Comparison of sequencing data processing pipelines and application to underrepresented human populations Additional File 1.

Comparison of high coverage whole genome processing pipelines (humans)

Study	Samples details	Pre-processing + mapping	Human genome build	BAM processing	SNPs calling	Indels calling	Callset filtering	Comments
(Wong et al. 2013)	100 Southeast Asian Malays Sequencing platform: Illumina HiSeq2000 ~ 30X	-Alignment: Illumina CASAVA GERALD module	hg19 with PAR masked	-optical duplicate flagging, likely with the CASAVA pipelines -exclusion of some individuals (ancestry; anomalous distribution of insert size)	Two methods: 1-single-sample variant caller in CASAVA, 2-SAMtools v0.1.17 multisample variant-caller mpileup	-Same as for SNPs -For deletions > 50 bp: BreakDancer v.1.1 and VariationHunter v0.3	-Hard filtering	-
(H. L. Kim et al. 2014)	five Khoe-San one Bantu-speaker Sequencing platform: Illumina HiSeq	-bwa v.0.5.9 with default parameters except for '-q 15' to soft-trim low quality bases at the 3' end of reads	hg19	-GATK IndelRealignment -Picard MarkDuplicates	-Obtained diploid consensus sequence with SAMtools mpileup -C 50 Then marked locations as missing data (see Callset filtering) and used SAMtools v0.1.18 to "identify the location of SNPs".	-Same as SNPs calling given the pipeline used (SAMtools).	-Hard filtering	-
(Auton et al. 2015)	High coverage genomes (30 individuals) from 1000 Genomes phase 3 Illumina HiSeq2500	-bwa mem -Adapter clipping with MarkIlluminaAdapters.jar	GRCh37 with decoy	-Indel realignment GATK -BQSR GATK	-NA	-NA	-NA	-
(Besenbacher et al. 2015)	30 Danes (ten trios) Sequencing platform: Illumina HiSeq2000	-Adapter-Removal to: trim the tails of the reads if the Phred quality dropped < 2. collapse 180 bp insert size libraries into longer single end reads -bwa-mem v.0.7.5.a	hg 19 decoy	-SAMtools v.0.1.19 and Picard v.1.96 to "process the alignment files and mark duplicate reads" (at lane level). -Indel realignment with GATK v.2.7.2 (at lane level and after merging) with Mills and 1000G gold standard and dbSNP138. -BQSR with GATK v.2.7.2 at lane level and with	-GATK v2.7.2: HaplotypeCaller for multi-sample genotyping	-Same as for SNPs.	GATK: VQSR -SNPs: datasets: HapMap, Omnichip, 1000G phase1 high confidence variants, dbSNP138 Annotations: QD, MQRankSum, ReadPosRankSum, FS, DP --numBadVariants	-

