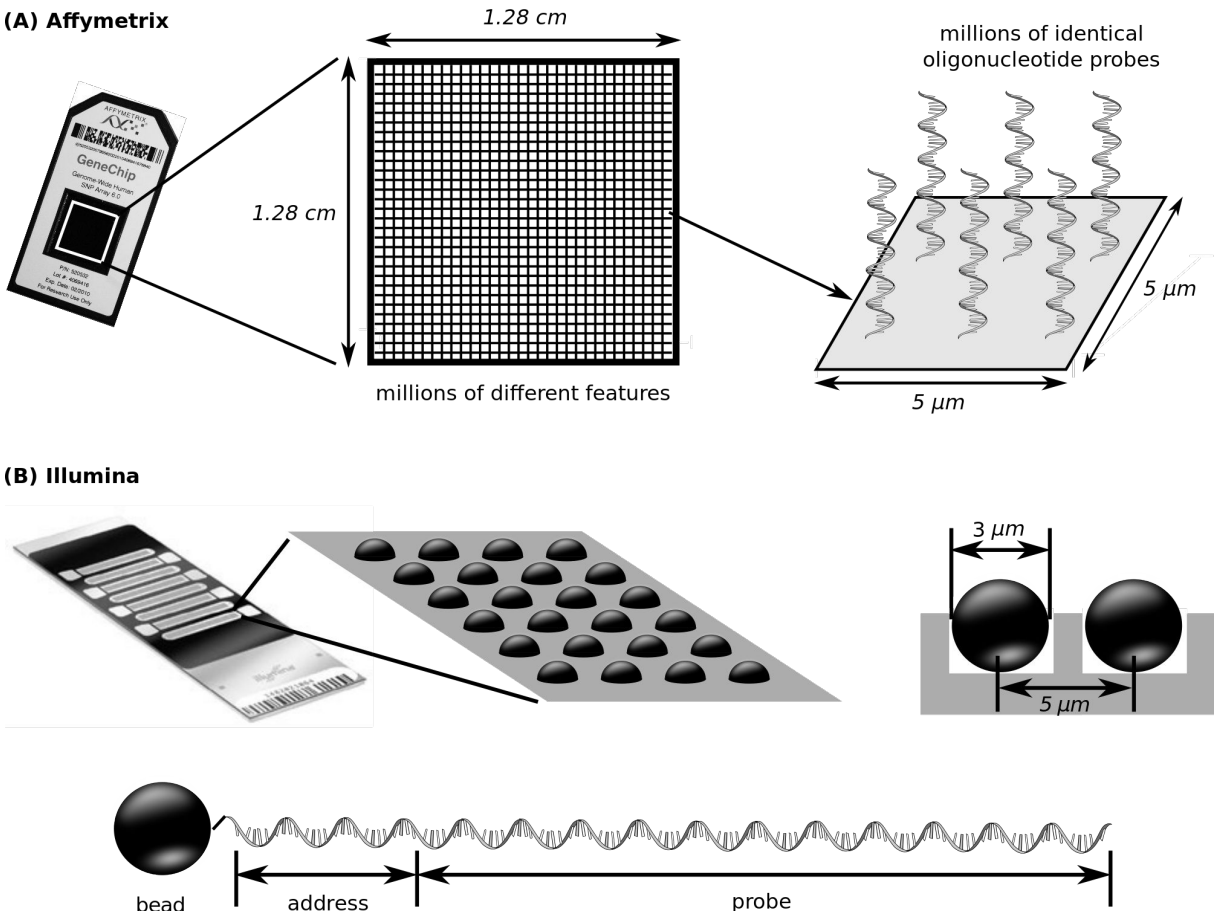
GWAS workshop

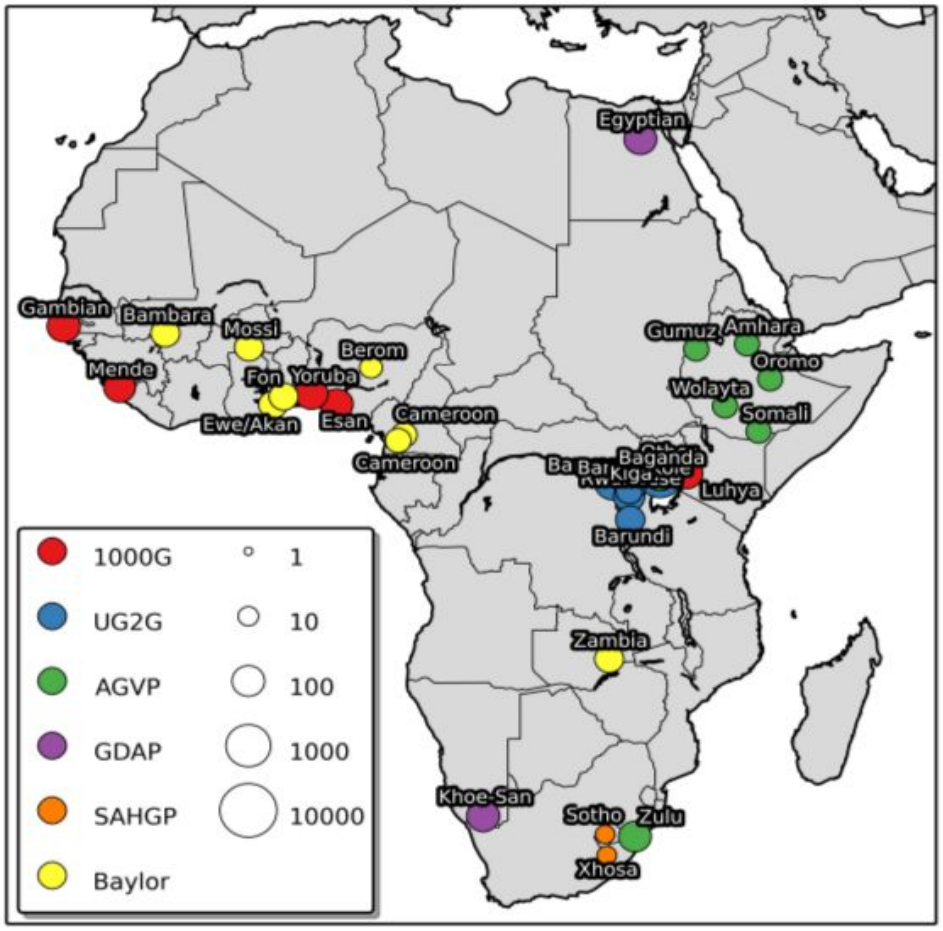
Processing array data with GenomeStudio

Intro - Genotyping Microarrays



Intro - H3A chip

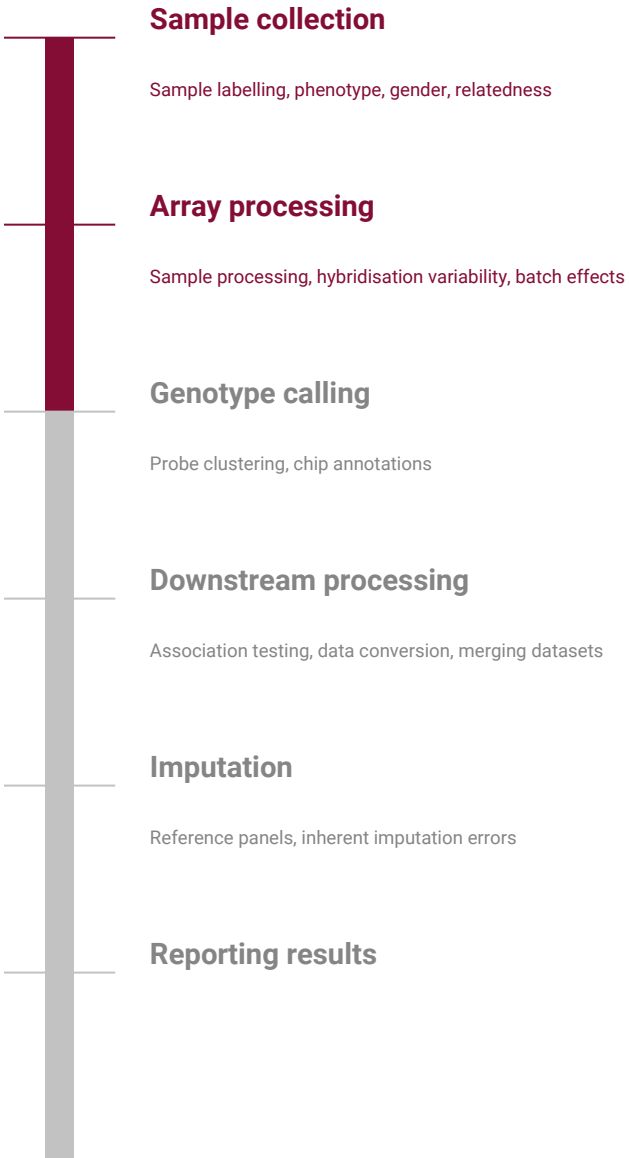
Source	Number of samples
GDAP	204
UG2G	2000
AGVP	320
1000G*	~2500
TrypanoGEN	168
Baylor	350
SAHGP	16
Total	**~5500**



