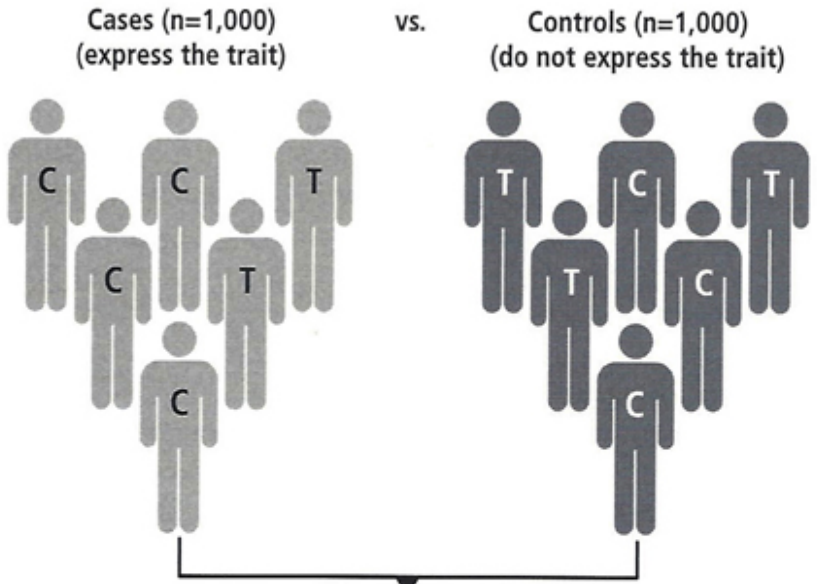
Quality control is essential to ensuring reproducibility in genotype and non-genetic data

