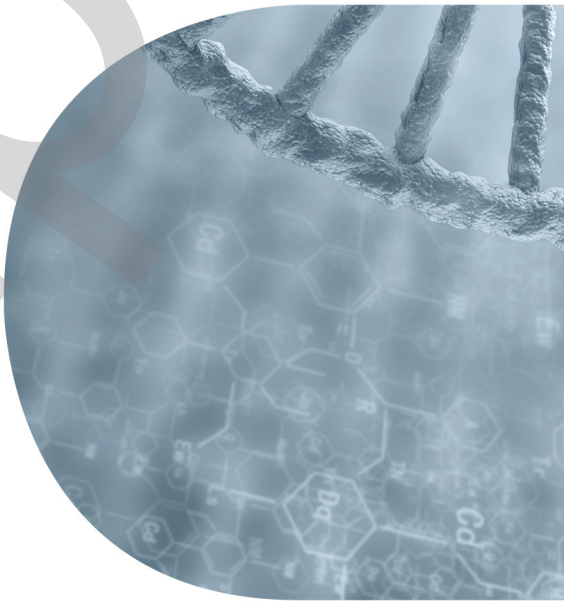
Research use only

Oxford Nanopore
De novo assembly
Report

November 2020



Humanizing Genomics
macrogen

Project Information

Client Name	Macrogen
Company / Institution	Macrogen
Order Number	HN12345678
Sample	Sample_nanopore
Type of Analysis	De novo assembly
Type of Sequencer	Oxford Nanopore

Table of Contents

1. Data Download

Download link	File size	md5sum
Sample_nanopore.fast5.tar	9.0G	048c2c9ee17dd417867826d0925ae1b2
Sample_nanopore.fastq.gz	7.82G	e9769de9b673fc5cdae248f2d3efb30f
Analysis Results	260M	4009a64f9e52a47eaaca01ceda52a3bf

- *.fast5.tar.gz : Compressed file of both raw signal data and base-called information.
- *.fastq.gz : Gzip compressed file of raw data used in analysis.
- md5sum : In order to verify the integrity of files, md5sum is used. If the values of md5sum are the same, there is no forgery, modification or omission.

Your data will be retained in our server for 3 months. Should you wish to extend the retention period, please contact us.

2. Sequencing and Analysis Workflow

2. 1. Experiment Overview

Figure 1. Experiment overview

1) Sample Preparation

For library construction, DNA/RNA is extracted from a sample. After performing quality control (QC), qualified samples proceed to library construction.

2) Library Construction

First, DNA was isolated then it was either fragmented or not. Next, the ends of DNA fragments were repaired by 5' end phosphorylation and dA-tailing. Sequencing adapters were ligated to DNA ends for sequencing. For barcoded library construction, the native barcodes were ligated to DNA ends first then the sequencing adapters were subsequently ligated to the DNA ends for sequencing.

3) Sequencing

For sequencing, the constructed library was inserted into the flow cell where the nanopores are embedded in membrane and immersed in an electrolyte solution. Nanopores are biological pores that are specifically designed by ONT. When electrical potential is applied across the membrane, the ionic current is generated through each nanopore. During the sequencing process, the motor protein that is ligated to the library binds to the nanopore. Then, motor protein attached strand passed through nanopore and cause disruption in current. The disruption patterns are unique for each base (G, A, T and C). These unique patterns were collected as sequencing data.