				dbSNP138.			5000 tranche: 99.5 -INDELS: dataset: Mills and 1000G --maxGaussians 4 --numBadVariants 1000 tranche: 95.0	
(Gudbjartsson et al. 2015)	2,636 Icelanders at ~ 20 X Sequencing platform: Illumina $GAII_x$ and/or HiSeq2000	-BWA v.0.5.7-0.5.9 then alignments merged into a single BAM	hg18	-Alignments for a given individual merged into a single BAM file -Picard v.1.55: duplicates marking -GATK v.1.2-29-g0acaf2d: realignment around indels and BQSR	-Multi-sample variant calling GATK v.2.3.9 (algorithm: NA, but likely UnifiedGenotyper)	-Multi-sample variant calling GATK v.2.3.9 (algorithm: NA, but likely UnifiedGenotyper)	-Hard filtering for SNVs and indels, using GATK recommended filters and additional filters. -Regional filter (simple-repeat regions).	-
(Nagasaki et al. 2015)	1,070 Japanese Sequencing platform: HiSeq 2500	Two strategies: 1-Bowtie2 v.2.1.0 with '-X 2000' option 2-bwa-mem v.0.7.5a-r405 with default option	hg19 decoy		Depending on the mapping strategy: 1-BCFtools v.0.1.17-dev (Bowtie alignment) 2-GATK v2.5-2 (bwa alignment) (algorithm: NA) For downstream analyses, used the outcome of the first strategy (bowtie+BCFtools).	-Short structural variants (<100 bp): GATK HaplotypeCaller v.2.5-2 on bwa alignment -Long structural variants: BreakDancer v.1.1 and Pindel v.0.2.5a3 -CNV: CNVnator v.0.3	-Hard filtering (genotype depth filter per individual, SNV depth filter per locus, genome complexity filter, tool bias filter, population genetics filter).	-
(UK10K Consortium et al. 2015)	~10,000 individuals of which: ~4,000: WG at ~7X from two British cohorts of European ancestry and ~6,000: exomes at ~80X Sequencing platform: Illumina HiSeq	-sequencing reads failing QC removed with the Illumina GA pipeline -bwa v.0.5.9-r16: bwa aln -q15	hg19	-Obtained sorted BAM from SAM files: Picard v1.36 and SAMtools v0.1.11 / or SAMtools v0.1.8 only. -Mark duplicates with Picard MarkDuplicates / or SAMtools rmdup -GATK v.1.1-5-g6f43284, Picard v.1.64, SAMtools v.0.1.16: Realignment around indels, BQSR, addition of BAQ tags, merging and	-SAMtools mpileup -C50 -d 8000/bcftools v. 0.1.18-r579 (pooling all individuals). -GATK v.1.3-21 UnifiedGenotyper for SNVs to recall the sites found by the SAMtools/BCFtools pipeline to generate	-SAMtools mpileup -C50 -d 8000/bcftools v. 0.1.18-r579 (pooling all individuals).	-SNV: GATK VQSR with default parameters and datasets (dbSNP132), tranche 99.5 -indels: pre-filtering due to technical issues + GATK VQSR, did not use dbSNP but used 1000 Genomes	-

				duplicate marking (“follows that used for Illumina low-coverage data in 1000GP”.)	annotations used in VQSR.		Phase I.	
(Fakhro et al. 2016)	108 unrelated Qataris Sequencing platform: Illumina HiSeq2500 (+ 1,268 exomes on one of three platforms) One individual sequenced on three platforms + HiSeqX	-BWA v.0.5.9 with maximum insert size 3 kb	hg19	-SAMtools: PCR duplicates removal -“mapped reads were prepared for variant calling using GATK Best Practices”: GATK realignment of indels and BQSR?	-Group-call with GATK (algorithm and version: NA, but likely UnifiedGenotyper - see Comments)	-Short indels (<300 bp) called with the CASAVA v1.9 pipeline by Illumina	-Hard filtering based on comparing different batches and results for the quadruple-sequenced Qatari.	-Also tested mapping to a Qatari reference genome. -Joint calling of exomes and genomes with GATK UnifiedGenotyper
(Haber et al. 2016)	four Lebaneses eleven Chadians four Greeks Sequencing platform: Illumina HiSeq X or HiSeq2500 Genomes were not the main focus of the paper.	NA	NA	NA	-SAMtools v.1.2 and BCFtools v.1.2: samtools mpileup -q 20 -Q 20 -C 50 | bcftools call -c -V indels	-Same as SNPs calling given the pipeline used (SAMtools + BCFtools)	NA	-
(Malaspinas et al. 2016)	83 Australians 25 Papuans Sequencing platform: Illumina HiSeq X	-adapter trimming (Adapter-Removal-1.5.4) + trimmed leading / trailing stretches of Ns + bases with quality 2. Kept reads of >= 30 bp -bwa-mem -re-aligned with stampy v.1.0.23 (option: keep good mapped reads from bwa)	hg19	-Picard v.1.127 -add read groups -sort reads -merge reads at the library level -duplicate reads marked and merged at the sample level -GATK: indel realignment (with Mills and 1000G gold standard) -Removed reads with mapping quality <30 -SAMtools calmd: recalculated the md tags and added extended BAQs	-Per sample: SAMtools v.0.1.18 and BCFtools v.0.1.17: mpileup -C50 | bcftools	-Same as SNPs calling given the pipeline used (SAMtools + BCFtools)	-Hard filtering	-
(Mallick et al. 2016)	300 samples from 142 worldwide populations (22 of which were produced earlier) Sequencing platform: Illumina HiSeq2000	-adapter trimming: extract and shuffle reads from BAM with htslib, run trimadapt -bwa-mem v0.7.10-r1005-dirty -RG added during alignment	hg19 decoy	-mark optical duplicates with *samblaster* -SAMtools: sort	-Per sample. Biallelic SNPs. Modification of UnifiedGenotyper developed with the GATK team to minimize reference bias.	-FermiKit-0.8 for the 263 fully public samples.	-Hard filtering	-

