Supplementary material for article: Whisper: Read sorting allows robust mapping of sequencing data

Sebastian Deorowicz Agnieszka Debudaj-Grabysz Adam Gudyś
Szymon Grabowski

July 13, 2018

Contents

1 Examined programs

The following programs were used in the experimental part. The running parameters for are also given.

- Bowtie2 v. 2.3.0
  `bowtie2 -x hg38 -p 12 -1 <r1.fastq.gz> -2 <r2.fastq.gz> -S <output.sam>`
  The pairs of files were mapped pair by pair.

- BWA-MEM v. 0.7.15-r1140
  `bwa mem -t 12 hg38 <r1.fastq> <r2.fastq> > <output.sam>`
  The pairs of files were mapped pair by pair.

- GEM v. 3.6.0-bundle-release
  `gem-mapper -t 12 -p -1 <r1.fastq.gz> -2 <r2.fastq.gz> -o <output.sam> -I hg38`
  The pairs of files were mapped pair by pair.

- Kart v. 2.1.0
  `kart aln -t 12 -i hg38 -f <r1.fastq> -f2 <r2.fastq> > <output.sam>`
  The pairs of files were mapped pair by pair. As Kart does not support gzipped input the files were initially decompressed before mapping:
  `gzip -d <r1.fastq.gz>`

- Whisper v. 1.0
  `whisper -t 12 -out <output.sam> hg38 @<reads.txt>`
  The file `<reads.txt>` contains pairs of names of files to be mapped separated by a semicolon, e.g., in case of 3 pairs they are processed in a single run. Sample contents of `<reads.txt>`:
  `r1_1.fastq.gz;r1_2.fastq.gz`
  `r2_1.fastq.gz;r2_2.fastq.gz`
  `r3_1.fastq.gz;r3_2.fastq.gz`

2 Evaluation

2.1 Real data

2.1.1 Datasets

Reference human genome HG38 was downloaded as a part of Genome Analysis Tookit bundle `ftp://ftp.broadinstitute.org/bundle/hg38`). In the analysis we removed alternative and decay assemblies retaining 25 main chromosomes (22 autosomes, 2 allosomes and a mitochondrial chromosome).

The reads from NA12878 sample were downloaded from the EMBL-EBI European Nucleotide Archive:

`ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174324/ERR174324_1.fastq.gz`
`ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174324/ERR174324_2.fastq.gz`
`ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174325/ERR174325_1.fastq.gz`
`ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174325/ERR174325_2.fastq.gz`
`ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174326/ERR174326_1.fastq.gz`
`ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174326/ERR174326_2.fastq.gz`
`ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174327/ERR174327_1.fastq.gz`
`ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174327/ERR174327_2.fastq.gz`
`ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174328/ERR174328_1.fastq.gz`
`ftp.sra.ebi.ac.uk/vol1/fastq/ERR174/ERR174328/ERR174328_2.fastq.gz`

The subsets for different coverages constituted of files:

- 14.4×:
  - ERR174324_1, ERR174324_2,
- 28.3×:
  - ERR174324_1, ERR174324_2,
  - ERR174325_1, ERR174325_2,
- 42.0×:
  - ERR174324_1, ERR174324_2,
  - ERR174325_1, ERR174325_2,
  - ERR174326_1, ERR174326_2,
- 55.6×:
  - ERR174324_1, ERR174324_2,
  - ERR174325_1, ERR174325_2,
  - ERR174326_1, ERR174326_2,
  - ERR174327_1, ERR174327_2,
- 69.2×:
  - ERR174324_1, ERR174324_2,
  - ERR174325_1, ERR174325_2,
  - ERR174326_1, ERR174326_2,
  - ERR174327_1, ERR174327_2,
  - ERR174328_1, ERR174328_2.

2.1.2 Assessment of the results

The variants were called according to the GATK Best Practice pipeline. We assumed mappings to be stored in `mappings.sam` file.