Molly Hall, PhD
Assistant Professor
Veterinary and Biomedical Sciences

June 18, 2018

Genetic Association

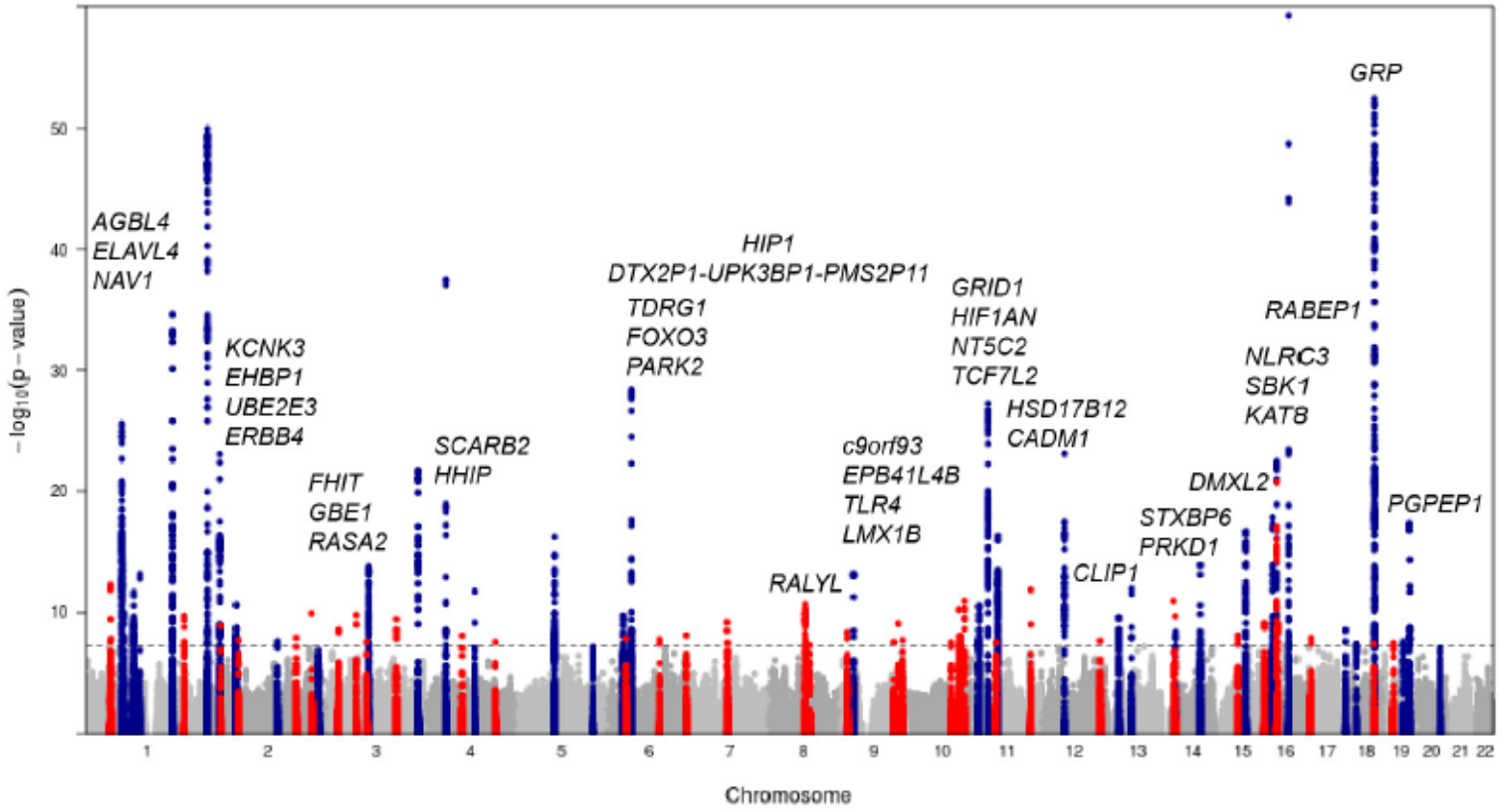


	C	T
Cases	62%	38%
Controls	49%	51%

$\chi^2 = 34.2$, $p\text{-}value = 4.9\text{x}10^{-9}$

http://www.discoveryandinnovation.com/

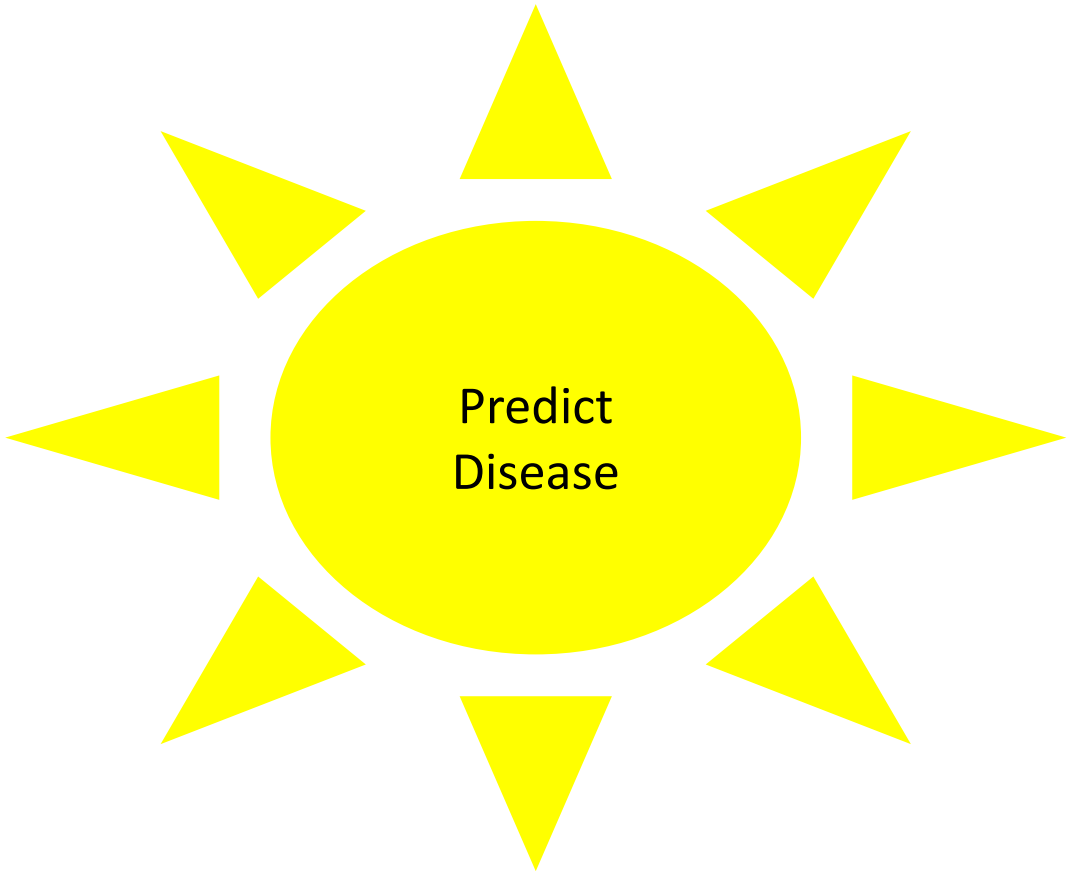
GWAS of Body Mass Index (BMI)