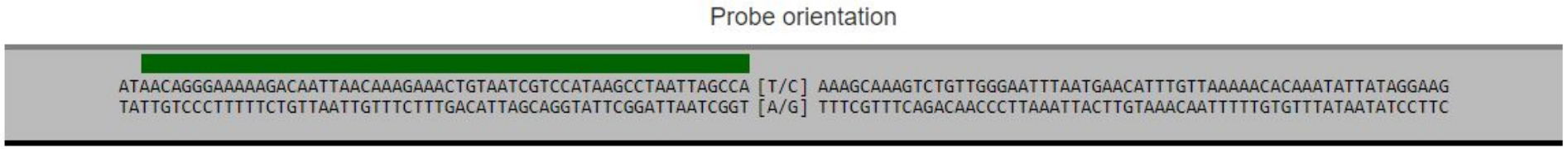
Intro - H3A chip

- 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



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   TCTACGTCTATTCCTACTG…
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   TCATCCTATTATTTATCGC…
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   GGAAATTTTTTGTTATGAT…
200610-107-0_B_F_1867864670     200610-107          BOT         [T/C]  0022767164   ACCCCCCAACTAACACATT…
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   ATAATTAAACTTTACTTCC…
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   GTAAAAAACTCCAGTTGAC…
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   CTCTGTTCTTTCATGGGGA…
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   AAATTCAAGTGGTTTGCAT…
h3a-req-37_13_37486207_T_C_ver2-… h3a-req-37_13_37486… TOP      [A/G]  0002800215   AAATTCAAGTGGTTTGCAT…
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

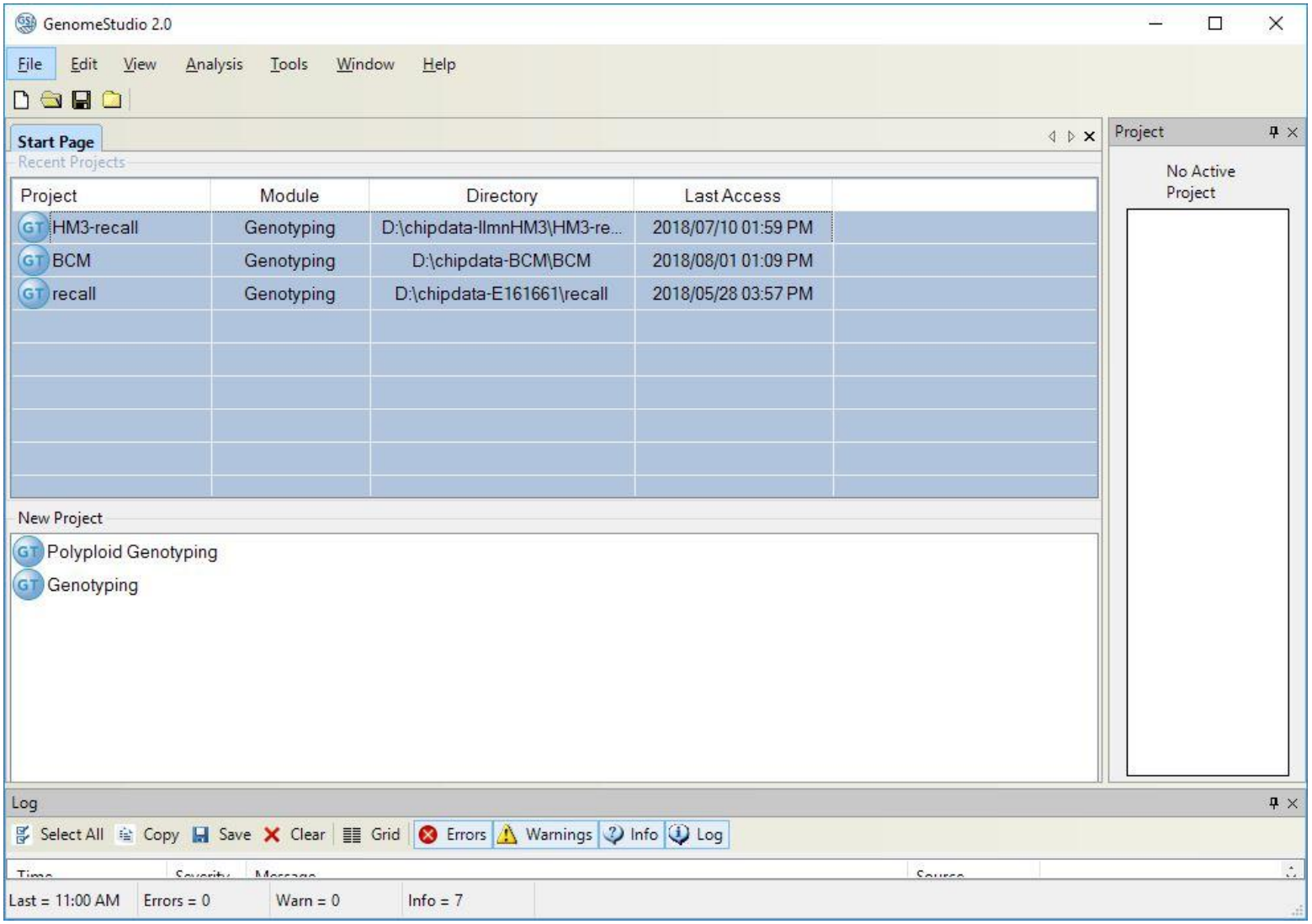
snp-known10000012-142_T_F_2556548951

SNP106-0_B_R_1852975650

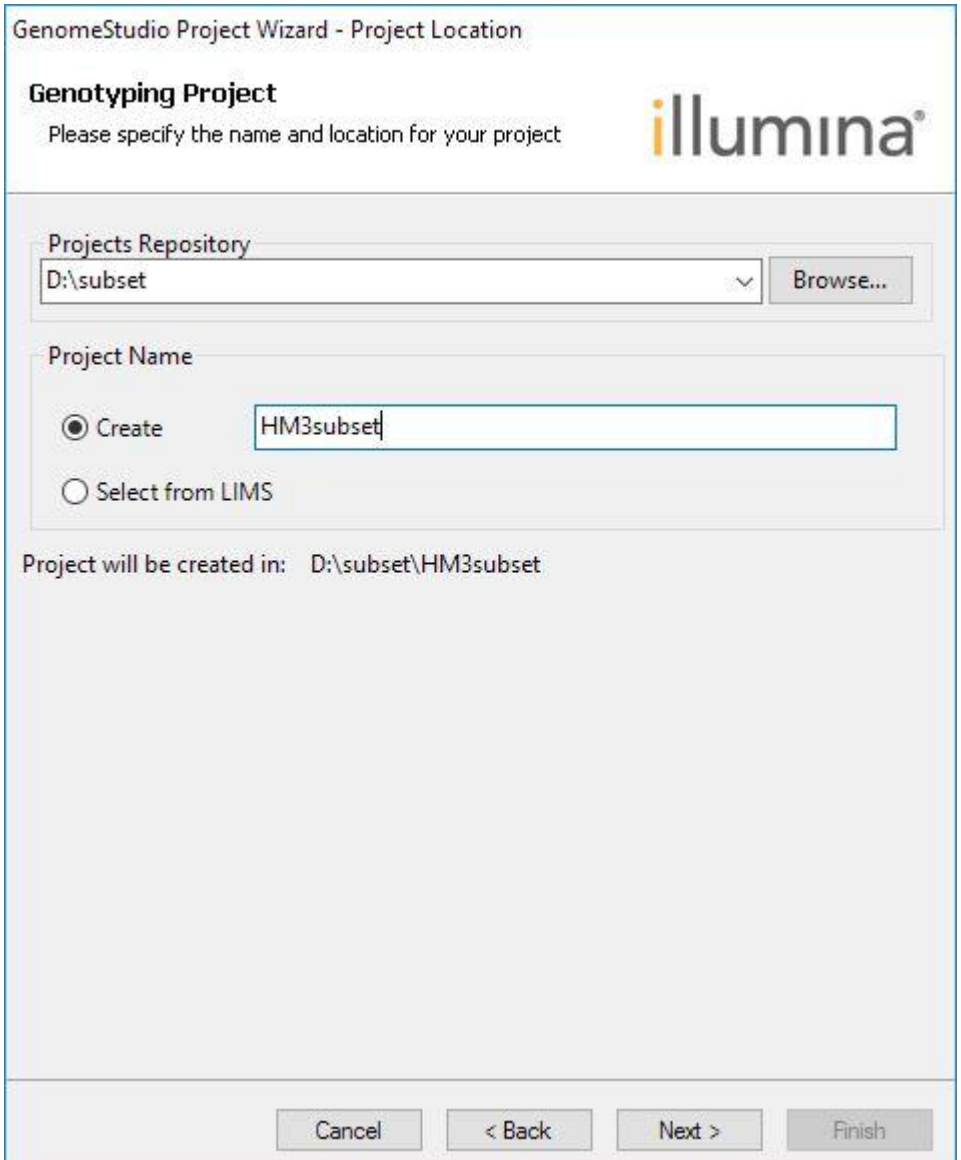
VG01S1133-0_B_R_1852963053

h3a-req-37_1_16534074_G_T_ver1-1_T_R_2598050152
h3a-req-37_1_16534074_G_T_ver2-1_T_R_2598067940

