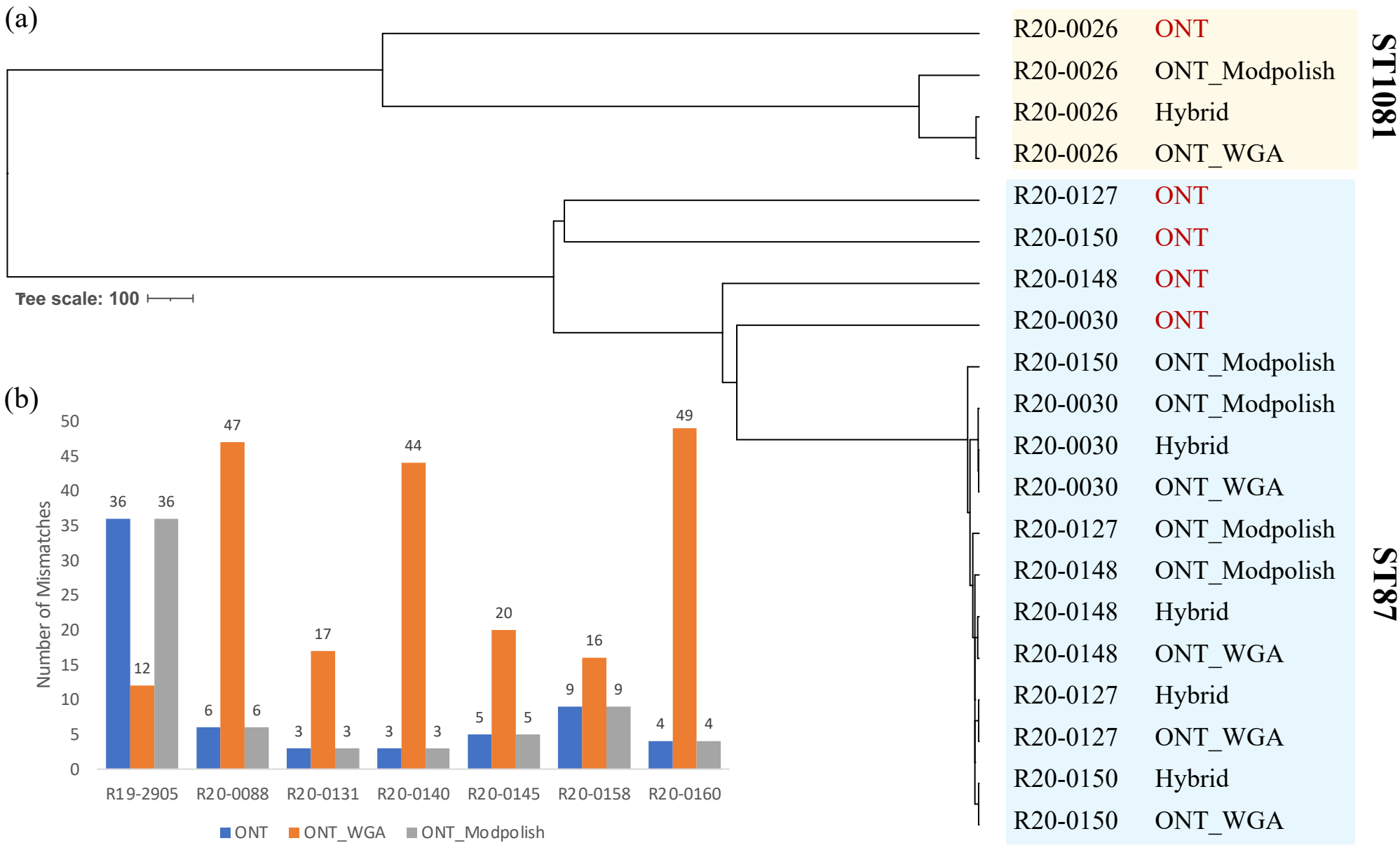

1,506,980 bp
1,506,990 bp
ONT reads
ONT genome
Illumina reads
Modification -mediated errors
A T G C C C A G C G A G G G T