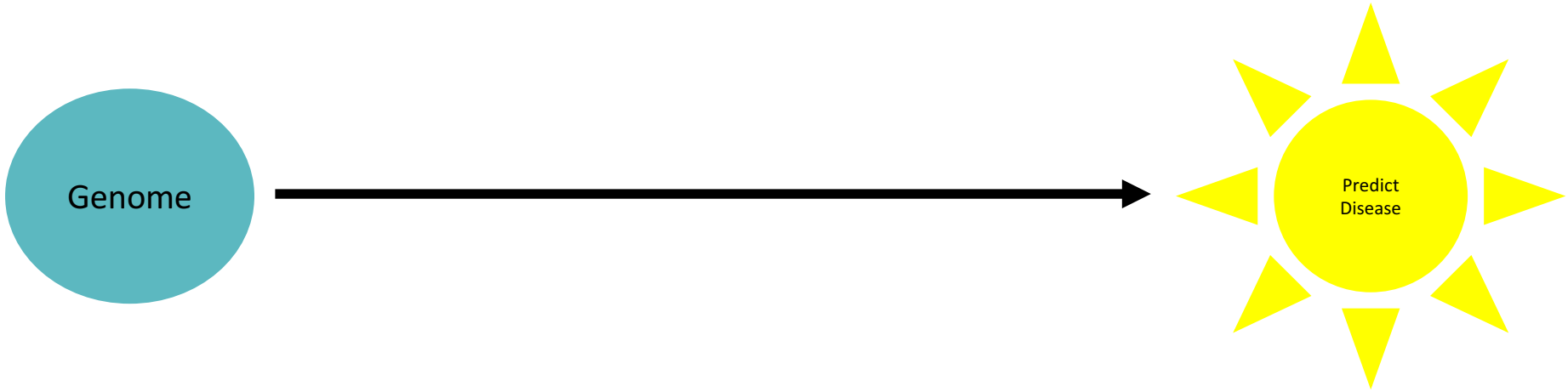


AE Locke *et al. Nature* **518**, 197-206 (2015) doi:10.1038/nature14177

nature

Biomedical Informatics




Genome
Predict
Disease

Why do we perform quality control (QC) methods?

