DNA Research, 2022, **29**, 1–10
https://doi.org/10.1093/dnares/dsac039
Advance access publication 29 October 2022
Research article

OXFORD

Research article

Assembly-free discovery of human novel sequences using long reads

Qiuhui Li†, Bin Yan†, Tak-Wah Lam, and Ruibang Luo*,

Department of Computer Science, The University of Hong Kong, Hong Kong, China

*To whom correspondence should be addressed. Tel. +852-2859-2186. Fax. +852-2559-8447.Email: rbluo@cs.hku.hk
†These authors contributed equally to this work.

Abstract

DNA sequences that are absent in the human reference genome are classified as novel sequences. The discovery of these missed sequences is crucial for exploring the genomic diversity of populations and understanding the genetic basis of human diseases. However, various DNA lengths of reads generated from different sequencing technologies can significantly affect the results of novel sequences. In this work, we designed an assembly-free novel sequence (AF-NS) approach to identify novel sequences from Oxford Nanopore Technology long reads. Among the newly detected sequences using AF-NS, more than 95% were omitted from those using long-read assemblers and 85% were not present in short reads of Illumina. We identified the common novel sequences among all the samples and revealed their association with the binding motifs of transcription factors. Regarding the placements of the novel sequences, we found about 70% enriched in repeat regions and generated 430 for one specific subpopulation that might be related to their evolution. Our study demonstrates the advance of the assembly-free approach to capture more novel sequences over other assembler based methods. Combining the long-read data with powerful analytical methods can be a robust way to improve the completeness of novel sequences.

Key words: long reads, novel sequences, assembly-free approach, human references

1. Introduction

Building a complete reference genome in humans is fundamental for decoding genetic variation and the associations with human diseases.[1] However, the current reference genome was derived from few individuals, resulting in an underrepresentation of the human population.[2,3] To increase its diversity, human genome projects have assembled genomes across and within subpopulations.[4,5] In general, DNA sequences missed in the human reference genomes are called novel sequences. These missed genomic sequences may contain thousands of unknown genetic variants implicated in biological functions. Therefore, identifying novel sequences can enrich the human reference genome and facilitate genome-based research.

Discovering novel sequences is dependent largely on the development of sequencing technologies. Initially, novel sequences were defined using fosmid end sequence pairs[6] and the entire fosmid clone.[7] However, the high expense of capillary sequencing made the creation of large-scale genomes impractical. In contrast, next-generation sequencing (NGS, also known as short-read sequencing) technologies enable the sequencing of thousands of samples with higher efficiency and at lower cost, resulting in more complete genomic information. Recently, an African pan-genome built with 910 African descents discovered 296 Mbp of novel sequences.[8] We built a Chinese pan-genome using 486 Chinese genomes and got 276 Mbp of novel sequences.[9] These studies have increased the human genome diversity and the related functional implications of novel sequences.

However, there is a major limitation for NGS short-reads <300 bases long that are too short to detect more variants, especially >70% of human genome structural variations.[10] To overcome this shortcoming, long-read sequencing has emerged, including Oxford Nanopore Technology (ONT), which generates long (10–100 kbp) and ultra-long (>100 kbp) DNA reads.[11] The average lengths of over 10,000 bp are helpful for analyzing structural variations and *de novo* assembly.[12–14] The first gapless human reference was created using PacBio HiFi and ONT, which incorporated 200 Mbp sequences absent in GRCh38.p13.[3] Currently, the detection of novel sequences relies mainly on genome assembling from long-read data. For example, 12.8 Mbp of novel sequences from a Chinese assembly[15] and over 10 Mbp of novel sequences from each individual in two Swedish genome projects.[16] In 2019, a landmark work used 15 samples to produce the largest long-read structural variant callset and reported 6.4 Mbp of novel sequences per individual.[17] These studies discovered a certain number of novel sequences, but considering varying lengths of DNA reads and different analytical approaches, it is necessary to evaluate their impact on building complete novel sequences.

In this study, we defined DNA sequences missed from the human reference and longer than 300 bp as "novel sequences." We proposed an assembly-free novel sequence

Received 7 May 2022; Revised 19 October 2022; Accepted 27 October 2022

(AF-NS) approach that performs quick identification of novel sequences without assembling processes. Derived from ONT long reads, the AF-NS detected novel sequences covered over 90% of Illumina novel sequences and contained more DNA information missing from the Illumina data. Our findings show the advantage of AF-NS in obtaining more intact DNA sequences while saving considerable computational resources, and the importance of long-read sequencing in building a complete picture of novel sequences. Furthermore, we uncovered the biological significance of the identified novel sequences and the population-specific novel insertions.

2. Materials and methods

2.1. Samples and data sources

To study the effects of sequencing technologies on detecting novel sequences, we selected both ONT long reads and Illumina paired-end short reads from six samples—an Ashkenazim trio (HG002, HG003, HG004) and a Chinese trio (HG005, HG006, HG007)—extracted from the Genome in a Bottle (GIAB) project.[18] The ONT reads were basecalled using Guppy v4.2.2. To keep the consistent sequencing depth for parent–offspring trios, we down-sampled the three Ashkenazim samples (HG002–HG004) to 50-fold and the three Chinese samples (HG005–HG007) to 40-fold. To guarantee the identified sequences/reads that are truly absent in the human reference, we combined the latest references, GRCh38.p13 and chm13.v1.1, as the new reference. In addition, we removed chromosome Y when identifying novel sequences of HG004 and HG007 because the two samples were female.

2.2. AF-NS: assembly-free novel sequences construction

As shown in Fig. 1, AF-NS carried out an assembly-free pipeline consisting of three main steps. First, we aimed to obtain high-quality unmapped reads. All ONT long reads were aligned to references using minimap 2.17-r941,[19] and reads with unmapped fragments longer than 300 bp were selected. We then discarded low-quality reads with <10 Q-score by NanoFilt[20] and cut the adaptors of reads using porechop v0.2.4.[21] Second, we collected the unmapped sequences by three-round alignment-filtering. The first round was to align the filtered reads to references using minimap2 (-map-ont) and to obtain the unmapped fragments. The second round was to align these unmapped sequences generated from the first round to references using NUCmer 4.0.0rc122[22] (-l 15 -c 31). This would further remove the sequences mapped to the references. In last round, we aligned the remaining sequences to references again using minimap2 with the default parameters. Any fragments aligned to references and shorter than 300 bp were removed. Third, we eliminated sequences labelled as archaea, bacteria, fungi, plasmid, viral and UniVec (confidence score > 0.05) by Kraken2[23] and used minimap2 to overlap the remaining sequences. Two sequences were clustered together if they aligned with each another with >80% coverage. We chose the longest sequence in a cluster as its representative. Since the long-read sequencing could be inaccurate in low-complexity regions,[24] we utilized RepeatMasker 4.1.2-p1[25] to annotate the novel sequences and removed sequences if 80% of them were low-complexity or simple repeats. The remaining ones were collected as the final set of novel sequences.

