Illumina GenomeStudio Analysis

Paris Veltsos
University of St Andrews

February 23, 2012

1 Introduction

GenomeStudio is software by Illumina used to score SNPs based on the Illumina BeadExpress platform.

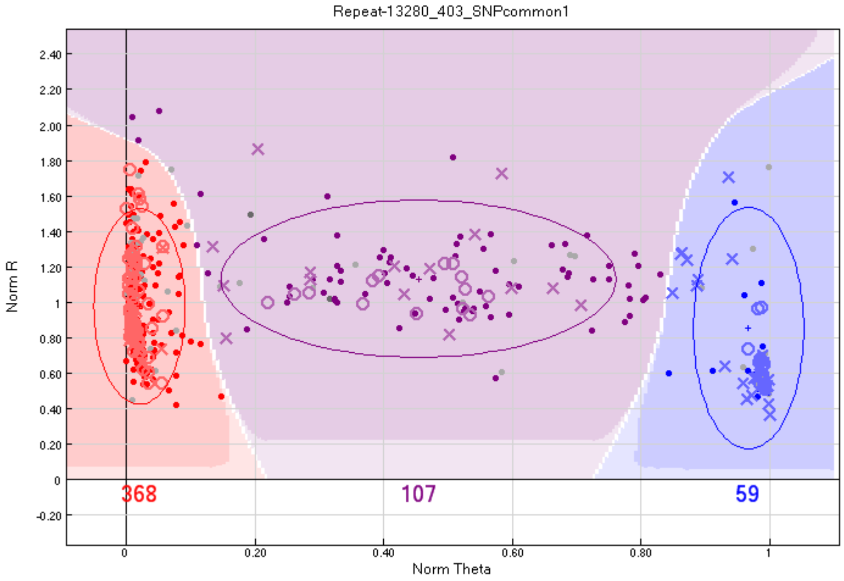


Figure 1: Example of a SNP that is typed well. There are three well-defined groups with limited overlap between them. X and O indicate calling errors based on imported pedigree relationships between the samples.

SNP calling is based on the ability to differentiate between three well-defined groups, two homozygotes and the heterozygotes. The less overlap there is between these groups, the better the SNP has been typed.

Populations should be scored separately. Different populations may have inversion differences, which will make mapping hard. If they have no common pedigrees, there is no advantage in scoring them together because no extra information becomes available. In contract, groups may be more obvious if scoring from a single populations is used.

Use all SNPs that you can, even if they are expected to be on the same genomic location. If they are, they can be removed at the mapping stage.

This is an iterative process, where each step affects previous steps and clarifies whether data look ok or not.

Remember to right click and select 'Recalculate statistics for all samples' in the samples table, after removing samples from the analysis.

2 Insert pedigree information

Pedigree information reveals scoring errors. Those errors may reflect errors in the pedigree file or the scoring, and can be used to clean up the data.

For the *D. montana* project, the names used in the pedigree file containing phenotypic information were bigger than the names on the eppendorfs containing the DNA, because the full name was too long to fit on eppendorfs. The pedigree name is a letter followed by a number and letters for each generation.

- For example K154DBA is a 'K' pedigree (Kuusamo-Finland, the other one would be 'V' for Vancouver-Canada), line 154, F3 generation (as DBA is three letters, one for each generation).

Once pedigree information is added, SNP typing incompatible with the pedigree is indicated on the SNP graph. Any samples run twice are also indicated:

- X indicates offspring with PC or PPC[1] errors based on a given pedigree.
- O indicates parents with PC or PPC errors.
- Square indicates errors in reproducibility, i.e. when the same sample is genotyped twice and the two replicates fall into different groups.

3 Replicate handling

Replicate samples give an idea of how well the typing works. Each sample must have an individual name, so the worse repeat is named to include 'repeat' in the end of its name.

3.1 Add replicate data

1. Put all files from a BeadExpress run into a folder
2. `File > Load Additional Samples > Load samples based on folder with intensities`

For the *D. montana* project there are two pedigrees and three GenomeStudio files, one for each pedigree and one containing both. The additional run with replicate data contained individuals from both replicates.

