Data processing and output of the Qatar Genome Project Pilot phase to be made available to the PPM program

For 3000 individuals, raw whole genome read data is generated by Illumina HiSeq X Ten[1] sequencers and converted from the native BCL format to paired end FASTQ[2] format using bcl2fastq[3][v2.16]. The quality of the raw data is then assessed using fastqc[4]. Data passing quality control is then aligned to the reference genome sequence (build GRCh37 (hs37d5)[5]) using the bwa-kit[6] aligner[v7.12]. Variant calling is performed using GATK[7] haplotype caller[v3.3] and annotation of the resulting VCF[8] is performed using snpeff[9][v4.1b] and the following databases(dbsnp[10] v138 and dbNSFP[11] v2.9) . In the future, we will also provide mapping and variant calling using the new GRCh38 reference genome[12]. Annotation using VEP[13] tool will be made available as well.

The data types which will be provided to the PPM investigators are BAM and VCF.

[1] http://www.illumina.com/systems/hiseq-x-sequencing-system/system.html
[2] https://en.wikipedia.org/wiki/FASTQ_format
[3] https://support.illumina.com/downloads/bcl2fastq-conversion-software-v216.html
[4] http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[5] ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/README_human_reference_20110707
[6] https://github.com/lh3/bwa/tree/master/bwakit
[7] https://www.broadinstitute.org/gatk/
[8] https://github.com/samtools/hts-specs
[9] http://snpeff.sourceforge.net/
[10] http://www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi?view+summary=view+summary&build_id=138
[11] https://sites.google.com/site/jpopgen/dbNSFP
[12] http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/
[13] http://www.ensembl.org/info/docs/tools/vep/index.html