(Telenti et al. 2016)	10,000 individuals: “representatives of major human populations and ancestries” (8,096 unrelated) Sequencing platform: Illumina HiSeqX	-ISIS Analysis Software v.2.5.26.13	hg38 with Y chr PAR masked	-ISIS Isaac Aligner v. 1.14.02.06: mark duplicates. -Picard v. 1.113-1.131: to “characterize[d]” bam files.	-ISIS Isaac Variant Caller v.2.0.17 with default settings.	-ISIS Isaac Variant Caller v.2.0.17 with default settings.	-Comparison of variants found in NA12878 to the results from the Genome in a Bottle Consortium.	-
(Ameur et al. 2017)	1000 Swedes	-bwa mem -M -R (added read group information)	hg19	-SAMtools: sort alignment -Picard: merge all BAM files corresponding to a sample -GATK IndelRealignment with Mills and 1000G gold standard and dbSNP138 -Picard MarkDuplicates -GATK BQSR with dbSNP138	-HaplotypeCaller CombineGVCF GenotypeGVCFs CatVariants (merge the non-overlapping intervals)	-Same as SNPs calling given the pipeline used.	-VQSR (SNPs and indels) GATK-recommended datasets	-
(Choudhury et al. 2017)	24 South Africans: -eight Coloured -seven Sotho-speakers -eight Xhosa-speakers (Nguni) -one Zulu-speaker (Nguni) Sequencing platform: Illumina HiSeq2000	-Isaac Analysis Pipeline v.2.0.2. During mapping, low quality 3’ ends and of adaptor sequences were trimmed.	hg19	-Isaac Analysis Pipeline v.2.0.2: marking of duplicates, realignment of indels. -GATK v.3.2.2: BQSR. -SAMtools v.1.1-26-g29b0367: flagstats. -Checked that all samples pass thresholds (quality of reads).	Three callsets: 1.Isaac variant caller 2.GATK v.3.2-2 HaplotypeCaller default parameters 3.GATK v.3.2-2 HaplotypeCaller stand_call_conf 50 For approaches 2 and 3: GenotypeGVCFs for the three groups (Coloured, Sotho, Xhosa+Zulu) indepently.	-Isaac variant caller (indels and CNV)	For SNV: > Isaac callset: hard filtering > GATK callsets: VQSR with default parameters and datasets, tranche 99.5 Then took overlap.	-
(Harris et al. 2018)	150 individuals from Peru (Native American and mestizo) ~35X Illumina HiSeqX10	-bwa mem	Hg19	-MarkDuplicates Picard -IndelRealignment GATK -BQSR GATK	-HaplotypeCaller GATK -GenotypeGVCF GATK It is written “variants called using GATK in each individual’s genome”. UnifiedGenotyper is run too.	-See for SNPs	-Hard filtering	Processing was performed at the sequencing center. Do not specify version. www.nygenome.org/wp-content/uploads/2018/01/Whole-Genome-Sequencing-Germline.pdf

