# Illumina Human Exome & OmniExpress Chip processing v1.1

* Below is a description of the criteria that will be used to generate genotype calls (.report) and cluster file (.egt) for Illumina Human Exome & OmniExpress Chips on the BRC Genomestudio Workstation and compute cluster.
* All projects require user decisions based on their own data and the guidelines set out below.
* We strongly recommend exploring all data in Genomestudio before applying any of the filters below.
* We do not recommend blindly applying any filters or exporting data until users have had a look at their own data.
* Genomestudio provides a wealth of data that can inform users on the general quality of their data and SNP clustering performance.
* We have created filters for all stages below that users can apply
* If you decide to modify any filter, please do not overwrite our pre-defined filters, but save with using the filter name plus your Initials ie FILTERNAME.flt > FILTERNAME\_SJN.flt).
* Filter extension are \*.flt files will need to be placed in C:\Users\<username>\AppData\Roaming\Illumina\GenomeStudio to be visible in Genome Studio
* Projects with < 100 samples will need to be manually edited in Genomestudio.
* We do not recommend using any SNP calling software for < 100 samples.
* Due to the low minor allele frequency of the ~250,000 markers on the Illumina Exome Chip, the Genome Studio clustering algorithm has limited ability to accurately detect and assign genotype calls.
* See “**GenomeStudio\_GT\_Module\_v1.0\_UG\_11319113\_RevA\_tcm240-21464.pdf** ” for detailed descriptions on using GenomeStudio.

1. Load data using sample sheet and IDATs.
2. Sample sheet templates will be provided for each genotyping project. Sample Sheet format = \*.cvs. Variables requested will include the following columns:
   1. Sample ID
   2. Cohort
   3. Sample Type (DNA or WGA)
   4. **Race (self-reported)\***
   5. **Gender\***
   6. Sample Plate
   7. Sample Well
   8. Chip (Sentrix) Barcode
   9. Chip (Sentrix) Position
   10. Replicate ID
   11. Parent 1 (Father ID)
   12. Parent 2 (Mother ID)
   13. Path (Full path to data directory containing IDAT files – to be filled in by BRC Lab)

**\*Without “Gender” and/or “Race”, processing of the project by BRC Bioinformaticians will not proceed, unless the user agrees on dropping the sex chromosome calling and basic quality control steps - this is ill advised for all GWAS studies.**

1. Clustering software used is Illumina GenomeStudio 2011.1 with the following parameters:
   1. Select Tools > Options > Project and uncheck “plot excluded samples”, check “Exclude Female Y-SNPs from SNP Statistics” and click “Use for all New Projects. Note, even though you have done this, Genomestudio will often reset this to default, you should check these settings every time you open a project.
   2. For < 100 samples use the \*.egt otherwise cluster using project data (this takes longer than using the \*egt, but lets the data speak for its self.
   3. Cluster algorithm GenTrain 2.0
   4. No-call threshold 0.15
   5. Select Calculate Sample & SNP statistics
   6. Exclude duplicate samples by best run (use p10GC).
   7. Exclude samples with call rate < 0.90
   8. Plot 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. This step will also help pick out obvious batch effects.
   9. Tip : to improve call rates the user could re-cluster all SNPs once poor samples have been removed. Then re- calculate Sample & SNP statistics.Poor samples and outliers throw GenCall out of whack. Re clustering can generate tighter and cleaner clusters once bad samples have been removed.
2. Save Project as: [PROJECT\_NAME]\_[DATE]\_01.bsc.
3. Create a cluster file in which the X chromosome loci and clustered using only the female samples.
   1. Select all the male samples in the Samples Tables.
   2. Right click the selected samples to launch the context menu
   3. Select Exclude Selected Samples
   4. Do not update SNP statistics
   5. Filter the SNP Table so that only X chromosome SNPs are displayed
   6. If there are Intensity only SNP, identified by a 1 in the Intensity only column, filter out these loci
   7. Select all displayed X chromosome SNPs
   8. Right click the selected SNPs to launch the context menu
   9. Select Cluster Selected SNPs
   10. Do not update SNP statistics because males need to be reintroduced
   11. Select all males previously excluded and select Include Selected Samples
   12. Do not update SNP statistics
   13. In the SNP table, select the displayed X loci
   14. Right click and select Update Selected SNP Satitstics
   15. To properly include these changes select the calculate button in the Samples table
4. Filter the SNP Table so that only Y and MT chromosome loci are displayed.
   1. Select all displayed Y & MT chromosome SNPs
   2. Right click the selected SNPs to launch the context menu
   3. Select SNP Properties
   4. Set Expected Number of Clusters to 2
   5. This step re-clusters Y and MT SNPs:
      * Select all displayed Y & MT chromosome SNPs
      * Right click the selected samples to launch the context menu
      * Select Cluster Selected SNPs
5. Review and visually inspect the following non-autosomal SNPs. Gender information is required for this step.
   1. Y SNPs to identify males.
   2. Y SNPs should show no male subjects as heterozygotes.
   3. X SNPs should show no male subjects as heterozygotes.
   4. Pseudoautosomal (PAR) SNPs (present on both X and Y) may show male heterozygotes. Keep if a SNP has male heterozygotes and all female homozygotes.
   5. XY SNPs
   6. MT SNPs
   7. Calculate Sample & SNP statistics
   8. From the Samples Table select Column Chooser and show Gender Est from Hidden Subcolumns
   9. Select all samples, right click and select Estimate Gender for selected samples
   10. Do not populate Gender column with result of the calculation
6. 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 5-7 above). Do not zero out SNPs until you have manually checked these SNPs and hare happy to proceed. **In some instances right clicking the filtered SNPs to launch the context menu, and selecting Cluster Selected SNPs may help rescue certain loci without having to manually re-cluster SNPs**. We have created filters for many of these settings that users can apply. Go to SNP Table view and click the filter button (funnel shaped button). Select from the dropdown menu the filter to apply > select load > select ok.. Do not blindly apply these filters. Manually edit filters guided by your specific data set. Often stepping through many of these filters one at a time works best. Tip: Filtering by call rate and re-clustering will save many variants (data and chip dependent).