For the individual pedigree files:

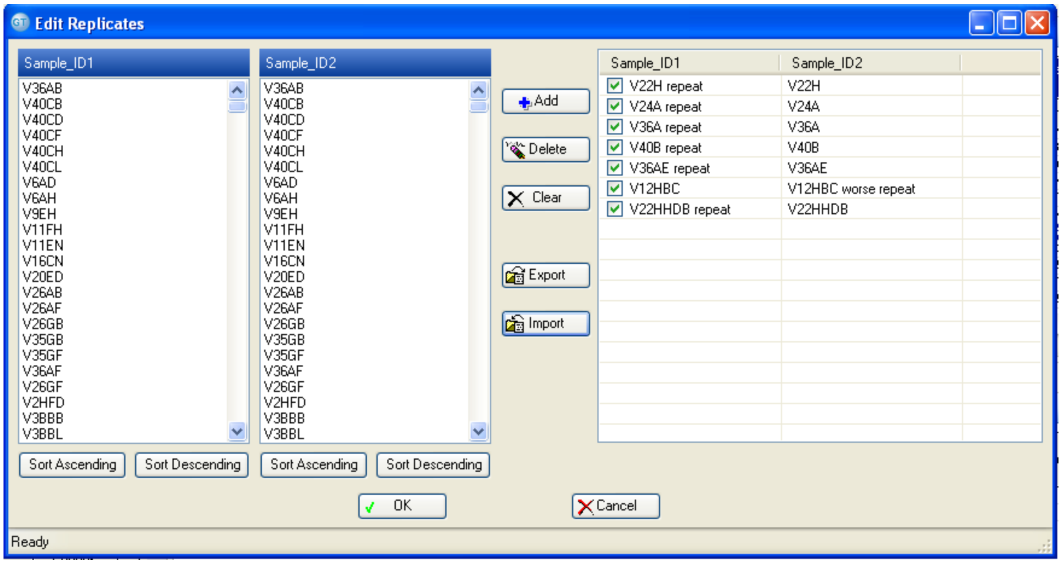
1. Mark the files from the wrong pedigree for exclusion.

[1] Parent-Child or Parent-Parent-Child

2. Save a copy of the file as.

3. Tick box to remove samples marked for exclusion.

4. Right click on each sample in the Samples Table and edit the names



Sample_ID1

V36AB
V40CB
V40CD
V40CF
V40CH
V40CL
V6AD
V6AH
V9EH
V11FH
V11EN
V16CN
V20ED
V26AB
V26AF
V26GB
V35GB
V35GF
V36AF
V26GF
V2HFD
V3BBB
V3BBL

Sample_ID2

V36AB
V40CB
V40CD
V40CF
V40CH
V40CL
V6AD
V6AH
V9EH
V11FH
V11EN
V16CN
V20ED
V26AB
V26AF
V26GB
V35GB
V35GF
V36AF
V26GF
V2HFD
V3BBB
V3BBL

Sample_ID1	Sample_ID2
☑ V22H repeat	V22H
☑ V24A repeat	V24A
☑ V36A repeat	V36A
☑ V40B repeat	V40B
☑ V36AE repeat	V36AE
☑ V12HBC	V12HBC worse repeat
☑ V22HHDB repeat	V22HHDB

Figure 2: The ‘Edit Replicates’ popup window.

Define the replicate samples. `Analysis` > `Edit Replicates`. Either import a text file where each row contains the names of the replicate samples, or define them from the ‘Edit Replicates’ popup window.

3.2 Remove worst replicate

Replicates could have been useful in defining groups to use for calling SNPs because they are expected to show the variance in calling the same sample.

Unfortunately the data were too messy in the *D. montana* project because the DNA extractions were not high quality. I decided to ignore the duplicate sample information and retained only the duplicate sample with most SNP calls:

1. Sort by Sample ID in the Samples Table and look for the name of sample with ‘repeat’ next to it.

2. Decide which run best by comparing the call frequency of these samples.

3. Add the sample with the least SNP calls into the group ‘worse repeat’. This is a group I made for the purpose.

4. Right click, choose the ‘worse repeat’ samples, right click on one of them and ‘exclude those samples from the analysis’.

5. If the originally run sample was the worst, rename it by appending ‘worse repeat’ to its name, and rename the best repeat to just the sample name, so that it matches the pedigree name.

4 Add gender information

Gender can be useful in defining groups and identifying sex linked samples.

Create a .csv file with two columns, `Sample ID` and `Gender`.

4.1 Create csv with sex information

The instructions assume there is an Excel file with the pedigree (ID, mother, father) and sex information, from which the correct columns need to be extracted.

1. Make `sex.csv` file

   To make the `sex.csv` file, copy the pedigree name along with the sex column. Copy both the pedigree individuals AND the parental generation, i.e. the generation with information on sex, but unknown parents. Paste to a new text file and convert to CSV[2]. Add `Sample ID,Gender` for header.

2. Import in GenomeStudio `File > Import Phenotype Information from File...`

Upon successful import, the Gender column of the samples should be populated. It may not be visible by default, but it is useful to be able to sort by this column. To make it visible using the interface from ‘Column Chooser’.

