Variant Calling Benchmark

Not Only Human

Variant call process refers to the identification of a nucleotide difference from reference sequences at a given position in an individual genome or transcriptome. It includes single nucleotide polymorphism (SNPs), insertion/deletions (indels) and structural variants. One of the most popular variant calling applications is GenomeAnalysisTK (GATK) from Broad Institute. Often this GATK is used with BWA to compose a variant calling workflow focusing on SNPs and indels. After we published Dell HPC System for Genomics White Paper last year, there were significant changes in GATK. The key process, variant call step UnifiedGenotyper is no longer recommended in their best practice. Hence, here we recreate BWA-GATK pipeline according to the recommended practice to test whole genome sequencing data from mammals and plants in addition to human's whole genome sequencing data. This is a part of Dell's effort to help customers estimating their infrastructure needs for their various genomics data loads by providing a comprehensive benchmark.

Variant Analysis for Whole Genome Sequencing data

System

The detailed configuration is in Dell HPC System for Genomics White Paper, and the summary of system configuration and software is in Table 2.

Table 1 Server configuration and software

Component	Detail
Server	40x PowerEdge FC430 in FX2 chassis
Processor	Total of 1120 cores: Intel® Xeon® Dual E5-2695 v3 - 14 cores
Memory	128GB - 8x 16GB RDIMM, 2133 MT/s, Dual Rank, x4 Data Width
Storage	480TB IEEL (Lustre)
Interconnect	InfiniBand FDR
OS	Red Hat Enterprise 6.6
Cluster Management tool	Bright Cluster Manager 7.1
Short Sequence Aligner	BWA 0.7.2-r1039
Variant Analysis	GATK 3.5
Utilities	sambamba 0.6.0, samtools 1.2.1

BWA-GATK pipeline

The current version of GATK is 3.5, and the actual workflow tested obtained from the workshop, 'GATK Best Practices and Beyond'. In this workshop, they introduce a new workflow with three phases.

- Best Practices Phase 1: Pre-processing
- Best Practices Phase 2A: Calling germline variants
- Best Practices Phase 2B: Calling somatic variants
- Best Practices Phase 3: Preliminary analyses

Here we tested out phase 1, phase 2A and phase3 for germline variant call pipeline. The details of commands used in benchmark are listed below.

Phase 1. Pre-processing

Step 1. Aligning and Sorting

bwa mem -c 250 -M -t [number of threads] -R '@RG\tID:noID\tPL:illumine\tLB:noLB\tSM:bar' [reference chromosome] [read fastq 1] [read fastq 2] | samtools view -bu - | sambamba sort -t [number of threads] -m 30G --tmpdir [path/to/temp] -o [sorted bam output] /dev/stdin

Step 2. Mark and Remove Duplicates

sambamba markdup -t [number of threads] --remove-duplicates --tmpdir=[path/to/temp] [input: sorted bam output] [output: bam without duplicates]

Step 3. Generate Realigning Targets

java -d64 -Xms4g -Xmx30g -jar GenomeAnalysisTK.jar -T RealignerTargetCreator -nt [number of threads] -R [reference chromosome] -o [target list file] -I [bam without duplicates] -known [reference vcf file]

Step 4. Realigning around InDel

java -d64 -Xms4g -Xmx30g -jar GenomeAnalysisTK.jar -T IndelRealigner -R [reference chromosome] -I [bam without duplicates] -targetIntervals [target list file] -known [reference vcf file] -o [realigned bam]

Step 5. Base Recalibration

java -d64 -Xms4g -Xmx30g -jar GenomeAnalysisTK.jar -T BaseRecalibrator -nct [number of threads] -l INFO -R [reference chromosome] -I [realigned bam] -known [reference vcf file] -o [recalibrated data table]

Step 6. Print Recalibrated Reads - Optional

java -d64 -Xms8g -Xmx30g -jar GenomeAnalysisTK.jar -T PrintReads -nct [number of threads] -R [reference chromosome] -I [realigned bam] -BQSR [recalibrated data table] -o [recalibrated bam]

Step 7. After Base Recalibration - Optional

java -d64 -Xms4g -Xmx30g -jar GenomeAnalysisTK.jar -T BaseRecalibrator -nct [number of threads] -l INFO -R [reference chromosome] -I [recalibrated bam] -known [reference vcf file] -o [post recalibrated data table]

Step 8. Analyze Covariates - Optional

java -d64 -Xms8g -Xmx30g -jar GenomeAnalysisTK.jar -T AnalyzeCovariates -R [reference chromosome] -before [recalibrated data table] -after [post recalibrated data table] -plots [recalibration report pdf] -csv [recalibration report csv]