2.3. novel_WG: extraction of novel sequences from the whole-genome assembly

We executed whole-genome assemblies using two popular assemblers, Shasta 0.7.0[26] and raven 1.7.0,[27] and aligned the assemblies to references using NUCmer. Any fragments mapped to the reference with ≥80% identity were removed. Next, we aligned the unmapped fragments >300 bp to references using BWA-mem v0.7.17[28] and further eliminated the mapped parts if the alignment identity reached 80%. Then, to ensure the novelty of the remaining sequences, we aligned them to references using BWA-mem again and discarded any sequences that had hits to references satisfying >80% identity and >50% coverage. Sequentially, we used Kraken2 to remove contaminated sequences to obtain the final set of novel sequences. For Illumina data, we performed *de novo* assembly using MEGAHIT v1.2.9.[29] Other procedures were the same as those used for long-read sequencing data.

2.4. unmap_ASM: assembly of unmapped reads

Illumina reads were mapped to human references using BWA-mem, and the unmapped parts were obtained and assembled into sequences using MEGAHIT. Next, we deleted novel sequences shorter than 300 bp and aligned the remaining ones to references using BWA-mem. Any sequences mapped to references with identity ≥ 80% and coverage ≥ 50% were filtered out. Finally, we eliminated contaminants identified by Kraken2 and generated a set of novel sequences.

2.5. Detection of novel placements

We aligned the ONT long reads to chm13 v1.1 using minimap2 and retained alignments with ≥80% identity. For reads with one or two consistent alignments, we extracted the corresponding unmapped fragments longer than 1,000 bp. These fragments were classified into two groups: (i) those with a single end aligned to the reference (SEP) and (ii) those with both ends mapped to the reference (BEP). Then we clustered two types of fragments. Two BEP sequences were merged if the length difference between their overlap and length was less than 100 bp. SEP sequences within 100 bp of each other were combined using BEDtools.[30] The longest sequence in a cluster was selected as its representative. We excluded clusters containing fewer than three sequences to obtain the final placement clusters.

3. Results

3.1. Procedure for detecting long-read novel sequences

To examine the effects of different length reads on the discovery of novel sequences, we chose six samples, an Ashkenazim trio (HG002, HG003, HG004) and a Chinese trio (HG005, HG006, HG007), both of which have ONT long-read and Illumina short-read data. First, we conducted two widely used strategies to acquire long-read novel sequences, named novel_WG and unmap_ASM. The first approach involved extracting novel sequences from the whole-genome assembly. We used Shasta and raven to generate the whole-genome assemblies. However, the resulting novel sequences differed greatly in size, and more than 50% of them could not be aligned against each other with ≥80% identity (Table 1, Supplementary Table 1). This inconsistency might be due to different whole-genome assemblers. The authenticity of these unmapped sequences requires further verification to

