Sample analysis of genotyping data.

Prepared by Zhe Zhang

October 27, 2014

Contents

1 Data summary

Sample cohort name: **gVCF6703**.

1.1 Genotype per sample

Table 1: Samples with the lowest and the highest call rates.

ID	A.A	A.B	B.B	call_rate
1-00722-01	115706.00	1489.00	794.00	36.91
11029.fa	181652.00	2651.00	1527.00	58.13
11691.fa	183477.00	2468.00	1619.00	58.68
1-06162-02	200745.00	2783.00	1567.00	64.16
1-06101-01	222733.00	3043.00	1688.00	71.16
SSC06977	312349.00	4807.00	2489.00	99.99
SSC07145	312540.00	4676.00	2430.00	99.99
1-03354-02	312711.00	4415.00	2521.00	99.99
SSC07151	312692.00	4310.00	2645.00	99.99
1-01256-02	312821.00	4167.00	2660.00	99.99

Table 2: Samples with the lowest and highest percents of heterozygous calls.

ID	A.A	A.B	B.B	het_rate
1-05243-01	309302.00	2861.00	3277.00	0.91
1-00894-02	312842.00	3414.00	3104.00	1.07
GT04012012-01	312777.00	3445.00	3123.00	1.08
1-04943	312782.00	3458.00	3104.00	1.08
1-00894	312586.00	3465.00	3065.00	1.09
13559.fa	310629.00	6305.00	2534.00	1.97
SSC06253	309905.00	6300.00	2532.00	1.98
1-00384-02	310588.00	6319.00	2528.00	1.98
14342.mo	310707.00	6321.00	2434.00	1.98
1-00018-02	309184.00	6529.00	1453.00	2.06

1.2 Pruned genotype calls

Table 3: Frequency of genotypes after SNP pruning.

Genotype	Percent
A/A	98.5622
A/B	0.5523
B/B	0.2862
Others	0.0000
Total	99.4007

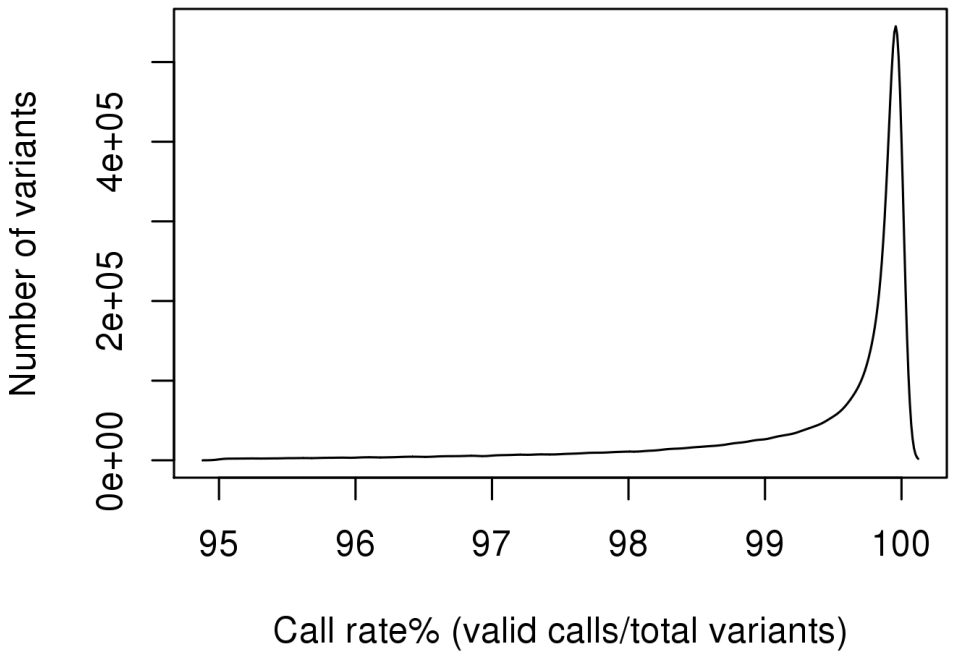

Fig 1 Distribution of call rates of individual variants after SNP pruning.

