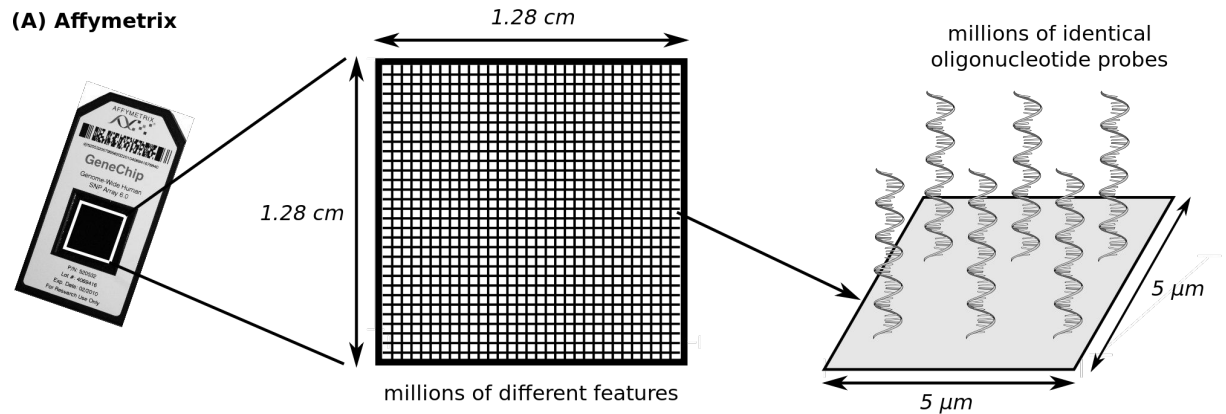


GWAS workshop

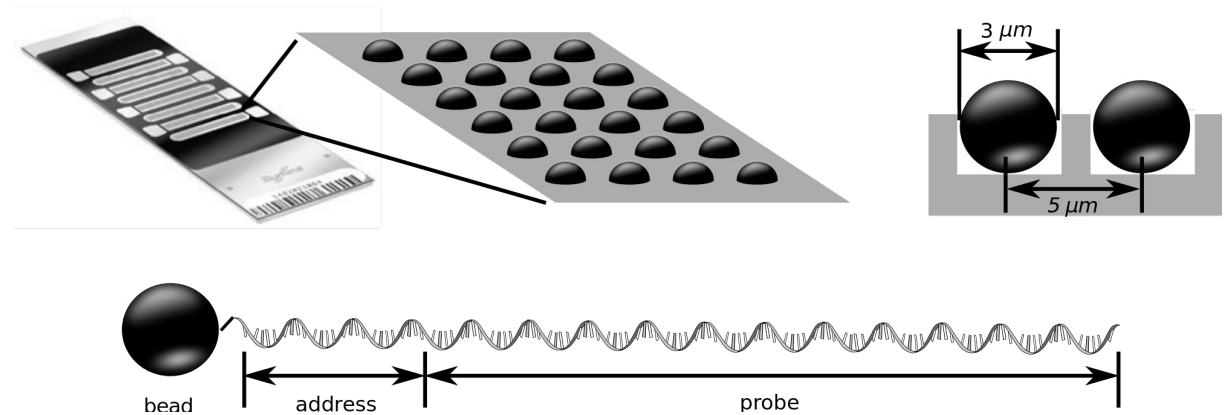
Processing array data with GenomeStudio

Intro - Genotyping Microarrays

(A) Affymetrix

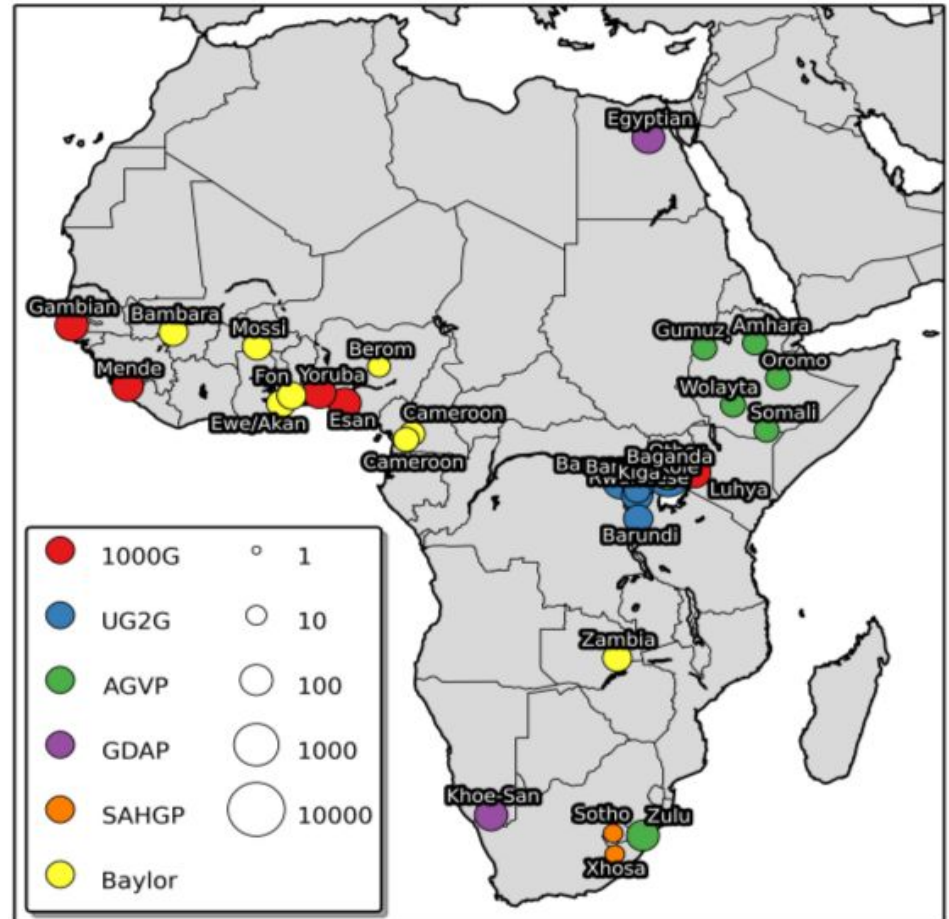


(B) Illumina



Intro - H3A chip

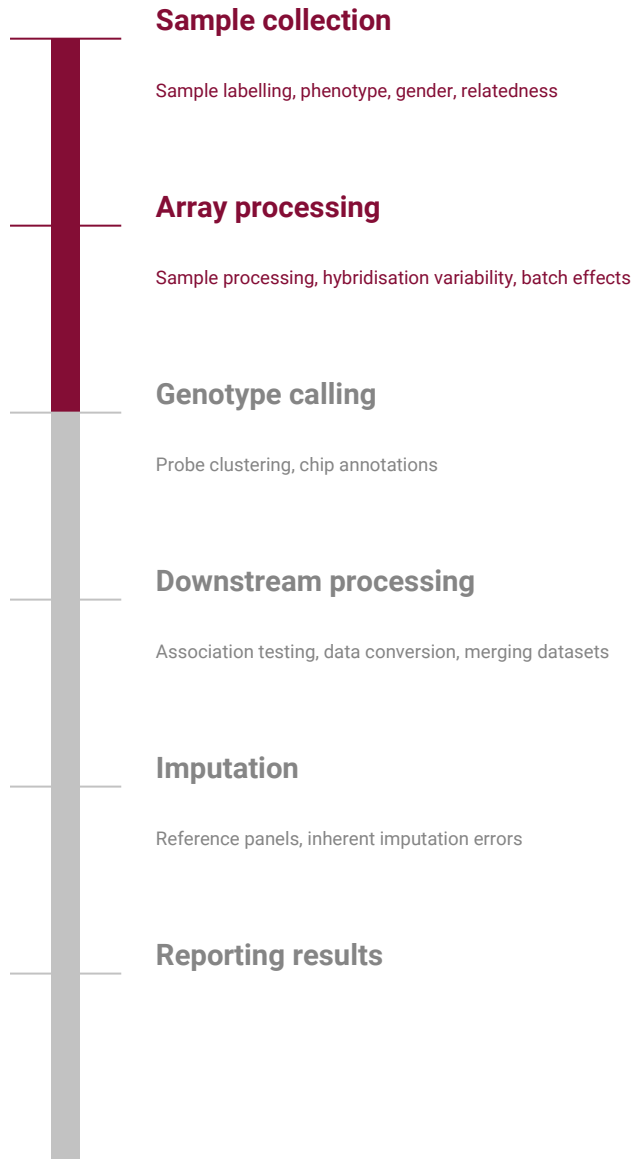
Source	Number of samples
GDAP	204
UG2G	2000
AGVP	320
1000G*	~2500
TrypanoGEN	168
Baylor	350
SAHGP	16
Total	~5500



Intro - H3A chip

- Existing catalog (~1.7M probes)
- Preselected content
 - H3Africa requested SNPs
 - ClinVar disease relevant SNPs
 - GWAS catalogue
 - AIMS
 - Y and MT
- Sequence content
 - from ~5000 sequence samples including ~350 H3Africa samples

Intro - Workflow



Concepts - SNP, alleles, genotypes, probes

Probe orientation

ATAACAGGGAAAAAGACAATTAACAAAGAACTGTAATCGTCCATAAGCCTAATTAGCCA [T/C] AAAGCAAAGTCTGTTGGGAATTTAATGAACATTTGTTAAAAACACAAATATTATAGGAAG
TATTGTCCCTTTTCTGTTAATTGTTTCTTTGACATTAGCAGGTATTCGGATTAATCGGT [A/G] TTTCGTTTCAGACAACCCTTAAATTACTTGTAACAATTTTGTGTTTATAATATCCTTC

Intro - Why GenomeStudio?

- Easy user experience
- Datasets are often distributed with GenomeStudio structure
- Several useful features
- Support from Illumina
- Good way to start

File formats

iDat

Cluster files

Plink

Annotation files

Concepts - Annotations

Illumina		Inc.					
[Heading]							
Descriptor File Name		H3Africa_2017_20021...					
Assay Format		Infinium LCG					
Date Manufactured		1/10/2018					
Loci Count		2267346					
[Assay]							
IlmnID	Name	IlmnStrand	SNP	AddressA_ID	AlleleA_ProbeSeq	AddressB_ID	AlleleB_ProbeSeq
200610-1-0 B R 1867864656	200610-1	BOT	[T/C]	0009803928	TTTTTAATGCAGGTTTGGT...		
200610-10-0 B R 1867864658	200610-10	BOT	[T/C]	0021792978	TCTACGTCTATTCTCTACTG...		
200610-100-0 B R 1867864660	200610-100	BOT	[T/C]	0069608915	TCTGTCCCAATGTATGGGA...		
200610-102-0 T R 1867864662	200610-102	TOP	[A/G]	0040640294	TGAACGTAGGTGCGATAAA...		
200610-103-0 B F 1867864663	200610-103	BOT	[T/C]	0096607568	TCATCCTATTATTATCGC...		
200610-105-0 T R 1867864667	200610-105	TOP	[A/G]	0086686956	TGTTATGATGTCTGTGTGG...		
200610-106-0 T R 1867864669	200610-106	TOP	[A/G]	0079751391	GGAAATTTTTTGTATGAT...		
200610-107-0 B F 1867864670	200610-107	BOT	[T/C]	0022767164	ACCCCCAACTAACACATT...		
200610-108-0 T R 1867864673	200610-108	TOP	[A/G]	0033808890	AGCTGTGGCTCGTAGTGTT...		
200610-109-0 T R 1867864675	200610-109	TOP	[A/G]	0069779106	CTTTGAGTTTTAAGCTGTG...		
2010-08-MT-143-0 T F 1867864939	2010-08-MT-143	TOP	[A/G]	0096801977	TCCATAATATTCATCCCTG...		
2010-08-MT-158-0 T R 1867864942	2010-08-MT-158	TOP	[A/G]	0092645470	GTTGTATAGGACTGCTTGA...		
2010-08-MT-226-0 T F 1867864947	2010-08-MT-226	TOP	[A/G]	0007668561	ATAATTAACCTTTTACTTCC...		