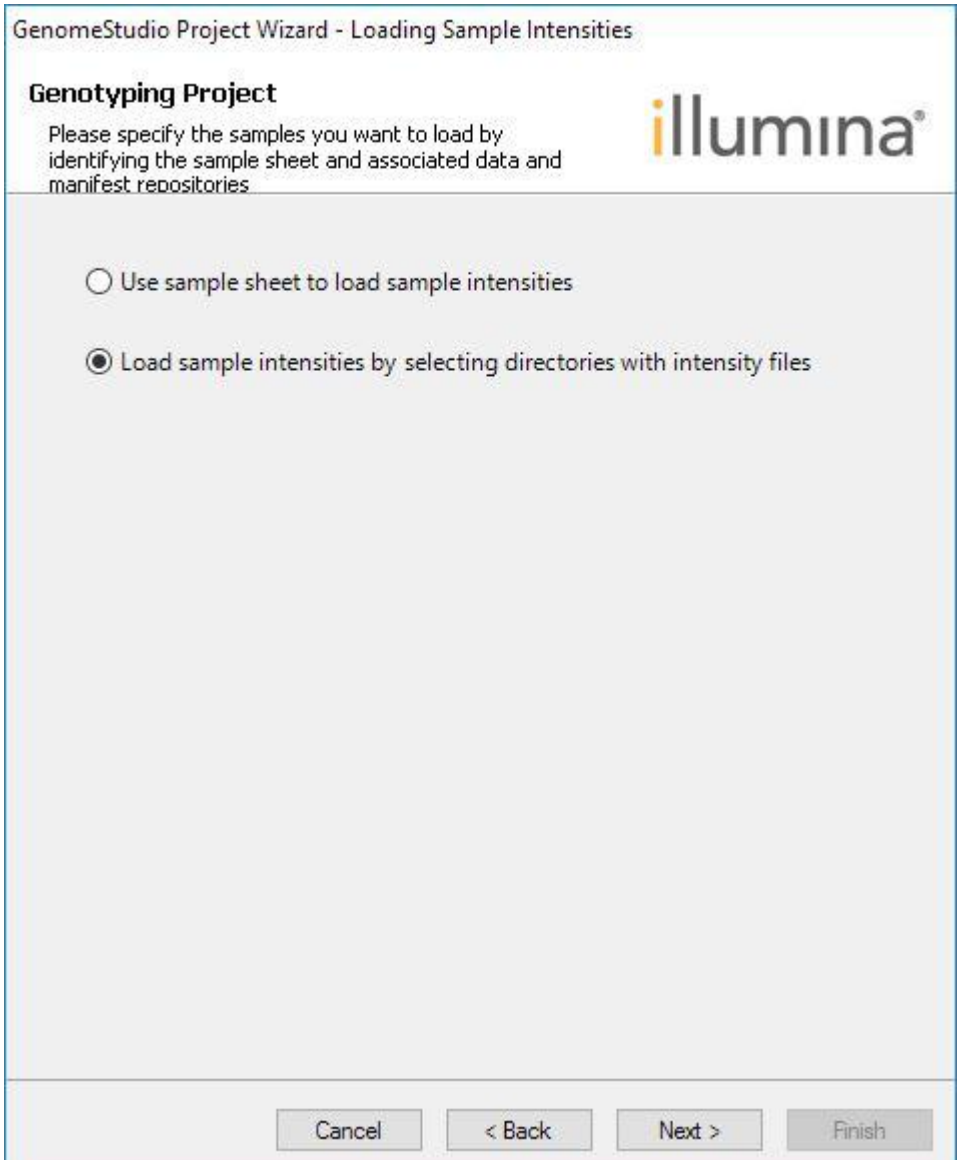
GenomeStudio - start project



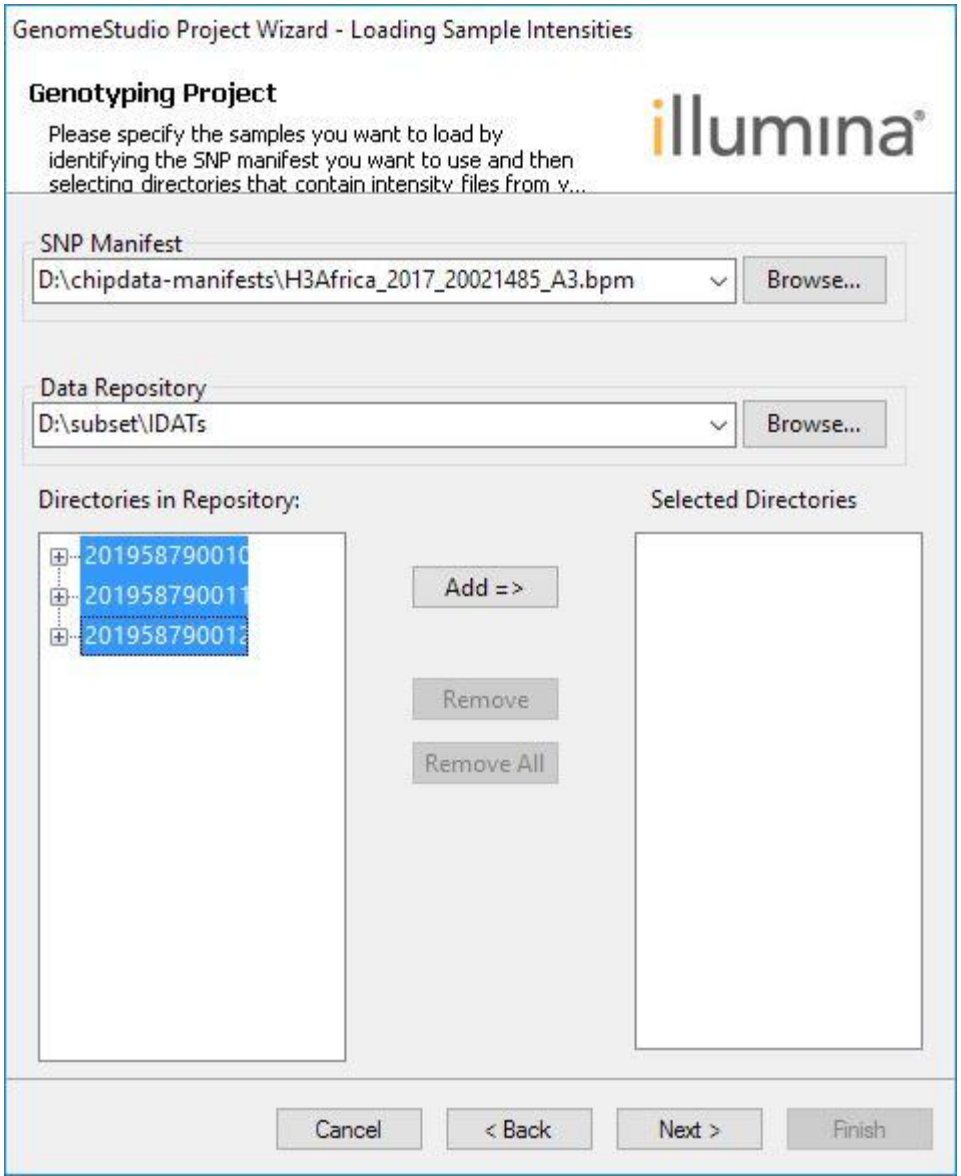
GenomeStudio - start project



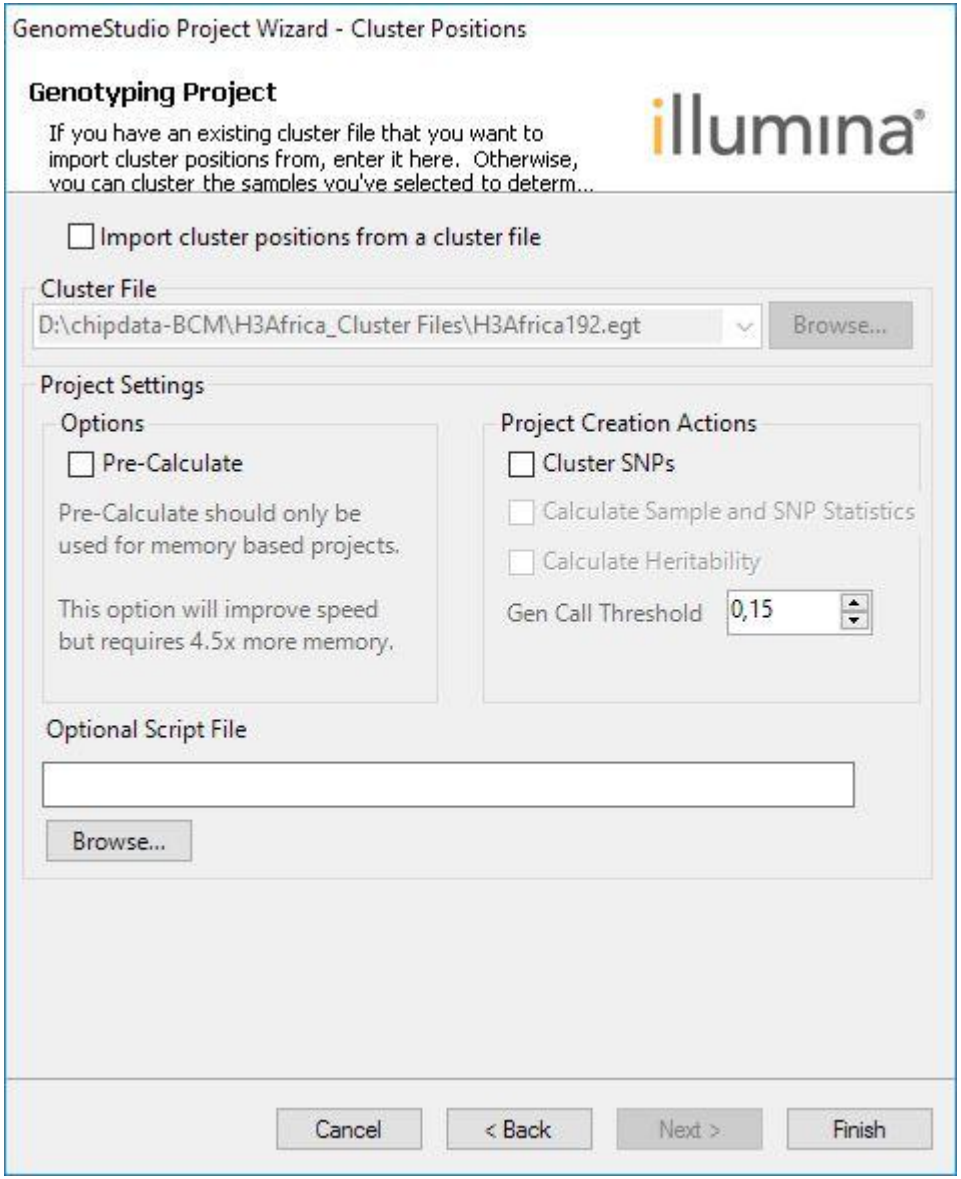
GenomeStudio - start project