5 Improve calling

5.1 Exclude all individuals with call rates < 0.65 and make clusters automatically

5.2 Manually re-cluster each locus

Zero SNPs if:

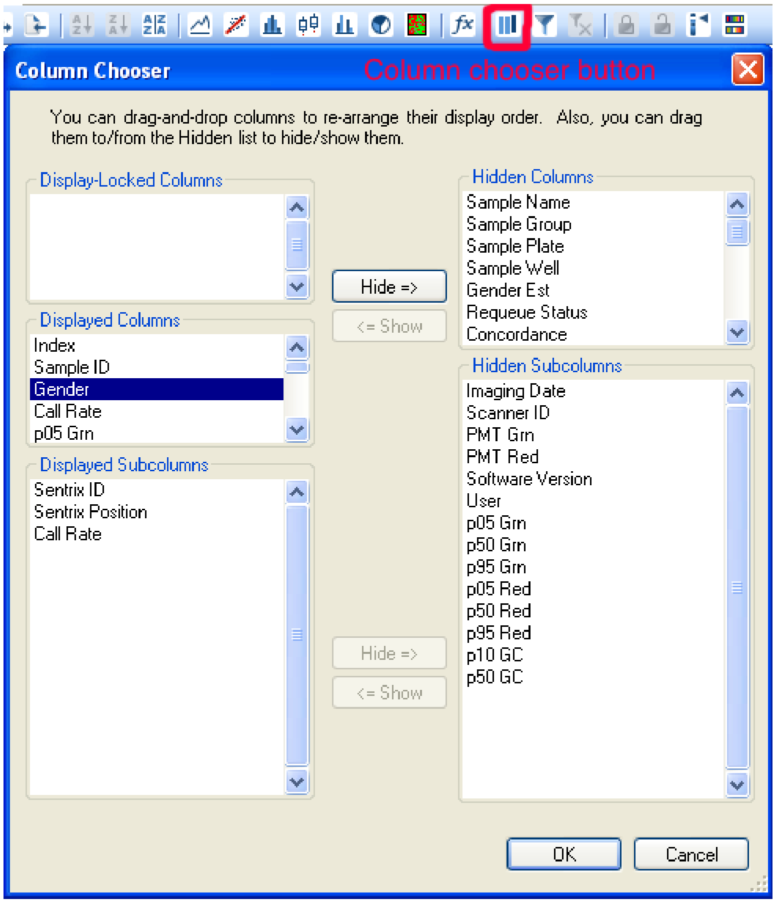
1. They look ugly
2. Have low normalised R (y axis on graph)
3. Have too low polymorphisms to be useful

X-linked genes are expected to have more pedigree errors.

The general rule is to move and resize the clusters to minimise the number of pedigree errors. However some errors may be there as a consequence of inherent bad calling in the particular study (for example due to low DNA quality). So also try to define three separate clusters.

Holding Shift, click and drag the centre of a cluster to move it and click and drag on its edges to resize it. The number of pedigree errors is visible in the bottom of the Errors Table.

[2] In the Terminal `perl -pe ’s/\t/,/g’ tabdelimitedFile.txt > sex.csv`


Column Chooser

Column chooser button

You can drag-and-drop columns to re-arrange their display order. Also, you can drag them to/from the Hidden list to hide/show them.

Display-Locked Columns

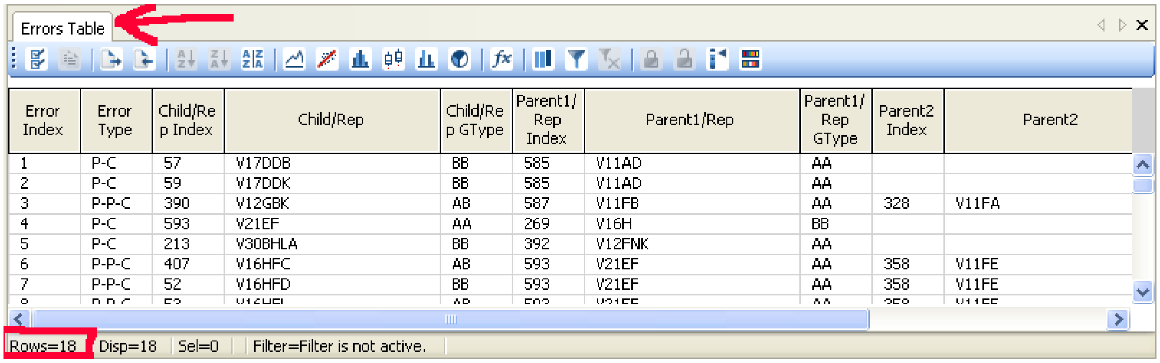
Displayed Columns
- Index
- Sample ID
- Gender
- Call Rate
- p05 Grn

Displayed Subcolumns
- Sentrix ID
- Sentrix Position
- Call Rate

Hidden Columns
- Sample Name
- Sample Group
- Sample Plate
- Sample Well
- Gender Est
- Requeue Status
- Concordance

Hidden Subcolumns
- Imaging Date
- Scanner ID
- PMT Grn
- PMT Red
- Software Version
- User
- p05 Grn
- p50 Grn
- p95 Grn
- p05 Red
- p50 Red
- p95 Red
- p10 GC
- p50 GC

Figure 3: Click the ‘Column Chooser’ button above the Samples Table. Add the Gender column in the Displayed columns box, close to the Sample ID so that sorting by either is easy.


