Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genome Analysis

LRTK: A unified and versatile toolkit for analyzing linked-read sequencing data

Chao Yang[1†], Zhenmiao Zhang[1†], Herui Liao[2], Lu Zhang[3,1*]

[1]Department of Computer Science, Hong Kong Baptist University, Hong Kong SAR, Hong Kong

[2]Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, Hong Kong

[3]Institute for Research and Continuing Education, Hong Kong Baptist University, China

[†]These authors contributed equally to this work.

*To whom correspondence should be addressed: E-mail: ericluzhang@hkbu.edu.hk.

Abstract

Summary:
Linked-read sequencing technologies offering reads with both high base quality and long-range DNA connectedness have shown great success in genomic studies. The mainstream platforms include 10x Genomics linked-read (10x), Single Tube Long Fragment Read (stLFR) and Transposase Enzyme-Linked Long-read Sequencing (TELL-Seq). The existing data analysis pipelines, e.g., Long Ranger, have been developed to process sequencing data from particular platforms and so are unable to fully utilize the unique characteristics of other platforms; thus, users have limited tools to choose for downstream analysis. To address these limitations, we present Linked-Read ToolKit (LRTK), a unified and versatile toolkit to process linked-read sequencing data from different platforms. LRTK provides flexible functions to perform data simulation, format conversion, data preprocessing, barcode-aware read alignment, variant calling and phasing. It also allows multi-sample batch processing and generates a HTML report with key statistics and plots. We applied LRTK to the linked-read data of NA24385 obtained from all three platforms, where the results showed the advancement of LRTK in structural variation recall rate for 10x linked-reads and in increasing phase block N50 for 10x and stLFR linked-reads.
Availability: Source codes are available at https://github.com/ericcombiolab/LRTK. Anaconda supports the installation of LRTK and its dependencies.
Contact: ericluzhang@hkbu.edu.hk
Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Linked-read sequencing provides data with high base quality and long-range DNA connectedness and has shown significant advancement in human genome and metagenome research (Eisenstein, 2015; Wang *et al.*, 2019; Chen *et al.*, 2020). It circumvents the lack of long-range DNA information by short-read sequencing; and the high error rates and large initial DNA load requirements of long-read sequencing (e.g., Oxford Nanopore and Pacific Bioscience). These advantages of linked-read sequencing are valuable in dealing with some challenging cases of low-input clinical samples, such as cancer tissues and infectious disease samples. Further, the low cost of linked-read sequencing enables its application to large cohort studies.

In the past decades, several commercially available linked-read sequencing platforms, such as 10x Genomics linked-read (10x; now discontinued) and the newly developed stLFR (Wang *et al.*, 2019) and TELL-Seq (Chen *et al.*, 2020), have been applied to many genomic studies. Some pipelines have been developed to analyze the linked-read sequencing data generated by these platforms. For example, Long Ranger (Zheng *et al.*, 2016) performs barcode-aware read alignment and implements modules for variant calling and phasing using 10x linked-reads. It requires large storage to save intermediate outputs. Tell-Sort (Chen *et al.*, 2020) is a Docker-based pipeline to process raw TELL-Seq reads, and its source codes are not publicly available. For stLFR (Wang *et al.*, 2019), a customized pipeline has been developed to first convert its raw reads into a 10x-compatible format, after which Long Ranger is applied for downstream analysis. This pipeline commonly requires a lot of RAM, and its data format conversion process is time consuming. There is a lack of a unified and open-source toolkit that works compatibly with different platforms.

Here, we present Linked-Read ToolKit (LRTK), a unified and versatile toolkit to analyze linked-read sequencing data from any of the three platforms. LRTK delivers a suite of utilities to perform data simulation, format conversion, data preprocessing, barcode-aware read alignment, quality control, variant detection and phasing. In particular, LRTK is open-source and can generate a HTML report to calculate the key parameters for library preparation and summarize the quality statistics of sequenced reads. We applied LRTK to analyze 10x linked-reads, stLFR and TELL-Seq from NA24385 and found that LRTK outperformed Long Ranger in structural variation (SV) detection recall rate for 10x linked-reads and increased phase block N50 for 10x linked-reads and stLFR.