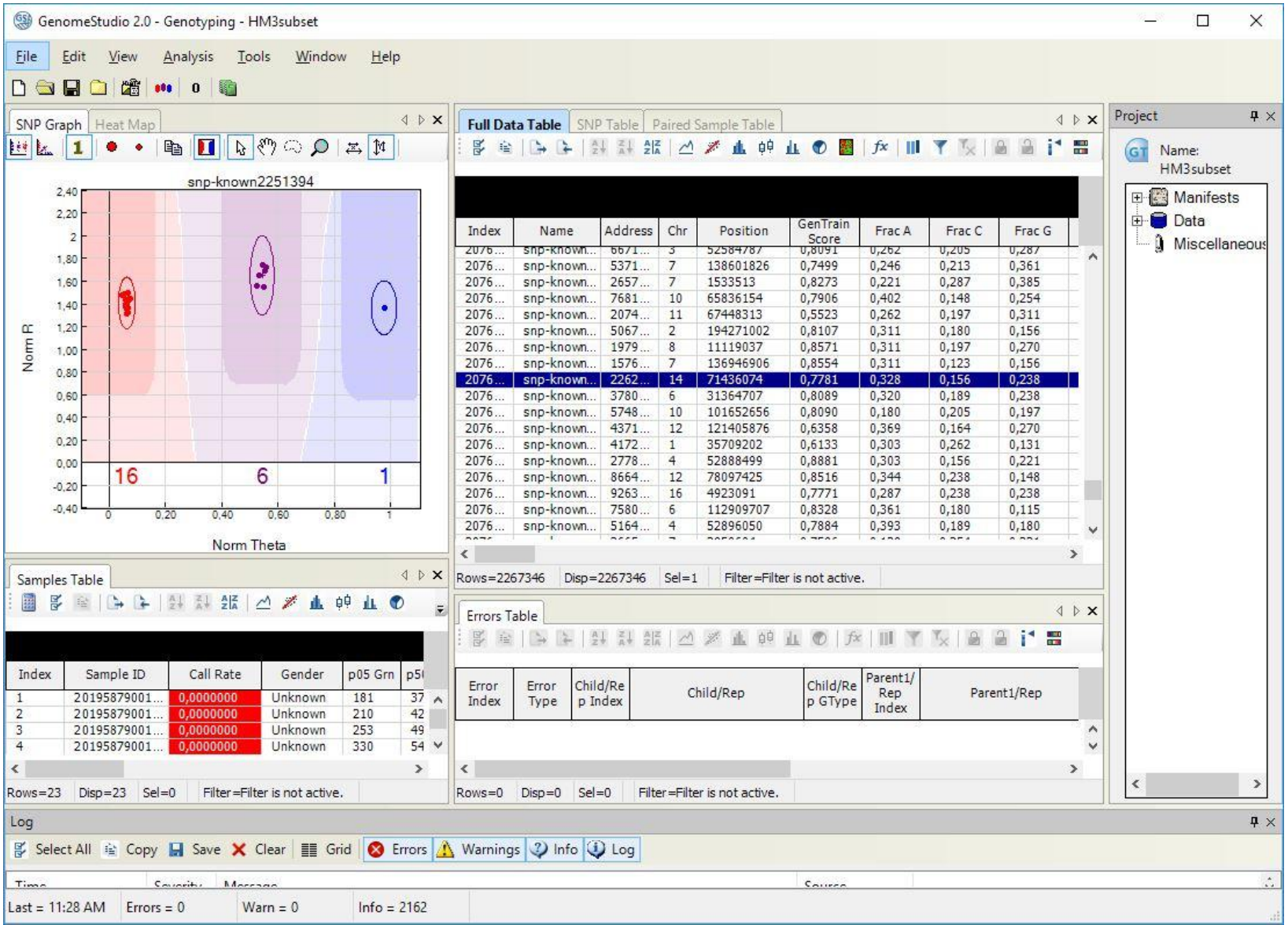
GenomeStudio - start project



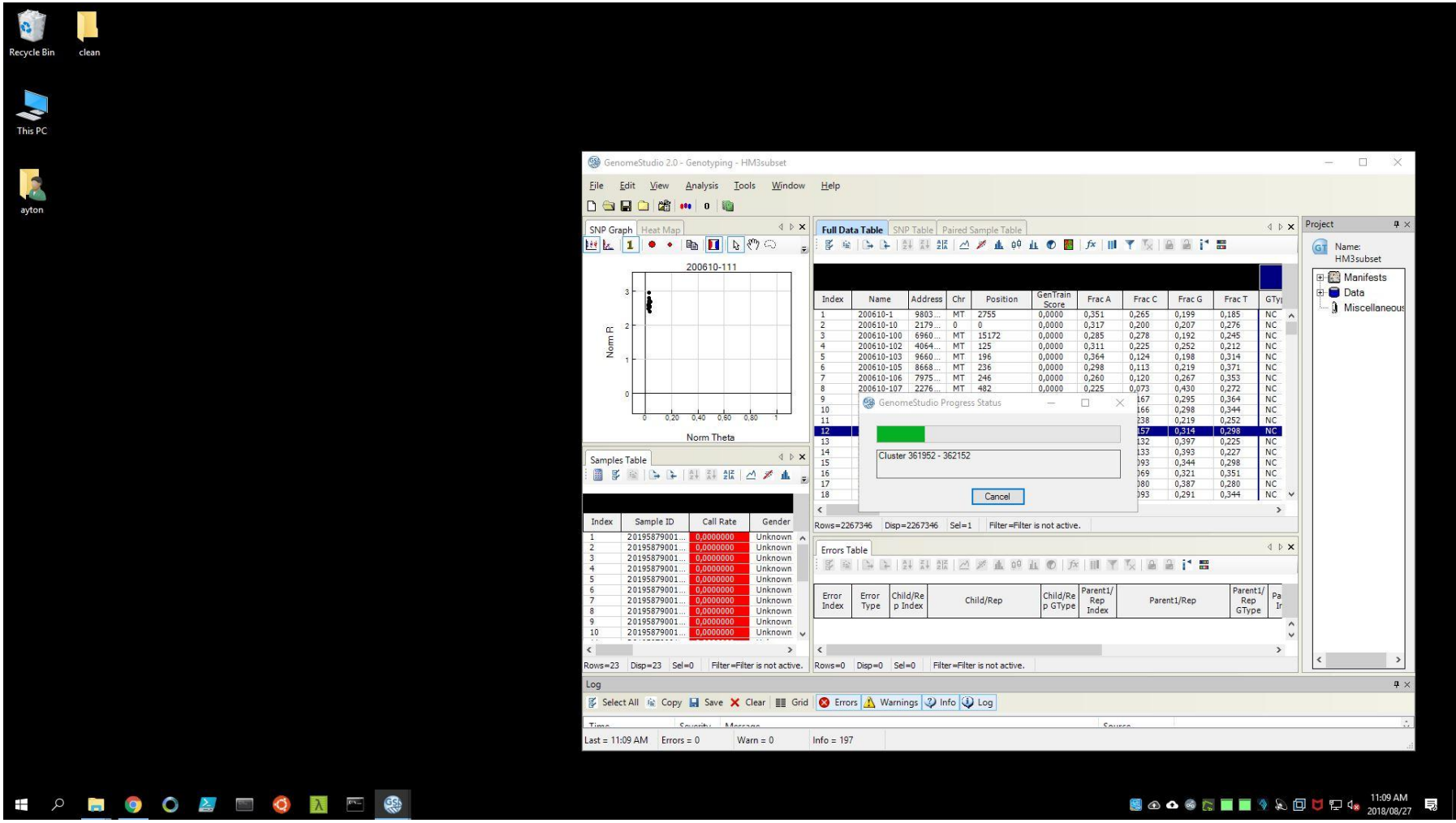
GenomeStudio - start project



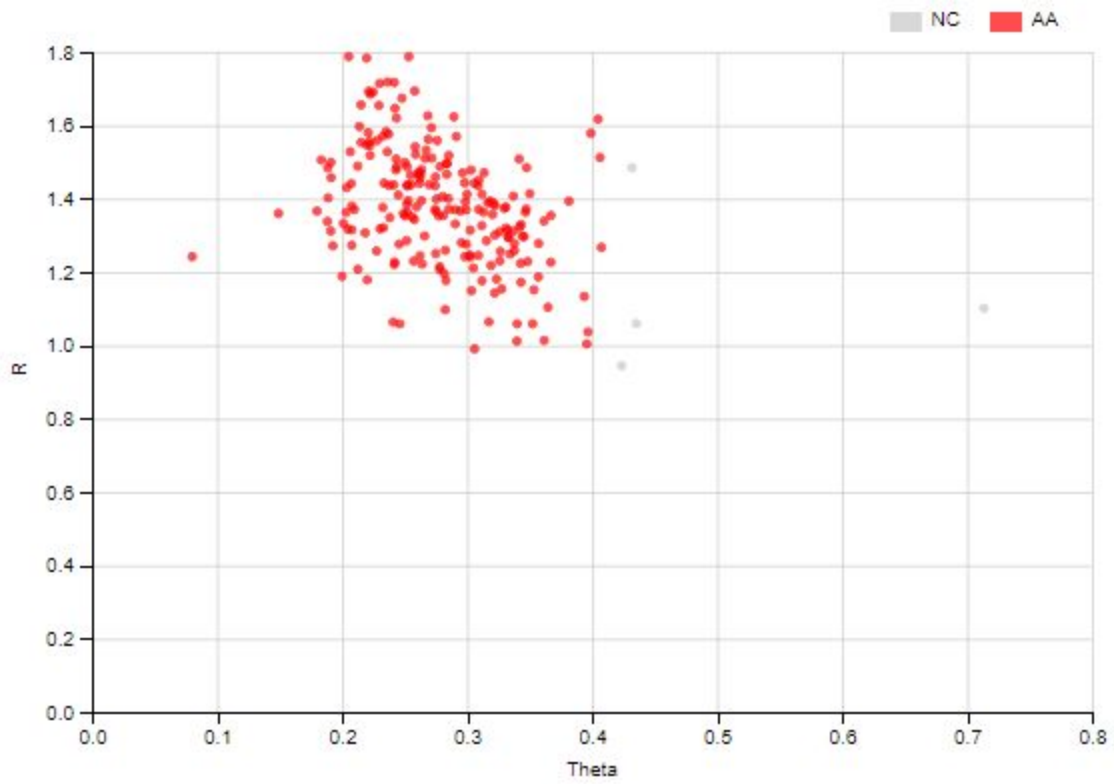
GenomeStudio - overview