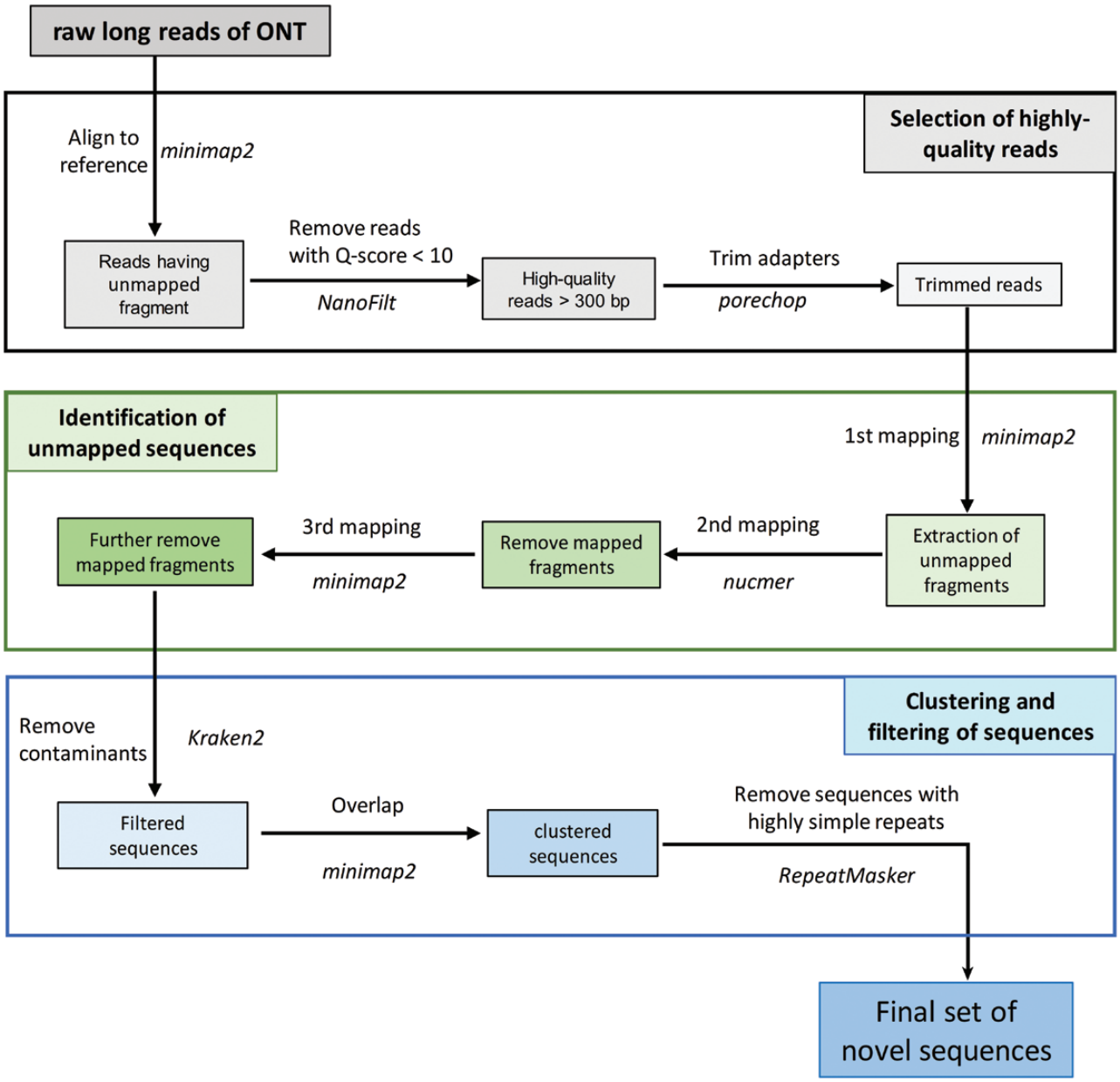

Figure 1. Workflow for identification of long-read novel sequences by assemble-free novel sequence (AF-NS) approach. Using long reads of Oxford Nanopore Technology (ONT), AF-NS carries out three main steps to discover the novel sequences missed from genome references in humans.

Table 1. Size (bp) of long-read novel sequences identified by AF-NS and novel_WG

Samples	AF-NS	novel_WG (Shasta)	novel_WG (raven)
HG002	19,055,769	182,032	1,328,545
HG003	23,950,540	207,148	728,462
HG004	22,731,658	240,347	2,141,062
HG005	15,973,628	224,995	1,361,765
HG006	16,549,640	167,900	1,201,172
HG007	15,853,142	238,687	1,024,175

AF-NS is a newly designed assembly-free approach. novel_WG detects novel sequences from the long-read whole-genome assembly generated using two assemblers, Shasta and raven, respectively.

determine whether there are assembly-caused artifacts. The mapped ones were treated as common novel_WG sequences with high confidence. Using the unmap_ASM approach, we obtained only a few novel sequences, indicating its ineffectiveness for long-read data (Supplementary Table 2).

To discover more complete long-read novel sequences, we designed a new pipeline AF-NS, consisting of a three-step procedure without using assemblers (Fig. 1). The first step obtained high-quality reads unmapped to the human reference and trimmed their adaptors. Then, through a three-round alignment-filtering process, the unmapped fragments longer than 300 bp were extracted from the trimmed reads as novel candidates. During the last step, after removing contaminants, similar candidates were clustered together to generate the final set of novel sequences. Compared with other methods, which assemble reads into contigs (contiguous sequences), AF-NS obtains novel genomic information from high-quality long reads. This ensures its efficiency in detecting novel sequences through minimizing sequence loss due to assembling.

3.2. Comparison of novel sequences

First, we compared AF-NS with the assembly-based approach novel_WG. AF-NS does not require computationally intensive whole-genome assembly and can detect novel sequences 15 times more than novel_WG (Table 1). We note that over 85% of the common novel_WG sequences were mapped to AF-NS-identified sequences with ≥80% identity (Table 2), showing the superiority of AF-NS in capturing more novel genomic information. Then we aligned the AF-NS sequences to whole-genome assemblies using NUCmer. The mapping percentage was less than 5% with ≥50% coverage (Table 3). This result indicates that current long-read assemblers lose most of the novel genomic information.

Next, we compared the novel sequences generated from different sequencers. We first constructed Illumina novel sequences of the six samples using two methods, unmap_ASM and novel_WG. The size of novel_WG sequences was slightly larger than that of unmap_ASM sequences (Table 4). We aligned novel_WG and unmap_ASM sequences with each other and found that ~85% of unmap_ASM sequences or ~95% of novel_WG sequences could be mapped together with ≥ 80% identity (Supplementary Table 3). This result shows the consistency of novel sequences obtained from the two approaches. To minimize assembly-induced artifacts, we retained only the novel sequences detected by both.

Table 2. Comparison of long-read novel sequences (bp) using AF-NS and novel_WG

Samples	Common novel_WG vs. AF-NS sequences aligned length (percentage of alignment)	AF-NS sequences vs. common novel_WG aligned length (percentage of alignment)
HG002	59,429 (86.56%)	480,109 (2.52%)
HG003	58,160 (95.7%)	416,477 (1.74%)
HG004	91,127 (93.85%)	769,500 (3.39%)
HG005	73,791 (85.76%)	360,576 (2.26%)
HG006	73,261 (91.51%)	405,168 (2.45%)
HG007	100,790 (90.99%)	557,672 (3.52%)

Common novel_WG: the novel sequences extracted from both Shasta and raven assemblies using novel_WG approach can be aligned with each other with ≥ 80% identity. AF-NS sequences: long-read novel sequences identified by the AF-NS approach.

Table 3. Alignment of AF-NS sequences to long-read whole genome assembly (bp)

Samples	AF-NS sequences vs. Shasta aligned length (percentage of alignment)	AF-NS sequences vs. raven aligned length (percentage of alignment)
HG002	570,521 (2.99%)	603,178 (3.17%)
HG003	735,000 (3.07%)	602,454 (2.52%)
HG004	1,330,784 (5.85%)	1,173,459 (5.16%)
HG005	491,228 (3.08%)	426,414 (2.67%)
HG006	510,282 (3.08%)	479,833 (2.9%)
HG007	707,499 (4.46%)	779,358 (4.92%)

The AF-NS sequences mapped to long-read whole genome assembly generated by two assemblers, Shasta and raven, with ≥50% coverage.

The HG005 sample was excluded because the size of novel sequences discovered from its Illumina reads was considerably higher than that of the other samples (Table 4).

As shown in Fig. 2A, the AF-NS identified sequences covered more than 91% of Illumina novel sequences under ≥80% identity and ≥50-bp alignment length. By contrast, only 58–80% of the Illumina novel sequences were mapped to the long-read whole genomes using an assembler Shasta or raven (Fig. 2A). This analysis verifies the effectiveness of the AF-NS approach in discovering intact novel sequences. Because the above two identification methods may lose some novel sequences, we mapped entire Illumina raw reads to the AF-NS sequences using BWA-mem. Under the same threshold, less than 15% of the AF-NS sequences were covered by Illumina data (Fig. 2B). We also ran BUSCO[31] on both ONT and Illumina whole-genome assemblies using the eukaryote set of orthologs from OrthoDB v.10.[32] As expected, the short-read whole genome was not as complete as the long-read assemblies (Supplementary Table 4). These results indicate that short-read sequencing does miss a lot of genomic information.