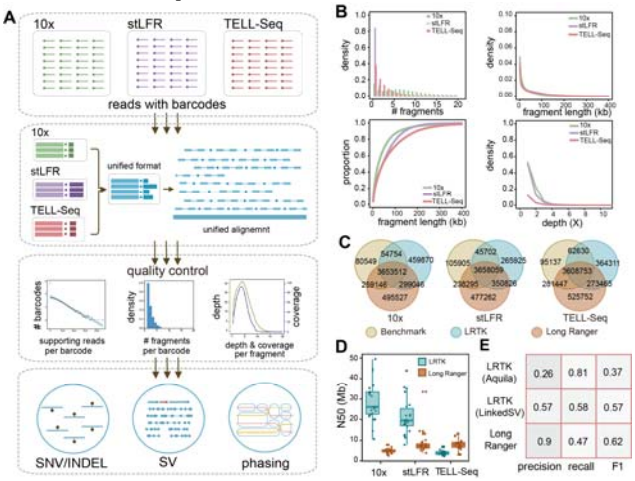


Figure 1. (**A**) Overview of LRTK; (**B**) Quality matrices for different platforms; (**C**) Performance of SNV and INDEL calls using FreeBayes; (**D**) Phase block N50; (**E**) SV detection in 10x linked-reads. The detailed descriptions are provided in **Supplementary Figure 1** and **Supplementary Notes**.

2 Methods

LRTK integrates several widely used off-the-shelf tools and implements utility scripts to facilitate linked-read analysis. It consists of two modules: (1) raw read analysis module that performs data simulation, format conversion, data preprocessing, barcode-aware read alignment and quality control; and (2) variant analysis module that performs detection and phasing of single nucleotide variations (SNVs), small insertions and deletions (INDELs) and SVs (**Supplementary Figure 1** and **Figure 1A**). The raw read analysis module converts the format of linked-reads from any of the three platforms into a unified FASTQ format (**Supplementary Figure 2**), which contains a new field "BX:Z:" to store 16 bp (10x linked-reads), 18 bp (TELL-Seq) and 30 bp (stLFR) barcode sequences. For 10x and stLFR linked-reads, the barcodes are compared to corresponding barcode whitelists to remove any potential errors. Due to the lack of a barcode whitelist for TELL-Seq, LRTK adopts the approach described by Chen et al. (Chen *et al.*, 2020) to correct the barcode errors. For data preprocessing, LRTK adopts fastp (Chen *et al.*, 2018) to remove low-quality reads and adapter contamination rapidly. We modified EMA (Shajii *et al.*, 2018) to perform barcode-aware read alignment and make it compatible for the barcodes with various lengths. The duplicated reads are marked using MarkDuplicates with the "BARCODE_TAG" parameter in Picard (https://broadinstitute.github.io/picard/). LRTK also supports linked-reads simulation from 10x and stLFR platforms.

The variant analysis module integrates six well-known variant callers and three phasing tools to detect and phase SNVs, INDELs and SVs **(Supplementary Figure 1)**. Users can either run these functions independently or use the "WGS" command to run the whole pipeline in an end-to-end manner, which has been optimized for multi-sample batch analysis. LRTK can generate a standalone HTML report to show the key statistics and essential plots for raw reads, read alignments, reconstructed physical long fragments and genomic variants **(Supplementary Figure 3)**. LRTK reconstructs long DNA fragments using the read coordinates from alignments and calculates several key statistics such as the number of fragments per barcode, weighted/unweighted average fragment length, and read depth per fragment **(Supplementary Figure 4)** (Zhang *et al.*, 2019). LRTK and its dependents are compatible and can be easily installed using Anaconda. The detailed user mannuals are added in the **Supplementary Notes**.

3 Results

We used LRTK to analyze 10x linked-reads, stLFR and TELL-Seq of NA24385 **(Supplementary Table 1)**. We obtained 4.5 M, 38 M, and 41 M error-corrected barcodes for 10x linked-reads, stLFR and TELL-Seq, respectively. Of these barcodes, 2.3 M, 13 M and 6.9 M barcodes were eligible to reconstruct long DNA fragments **(Supplementary Table 2)**. LRTK detected approximately 7.36, 1.22 and 3.09 fragments per barcode and achieved an average fragment length of 50.19 kb, 62.28 kb and 80.04 kb for 10x linked-reads, stLFR and TELL-Seq, respectively **(Figure 1B and Supplementary Table 3)**. We further evaluated the variants detected by LRTK benchmarked with the gold standard from Genome in a Bottle (Zook *et al.*, 2020). LRTK (FreeBayes) achieved average recall rates of 94% for SNVs and 73% for INDELs **(Figure 1C and Supplementary Table 4)**. We observed that nearly 98% (10x linked-reads), 96% (stLFR), and 95% (TELL-Seq) of the SNVs and INDELs could be phased by LRTK (HapCUT2), suggesting that the linked-reads from the three platforms had comparable variant phasing performance **(Supplementary Figure 5C)**. Compared to Long Ranger, LRTK increased phase block N50 up to 26.1 Mb (Long Ranger: 4.9Mb) and 19.4 Mb for 10x linked-reads and stLFR (Long Ranger: 7.4 Mb, **Figure 1D**). LRTK (Aquila) outperformed Long Ranger with respect to the recall of SVs, especially the deletions (Aquila: 81%, Long Ranger: 47%) **(Figure 1E)**.