2 PCA (Principal Components Analysis)

PCA calculates the genetic covariance matrix from genotypes, computes the correlation coefficients between sample loadings and geno- types for each SNP, calculates SNP eigenvectors (loadings), and estimates the sample loadings of a new dataset from specified SNP eigenvectors.

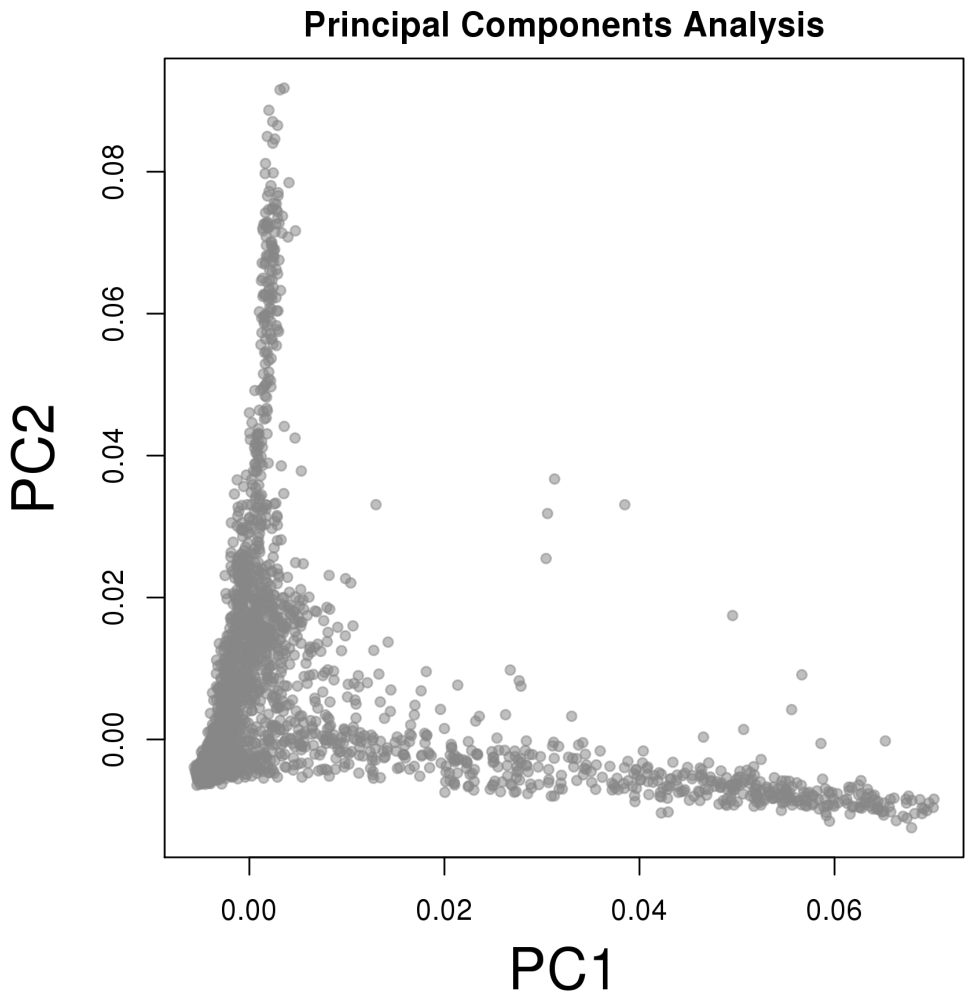

Fig 2 Categorization of samples by principal components.

Table 4: PC-SNP pairs (top 20 overall, top 5 each PC)

PC	SNP	Corr	Chromosome	Location
PC1	rs710079	0.67	16	129223
PC1	rs386607415	0.66	16	129277
PC1	rs771205	-0.64	1	150975108
PC1	rs10796961	-0.63	1	156556321
PC1	rs331537	0.62	11	4471276
PC2	rs3857809	0.60	7	100416139
PC2	rs2285044	0.58	3	50336661
PC2	rs2071203	0.57	3	50311900
PC2	rs3738591	0.53	1	155764808
PC2	rs2303893	0.53	2	26507076
PC4	rs2288518	-0.38	19	55710021
PC4	rs4684677	-0.38	3	10328453
PC4	rs2288420	-0.37	19	55693123
PC4	rs2071572	-0.37	19	55686230
PC4	rs312470	0.36	17	6902179
PC5	rs10186233	0.23	2	97877399