**SNP Review Filters**

* 1. AB R mean < 0.2 (will identify low intensity SNPs)
  2. Call frequency !=0 and between 0.95 and 0.99
  3. Cluster separation < 0.4
  4. AB frequency > 0.6
  5. Het excess > 0.1
  6. Het excess < -0.9
  7. AA theta mean between 0.2 and 0.3
  8. BB theta mean between 0.7 and 0.8
  9. AB theta mean between 0.2 and 0.3
  10. AB theta mean between 0.7 and 0.8
  11. AA theta deviation > 0.025 (determined by histogram of data)
  12. AB theta deviation ≥ 0.07
  13. BB theta deviation > 0.025 (determined by histogram of data)
  14. AB frequency = 0 and minor allele frequency > 0 (will identify missed AB clusters)
  15. AA frequency = 1 and call rate < 1 (will identify missed AB clusters)
  16. BB frequency = 1 and call rate < 1 (will identify missed AB clusters)
  17. MAF < 0.0001 and call rate ≠ 0.

1. Update SNP and Sample statistics
2. Save Project as : [PROJECT\_NAME]\_[DATE]\_02.bsc.
3. Save Samples Table: [PROJECT\_NAME]\_[DATE\_Sample\_Table\_02.txt
4. Save Paired Sample Table : [PROJECT\_NAME]\_[DATE]\_ Sample\_Table\_02.txt
5. Open SNP Table tab and use column chooser to add LMN\_STRAND to SNP Table
6. Save SNP Table : [PROJECT\_NAME]\_[DATE]\_ SNP\_Table\_02.txt
7. Select File > Export Cluster Positions > save as [PROJECT\_NAME]\_[DATE].egt
8. Create input for zCall/Opticall pipeline.
   1. Remove all filters to display all remaining SNPs and Samples
   2. Select Full Data Table
   3. Select Column Chooser
   4. From the Displayed Columns window select
      * Index
      * Address
      * Gen Train Score
      * Frac A
      * Frac C
      * Frac G
      * Frac T
   5. Click Hide
   6. From the Displayed Subcolumns Window select
      * Score
      * Theta
      * R
   7. Click Hide
   8. From the Hidden Subcolumns Window select
      * X
      * Y
   9. Click Show
   10. Select all SNPs
   11. Select Export displayed data to file
   12. Save as [PROJECT\_NAME]\_[DATE]\_intesity\_data.report