# Novel long-read sequencing technology in practise

Martin Kašný [1*], Alžběta Hamplová [1], Karolína Stivínová [1], Eliška Hladíková [1], Nela Chalupníková [1], Ondřej Brzoň [1], Petr Kvapil [1]

[1] Institute of Applied Biotechnologies, Služeb 3056/4, 108 00 Prague, Czech Republic
* Correspondence: kasny@iabio.eu; Tel.: +420 739 394 364

Whole genome short-read sequencing is currently commonly used in number of diagnostic applications. The limitations are mostly related to difficulties in mapping of short reads to the challenging regions of the reference genome such as highly homologous regions and repetitive regions. Long-read sequencing is gaining ground in clinical applications as it overcomes some of these issues. However, many long-read sequencing solutions have been plagued by high DNA input requirements, complex workflows with low throughput and highly variable results and these have limited their utility and adoption. Illumina Complete Long Reads (ILCR) makes long-read sequencing accessible and streamlined for genomic labs. Additionally, this protocol enables generation of comprehensive human genomic data with combination of both long and short reads sequencing.

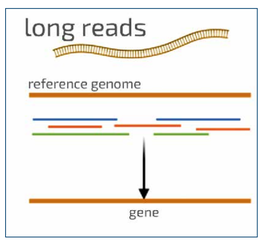
## Material and methods

**8 DNA samples**
DNA isolates from blood

**Quality control of DNA** (Supp. Table)
- Purity – NanoPhotometer
- Quantity – Qubit 2.0 (Qubit™ 1X dsDNA High Sensitivity (HS) Assay Kit)
- Quality – Agarose gel electrophoresis

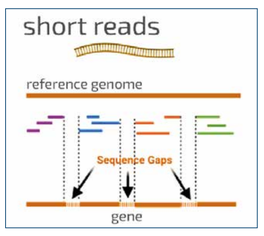
The concentration of DNA ranged from 63 to 142 ng/µl, based on Qubit measurement (Supp. Table). NGS libraries for short reads we prepared from these samples.
Before long-read library preparation, buffer from the original samples was exchanged for 30 µl of Resuspension Buffer (RSB) and another QC step was performed (Supp. Table).

### NGS library preparation




**1. Long Read**
Illumina Complete Long Read Prep Kit, Human (Illumina, USA) with Illumina Unique Dual Indexes (Illumina, USA)
**Input: 50 ng** (protocol recommendation)

Short read libraries were prepared from the same DNA samples.




**2. Short Read**
TruSeq DNA PCR-Free High Throughput Library Prep Kit (Illumina, USA) with IDT for Illumina – TruSeq DNA UD Indexes (Illumina, USA)
**Input: 1050 ng** (protocol recommendation)

### Sequencing
Illumina NovaSeq X Plus, 10B Reagent Kit

### Bioinformatic analysis
Sequence quality control, secondary analysis:
DRAGEN ICLR WGS
(Illumina Complete Long Reads v1.1.2)

Easy-to–use BaseSpace application for complete secondary analysis. hg38 human genome was used for reference.

### Workflow
1. Identification of marked sites in each read.
2. Location of read pairs that share marks.
3. Location of groups of reads from the same template.
4. Assembly of each group of marked reads.
5. Removal of long read marks.
6. Variant calling using long reads and unmarked short reads.

## Results

- Preparation of all 8 NGS libraries for long reads was successful (Supp. Table):
  - Concentration: 8.7–9.7 ng/µl; 14.1–15.4 nM
  - Average fragment length: 949 bp (932 to 985 bp) (Supp. Table, Figure 1)
  - No adapter dimers were recorded

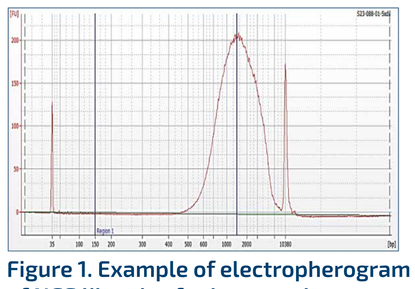


Figure 1. Example of electropherogram of NGS libraries for long reads.

- Preparation of all 8 NGS libraries for short reads was successful (Supp. Table):
  - Concentration: 4.0–11.7 ng/µl; 6.8–23.7 nM
  - Average fragment length: 795 bp (722 to 890 bp) (Supp. Table, Figure 2)
  - No adapter dimers were recorded

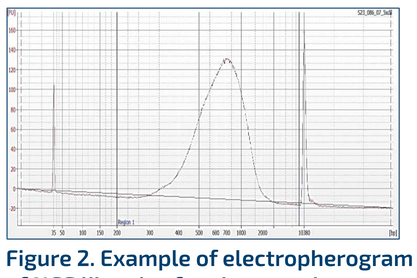


Figure 2. Example of electropherogram of NGS libraries for short reads.