GenomeStudio - clustering

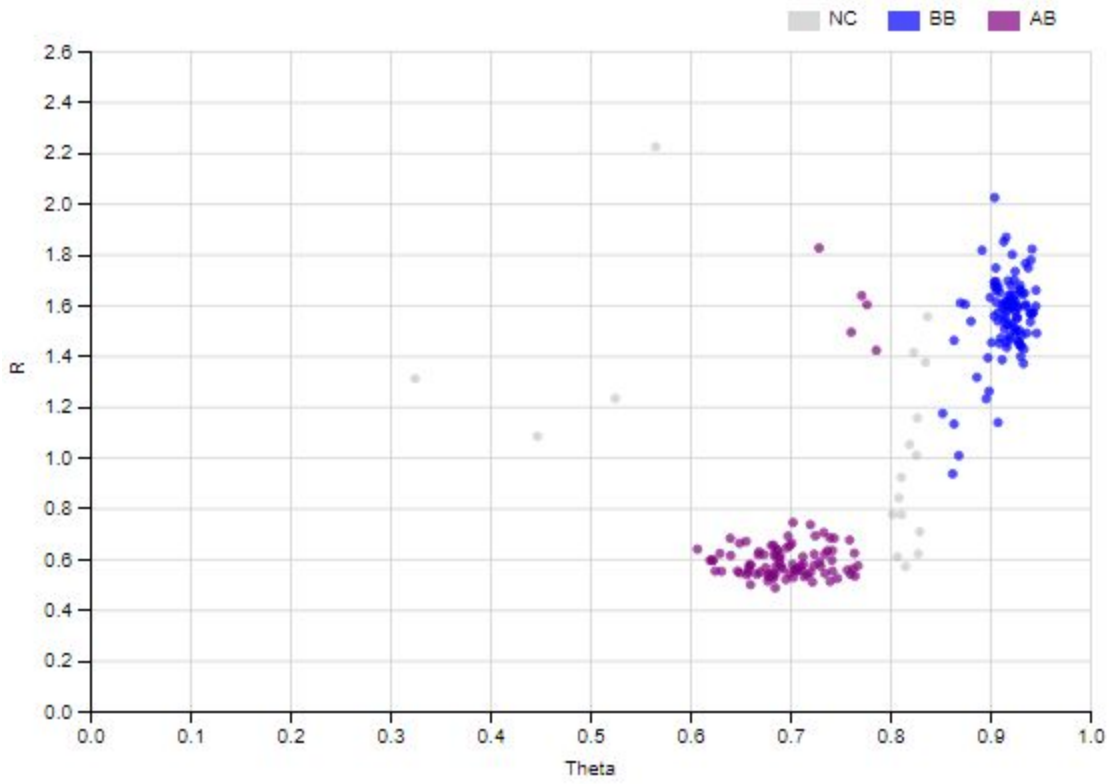


Troubleshooting - Non-polymorphic



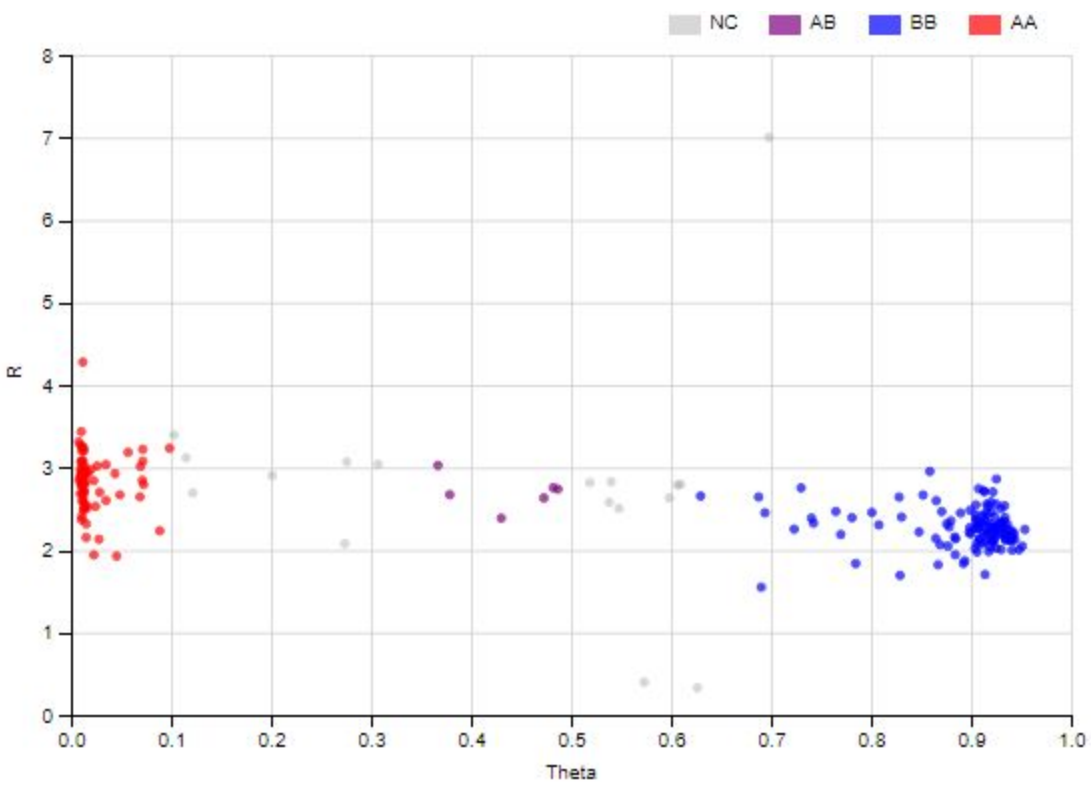
IlmnID	snp-known35339498-142_B_R_2367222456
Name	snp-known35339498
NUID	1:1150211