Errors Table

Error Index	Error Type	Child/Rep Index	Child/Rep	Child/Rep GType	Parent1/ Rep Index	Parent1/Rep	Parent1/ Rep GType	Parent2 Index	Parent2
1	P-C	57	V17DDB	BB	585	V11AD	AA		
2	P-C	59	V17DDK	BB	585	V11AD	AA		
3	P-P-C	390	V12GBK	AB	587	V11FB	AA	328	V11FA
4	P-C	593	V21EF	AA	269	V16H	BB		
5	P-C	213	V30BHLA	BB	392	V12FNK	AA		
6	P-P-C	407	V16HFC	AB	593	V21EF	AA	358	V11FE
7	P-P-C	52	V16HFD	BB	593	V21EF	AA	358	V11FE

Rows=18 Disp=18 Sel=0 Filter=Filter is not active.

Figure 4: The Errors Table. If duplicate samples remain, they are also indicated and count towards the total.

5.3 Remove individuals with low call rates relative to others

Once re-clustering is done, rank by call rate. Remove any individuals with relatively low call rates when compared to other individuals.

5.4 Identify sex linkage

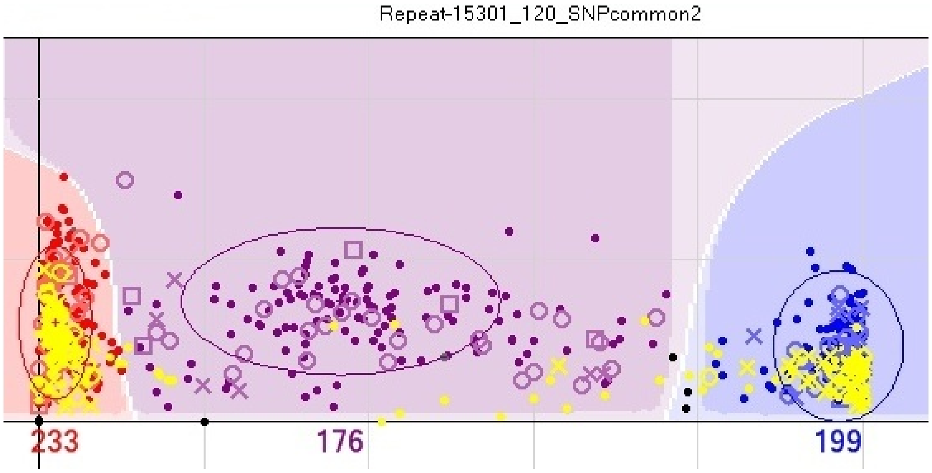


Figure 5: Typical indicator of sex linkage is for males to be under-represented amongst the heterozygotes, and have lower than average signal intensity.

To identify potential X linkage of a SNP, sort by Gender and then Het Excess. X-linked SNPs will have large and negative values for Het Excess. Visually, no male heterozygotes should exist on the graph showing the clusters.

If the DNA concentration of all samples was similar, males should have half the signal intensity than females, on X linked markers, because they contain a single copy of the SNP.

5.5 Remove mistyped males

Remove individual males that are heterozygous for X-linked loci, as judged by all other male data. If they are in HWE the loci may be autosomal after all.

5.6 Use duplicate samples to identify clusters

This approach did not work for the *D. montana* project. I kept the SNP run with the most calls amongst samples, see Remove worst replicate (section 3.2).

5.7 Minimise PPC errors

Exclude X-linked loci when estimating PPC errors.

Remove all individuals with high error rates, or check their pedigrees.

Remove loci with unusually high error rates.

5.8 Make a list with the names of X-linked markers

This is used later to modify genotyping information of the males for these markers, to help with genetic map construction.

Add the X-linked markers back to the data

5.9 Remove loci with high error rates

If any loci still have high error rates after excluding X-linked markers and pedigree errors, remove them.

6 Output

6.1 In GenomeStudio

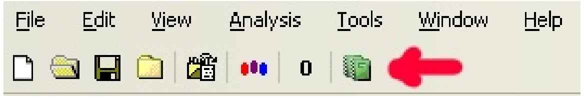


Figure 6: The report wizard button

1. Click on 'Report Wizard. . . '
2. Choose `Final Report` > `All Samples` > `50th percentile GC score ?` > `Remove excluded samples from report` > `Remove zeroed SNPs from report` > `Matrix - do not Include GenCall score` > `Finish`.