2010-08-MT-232-0 B F 1867864949	2010-08-MT-232	BOT	[T/C]	0087788852	TACTCCTAATCACATAACC...		
2010-08-MT-251-0 T F 1867864951	2010-08-MT-251	TOP	[A/G]	0015776530	CCAATCCTACCTCCATCGC...		
2010-08-MT-27-0 T F 1867864953	2010-08-MT-27	TOP	[A/G]	0040809554	GTAAAAAATCCAGTTGAC...		
2010-08-MT-288-0 B F 1867864957	2010-08-MT-288	BOT	[T/C]	0058794242	AACATCTCCGCATGATGAA...		
2010-08-MT-395-0 B F 1867864963	2010-08-MT-395	BOT	[T/C]	0031677820	CTCTGTTCTTTTCATGGGGA...		
2010-08-MT-398-0 B F 1867864965	2010-08-MT-398	BOT	[T/C]	0002685597	GAAGCAGATTTGGGTACCA...		
2010-08-MT-482-0 T R 1867864992	2010-08-MT-482	TOP	[A/G]	0075624251	ACTGTAATGTGCTATGTAC...		
h3a-req-37_13_113742667_G_A_ver1...	h3a-req-37_13_11374...	BOT	[T/C]	0010720249	CTACAAGCCTGCAGCTGGC...		
h3a-req-37_13_113742667_G_A_ver2...	h3a-req-37_13_11374...	BOT	[T/C]	0027668141	CTACAAGCCTGCAGCTGGC...		
h3a-req-37_13_37431791_C_A_ver1...	h3a-req-37_13_37431...	TOP	[A/C]	0082632541	CTTCTTTTGCCTTGAATGA...		
h3a-req-37_13_37431791_C_A_ver2...	h3a-req-37_13_37431...	TOP	[A/C]	0004807537	CTTCTTTTGCCTTGAATGA...		
h3a-req-37_13_37447910_G_A_ver1...	h3a-req-37_13_37447...	BOT	[T/C]	0083767327	TGGTATGATCTCGGCTCAC...		
h3a-req-37_13_37461874_A_T_ver1...	h3a-req-37_13_37461...	BOT	[T/A]	0093757865	CTTCCCTTCTGTTGTCCCA...	0090695972	CTTCCCTTCTGTTGTCCCA...
h3a-req-37_13_37486207_T_C_ver1...	h3a-req-37_13_37486...	TOP	[A/G]	0076776939	AAATTCGAAGTGGTTTGCAT...		
h3a-req-37_13_37486207_T_C_ver2...	h3a-req-37_13_37486...	TOP	[A/G]	0002800215	AAATTCGAAGTGGTTTGCAT...		
h3a-req-37_13_37489493_G_A_ver1...	h3a-req-37_13_37489...	TOP	[A/G]	0007636844	AATGGGCATCCCTGGCACT...		
h3a-req-37_13_37489493_G_A_ver2...	h3a-req-37_13_37489...	TOP	[A/G]	0046647155	AATGGGCATCCCTGGCACT...		
h3a-req-kgp1258864_ver2-1_B_F_25...	h3a-req-kgp1258864_...	BOT	[T/C]	0071775295	TGTGCCAGATACAAATGTA...		
h3a-req-kgp7369827_ver1-1_T_F_25...	h3a-req-kgp7369827_...	TOP	[A/G]	0087765124	ACCTGATGTAAGCCTGCAT...		
h3a-req-kgp7369827_ver2-1_T_F_25...	h3a-req-kgp7369827_...	TOP	[A/G]	0035750249	ACCTGATGTAAGCCTGCAT...		
kgp10000048-0 B R 1811117133	kgp10000048	BOT	[T/C]	0029768312	TCCCTGTGGCTTCGTAGCA...		
kgp100001-0 B R 1817044898	kgp100001	BOT	[T/C]	0068694355	CAATTGGAAAGAAGTGGGT...		
kgp10000114-0 B R 1802430487	kgp10000114	BOT	[T/C]	0093648476	GCCCCAACTAGACTGTGAG...		
kgp10000117-0 T F 1816868468	kgp10000117	TOP	[A/G]	0013803351	TGCTTTTTCTCTCGGAAGA...		
kgp10000118-0 B R 1814768576	kgp10000118	BOT	[T/A]	0033774380	ACAGTAGTTACACCAAGGT...	0023714370	ACAGTAGTTACACCAAGGT...
kgp10000146-0 B F 1800806438	kgp10000146	BOT	[T/C]	0035740320	CATCCCTTCAAATGAAAAG...		
kgp10000161-0 B R 1809453492	kgp10000161	BOT	[T/C]	0035723312	TCTGTATGTGTTGAGTGTG...		