Phase 2. Variant Discovery – Calling germline variants

Step 1. Haplotype Caller

java -d64 -Xms8g -Xmx30g -jar GenomeAnalysisTK.jar -T HaplotypeCaller -nct [number of threads] -R [reference chromosome] -ERC GVCF -BQSR [recalibrated data table] -L [reference vcf file] -I [recalibrated bam] -o [gvcf output]

Step 2. GenotypeGVCFs

java -d64 -Xms8g -Xmx30g -jar GenomeAnalysisTK.jar -T GenotypeGVCFs -nt [number of threads] -R [reference chromosome] -V [gvcf output] -o [raw vcf]

Phase 3. Preliminary Analyses

Step 1. Variant Recalibration

java -d64 -Xms512m -Xmx2g -jar GenomeAnalysisTK.jar -T VariantRecalibrator -R [reference chromosome] --input [raw vcf] -an QD -an DP -an FS -an ReadPosRankSum -U LENIENT_VCF_PROCESSING --mode SNP --recal_file [raw vcf recalibration] --tranches_file [raw vcf tranches]

Step 2. Apply Recalibration

java -d64 -Xms512m -Xmx2g -jar GenomeAnalysisTK.jar -T ApplyRecalibration -R [reference chromosome] -input [raw vcf] -o [recalibrated filtered vcf] --ts_filter_level 99.97 --tranches_file [raw vcf tranches] --recal_file [raw vcf recalibration] --mode SNP -U LENIENT_VCF_PROCESSING

Job Scheduling

Torque/Maui is used to manage a large number of jobs to process sequencing samples simultaneously. Optional steps, 6, 7 and 8 in phase 1 were not included in the benchmark since Step 6 PrintRead took 12.5 hours with 9 threads for *Bos Taurus* sample (18 hours with single thread). These optional steps are

not required, but these steps are useful for the reporting purpose. If it is necessary, it can be added as a side workflow to the main procedure. For each job, 9 cores were assigned when 120 concurrent jobs were processed concurrently and 13 cores were used for the test of 80 concurrent jobs.

Data

In addition to the benchmark for human whole genome sequencing data published in the whitepaper, we gathered cow, pig, two sub-species of rice (*japonica* and *indica*) and corn reference genomes from Illumina's iGenome site and Ensembl database. Fortunately, reference variant call data exist as a standard VCF file format for human, cow and pig. A variant data for *japonica* rice were obtained from 3000 Rice Genome on AWS and was modified according to the standard VCF file format. However, the chromosome coordinates in this VCF file do not match to the actual reference chromosome sequences, and we were not able to find matching version of reference variant information from public databases. For *indica* rice and corn, we gathered the variant information from Ensembl and converted them into a compatible VCF format. Whole genome sequencing data were obtained from European Nucleotide Archive (ENA). ENA Sample IDs in Table 1 are the identifiers allow to retrieve sequence data from the site. Although it is not ideal to test an identical input for large number of processes, it is not feasible to obtain large number of similar sample data from public databases.

*Table 2 WGS test data for the different species: * x2 indicates the data is paired end reads. [‡]Test ID column represent identifiers for the sequence data used throughout the test.*

Species		Test ID[‡]	ENA Sample ID	Sample Base Count	Single file Size x2*	Reference Genome size (bp)	Depth of Coverage	Number of variants in Ref
***Homo sapiens* (human)**		Hs1	ERR091571	42,710,459,638	17 GB x2	3,326,743,047	13x	3,152,430
		Hs2	ERR194161	171,588,070,386	54 GB x2	3,326,743,047	52x	
***Bos Taurus* (cow)**		Bt1	SRR1706031	82,272,305,762	35 GB x2	2,649,685,036	31x	93,347,258
		Bt2	SRR1805809	32,681,063,800	12 GB x2	2,649,685,036	12x	
***Sus scrofa* (pig)**		Ss1	SRR1178925	41,802,035,944	19 GB x2	3,024,658,544	14x	52,573,286
		Ss2	SRR1056427	24,901,150,040	10 GB x2	3,024,658,544	8x	
***Oryza sativa* (rice)**	*japonica*	Osj	SRR1450198	49,676,959,200	22 GB x2	374,424,240	132x	19,409,227
	indica	Osi	SRR3098100	12,191,702,544	4 GB x2	411,710,190	30x	4,538,869
Zea mays (corn)		Zm	SRR1575496	36,192,217,200	14 GB x2	3,233,616,351	11x	51,151,183

Benchmark Results

Data Quality

After mapping and sorting of the sequence input files, quality statistics were obtained from the output files of Phase 1, Step 1. SRR17060031 sample is from bovine gut metagenomics study and was not well mapped onto *Bos taurus* UMD3.1 reference genome from Ensembl as expected. The majority of DNAs from bovine gut is foreign and has different sequence composition.