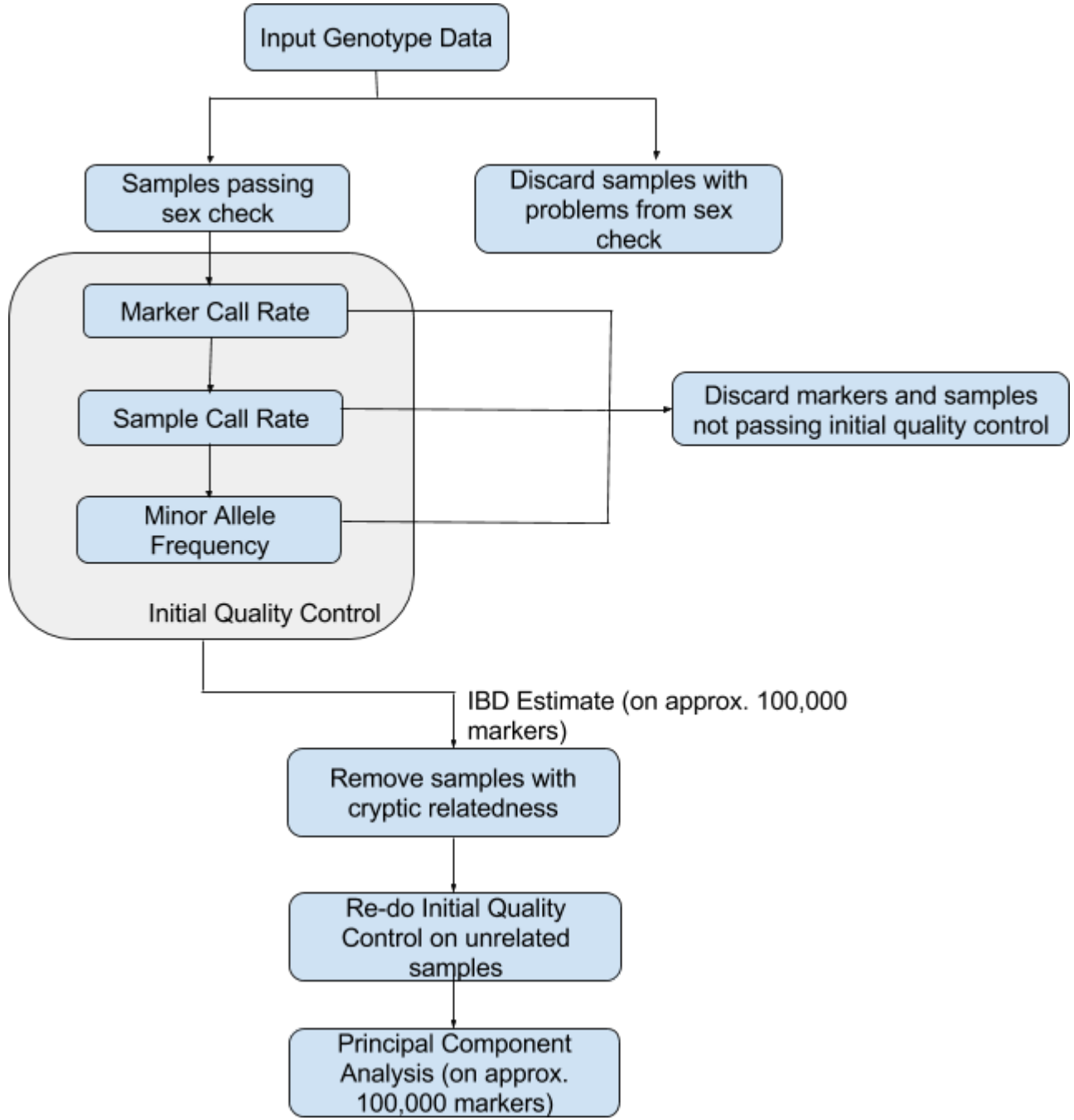
- We want to make sure signals we find in GWAS (or other analyses) are not due to:
  - Poor quality samples (sample call rate)
  - Poor quality markers (marker call rate)
  - Relatedness in your sample set (Identity by decent (IBD))
  - Differences due to population structure (principal component analysis (PCA))