Concepts - ID types

200610-1-0_B_R_1867864656

exm-rs10032423-131_B_F_1990477642

GA000172-0_T_F_1853021140

hCV32294341-0_T_F_1853033970

kgp10000048-0_B_R_1811117133

MitoA10045G-0_B_R_1850911267

pgxUn0004-0_T_F_1853105790

rs1000000-131_B_F_1894802339

seq-h3a_37_1_100726313_G_A-1_T_F_2510603673

seq-rs10498672-147_T_F_2585303702

snp-known10000012-142_T_F_2556548951

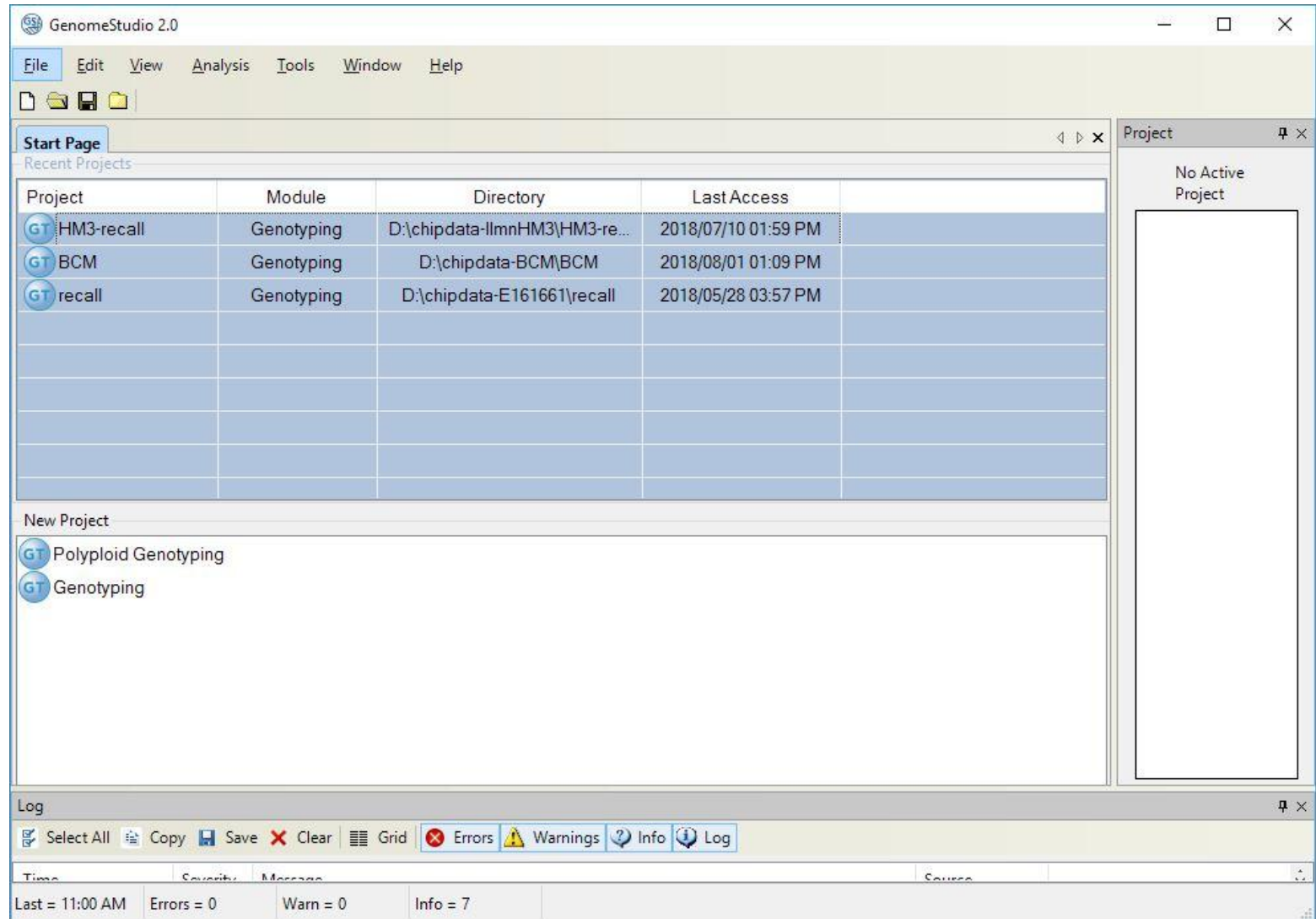
SNP106-0_B_R_1852975650

VG01S1133-0_B_R_1852963053

h3a-req-37_1_16534074_G_T_ver1-1_T_R_2598050152

h3a-req-37_1_16534074_G_T_ver2-1_T_R_2598067940

GenomeStudio - start project



GenomeStudio - start project

GenomeStudio Project Wizard - Project Location

Genotyping Project
Please specify the name and location for your project



Projects Repository
D:\subset

Project Name
☒ Create
☐ Select from LIMS

Project will be created in: D:\subset\HM3subset

GenomeStudio - start project

GenomeStudio Project Wizard - Loading Sample Intensities

Genotyping Project

Please specify the samples you want to load by identifying the sample sheet and associated data and manifest repositories



☐ Use sample sheet to load sample intensities

☒ Load sample intensities by selecting directories with intensity files

Cancel < Back Next > Finish

GenomeStudio - start project

GenomeStudio Project Wizard - Loading Sample Intensities

Genotyping Project

Please specify the samples you want to load by identifying the SNP manifest you want to use and then selecting directories that contain intensity files from v...