As the first step SAM files created by mapping software were converted to BAM format and sorted using `samtools 1.3.1-42-g0a15035` (`http://www.htslib.org`).

```
samtools view -@ <num-threads> -b -h mappings.sam > mappings.bam

samtools sort -T <temp-dir> -@ <num-threads> -O bam mappings.bam
   > mappings.sorted.bam
```

After that, `Picard 2.9.2` (`https://broadinstitute.github.io/picard/`) was employed for marking duplicates and indexing BAM file:

```
java -jar picard-2.9.2.jar MarkDuplicates I=mappings.sorted.bam
   O=mappings.marked.bam M=mappings.metrics.txt ASSUME_SORTED=true

java -jar picard-2.9.2.jar AddOrReplaceReadGroups I=mappings.marked.bam
   O=mappings.rg.bam RGID=1 RGLB=lib1 RGPL=illumina RGPU=unit1 RGSM=NA12878

java -jar picard-2.9.2.jar BuildBamIndex I=mappings.rg.bam
```

These steps were followed by recalibrating quality scores with a use of `Genome Analysis Toolkit v3.7-0-gcfedb67` (`https://software.broadinstitute.org/gatk`). As a source of known SNPs and indels we used dbSNP 1.38 and Mills and 1000G gold standard indels from GATK bundle (`ftp://ftp.broadinstitute.org/bundle/hg38`).

```
java -jar GenomeAnalysisTK.jar -nct <num-threads> -T BaseRecalibrator
   -R Homo_sapiens_assembly38.fasta -I mappings.rg.bam
   -knownSites dbsnp_138.hg38.vcf
   -knownSites Mills_and_1000G_gold_standard.indels.hg38.vcf
   -o mappings.bqsr.table

java -jar GenomeAnalysisTK.jar -nct <num-threads> -T PrintReads
   -R Homo_sapiens_assembly38.fasta -I mappings.rg.bam
   -BQSR mappings.bqsr.table -o mappings.recalibrated.bam
```

After recalibration, variants were called using `GATK HaplotypeCaller`.

```
java -jar GenomeAnalysisTK.jar -nct <num-threads> -T HaplotypeCaller
   -R Homo_sapiens_assembly38.fasta -I mappings.recalibrated.bam
   --genotyping\_mode DISCOVERY -o mappings.raw.vcf
```

The accuracy of variant calling was assesed by `hap.py 0.2.12` package (`https://github.com/Illumina/hap.py`) on the basis of "true" variants obtained as a part of Genome in a Bottle project v3.3.2 (`ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.2/GRCh38`)

```
python hap.py HG001_GRCh38_...vcf.gz mappings.raw.vcf -f HG001_GRCh38_...bed
   -o mappings.happy.nist -r Homo_sapiens_assembly38.fasta --roc VQLSOD
```

2.2 Simulated data

2.2.1 Datasets

The datasets were generated using wgsim (`https://github.com/lh3/wgsim`) executed with default parameters. The 4 paired-end datasets for read lengths 75 bp, 100 bp, 125 bp, and 150 bp were obtained. Each containing 200 million pairs of reads. The used commands:

```
wgsim -N 200000000 -1 75 -2 75 hg38 sim_075_1.fq sim_075_2.fq
wgsim -N 200000000 -1 100 -2 100 hg38 sim_100_1.fq sim_100_2.fq
wgsim -N 200000000 -1 125 -2 125 hg38 sim_125_1.fq sim_125_2.fq
wgsim -N 200000000 -1 150 -2 150 hg38 sim_150_1.fq sim_150_2.fq
```

2.3 Assessment of the results

To evaluate the quality of mappings we used `wgsim_eval.py` included in the wgsim package in the following ways:

```
wgsim_eval.pl unique output.sam |
   ../wgsim_eval.pl alneval -g 30 -p > results_a1 2> results_a2
wgsim_eval.pl unique output.sam |
   ../wgsim_eval.pl alneval -g 30 -p -a > results_a1 2> results_a2
```

3 Environment

The computer used in test were of the following configuration:

- 2 Intel Xeon E5-2670 v3 CPU, 12 cores per CPU, each clocked at 2.3 GHz,
- 128 GiB RAM,
- 2 Seagate NAS HDD of size 6 TB each in RAID-0 configuration, hdparm -t reported read speed 360 MB/s.