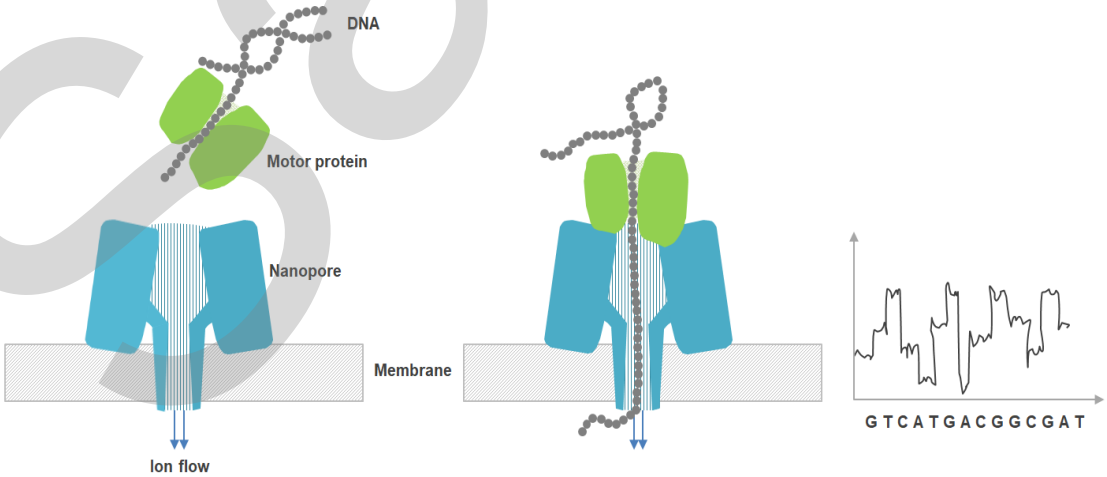


Figure 2. Schematic workflow of Oxford Nanopore sequencing

2. 2. Generation of Raw Data

The Nanopore sequencer generates the electronic raw signal data. They are converted into FAST5 (HDF5) files and FASTQ files using the neural network based basecaller called Guppy. FAST5 files contain raw signal data that can be used for basecalling. FASTQ files are generated after trimming adapters from sequenced reads using the software, Porechop.

2. 3. Analysis Overview

1) De novo assembly

De novo assembly is performed using the trimmed data. Assembler having algorithms enable to decrease assembly noise due to read errors was used for the analysis. And then, the consensus sequences generated from assembler are polished because assemblies using Nanopore reads usually have low base-level quality.

2) Assembly validation

The assembled genome is validated using BUSCO analysis. BUSCO analysis is performed to evaluate genome assemblies based on evolutionarily-informed expectations of gene contents.

3. Summary of Data Production

3. 1. Raw Data Statistics

The total number of bases, reads, N50 value, average length and quality were calculated for the Sample_nanopore. Totally, 165 reads were produced, and the total bases are 0.00Gbp. Porechop and NanoLyse were used to trim adapter sequences and remove lambda DNA.

Table 1. Raw Data Stats

Sample Name	Total Read Bases (bp)	Total Reads	N50	Average Length	Quality
Sample_nanopore	371,479	165	2864.0	2,251	9.5

- Total Read Bases : The number of bases sequenced.
- Total Reads : The number of reads produced.
- N50 : 50% of all bases come from reads equal to or longer than this value.
- Average Length : The average length of the produced reads.
- Quality : Phred Quality score. Quality scores of ONT data usually range from 8-15.

