WGS QC PIPELINE V.0.3

1. Pre-VCF QC – Done in production. Check exclusion procedures.

- Contamination estimates by sample
- Faction of chimeric reads by sample
- GC content by coverage
- GC/AT dropout relative to mean coverage
- Insert size distribution per sample

2. General Sample QC – Use to exclude potential problematic samples. Use well-behaving SNPs (prune Omin chip data). Exclude problematic regions (chr6: 25.8-36 Mbp, chr8: 6-16 Mbp, chr17: 40-45 Mbp).

- Identify duplicates and relatedness via kinship matrix
- Check sex concordance (done on the non-filtered dataset)
- MDS to determine population structure (exclude potential problematic samples)

General stats: N. of SNPs, N. of indels, N. of novel SNPs (not in dbSNP), N. multiallelic sites, dbSNP rate, % of variants with MQ = 0.

3. Check VQSLOD cut-offs

- Plots for different levels of VQSLOD threshold: FS, QD, DP, Ti/Tv, MQRankSUM, ReadPosRankSum, MQRankSum, inbreedingcoef, insertion/deletion ratio.

4. Sample-specific QC – All stats and plot are reported for four subgroups: SNP-Exomes / INDEL-Exomes / SNP-WG / INDELS – WG. only PASS.

- Indels-related plots: deletions-duplications size distribution, insertion/deletion ratio per sample. N. insertion and N. deletion per sample.
- Sequencing depth (DP)-related plots: mean/median DP per sample, % of the genome > 30x per sample
- Ti/Tv per sample
- Heterozygous/Homozygous ratio per sample.
- N. of singletons per sample.

5. Variant-specific QC – Some metrics are used to exclude variants before association analysis. Stats and plot are reported for four subgroups: SNP-Exomes / INDEL-Exomes / SNP-WGS / INDELS – WGS. ONLY PASS.

- Distribution Hardy-Weinberg P-value per variants
- Genotype quality per variants
- Call rate distribution per variant.
- % of variants with GQ > 20 by variant-specific DP.
- Variant-specific GQ by variant-specific DP (exclude variants with low DP but high GQ).
- Ti/Tv by DP, Ti/Tv by allele frequency, Ti/Tv by VQSLOD.
- Allelic balance plot: % of samples with heterozygous deviating from the 30/70 ratio per variant., allelic balance distribution, allelic balance distribution by allele frequency.
- Mendelian errors in trios
- Differences in missing calls in cases vs. controls per variant.

6. PCA analysis. Use well-behaving SNPs (prune Omin chip data). Exclude problematic regions.

Overlay PCs from actual data with PCs from EXAC and SISU exomes.

7. Concordance analysis

Check concordance with chip data. Plot non-reference sensitivity and non-reference discrepancy rate per sample.