SNP Manifest

D:\chipdata-manifests\H3Africa_2017_20021485_A3.bpm

Data Repository

D:\subset\IDATs

Directories in Repository:

- + 201958790010
- + 201958790011
- + 201958790012

Selected Directories

GenomeStudio - start project

GenomeStudio Project Wizard - Cluster Positions

Genotyping Project

If you have an existing cluster file that you want to import cluster positions from, enter it here. Otherwise, you can cluster the samples you've selected to determine...



☐ Import cluster positions from a cluster file

Cluster File

D:\chipdata-BCM\H3Africa_Cluster Files\H3Africa192.egt

Project Settings

Options

☐ Pre-Calculate

Pre-Calculate should only be used for memory based projects.

This option will improve speed but requires 4.5x more memory.

Project Creation Actions

☐ Cluster SNPs

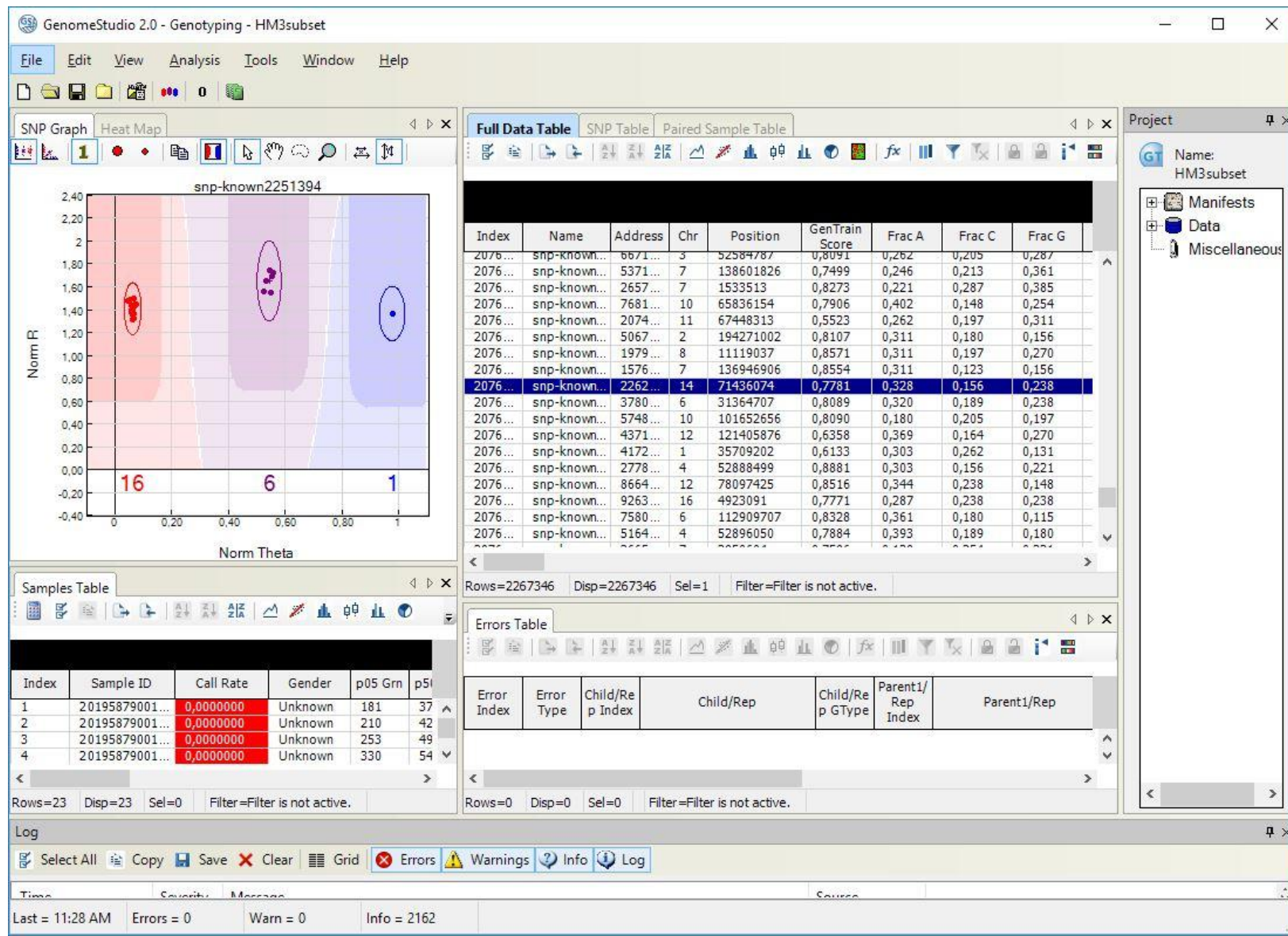
☐ Calculate Sample and SNP Statistics

☐ Calculate Heritability

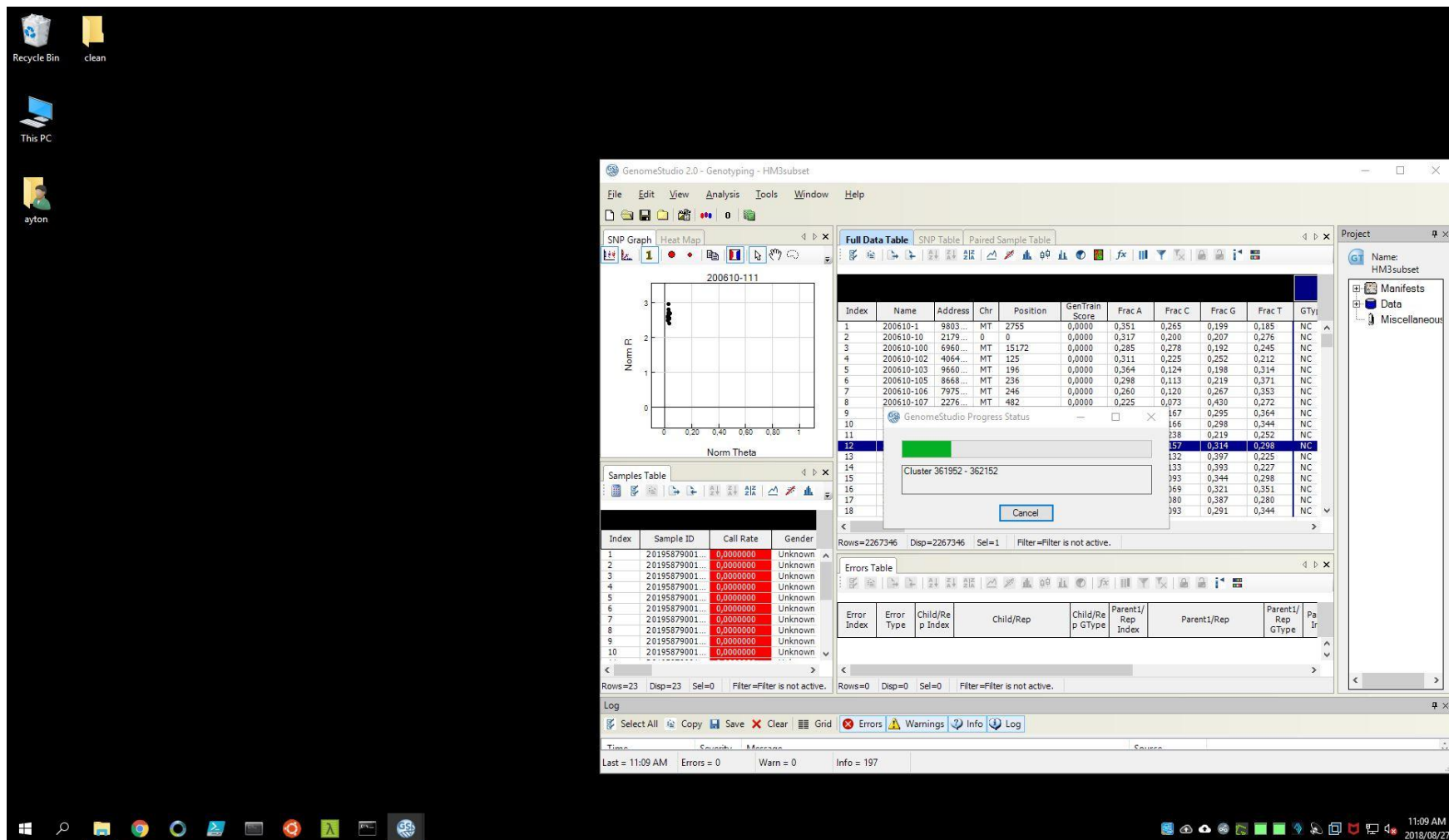