6.2 In the Terminal/Text editor

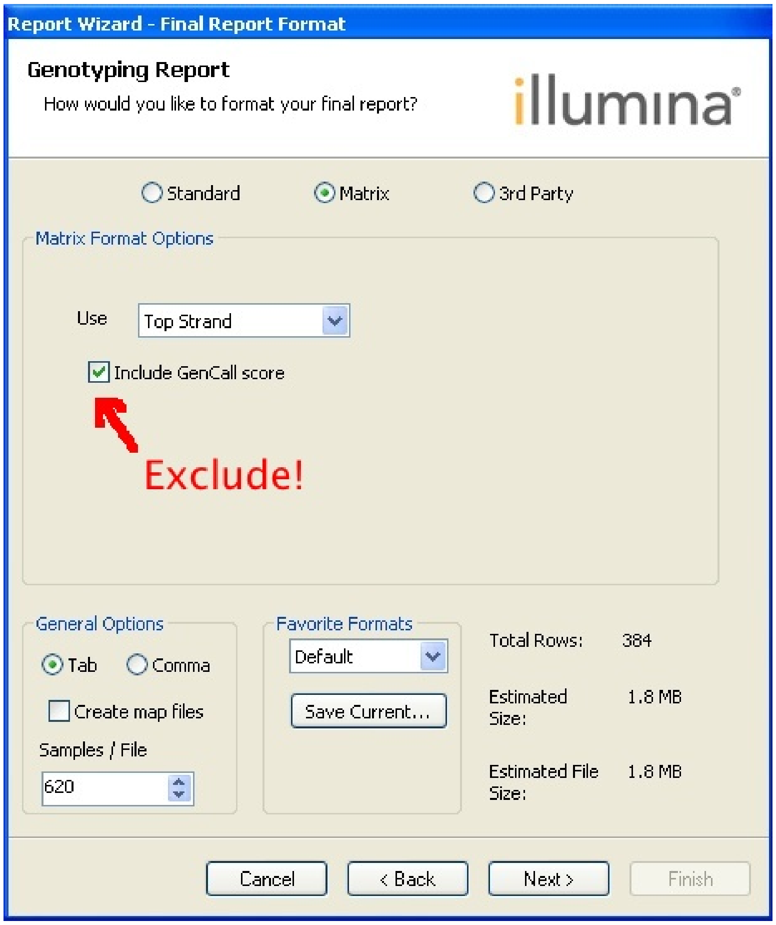
Make a SNP scoring file for R:

- Delete the first SNP report lines in a text editor.
- Transpose the file[3] `perl transposeTabDelimited.pl inputFile.txt > SNPData.txt`.
- Add 'ID' in the very beginning before the tab of the first line, to make it compatible for R import.

Make a `pedData.txt` file. This should be tab-delimited, contain the same name of individuals as the `SNPData.txt` and contain the headers:

```
ANIMAL_ID   SIRE_ID DAM_ID
```

[3] Use the perl script transposeTabDelimited.pl[4]. It is possible to transpose in Excel too (select all) `copy > paste special > (tick) transpose`



Report Wizard - Final Report Format

Genotyping Report

How would you like to format your final report?

○ Standard ● Matrix ○ 3rd Party

Matrix Format Options

Use Top Strand

☑ Include GenCall score

Exclude!

General Options

● Tab ○ Comma

□ Create map files

Samples / File 620

Favorite Formats

Default

Total Rows: 384

Estimated Size: 1.8 MB

Estimated File Size: 1.8 MB

Figure 7: Final report 3rd window. Choose ‘Matrix’.

6.3 In R

The following commands read the files and combine them based on the column names ‘ANIMAL_ID’ and ‘ID’, and export the merged information.

```
KData<-read.table("/Users/zabameos/Dropbox/GTD/Inbox/kdataManip/Ktransposed.txt", header=T)
VData<-read.table("/Users/zabameos/Dropbox/GTD/Inbox/kdataManip/Vtransposed.txt", header=T)
pedData<-read.table("/Users/zabameos/Dropbox/GTD/Inbox/kdataManip/pedData.txt", header=T)
sexData<-read.table("/Users/zabameos/Dropbox/GTD/Inbox/kdataManip/aPedSex.txt", header=T)
pedSexData=merge(pedData, sexData, by.x="ANIMAL_ID", by.y="Sample_ID", all=F )
mergedKData=merge(pedSexData, KData, by.x="ANIMAL_ID", by.y="ID", all=F )
mergedVData=merge(pedSexData, VData, by.x="ANIMAL_ID", by.y="ID", all=F )
write.table(mergedVData, file="VMerged.txt", quote=F, row.names=F, sep="\t")
write.table(mergedKData, file="KMerged.txt", quote=F, row.names=F, sep="\t")
```

6.4 Back in the Terminal/Text editor

Generate a tab between each allele of a SNP and convert missing data to 0s:

```
perl -pe ’s/\-\-/0\t0/g ; s/\t(\D)(\D)/\t$1\t$2/g’ pedigreeMerged.txt > out.txt
```

In a text editor:

- Delete the first rows up to the sample names, the names of which have been scrambled.
- Copy the sample names from the `pedigreeMerged.txt` file.
- For the first row (just copied) replace tabs with double tabs for the columns with marker names, and paste on the top of the file.