4 Additional results

4.1 Mapping times

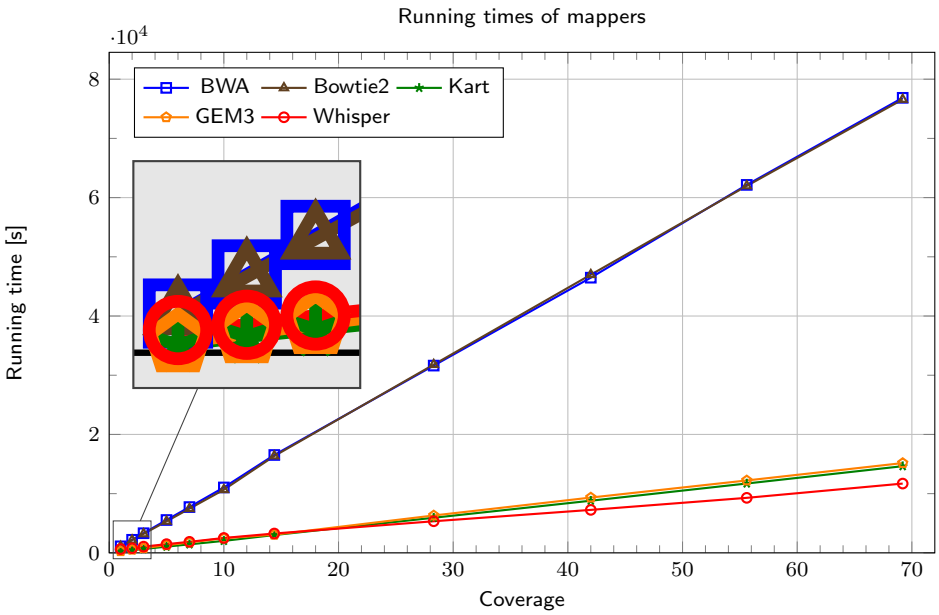

Figure 1: Comparison of mapping times for various coverages

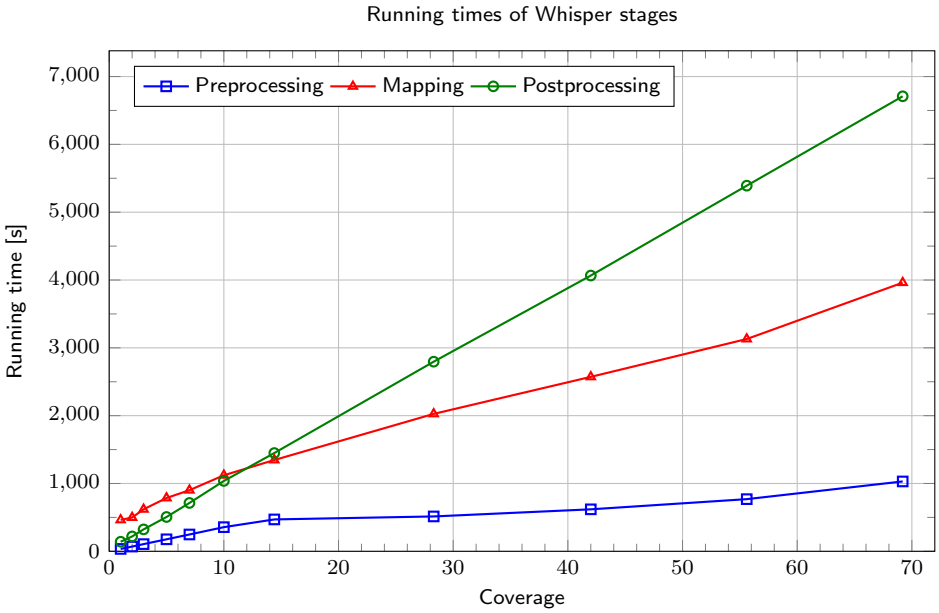

Figure 2: Comparison of running times of successive stages of Whisper

Table 1: Comparison of mapping times (in seconds) for various mappers for *H. sapiens* data. All reads are of length 101 bp.

Coverage	No. of reads [M]	BWA	Bowtie2	Kart	GEM3	Whisper
1.0	31	1,103	1,078	212	283	639
2.0	62	2,204	2,142	453	539	787
3.0	93	3,316	3,206	651	796	1,048
5.0	155	5,547	5,324	1,048	1,289	1,469
7.0	217	7,732	7,462	1,437	1,762	1,862
10.0	310	11,046	10,667	2,022	2,459	2,514
14.4	447	16,529	16,354	3,009	3,072	3,265
28.3	879	31,627	31,790	5,928	6,309	5,336
42.0	1,305	46,475	46,979	8,805	9,336	7,257
55.6	1,728	62,147	61,974	11,712	12,226	9,290
69.2	2,151	76,852	76,581	14,637	15,162	11,700

Table 2: Comparison of running times (in seconds) and RAM usage for various stages of Whisper for *H. sapiens* data

