Long read de novo sequencing from blood using SQK-LSK114 – scalable (near) telomere-to-telomere method

Downstream analysis recommendations

To carry out the scalable long read assembly downstream analysis, we recommend previous bioinformatics experience.

The commands below enable dual assembly using only the data generated above from LSK114 sequencing data to produce a collapsed (i.e. pseudo-haploid) or "dual" (i.e. partially phased i.e. pseudo-diploid) assembly.

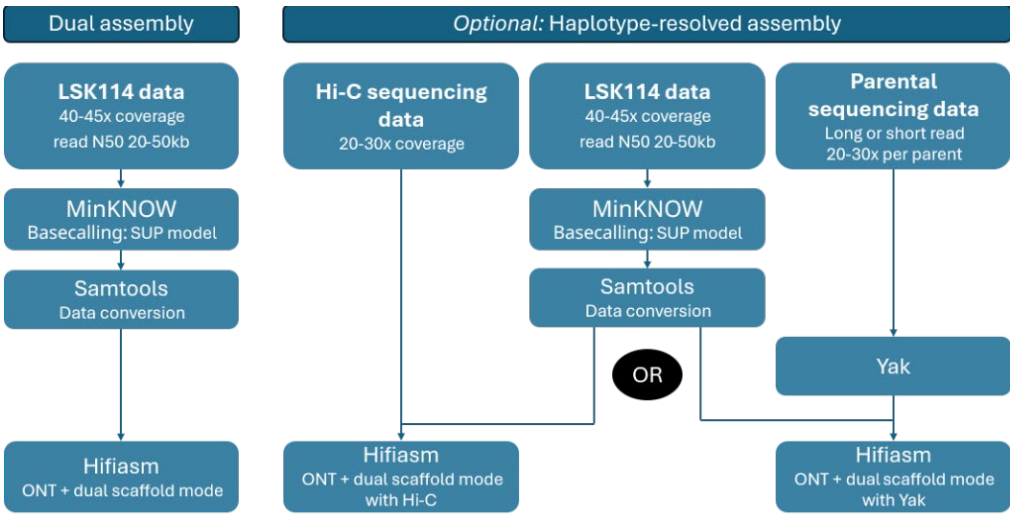
Please note: The information outlined in this document are recommendations for 3rd party analysis tools and methods to use in conjunction with the library preparation method. Oxford Nanopore Technologies do maintain or support these analysis tools. For further information and support please refer to the tools repositories.

Analysis considerations

The library preparation method aimed to generate 40–45x coverage long read data, with an N50 of approximately 30kb. The downstream analysis handles read lengths in the range of N50 20–50kb. Reasonably complete assemblies can be achieved down to circa 30x coverage and an N50 of 10kb, but contiguity will be severely diminished (see the telomere-to-telomere sequencing (T2T) know-how document for further details).

Phased data are required to produce a haplotype-resolved assembly, and we therefore include an optional workflow to combine the LSK114 sequencing data produced above with either Hi-C sequencing data (Hi-C assembly) **or** long- or short-read parental sequencing data (trio assembly).

Note: Hi-C (20–30x) or parental sequencing data (long or short read, 20–30x per parent) need to be generated in addition to that generated by the sequencing workflow described above, and must be obtained from the same sample for Hi-C or parents of the sample used for the long read data for parental sequencing.



Note: The analysis workflow described here are recommended guidelines using existing 3rd party tools.

Note: All basecalling must have been completed before starting downstream analysis.

Oxford Nanopore Technologies, the Wheel icon, AmPORE-TB, ElysION, EPI2ME, Flongle, GridION, MinION, MinKNOW, PromethION, and TraxION are registered trademarks or the subject of trademark applications of Oxford Nanopore Technologies plc in various countries. Information contained herein may be protected by patents or patents pending of Oxford Nanopore Technologies plc. All other brands and names contained are the property of their respective owners. © 2025 Oxford Nanopore Technologies plc. All rights reserved.

Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition.