Check all is well by opening in Excel. The input crigen expects is of the form:

Table 1: Example of crigen input.

ID	Sire	Dam	Marker1		Marker2		Marker3		Marker4		Marker5	
1	0	0	4	4	3	4	2	1	3	1	2	1
2	0	0	1	4	4	4	1	2	2	3	6	1
3	0	0	1	1	3	4	1	1	1	1	4	2

6.5 Change typing of males to make the X-chromosome mappable

Males should all be homozygotes for X markers, which is obvious for some markers in Genomestudio.

These males should be manually made heterozygote with another allele. However it must be one of the DNA letters, e.g. `T`. Use anything other than what the males already are. The Pedigree file should also contain sex information.

Also take the opportunity to zero any heterozygote males.

6.6 Format the pedigree for crimap

Crimap does not accept letters on the pedigree names. Originally I used a perl script, however it did not result in unique names. I then manually changed the names by find and replace for all individuals in Excel to make sure the names are unique.

This also allowed to visualise how many of the parents are not scored, because I renamed sequentially based on the offspring row, so if some offspring were not there but were parents of other individuals, they would have kept their original names in the parent columns. I then sequentially renamed the parent columns.

It is useful to get a feel of the data by doing it manually, but it is much faster to do it in R.

Rename factors to numbers in R

First collect all columns from the pedigree files into a single column and sort alphabetically in Excel. Output as text as `pedigreeNames.txt`, no header is needed. Also copy the first three columns of the excel file into a `correctVPed.txt` file. Import both in R.

```
pedData<-read.table("/Users/zabameos/Desktop/temp/correctVPed.txt", header=T)
pedigreeNames<-read.table("/Users/zabameos/Desktop/temp/pedigreeNames.txt", header=F)
```

The following code renames the individual names to numbers and applies the naming to the pedigree file.

```
oldNames <- unique(pedigreeNames$V1)
newNames <- seq(1:length(oldNames))
nameData <- data.frame(oldNames, newNames)

newID = merge(pedData, nameData, by.x="ANIMAL_ID", by.y="oldNames", all=F )
names(newID)[names(newID)=="newNames"] = "newID"
newSireID = merge(newID, nameData, by.x="SIRE_ID", by.y="oldNames", all=F )
names(newSireID)[names(newSireID)=="newNames"] = "newSireID"
newDamID = merge(newSireID, nameData, by.x="DAM_ID", by.y="oldNames", all=F )
names(newDamID)[names(newDamID)=="newNames"] = "newDamID"

allPedigreeNames <- data.frame(newDamID)
write.table(allPedigreeNames, file="allPedigreeNames.txt", quote=F, row.names=F, sep="\t")
```

The output file contains the original input pedigree and the new names.

1. Replace the number that replaced `0` in the original pedigree with `0`.
2. Sort it by ID in Excel and make sure the Excel file is sorted by ID. Copy and paste the new names onto the old ones in Excel.
3. Copy all data to a text file[5] and save with a `.gen` extension.

Congratulations! You have a file that can be used by Crigen.

7 Unexplained errors - Thoughts

7.1 Program bugs

It is not possible to unmark a sequence, so be careful of what markings you use.

Once some sequences are removed from the analysis, they need to be manually processed again. I encountered this when excluding the X linked markers for identifying real pedigree errors: they have to be clustered again manually.

[5] Don't save from Excel, because the `.gen` file needs to have Linux line format line endings.

7.2 General

Pedigree names were complicated to keep track of, and resulted in duplicate shortened names on the eppendorfs which were tricky to track down and had to be excluded. A better approach would have been to just name the individuals sequentially.

7.3 Male-biased failure

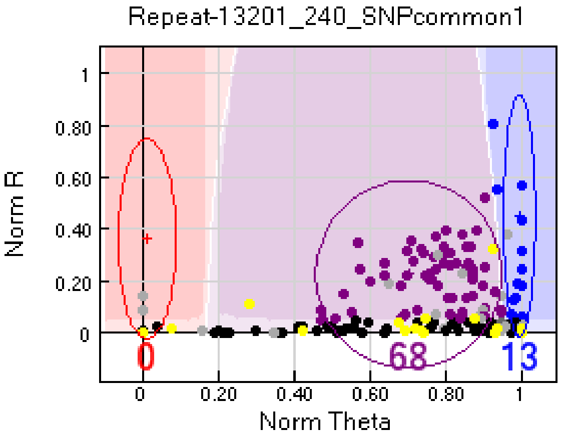


Figure 8: Example where male samples failed to amplify, while some female samples worked.