References

Chen,S. *et al.* (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

Chen,Z. *et al.* (2020) Ultra-low input single tube linked-read library method enables short-read second-generation sequencing systems to generate highly accurate and economical long-range sequencing information routinely. Genome Res.,**30**(6): 898-909.

Eisenstein,M. (2015) Startups use short-read data to expand long-read sequencing market. *Nat. Biotechnol.*, **33**, 433–435.

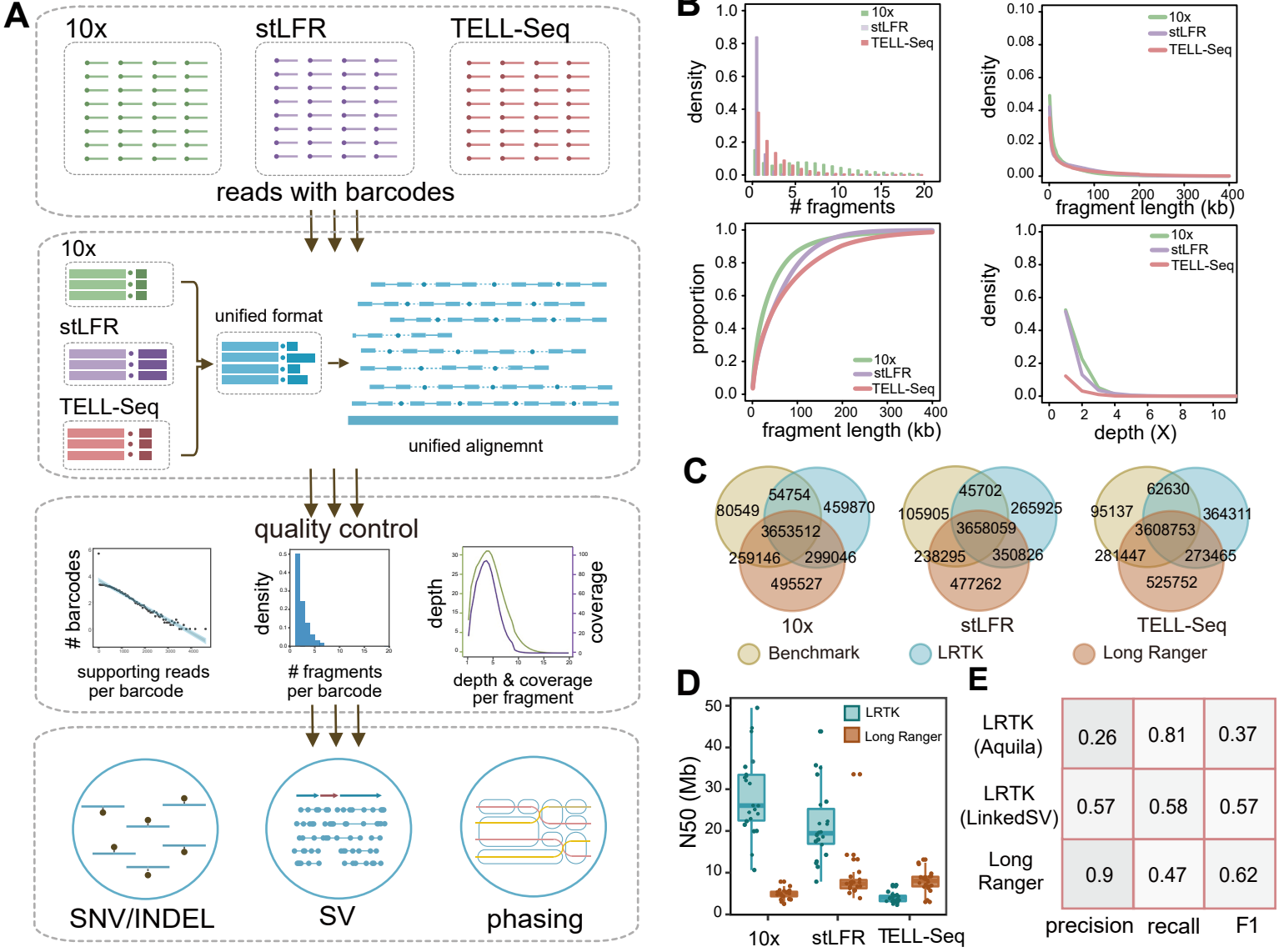
Shajii,A. *et al.* (2018) Statistical binning for barcoded reads improves downstream analyses. *Cell Syst.*, **7**, 219-226.e5.