Table 3 Mapping qualities of sequence reads data; obtained by using 'samtools flagstat'. 'Total QC-passed reads' is the number of reads passed the criteria of sequencing quality. Among all QC-passed reads, the number of reads actually mapped on a reference genome and its percentage is on 'Mapped reads (%)' column. 'Paired in sequencing' column is the number of paired reads properly paired by a sequencer. Among the reads properly paired in sequencing, the number of those paired reads mapped on a reference genome as paired reads is listed in 'Properly paired (%) in mapping.

Species	Sequencing Reads	Test ID	Total QC-passed reads	Mapped reads (%)	Paired in sequencing	Properly paired (%) in mapping
Human	ERR091571	Hs1	424,118,221	421,339,198 (99.34%)	422,875,838	412,370,120 (97.52%)
	ERR194161	Hs2	1,691,135,957	1,666,486,126 (98.54%)	1,686,908,514	1,621,073,394 (96.10%)
Cow	SRR1706031	Bt1	813,545,863	29,291,792 (3.60%)	813,520,998	28,813,072 (3.54%)
	SRR1805809	Bt2	327,304,866	316,654,265 (96.75%)	326,810,638	308,600,196 (94.43%)
Pig	SRR1178925	Ss1	416,854,287	379,784,341 (91.11%)	413,881,544	344,614,170 (83.26%)
	SRR1056427	Ss2	249,096,674	228,015,545 (91.54%)	246,546,040	212,404,874 (86.15%)
Rice	SRR1450198	Osj	499,853,597	486,527,154 (97.33%)	496,769,592	459,665,726 (92.53%)
	SRR3098100	Osi	97,611,519	95,332,114 (97.66%)	96,759,544	86,156,978 (89.04%)
Corn	SRR1575496	Zm	364,636,704	358,393,982 (98.29%)	361,922,172	315,560,140 (87.19%)

The rest of samples were properly aligned on the reference genome with high quality; more than 80% of reads paired in sequencing data properly mapped as pairs on reference genomes.

It is also important to check what level of mismatches exists in the aligning results. The estimated variance in human genome is one in every 1,200 to 1,500 bases. This makes 3 million base differences between any two people randomly picked. However, as shown in Table 4, the results are not quite matched to the 3 million base estimation. Ideally, 36 million mismatches should be shown in Hs1 data set since it covers the human reference genome 13 times. However, the rate of mismatches is quite higher than the estimation, and at least one out of two variants reported by the sequencing might be an error.

Table 4 The number of reads are perfectly mapped on a reference genome and the number of reads do not

Test ID	Depth	Mapped reads	Number of reads mapped with mismatches (mm)					
			Perfect match (%)	One mm (%)	Two mm (%)	Three mm (%)	Four mm (%)	Five mm (%)
Hs1	13x	421,339,198	328,815,216 (78.0)	53,425,338 (12.7)	13,284,425 (3.2)	6,842,191 (1.6)	5,140,438 (1.2)	4,082,446 (1.0)
Hs2	52x	1,666,486,126	1,319,421,905 (79.2)	201,568,633 (12.1)	47,831,915 (2.9)	24,862,727 (1.5)	19,052,800 (1.1)	15,568,114 (0.9)
Bt1	31x	29,291,792	25,835,536 (88.2)	2,684,650 (9.2)	338,781 (1.2)	147,841 (0.5)	89,706 (0.3)	70,789 (0.24)
Bt2	12x	316,654,265	158,463,463 (50.0)	68,754,190 (21.7)	29,544,252 (9.3)	17,337,205 (5.5)	12,639,289 (4.0)	10,015,029 (3.2)
Ss1	14x	379,784,341	228,627,231 (60.2)	69,912,403 (18.4)	29,142,572 (7.7)	16,701,248 (4.4)	11,036,852 (2.9)	7,652,513 (2.0)
Ss2	8x	228,015,545	112,216,441 (49.2)	53,739,562 (23.6)	25,132,226 (11.0)	13,874,636 (6.1)	8,431,144 (3.7)	5,375,833 (2.4)
Osj	132x	486,527,154	208,387,077 (42.8)	113,948,043 (23.4)	61,697,586 (12.7)	37,520,642 (7.7)	23,761,302 (4.9)	15,370,422 (3.2)
Osi	30x	95,332,114	54,462,837 (57.1)	17,325,526 (18.2)	8,190,929 (8.6)	5,146,096 (5.4)	3,394,245 (3.6)	2,322,355 (2.4)
Zm	11x	358,393,982	150,686,819 (42.1)	82,912,817 (23.1)	44,823,583 (12.5)	28,375,226 (7.9)	19,093,235 (5.3)	12,503,856 (3.5)

Time Measurement

Total run time is the elapsed wall time from the earliest start of Phase 1, Step 1 to the latest completion of Phase 3, Step 2. Time measurement for each step is from the latest completion time of the previous step to the latest completion time of the current step as described in Figure 1.