In some cases, mostly in the Vancouver pedigree (40 occasions compared to 8 in Oulanka) the SNPs of males would fail, while some female SNPs would work.

One possible explanation is that these are sex-linked markers and females have more targets that can amplify, and are therefore biased to work. In any case most of the samples did not work, so the SNPs were considered unscorable.

Action taken - SNPs considered unscorable, as some females would also fail. For example in the figure only 68 + 13 samples worked, so most samples failed.

7.4 Above-average pedigree errors

Both Vancouver and Oulanka had some pedigree errors even after carefully checking the pedigrees for duplicates, wrong names etc. The number of errors was similar for most SNPs and can be traced to its source. However some SNPs showed an uncharacteristically high number of pedigree errors. These may sometimes be removed by changing the group assignment manually.

There was usually no alternative group calling that reduced the number of errors. One possible source is multiple targets to the same SNP, for example if the SNPs are on transposable elements.

One possible explanation is that sequences with these SNPs are repetitive DNA, so there were multiple targets for them in the genome.

Action taken - Removed from the analysis.

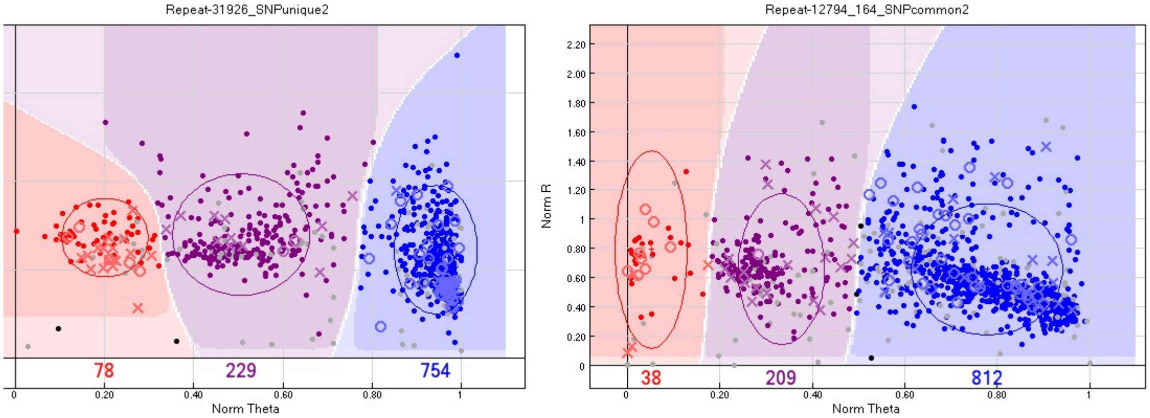


Figure 9: Examples where there were more pedigree errors than an average SNP for the particular pedigree. Moving the groups did not result in fewer errors in these cases.

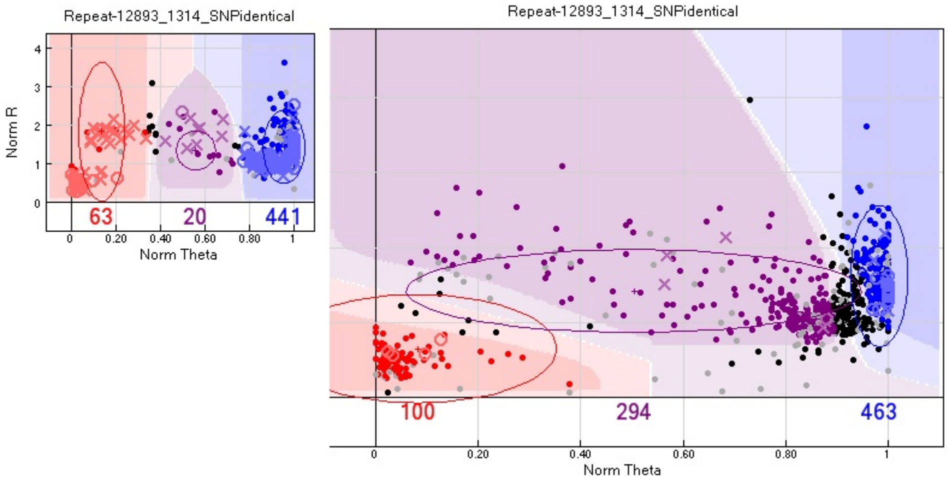


Figure 10: Example where pedigree errors have resulted from bad group calling.

7.5 Heterozygote-biased pedigree errors

For some samples most errors seemed to be on female heterozygotes (if there were male heterozygote errors the SNP might be sex-linked).

