



Jan 16, 2019

🌐 Instructions for recreating elPrep 4.0.0 WGS benchmarks

📖 [PLOS One](#)

DOI

dx.doi.org/10.17504/protocols.io.w35fgq6

Charlotte Herzeel¹

¹ExaScience Life Lab, Imec, Leuven, Belgium



Charlotte Herzeel

OPEN  ACCESS



DOI: dx.doi.org/10.17504/protocols.io.w35fgq6

External link: <https://www.biorxiv.org/content/early/2018/12/10/492249>

Protocol Citation: Charlotte Herzeel 2019. Instructions for recreating elPrep 4.0.0 WGS benchmarks. **protocols.io**
<https://dx.doi.org/10.17504/protocols.io.w35fgq6>

Manuscript citation:

Herzeel C, Costanza P, Decap D, Fostier J, Verachtert W (2019) elPrep 4: A multithreaded framework for sequence analysis. PLoS ONE 14(2): e0209523. doi: [10.1371/journal.pone.0209523](https://doi.org/10.1371/journal.pone.0209523)

License: This is an open access protocol distributed under the terms of the **[Creative Commons Attribution License](#)**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited

Protocol status: Working

We use this protocol and it's working

Created: January 15, 2019

Last Modified: January 16, 2019

Protocol Integer ID: 19293

Keywords: elprep, markduplicates, bqsr, sam, bam

Abstract

Instructions for recreating the elPrep4.0.0 WGS benchmarks used in the following paper:

Herzeel C, Costanza P, Decap D, Fostier J, Verachtert W. elPrep: A multithreaded framework for sequence analysis. BioRxv <https://doi.org/10.1101/492249>



1 Configuration

Note

These instructions have been tested with elPrep v.4.0.0. The following assumes that everything is performed from a working directory WORKDIR.

1.1 Hardware

Note

- * 2x18-core Intel Xeon processor E5-2699v3 Haswell @ 2.3GHz
- * 256 GB RAM
- * 2x400 GB SSD

1.2 Software

Note

- * Ubuntu 14.04.5 LTS
- * elPrep 4.0.0

2 Installation



Software

elPrep	NAME
Linux	OS
imec	DEVELOPER
https://github.com/ExaScience/elprep	SOURCE LINK

Note

The following steps are required to run elPrep:

1. Download the elPrep binary distribution from <https://github.com/ExaScience/elprep>
Direct download link:
<https://github.com/ExaScience/elprep/releases/download/v4.0.0/elprep-v4.0.0.tar.gz>
2. `mkdir elprep-v4.0.0`
3. `mv elprep-v4.0.1.tar.gz elprep-v4.0.0`
4. `cd elprep-v4.0.0`
5. `tar xvf elprep-v4.0.0.tar.gz`
6. `PATH=$WORKDIR/elprep-v4.0.0:$PATH`

3 Data preparation

Note

Our WGS benchmark uses the public data provided by Illumina (<https://www.illumina.com/platinumgenomes.html>). This data consists of unaligned FASTQ files, which can be downloaded from the European Nucleotide Archive (<https://www.ebi.ac.uk/ena/data/view/PRJEB3381>). The following steps describe how to download, prepare, and align the data using BWA mem (version 0.7.17). Similarly, our benchmark requires the reference genome and databases with known SNPs to be converted into elPrep-specific formats. The following steps also describe how to download the data from public repositories and creating the elPrep-specific conversions.

3.1 Required tools



Software

BWA

NAME

Linux

OS

Heng Li

DEVELOPER

<https://github.com/lh3/bwa>

SOURCE LINK

Note

1. Ensure GCC installed (version 4.8.4 recommended)
2. Download BWA source code from <https://github.com/lh3/bwa> Direct link: <https://github.com/lh3/bwa/releases/download/v0.7.17/bwa-0.7.17.tar.bz2>
3. tar xvf bwa-0.7.17.tar.bz2
4. cd bwa-0.7.17
5. make
6. cd \$WORKDIR

Software

correct-platinum-fastq-sequence-identifier

NAME

Linux

OS

imec

DEVELOPER

<https://github.com/ExaScience/correct-platinum-fastq-sequence-identifier>SOURCE
LINK



Note

The FASTQ files at the ENA provide the Illumina sequence identifiers only as comments, but for optical duplicate marking to be done properly in elPrep, Picard, and GATK, they need to be used as actual sequence identifiers in the FASTQ files before they are aligned with BWA mem. This can be arranged with a small tool we provide (<https://github.com/ExaScience/correct-platinum-fastq-sequence-identifier>) or another tool.