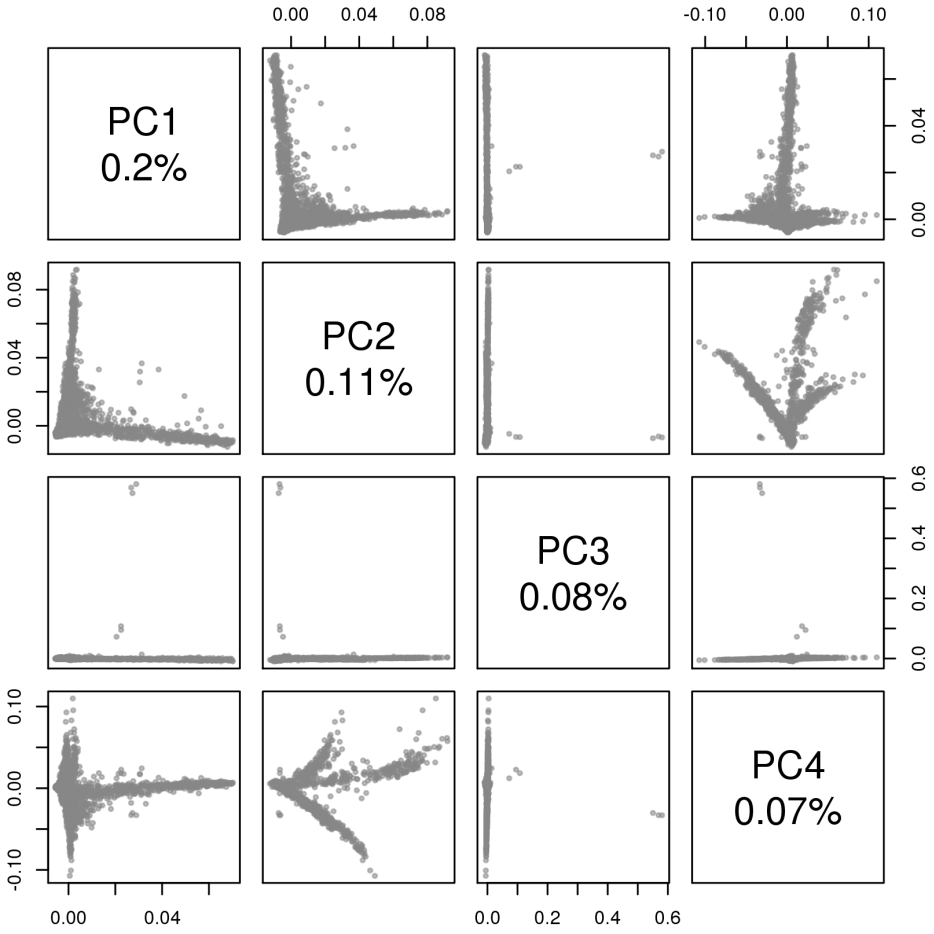

Fig 3 Pairs of top four principal components

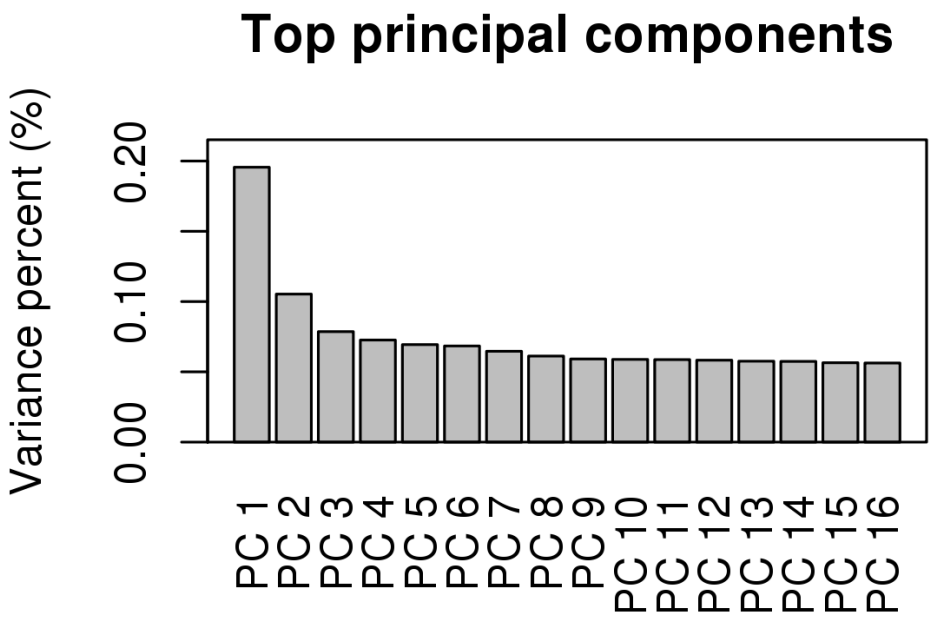

Fig 4 Percentage of variance accounted for by top PCs

3 IBD (Identity By Descent)

For relatedness analysis, identity-by-descent (IBD) can be done by either the method of moments (MoM) (Purcell et al., 2007) or maximum likelihood