EXOME-CHIP QUALITY CONTROL SOP
Version 5, 2012-11-20

This SOP specifies the use of zCall to post-process Gencall results. This SOP uses Version 3 “GenomeStudio - Thresholds derived from GenomeStudio report”. The zCall distribution is available for download at:

https://github.com/jigold/zCall

As of writing the current zCall version is 3.3. There have now been 4 distributed versions of zCall (0 to 3). Since version 1 there have been no significant source code changes, therefore there should be no need to recall data run on versions 1 or later. Otherwise our advice is to use the latest available version.

1) Capture IDATs from iScan scanner, along with sample sheet.

- Multi-Scanner protocol. If multiple scanners are used and calibration of scanners is unknown, then it will be necessary to record which scanner each chip was run on.

2) Call using Illumina Genome Studio (Gencall module in the BeadStudio/GenomeStudio software).

- Manifest file (HumanExome-12v1_A.bpm).
- Import cluster positions from the standard Illumina Cluster file (HumanExome-12v1.egt).
- Use the default Gen Call Threshold of 0.15.
- Output data to a SNP-wise zCall matrix as required by zCall. See zCall file InstructionsForGeneratingGSreport.txt.
- Multi-Scanner protocol. Genome Studio Projects should be generated and matrices output per scanner.

3) Extract raw Gencall calls to PED. This can be done numerous ways. The zCall distribution includes a script for this purpose (convertReportToTPED.py).

The zCall README specifies the use of Plink (--update-alleles) to update the alleles from the A/B calls to the TOP allele calls. In order to harmonise allele specification, we have made available a file that correctly encodes the top alleles:

http://www.well.ox.ac.uk/~wrayner/strand/HumanExome.A.update_alleles.txt.gz

4) Run QC on Gencall data.

Firstly, perform an initial removal of obvious sample and site failures [e.g. call rate <90% at an individual or SNP level].

Then perform QC with stringent exclusion criteria:

- Call rate (plot call rates per sample and use an appropriate cut off- Based on our experience we would recommend 98% threshold)

- Heterozygosity (we strongly recommend looking at global heterozygosity for >1% SNPs, and <1% SNPs for excluding outliers)
- Gender discordance
- GWAS discordance (if GWAS data available)
- Fingerprint concordance
- PCA outliers (using the Ancestry Informative Markers)

NOTE: We have provided recommendations based on our experience, but exclusion criteria are variable. Visual inspection of call rate/heterozygosity plots should help in deciding these. Also, the number of variable sites in the total dataset makes for a good indicator as to whether individual QC steps are working.

There is no need to run SNP exclusions at this stage – zCall internally performs a strict SNP QC as part of the script findMeanSD.py (incl. call rate > 99%, MAF > 5%, HWE > 0.00001).

5) Exclude samples from the zCall matrices. The zCall distribution includes a script for this purpose (dropSamplesFromReport.py).

6) Calibrate zCall as per zCall README. This will need to be done per scanner. The best advice is to use at least 1000 samples for zCall calibration.

The z-value calibration may be skipped by using a default z-value of 7. According to the zCall README the value 7 works best “when applied to real data”.

7) Run zCall using chosen z-values and calibrations. zCall only adds calls where the original Gencall did not call, therefore valid Gencall data will be left unaltered. It will be necessary to update the alleles from the A/B calls to the TOP allele calls.

8) Chip TOP allele annotations need to be updated to the forward strand of build 37. The strand file is available at:

http://www.well.ox.ac.uk/~wrayner/strand/HumanExome-12v1_A-b37-strand.zip

Usage instructions, including scripts, are available here:

http://www.well.ox.ac.uk/~wrayner/strand/

9) Apply sample exclusions from Gencall (step 4) to the zCall dataset. Run a SNP QC on Gencall data and apply any exclusions to the zCall dataset:

- Call rate (recommend 95% threshold)
- HWE (recommend removing variants with HWE p value < 1e^{-4})
- Cluster separation score (variable according to dataset, example cut-off 0.4)

10) Run a secondary QC on zCall data. This step is primarily designed to verify your QC and should not result in significant exclusions.

- Sample based:

- Call rate (recommend 99% threshold)

- Heterozygosity

- Variant based:

- Call rate (recommend 99% threshold)

11) Post calling QC

- Strategy for dealing with population outliers
- Strategy for dealing with cryptic relatedness
- Compare allele frequencies with 1000 Genomes
- Metrics for identifying poorly performing variants surviving zCall (e.g. HWE, intensity data, visualization of cluster plots)
- Generation of “traffic light” system for poorly performing variants

---

For any queries/feedback please contact any of us from the Oxford team:

Anubha Mahajan: anubha@well.ox.ac.uk
Neil Robertson: neilr@well.ox.ac.uk
Will Rayner: wrayner@well.ox.ac.uk