Jan 16, 2019

Instructions for recreating elPrep 4.0.0 WES benchmarks

PLOS One

DOI
dx.doi.org/10.17504/protocols.io.w65fhg6

Charlotte Herzeel[1]

[1]ExaScience Life Lab, imec, Leuven, Belgium

Charlotte Herzeel

DOI: **dx.doi.org/10.17504/protocols.io.w65fhg6**

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

Protocol Citation: Charlotte Herzeel 2019. Instructions for recreating elPrep 4.0.0 WES benchmarks. **protocols.io** **https://dx.doi.org/10.17504/protocols.io.w65fhg6**

Manuscript citation:

Protocol status: Working
We use this protocol and it's working

Created: January 16, 2019

Last Modified: January 16, 2019

Protocol Integer ID: 19389

Keywords: elprep, markduplicates, bqsr, sam, bam

Abstract

Instructions for recreating the elPrep4.0.0 WES 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 4.0.1

2 Installation

Software

elPrep	NAME
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. mdkir 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 WES benchmark uses the public data provided by the Genome in a Bottle Consortium (GIAB). This data consists of unaligned FASTQ files, but we offer an aligned BAM file for this data on our demo repository (see https://github.com/ExaScience/elprep/tree/master/demo). Otherwise, the following steps describe how to download and align the data yourself using BWA mem (version 0.7.17). Similarly, our benchmark requires the reference genome, databases with known SNPs and BED files to be converted into an elPrep-specific format. Again, these files can be downloaded from our demo repository. Otherwise, the following steps describe how to download the data from public repositories and creating the elPrep-specific conversions.

3.1 Required tools

Software

BWA	NAME
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

3.2 Required data

Note

FASTQ and BED files
* Download GIAB whole-exome NA12878, FASTQ and BED files from https://github.com/genome-in-a-bottle
Direct links:

ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/NIST7035_TAAGGCGA_L001_R1_001.fasq.gz

ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/NIST7035_TAAGGCGA_L001_R2_001.fastq.gz

ftp://ftp-trace.ncbi.nih.gov/giab/ftp/data/NA12878/Garvan_NA12878_HG001_HiSeq_Exome/nexterarapidcapture_expandedexome_targetedregions.bed.gz

Reference files
* Download the hg19 reference files from https://software.broadinstitute.org/gatk/download/bundle
Direct links:

ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg19/ucsc.hg19.*

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.

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/hg19/Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.gz

ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/hg19/dbsnp_138.hg19.vcf.gz

3.3 Data preparation steps

3.3.1 Create the reference index:

Command

Required time: ca. 1h
Result: ucsc.hg19.fasta.gz.*

```
bwa-0.7.17/bwa index ucsc.hg19.fasta.gz
```

3.3.2 Align the FASTQ files to create a BAM file:

Command

Required time: ca. 5 minutes
Result: NA12878.bam

```
bwa-0.7.17/bwa mem -t 72 -R
'@RG\tID:Group1\tLB:lib1\tPL:illumina\tSM:sample1' ucsc.hg19.fasta.gz
NIST7035_TAAGGCGA_L001_R1_001.fastq.gz
NIST7035_TAAGGCGA_L001_R2_001.fastq.gz | elprep /dev/stdin NA12878.bam
```

3.3.3 Create hg19 elfasta file:

Command

Required time: ca. 1 minute
Result: ucsc.hg19.elfasta

```
cp ucsc.hg19.fasta.gz hg19.fasta.gz
gunzip hg19.fasta.gz
elprep fasta-to-elfasta hg19.fasta ucsc.hg19.elfasta
```

3.3.4 Create elsites files from vcf files:

Command

Required time: ca. 1 minute
Result: dbsnp_138.hg19.elsites

```
gunzip dbsnp_138.hg19.vcf.gz
elprep vcf-to-elsites dbsnp_138.hg19.vcf dbsnp_138.hg19.elsites
```

Command

Required time: ca. 10 seconds
Result: Mills_and_1000G_gold_standard.indels.hg19.elsites

```
gunzip Mills_and_1000G_gold_standard.indels.hg19.sites.vcf.gz
elprep vcf-to-elsites
Mills_and_1000G_gold_standard.indels.hg19.sites.vcf
Mills_and_1000G_gold_standard.indels.hg19.elsites
```

3.3.5 Unzip the BED file with captured regions:

Command

```
gunzip nexterarapidcapture_expandedexome_targetedregions.bed.gz
```

4 Benchmarking elPrep

Note

elPrep provides a lot of filtering options, as well as two modes to execute it. 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.

As for the execution options, elPrep can be run in an in-memory mode (filter) or a mode that first splits the input into smaller chunks, operates on each chunk to produce partial results, and merges the partial results into a final output file (sfm). The filter mode uses significantly more RAM than the sfm mode, but also runs significantly faster.

Each invocation produces the same BAM file as output. See the Statistics section below to double-check whether the BAM file is correctly processed. Please delete the output files before each rerun of elPrep.

4.1 elPrep filter mode

Command

Required time: ca. 5 minutes
Required RAM: 80 GB RAM
Result: NA12878.filter.bam, NA12878.filter.metrics, NA12878.filter.recal

```
elprep filter NA12878.bam NA12878.filter.bam --mark-duplicates --mark-optical-duplicates NA12878.filter.metrics --sorting-order coordinate --bqsr NA12878.filter.recal --known-sites Mills_and_1000G_gold_standard.indels.hg19.elsites,dbsnp_138.hg19.elsites --bqsr-reference ucsc.hg19.elfasta --filter-non-overlapping-reads nexterarapidcapture_expandedexome_targetedregions.bed
```

4.2 elPrep sfm mode

Command

Required time: ca. 11 minutes
Required RAM: 22 GB RAM
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 Mills_and_1000G_gold_standard.indels.hg19.elsites,dbsnp_138.hg19.elsites --bqsr-reference ucsc.hg19.elfasta --filter-non-overlapping-reads nexterarapidcapture_expandedexome_targetedregions.bed
```