estimation (MLE) (Milligan, 2003; Choi et al., 2009). Although MLE estimates are more reliable than MoM, MLE is significantly more computationally intensive. For both of these methods it is preffered to use a LD pruned SNP set.' This report used MLE.

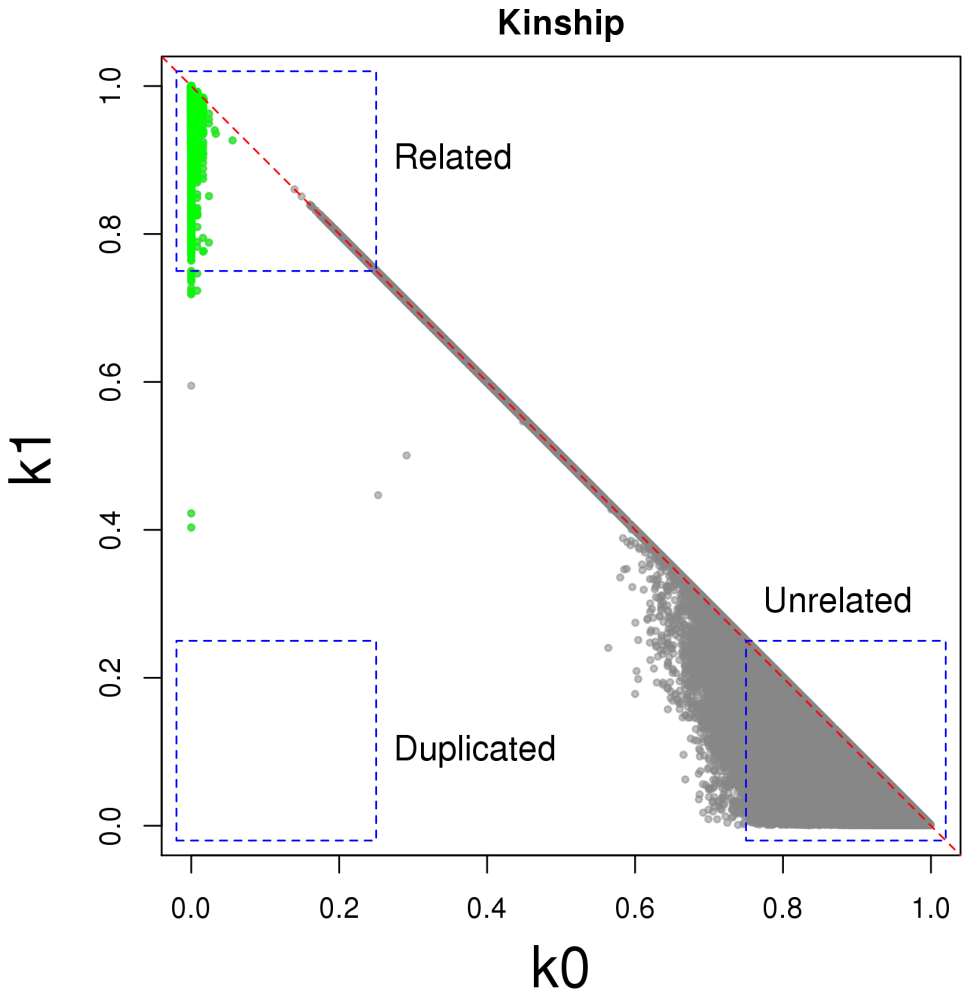

Fig 5 Kinship of sample pairs based on identity-by-descent.