3. 2. Read Length versus Average Read Quality

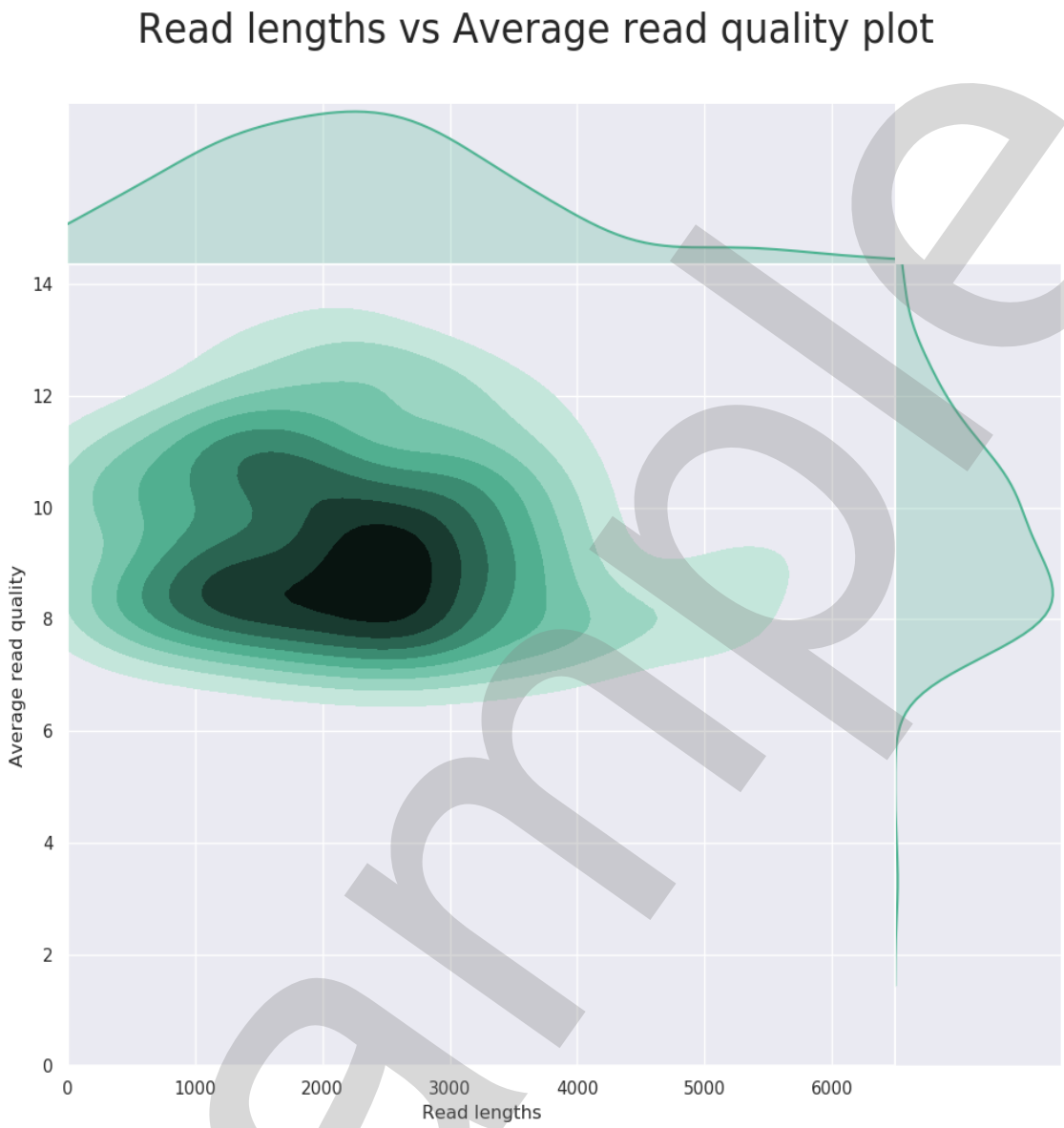


Figure 3. Read Length vs Average Read Quality : Sample_nanopore

4. Analysis Results

4. 1. De novo Assembly

De novo assembly was performed by Shasta. After then, the consensus sequences are polished by HELEN. The assembly results are summarized in the table below.

Table 2. Summary of assembly

Contigs	Total Contig Bases	N50	Max Length	Min Length	Mean Length
1,234	890,123,456	3,456,789	23,456,789	345	678,901

- Contigs : The number of contigs assembled.
- Total contig bases : The total length of contigs.
- N50 : 50% of all bases come from reads equal to or longer than this value.
- Max length : The length of longest contig.
- Min length : The length of shortest contig.
- Mean length : The average length of contigs assembled.

5. Assembly Validation

5. 1. BLAST Results

After complete genome or draft genome was assembled, BLAST analysis was carried out to identify to which species each scaffold show similarity. Best hit and top 5 hit results were identified using NCBI NT database. Each result was prepared separately by the sheet of excel. Following is the example.