Coverage	Running time				RAM
	Preprocessing	Mapping	Postprocessing	Total	Total
1.0	35	463	141	639	10.5
2.0	70	498	219	787	10.7
3.0	106	619	323	1,048	10.7
5.0	177	786	506	1,469	11.9
7.0	248	901	713	1,862	12.0
10.0	356	1,121	1,037	2,514	12.6
14.4	470	1,346	1,449	3,265	13.5
28.3	514	2,026	2,796	5,336	14.6
42.0	619	2,572	4,066	7,257	15.8
55.6	769	3,130	5,391	9,290	16.0
69.2	1,030	3,961	6,709	11,700	15.8

4.2 Variant calling

Table 3: SNP variant calling for various coverages. 'Rec.' is for 'Recall', 'Prec.' is for 'Precision', 'F1' is for 'F1 score'.

Coverage	BWA-MEM			Bowtie2			Kart			GEM3			Whisper		
	Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1
14.4	0.9809	0.9908	0.9858	0.9646	0.9950	0.9796	0.9589	0.9587	0.9588	0.9630	0.9938	0.9782	0.9805	0.9894	0.9849
28.3	0.9965	0.9911	0.9938	0.9837	0.9958	0.9897	0.9923	0.9460	0.9686	0.9840	0.9948	0.9894	0.9965	0.9892	0.9929
42.0	0.9977	0.9901	0.9939	0.9863	0.9950	0.9906	0.9947	0.9525	0.9732	0.9873	0.9943	0.9907	0.9976	0.9881	0.9928
55.6	0.9981	0.9894	0.9937	0.9876	0.9943	0.9909	0.9954	0.9337	0.9636	0.9889	0.9938	0.9913	0.9979	0.9872	0.9925
69.2	0.9983	0.9889	0.9935	0.9883	0.9939	0.9911	0.9957	0.9278	0.9605	0.9898	0.9935	0.9916	0.9980	0.9866	0.9923

Table 4: Indel variant calling for various coverages. 'Rec.' is for 'Recall', 'Prec.' is for 'Precision', 'F1' is for 'F1 score'.

Coverage	BWA-MEM			Bowtie2			Kart			GEM3			Whisper		
	Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1	Rec.	Prec.	F1
14.4	0.8004	0.9466	0.8674	0.7867	0.9479	0.8598	0.7639	0.9298	0.8387	0.7794	0.9476	0.8553	0.7996	0.9464	0.8669
28.3	0.9049	0.9678	0.9353	0.8941	0.9701	0.9305	0.8873	0.9513	0.9182	0.8898	0.9699	0.9282	0.9046	0.9678	0.9351
42.0	0.9359	0.9737	0.9544	0.9273	0.9765	0.9513	0.9248	0.9597	0.9419	0.9249	0.9767	0.9501	0.9361	0.9736	0.9545
55.6	0.9501	0.9758	0.9628	0.9421	0.9790	0.9602	0.9422	0.9627	0.9523	0.9411	0.9796	0.9600	0.9501	0.9758	0.9628
69.2	0.9571	0.9764	0.9667	0.9499	0.9800	0.9647	0.9500	0.9638	0.9568	0.9497	0.9810	0.9651	0.9571	0.9762	0.9666

4.3 Simulated data results

4.3.1 Various read lengths

Table 5: Results for 200 million pairs of reads of length 75 bp with the default base error rate (0.020)

Mapper	Time [s]	All mappings		'Good' mappings (MAPQ $\geq$ 20)	
		Unmapped	Incorrect	Unmapped	Incorrect
BWA-MEM	16,234	0.00%	2.80%	4.50%	0.02%
Bowtie2	11,484	1.01%	6.68%	14.91%	0.31%
GEM	2,596	0.20%	3.03%	9.66%	0.02%
Kart	2,407	0.19%	3.84%	3.99%	1.35%
Whisper	3,026	0.01%	2.84%	3.74%	0.23%

Table 6: Results for 200 million pairs of reads of length 100 bp with the default base error rate (0.020)

Mapper	Time [s]	All mappings		'Good' mappings (MAPQ $\geq$ 20)	
		Unmapped	Incorrect	Unmapped	Incorrect
BWA-MEM	22,586	0.00%	2.29%	3.75%	0.02%
Bowtie2	13,901	0.53%	5.66%	13.19%	0.23%
GEM	3,205	0.15%	2.44%	7.26%	0.04%
Kart	2,762	0.19%	2.97%	3.25%	1.08%
Whisper	4,400	0.00%	2.31%	3.19%	0.14%