4 IBS (Identity By State)

For the n individuals in a sample, IBS creates a n by m matrix of genome-wide average IBS pairwise identities, performs multidimensional scaling (MDS) analysis on the n × n matrix of pairwise distances, perform cluster analysis, and determine the groups by a permutation score.'

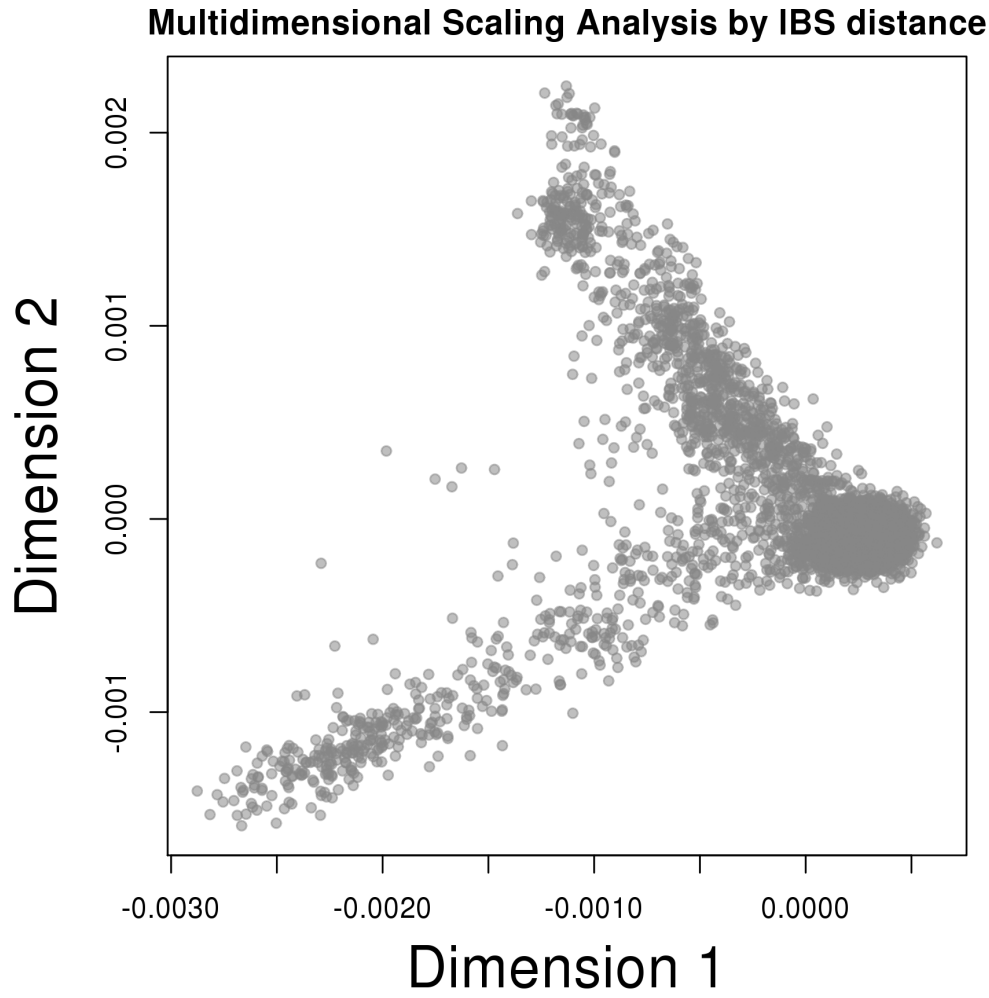

Fig 6 Sample categorization based on identity-by-state

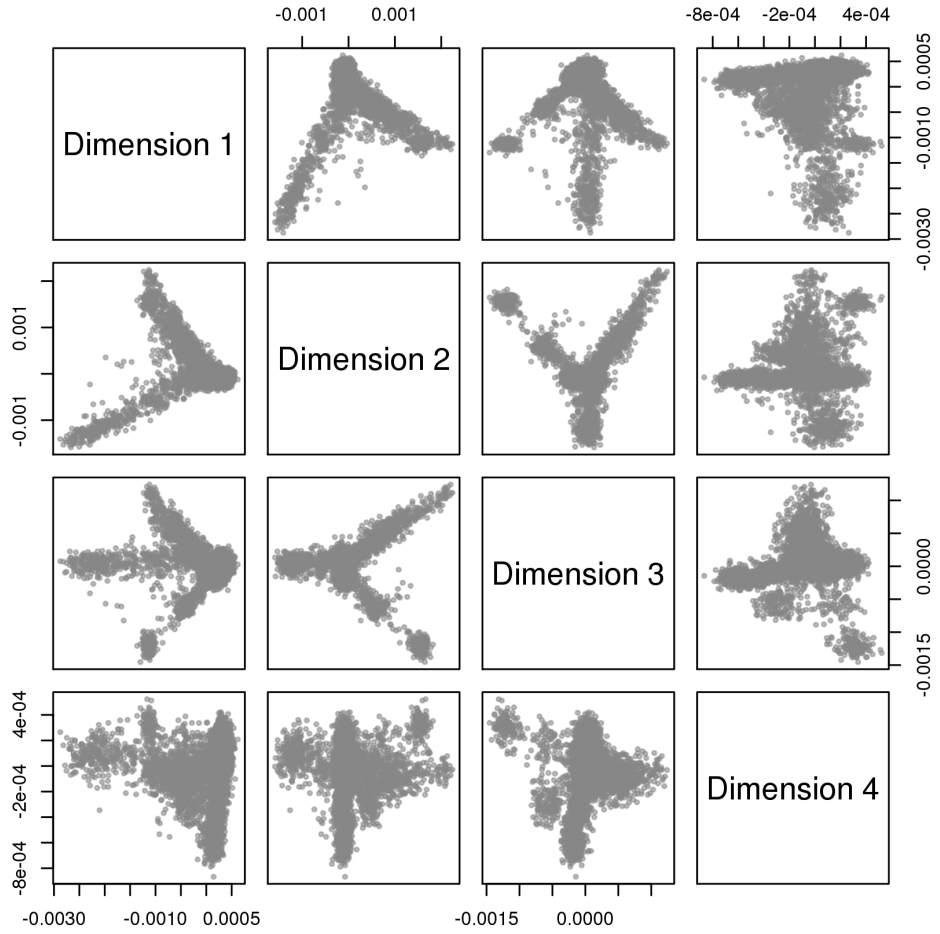

Fig 7 Pairs of top four dimensions.

5 Gender analysis

When there are enough variants and genotype calls from both X and Y chromosomes, samples' gender can be identified based on their percent of heterozygous calls from X and number of valid calls from Y. When the samples' gender was previously known, it can be compared to the data-based gender classification to identify potentially mislabelled samples.

