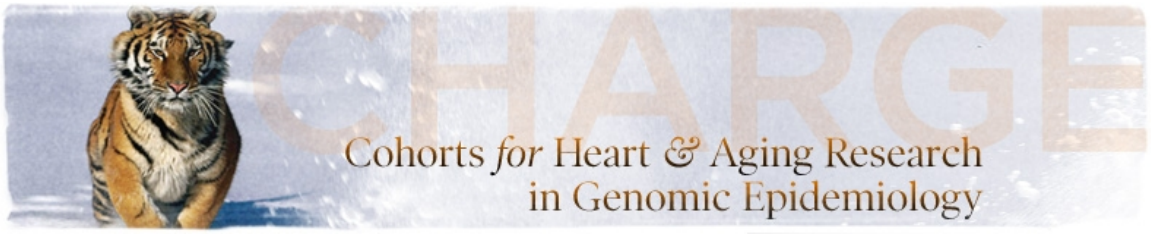


**CHARGE**
**Illumina Human Exome Chip**
**Best Practices and Joint Calling Protocol**

**Objective:** Due to the low minor allele frequency of the ~250,000 markers on the Illumina ExomeChip, the Genome Studio clustering algorithm has limited ability to accurately detect and assign genotype calls. Thus, cohorts primarily from the CHARGE consortium have agreed to have the raw data analyzed collectively at The University of Texas Health Science Center at Houston (UT Houston). Below is a description of the criteria that will be used to generate the joint genotype calls and cluster file (.egt). Importantly, all projects require user decisions based on the data and metrics will be updated accordingly.

Laboratory staff performs the following steps:

1) Clustering software used is Illumina GenomeStudio 2011.1 with the following parameters:
    a. Cluster algorithm GenTrain 2.0
    b. No-call threshold 0.15
    c. "Exclude Female Y- SNPs from SNP Statistics" checked

2) Load data using sample sheet and IDATs (2 per sample: Grn.idat and Red.idat) provided from each cohort via sftp. Sample sheet templates will be provided to each genotyping center. Variables requested will include the following:
    a. Sample ID
    b. Cohort
    c. Sample Type (DNA or WGA)
    d. Race (self-reported)
    e. Gender
    f. Sample Plate
    g. Sample Well
    h. Chip (Sentrix) Barcode
    i. Chip (Sentrix) Position
    j. Replicate ID
    k. Parent 1 (Father ID)
    l. Parent 2 (Mother ID)

    *A unique sample identifier for the aggregate project will be created by combining the cohort abbreviation and sample ID (ex: ARIC_J123456).
    **Cluster file (.egt) provided by Illumina was used when first importing samples into the project.

Creating the cluster file:

3) Exclude duplicate samples by best run (use p10GC).

4) Exclude samples with call rate < 0.99.

5) Plot histogram of Index vs. p10GC and look for outliers to determine if additional poor performing samples should be excluded. This step is strongly recommended if using WGA material.

6) Cluster using project data (only samples with call rate ≥ 0.99) and calculate SNP, heritability and reproducibility, and sample statistics.

7) In the SNP Table, set the following conditions:
    a. "SNP Properties" > "Expected Number of Clusters of Y SNPs" to 2
    b. "SNP Properties" > "Expected Number of Clusters of mtSNPs" to 2.
    *This step automatically re-clusters Y and MT SNPs when exclusion criteria above have been applied.

8) Review and visually inspect the following non-autosomal SNPs. Manually recluster as needed, but do not zero out SNPs or samples for just a few errors.
    a. Y SNPs to identify males
    b. X SNPs should show no male subjects as heterozygotes
    Pseudoautosomal (PAR) SNPs (present on both X and Y) may show male heterozygotes. Keep if a SNP has male heterozygotes and all female homozygotes.
    c. XY SNPs
    d. MT SNPs

9) Apply filter criteria to the project based on the following conditions and visually inspect and manually re-cluster autosomal SNPs when possible (exclude X, Y, XY and MT loci since all were reviewed in step 8 above). Do not zero out SNPs.

    a. Call frequency between 0.95 and 0.99
    b. Cluster separation < 0.4
    c. AB frequency > 0.6
    d. AB R mean < 0.2 (will identify low intensity SNPs)
    e. Het excess > 0.1
    f. Het excess < -0.9
    g. AA theta mean between 0.2 and 0.3
    h. BB theta mean between 0.7 and 0.8
    i. AB theta mean between 0.2 and 0.3
    j. AB theta mean between 0.7 and 0.8
    k. AA theta deviation > 0.025 (determined by histogram of data)
    l. AB theta deviation ≥ 0.07
    m. BB theta deviation > 0.025 (determined by histogram of data)
    n. AB frequency = 0 and minor allele frequency > 0 (will identify missed AB clusters)
    o. AA frequency = 1 and call rate < 1 (will identify missed AB clusters)
    p. BB frequency = 1 and call rate < 1 (will identify missed AB clusters)
    q. MAF < 0.0001 and call rate ≠ 0.

10) Calculate SNP, heritability and reproducibility, and sample statistics.

11) Visually inspect SNPs with the following criteria and re-cluster those that look recoverable.
    a. Parent-Parent-Child (PPC) error > 1.
    b. Parent-Child (PC) error > 1.
    c. Replicate error > 2.

12) Calculate SNP, heritability and reproducibility, and sample statistics.

    Reviewing aggregate project data:

13) Bring back all samples that had been previously excluded based on call rate < 0.99.

