Module 0. Harmonization of genotype data

The steps in the “Harmonization of genotype data” are intended to be executed on each genotype batch to standardize the data.

**Table 1.** Overview of the genotype batches

| Batch | Genotyping array | Current genotyping build |
| --- | --- | --- |
| HARVEST12good (m12a) | HumanCoreExome12v1.1 | GRCh37 (hg19) |
| HARVEST12bad (m12b) | HumanCoreExome12v1.1 | GRCh37 (hg19) |
| HARVEST24 | HumanCoreExome24v1.0 | GRCh37 (hg19) |
| NORMENT-Jan-15 | OmniExpress24v1.0 | GRCh37 (hg19) |
| NORMENT-Jan-15 | OmniExpress24v1.0 | GRCh37 (hg19) |
| NORMENT-May-16 | OmniExpress24v1.2 | GRCh38 (hg38) |
| ADHD | OmniExpress24v1.2 | GRCh37 (hg19) |
| NORMENT-Feb-18 | Global Screening Array v.1.0 | GRCh38 (hg38) |
| Rotterdam1 | Global Screening Array v.1.0 | GRCh37 (hg19) |
| Rotterdam2 | Global Screening Array v.1.0 | GRCh37 (hg19) |
| ADHD2 | Global Screening Array v.1.0 | GRCh38 (hg38) |
| NORMENT-Feb-20-V1 | Global Screening Array v.1.0 | GRCh38 (hg38) |
| NORMENT-Feb-20-V3 | Global Screening Array v.3.0 | GRCh38 (hg38) |
| NORMENT-Aug-20-996 | Global Screening Array v.3.0 | GRCh38 (hg38) |
| NORMENT-Aug-20-1029 | Global Screening Array v.3.0 | GRCh38 (hg38) |
| NORMENT-Nov-20-1066 | Global Screening Array v.3.0 | GRCh38 (hg38) |
| NORMENT-Nov-20-1077 | Global Screening Array v.3.0 | GRCh38 (hg38) |
| NORMENT-Nov-20-1108 | Global Screening Array v.3.0 | GRCh38 (hg38) |
| NORMENT-Nov-20-1109 | Global Screening Array v.3.0 | GRCh38 (hg38) |
| NORMENT-Nov-20-1135 | Global Screening Array v.3.0 | GRCh38 (hg38) |
| NORMENT-Nov-20-1146 | Global Screening Array v.3.0 | GRCh38 (hg38) |
| NORMENT-Mar-21-1273 | Global Screening Array v.3.0 | GRCh38 (hg38) |
| NORMENT-Mar-21-1409 | Global Screening Array v.3.0 | GRCh38 (hg38) |
| NORMENT-Mar-21-1413 | Global Screening Array v.3.0 | GRCh38 (hg38) |
| NORMENT-Mar-21-1531 | Global Screening Array v.3.0 | GRCh38 (hg38) |
| NORMENT-Mar-21-1532 | Global Screening Array v.3.0 | GRCh38 (hg38) |

# Quality Control (QC) steps

Load PLINK module

Set up the GITHUB environmental variable

GITHUB=/tsd/p697/data/durable/s3-api/github/norment/moba\_qc\_imputation or GITHUB=/cluster/projects/p697/github/norment/moba\_qc\_imputation.

To check the GITHUB variable is set correctly, execute echo $GITHUB command (it should print the full path to the “moba\_qc\_imputation” folder).

Unless otherwise specified all commands are supposed to be run in your working directory (named with your initials) in DATA/DURABLE. Please refer to the File naming and Folder system.docx for the naming conventions.

When you read “record” in the instructions below, it means to record the specified number in the google drive MoBa\_QC\_numbers spreadsheet (each tab is devoted to a batch).

## Initial batch files and information

* 1. Copy the set of binary PLINK files (bed, bim and fam files) of the batch assigned to you from “original\_data” folder to your working directory (the one named with your initials). Never work on the original data directly!
  2. Record the number of SNPs in your batch  
     wc -l original.bim
  3. Record the number of individuals in your batch  
     wc -l original.fam

## Update SNP alleles

* 1. For the HumanCoreExome batches only: update SNP alleles (convert the Illumina A/B alleles to the genetic A,C,T,G alleles)
  2. Using the information on the following website - <https://www.well.ox.ac.uk/~wrayner/strand/ABtoTOPstrand.html>, in the “resources” folder in DATA/DURABLE there are the following files to use to update alleles:  
     humancoreexome-12v1-1\_A.update\_alleles  
     humancoreexome-24v1-0\_A.update\_alleles  
       
     Copy the file corresponding to your batch to your working directory in DATA/DURABLE and record the total number of SNPs for which there is allele information.
  3. Run the following command to update the alleles:  
     plink --bfile original --update-alleles name.of.the.file.with.alleles.of.your.array --make-bed --out original-initials-alleles-update  
       
