QC Uganda

09/08/2012 data release

1 Summary

The data QC process described here consists of a series of sequential steps arranged in a logical order which must not be altered as this may affect the final outcome. The steps are grouped under sample QC and SNP QC.

We started from 4,872 samples and 2,269,626 autosomal markers. During data preparation 44,548 autosomal markers were excluded (Section 3). A total of 91 samples were dropped during sample QC as they did not pass the quality thresholds for sample call rate (>97%) or heterozygosity (H_O=0.209333±0.007416 corresponding to the mean±3SD), or the gender inferred from the X-chromosome data did not match the supplied gender. Three additional samples were dropped because of high relatedness (i.e. IBD>0.90). No samples were identified as population/ancestry outliers. During SNP QC 39,368 additional autosomal markers were excluded because they did not pass the quality thresholds for SNP call rate (>97%, 25,037 SNPs) and HWE ($p<10^{-8}$, 14,331 SNPs). The final post-QC dataset comprised 4,778 samples and 2,230,258 autosomal markers.

2 Released data description

2.1 Samples

- Number of distinct samples sent to genotyping: 5,000
  - 140[1] samples replaced by repeated genotyping
  - 103 samples excluded as 1st round call-rate fails
  - 4 samples excluded as laboratory fails
  - 2 samples withdrawn due to consent issues
  - 1 sample included in duplicate
- **Number of samples released** after genotype calling by GAPI: **4,892**
  - Number of males : 2,106
  - Number of females : 2,786

2.2 Markers

- 2,379,855 markers genotyped on the HumanOmni2.5-8 chip and on .bim file
  - **2,314,174 autosomal markers**
  - 55,208 X-chromosome markers
  - 2,561 Y-chromosome markers
  - 418 pseudo-autosomal markers
  - 256 mitochondrial markers
  - 7,238 unplaced markers assigned to chromosome 0

[1] 140 samples with low call rates were replaced by re-runs. One additional sample was re-genotyped by mistake and was excluded from the release before analyses.

3 Pre-QC Data preparation

3.1 Sample exclusions

- 19 samples withdrawn due to consent issues
- 1 duplicate removed
- **Number of samples retained**: **4,872**
  - Number of males : 2,098
  - Number of females : 2,774

3.2 Marker exclusions

- Extract autosomal SNPs from unfiltered .bim files:
  - **2,314,174** autosomal markers extracted
- Remove SNPs with rsID/chromosome/position/allele mismatches between .strand[2] and .bim files
- Remove SNPs found in .miss file[2]
- Remove SNPs found in .multiple file[2]
- Remove any unique SNPs found in .bim file but not in .strand files
  - The union of the above sets includes 41,406 autosomal SNPs (Figure 1).
- Remove duplicate SNPs in terms of chromosome/positions. Keep the SNP with rs id preferentially; where no rsid is present, choose the first one; where both SNPs have an rs id, choose the first one
  - 3,142 autosomal markers to be excluded
- **44,548 autosomal** markers excluded in total
- **2,269,626 autosomal markers retained**
- **50,130 X chromosome markers retained** for sex check

	all	autosomal	X
.bim	2,379,855	2,314,174	55,208
.strand	2,379,514	2,320,543	55,902
position mismatch	10,796	2,198	102
.miss	153	117	0
.multiple	44,558	39,307	2,667
unique to .bim	341	134	0
union	53,288	41,406	2,695
duplicates	5,525	3,142	2,383
excluded	58,813	44,548	5,078
retained	2,321,042	2,269,626	50,130

Table 1 – Summary of SNP exclusion prior to sample and SNP QC.

[2] http://www.well.ox.ac.uk/~wrayner/strand/HumanOmni2.5-8v1_A-b37-strand.zip
The .**strand** file contains six columns, SNP id, chromosome, position, %match to genome, strand and alleles. The SNP ids used are those from the annotation file and so are not necessarily the latest from dbSNP. The alleles listed are the Illumina TOP alleles. The **.miss** file gives the ids of the SNPs that did not reach the required threshold for mapping to the genome, the position and strand of the best match are given. The **.multiple** file contains SNPs that had more than 1 high quality match (>90%) to the genome, in this instance the better match is taken for the .strand file.

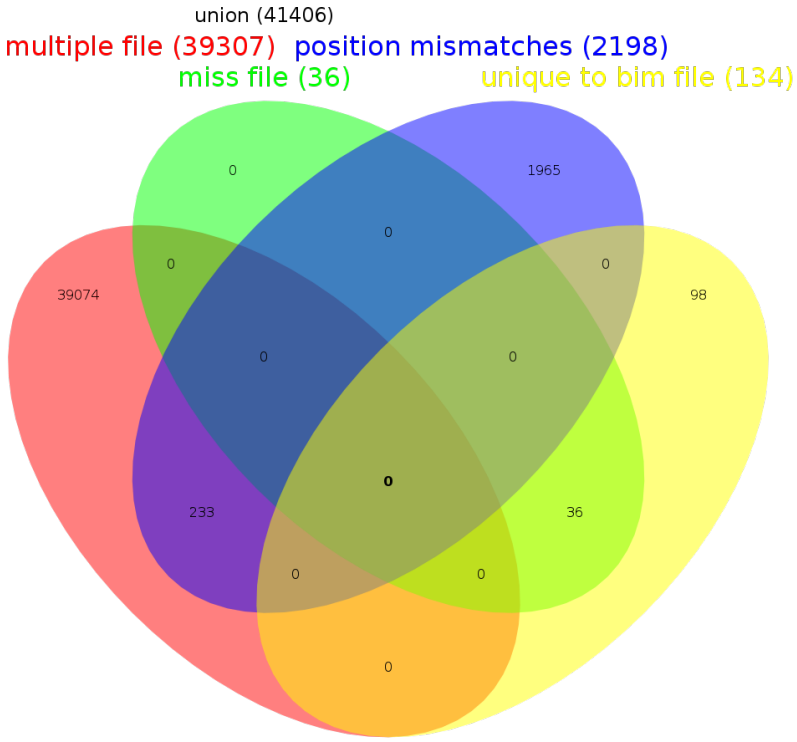


Figure 1 – Intersection between sets of autosomal SNPs excluded prior to the identification of duplicates.