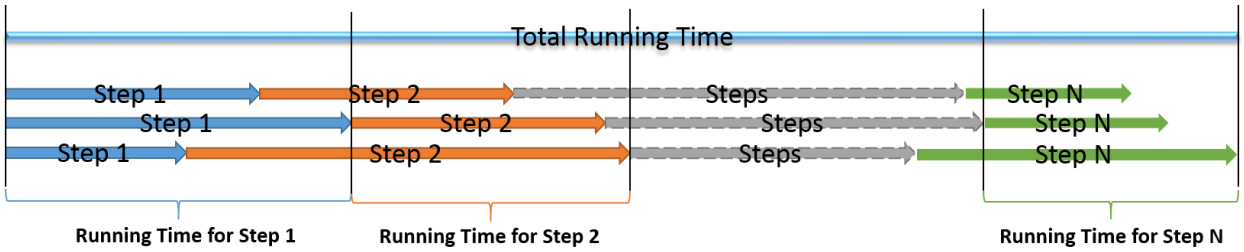

Figure 1 measuring the running time

The running time for each data set is summarized in Table 4. Clearly the input size, size of sequence read files and reference genomes are the major factors affecting to the running time. The reference genome size is a major player for 'Aligning & Sorting' step while the size of variant reference affects most on 'HaplotypeCaller' step.

Table 5 running time for BWA-GATK pipeline

Species		***Homo sapiens* (human)**		***Bos taurus* (cow)**		***Sus scrofa* (pig)**		***Oryza sativa* (rice)**		***Zea mays* (corn)**
								japonica	***Indica***	
Depth of Coverage		13x	52x	31x	12x	14x	8x	132x	30x	11x
Test ID		Hs1	Hs2	Bt1	Bt2	Ss1	Ss2	Osj	Osi	Zm
Total read size, gzip compressed (GB)		34	108	70	22	38	20	44	8	28
Number of samples ran concurrently		80	80	120	80	120	80	120	80	80
Run Time (hours)	**Aligning & Sorting**	3.93	15.79	7.25	5.77	7.53	3.04	9.50	1.18	11.16
	Mark/Remove Duplicates	0.66	2.62	3.45	0.73	1.07	0.27	1.27	0.12	0.72
	Generate Realigning Targets	0.29	1.08	3.12	1.57	0.47	0.27	0.22	0.05	0.26
	Realign around InDel	2.50	8.90	4.00	3.15	2.87	1.83	7.37	1.25	3.18
	Base Recalibration	1.82	6.80	1.39	1.96	2.37	1.01	3.16	0.36	1.91
	HaplotypeCaller	4.52	10.28	2.75	9.33	26.21	14.65	8.95	1.77	16.72
	GenotypeGVCFs	0.03	0.03	0.20	0.05	0.34	0.06	1.12	0.01	0.04
	Variant Recalibration	0.67	0.37	0.32	0.86	0.58	0.56	0.92	0.04	0.46
	Apply Recalibration	0.04	0.04	0.03	0.06	0.03	0.08	0.03	0.01	0.05
	Total Run Time	14.5	45.9	22.5	23.5	41.5	21.8	32.5	4.78	34.5
Number of Genomes per day		133	42	128	82	69	88	89	402	56

Discussion

The running time of the current version, GATK 3.5 is overly slower than the version of 2.8-1 we tested in our white paper. Particularly, HaplotypeCaller in the new workflow took 4.52 hours while UnifiedGenotyper in the older version took about 1 hour. Despite of the significant slow-down, GATK team believes HaplotypeCaller brings a better result, and that is worthy for the five times longer run.

There are data issues in non-human species. As shown in Table 4, for the similar size of inputs, Hs1 and Ss1 show large difference in the running time. The longer running time in non-human species can be explained by the quality of reference data. Aligning and sorting process takes more than twice times in

other mammals, and it became worse in plants. It is known that plants genomes contain large number of repeat sequences which make mapping process difficult. It is important to note that the shorter running time for HaplotypeCaller in rice does not reflect a real time since the size of the reference variant file was reduced significantly due to the chromosome length/position mismatches in the data. All the variant records outside of chromosome range were removed, but position mismatches were used without corrections. The smaller size of the reference variant information and wrong position information the running time of HaplotypeCaller shorter. Corn's reference data is not any better in terms of the accuracy of these benchmark. These data errors are the major causes of longer processing time.

Nonetheless, the results shown here could serve as good reference points for the worst case running time. Once reference data are cleaned up by researchers, the overall running time for other mammals should be similar to the one from Hs1 in Table 4 with a proper scaling of input size. However, it is hard to estimate an accurate running times for non-human species at this moment.