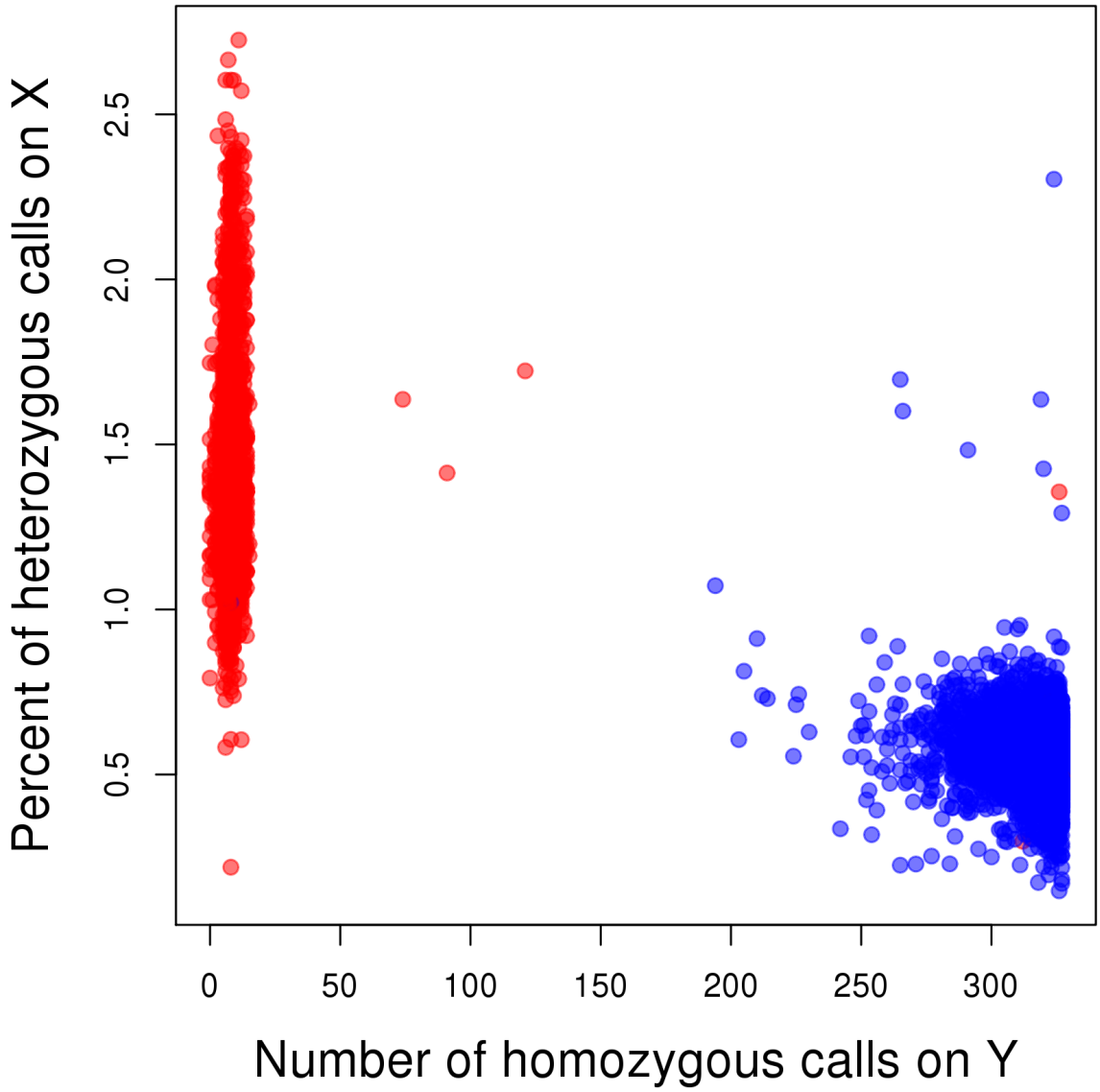

Fig 8 Gender prediction based on X and Y chromosome calls.

6 Summary

6.1 Summary statistics

* There are totally **6703** samples.
* There are totally **328254** variants from inputs.
* **36.91** to **99.99** percents of variants had valid calls per sample (mean = **99.12**).
* **0.907** to **2.059** percents of genotypes calls are heterozygous per sample (mean = **1.435**).
* **166630** autosomal variants were selected by SNP pruning for sample analysis.
* The top 2 principal components account for **0.3** percents of total variance.
* The IBD analysis identified **4** pairs of potentially duplicated samples.
* The IBD analysis identified **17889** pairs of potentially related samples.
* The IBS analysis classified samples into **121** groups.

6.2 Alerts.

Previous analyses generated the following alerts that might suggest issues such as low quality samples, sample mislabeling, unknown kinship, duplicated samples, and so on.

1. The first 2 principal components account for less than 5 percents of total variance.
2. Range of sample call rates is too big (max/min > 1.5)
3. Range of heterozygous percents is too big (max/min > 1.5)
4. The reported gender of 15 samples was not agreed by variants on X and Y chromosomes.
5. IBD analysis identified 23396 sample pairs with unreported kinship.
6. IBD analysis rejected the reported kinship of 2 sample pairs.
7. IBD analysis identified 4 unreported pair(s) of duplicated samples.