	Query					Subject					Score		Identities		Gap	
Name	Q_Leng	Q_Start	Q_End	Q_Cove	Description	Accession	S_Leng	S_Start	S_End	S_Cove	Bit	E-value	I_Matcl	I_Pct.(%	G_Matc	G_Pct.(%
contig1	4531255	121	367596	5.98	Escherichia coli str. K-12 substr. MG1655	CP003291.1	4617381	99334	181243	6.08	151115	0.0	115657/11	96	91/119601	0
contig2	55965	7321	463	7.7	Lactobacillus paracasei N1115 plasmid, complete sequence	CP007126.1	8755	7321	463	7.77	6867	0.0	55285/573	96	99/57338	0
contig3	31102	21999	22654	2.11	Streptomyces chartreusis strain WZS021 chromosome 7	CP028498.1	1402960	301599	300926	0.05	481	8e-130	565/700	80	70/700	10

Figure 4. Best hit result example

	Query					Subject					Score		Identities		Gap	
Name	Q_Leng	Q_Start	Q_End	Q_Cove	Description	Accession	S_Leng	S_Start	S_End	S_Cove	Bit	E-value	I_Matcl	I_Pct.(%	G_Matc	G_Pct.(%
contig1	4531255	121	367596	5.98	Escherichia coli str. K-12 substr. MG1655	CP003291.1	4617381	99334	181243	6.08	151115	0.0	115657/11	96	91/119601	0
contig1	4531255	11446	367603	5.6	Escherichia coli O157:H7 str. Sakai	CU928148.1	5064201	441603	520377	5.43	133154	0.0	108685/11	96	47/112047	2
contig1	4531255	1749458	1851600	5.11	Escherichia coli O104:H4 str. 2011C-3493	CP019302.1	4841212	1555277	1623489	4.82	84236	0.0	98848/102	96	75/102171	4
contig1	4531255	1857455	14560	5.11	Escherichia coli O83:H1 str. NRG 857C	CP001855.1	5119790	1855901	1921622	4.82	83837	0.0	98844/102	96	79/102171	6
contig1	4531255	3644512	1235621	4.31	Escherichia coli IAI39	CU928164.2	5131046	1097802	1163163	4.04	81923	0.0	83481/861	96	57/86182	0
contig2	55965	7321	463	7.7	Lactobacillus paracasei N1115 plasmid, complete sequence	CP007126.1	8755	7321	463	7.77	6867	0.0	55285/573	96	99/57338	0
contig2	55965	8755	8049	7.25	Lactobacillus paracasei N1115 plasmid, complete sequence	CP007126.1	8755	8755	8049	7.89	715	0.0	52195/539	96	62/53948	0
contig2	55965	1246426	1239620	7.25	Lactobacillus kefiranofaciens ZW3, complete genome	CP002764.1	2113023	1246426	1239620	7.16	6820	0.0	52201/539	96	101/53967	0
contig2	55965	1907493	1909660	7.29	Lactobacillus kefiranofaciens ZW3, complete genome	CP002764.1	2113023	1907493	1909660	9.03	2179	0.0	52111/542	96	78/54256	0
contig2	55965	1916443	1915335	6.04	Lactobacillus kefiranofaciens ZW3, complete genome	CP002764.1	2113023	1916443	1915335	7.47	1109	0.0	43152/448	96	59/44896	0
contig3	31102	21999	22654	2.11	Streptomyces chartreusis strain WZS021 chromosome 7	CP028498.1	1402960	301599	300926	0.05	481	8e-130	565/700	80	70/700	10
contig3	31102	21999	22654	2.11	CP026121.1 Streptomyces sp. Go-475 chromosome, comple	CP026121.1	8570609	2528836	2526237	0.03	929	0.0	2091/2811	74	296/2811	10
contig3	31102	29192	30719	4.97	CP026652.1 Streptomyces sp. XZHG99 chromosome, compl	CP026652.1	8541354	354581	353109	0.02	1037	0.0	1245/1558	79	115/1558	7
contig3	31102	1260	1565	1.02	CP011799.1 Streptomyces sp. PBH53 genome	CP011799.1	9153597	3057639	3057340	0	111	1e-18	241/322	74	38/322	11
contig3	31102	20233	25764	18.43	CP022744.1 Streptomyces lincolnensis strain LC-G chromos	CP022744.1	9513637	9239548	9234370	0.05	2549	0.0	4388/5713	76	715/5713	12