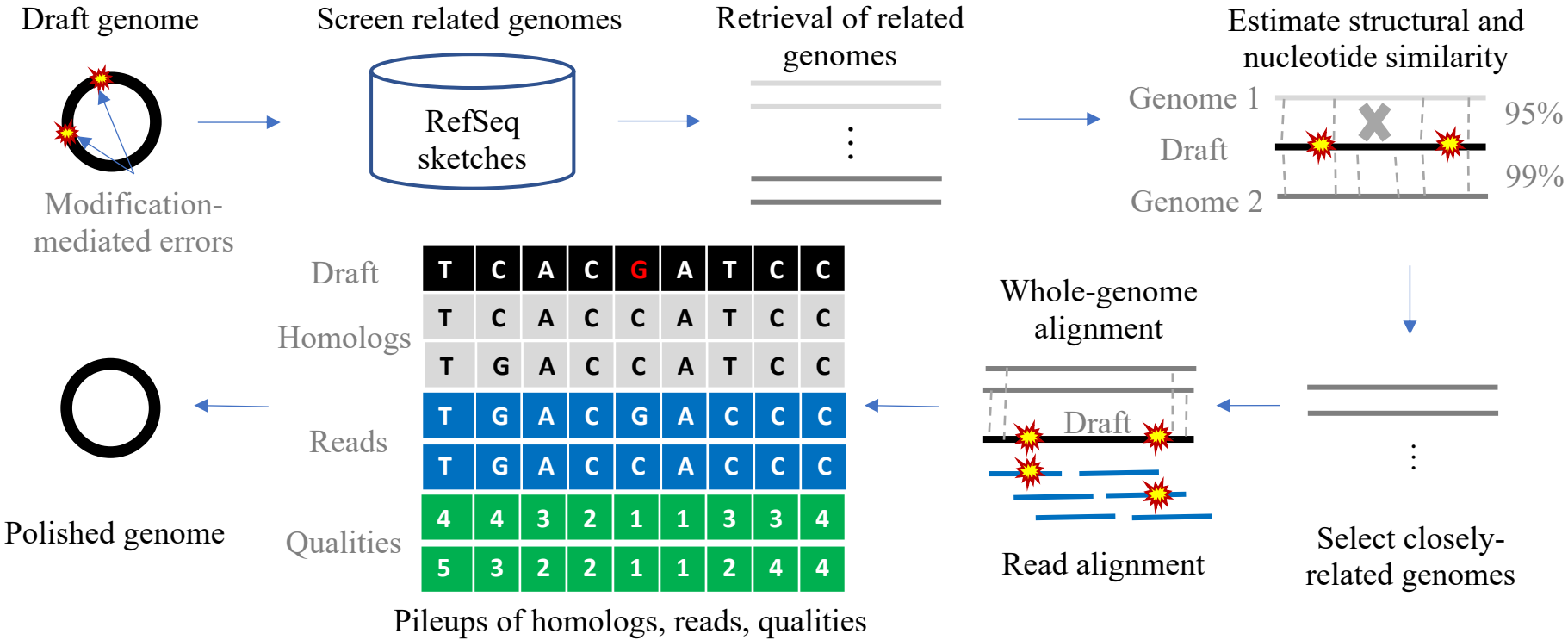
(e)


5mC
P=0.89
Number of Sites
360000
340000
320000
300000
280000
260000
240000
220000
200000
HQ
LQ
6mA
P=0.38
Numboer of Sites
240000
220000
200000
180000
160000
140000
120000
100000
80000
HQ
LQ
Mismatch
P=0.005
Number of Sites
6000
5000
4000
3000
2000
1000
0
HQ
LQ


(a)
Original genome
Demodification by WGA
ONT sequencing
WGA-demodified genome
ONT/Illumina hybrid assembly
Comparison
(b)
Q score
R19-2905
R20-0026
R20-0030
R20-0088
R20-0127
R20-0131
R20-0140
R20-0145
R20-0148
R20-0150
R20-0158
R20-0160
ONT
WGA ONT
(c)
Mismatch
ONT
WGA ONT
(d)
Q score
Depth
Method
WGS
WGA
(e)
Number of Pores
Avail. Pore (ONT)
Active Pore (ONT)
Avail. Pore (WGA)
Active Pore (WGA)
Tme (hr)


(a)
Original genome
ONT sequencing
Initial genome
Modpolish correction
Homologs
Reads
Modpolished genome
ONT/Illumina hybrid assembly
Comparison
(b)
Q score
R19-2905
R20-0026
R20-0030
R20-0088
R20-0127
R20-0131
R20-0140
R20-0145
R20-0148
R20-0150
R20-0158
R20-0160
Before Modpolish
After Modpolish
(c)
Mismatch
Before Modpolish
After Modpolish
(d)
bits
(e)
bits


(a)
Tee scale: 100
R20-0026 ONT
R20-0026 ONT_Modpolish
R20-0026 Hybrid
R20-0026 ONT_WGA
R20-0127 ONT
R20-0150 ONT
R20-0148 ONT
R20-0030 ONT
R20-0150 ONT_Modpolish
R20-0030 ONT_Modpolish
R20-0030 Hybrid
R20-0030 ONT_WGA
R20-0127 ONT_Modpolish
R20-0148 ONT_Modpolish
R20-0148 Hybrid
R20-0148 ONT_WGA
R20-0127 Hybrid
R20-0127 ONT_WGA
R20-0150 Hybrid
R20-0150 ONT_WGA
ST1081
ST87
(b)
Number of Mismatches
50
45
40
35
30
25
20
15
10
5
0
36
12
36
6
47
6
3
17
3
3
44
3
5
20
5
9
16
9
4
49
4
R19-2905
R20-0088
R20-0131
R20-0140
R20-0145
R20-0158
R20-0160
ONT
ONT_WGA
ONT_Modpolish


Draft genome
Modification-mediated errors
Screen related genomes
RefSeq sketches
Retrieval of related genomes
Estimate structural and nucleotide similarity
Genome 1
Draft
Genome 2
95%
99%
Select closely-related genomes
Whole-genome alignment
Draft
Read alignment
Draft
Homologs
Reads
Qualities
T C A C G A T C C
T C A C C A T C C
T G A C C A T C C
T G A C G A C C C
T G A C C A C C C
4 4 3 2 1 1 3 3 4
5 3 2 2 1 1 2 4 4
Pileups of homologs, reads, qualities
Polished genome