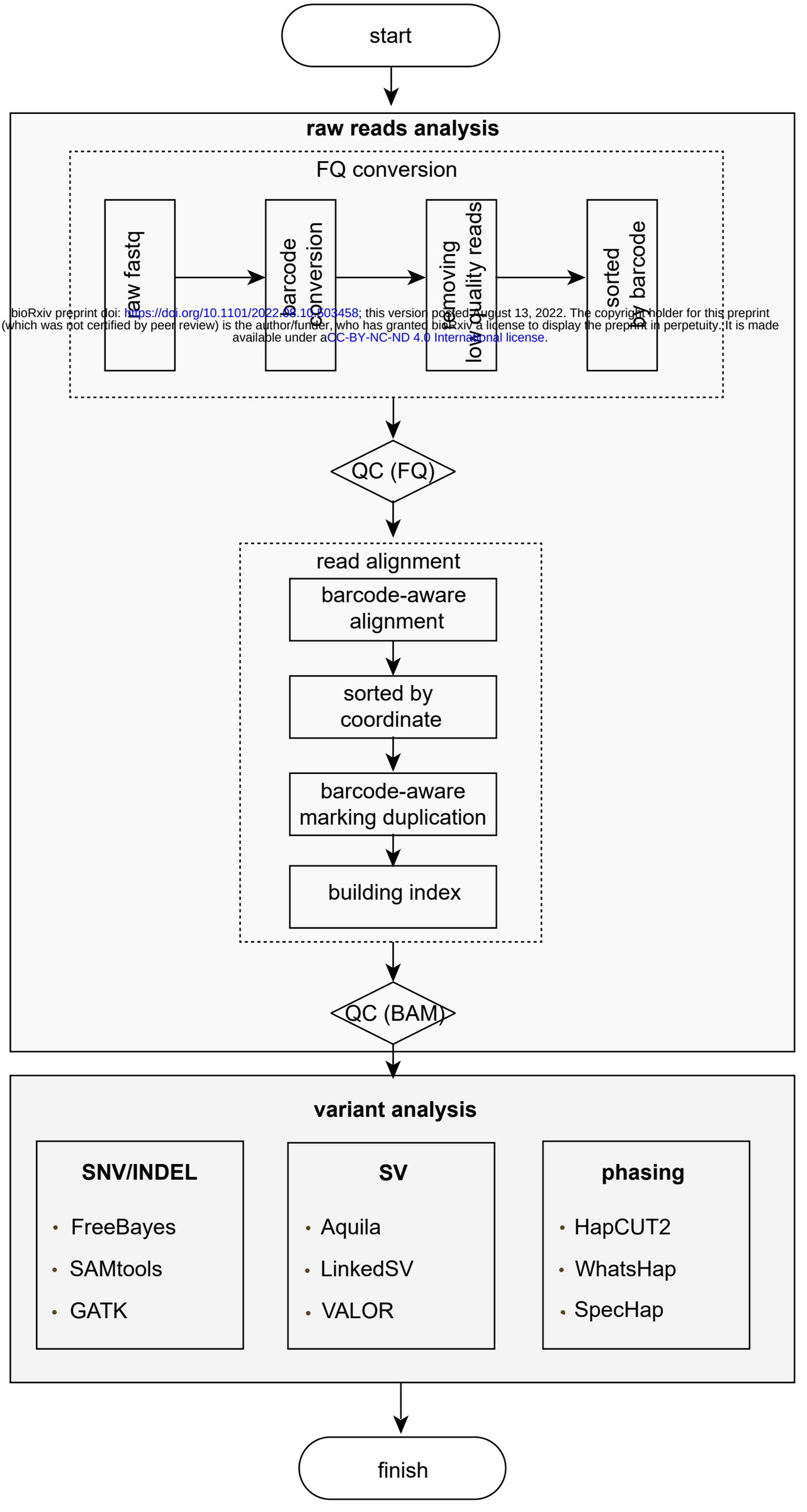
Wang,O. *et al.* (2019) Efficient and unique co-barcoding of second-generation sequencing reads from long DNA molecules enabling cost effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res.*, **29**(5): 798-808.

Zhang,L. *et al.* (2019) Assessment of human diploid genome assembly with 10x Linked-Reads data. *Gigascience*, **8**, 1–11.