3.3. Characteristics of novel sequences

According to our previous definition of the novel sequence composition,[9] we classified the AF-NS detected sequences as common and individual-specific components. A sequence was considered common if it could be mapped to other samples' sequences by satisfying ≥80% identity. Each of the six samples contained 1–2 Mbp common sequences, accounting for 6–8% of total sequences (Supplementary Table 5). To search for the origin of the AF-NS sequences, we aligned them to the chimpanzee genome (GCF_002880755.1) using NUCmer. Under the minimum requirement of 80% identity, about 7% of the novel sequences were present in the chimpanzee genome. By contrast, the common sequences of each sample obtained much higher alignment percentages of ~60% compared with individual-specific sequences of 2–3% (Fig. 3A).

To determine the novelty of the AF-NS sequences, we mapped them to existing novel sequences of three long-read sequencing samples, HX1 (Chinese) and two Swedes.[15,16] Few AF-NS sequences were aligned with ≥80% identity, indicating that most of them have not been reported (Supplementary Table 6). Next, we compared the AF-NS sequences with two pan-genomes, the African pan-genome and the Chinese pan-genome.[8,33] Only about 3–7% of the AF-NS sequences could be matched to either one, implying a great number of novel sequences lost in the pan-genomes (Fig. 3B and C). The low

Table 4. Size (bp) of short-read novel sequences identified by two methods

Samples	novel_WG	unmap_ASM
HG002	130,598	122,250
HG003	133,065	124,995
HG004	161,827	139,584
HG005	3,018,105	445,921
HG006	137,778	123,754
HG007	181,782	157,256

Novel_WG represents a method that identifies novel sequences from the whole-genome assembly. unmap_ASM represents a method for assembling reads unmapped to the human reference into sequences.

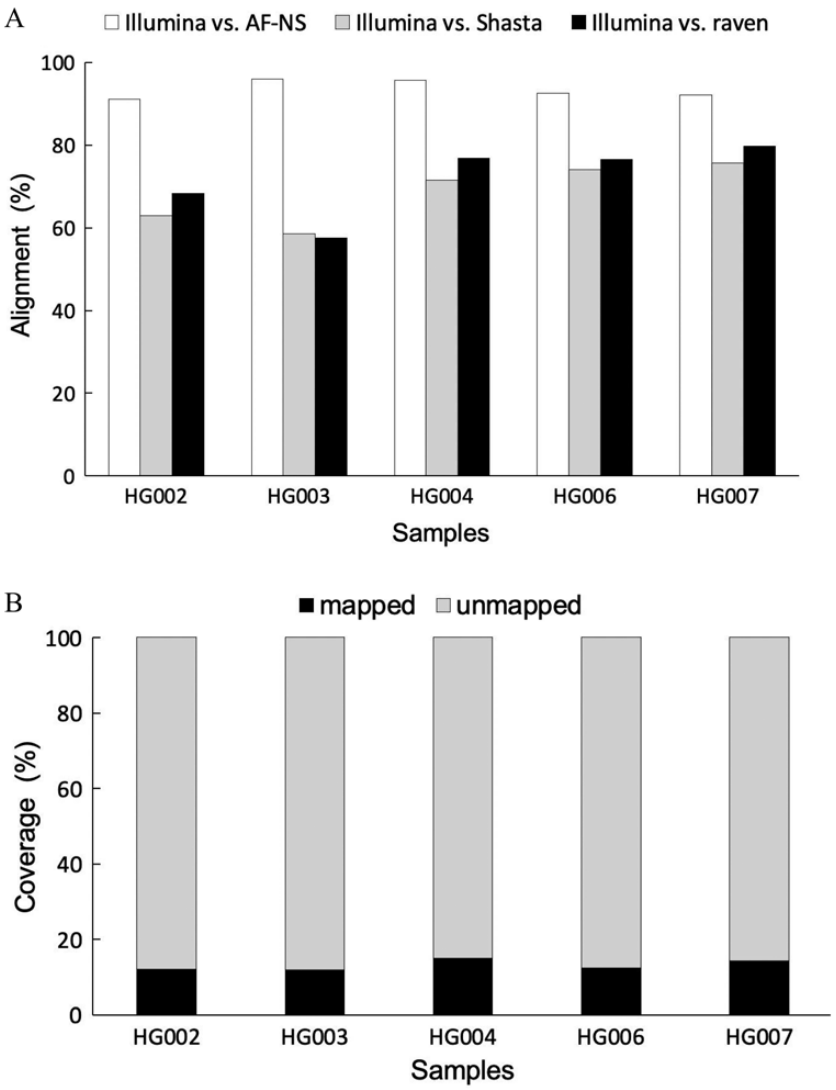

Figure 2. Comparison of short-read and long-read novel sequences. (A) The percentage of short-read novel sequences of Illumina that were aligned to long-read sequences; AS-NS: long-read novel sequences identified by the AS-NS method; Shasta: long-read whole-genome assemblies generated using Shasta; raven: long-read whole-genome assemblies generated using raven. (B) The percentage of long-read novel sequences that were mapped to short-read raw reads; mapped: long-read novel sequences were mapped to short-read raw reads; unmapped: long-read novel sequences were not mapped to short-read raw reads. All alignments were based on at least 80% identity and 50-bp alignment length.

alignment rate is possibly because the current pan-genomes were assembled using short-read data. However, the percentage of alignment to pan-genomes was still three to four times higher than that to the long-read novel sequences of an individual, emphasizing the importance of large-scale sequencing and assembling for obtaining complete genomic information.

In order to reveal the biological significance of the identified novel sequences, we, respectively, extracted the highly common sequences shared among at least four, five, and all six samples. Transcription factors (TFs) represent the main regulators that control gene transcription by binding to the DNA sequences. We performed TF binding motif searching on the highly common and individual-specific sequences using fimo[34] and based it on the TRANSFAC human dataset.[35] With a cutoff of *P*-value < 5e-8, we searched for TF binding to these highly common sequences. There were 71 TF motifs identified to bind to the sequences shared by all six samples (Fig. 3D, Supplementary Table 7). We display binding patterns of 15 main TFs that are believed to have a broad set of biological functions (Fig. 3E), such as transcriptional regulation (CEBPB, P300), immune or inflammatory response (AP1, IRF1, STAT1, PPARG, NFE2L2), tumours (AP1, SP1, EGR1, STAT1, PAX2, etc.), metabolism (FOXO1, PPARG), and growth and development of cells or organs (E2F1, SP1, MEF2A, NR2F2, HNF4A, PAX2, NEF2L2, PPARG, etc.). These proposed TFs suggest possible functional implications of our novel sequences.