One possibility is that those samples are sex-linked, but were not called as such because there are data on only one homozygote. However in many cases they included females. The samples are particularly dodgy if one of the homozygotes is completely missing.

Action taken - Removed from the analysis.

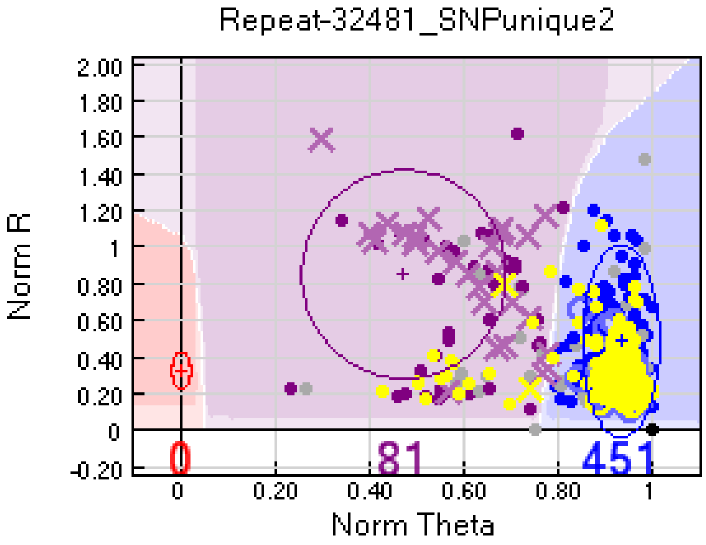


Figure 11: Example where most pedigree errors occur amongst heterozygotes.

7.6 X-linked marker deficit

In *D. montana* the X is the largest chromosome so about 25% of the markers were expected to be on it. Much fewer markers are X-linked. This may be because SNPs were called from both sexes, so there would be 75% coverage on the X compared to the autosomes, which may have biased against X-markers due to low reads achieved.

7.7 Be conservative

The golden rule of mapping is that 'No data is better than bad data'.

Pedigree errors show only some cases of bad SNP calls. Assuming that there are no pedigree errors[6], because the pedigree file has been checked specifically for them, the errors seen are due to wrong assignment of SNPs to groups. In general for the first attempts at mapping it is better to make the groups smaller and exclude individuals further away from the region showing pedigree errors. This ensures that the individuals with wrong genotypes are minimised.

7.8 Keep only pedigree individuals kept in scoring

Some individuals failed to score, even though there is pedigree information in them. It is a good idea to manually check which these individuals are to detect any mistakes.

In Excel, copy the original pedigree below the names of the scored samples. Make the original pedigree bold and sort by individual ID.

[6] Real pedigree errors can be picked up during mapping.

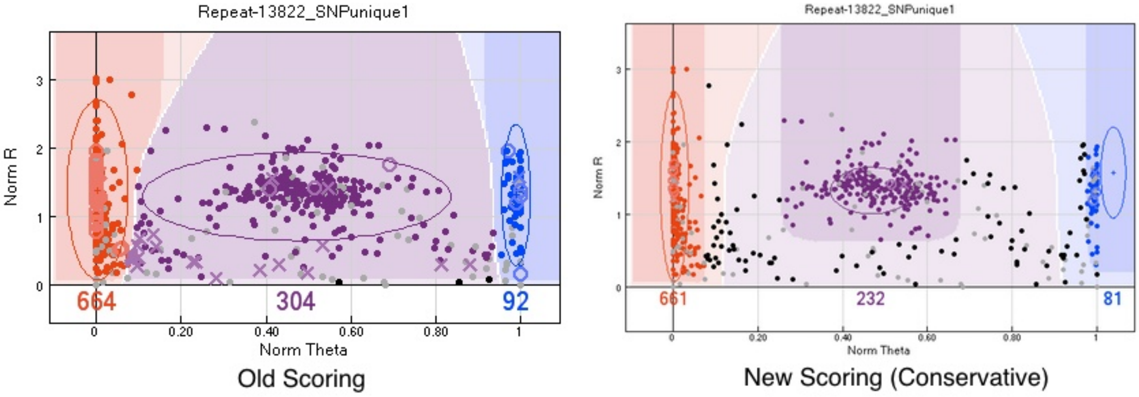


Figure 12: Example of conservative scoring. The old scoring tries to minimise pedigree errors while scoring as many individuals as possible, while the conservative scoring is trying to minimise pedigree errors.

Cut and pasted any parents of rows that existed in the scored pedigree (they are one row below or above). This is a manual step but important to identify any leftover mistakes in the pedigree structure, for example the same individual ID having multiple parent pairs.

This needs to be done once. The final output is the `pedData.txt`, which can be used to automatically add pedigree information to other files, such as SNP scoring from GenomeStudio or individual phenotypic information from other text files.