     “name.of.the.file.with.alleles.of.your.array” is the name of the file that contains allele information for the genotype array corresponding to your batch.
  4. Record the number of number of SNPs whose alleles were updated
  5. Remove SNPs whose alleles were not updated using the following command:  
     plink --bfile original-initials-alleles-update --exclude original-initials-alleles-update.allele.no.snp --make-bed --out original-initials-alleles
  6. Record the number of SNPs removed.
  7. Create a list of SNPs that were removed  
     ./match.pl -f original-initials-alleles.bim -g original-initials-alleles-update.bim -k 2 -l 2 -v 1 | awk ‘$7==”-” {print $2,”no-allele-update”}’ > original-initials-bad-snps.txt

## Blacklisted SNPs

When available, a list of known badly performing SNPs (black-listed SNPs) will be excluded from the data. There is such a list for HumanExome array and because it is rather similar to the content on the HumanCoreExome array, data genotypes on that array will be filtered for those badly performing SNPs. Because HARVEST is genotypes on HumanCoreExome array, we will removing black-listed SNPs from these data, using the steps below.

* 1. Copy the list of badly performing SNPs from the “resources” folder to your working directory, the file is named pchip\_blackList\_dec2015\_stripped.txt.
  2. Exclude blacklisted SNPs using the following PLINK command:  
     plink --bfile original-initials-alleles --exclude pchip\_blackList\_dec2015\_stripped.txt --make-bed --out original-initials-no-black-list
  3. Add the IDs of the removed SNPs to the original-initials-bad-snps.txt list.  
     ./match.pl -f original-initials-no-black-list.bim -g original-initials-alleles.bim -k 2 -l 2 -v 1 | awk ‘$7==”-” {print $2,”black-listed”}’ >> original-initials-bad-snps.txt

## Update SNP names to rsIDs

* 1. For the Global Screening Array batches only: update SNP names (to rsIDs).
  2. Using the information on the following website - <https://support.illumina.com/downloads/infinium-global-screening-array-v1-0-support-files.html>, in the “resources” folder in DATA/DURABLE there are the following files to use to update SNP names:  
     GSA-24v1-0\_C2\_b150\_rsids\_unique.txt  
     GSA-24v3-0\_A1\_b151\_rsids\_unique.txt  
       
     Copy the file that matches the chip version of your batch (this information is available in the table at the beginning of the document) to your working directory in DATA/DURABLE and record the total number of SNPs for which there rsID SNP information.
  3. Update the SNP names to correspond to the rsID using the below command:  
     plink --bfile original --update-name name.of.the.file.with.rsIDs.of.your.chip --make-bed --out original-initials-rsids

## Update the chromosomal positions to build GRCh37/hg19

For the OmniExpress and Global Screening Array batches: update the chromosomal positions (to match the build GRCh37/hg19) where needed (this will be needed for the NORMENT batches received from 2016 onwards and the ADHD2 batch - please see the initial table to confirm if a batch needs to be converted).

* 1. Copy the liftover executable (“software” folder in DATA/DURABLE) and file to update the chromosomal positions (name: hg38ToHg19.over.chain.gz, “software” folder in DATA/DURABLE) to your working folder.
  2. Create the liftover input file using the $GITHUB/lib/create-liftover-input.R script (<https://github.com/norment/moba_qc_imputation/tree/master/lib#create-liftover-inputr>).  
     Usage: Rscript $GITHUB/lib/create-liftover-input.R dataprefix outprefix
     + 1. For OmniExpress batches use the dataprefix: original
       2. For Global Screening Array batches use the dataprefix: original-initials-rsids
       3. Use the outprefix: original-initials-liftover\_input
  3. Run liftover:  
     ./liftOver original-initials-liftover\_input.bed hg38ToHg19.over.chain.gz original-initials-liftover\_output.bed original-initials-liftover\_unlifted.bed
  4. Create PLINK update map file:  
     awk ‘{print $4, $2}’ original-initials-liftover\_output.bed > original-initials-liftover-update-map.txt
  5. Create list of unlifted SNPs:  
     awk '{print $4}' original-initials-liftover\_unlifted.bed > original-initials-liftover\_unlifted.snps
  6. Use PLINK to update the chromosomal positions and exclude SNPs whose chromosomal positions were not able to be updated.
     1. For OmniExpress batches:  
        plink --bfile original --update-map original-initials-liftover-update-map.txt --exclude original-initials-liftover\_unlifted.snps --make-bed --out original-initials-liftover
     2. For Global Screening Array batches:  
        plink --bfile original-initials-rsids --update-map original-initials-liftover-update-map.txt --exclude original-initials-liftover\_unlifted.snps --make-bed --out original-initials-liftover
  7. Add the IDs of the removed SNPs to the original-initials-bad-snps.txt list.
     1. For OmniExpress batches:  
        ./match.pl -f original-initials-liftover.bim -g original.bim -k 2 -l 2 -v 1 | awk ‘$7==”-” {print $2,”no-build-liftover”}’ >> original-initials-bad-snps.txt
     2. For Global Screening Array batches:  
        ./match.pl -f original-initials-liftover.bim -g original-initials-rsids.bim -k 2 -l 2 -v 1 | awk ‘$7==”-” {print $2,”no-build-liftover”}’ >> original-initials-bad-snps.txt