Gen Call Threshold

Optional Script File

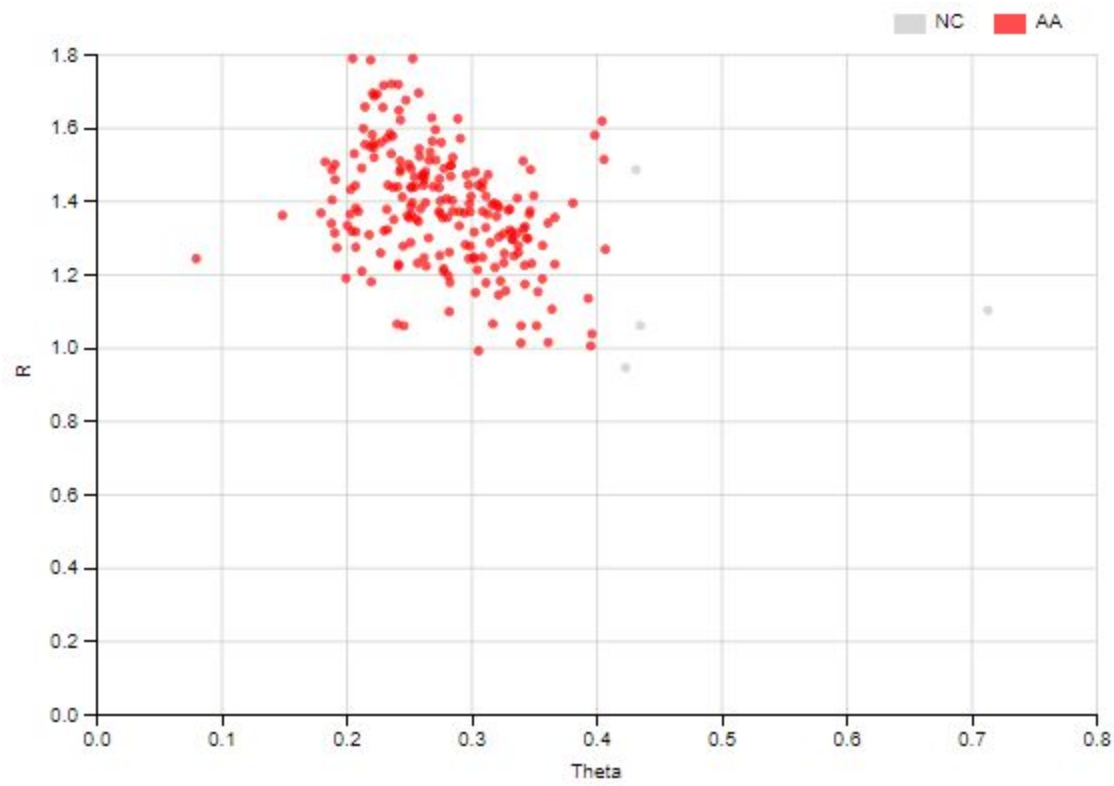
GenomeStudio - overview



GenomeStudio - clustering

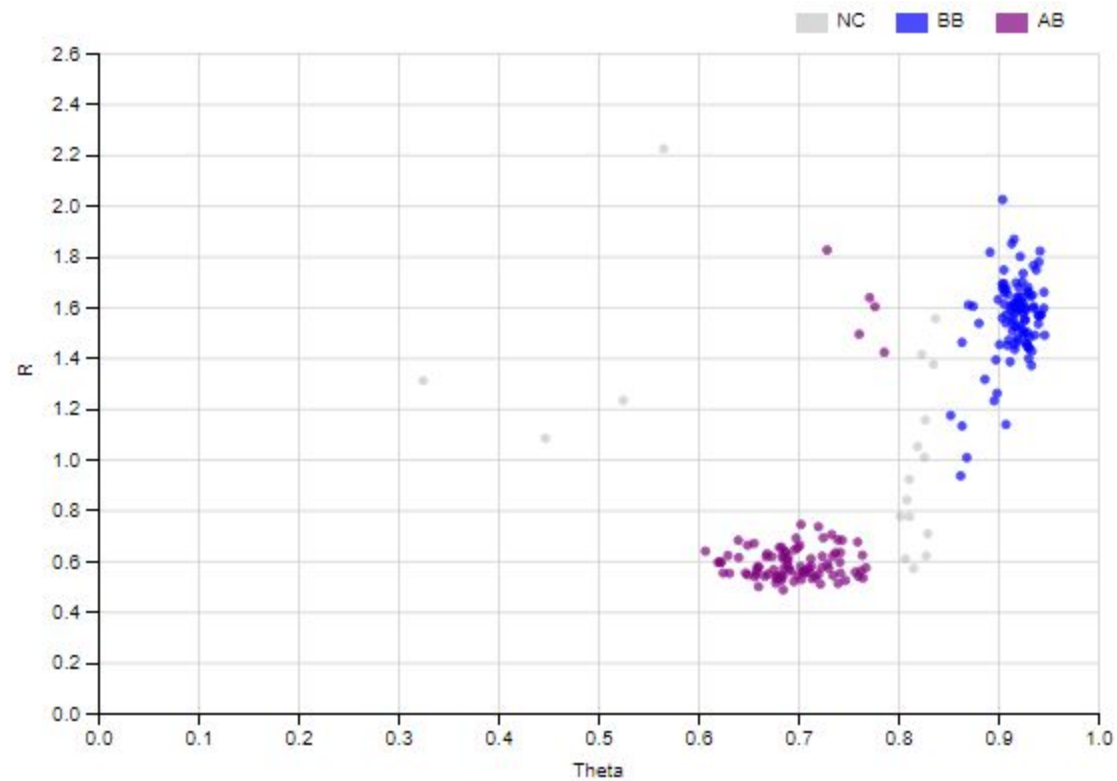


Troubleshooting - Non-polymorphic



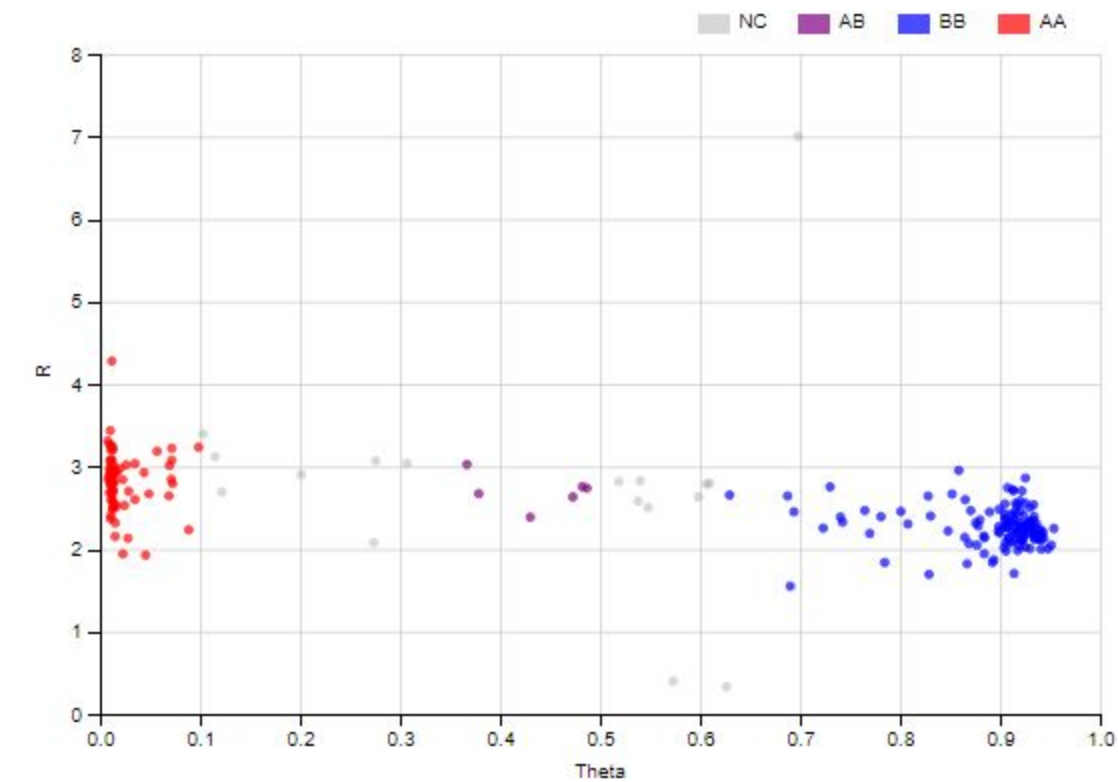
IlmnID	snp-known35339498-142_B_R_2367222456
Name	snp-known35339498
NUID	1:1150211

Troubleshooting - incorrect ploidy Y



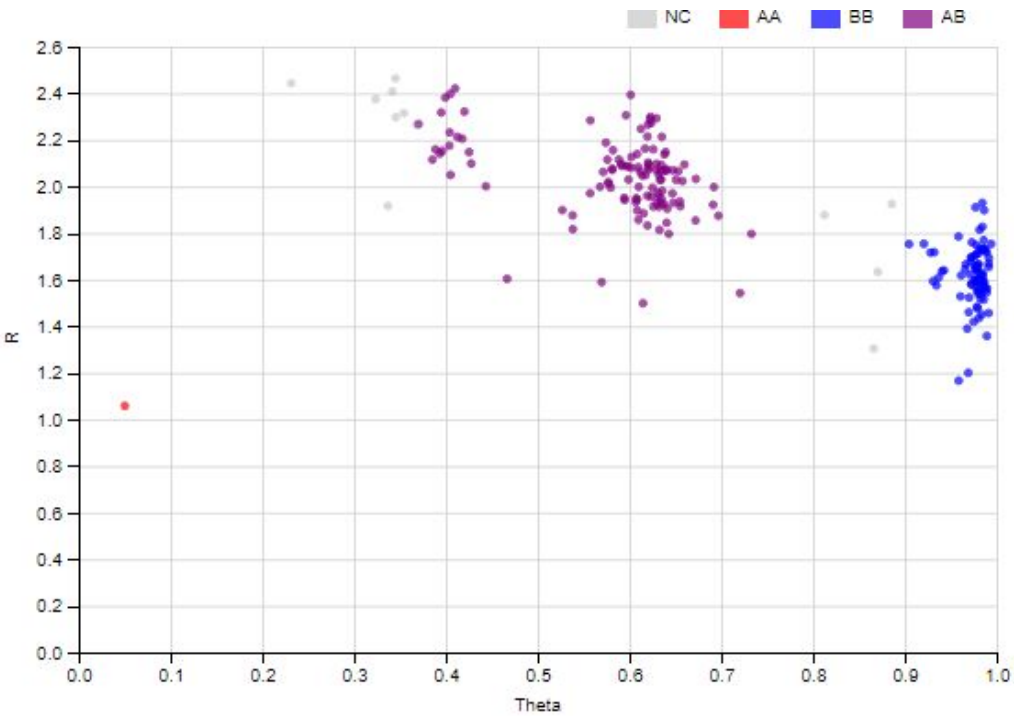
IlmnID	2010-08-Y-1952-0_T_F_1867866099
Name	2010-08-Y-1952
NUID	0:0