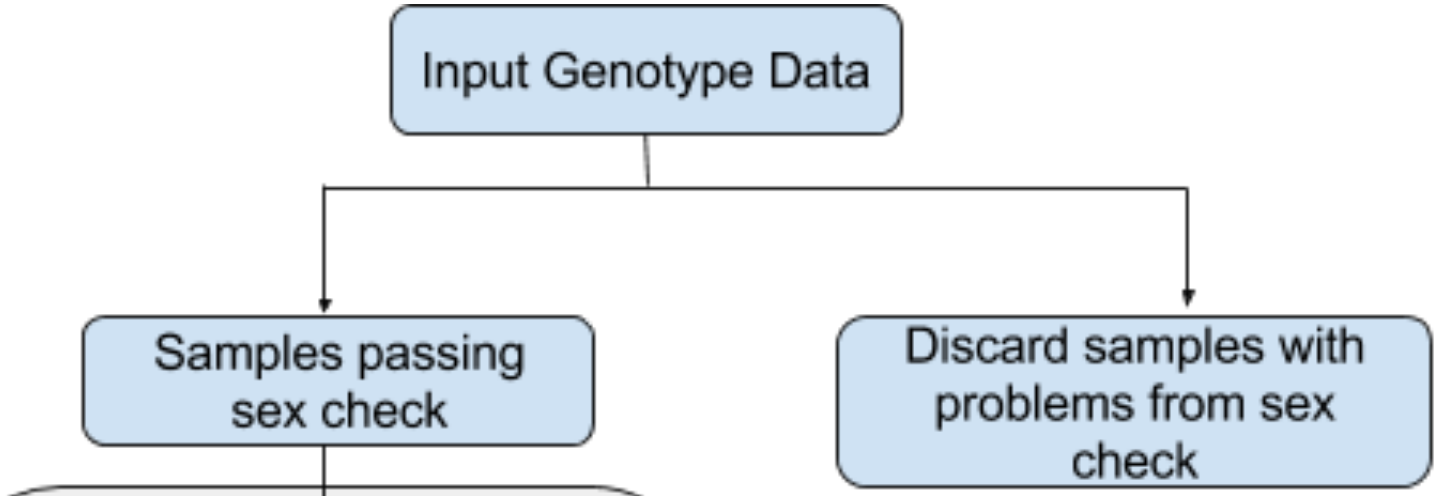
- We want to reduce the number of tests we need to adjust for:
  - Minor allele frequency (MAF)
  - Linkage disequilibrium (LD)

PLINK Files

- *.fam files contain information about samples, one sample per line
- *.bim files contain information about markers, one marker per line
- *.bed files contain binary genotype information. You should not be viewing this file directly.

- Here is an example command to read these files in PLINK:
  plink --bfile myfile


Input Genotype Data
Samples passing sex check
Discard samples with problems from sex check
Marker Call Rate
Sample Call Rate
Minor Allele Frequency
Initial Quality Control
Discard markers and samples not passing initial quality control
IBD Estimate (on approx. 100,000 markers)
Remove samples with cryptic relatedness
Re-do Initial Quality Control on unrelated samples
Principal Component Analysis (on approx. 100,000 markers)


Input Genotype Data
Samples passing sex check
Discard samples with problems from sex check

```
Location
Datasets: /gpfs/group1/m/mdr23/projects/eMERGE/Marshfield/PLATO_Paper_an
```

QC:

Pre-QC:

Samples: 3,896
Markers: 561,490

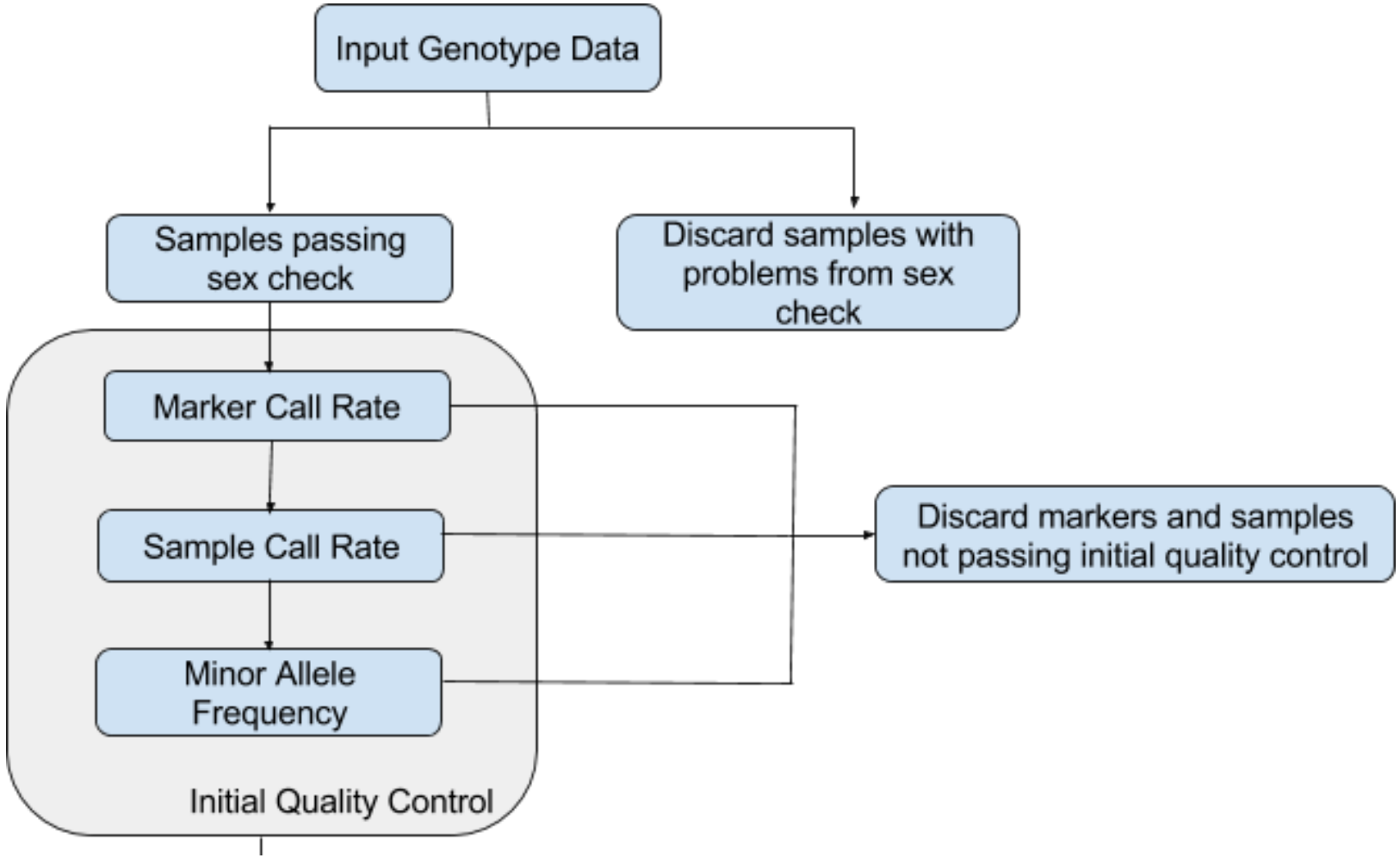
SEX CHECK:

```
plink --bfile T2D --check-sex
```

```
awk '{if ($5=="PROBLEM")print}' plink.sexcheck
```

16230834 111584@1018348317 2 0 PROBLEM 0.4688
16228083 119785@1018342676 2 0 PROBLEM 0.2364
16222319 137237@1018348183 2 0 PROBLEM 0.7259
16231930 108172@1018342658 2 0 PROBLEM 0.4507
16228204 119481@1018298703 2 1 PROBLEM 1
16233113 104569@1018301292 1 0 PROBLEM 0.4823
16214881 159853@1018299011 1 2 PROBLEM 0.1067

```
Dropped 2 samples: plink --bfile T2D --remove drop_sex-check --make-bed
```


Input Genotype Data
Samples passing sex check
Discard samples with problems from sex check
Marker Call Rate
Sample Call Rate
Minor Allele Frequency
Initial Quality Control
Discard markers and samples not passing initial quality control

MARKER:

```
plink --bfile T2D_sex-check --geno 0.01 --make-bed --out T2D_sex-check_g
```

Dropped 8416 markers.

Remaining:
Markers: 553,074
Samples: 3,894

SAMPLE:

```
plink --bfile T2D_sex-check_geno99 --mind 0.01 --make-bed --out T2D_sex-
```

Dropped 7 samples.