14) Calculate SNP, heritability and reproducibility, and sample statistics.

15) Exclude samples by best run (use p10GC).

16) Plot Index vs. p10GC. Visually determine cutoff for outliers. For this project, we identified outlier samples with p10GC < 0.38.

17) Calculate log R deviation using CN Metrics Report and import into sample table. Plot LogRDev by sample type (DNA and WGA), visually inspect and flag outliers. Other reports, such as Heritability and Reproducibility, DNA report, etc. can also be exported.

18) Gender estimations are not accurate due to low MAF and will not be used as a QC criterion within the Genome Studio project.

19) Exclude samples with call rate < 0.97 or p10GC < 0.38 (as determined by project data).

20) Calculate SNP, heritability and reproducibility, and sample statistics.

21) Repeat visual inspection of SNP filter criteria from steps 8-11. If any re-clustering was done, calculate SNP, heritability and reproducibility, and sample statistics.

    *Steps 19-21are necessary in order to observe and correct for batch effects in aggregate projects that contain samples from multiple cohorts with various ethnicities that were genotyped at several core laboratories. These steps may not be necessary for single cohort projects. Exclude SNPs with obvious batch effects.

22) Export list of SNP IDs for the following criteria to create SNP exclusion flag file (applicable to all cohorts).
    a. Parent-Parent-Child (PPC) error > 1.
    b. Parent-Child (PC) error > 1.
    c. Replicate error > 2.

23) Optionally, export SNP table prior to exclusion of SNPs in final project.

24) Exclude (zero out) SNPs based on the following criteria (applicable to all cohorts in project):
    a. Call frequency < 0.95
    b. Cluster separation < 0.4
    c. AB R mean < 0.2
    d. Het excess > 0.1
    e. Het excess < -0.9
    f. AA theta mean > 0.3
    g. BB theta mean < 0.7
    h. AB theta mean < 0.2 or > 0.8
    i. AA theta deviation > 0.06
    j. AB theta deviation ≥ 0.07
    k. BB theta deviation > 0.06

25) Export cluster file (.egt) and list of SNPs excluded for distribution.

26) Bring back all samples that had been previously excluded. Exclude by best run (p10GC) if necessary. Do not update any statistics. Current statistics will be based on removal of samples with call rate < 0.97 or p10GC < 0.38. Export sample table report so that a sample exclusion flag file with p10GC metrics can be created to provide to the individual cohorts.

27) Export the following reports:
    a. Final Report (by cohort) – genotype data references forward strand based on Ilumina annotation file. Additional annotation data based on HG19 available separately.
    b. Intensity Data Report (all cohorts) – X, Y, R and theta (for alternate clustering algorithm analyses).
    c. Genome Studio File (by cohort) – can only be viewed with Illumina software.
    d. Heritability and reproducibility (by cohort) – can be used to identify mispaired samples and plating errors.

28) Use final reports exported in step 27 to format data in PLINK (.bed, .bim, and .fam).
    a. 831 duplicate SNPs separated into second set of PLINK files.
    b. Exclude intensity only markers.

29) Provide PLINK files to statisticians for further quality control

    <u>Statistician performs the following steps:</u>
    - AGES will be excluded from the following QA analyses.
    - CHS will have additional SNPs excluded since WGA was used. CHS was called separately with similar best practices criteria.

30) Races pooled, CHS excluded, flag gender mismatch if minor allele frequency (MAF) ≥ 0.01 and MAF ≥ 0.05. Use same list of SNPs to flag gender mismatch in CHS.

31) Races pooled, flag monomorphic SNPs (must be monomorphic in all races).

32) Races pooled, flag individuals with missing data rate > 5%, > 3% and > 1%.

33) Races pooled, CHS excluded, flag SNPs with missing data rate > 5%.

34) Calculate identity-by-descent (IBD) allele sharing for all pairs after removing SNPs with missing data > 1%, LD $r^2 \geq 0.3$, and MAF ≤ 0.05. Flag individuals with high degrees of relatedness (pi-hat > 0.125). If twins are present, keep the individual with the least amount of missing SNP data and remove the twin with the most amount of missing SNP data. Plot HapMap samples separately to identify outliers.

35) Race specific. Restrict Hardy Weinberg calculation to autosomal SNPs only with MAF≥0.05 and unrelated individuals. Use QQ plot (ggplot2: R package) to determine acceptable threshold.

36) Race specific, EA and AA only Determine allele frequency differences among all cohorts (batches) for each SNP using Fisher exact tests (2 x 10 table for EAs; 2 x 8 table for AAs). For SNPs with $p<0.05$ in previous step, perform pair-wise Fisher exact tests to determine SNPs with significantly different allele frequencies by cohort ($p<10^{-6}$). Cohort-specific SNP flags will be provided.

37) Principal components (PCs) analysis of all individuals (races pooled) using EIGENSTRAT. Exclude SNPs if missing data > 1%, LD $r^2 \geq 0.3$, MAF ≤ 0.05 and HWE cutoff as determined in step 35. View scatter plot of PC1 vs. PC2, and PC1 vs. PC3. View PC1 - PC5 using parallel coordinates plot. Data will be discussed with project investigators to determine flagging. Will provide final SNP list used to cohorts.

38) Estimate F (inbreeding coefficient estimate) to determine sample contamination events. Generate histogram of data distribution and discuss with project investigators to determine cutoff. This suggests an individual has *fewer* homozygous genotypes than one would expect by chance at the genome-wide level.