(J. Kim et al. 2018)	50 unrelated Korean individuals	-cleaning of raw reads by Sickle: keep quality score >20 and read length >50bp -bwa	hg19	-Removal of PCR duplicates -Indels realignment and recalibration (software: NA)	-GATK-Lite-2.3-9: UnifiedGenotyper by individual	-GATK-Lite-2.3-9: UnifiedGenotyper by individual	-Hard filtering for SNVs and indels.	-
(Jeong et al. 2018)	17 individuals (four Tibetans mother-daughter duos and three Sherpa trios), ~20X (plus 36 low coverage). Illumina HiSeq2500 and 4000	-bwa backtrack v0.7.4 with -q 15 option	hg19	-MarkDuplicates with Picard v1.98 -IndelRealignment GATK v2.8-1 -BQSR GATK v2.8-1 -filtered reads with Phred-scaled mapping quality lower than 30 with SAMtools v1.2	-GotCloud	-GotCloud	-GotCloud	-
(Natarajan et al. 2018)	"TOPMED phase 1" cohort: 8,394 individuals from the USA (various ancestries, three studies: JHS, FHS, OOA)	-Not specified	1000 Genomes hs37d5 build 37 decoy reference sequence	-TOPMED phase 1: GotCloud	-TOPMED phase 1: GotCloud "freeze 3a" (vt discover2 software)	-TOPMED phase 1: GotCloud "freeze 3a"	-TOPMED phase 1: GotCloud "freeze 3a" (hard filtering)	-
	"TOPMED phase 2" (MESA study, 4510 individuals) and FINRISK (n=1165) and EST (n=2255) Illumina HiSeqX		hg19	-Not specified	-HaplotypeCaller GATK v3 -GenotypeGVCF	-HaplotypeCaller GATK v3 -GenotypeGVCF	-VQSR GATK -hard filtering	-
(Okada et al. 2018)	2234 individuals from Japan 20-35X Illumina HiSeq2500 or Illumina HiSeq X Five	-Adaptor trimming with Trimmomatic (v0.36 for datasets 2 and 3) -bwa mem v0.7.5a	GRCh37/ hg19, hs37d5 - decoy	-MarkDuplicates Picard (v1.106 for datasets 1 and 2, v2.5.0 for dataset 3) -Indel realignment and BQSR GATK (v3.2-2, 3.5-0 and 3.6 respectively)	-HaplotypeCaller GATK -GenotypeGVCF GATK (?)	-Same as for SNPs	-Hard filtering -then VQSR GATK -genotype refinement with beagle -and more...	-
(S. Fan et al. 2019)	43 individuals from 22 African groups (analyzed together with 29 previously sequenced – Mallick 2016). Illumina HiSeq2000	-Adaptor trimming -bwa mem version 0.7.12	Hs37d5	-mark optical duplicates with *samblaster* -SAMtools: sort	Per sample. Biallelic SNPs. Modification of UnifiedGenotyper developed with the GATK team for GATK UnifiedGenotyper to minimize reference bias.	-NA	-Hard filtering	Same as (Mallick et al. 2016)
(Lorente-	Nine individuals from	-bwa v0.6.1 aln -n 6 -q	GRCh37	-MarkDuplicates Picard v1.70	-UnifiedGenotyper	-Not called	-VQSR GATK v2.5-	-

Galdos et al. 2019)	African origin. 21-47X Illumina HiSeq2000	15 and sampe	with Epstein-Barr virus	-IndelRealignment GATK v2.5-2 -BQSR GATK v2.5-2	GATK v2.5-2		2 -Defined a callable genome filter (based on their data and other information e.g. TandemRepeatMarker)	
(Serra-Vidal et al. 2019)	21 male individuals (17 North Africans, 2 Basque and 2 Iraqi) ~26X Illumina HiSeq2000	-Assess read quality with fastqc -bwa v0.7.7	hg19	-Merged mapped reads -MarkDuplicates Picard v2.8.3 -IndelRealignment GATK v4.0.12 -BQSR GATK v4.0.12	-UnifiedGenotyper GATK GATK v4.0.12	-Not called	-VQSR GATK GATK v4.0.12 -Defined a callable genome filter (based on their data and other information e.g. TandemRepeatMarker)	-
(Bergström et al. 2019)	787 individuals fro 54 populations (analyzed with 142 previously sequenced – Mallick 2016, Meyer 2012, Raghavan 2015) of which 26: linked-read technology (physical phasing). Illumina HiSeqX ~35X	-Adaptor trimming with bamadapterclip (biobambam package) -Mapping with bwa mem v0.7.12 with the -T 0 parameter -Marking of duplicates with bamstreamingmarkduplicates (biobambam package).	GRCh38 with decoy	-None.	-HaplotypeCaller version 3.5.0. genotype priors without bias towards the reference allele through the “--input-prior 0.001 –input-prior 0.4995” arguments; “--pcr_indel_model NON” for the PCR-free libraries; “--includeNonVariantSites”. -GenotypeGCVF (guessed from the text).	-Same as for SNPs.	-Hard filtering because they wanted to filter both variant and non-variant sites (VQSR: only for variant sites). Filter on GQ/RGQ and DP. -VQSR -excess heterozygosity with BCFtools fill-tags plugin.	Wellcome Sanger Institute sequencing facility automated pipeline