3.4. Analysis of novel placements

We positioned novel sequences in chm13.v1.1 and allowed the placed sequences within 100 bp of each other to be clustered together. We obtained 1,637 placed clusters among the six samples and used the corresponding placements to analyze the inserting preferences. The majority of placements (69%) were located in repeat regions, especially satellite and microsatellite regions (Supplementary Table 8, Fig. 4A). The enrichment of novel sequences in repeats might be due to the

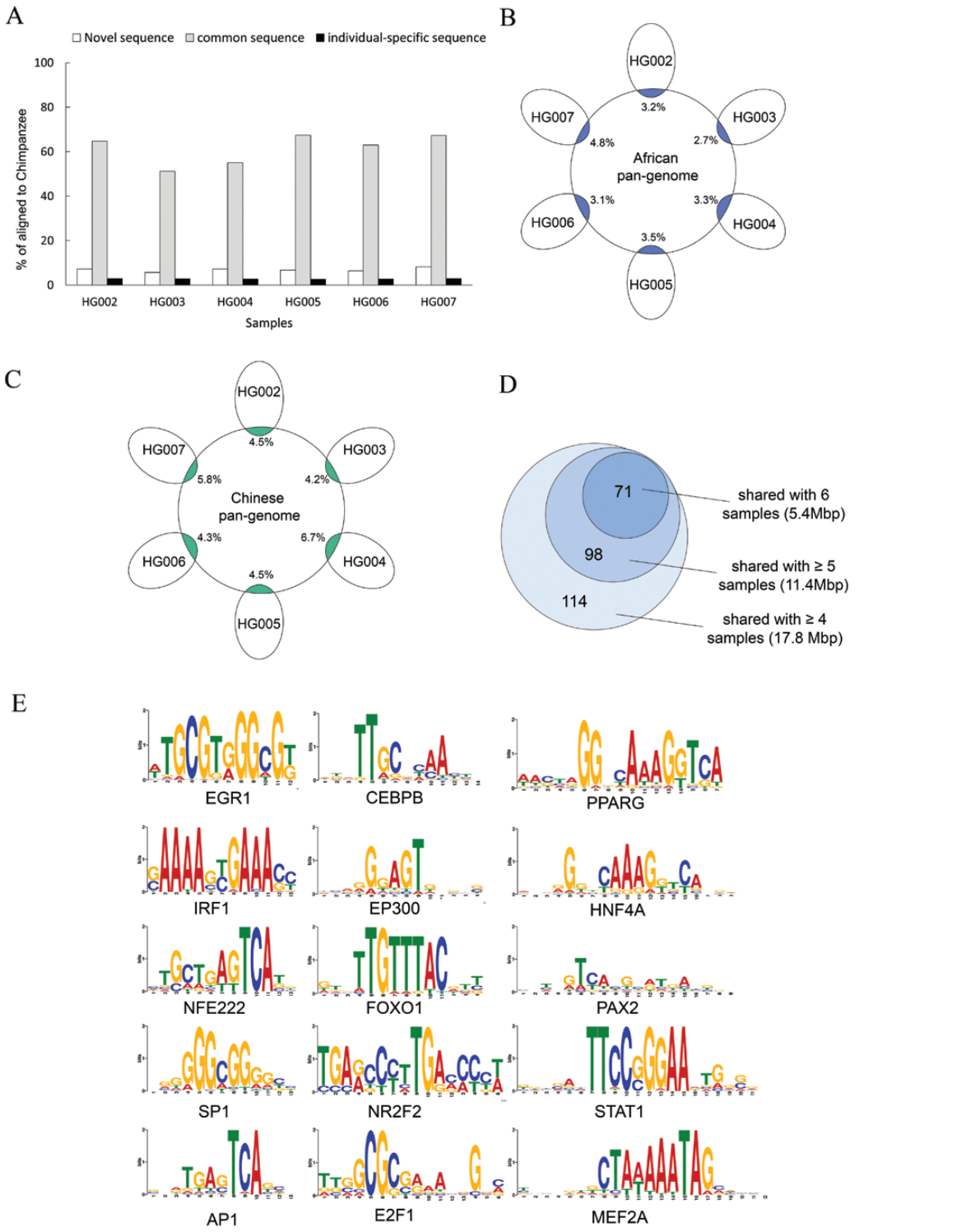

Figure 3. Analysis of the characteristics of long-read novel sequences. (A) Percentage of the AF-NS identified novel sequences mapped to the chimpanzee genome with ≥80% identity. Common sequence: novel sequences shared by at least two samples. Individual-specific sequence: novel sequences found in one individual sample. (B, C) Percentage of the AF-NS identified sequences mapped to African pan-genome (B) and Chinese pan-genome (C) with ≥80% identity. (D) Number of transcription factor (TF) binding motifs identified from the highly common sequences that are shared among at least four, five and six samples, respectively. The lengths of highly common sequences are also shown. (E) Binding motif logos of 15 main TFs that were predicted to bind to the highly common sequences shared by all six samples.

highly unstable nature of these regions, which are prone to generating variants. These variations probably drive genome plasticity and promote evolution.[36,37] In addition, we found 54 and 68 novel placements were in centromeres and telomeres, respectively, where genomic information is difficult to obtain using short-read data.[11] Long-read sequencing enables the mapping of novel sequences in highly condensed heterochromatin regions.

Then, we investigated the functional significance of the placed insertions using chm13 gene annotation (version 4).