Table 5: Sample clustering by IBS analysis.

Group_ID	Num_Samples
G001	6
G002	388
G003	216
G004	132
G005	5421
G006	210
Outlier001	3
Outlier002	1
Outlier003	2
Outlier004	3
Outlier005	3
Outlier006	3
Outlier007	3
Outlier008	3
Outlier009	3
Outlier010	2
Outlier011	2
Outlier012	3
Outlier013	3
Outlier014	3
Outlier015	3
Outlier016	3
Outlier017	3
Outlier018	3
Outlier019	3
Outlier020	1
Outlier021	3
Outlier022	3
Outlier023	3
Outlier024	2
Outlier025	3
Outlier026	3
Outlier027	2
Outlier028	3
Outlier029	3
Outlier030	1
Outlier031	3
Outlier032	3
Outlier033	1
Outlier034	3
Outlier035	3
Outlier036	3
Outlier037	3
Outlier038	3
Outlier039	3
Outlier040	3
Outlier041	3
Outlier042	3
Outlier043	3
Outlier044	3
Outlier045	3
Outlier046	3

Table 6: Predicted vs. reported gender

	F	M
F	3200	7
M	8	3486

Table 7: Reported gender not supported by X/Y variants.

ID	Xhet_Ratio	Y_Count	Decision	Predicted	Reported
1-00424-01	0.01	318	1.33	M	F
1-00424-02	0.02	6	-1.24	F	M
1-00722	0.01	9	-1.17	F	M
1-01130-01	0.00	312	1.17	M	F
1-01130-02	0.01	6	-1.18	F	M
1-05520-01	0.01	322	1.34	M	F
1-05520-02	0.01	8	-1.21	F	M
11104.s1	0.01	326	1.23	M	F
11347.fa	0.01	9	-1.24	F	M
11347.mo	0.01	326	1.29	M	F
11372.fa	0.01	12	-1.12	F	M
11372.mo	0.00	326	1.28	M	F
GT04006072-01	0.00	327	1.28	M	F
GT04006072-02	0.01	8	-1.06	F	M
GT04006123	0.01	319	1.33	M	F

Table 8: Unreported kinship identified by IBD. (20 of 23396)

ID1	ID2	k0	k1	kinship
11352.fa	11425.fa	0.0000	0.8852	0.2787
1-03347-02	SSC06670	0.0000	0.9317	0.2671
1-00924-02	1-01130-02	0.0000	0.9418	0.2646
1-03535-02	GT04006123	0.0000	0.9711	0.2572
1-03347	SSC06669	0.0000	0.9714	0.2572
1-03535-01	GT04006123	0.0000	0.9734	0.2567
1-03347	SSC06661	0.0000	0.9741	0.2565
1-03535	GT04006123-02	0.0000	0.9853	0.2537
1-03535	GT04006123-01	0.0000	0.9914	0.2521
1-03347-01	SSC06670	0.0000	0.9931	0.2517
1-00018	1-01525	0.0000	1.0000	0.2500
1-00018-01	1-00018-02	0.0000	1.0000	0.2500
1-00018-01	1-01525	0.0000	1.0000	0.2500
1-00018-02	1-00025-01	0.0000	1.0000	0.2500
1-00018-02	1-00025-02	0.0000	1.0000	0.2500
1-00018-02	1-00034	0.0000	1.0000	0.2500
1-00018-02	1-00034-01	0.0000	1.0000	0.2500
1-00018-02	1-00047-02	0.0000	1.0000	0.2500
1-00018-02	1-00051-02	0.0000	1.0000	0.2500
1-00018-02	1-00064-02	0.0000	1.0000	0.2500

Table 9: Reported kinship rejected by IBD.

ID1	ID2	k0	k1	kinship
1-06058	1-06058-02	0	0.4032	0.3992
GT04014641	GT04014641-02	0	0.4223	0.3944

Table 10: Unreported pairs of duplicated samples.

ID1	ID2	k0	k1	kinship
1-03347-01	SSC06661	0	0	0.5
1-03347-02	SSC06669	0	0	0.5
1-03535-01	GT04006123-01	0	0	0.5
1-03535-02	GT04006123-02	0	0	0.5