Download and install our tool for fixing FASTQ read names. These instructions assume you have a working Golang installation (see <https://golang.org/doc/install>):

1. `go get github.com/ExaScience/correct-platinum-fastq-sequence-identifier`
2. `go build github.com/ExaScience/correct-platinum-fastq-sequence-identifier`

3.2 Required data

**Note****FASTQ files**

* Download Illumina Platinum whole-genome NA12878 FASTQ files from <https://www.ebi.ac.uk/ena/data/view/PRJEB3381>

Direct links:

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR194/ERR194147/ERR194147_1.fastq.gz

ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR194/ERR194147/ERR194147_2.fastq.gz

Reference files

* Download the hg38 reference files from <http://lh3.github.io/2017/11/13/which-human-reference-genome-to-use>

Direct link:

ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/001/405/GCA_000001405.15_GRCh38/s eqs_for_alignment_pipelines.ucsc_ids/GCA_000001405.15_GRCh38_no_alt_analysis_set.f na.gz

Known variants

* Download the database with known SNPs from <https://software.broadinstitute.org/gatk/download/bundle>

Direct links:

ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/Mills_and_1000G_gold_standard.indels.hg38.vcf.gz

ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg38/dbsnp_138.hg38.vcf.gz

When attempting a download, this may result in an error message that the login is incorrect. This is because the ftp site only allows a maximum of 25 users at the same time. If this happens, try again.

3.3 Data preparation steps**3.3.1 Create the reference index**

**Command**

Required time: ca. 60 minutes

Result: GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.*

```
gunzip GCA_000001405.15_GRCh38_no_alt_analysis_set.fna.gz  
bwa-0.7.17/bwa index GCA_000001405.15_GRCh38_no_alt_analysis_set.fna
```

3.3.2 Clean up the FASTQ files

Command

This invocation runs the tool in parallel.

Required time: ca. 3.5h

Result: ERR194147OPT_1.fastq.gz

```
./correct-platinum-fastq-sequence-identifier par ERR194147_1.fastq.gz  
ERR194147OPT_1.fastq.gz
```

Command

Required time: ca. 3.5h

Result: ERR194147OPT_2.fastq.gz

```
./correct-platinum-fastq-sequence-identifier par ERR194147_2.fastq.gz  
ERR194147OPT_2.fastq.gz
```




3.3.3 Align the FASTQ files to create a BAM file

Command

Required time: 6h10m

Result: NA12878.bam

```
bwa-0.7.17/ bwa mem -t 72 -R
'@RG\tID:Group1\tLB:lib1\tPL:illumina\tSM:sample1'
GCA_000001405.15_GRCh38_no_alt_analysis_set.fna
ERR1941470PT_1.fastq.gz ERR1941470PT_2.fastq.gz | elprep filter
/dev/stdin NA12878.bam
```

3.3.4 Create the hg38 elfasta file

Command

Required time: ca. 1 minute

Result: hg38.elfasta

```
elprep fasta-to-elfasta
GCA_000001405.15_GRCh38_no_alt_analysis_set.fna hg38.elfasta
```

3.3.5 Create elsites files from vcf files

**Command****Required time: ca. 1 minute****Result: dbsnp_138.hg38.elsites**

```
gunzip dbsnp_138.hg38.vcf.gz  
elprep vcf-to-elsites dbsnp_138.hg38.vcf dbsnp_138.hg38.elsites
```

Command**Required time: ca. 10 seconds****Result: Mills_and_1000G_gold_standard.indels.hg38.elsites**

```
gunzip Mills_and_1000G_gold_standard.indels.hg38.vcf  
elprep vcf-to-elsites Mills_and_1000G_gold_standard.indels.hg38.vcf  
Mills_and_1000G_gold_standard.indels.hg38.elsites
```

4 Benchmarking elPrep



Note

elPrep provides a lot of filtering options. The following benchmark implements a pipeline that executes the following four steps:

1. Sorting by coordinate order (equivalent to, for example https://software.broadinstitute.org/gatk/documentation/tooldocs/current/picard_sam_SortSam.php)
2. Marking PCR and optical duplicates (equivalent to, for example, https://software.broadinstitute.org/gatk/documentation/tooldocs/current/picard_sam_markduplicates_MarkDuplicates.php)
3. Base quality score recalibration (equivalent to, for example, https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_hellbender_tools_walkers_bqsr_BaseRecalibrator.php)
4. Applying base quality score recalibration (equivalent to, for example, https://software.broadinstitute.org/gatk/documentation/tooldocs/current/org_broadinstitute_hellbender_tools_walkers_bqsr_ApplyBQSR.php)

Please see the elPrep documentation at <https://github.com/ExaScience/elprep> for further filtering options.

Command

Required time: 3h37m

Required RAM: 192GB

Result: NA12878.sfm.bam, NA12878.sfm.metrics, NA12878.sfm.recal

```
elprep sfm NA12878.bam NA12878.sfm.bam --mark-duplicates --mark-  
optical-duplicates NA12878.sfm.metrics --sorting-order coordinate --  
bqsr NA12878.sfm.recal --known-sites  
dbSNP_138.hg38.elSITES,Mills_and_1000G_gold_standard.indels.hg38.elSIT  
es --bqsr-reference hg38.elfasta
```