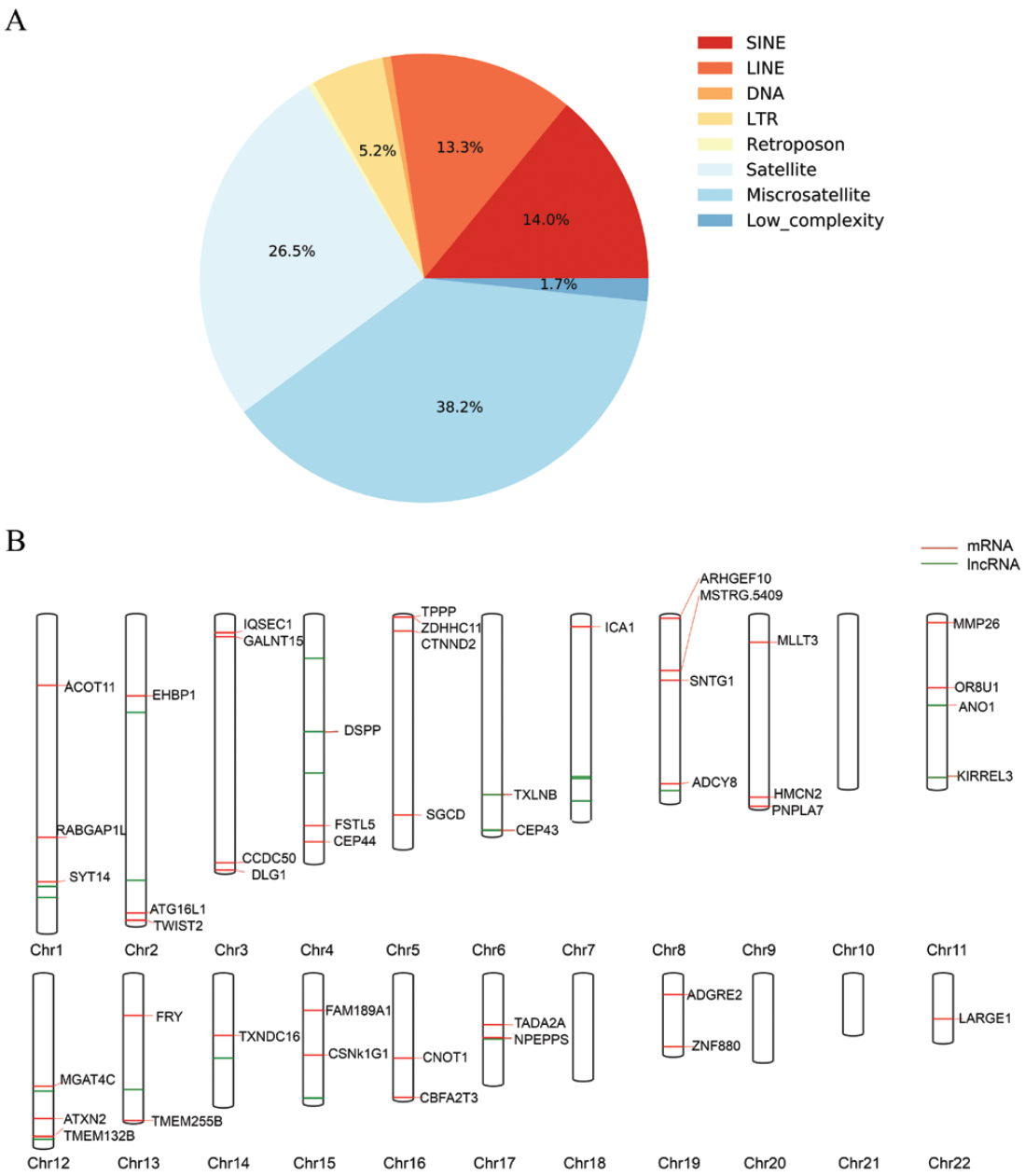

Figure 4. Properties of novel placements. (A) Percentage of novel sequence placements located in interspersed repeat and low-complexity regions. Satellites and microsatellites account for the largest proportions. (B) In the genome map, the red and green lines represent the novel sequence placements located in protein-coding and lncRNA regions on autosomes, respectively. These placements are shared among all samples. The corresponding protein-coding gene symbols are shown. Detailed annotations are reported in Supplementary Table 8.

A total of 613 insertions fell within 503 genes, including 301 coding and 162 long non-coding (lncRNA) regions (Supplementary Table 8). We reported 46 protein-coding and 22 lncRNA genes shared among all samples (Fig. 4B). The 46 protein-coding genes involved mainly cell adhesion (*ADGRE2*, *CTNND2*, *HMCN2*, *DSPP*, *KIRREL3* and *DLG*), actin cytoskeleton organization (*ARHGEF10*, *IQSEC1* and *EHBP1*), endocytosis (*RABGAP1L*, *EHBP1* and *CSNK1G1*) and multicellular organism development (*ANO1*, *FSTL5*, *TWIST2* and *DSPP*).

Furthermore, we noticed among the novel placements, 1,514 common ones were present in at least two samples (Fig. 5A). Since the GIAB samples were from Ashkenazim and Chinese subpopulations, we further divided the common novel placements into population-shared (across subpopulations) and population-specific (within a subpopulation). The 1,084 population-shared insertions as patches can supplement the latest reference genome. We obtained 430 population-specific insertions, including 198 Chinese-specific and 232 Ashkenazim-specific (Fig. 5B) insertions. Then, we found 260 population-specific insertions present in only one parent and its offspring. If these mutations are newly generated in the parents, they may serve as new heritable variants.

4. Discussion

The identification of novel sequences is a basis for complementing the existing reference genome in humans.[38] However,

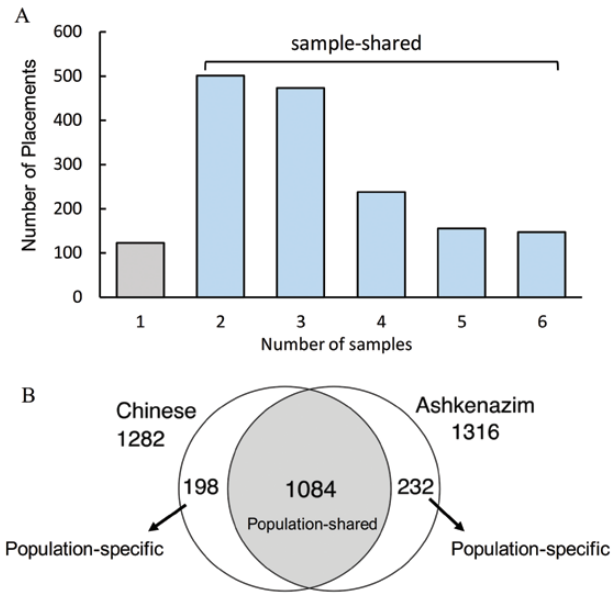

Figure 5. Novel sequence placements detected in six samples. (A) Number of novel sequence placements detected in only one or shared by two to six samples. (B) Intersection of the novel sequence placements shared by at least two samples within the Ashkenazim or Chinese subpopulation.