3.3 Reporting markers to forward strand

- Flip strand of remaining SNPs based on strand file
  - 1,113,605 autosomal markers flipped to the forward strand

Alleles	All SNPs before flip	All SNPs after flip
AC	209,408	104,603
TG	0	104,791
AG	913,819	456,271
TC	0	457,556
CA	174,105	86,765
GT	0	87,354
GA	696,396	347,673
CT	0	348,715
AT	14,326	13,114
TA	11,973	13,185
CG	16,500	15,447
GC	14,443	15,496
0A	101,480	53,460
0T	5,190	53,210
0C	47,111	81,535
0G	116,024	81,600
00	267	267
Sum	2,321,042	2,321,042

Table 2 – Allele counts before and after strand flipping (i.e. A–>T, C–>G, G–>C, T–>A).

4 Sample QC

4.1 Sample call rate

- Calculate call rate of samples, and exclude samples with a call rate below the agreed threshold
- A threshold of 0.97 was chosen and samples with call rate below this value were excluded.
  - 64 samples were removed
  - Before and after sample removal the average sample call rate is 99.6% and 99.7%, respectively.

Sample call rate threshold	0.970	0.975	0.980	0.985	0.990	0.995	1.000
Samples below threshold	64	80	105	123	181	581	4,872

Table 3 – Cumulated frequency table of excluded samples at different call rate thresholds.

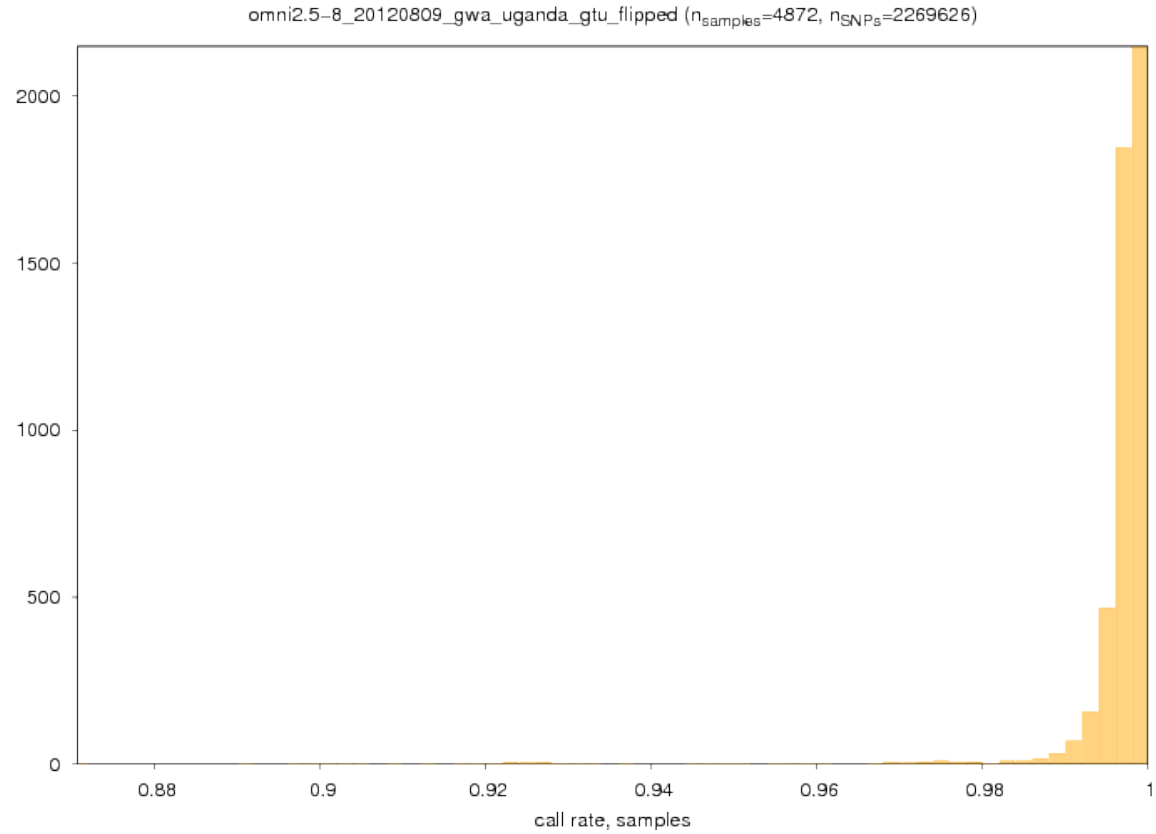


Figure 2 – Sample call rate histogram based on autosomal SNPs.

4.2 Heterozygosity

- Calculate heterozygosity in remaining samples[3]
- Exclusion threshold for heterozygosity is ±3SD from the mean
  - 0.209333±0.007416
  - 23 samples excluded

[3] Heterozygosity was not calculated in Plink but manually

Heterozygosity SD from mean	[∞,-3)	[-3,-2)	[-2,-1)	[-1,0)	[0,1)	[1,2)	[2,3)	[3,∞]	Sum
Samples within range	19	14	129	2,323	1,907	401	11	4	4,808

Table 4 – Sample counts in heterozygosity bins around the average heterozygosity.