Comparison of high coverage whole genome processing pipelines (non-humans)

Study	Samples details	Pre-processing + mapping	reference	BAM processing	SNPs calling	Indels calling	Variant recalibration	Comments
(Z. Fan et al. 2014)	one Tibetan macaque analyzed together with four comparative genomes Sequencing platform: Illumina HiSeq2000	-Bowtie2 under local alignment algorithm with very sensitive model	rheMac2 (Indian rhesus macaque)	-Picard: mark duplicate reads -GATK: indel realignment followed by fixing the mate pair information with FixMateInformation.jar -GATK: “empirical BQSR”. Started by doing a variant call on the Tibetan macaque to get a list of SNPs to exclude from BQSR (to replace the dbSNP dataset used for humans).	-GATK: UnifiedGenotyper (each sample separately). Standard minimum confidence thresholds to 0.	-Same as SNPs calling.	-Hard filtering.	-
(Z. Fan et al. 2016)	nine wolves one coyote one golden jackal analyzed together with 23 comparative canine genomes Sequencing platform: Illumina HiSeq2000	-Bowtie2 under local alignment algorithm with very sensitive model	CanFam3.1	-Picard: mark duplicate reads -GATK: indel realignment followed by fixing the mate pair information with FixMateInformation.jar -GATK: “empirical BQSR”. Started by doing a variant call on the Tibetan macaque to get a list of SNPs to exclude from BQSR (to replace the dbSNP dataset used for humans).	-GATK: UnifiedGenotyper (each sample separately). Standard minimum confidence thresholds to 0.	-Same as SNPs calling.	-Hard filtering.	-Same as (Z. Fan et al. 2014). The same pipeline was also used in studies not reviewed here (Zhang et al. 2014; Freedman et al. 2014)
(Friedenberg, Meurs, and Mackay 2016)	64 dogs across eight breeds Sequencing platform: Illumina HiSeq 2000 or 2500	-Trimmomatic v0.32: trim reads -bwa v.0.7.10	canFam3	-Picard v.1.115 (analysis performed not specified – mark duplicates?) -GATK v.3.4: BQSR and indel realignment	-GATK: HaplotypeCaller	-GATK: HaplotypeCaller	-VQSR: -dbSNP139 -Illumina CanineHD BeadChip -SNPs: tranche 99.9, indels: 99	-
(Friedenberg, Lunn, and Meurs 2017)	20 dogs (Standard Poodles)	-Trimmomatic v0.32: trim reads -bwa v.0.7.10	canFam3	-Picard v.1.115 (analysis performed not specified – mark duplicates?) -GATK v.3.4: BQSR and indel realignment	-GATK: HaplotypeCaller	-GATK: HaplotypeCaller	-VQSR: -dbSNP139 -Illumina CanineHD BeadChip -SNPs: tranche 99.9, indels: 99	-Same as (Friedenberg, Meurs, and Mackay 2016)
(Bimber et	21 rhesus macaques	-Trimmomatic adaptive	MacaM	-GATK: indel realignment	-GATK:	-Same as SNPs	-Hard filtering	-

al. 2017)	including two trios and multiple parent/child pairs Sequencing platform: HiSeq2000 or 3000	quality trimming -bwamem		-Picard: marking of duplicates	HaplotypeCaller + GenotypeGVCFs (emitting and calling threshold: 20)	calling		
(Pfeifer 2017)	seven publicly available green monkey (three generations pedigree) Sequencing platform: Illumina HiSeq2000	-bwa-mem v.0.7.13	Repeat masked *Chlorocebus sabaeus* (green monkey) v.1.1 + Epstein-Barr virus	-Per read group: -Picard v.2.1.1: Marking duplicates -GATK v.3.5: indel realignment -GATK v3.5: BQSR with ~500 k variants from the genome-wide SNPs panel of the Vervet Genetic Mapping Project -Per sample: -Picard v.2.1.1: Marking duplicates	-GATK v3.5: HaplotypeCaller, default heterozygosity rate 0.001 + GenotypeGVCF: 1-single-sample, 2-joint genotyping.	-Same as SNPs calling	-Hard filtering	-