Troubleshooting - incorrect ploidy Y



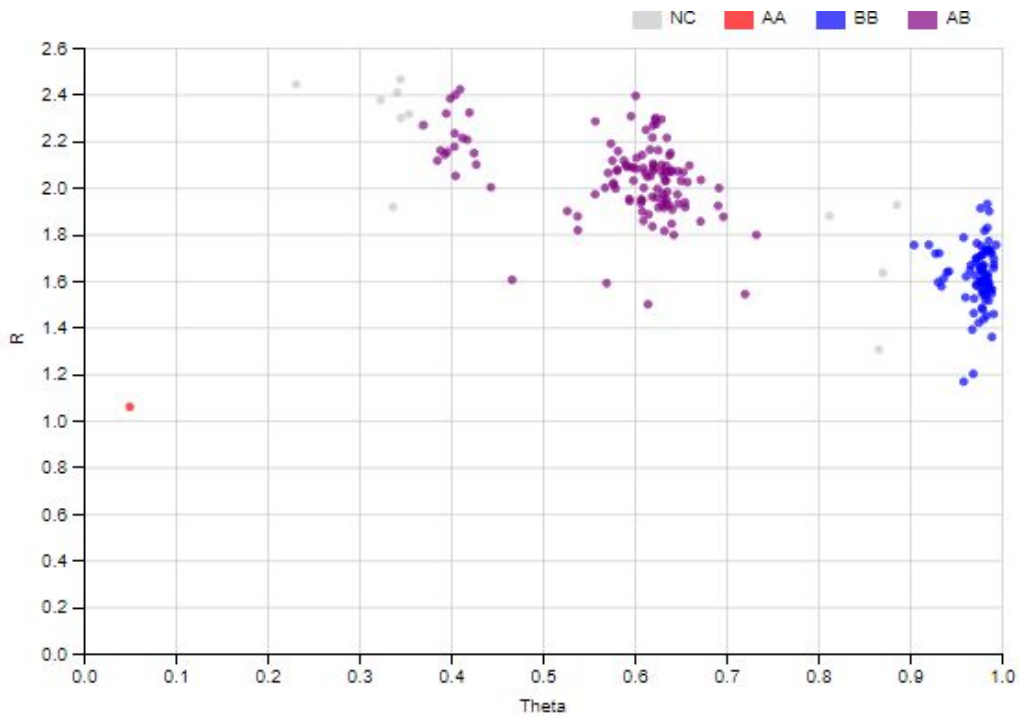
IlmnID	2010-08-Y-1952-0_T_F_1867866099
Name	2010-08-Y-1952
NUID	0:0

Troubleshooting - Incorrect ploidy MT



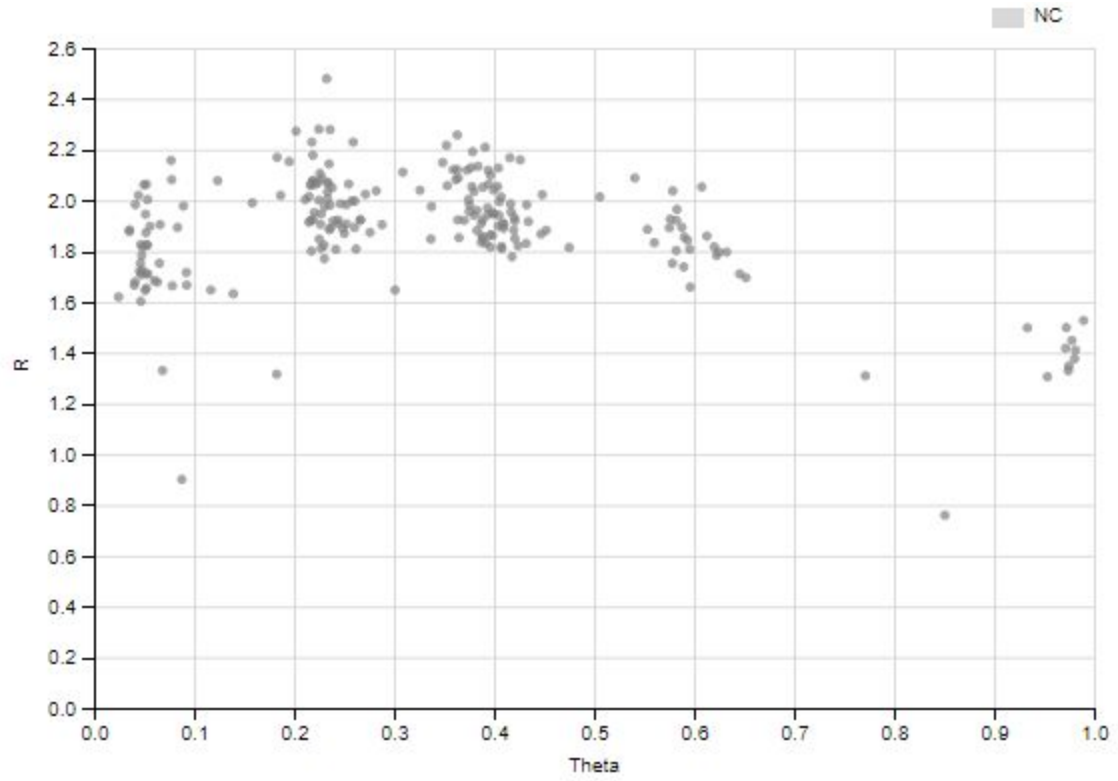
IlmnID	MitoG247A-1.32735e+007_B_R_1501342270
Name	MitoG247A
NUID	MT:247

Troubleshooting - False clustering



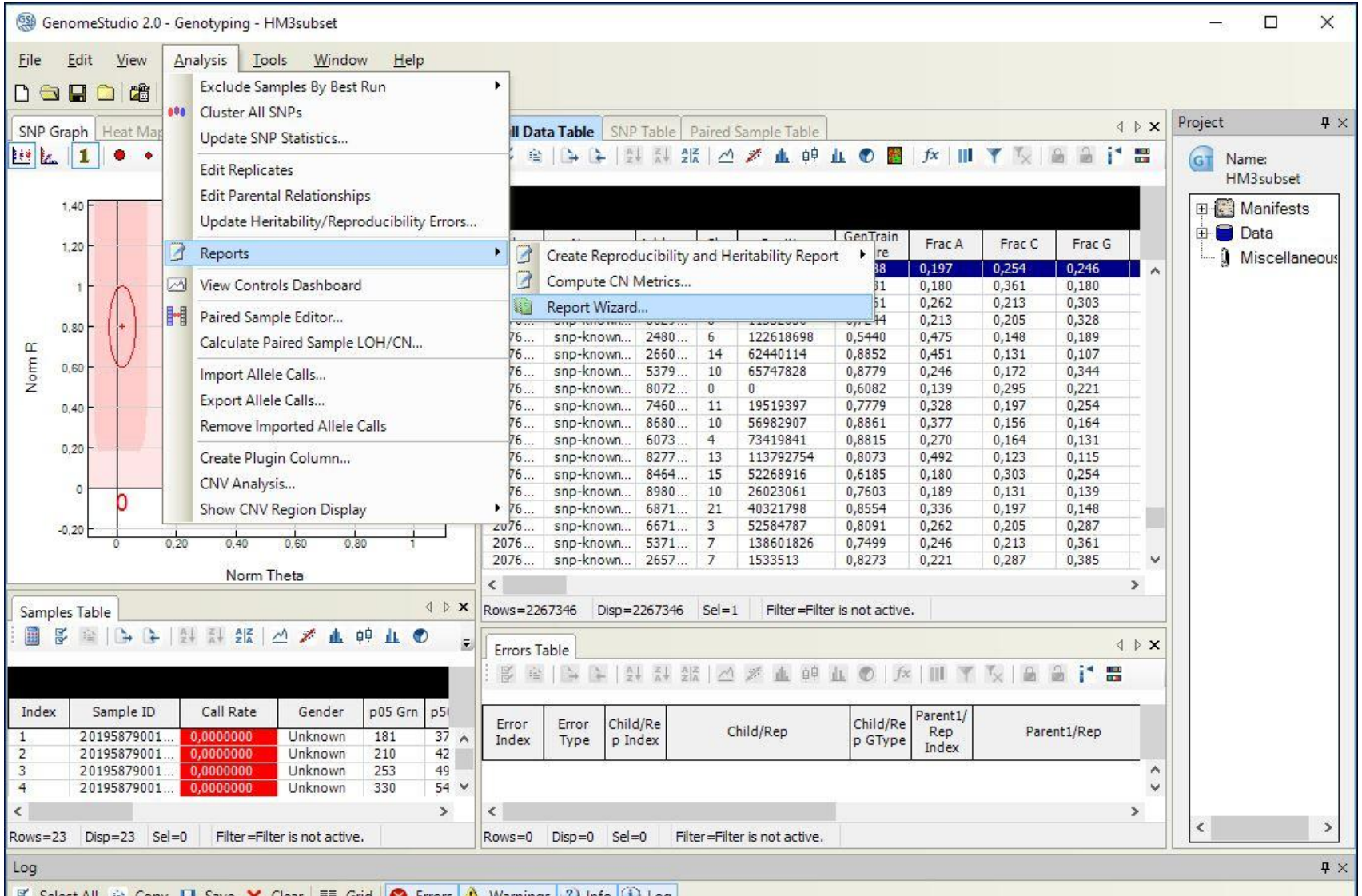
IlmnID	h3a_37_1_754964_C_T-6_T_R_2594103111
Name	h3a_37_1_754964_C_T
NUID	1:754964