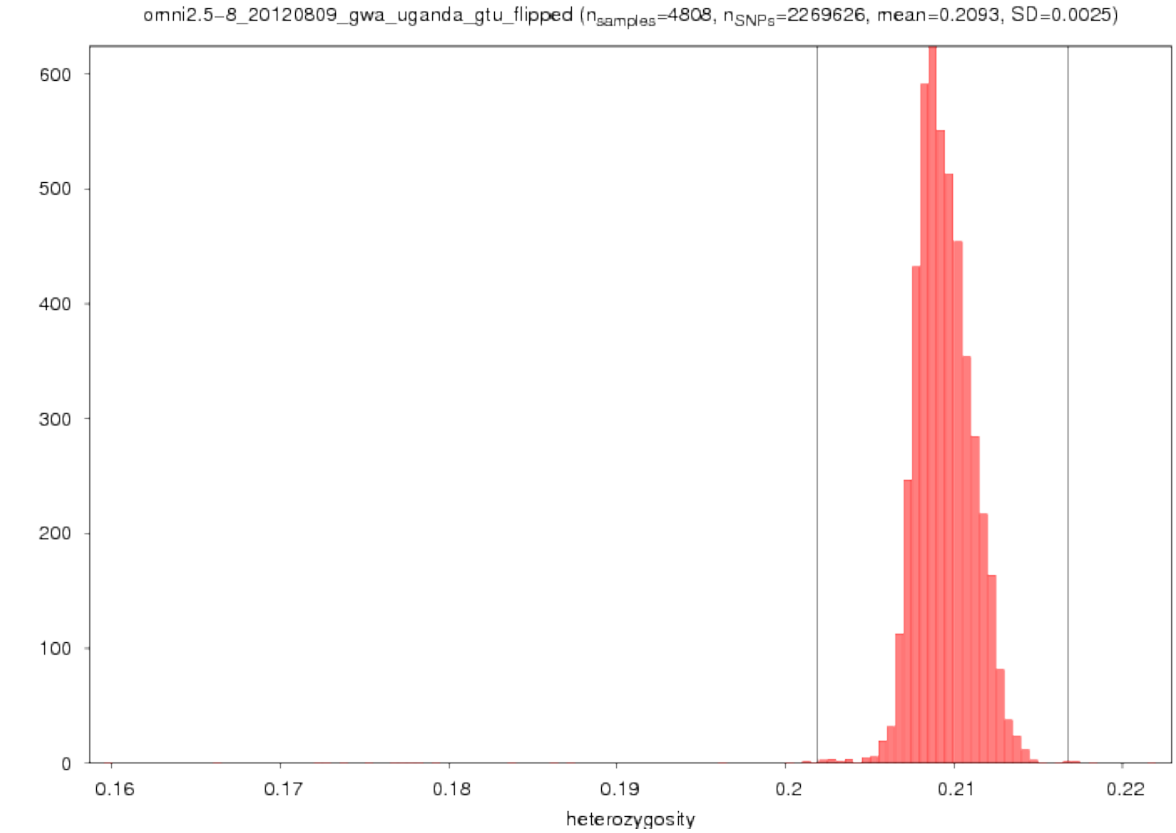


Figure 3 – Heterozygosity histogram based on autosomal SNPs.

4.3 Gender concordance

- Carry out gender checks on X chromosome (F>0.8 male, F <0.2 female)
  - 4 samples excluded

FID	Supplied gender	Inferred gender	F	comment
216767_H11_APP5211886	M	F	0.0567	unexplained[4]
216768_E11_APP5212097	M	0	0.7943	close to >0.8
232626_A02_APP5292757	M	F	0.1661	unexplained[3]
232628_A09_APP5292795	F	0	0.5810	unexplained[3]

Table 5 – Samples failing sex check. Thresholds are >0.8 for men and <0.2 for women.

4.4 Sample QC summary

- 64 samples that did not pass the sample call rate check were excluded
- 23 samples that did not pass heterozygosity check were excluded
- 4 samples that did not pass gender check were excluded

[4] All 23 SNPs genotyped on both Sequenom and Illumina platforms match for these samples.

- In total 91 samples were excluded during sample QC
- **4,781 samples retained after sample QC**

