

# 1 **PGP1 personal genome assembly - a hybrid assembly dataset using** 2 **ONT's PromethION and PacBio's HiFi sequencing**

## 3 **Authors:**

4 Hui-Su Kim<sup>1</sup>, Changjae Kim<sup>2</sup>, George Church<sup>3</sup>, and Jong Bhak<sup>1,2</sup>

5 <sup>1</sup>Korean Genomics Center (KOGIC), Ulsan National Institute of Science and Technology  
6 (UNIST), Ulsan, Ulju-gun, Eonyang-eup, 44919, Republic of Korea

7 <sup>2</sup>Clinomics LTD, Ulsan National Institute of Science and Technology (UNIST), Ulsan, Ulju-  
8 gun, Eonyang-eup, 44919, Republic of Korea

9 <sup>3</sup>Department of Genetics, New Research Building (NRB), 77 Avenue Louis Pasteur, Boston,  
10 MA 02115 USA

11

12 Corresponding author:

13 Jong Bhak<sup>1,2</sup>

14 50 UNIST-gil, Ulsan, Ulju-gun, Eonyang-eup, 44919, Republic of Korea

15 Email address [jongbhak@genomics.org](mailto:jongbhak@genomics.org)

## 16 **Abstract**

17 PGP1 is the first participant of Personal Genome Project. We present the PGP1's chromosome-  
18 scale genome assembly. It was constructed using 255 Gb ultra-long PromethION reads and 97  
19 Gb short paired-end reads. For reducing base calling errors, we corrected PromethION reads  
20 using 72 Gb PacBio HiFi reads. 327 Gb Hi-C chromosomal mapping data were utilized to  
21 maximize the assembly's contiguity. PGP1's contig assembly was 3.01 Gb in length comprising  
22 of 4,234 contigs with an N50 value of 33.8 Mb. After scaffolding with Hi-C data and extensive  
23 manual curation, we obtained a chromosome-scale assembly that represents 3,880 scaffolds with  
24 an N50 value of 142 Mb. From the Merqury assessment, PGP1 assembly achieved a high QV  
25 score of Q45.45. For a gene annotation, we predicted 106,789 genes with a liftover from the  
26 Gencode 38 and an assembly of transcriptome data.

27

## 28 **Keywords**

29 PGP1 genome, Long-read sequencing, Human genome assembly, ONT, PacBio, Hi-C

30

## 31 Specifications Table

32

<b>Subject</b>	Biology
<b>Specific subject area</b>	Genomics
<b>Type of data</b>	Sequencing raw reads, Assembly, Tables, Figure
<b>How data were acquired</b>	PromethION flow-cell R9.4.1 (Oxford Nanopore Technologies) PacBio HiFi (Pacific Bioscience) NovaSeq (Illumina) Hi-C (Arima-Genomics)
<b>Data format</b>	Raw reads (fastq), Assembly (fasta), Protein and Transcript sequences (fasta), Genome annotation (gff3)
<b>Parameters for data collection</b>	DNA from the PGP1 cell line used for library preparation and sequencing.
<b>Description of data collection</b>	Total genomics DNA extraction was performed using DNeasy Blood & Tissue Kit from QIAGEN. The library construction and whole genome sequencing were performed using Illumina's NoveSeq (100bp x2, short reads), ONT's PromethION (long reads), and Illumina's NoveSeq platform (151bp x2, Hi-C).
<b>Data source location</b>	Institution: Korean Genomics Center (KOGIC) City/Town/Region: Ulsan city Country: Korea, republic of
<b>Data accessibility</b>	Raw data was deposited in the NCBI database under BioProject: PRJNA734849 ( <a href="https://www.ncbi.nlm.nih.gov/bioproject/PRJNA734849">https://www.ncbi.nlm.nih.gov/bioproject/PRJNA734849</a> ), SRA: SRX11055733, SRX11055734, SRX11055735, SRX11055736, SRX11055737, SRX11055738 ( <a href="https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&amp;from_uid=734849">https://www.ncbi.nlm.nih.gov/sra?linkname=bioproject_sra_all&amp;from_uid=734849</a> ). Its description, the final assembly and data information are also found at <a href="http://genomics.org/PGP1">http://genomics.org/PGP1</a> .

33

## 34 Value of the Data

- 35 • This is a *de novo* genome assembly of PGP1 of Personal  
36 Genome Project of Harvard Medical school.
- 37 • The genome assembly of a male Caucasian and sequencing data add to current genomic  
38 representation of North-Eastern Europeans and is a useful resource for further in-depth  
39 analyses of European genomic structure and diversity in higher resolution.
- 40 • We share a hybrid assembly pipeline used in this study for constructing a high-quality  
41 chromosome-scale assembly from PromethION, PacBio-HiFi, and Hi-C data which can be a  
42 useful approach for the bioinformatics community specializing in genome assembly.