Figure 5. Top 5 hit result example

Because the BLAST analysis is based on registered information, it is difficult to determine the information of the species that is not registered. In particular, the assembly results could be matched with a relative species or an evolutionarily distant species due to sequence differences or error that may occur during the assembly process. Therefore, it would be more appropriate to use the analysis results to identify patterns rather than to use it as an absolute criterion for species determination. The BLAST results are in the "Analysis Result" file.

5. 2. BUSCO Results

In order to assess the completeness of the genome assembly, BUSCO analysis was performed based on evolutionarily-informed expectations of gene content from near-universal single-copy orthologs.

The recovered matches are classified as 'Complete' if their lengths are within the expectation of the BUSCO profile match lengths. If these are found more than once, they are classified as 'duplicated'. The matches that are only partially recovered are classified as 'Fragmented', and BUSCO groups for which there are no matches that pass the tests of orthology are classified as 'Missing'.

Higher complete BUSCOs may indicate good assembly. However for species other than model organisms, relatively low BUSCOs can be appears due to characteristics of the sample as well as the incompleteness of the assembly.

By default, bacteria or eukaryota DB was used for analysis.

Table 3. BUSCO analysis result

Used Lineage : eukaryota_odb9 (number of species: 100, number of BUSCOs: 303)

Status	# of BUSCOs	Percentage
Complete BUSCOs (C)		
Complete and single-copy BUSCOs (S)	81	26.73 %
Complete and duplicated BUSCOs (D)	189	62.38 %
Fragmented BUSCOs (F)	5	1.65 %
Missing BUSCOs (M)	28	9.24 %
Total BUSCO groups searched	303	100.00 %

- Status : A quantitative assessment list of the completeness in terms of expected gene content
  The following tow conditions are used t ocreate a status:
  a. Expected range of scores
  b. Expected range of length alignments
  If both conditions are met, it is classified as Complete (These complete busco matches are either single-copy or duplicated). If length alignments is not met, it is classified as Fragmented. If both conditions are not met, it is classified as Missing
- # of BUSCOs : Identified count in sample.
- Percentage : Identified percentage in sample.