Post basecalling analysis compute requirements

≥ 300 Gb RAM
1,000–3,000 CPU hours depending on CPU speed, I/O rate. Add 500–1,000 CPU hours if using Hi-C (parental sequencing adds negligible additional time). Efficiently parallelizable: actual run time on 96-thread AWS machine was 9 hours with no Hi-C

Setting up the environment

The scalable long read assembly workflow is summarised in the figure above and requires the following tools and their dependencies to be installed:

- Samtools (≥v1.2) for format conversion
- Hifiasm (≥ v0.25.0) for assembly
- Awk (any version) file conversion
- (if using trio mode i.e. parental sequencing for phasing) Yak (>= v0.1)

Fastq conversion (if data is currently in BAM format)

Hifiasm (ONT) requires long read data in fastq format rather than fasta as quality scores are used to distinguish informative polymorphisms from sequencing errors during read self-correction. Samtools is used to convert from bam to fastq format:

```
samtools fastq <input.bam> > longreads.fastq
```

Dual assembly

Hifiasm (ONT) is run by specifying the flag “--ont". Pass in the human telomere motif “CCCTAA” (check for your species if not a vertebrate) to enable slightly better assembly at the edges of chromosomes, and enable scaffolding across gaps that occur in one haplotype but not the other with the “--dual-scaf” option. Finally, set the output prefix to “hifiasmONT_asm”.

```
 hifiasm --ont --threads <threads> --telo-m CCCTAA --dual-scaf -o hifiasmONT_asm
longreads.fastq
```


Oxford Nanopore Technologies, the Wheel icon, AmPORE-TB, ElysION, EPI2ME, Flongle, GridION, MinION, MinKNOW, PromethION, and TraxION are registered trademarks or the subject of trademark applications of Oxford Nanopore Technologies plc in various countries. Information contained herein may be protected by patents or patents pending of Oxford Nanopore Technologies plc. All other brands and names contained are the property of their respective owners. © 2025 Oxford Nanopore Technologies plc. All rights reserved.

Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition.

Outputs and conversion to FASTA format

This pipeline produces several output files, most relevantly a collapsed assembly (hifiasmONT_asm.bp.p_ctg.gfa) and two dual assembly haplotypes (hifiasmONT_asm.bp.hap1.p_ctg.gfa, hifiasmONT_asm.bp.hap2.p_ctg.gfa), all in GFA format. To convert to fasta format, you can use many tools, but awk is the simplest. For example, for the collapsed assembly, run:

```
 cat hifiasmONT_asm.bp.p_ctg.gfa | awk '/^S/{print ">" $2 "\n" $3}' > hifiasmONT_asm.bp.p_ctg.fasta
```

Analysis modification: phased assembly

As described above, Hifiasm is able to generate dual assemblies, or partially phased diploid assemblies using the data from the target individual alone, generating a primary and alternate haplotype. Neither of these haplotypes, however, correspond to to the true chromosomal sequences actually present in the sample as the haplotypes are randomly shuffled on either side of each homozygous region. Every single allele will be present in one or the other assembly, but each assembly will have a mix of paternal and maternal alleles within the same contigs.

If you are able to obtain/generate Hi-C data or parental sequencing data, you can build haplotype resolved contigs which faithfully represent the sequences of the maternal and paternal chromosomes on separate contigs. These data can be used to figure out which allele in a polymorphic region belongs to the same physical chromosome (i.e. haplotype) as an allele in another polymorphic region – indeed as all alleles in all polymorphic regions on the same contig. This process is known as assembly phasing, and it yields a phased or **haplotype-resolved assembly**.

Note that with Hi-C data, each haplotype assembly will still be a mix of paternal and maternal chromosomes, but for each chromosome each haplotype assembly will have only maternal or paternal sequence. See Appendix for more details on haplotype resolved assemblies vs. dual assembly.

Hi-C assembly

Assuming your Hi-C data is in paired-end Fastq format (the usual delivery format from the sequencer/sequencing provider), run hifiasm by adding the Hi-C data into your above assembly command:

```
  hifiasm --ont --threads <threads> --telo-m CCCTAA --dual-scaf -o hifiasmONT_asm --h1 <hic_reads_r1.fastq> --h2 <hic_reads_r2.fastq> longreads.fastq
```

You will generate the essentially the same collapsed assembly, named slightly differently (hifiasmONT_asm.hic.p_ctg.gfa), but now you will also generate phased haplotype assemblies: hifiasmONT_asm.hic.hap1.p_ctg.gfa and hifiasmONT_asm.hic.hap2.p_ctg.gfa. See “Outputs and conversion to FASTA format” above for instructions on converting to FASTA format.