## 44 1. Data description

45 We generated sequencing data of long-reads by ONT PromethION, short-reads by Illumina  
46 NovaSeq and Hi-C reads (Table 1). The sequencing data have been deposited in the NCBI  
47 database under BioProject: PRJNA734849. PGP1's description is found also at  
48 <http://genomics.org/PGP1>. We collected PacBio HiFi reads from NCBI SRA accession  
49 SRX7671688. The *de novo* assembly of PGP1 genome is at chromosome-scale with a total  
50 length of 3.02 Gb, which consists of 3,880 scaffolds with 24 chromosomes and unplaced  
51 sequences (Table 2). Detailed features of the genome annotation are described in table 3. The  
52 sequence data and description are available at <http://genomics.org/PGP1>.

53 **Table 1. Statistics of long and short reads whole genome sequencing for PGP1**

Library type	Sequencing techs.	Library name	No. of reads	Total length of reads (bp)	N50 (bp)
Long reads	ONT PromethION	PGP1_PT	19,538,795	254,994,082,784	23,147
	PacBio HiFi	PGP1_PBCCS.Q20	5,701,695	71,831,314,346	12,947
Short reads	Illumina NovaSeq	PGP1_PE500	715,404,966	96,846,095,400	135
Hi-C	Illumina NovaSeq	PGP1_HiC	2,166,523,472	327,145,044,272	151

54

55 **Table 2. Statistics of PGP1 assembly**

	Contig assembly	Chromosome-scale assembly
Contigs No.	4,234	3,880
Total length (bp)	3,015,852,063	3,016,802,955
N50 (bp)	33,790,496	141,933,136
Max contig length (bp)	110,121,243	236,082,540
Gap	0.004%	0.035%
GC contents	40.87%	40.87%
QV (from Merqury)		45.4982
Error rate (from Merqury)		0.0000282

56

### 57 **Table 3. PGP1 genome annotation**

PGP1 gene	
Transcripts No.	106,789
Total length of transcripts (bp)	209,078,868
N50 (bp)	3,358
Max transcript length (bp)	95,488
Gap	0.000%
GC contents	48.75%

58

## 59 **2. Experimental Design, Materials, and Methods**

### 60 2.1. Sample preparation and whole-genome sequencing

61 DNA was extracted from samples from the PGP1 cell line from Coriell. For short-read  
62 sequencing, a 135 bp library was constructed, and the sequencing was conducted by Illumina's  
63 NovaSeq platform. For long-read sequencing, we constructed libraries using the 1D ligation  
64 sequencing kit (SQK-LSK109), and the sequencing data was generated using ONT's  
65 PromethION R9.4.1 platform. Base-calling was carried out using Guppy v3.5.4 with the Flip-  
66 Flop hac model. Libraries for Hi-C, the chromosome conformation capture data were generated

67 using the Arima-Hi-C kit. The sequencing of Hi-C was performed using Illumina's NovaSeq  
68 sequencer with a read length of 150 bp by Novogene.

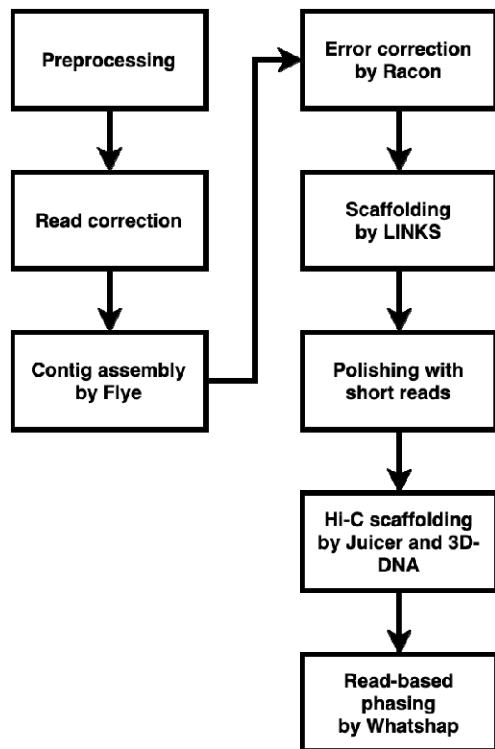
69

## 70 2.2. Read preprocessing and whole genome assembly

71 Procedures are described in figure 1. Trimming adapter sequences and low-quality sequences  
72 in short reads were performed using Trimmomatic v0.36[1]. We used tadpole.sh program of  
73 BBtools suite v38.96 (<https://sourceforge.net/projects/bbmap>) for an error correction. Adapter  
74 sequences in PromethION reads were removed using Porechop v.0.2.4  
75 (<https://github.com/rrwick/Porechop>), and we corrected PromethION reads against PacBio HiFi  
76 reads using Racon v1.4.3 program (<https://github.com/isovic/racon>).

77 *A de novo* assembly was performed using Flye v2.5[2] program. Correcting base-errors in  
78 assembled contigs was conducted using Racon, and an extension of contigs using ultra-long  
79 PromethION reads was carried out using LINKS v1.8.7 (<https://github.com/bcgsc/LINKS>).  
80 Polishing the assembled contigs with short reads was performed using Pilon v1.23[3] twice.