Troubleshooting - Incorrect ploidy MT

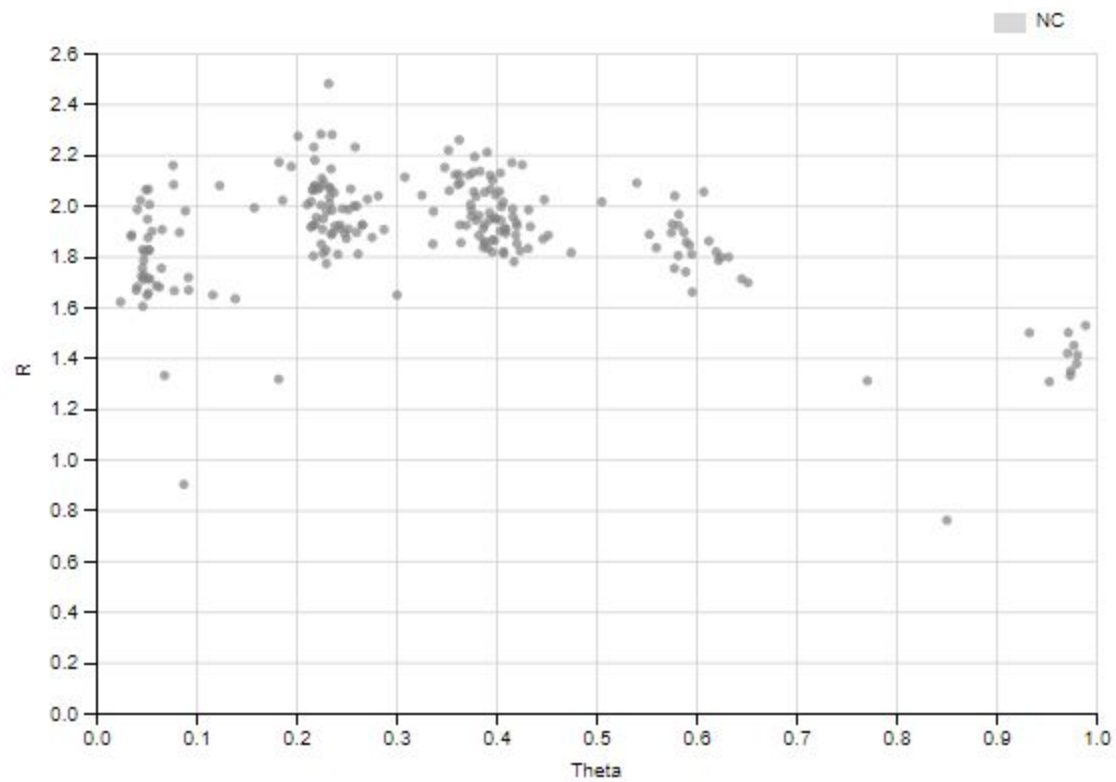


IlmnID	MitoG247A-1.32735e+007_B_R_1501342270
Name	MitoG247A
NUID	MT:247

Troubleshooting - False clustering



Troubleshooting - multiple mappings



IlmnID

h3a_37_19_22704447_T_C-6_T_R_2594837377

Name

h3a_37_19_22704447_T_C

NUID

19:22704447

GenomeStudio - Exporting

The screenshot displays the GenomeStudio 2.0 interface for a project named 'HM3subset'. The 'Analysis' menu is open, showing options like 'Exclude Samples By Best Run', 'Cluster All SNPs', 'Update SNP Statistics...', 'Edit Replicates', 'Edit Parental Relationships', 'Update Heritability/Reproducibility Errors...', 'Reports', 'View Controls Dashboard', 'Paired Sample Editor...', 'Calculate Paired Sample LOH/CN...', 'Import Allele Calls...', 'Export Allele Calls...', 'Remove Imported Allele Calls', 'Create Plugin Column...', 'CNV Analysis...', and 'Show CNV Region Display'. The 'Reports' submenu is also open, showing 'Create Reproducibility and Heritability Report', 'Compute CN Metrics...', and 'Report Wizard...'. The 'All Data Table' is visible, showing columns for 'GenTrain', 'Frac A', 'Frac C', and 'Frac G'. The 'Samples Table' is also visible, showing columns for 'Index', 'Sample ID', 'Call Rate', 'Gender', 'p05 Grn', and 'p50 Grn'. The 'Errors Table' is also visible, showing columns for 'Error Index', 'Error Type', 'Child/Rep Index', 'Child/Rep', 'Child/Rep GType', 'Parent1/Rep Index', and 'Parent1/Rep'. The 'Project' panel on the right shows the project structure with 'Manifests', 'Data', and 'Miscellaneous' folders.

GenomeStudio 2.0 - Genotyping - HM3subset

File Edit View Analysis Tools Window Help

SNP Graph Heat Map

Norm R

Norm Theta

Analysis Menu:

- Exclude Samples By Best Run
- Cluster All SNPs
- Update SNP Statistics...
- Edit Replicates
- Edit Parental Relationships
- Update Heritability/Reproducibility Errors...
- Reports
 - Create Reproducibility and Heritability Report
 - Compute CN Metrics...
 - Report Wizard...
- View Controls Dashboard
- Paired Sample Editor...
- Calculate Paired Sample LOH/CN...
- Import Allele Calls...
- Export Allele Calls...
- Remove Imported Allele Calls
- Create Plugin Column...
- CNV Analysis...
- Show CNV Region Display

Project Panel:

- Name: HM3subset
- Manifests
- Data
- Miscellaneous

Table 1: All Data Table (Partial View)

GenTrain	Frac A	Frac C	Frac G
0,197	0,254	0,246	
0,180	0,361	0,180	
0,262	0,213	0,303	
0,213	0,205	0,328	
0,475	0,148	0,189	
0,451	0,131	0,107	
0,8779	0,246	0,172	0,344
0,6082	0,139	0,295	0,221
0,7779	0,328	0,197	0,254
0,8861	0,377	0,156	0,164
0,8815	0,270	0,164	0,131
0,8073	0,492	0,123	0,115
0,6185	0,180	0,303	0,254
0,7603	0,189	0,131	0,139
0,8554	0,336	0,197	0,148
0,8091	0,262	0,205	0,287
0,7499	0,246	0,213	0,361
0,8273	0,221	0,287	0,385

Table 2: Samples Table

Index	Sample ID	Call Rate	Gender	p05 Grn	p50 Grn
1	20195879001...	0,0000000	Unknown	181	37
2	20195879001...	0,0000000	Unknown	210	42
3	20195879001...	0,0000000	Unknown	253	49
4	20195879001...	0,0000000	Unknown	330	54

Table 3: Errors Table

Error Index	Error Type	Child/Rep Index	Child/Rep	Child/Rep GType	Parent1/Rep Index	Parent1/Rep
-------------	------------	-----------------	-----------	-----------------	-------------------	-------------

Log

GenomeStudio - Exporting

Report Wizard - Report Type

Genotyping Report
What type of report would you like to generate?



☒ Final Report

☐ Locus Summary

☐ DNA Report

☐ Locus x DNA

☐ Custom Report

PLINK Input Report 2.1.4 by Illumina, Inc. from Illumina, Inc. ▾

Cancel

< Back

Next >

Finish

GenomeStudio - Exporting

Report Wizard - Final Report Format

Genotyping Report

How would you like to format your final report?



☒ Standard ☐ Matrix ☐ 3rd Party

Standard Format Options

Displayed Fields

- SNP Name
- Sample ID
- Allele1 - Top
- Allele2 - Top
- GC Score

Available Fields

- Sample Name
- Sample Group
- Sample Index
- SNP Index
- SNP Aux
- Allele1 - Forward
- Allele2 - Forward
- Allele1 - Design
- Allele2 - Design

Hide => <= Show

Group by ☒ sample ☐ SNP

General Options

☒ Tab ☐ Comma

☐ Create map files

Samples / File
23

Favorite Formats

Default

Save Current...

Total 52148958

Estimated Size: 1243,3 MB

Estimated File Size: 1243,3 MB

Cancel < Back Next > Finish

GenomeStudio - Exporting

Report Wizard - Destination

Genotyping Report

Where would you like to save your report?