| Short reads | average |
|---|---|
| Number of reads | 674 035 502 |
| Median coverage | 34.88 |
| Insert size average | 434.33 |
| Percentage duplicated reads | 17.80 |
| **ICLR** | |
| N50 | 4 679 |
| Total_input_reads (raw) | 4 932 565 534 |
| Number of reads (assembled) | 23 845 198.00 |
| Insert size average (raw) | 548.88 |
| Percentage duplicated reads (raw) | 22.01 |
| Median coverage | 32.38 |
| Percentage of reads mapped > 10kb | 3.45 |
| Percentage of bases in reads mapped > 10kb | 11.03 |
| **Combination short reads + ICLR** | |
| Number mapped reads | 693 165 348 |
| Median coverage | 67.5 |
| **VCF metrics** | |
| SNPs | 4 001 872 |
| Insertions | 461 168 |
| Deletions | 468 222 |
| Indels | 23 877 |
| Total_Het/Hom_ratio | 1.41 |
| Insertion/Deletion_ratio | 0.98 |
| Total_SVs | 28 171 |
| **Phasing statistics** | |
| Number biallelic. het variants phased | 2 720 055 |
| Percentage biallelic het variants phased | 98.09 |
| Number of phase blocks | 41 200 |
| Phase block N50 | 153 342 |

Table 1. Selected sequencing and analysis metrics. Average values from 8 samples are presented.

## Conclusion

We successfully prepared, sequenced and analyzed blood samples using Illumina Complete Long Reads technology. All samples achieved the required sequencing output and their QC parameters matched the manufacturer's specification. The average N50 value achieved for our samples was 4.6 kb for long reads. More than 11 % of data came from reads longer than 10 kb. Median coverage was 67.5 at average (short and long reads combined) and phased N50 value was more than 153 kb. The adoption of long-read technology has a potential to improve the detection of complex structural variants, which may be problematic to be determined by short reads; such as large inversions, deletions or translocations. Our data suggest a benefit in calling these types of structural variants.

# Novel long-read sequencing technology in practise

Martin Kašný [1*], Alžběta Hamplová [1], Karolína Stivínová [1], Eliška Hladíková [1], Nela Chalupníková [1], Ondřej Brzoň [1], Petr Kvapil [1]

[1] Institute of Applied Biotechnologies, Služeb 3056/4, 108 00 Prague, Czech Republic
* Correspondence: kasny@iabio.eu; Tel.: +420 739 394 364

| Sample ID | DNA in | NanoPhotometer Results Qubit Results | | | Volume [µl] | DNA Amount [ng] | Library Preparation | | | Qubit Results | | Bioanalyzer 2100 Results | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A260/280 | c [ng/µl] | c [ng/µl] | | | Type | Input [µl] | Input [ng] | c [ng/µl] | c [nM] | c [ng/µl] | Average Fragment Lenght [bp] |
| 1 | RSB | 1.7 | 12.5 | 13.5 | 30 | 404 | Long Read | 3.7 | 50 | 9.2 | 14.1 | 4.4 | 985 |
| 2 | RSB | 2.0 | 20.0 | 19.0 | 30 | 569 | Long Read | 2.6 | 50 | 8.7 | 14.1 | 4.2 | 939 |
| 3 | RSB | 2.0 | 20.0 | 23.6 | 30 | 708 | Long Read | 2.1 | 50 | 9.3 | 15.0 | 4.5 | 935 |
| 4 | RSB | 2.0 | 25.0 | 30.9 | 30 | 927 | Long Read | 1.6 | 50 | 9.7 | 15.4 | 4.5 | 957 |
| 5 | RSB | 1.7 | 12.5 | 16.9 | 30 | 506 | Long Read | 3.0 | 50 | 9.5 | 15.2 | 4.2 | 950 |
| 6 | RSB | 1.8 | 27.5 | 39.1 | 30 | 1173 | Long Read | 1.3 | 50 | 9.4 | 15.2 | 4.8 | 932 |
| 7 | RSB | 1.8 | 90.0 | 98.2 | 30 | 2946 | Long Read | 3.1 | 50 | 9.7 | 15.4 | 4.8 | 948 |
| 8 | RSB | 1.8 | 27.5 | 34.7 | 30 | 1041 | Long Read | 1.4 | 50 | 8.9 | 14.3 | 5.0 | 945 |
| 1 | water | 1.8 | 128.0 | 67.5 | 30 | 2025 | Short Read | 15.6 | 1050 | 7.2 | 13.5 | 7.3 | 806 |
| 2 | water | 1.8 | 135.0 | 74.6 | 33 | 2462 | Short Read | 14.1 | 1050 | 7.4 | 14.7 | 8.5 | 769 |
| 3 | water | 1.7 | 120.0 | 62.0 | 35 | 2170 | Short Read | 16.9 | 1050 | 6.8 | 12.7 | 7.9 | 809 |
| 4 | water | 1.7 | 118.0 | 62.6 | 48 | 3005 | Short Read | 16.8 | 1050 | 6.5 | 11.8 | 6.8 | 829 |
| 5 | water | 1.8 | 45.0 | 67.4 | 20 | 1348 | Short Read | 15.6 | 1050 | 6.0 | 11.4 | 6.7 | 792 |
| 6 | water | 1.6 | 70.0 | 86.9 | 20 | 1738 | Short Read | 12.1 | 1050 | 4.0 | 6.8 | 4.1 | 890 |
| 7 | water | 1.9 | 218.0 | 142.0 | 20 | 2840 | Short Read | 7.4 | 1050 | 11.7 | 23.7 | 8.1 | 746 |
| 8 | water | 1.8 | 95.0 | 71.0 | 20 | 1420 | Short Read | 14.8 | 1050 | 6.1 | 12.7 | 5.7 | 722 |

**Supplementary table. The selected quantitative and qualitative metric related to the DNA samples and NGS libraries.**