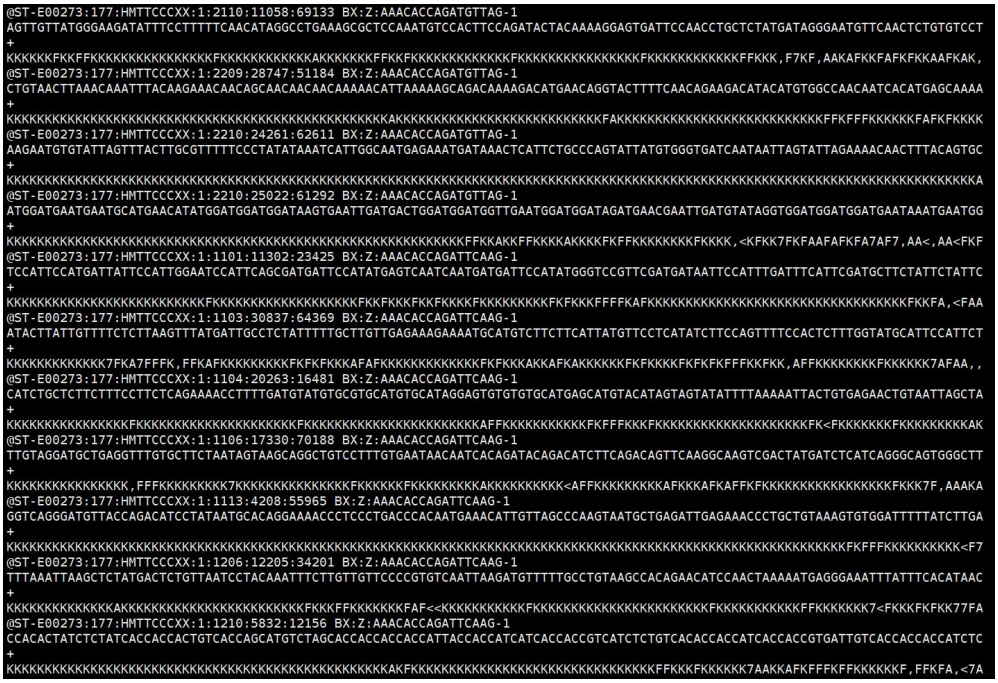
Zheng,G.X.Y. *et al.* (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.*, **34**, 303–311.

Zook,J.M. *et al.* (2020) A robust benchmark for detection of germline large deletions and insertions. *Nat. Biotechnol.*, **38**, 1347–1355.