Oxford Nanopore Technologies, the Wheel icon, AmPORE-TB, ElysION, EPI2ME, Flongle, GridION, MinION, MinKNOW, PromethION, and TraxION are registered trademarks or the subject of trademark applications of Oxford Nanopore Technologies plc in various countries. Information contained herein may be protected by patents or patents pending of Oxford Nanopore Technologies plc. All other brands and names contained are the property of their respective owners. © 2025 Oxford Nanopore Technologies plc. All rights reserved.

Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition.

Trio assembly

First, you must build YAK-format indexes of kmers present in the maternal and paternal sequences. Sequences can be in either FASTA or FASTQ format. "-b37" is used to instantiate a bloom filter to exclude singleton kmers.

If you have single-end short read or long read parental sequencing data, run:

```
  yak count -b37 -t32 -o mat.yak <maternal_seqs.fast[aq]>

  yak count -b37 -t32 -o pat.yak <paternal_seqs.fast[aq]>
```

If you have paired-end short read data, run:

```
  yak count -b37 -t32 -o mat.yak <(cat <maternal_seqs_r1.fast[aq]>
<maternal_seqs_r2.fast[aq]> ) <(cat <maternal_seqs_r1.fast[aq]>
<maternal_seqs_r2.fast[aq]>)

  yak count -b37 -t32 -o pat.yak <(cat <paternal_seqs_r1.fast[aq]>
<paternal_seqs_r2.fast[aq]> ) <(cat <paternal_seqs_r1.fast[aq]>
<paternal_seqs_r2.fast[aq]>)
```

Note: Change "cat" to "zcat" if gzipped.

Note: Yak protocol does include two identical process substitutions, each concatenating in turn the forward then reverse reads.

Next, run hifiasm by adding the yak data to the above assembly command:

```
hifiasm --ont --threads <threads> --telo-m CCCTAA --dual-scaf -o hifiasmONT_asm -1
pat.yak -2 mat.yak longreads.fastq
```

Again, you will generate essentially the same collapsed assembly as all the other methods, named slightly differently (hifiasmONT_asm.dip.p_ctg.gfa), but now you will also generate *phased* haplotype assemblies: hifiasmONT_asm.dip.hap1.p_ctg.gfa and hifiasmONT_asm.dip.hap2.p_ctg.gfa. See "Outputs and conversion to FASTA format" above for instructions on converting to FASTA format.

Note: if paternal sequences are passed as "-1" and maternal as "-2", "hap1" will be paternal and "hap2" will be maternal – otherwise *vice versa*.

Oxford Nanopore Technologies, the Wheel icon, AmPORE-TB, ElysION, EPI2ME, Flongle, GridION, MinION, MinKNOW, PromethION, and TraxION are registered trademarks or the subject of trademark applications of Oxford Nanopore Technologies plc in various countries. Information contained herein may be protected by patents or patents pending of Oxford Nanopore Technologies plc. All other brands and names contained are the property of their respective owners. © 2025 Oxford Nanopore Technologies plc. All rights reserved.

Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition.

Change log

Date	Version	Changes
July 2026	V1	Document release

Oxford Nanopore Technologies, the Wheel icon, AmPORE-TB, ElysION, EPI2ME, Flongle, GridION, MinION, MinKNOW, PromethION, and TraxION are registered trademarks or the subject of trademark applications of Oxford Nanopore Technologies plc in various countries. Information contained herein may be protected by patents or patents pending of Oxford Nanopore Technologies plc. All other brands and names contained are the property of their respective owners. © 2025 Oxford Nanopore Technologies plc. All rights reserved.

Oxford Nanopore Technologies products are not intended for use for health assessment or to diagnose, treat, mitigate, cure, or prevent any disease or condition.