Troubleshooting - multiple mappings

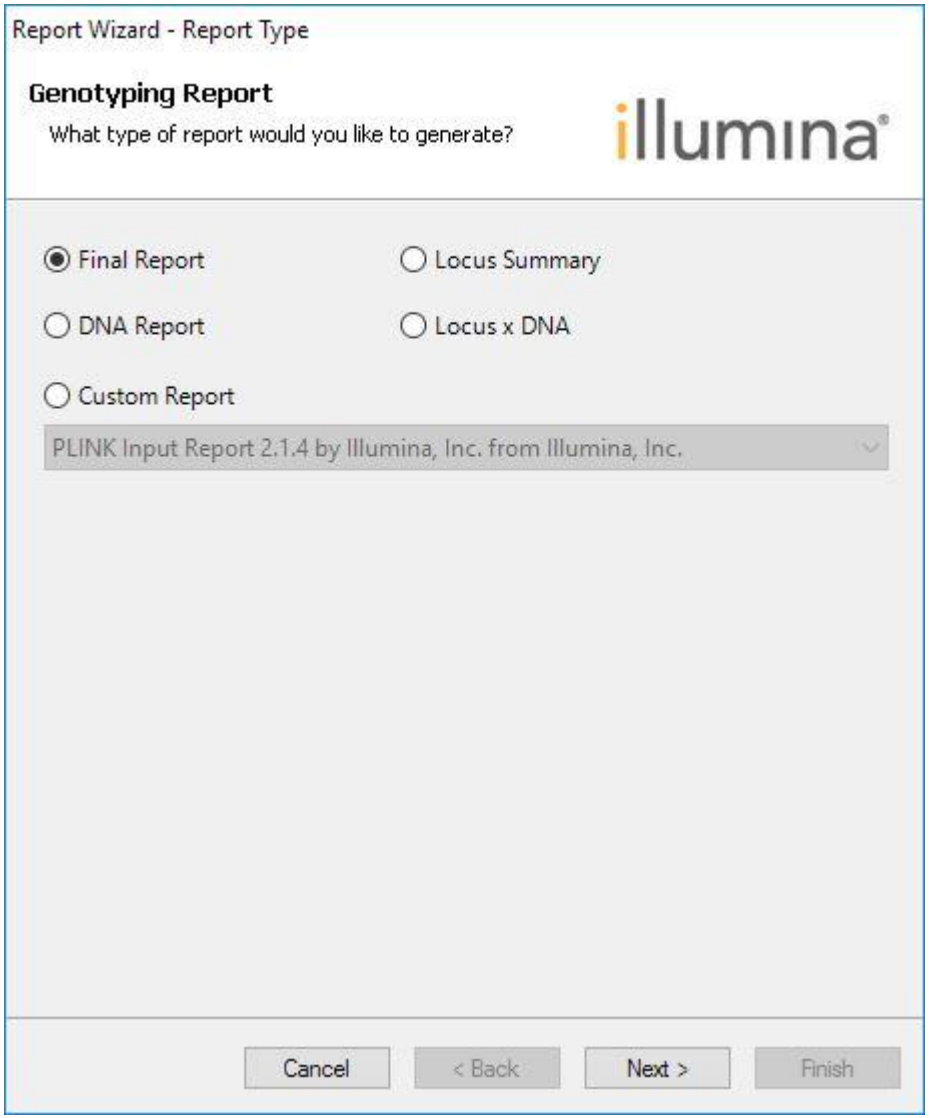


IlmnID	h3a_37_19_22704447_T_C-6_T_R_2594837377
Name	h3a_37_19_22704447_T_C
NUID	19:22704447

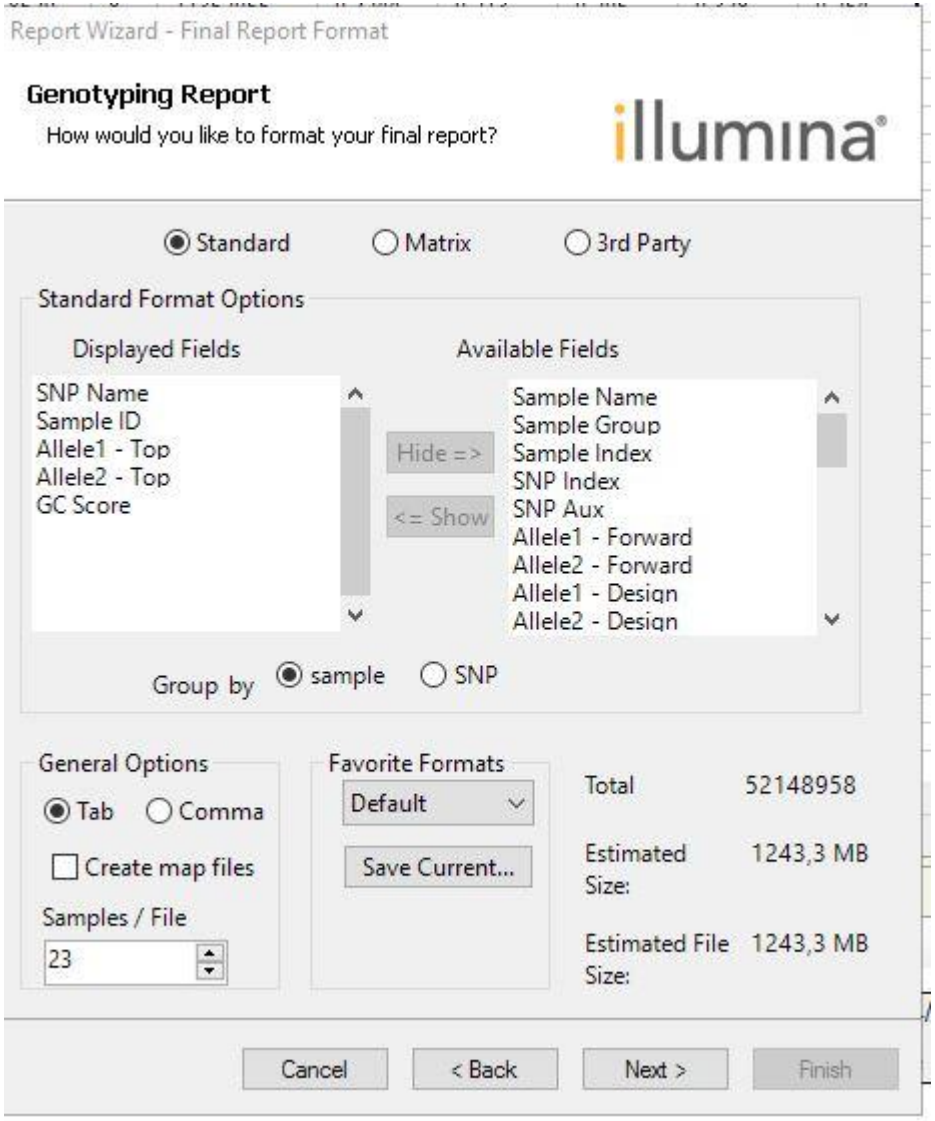
GenomeStudio - Exporting



GenomeStudio - Exporting



GenomeStudio - Exporting



GenomeStudio - Exporting

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	...ATGAGTGAAT[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

[1]SNP name and sequence are from dbSNP version 126
[2]Sequence 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.