81 For generating a chromosome-scale assembly, scaffolding contigs with Hi-C was performed  
82 using Juicer v1.5[4] and 3D-DNA pipeline[5]. For correcting mis-assemblies in the scaffolds, we  
83 used JBAT v1.11.08 program ([https://github.com/aidenlab/Juicebox/wiki/Juicebox-Assembly-  
84 Tools](https://github.com/aidenlab/Juicebox/wiki/Juicebox-Assembly-Tools)) and corrected them manually. An assessment of PGP1 genome assembly was carried out  
85 using Merqury v1.3 program[6].



86

87 Fig. 1. A bioinformatics pipeline for PGP1 genome assembly.

88

### 89 2.3. Read-based phasing and genome annotation

90 A read-based phasing of the assembly was performed using DeepVariant v1.1.0  
91 (<https://github.com/google/deepvariant>) and WhatsHap v1.0[7], and we generated phased  
92 genome sequences from the phased variant-information using Bcftools v1.9  
93 (<http://github.com/samtools/bcftools>). For gene annotation, a liftover of an annotated gene set  
94 from Gencode release 38 (<https://www.encodegenes.org/human/>) using Liftoff v1.6.1[8] and a  
95 reference-guided transcriptome assembly using Stringtie v2.1.5 program[9] were conducted. The  
96 RNASeq data was obtained from SRA no. SRX683721, SRX683722, SRX683723.

97

## 98 **Declarations**

## 99 **Ethics Statement**

100 This study was a part of Korean Personal Genome Project (KPGP also known as PGP-Korea)  
101 and was approved by the Institutional Review Board at Genome Research Foundation with IRB-  
102 REC- 20101202 – 001. The anonymous sample donor has signed a written informed consent to  
103 participate in the whole genome sequencing and following analysis in compliance with the  
104 Declaration of Helsinki.

105

## 106 **Consent for publication**

107 The (KPGP) informed consent included a section about data publication, which was consented to.

108

109 **Competing interest:** C. K. is an employee in Clinomics Inc., where J.B. is a founder and  
110 a CEO of Clinomics USA and Clinomics Inc., Korea. They have an equity interest in the  
111 company. All other authors declare they have no competing interests.

112

113

## 114 **CRedit author statement**

115 **Hui-Su Kim:** Methodology, Formal analysis, Writing - Original Draft, Data Curation,  
116 Visualization, and Editing. **Jong Bhak:** Funding acquisition, Project administration, Supervision,  
117 Resources, Conceptualization, and Writing - Review & Editing. **Changjae Kim:** Performed



118 DNA preparation, sequencing, and data quality checking. **George Church:** Initiated and  
119 supervised PGP, provided cell-line. All authors read and approved the finalized manuscript.

120

## 121 **Acknowledgements**

122 We thank GenomeLab, PGI of GRF, and KOGIC members for providing technical assistance  
123 and discussions. We also thank the Korea Institute of Science and Technology Information  
124 (KISTI) that provided us with the Korea Research Environment Open NETwork (KREONET).

125

126

## 127 **References**

- 128 1. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina*  
129 *sequence data*. *Bioinformatics*, 2014. **30**(15): p. 2114-20.
- 130 2. Kolmogorov, M., et al., *Assembly of long, error-prone reads using repeat graphs*. *Nature*  
131 *Biotechnology*, 2019. **37**(5): p. 540-+.
- 132 3. Walker, B.J., et al., *Pilon: An Integrated Tool for Comprehensive Microbial Variant*  
133 *Detection and Genome Assembly Improvement*. *Plos One*, 2014. **9**(11).
- 134 4. Durand, N.C., et al., *Juicer Provides a One-Click System for Analyzing Loop-Resolution*  
135 *Hi-C Experiments*. *Cell Systems*, 2016. **3**(1): p. 95-98.
- 136 5. Dudchenko, O., et al., *De novo assembly of the Aedes aegypti genome using Hi-C yields*  
137 *chromosome-length scaffolds*. *Science*, 2017. **356**(6333): p. 92-95.
- 138 6. Rhie, A., et al., *Merqury: reference-free quality, completeness, and phasing assessment*  
139 *for genome assemblies*. *Genome Biol*, 2020. **21**(1): p. 245.
- 140 7. Patterson, M., et al., *WhatsHap: Weighted Haplotype Assembly for Future-Generation*  
141 *Sequencing Reads*. *J Comput Biol*, 2015. **22**(6): p. 498-509.
- 142 8. Shumate, A. and S.L. Salzberg, *Liftoff: accurate mapping of gene annotations*.  
143 *Bioinformatics*, 2020.
- 144 9. Pertea, M., et al., *StringTie enables improved reconstruction of a transcriptome from*  
145 *RNA-seq reads*. *Nat Biotechnol*, 2015. **33**(3): p. 290-5.

146