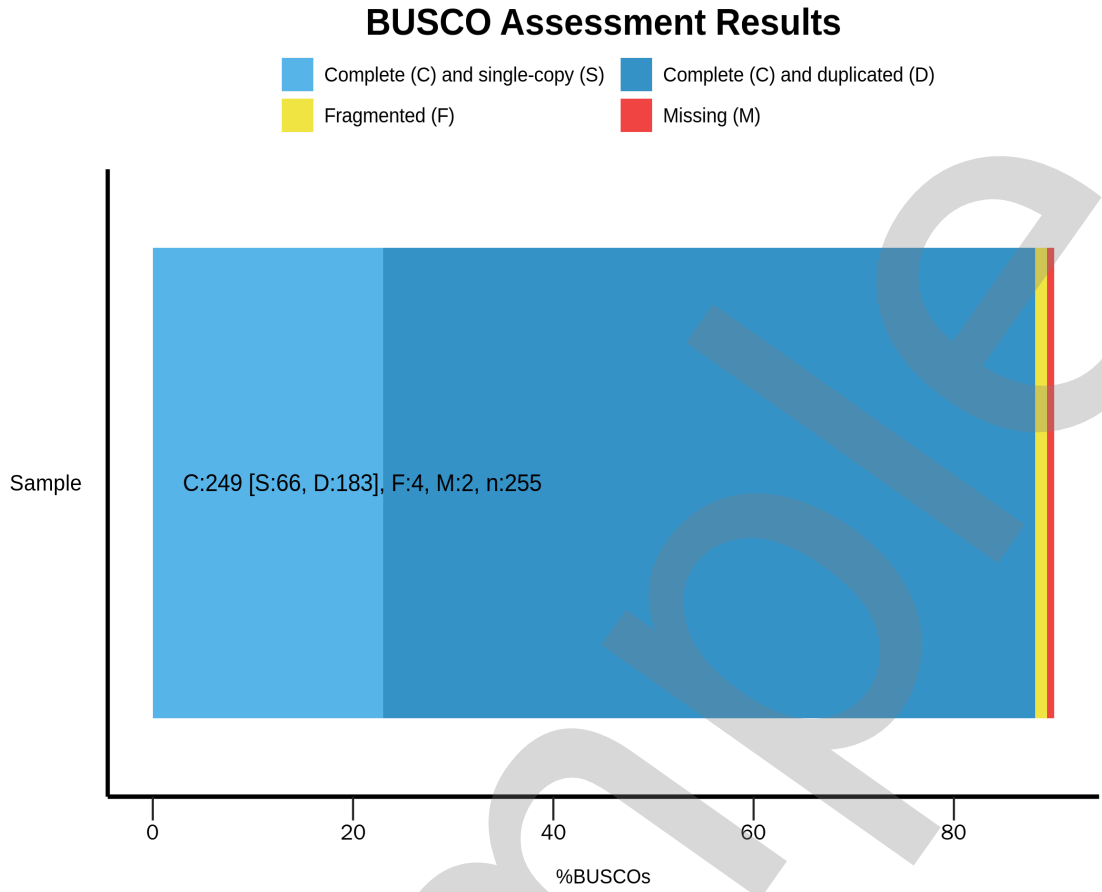


Figure 6. BUSCO result plot

6. Details of File Extensions

6. 1. Raw Data

File Extensions	Description
*.fast5.tar	A variant of the HDF5 file format, containing both raw signal and base-called information obtained from the nanopore. One fast5 file is produced per read and compressed with tar format.
*.fastq.gz	The files contain gzip compressed reads sequence in FASTQ format.

6. 2. Analysis Results

File Extensions	Description
*contigs.fasta	Sequences of assembled contigs.
*_BLAST.xlsx	BLAST analysis result against NCBI NT database.

7. Appendix

7. 1. Glossary of Terms

- **Fast5**

  Fast5 is the native raw data format of ONT sequencer. Being fundamentally a HDF5 file, it allows flexible storing of various data types. Pre-basecalled Fast5 files contain only the raw electric signal data produced by the sequencer. These raw signals can be translated into actual base information using a basecaller, and produce a post-basecalled Fast5 format (which now contains both signal and sequence data).

- **Nanopore**

  Nanopore is a nanometer-sized pore. These pores can either be embedded in a biological membrane. In ONT devices, an ionic current is passed through the nanopores and the changes in current as biological molecules pass through the nanopore are measured to obtain information about the nucleobase passing through the pore at a given time.

7. 2. FAQ

Q : I would like to see the produced data. How can I open those files?

A : After unzipping the file, the data can be opend with any kind of text editor, however, large volume zip file that is provided by our company is not user-friendly in Windows environment, so it is recommended to use Linux environment for smooth operation.
(Using Vim (https://www.vim.org/) or Notepad++ (http://notepad-plus-plus.org/))

7. 3. Phred Quality Score Chart

Phred quality score numerically expresses the accuracy of each nucleotide. Higher Q number signifies higher accuracy. For example, if Phred assigns a quality score of 30 to a base, the chances of having base call error are 1 in 1000.
Phred Quality Score Q is calculated with -10logP, where P is probability of erroneous base call.

Quality of phred score	Probability of incorrect base call	Base call accuracy	Characters
10	1 in 10	90%	!"#$%&'()*+
20	1 in 100	99%	,-./012345
30	1 in 1000	99.9%	6789:;h=i?
40	1 in 10000	99.99%	@ABCDEFGHIJ

Encoding: Sanger Quality (ASCII Character Code=Phred Quality Value + 33)

7. 4. Programs Used in Anlysis

7. 4. 1. Porechop

LINK https://github.com/rrwick/Porechop

Porechop (v0.2.4) is a tool for trimming adapters from Nanopore reads. The adapters at the end of the reads are trimmed and the adapters at the middle of the reads are considered as chimeric and chopped into separate reads.