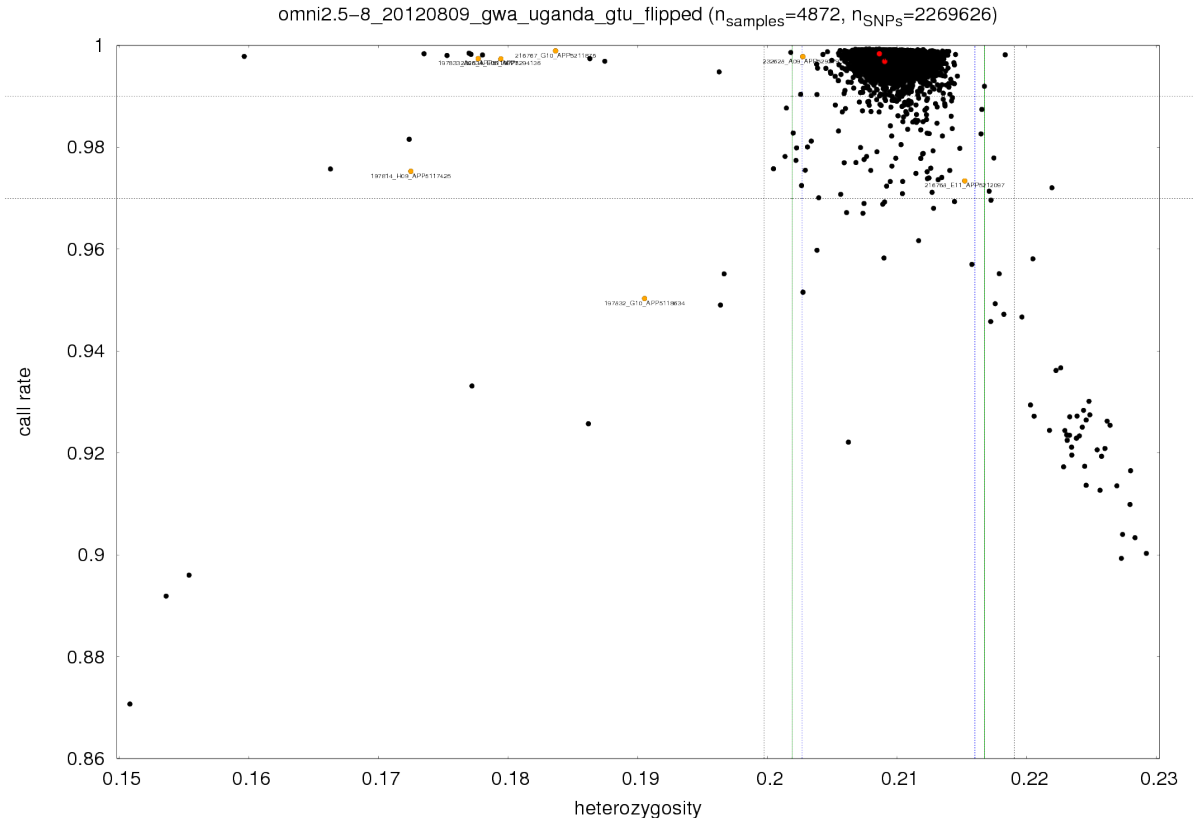


Figure 4 – Scatter plot of heterozygosity and call rate for all samples. Samples that failed gender check shown in red (gender opposite than supplied) and orange (gender not inferred). Horizontal lines are call rate thresholds. Vertical lines are 3SD from the mean het calculated after applying sample call rate thresholds of 0% (black), 97% (green) and 99% (blue).

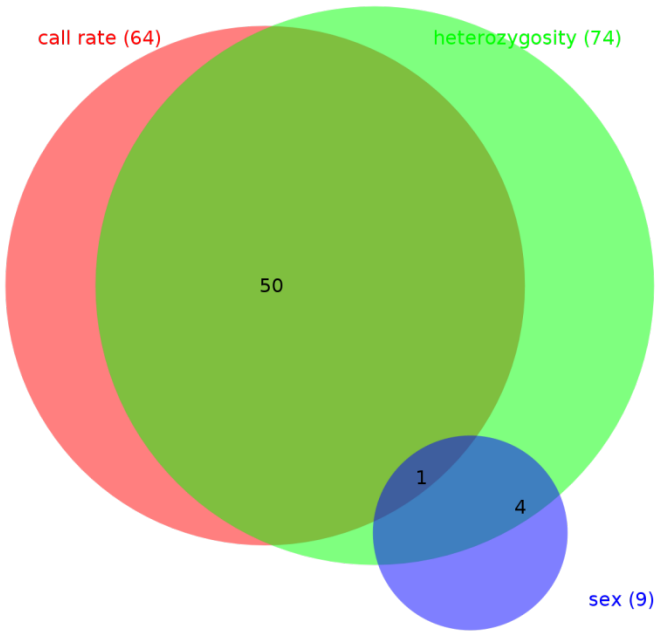


Figure 5 – Area proportional Venn diagram. Overlap is shown between all samples prior to sample QC that fail checks of call rate (red), heterozygosity (green) and sex (blue).

5 SNP QC

5.1 SNP call rate for autosomal markers

- Count SNP call rate, and mark SNPs for exclusion based on the agreed thresholds
- Exclude SNPs based on a call rate threshold of 97%
  - 25,037 autosomal SNPs are excluded
  - Before and after SNP removal the average SNP call rate is 99.62% and 99.70%, respectively.

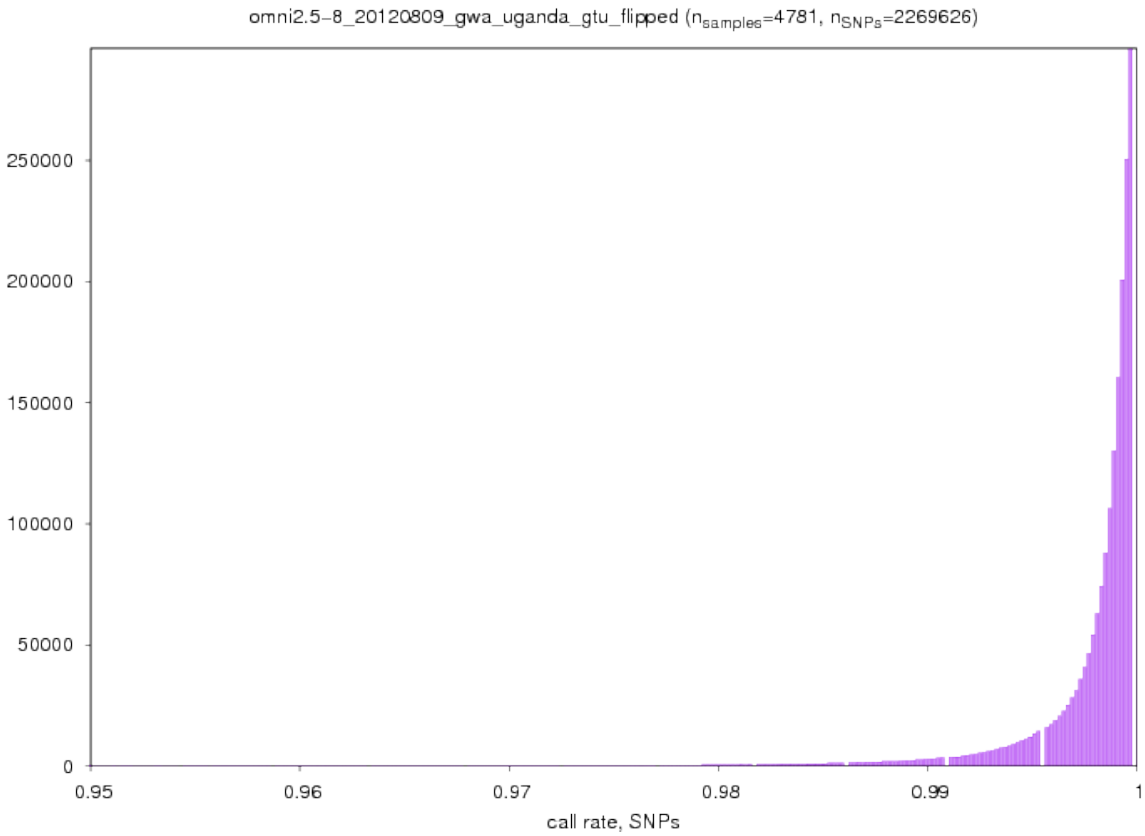


Figure 6 – Autosomal SNP call rate in the range 0.97 to 1.