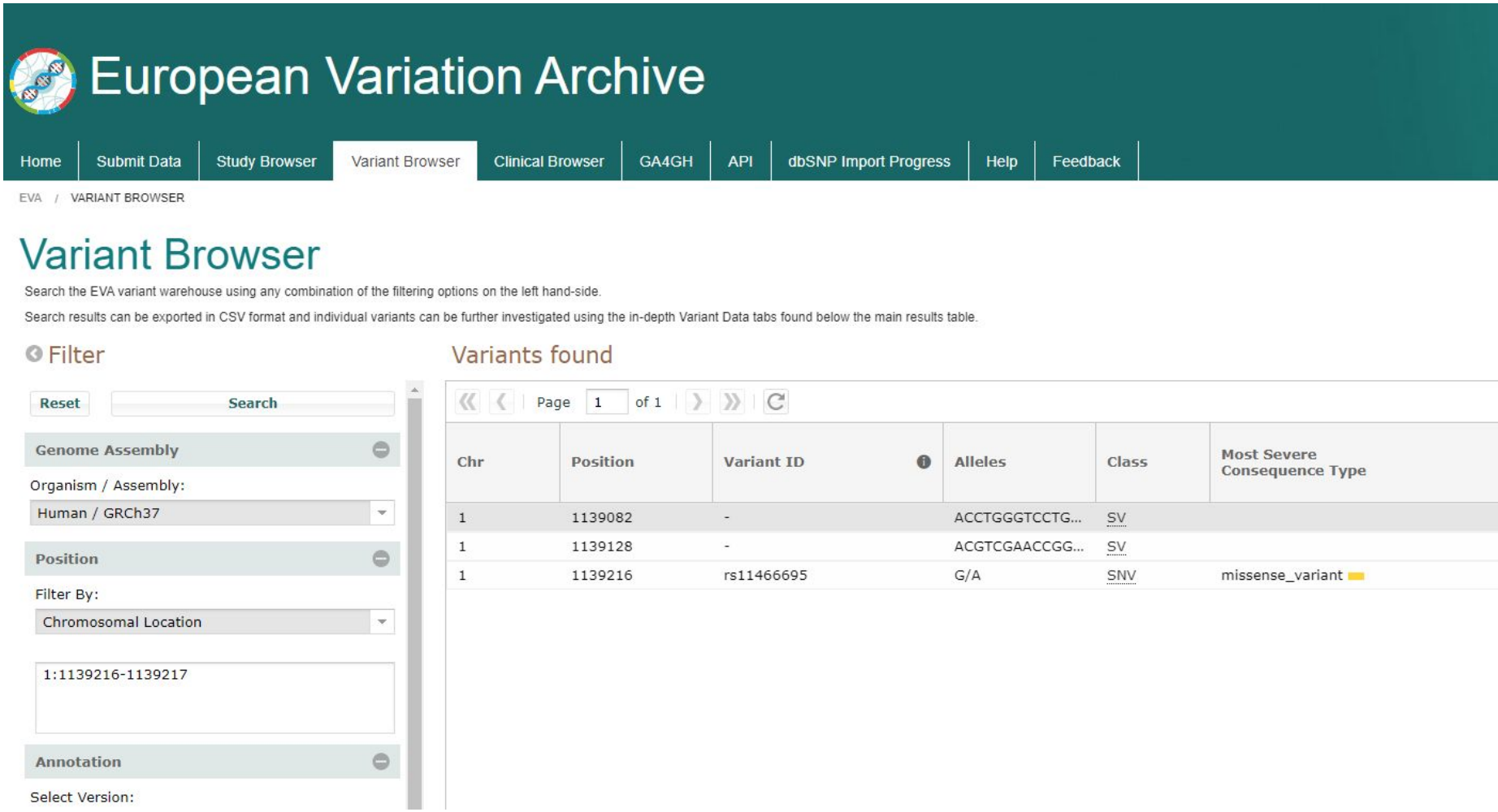
*An unambiguous pairing refers to a pair of nucleotides that are of the following combinations: A/G, A/C, T/C, or T/G. The pair is considered to be the nucleotides that are equidistant on either side of the SNP (e.g., n-1/n+1 or n-5/n+5).

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

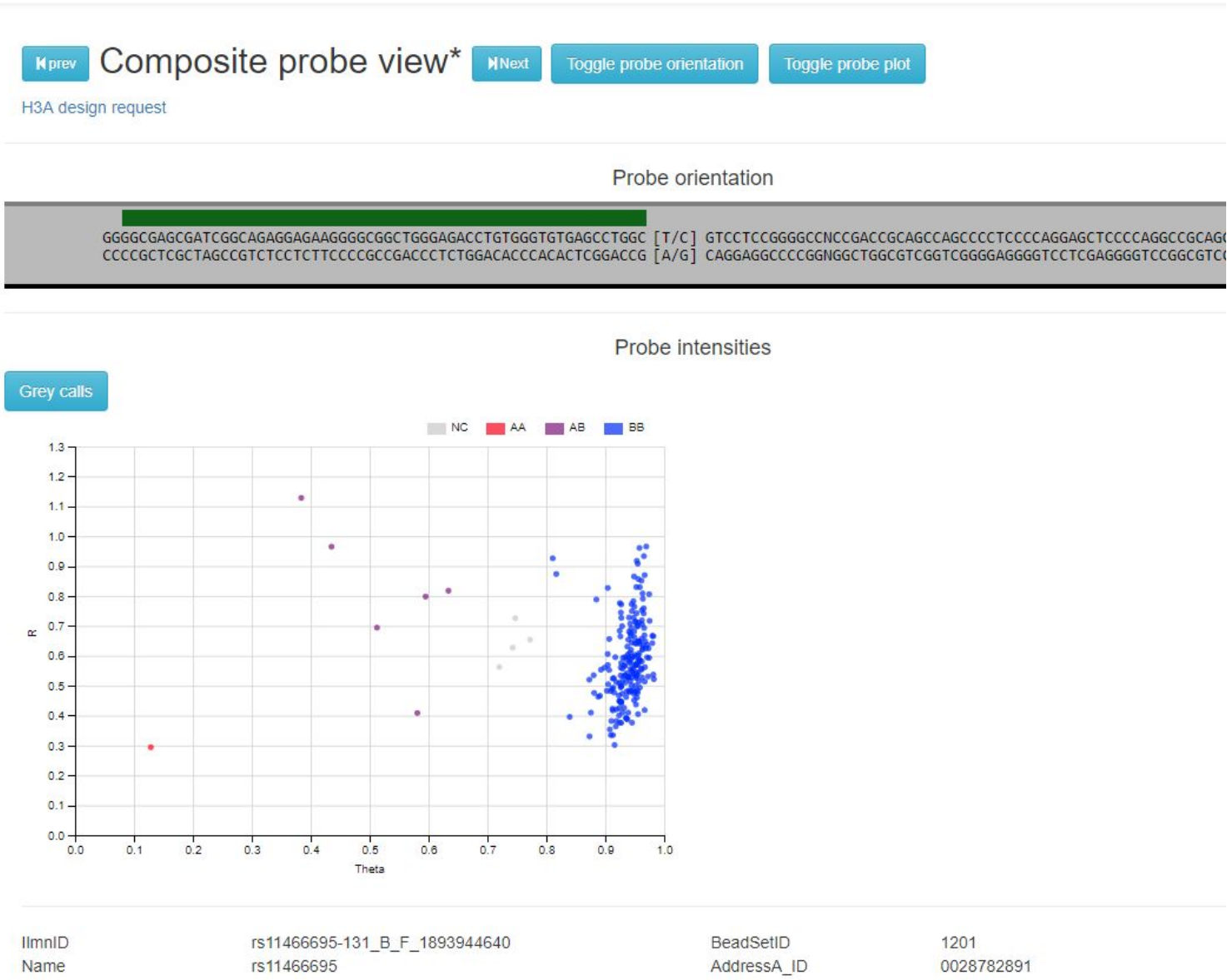
Troubleshooting

- 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] GTCCTCCGGGGCCNCCGACCGCAGCCAGCCCCTCCCCAGGAGCTCCCCAGGCCGCAG
CCCCGCTCGCTAGCCGTCTCCTCTTCCCCGCCGACCCTCTGGACACCCACACTCGGACCG [A/G] CAGGAGGCCCCGGNGGCTGGCGTCGGTCGGGGAGGGGTCCTCGAGGGGTCCGGCGTC
Probe intensities
Grey calls
NC
AA
AB
BB
R
Theta
IlmnID rs11466695-131_B_F_1893944640
Name rs11466695
BeadSetID 1201
AddressA_ID 0028782891