7. 4. 2. NanoLyse

LINK https://github.com/wdecoster/nanolyse

NanoLyse (v1.1.0) remove lambda phage reads from raw data using Minimap2 aligner. Lambda phage is a control DNA fragment supplied by Oxford Nanopore Technologies (ONT).
Caution : This approach can lead to loss of reads that is similar to the lambda phage genome.

Citation

Wouter De Coster, Svenn D'Hert, Darrin T Schultz, Marc Cruts, Christine Van Broeckhoven, NanoPack: visualizing and processing long-read sequencing data, Bioinformatics, Volume 34, Issue 15, 01 August 2018, Pages 2666-2669, https://doi.org/10.1093/bioinformatics/bty149

7. 4. 3. Shasta

LINK https://github.com/chanzuckerberg/shasta

Shasta (v0.4.0) is an assembler for Oxford Nanopore Techonologies (ONT) long reads. Shasta applied Run-Length Encoding which makes reads as homopolymer-compressed (HPC) form. It can solved homopolymer error of ONT long reads and improve assembly quality due to significantly higher identity alignments between reads.

Citation

Shafin, K., Pesout, T., Lorig-Roach, R. et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol 38, 1044-1053 (2020). https://doi.org/10.1038/s41587-020-0503-6

7. 4. 4. MarginPolish and HELEN

LINK https://github.com/kishwarshafin/helen

MarginPolish (v1.3.0) and HELEN (v0.0.22) are deep neural network-based polishing tool for Nanopore assembly. Based on weighted graph generated from MarginPolish, HELEN corrects the Nanopore assembly base. Default option is used for this anlaysis.

7. 4. 5. BLAST

LINK https://blast.ncbi.nlm.nih.gov/Blast.cgi

The Basic Local Alignment Search Tool (BLAST, v2.7.1+) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Citation

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.

7. 4. 6. BUSCO

LINK https://busco.ezlab.org/

BUSCO (v3.0.2) evaluates genome assemblies based on evolutionarily-informed expectiations of gene content from near-universal single-copy orthologs selected from OrthoDB v9.

Citation

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210-3212.

Research use only

HEADQUARTER

Macrogen, Inc.

Laboratory, IT and Business Headquarter & Support Center

[08511] 1001, 10F, 254, Beotkkot-ro, Geumcheon-gu, Seoul, Republic of Korea (Gasan-dong, World Meridian 1)
Tel: +82-2-2180-7000
Email1: ngs@macrogen.com(Overseas)
Email2: ngskr@macrogen.com (Republic of Korea)
Web: www.macrogen.com
LIMS: dna.macrogen.com

SUBSIDIARY

Macrogen Europe

Laboratory, Business & Support Center

Meibergdreef 31, 1105 AZ, Amsterdam, the Netherlands
Tel: +31-20-333-7563
Email: ngs@macrogen.eu

Macrogen Singapore

Laboratory, Business & Support Center

3 Biopolis Drive #05-18, Synapse, Singapore 138623
Tel: +65-6339-0927
Email: info-sg@macrogen.com

Psomagen (Macrogen USA)

Laboratory, Business & Support Center

1330 Piccard Drive, Suite 103, Rockville, MD 20850, United States
Tel: +1-301-251-1007
Email: inquiry@psomagen.com

Macrogen Japan

Laboratory, Business & Support Center

16F Time24 Building, 2-4-32 Aomi, Koto-ku, Tokyo 135-0064 JAPAN
Tel: +81-3-5962-1124
Email: ngs@macrogen-japan.co.jp

BRANCH

Macrogen Spain

Laboratory, Business & Support Center

Av. Sur del Aeropuerto de Barajas, 28. Office B-2, 28042 Madrid, Spain
Tel: +34-911-138-378
Email: info-spain@macrogen.com

Humanizing Genomics
macrogen