F	0.950	0.955	0.960	0.965	0.970	0.975	0.980	0.985	0.990	0.995	1.000
n	11,334	13,385	16,095	19,770	25,037	33,128	45,636	67,961	114,496	266,534	2,269,626

Table 6 – Cumulated frequency table of autosomal SNPs falling below a given SNP call rate threshold.

5.2 Identity-by-descent (IBD)

- To calculate pairwise IBD in PLINK create a 'LD-pruned' dataset by including common (MAF>0.05) SNPs with r^2<0.2.
  - 323,831 SNPs included in the 'LD-pruned' dataset
- Check for identical or highly related individuals (i.e. IBD > 0.90) and for each pair exclude the sample with the lower sample call rate.
  - 3 samples excluded

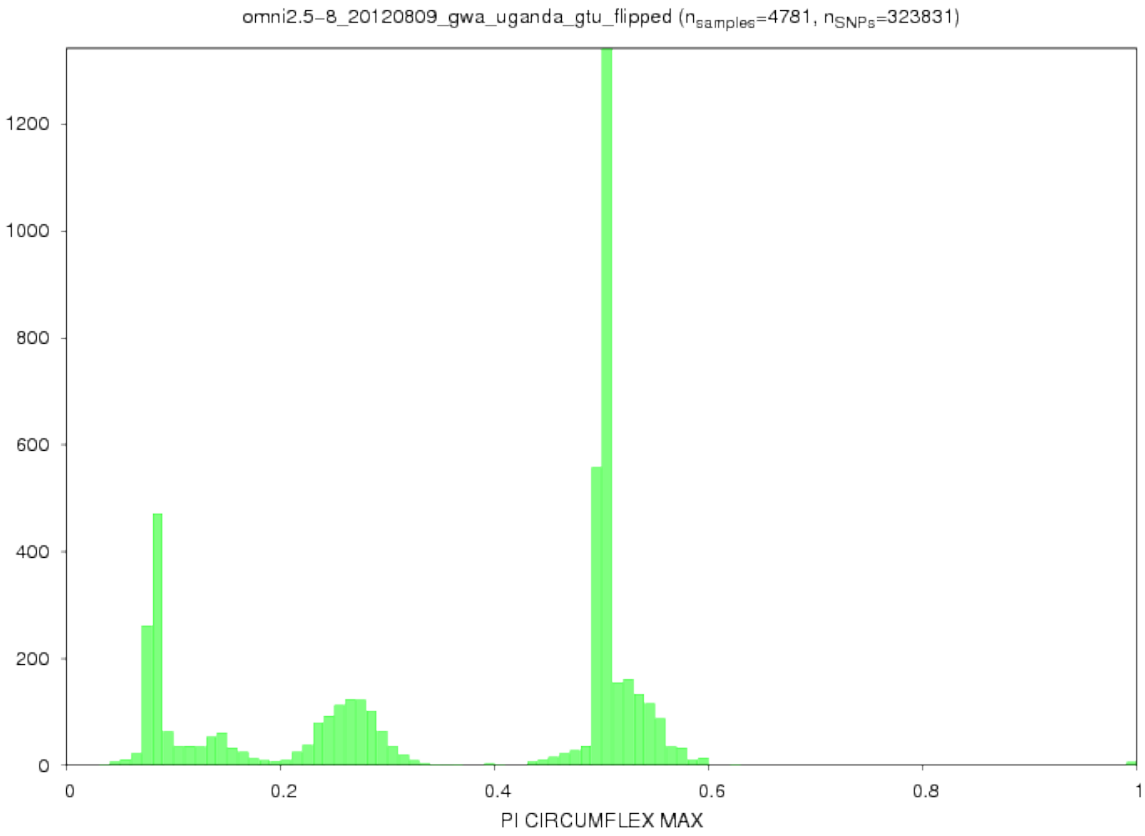


Figure 7 – Distribution of the maximum pairwise $\hat{\pi}$ for each sample.

ID1	ID2	PI_HAT	comment
APP5118823	APP5294270	0.9991	different house, different dob/age
APP5118028	APP5212152	0.9989	different house, dob approximated 1 year difference
APP5119248	APP5119212	0.9987	same house, same age/dob, twins?

Table 7 – Sample pairs with $\hat{\pi}$ above 0.90. From each sample pair the sample with the least associated phenotypes or alternatively lower sample call rate is removed.

5.3 Hardy-Weinberg equilibrium (HWE) for autosomal markers

- Use the pairwise IBD matrix generated at the previous step and extract the maximum number of unrelated individuals (founders) from main dataset (Table 8)
- Carry out HWE check only on founders (i.e. IBD <0.05) and exclude SNPs below a threshold of $p_{HWE}<10^{-8}$
- Exclude SNPs from main dataset
    - 14,331 autosomal SNPs excluded

$\hat{\pi}$	0.05	0.10	0.15	0.20	0.50	0.90	1.00
n	1,899	2,202	2,541	2,652	3,800	4,778	4,781

Table 8 – Maximum number of samples all falling below a given $\hat{\pi}$ threshold. The samples above a 0.05 threshold are excluded prior to the HWE check.