influenced by different sequencing technologies and computational methods, the derived novel sequences vary considerably.[16,39,40] Finding a widely accepted standard for constructing novel sequences is a challenging issue in genomics research. Currently, there are two strategies for building novel sequences. The first one involves filtering out reads mapped to references and assembling the unmapped reads into novel sequence sets.[8,9,41] Another involves performing the whole-genome *de novo* assembly and extracting the unmapped sequences.[16,40] However, the quality of the identified novel sequences using these strategies is highly dependent on the performance of assemblers. When assembling unmapped reads, some novel sequences can be lost because of insufficient supporting reads or overlaps between reads that fail to meet assembly requirements.[42] Owing to the short length of NGS data, genome assembly is essential for identifying novel sequences that exceed the read lengths. By contrast, long-read sequencing can generate reads longer than 100 kbp. Therefore, a single read can decipher large structural variations, which guarantees the feasibility of AF-NS to discover novel sequences at read level. Our assembly-free strategy includes a series of processes to ensure the completeness and accuracy of the identified novel sequences. First, the selected ONT reads during the first step should be of high quality. Second, the contaminants in the clustered candidate sequences must be filtered out. Last, we removed highly simple repetitive sequences from the final set due to the high error rate of long-read sequencing in low-complexity regions.[24] To obtain more complete novel sequences, we did not use the self-correction approaches that are expected to correct the sequencing errors because their performances are largely affected by relatively low sequencing depths,[43] this possibly causes missing of novel sequences due to no sufficient long-read coverages. The performance of AF-NS demonstrates its advances in identifying novel sequences that have a larger size and better characteristics over other methods and saves considerable computational resources.

With the advancement of sequencing technologies, different ones have been applied to detect novel sequences.[15,39,44] However, the corresponding analyses focused only on specific data sets. Here, we compared the novel sequences generated using short- and long-read data and found a big difference, i.e., that most of the long-read sequences cannot be covered by short-read ones. Also, most of the long-read novel sequences were absent in Illumina pan-genomes even though the pan-genomes were built using hundreds of samples. These findings demonstrate that utilizing long-read technologies is the inevitable trend for creating complete novel DNA sequences.

Interaction between TFs and TF binding DNA sites on promotors is a key for transcriptional gene regulation. To analyze biological associations of the AF-NS novel sequences, we identified 71 TF binding motifs on the highly common sequences shared by all samples. These TFs are assumed to play regulatory functions involving many basic biological processes and pathways. Although the TF binding sites were not confirmed to promoters or non-coding regions, this finding suggests a wide biological connection to the newly detected sequences. Moreover, we discovered the population-shared and population-specific insertions by clustering placed novel sequences of six samples. The population-shared insertions indicate that the latest human reference genome is still underestimated, possibly because chm13 was derived from one genome.[3] The population-specific insertions may be the force of the subpopulation evolution. Considering our classification using only six samples, increasing individuals is expected to assist in establishing more representative population-specific or common insertions. These population-specific insertions will probably allow us to determine the relationships of novel sequences with human evolution.

In summary, AF-NS represents a rapid sequence detection method from long reads and outperforms the existing assemblers in discovering intact novel sequences. The assembly-free design considerably reduces computation costs and facilitates the construction of large-scale novel sequence genomes. The superiority of long-read sequencing in identifying more novel sequences and positioning them in difficult-to-assemble regions, like centromeres and telomeres, shows its great potential for future novel sequence research. The high cost of long-read sequencing is still a challenge for generating numerous genomes. However, as more samples using advanced sequencers become available, we believe that we can establish a more comprehensive landscape of novel sequences.

Conflict of interest

R.L. receives research funding from Oxford Nanopore Technologies. The remaining authors declare no competing interests.

Acknowledgements

R.L. was supported by the General Research Funding [17113721] of the Hong Kong SAR government, General Program [JCYJ20210324134405015] of the Shenzhen municipal government, China, and URC fund by the University of Hong Kong.

Author contributions

Q.L. and B.Y. performed data processing and analysis. R.L. conceived and advised the project. All authors wrote and approved the final manuscript.