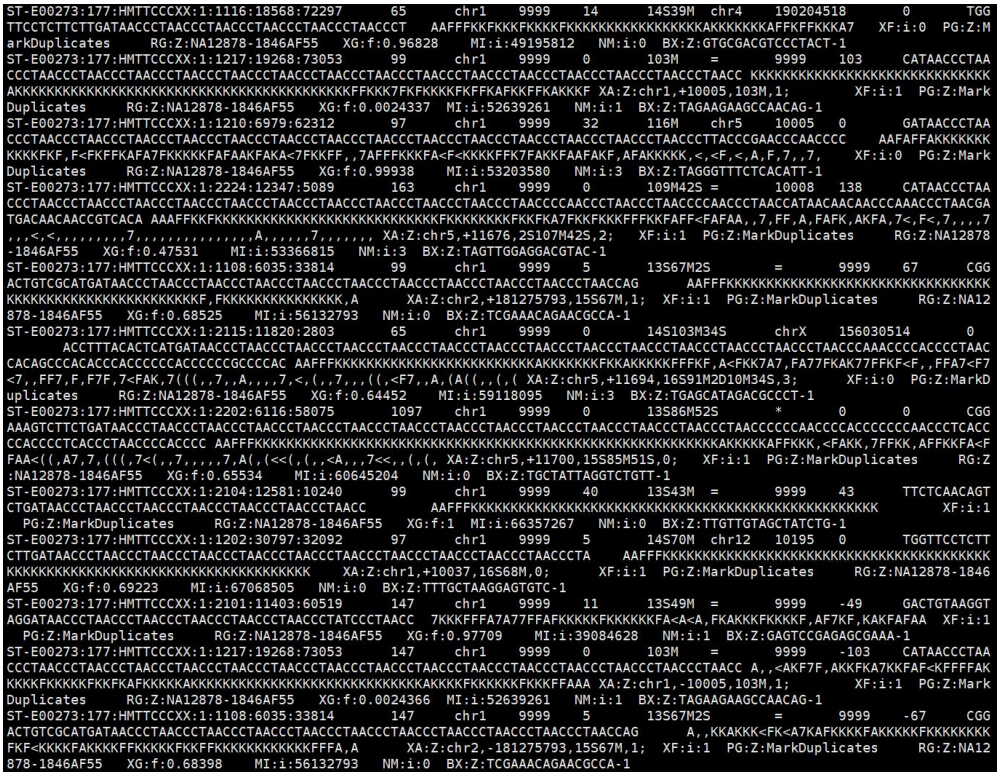

A
10x
stLFR
TELL-Seq
reads with barcodes
unified format
unified alignemnt
quality control
barcodes
supporting reads per barcode
density
fragments per barcode
depth
coverage
depth & coverage per fragment
SNV/INDEL
SV
phasing
B
density
fragments
fragment length (kb)
proportion
depth (X)
C
80549
54754
459870
3653512
259146
299046
495527
105905
45702
265925
3658059
238295
350826
477262
95137
62630
364311
3608753
281447
273465
525752
Benchmark
LRTK
Long Ranger
D
N50 (Mb)
LRTK
Long Ranger
E
LRTK (Aquila)
LRTK (LinkedSV)
Long Ranger
0.26
0.81
0.37
0.57
0.58
0.57
0.9
0.47
0.62
precision
recall
F1



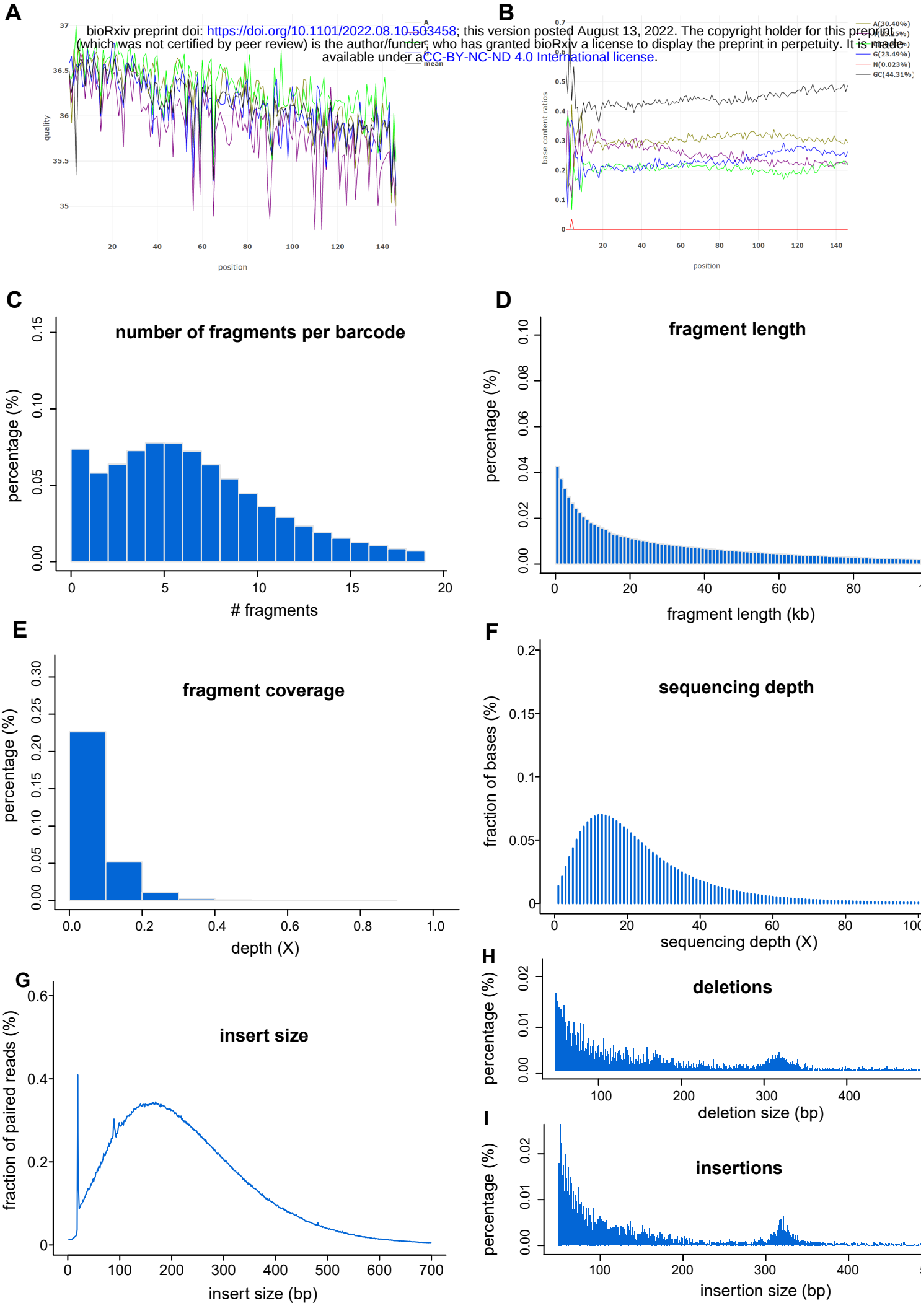
A.