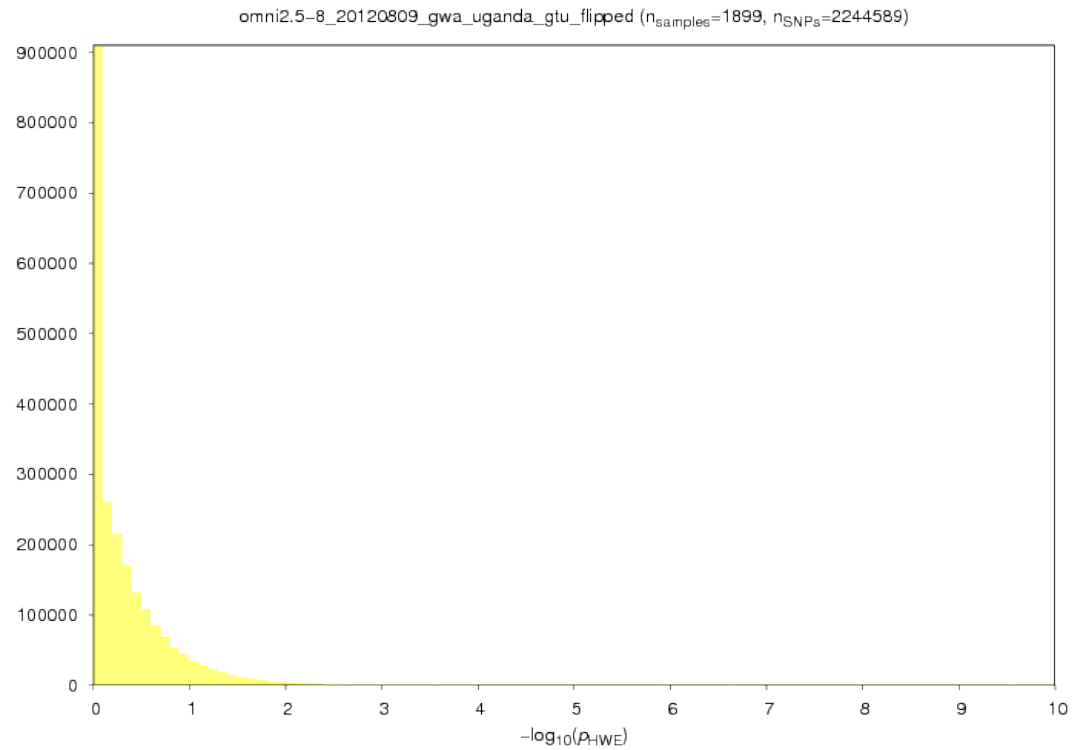


Figure 8 – Distribution of $-\log_{10}(p_{HWE})$ in the range 0-10 for autosomal SNPs in the founder set.

p_{max}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
n	20,649	18,560	17,002	15,075	14,331

Table 9 - Cumulated frequency table of autosomal SNPs falling below a given p_{HWE} threshold.

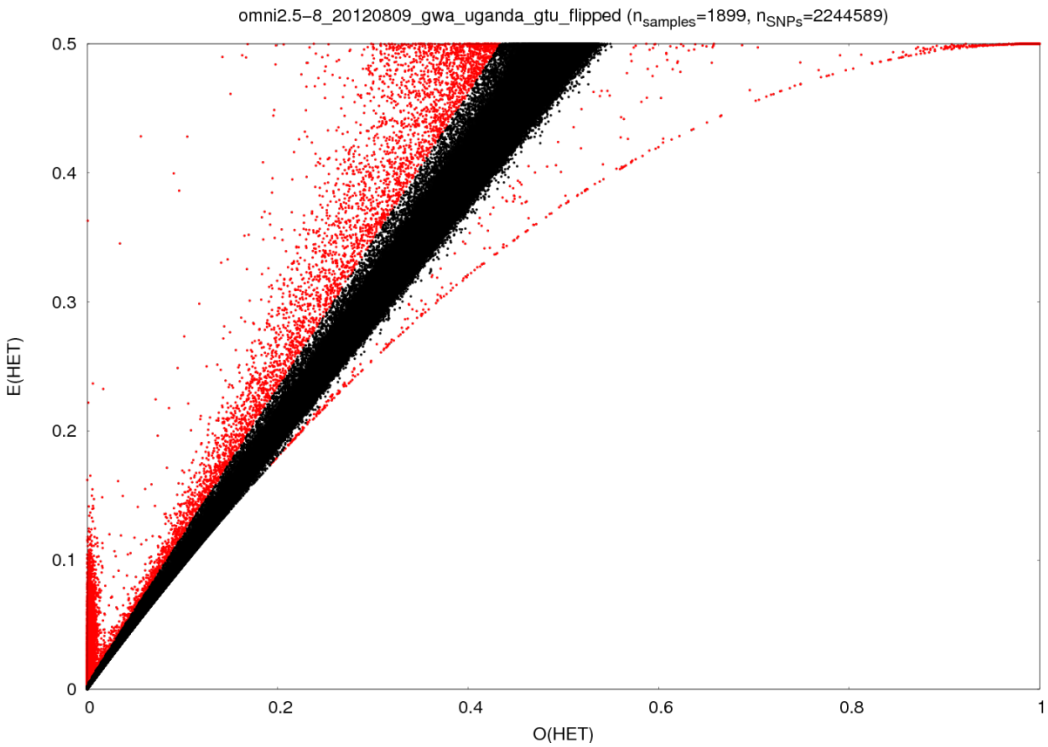


Figure 9 –Observed heterozygosity (x-axis) and expected heterozygosity (y-axis) for founders and autosomal SNPs (SNP call rate greater than 97%). SNPs in HWE are shown in black and all other SNPs are shown in red.

5.4 Principal components analysis (PCA)

- PCA plots are generated using EIGENSOFT.(Patterson, Price et al. 2006; Price, Patterson et al. 2006)
- After the removal of samples that fail the IBD check and exclusion of SNPs that fail the HWE check, the MAFs are recalculated and a new LD-pruned dataset with 316,819 SNPs and the 1,899 founders from previously is generated.
    - Exclude also regions of long range LD (i.e. 66,633 SNPs), as these can lead to artificial associations in association analyses[5]
- Clear ancestry outliers along PCs 1 and 2 are removed.
    - Based on the PCA plots (Figures 12 and 13) no additional samples were removed.