Table 7: Results for 200 million pairs of reads of length 125 bp with the default base error rate (0.020)

Mapper	Time [s]	All mappings		'Good' mappings (MAPQ $\geq$ 20)	
		Unmapped	Incorrect	Unmapped	Incorrect
BWA-MEM	25,685	0.00%	1.97%	3,27%	0.01%
Bowtie2	16,153	0.22%	4.89%	11.98%	0.18%
GEM	3,926	0.12%	2.09%	6.02%	0.05%
Kart	3,312	0.17%	2.54%	2.80%	0.95%
Whisper	4,819	0.02%	1.99%	2.83%	0.12%

Table 8: Results for 200 million pairs of reads of length 150 bp with the default base error rate (0.020)

Mapper	Time [s]	All mappings		'Good' mappings (MAPQ $\geq$ 20)	
		Unmapped	Incorrect	Unmapped	Incorrect
BWA-MEM	33,565	0.00%	1.73%	2.95%	0.01%
Bowtie2	20,755	0.08%	4.29%	11.34%	0.15%
GEM	4,374	0.09%	1.85%	5.18%	0.06%
Kart	3,660	0.16%	2.27%	2.51%	0.88%
Whisper	5,084	0.06%	1.81%	2.62%	0.14%

4.3.2 Various base error rates

Table 9: Results for 200 million pairs of reads of length 100 bp, base error rate 0.010

Mapper	Time [s]	All mappings		'Good' mappings (MAPQ $\geq$ 20)	
		Unmapped	Incorrect	Unmapped	Incorrect
BWA-MEM	15,038	0.00%	2.22%	3.69%	0.01%
Bowtie2	14,949	0.03%	4.60%	10.60%	0.25%
GEM	2,521	0.12%	2.22%	6.74%	0.02%
Kart	2,421	0.33%	2.36%	3.19%	0.62%
Whisper	2,692	0.00%	2.20%	3.11%	0.08%

Table 10: Results for 200 million pairs of reads of length 100 bp with, base error rate 0.015

Mapper	Time [s]	All mappings		'Good' mappings (MAPQ $\geq$ 20)	
		Unmapped	Incorrect	Unmapped	Incorrect
BWA-MEM	17,871	0.00%	2.26%	3.72%	0.01%
Bowtie2	14,285	0.16%	5.38%	12.45%	0.25%
GEM	2,983	0.15%	2.31%	6,92%	0.03%
Kart	2,692	0.25%	2.67%	3.21%	0.85%
Whisper	3,554	0.00%	2.24%	3.14%	0.11%

Table 11: Results for 200 million pairs of reads of length 100 bp, base error rate 0.020

Mapper	Time [s]	All mappings		'Good' mappings (MAPQ $\geq$ 20)	
		Unmapped	Incorrect	Unmapped	Incorrect
BWA-MEM	22,586	0.00%	2.29%	3.75%	0.02%
Bowtie2	13,901	0.53%	5.66%	13.19%	0.23%
GEM	3,205	0.15%	2.44%	7.26%	0.04%
Kart	2,762	0.19%	2.97%	3.25%	1.08%
Whisper	4,400	0.00%	2.31%	3.19%	0.14%

Table 12: Results for 200 million pairs of reads of length 100 bp, base error rate 0.025

Mapper	Time [s]	All mappings		'Good' mappings (MAPQ $\geq$ 20)	
		Unmapped	Incorrect	Unmapped	Incorrect
BWA-MEM	23,157	0.00%	2.34%	3.78%	0.02%
Bowtie2	13,800	1.26%	5.97%	14.31%	0.21%
GEM	3,566	0.14%	2.60%	7.77%	0.05%
Kart	3,076	0.15%	3.31%	3.31%	1.33%
Whisper	5,260	0.02%	2.44%	3.28%	0.21%

Table 13: Results for 200 million pairs of reads of length 100 bp, base error rate 0.030

Mapper	Time [s]	All mappings		'Good' mappings (MAPQ $\geq$ 20)	
		Unmapped	Incorrect	Unmapped	Incorrect
BWA-MEM	25,062	0.00%	2.38%	3.82%	0.02%
Bowtie2	13,192	2.44%	6.28%	15.88%	0.19%
GEM	3,768	0.14%	2.78%	8.45%	0.06%
Kart	3,346	0.14%	3.69%	3.41%	1.62%
Whisper	6,088	0.07%	2.65%	3.44%	0.31%