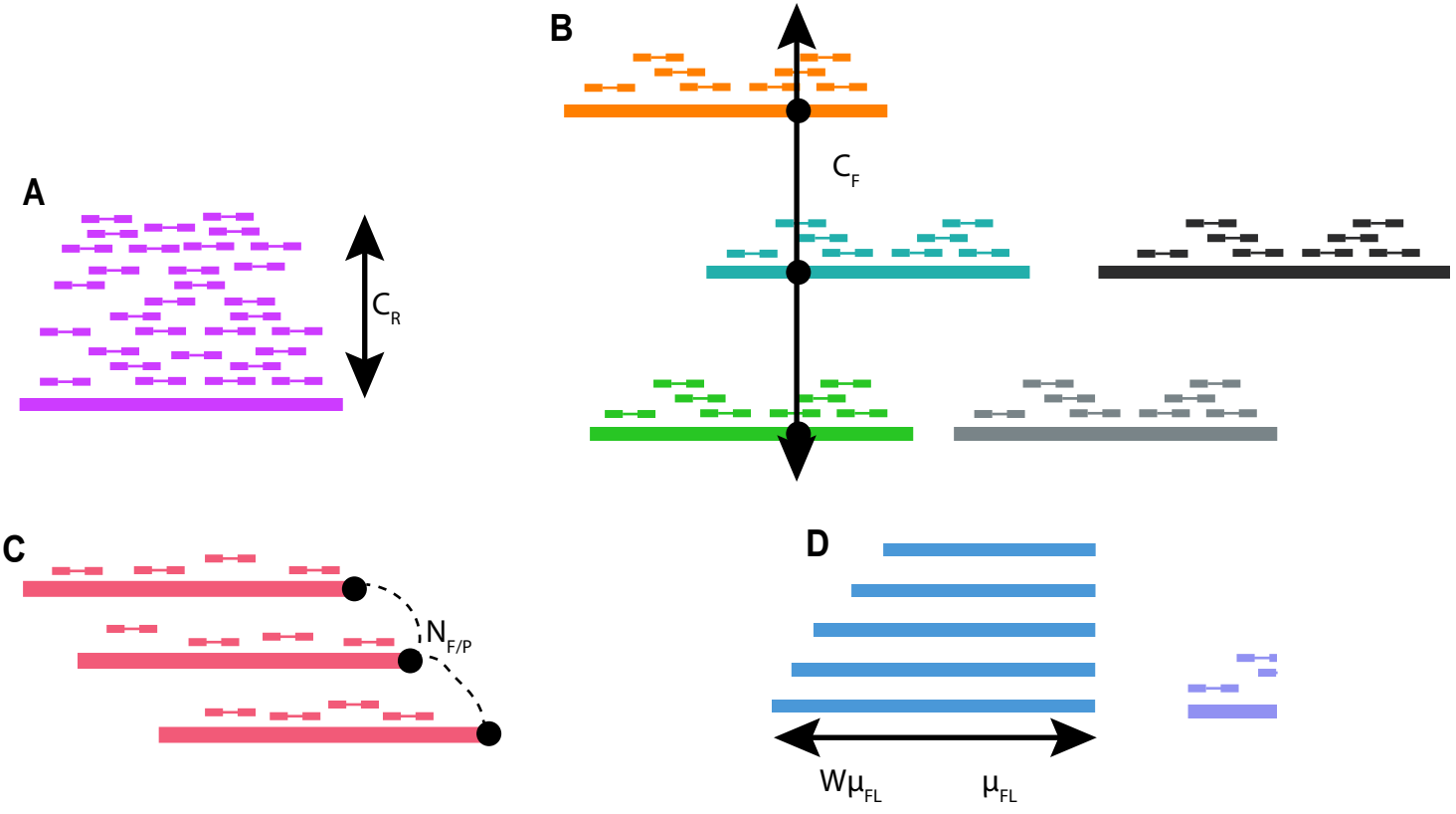


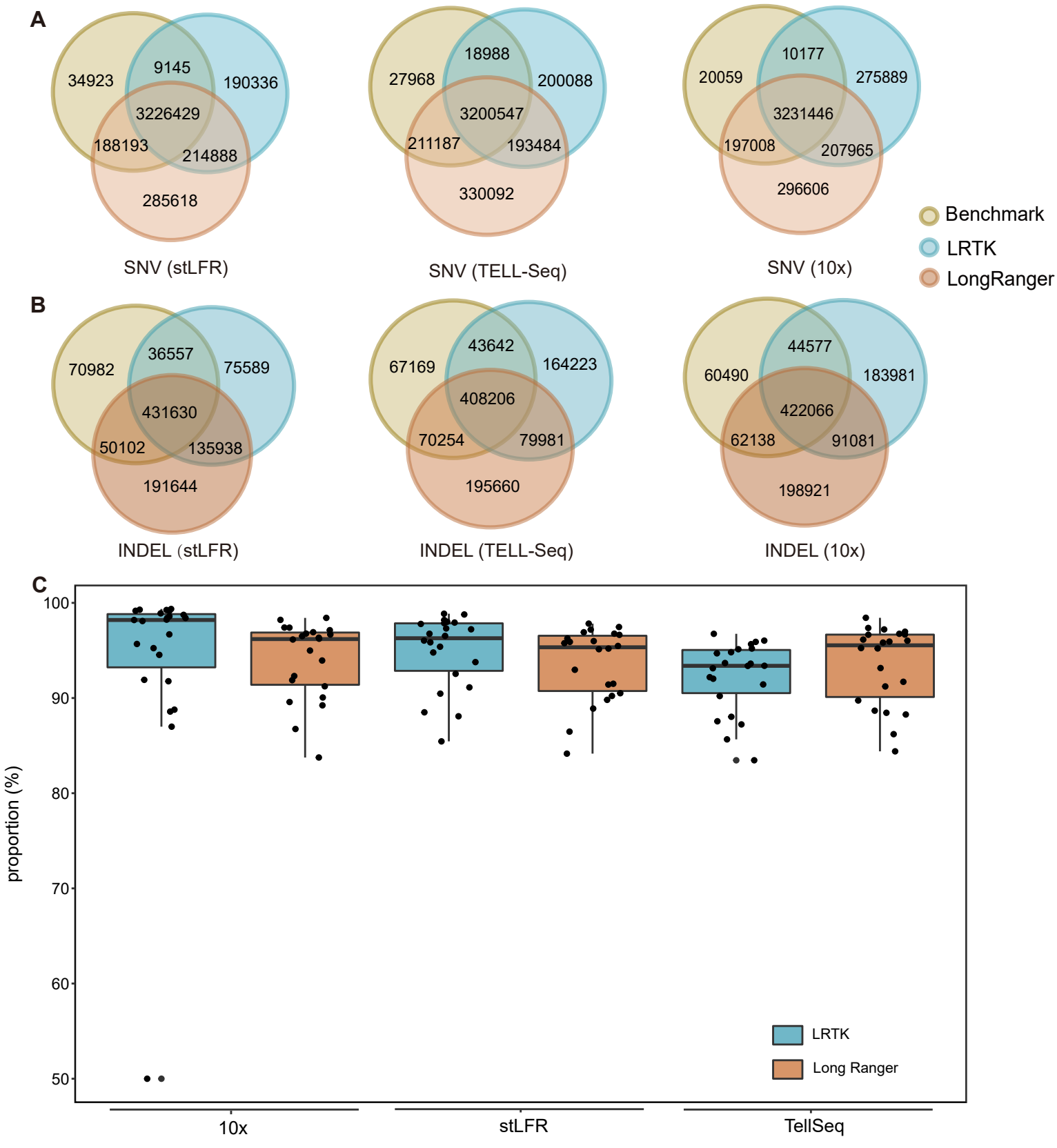
B.



Supplemental Figure 2. (A). Illustration of the unified FASTQ format. Each read contains the barcode field “BX:Z:barcode” following the read id. The length of barcode sequence are 16 bp, 30bp and 18bp for 10x, stLFR and TELL-Seq, respectively. **(B)**. Illustration of the barcode information in the alignment file. The barcode information is presented using the “BX:Z:barcode” format in the optional filed of alignment files.




A
C_R
B
C_F
C
$N_{F/P}$
D
$W\mu_{FL}$
μ_{FL}


A
34923
9145
190336
3226429
188193
214888
285618
SNV (stLFR)
27968
18988
200088
3200547
211187
193484
330092
SNV (TELL-Seq)
20059
10177
275889
3231446
197008
207965
296606
SNV (10x)
Benchmark
LRTK
LongRanger
B
70982
36557
75589
431630
50102
135938
191644
INDEL (stLFR)
67169
43642
164223
408206
70254
79981
195660
INDEL (TELL-Seq)
60490
44577
183981
422066
62138
91081
198921
INDEL (10x)
C
proportion (%)
100
90
80
70
60
50
LRTK
Long Ranger
10x
stLFR
TellSeq