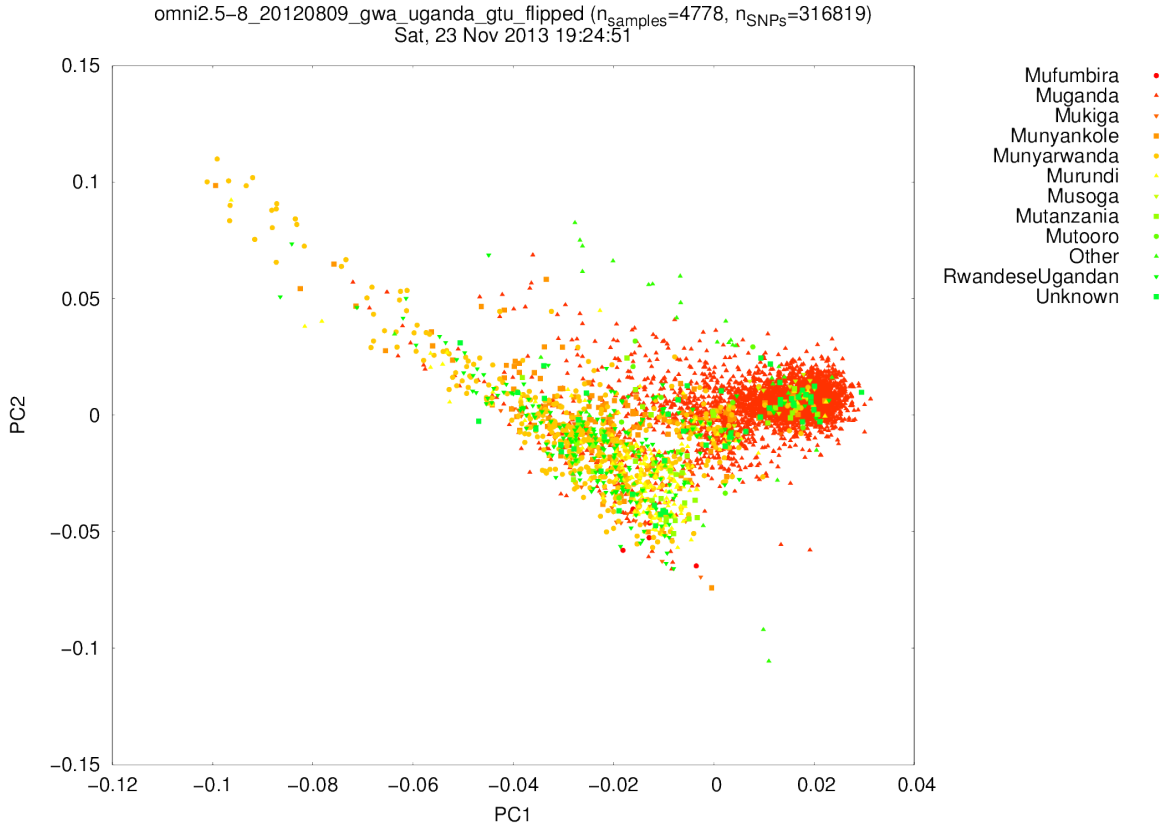


Figure 12 – Plot of PCs 1 and 2 within the Ugandan population.

[5] Price et al. (2008) Long-Range LD Can Confound Genome Scans in Admixed Populations. Am J Hum Genet. 83(1): 132–135.

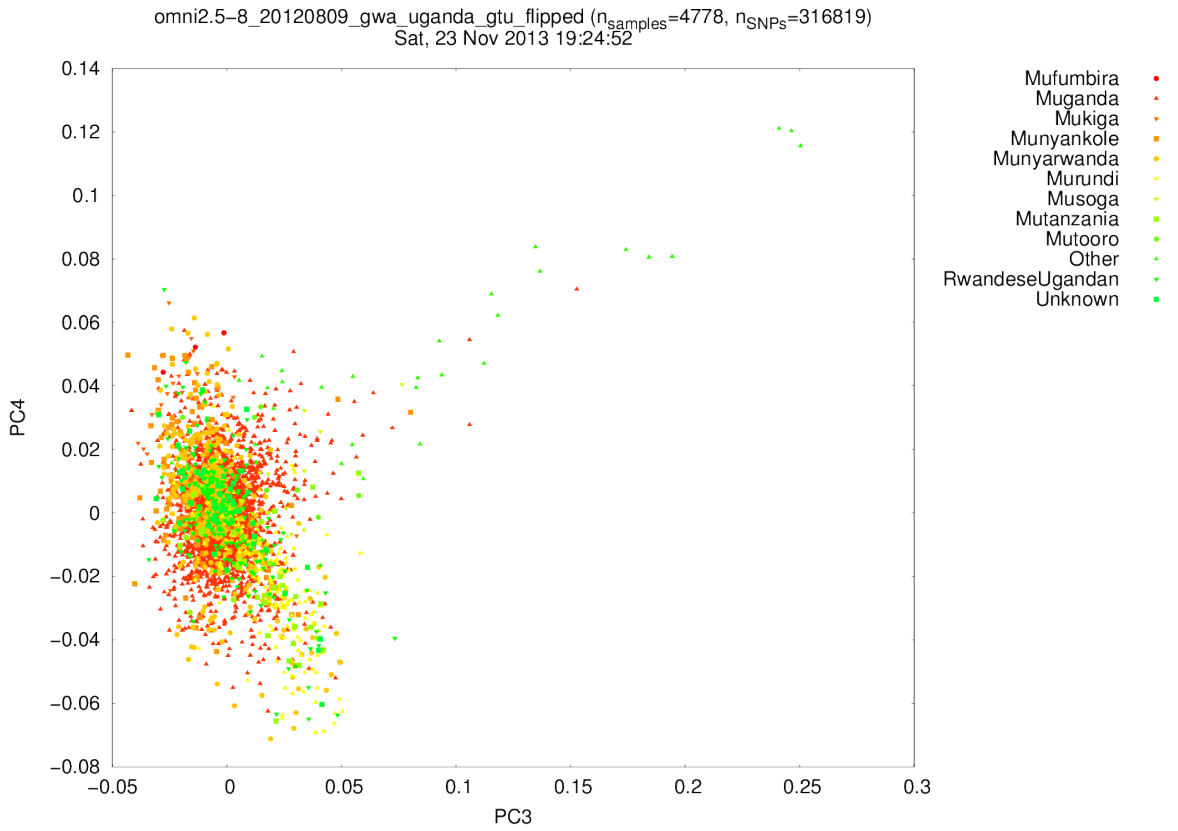
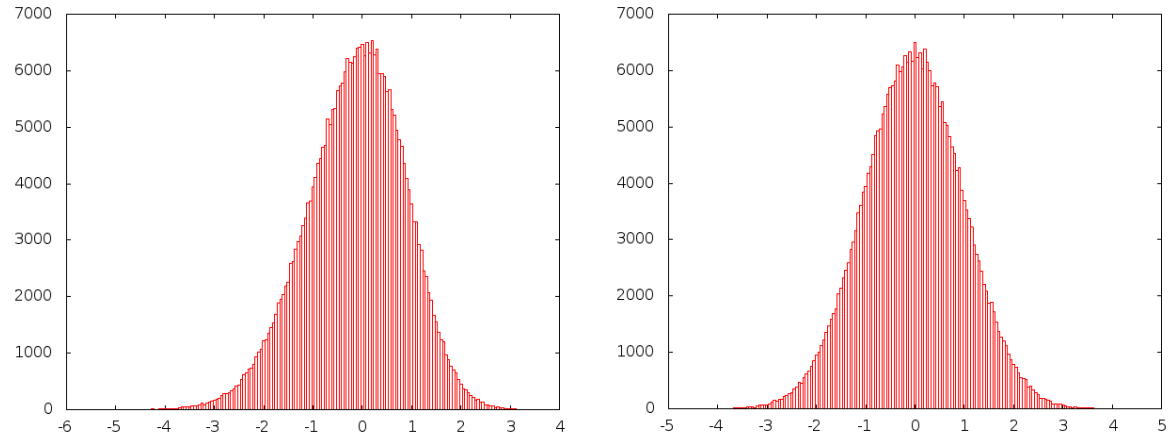