Output Path
D:\subset\HM3subset

Report Name
HM3subset_FinalReport

GenomeStudio - Exporting

```
[Header]
GSGT Version      2.0.3
Processing Date    8/27/2018 11:48 AM
Content           H3Africa_2017_20021485_A3.bpm
Num SNPs          2267346
Total SNPs        2267346
Num Samples       23
Total Samples     23
[Data]
SNP Name          Sample ID          Allele1 - Top    Allele2 - Top    GC Score
200610-1          201958790012_R02C01    A               A               0.3482
200610-10         201958790012_R02C01    A               A               0.3616
200610-100        201958790012_R02C01    G               G               0.2703
200610-102        201958790012_R02C01    A               A               0.3483
200610-103        201958790012_R02C01    -               -               0.0000
200610-105        201958790012_R02C01    A               A               0.6442
200610-106        201958790012_R02C01    A               A               0.3100
200610-107        201958790012_R02C01    A               A               0.3737
200610-108        201958790012_R02C01    A               A               0.3482
200610-109        201958790012_R02C01    A               A               0.3482
200610-110        201958790012_R02C01    A               A               0.3462
200610-111        201958790012_R02C01    A               A               0.3868
200610-112        201958790012_R02C01    A               A               0.3716
200610-113        201958790012_R02C01    A               A               0.3716
200610-114        201958790012_R02C01    A               A               0.3716
200610-115        201958790012_R02C01    A               A               0.3010
200610-116        201958790012_R02C01    A               A               0.3716
200610-117        201958790012_R02C01    A               A               0.3462
200610-118        201958790012_R02C01    A               A               0.3833
200610-119        201958790012_R02C01    A               A               0.3583
200610-12         201958790012_R02C01    G               G               0.6883
200610-120        201958790012_R02C01    A               A               0.2815
200610-121        201958790012_R02C01    A               A               0.3737
200610-122        201958790012_R02C01    A               A               0.3333
200610-123        201958790012_R02C01    A               A               0.3833
200610-124        201958790012_R02C01    A               A               0.5062
200610-125        201958790012_R02C01    A               A               0.2713
200610-126        201958790012_R02C01    A               A               0.3333
200610-127        201958790012_R02C01    A               A               0.2637
```

TOP/BOT

TABLE 2: STRAND AND ALLELE DESIGNATIONS FOR AMBIGUOUS SNPS

SNP Name ¹	Sequence ²	Strand Designation	Allele A	Allele B
rs1535632	...ACGGGGACAG[A/T]TATGTTAACT...	BOT	T	A
rs363334	...ATGAGTGAAAT[C/G]AAGCACTATT...	TOP	C	G
rs7101540	...AAATTCAGAT[A/T]CAGAATCTTT...	TOP	A	T
rs7113791	...GATGGACAGG[A/T]TGACCTCTAG...	BOT	T	A
rs778833	...GGTTAAATGT[C/G]AAGGTGAGCT...	BOT	G	C
rs4933195	...ATGCTAATAA[A/T]ACATTAAAGT...	TOP	A	T
rs903997	...ATGAGAAAGT[C/G]TGAGAGTGCA...	TOP	C	G
rs1942968	...CTACATGACT[C/G]TTTATGTTAC...	BOT	G	C

¹SNP name and sequence are from dbSNP version 126

²Sequence shortened to 10 base pairs on each side of SNP as noted by the ellipses (...) for illustrative purposes. The unambiguous pairings used to determine Strand and Allele are noted in red.

*An unambiguous pairing refers to a pair of nucleotides that are of the following combinations: A/G, A/C, T/C, or T/G. The pair is considered to be the nucleotides that are equidistant on either side of the SNP (e.g., n-1/n+1 or n-5/n+5).

- 1) https://www.illumina.com/documents/products/technotes/technote_topbot.pdf
- 2) S. C. Nelson, K. F. Doheny, C. C. Laurie, and D. B. Mirel, "Is 'forward' the same as 'plus'?...and other adventures in SNP allele nomenclature." *Trends in Genetics*, vol. 28, no. 8, pp. 361-363, 2012. <http://dx.doi.org/10.1016/j.tig.2012.05.002>

Troubleshooting

- Bad annotations
- Legacy probes
- Incorrect location
- TOP/BOT vs FWD/REV
- Plink swapping alleles
- Be critical

Resources



European Variation Archive

[Home](#)[Submit Data](#)[Study Browser](#)[Variant Browser](#)[Clinical Browser](#)[GA4GH](#)[API](#)[dbSNP Import Progress](#)[Help](#)[Feedback](#)

EVA / VARIANT BROWSER

Variant Browser

Search the EVA variant warehouse using any combination of the filtering options on the left hand-side.
Search results can be exported in CSV format and individual variants can be further investigated using the in-depth Variant Data tabs found below the main results table.

Filter

Reset

Search

Genome Assembly

Organism / Assembly:

Human / GRCh37

Position

Filter By:

Chromosomal Location

1:1139216-1139217

Annotation

Select Version:

Variants found

« ‹ Page 1 of 1 › » ↻

Chr	Position	Variant ID	Alleles	Class	Most Severe Consequence Type
1	1139082	-	ACCTGGGTCCTG...	SV	
1	1139128	-	ACGTCGAACCGG...	SV	
1	1139216	rs11466695	G/A	SNV	missense_variant

Resources

[< prev](#)

Composite probe view*

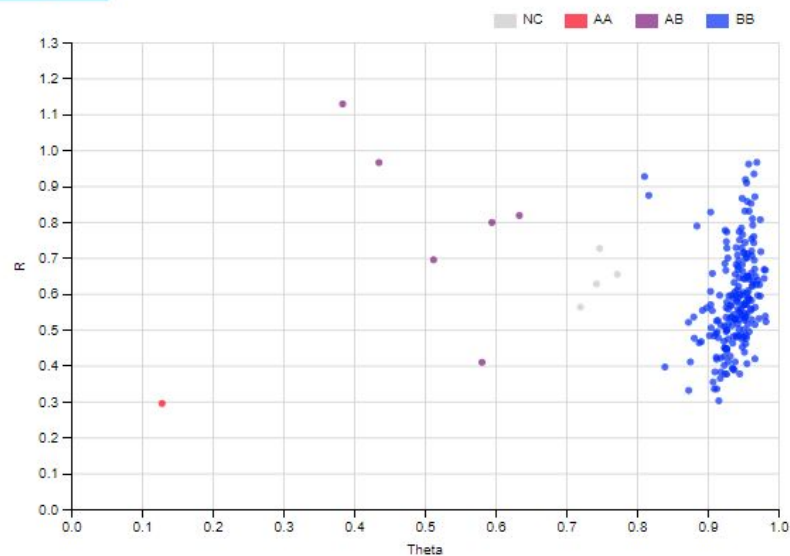
[Next >](#)[Toggle probe orientation](#)[Toggle probe plot](#)

H3A design request

Probe orientation

```
GGGGCGAGCGATCGGCAGAGGAGAAGGGGCGGCTGGGAGACCTGTGGGTGTGAGCCTGGC [T/C] GTCCTCCGGGGCCNCCGACCGCAGCCAGCCCTCCCCAGGAGCTCCCCAGGCCGCAGC  
CCCCGCTCGCTAGCCGTCTCCTCTTCCCCGCCGACCTCTGGACACCCACACTCGGACCG [A/G] CAGGAGGCCCCGGNGGCTGGCGTCGGTCGGGGAGGGGTCTTCGAGGGGTCCGGCGTC
```

Probe intensities

[Grey calls](#)

IlmnID
Name

rs11466695-131_B_F_1893944640
rs11466695

BeadSetID
AddressA_ID

1201
0028782891