Data availability

The AF-NS identified sequences can be found at https://www.bio8.cs.hku.hk/novel/AF-NS/. The implementation of the AF-NS method is available at https://github.com/HKU-BAL/AF-NS. The Illumina paired-end short reads of HG002-HG007 were downloaded from https://github.com/genome-in-a-bottle/giab_data_indexes. The ONT long reads of HG002-HG007 were downloaded from https://s3-us-west-2.amazonaws.com/human-pangenomics/index.html?prefix=NHGRI_UCSC_panel/. The novel sequences of 910 Africans in APG were downloaded from NCBI with accession number PDBU01000000. The novel sequences of 486 Chinese in CPG were downloaded from the Genome Sequence Archive for Human with the accession number PRJCA003657. The assembly of HX1 was downloaded from http://www.openbioinformatics.org/hx1/data/nonhg38.fa.gz.The assemblies of the two Swedes were downloaded from https://www.mdpi.com/2073-4425/9/10/486 in Supplementary Data S1.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Sherman, R.M. and Salzberg, S.L. 2020, Pan-genomics in the human genome era, *Nat. Rev. Genet.*, **21**, 243–54.
2. Ballouz, S., Dobin, A. and Gillis, J.A. 2019, Is it time to change the reference genome?, *Genome Biol.*, **20**, 159.
3. Nurk, S., Koren, S., Rhie, A., et al. 2021, The complete sequence of a human genome, *bioRxiv*, doi:10.1101/2021.05.26.445798.
4. Taliun, D., Harris, D.N., Kessler, M.D., et al. 2019, Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program, *BioRxiv*, doi:10.1101/563866.
5. Hehir-Kwa, J.Y., Marschall, T., Kloosterman, W.P., et al.; Genome of the Netherlands Consortium. 2016, A high-quality human reference panel reveals the complexity and distribution of genomic structural variants, *Nat. Commun.*, **7**, 12989.
6. Kidd, J.M., Cooper, G.M., Donahue, W.F., et al. 2008, Mapping and sequencing of structural variation from eight human genomes, *Nature*, **453**, 56–64.
7. Kidd, J.M., Sampas, N., Antonacci, F., et al. 2010, Characterization of missing human genome sequences and copy-number polymorphic insertions, *Nat. Methods*, **7**, 365–71.
8. Sherman, R.M., Forman, J., Antonescu, V., et al. 2019, Assembly of a pan-genome from deep sequencing of 910 humans of African descent, *Nat. Genet.*, **51**, 30–5.
9. Li, Q., Tian, S., Yan, B., et al. 2021, Building a Chinese pan-genome of 486 individuals, *Commun. Biol.*, **4**, 1–14.
10. Chaisson, M.J., Sanders, A.D., Zhao, X., et al. 2019, Multi-platform discovery of haplotype-resolved structural variation in human genomes, *Nat. Commun.*, **10**, 1–16.
11. Miga, K.H., Koren, S., Rhie, A., et al. 2020, Telomere-to-telomere assembly of a complete human X chromosome, *Nature*, **585**, 79–84.
12. Sedlazeck, F.J., Lee, H., Darby, C.A. and Schatz, M.C. 2018, Piercing the dark matter: bioinformatics of long-range sequencing and mapping, *Nat. Rev. Genet.*, **19**, 329–46.
13. Lee, H., Gurtowski, J., Yoo, S., et al. 2016, Third-generation sequencing and the future of genomics, *BioRxiv*, doi:10.1101/048603.
14. Jain, M., Koren, S., Miga, K.H., et al. 2018, Nanopore sequencing and assembly of a human genome with ultra-long reads, *Nat. Biotechnol.*, **36**, 338–45.
15. Shi, L., Guo, Y., Dong, C., et al. 2016, Long-read sequencing and de novo assembly of a Chinese genome, *Nat. Commun.*, **7**, 1–10.
16. Ameur, A., Che, H., Martin, M., et al. 2018, De novo assembly of two Swedish genomes reveals missing segments from the human GRCh38 reference and improves variant calling of population-scale sequencing data, *Genes (Basel)*, **9**, 486.
17. Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., et al. 2019, Characterizing the major structural variant alleles of the human genome, *Cell*, **176**, 663. e619–75.e19.
18. Zook, J.M., Catoe, D., McDaniel, J., et al. 2016, Extensive sequencing of seven human genomes to characterize benchmark reference materials, *Sci. Data*, **3**, 160025.
19. Li, H. 2018, Minimap2: pairwise alignment for nucleotide sequences, *Bioinformatics*, **34**, 3094–100.
20. De Coster, W., D'Hert, S., Schultz, D.T., Cruts, M. and Van Broeckhoven, C. 2018, NanoPack: visualizing and processing long-read sequencing data, *Bioinformatics*, **34**, 2666–9.
21. Wick, R., Volkening, J. and Loman, N. 2017, *Porechop*. Github. Available at: https://github.com/rrwick/Porechop.
22. Marçais, G., Delcher, A.L., Phillippy, A.M., et al. 2018, MUMmer4: a fast and versatile genome alignment system, *PLoS Comput. Biol.*, **14**, e1005944.
23. Wood, D.E., Lu, J. and Langmead, B. 2019, Improved metagenomic analysis with Kraken 2, *Genome Biol.*, **20**, 257.
24. Delahaye, C. and Nicolas, J. 2021, Sequencing DNA with nanopores: Troubles and biases, *PLoS One*, **16**, e0257521.
25. Tarailo-Graovac, M. and Chen, N. 2009, Using RepeatMasker to identify repetitive elements in genomic sequences., *Curr. Protoc. Bioinformatics*, **25**, 4.10. 11–14.10. 14.
26. Shafin, K., Pesout, T., Lorig-Roach, R., et al. 2020, Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes, *Nat. Biotechnol.*, **38**, 1044–53.
27. Vaser, R. and Šikić, M. 2021, Raven: a de novo genome assembler for long reads, *BioRxiv*, doi:10.1101/2020.08.07.242461.
28. Li, H. 2013, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM., *arXiv*, doi:10.48550/arXiv.1303.3997.
29. Li, D., Liu, C.-M., Luo, R., Sadakane, K. and Lam, T.-W. 2015, MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph, *Bioinformatics*, **31**, 1674–6.
30. Quinlan, A.R. and Hall, I.M. 2010, BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics*, **26**, 841–2.
31. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–2.
32. Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., et al. 2019, OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs, *Nucleic Acids Res.*, **47**, D807–11.
33. Li, Q., Tian, S., Yan, B., et al. 2021, Building a Chinese pan-genome of 486 individuals, *Commun. Biol.*, **4**, 1016.
34. Grant, C.E., Bailey, T.L. and Noble, W.S. 2011, FIMO: scanning for occurrences of a given motif, *Bioinformatics*, **27**, 1017–8.
35. Matys, V., Kel-Margoulis, O.V., Fricke, E., et al. 2006, TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes, *Nucleic Acids Res.*, **34**, D108–10.
36. Course, M.M., Gudsnuk, K., Smukowski, S.N., et al. 2020, Evolution of a human-specific tandem repeat associated with ALS, *Am. J. Human Genetics*, **107**, 445–60.
37. Kashi, Y. and King, D.G. 2006, Simple sequence repeats as advantageous mutators in evolution, *Trends Genet.*, **22**, 253–9.
38. Hajirasouliha, I., Hormozdiari, F., Alkan, C., et al. 2010, Detection and characterization of novel sequence insertions using paired-end next-generation sequencing, *Bioinformatics (Oxford, England)*, **26**, 1277–83.

39. Kehr, B., Helgadottir, A., Melsted, P., et al. 2017, Diversity in non-repetitive human sequences not found in the reference genome, *Nat. Genet.*, **49**, 588–93.
40. Eisfeldt, J., Mårtensson, G., Ameur, A., Nilsson, D. and Lindstrand, A. 2020, Discovery of novel sequences in 1,000 Swedish genomes, *Mol. Biol. Evol.*, **37**, 18–30.
41. Li, R., Li, Y., Zheng, H., et al. 2010, Building the sequence map of the human pan-genome, *Nat. Biotechnol.*, **28**, 57–63.
42. Jiang, T., Fu, Y., Liu, B. and Wang, Y. 2019, Long-read based novel sequence insertion detection with rCANID, *IEEE Trans. Nanobiosci.*, **18**, 343–52.
43. Zhang, H., Jain, C. and Aluru, S. 2020, A comprehensive evaluation of long read error correction methods, *BMC Genomics*, **21**, 889.
44. Huddleston, J., Chaisson, M.J., Steinberg, K.M., et al. 2017, Discovery and genotyping of structural variation from long-read haploid genome sequence data, *Genome Res.*, **27**, 677–85.