Figure 12 – Plot of PCs 3 and 4 within the Ugandan population.



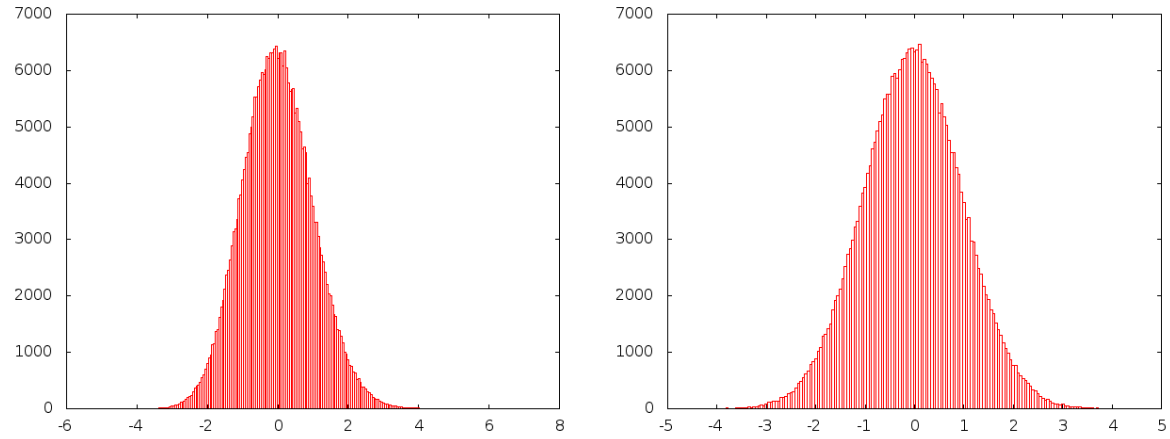


Figure 10 – Distribution of snpweights for PCs 1-4, which is normal, and shows that local LD effects have been removed during LD pruning prior to PC analysis.

5.5 SNP call rate for X-chromosome markers

- Exclude SNPs based on a call rate threshold of 97%
  - 1,268 X chromosome SNPs identified from 2,061 males
  - 5,578 X chromosome SNPs identified from 2,717 females
  - 6,392 X chromosome SNPs in the union set excluded from both sexes

sex	*F*	0.950	0.955	0.960	0.965	0.970	0.975	0.980	0.985	0.990	0.995	1.000
m	*n*	941	985	1,060	1,125	1,268	1,452	1,741	2,360	3,655	8,529	50,130
f	*n*	4,893	5,030	5,186	5,376	5,578	5,844	6,186	6,725	7,638	10,473	50,130

Table 10 – Cumulated frequency table of X chromosome SNPs falling below a given SNP call rate threshold.

5.6 Hardy-Weinberg equilibrium (HWE) for X-chromosome markers

- Carry out HWE check on X-chromosome SNPs for only female founders (i.e. IBD<0.05, 1,225 females) and exclude SNPs below a threshold of $p_{HWE}<10^{-8}$
- Exclude SNPs from main dataset
  - 560 X chromosome SNPs excluded

p_{max}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}
n	733	681	636	588	560

Table 11 – Cumulated frequency table of X chromosome SNPs falling below a given p_{HWE} threshold in females.

5.7 SNP QC summary

- 3 additional samples excluded based on high relatedness (i.e. high IBD)
- 39,368 autosomal SNPs excluded (call rate and HWE)
- 6,952 X chromosome SNPs excluded (call rate in males and females and HWE in females)
- **4,778 samples retained**
- **2,230,258 autosomal SNPs retained**

- **43,178 X chromosome SNPs retained**

6 Summary

	Released	Excluded			Post-QC
		Pre-QC	Sample QC	SNP QC	
Samples	4,892	20	91	3	4,778
SNPs					
Autosomes	2,314,174	44,548	0	39,368	2,230,258
X chromosome	55,208	5,078	0	6,952	43,178

Table 12 – Summary of QC

7 References

Patterson, N., A. L. Price, et al. (2006). "Population Structure and Eigenanalysis." PLoS Genet 2(12): e190.

Price, A. L., N. J. Patterson, et al. (2006). "Principal components analysis corrects for stratification in genome-wide association studies." Nat Genet 38(8): 904-909.