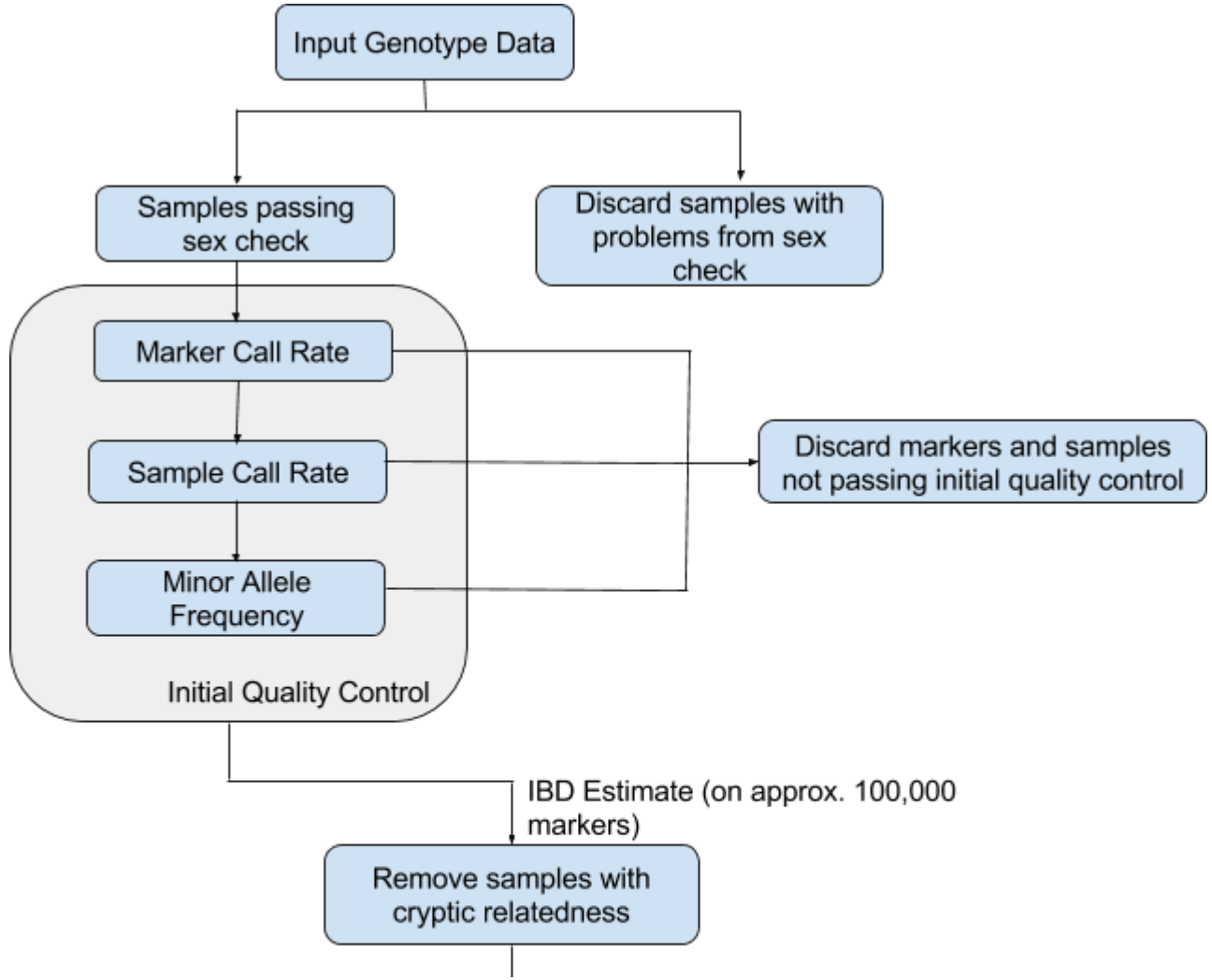
Remaining:
Markers: 553,074
Samples: 3,887

MAF:

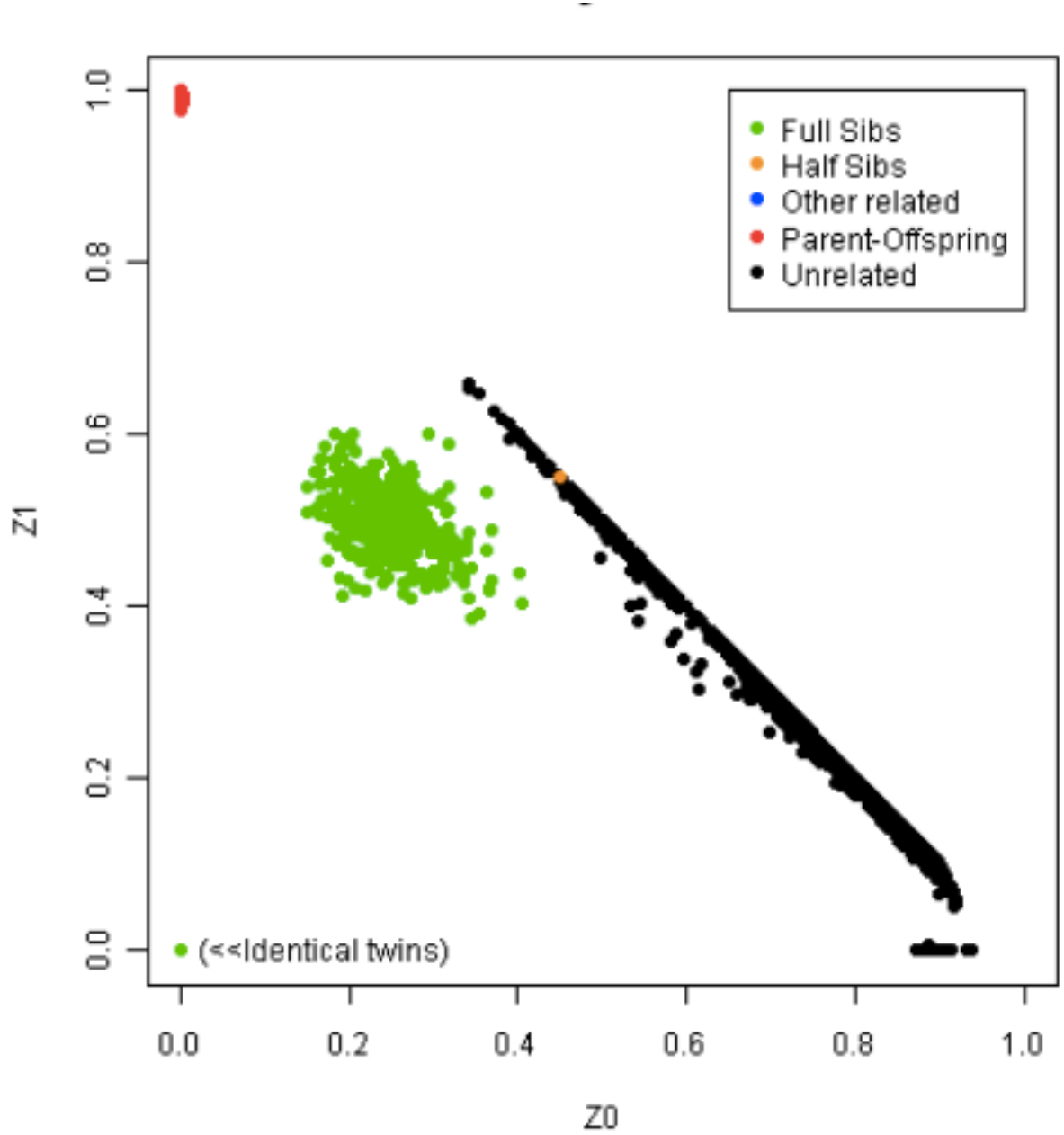
```
plink --bfile T2D_sex-check_geno99_mind99 --maf 0.05 --make-bed --out T2
```

52,401 markers dropped.

Remaining:
Markers: 500,673
Samples: 3,887


Input Genotype Data
Samples passing sex check
Discard samples with problems from sex check
Marker Call Rate
Sample Call Rate
Minor Allele Frequency
Initial Quality Control
Discard markers and samples not passing initial quality control
IBD Estimate (on approx. 100,000 markers)
Remove samples with cryptic relatedness

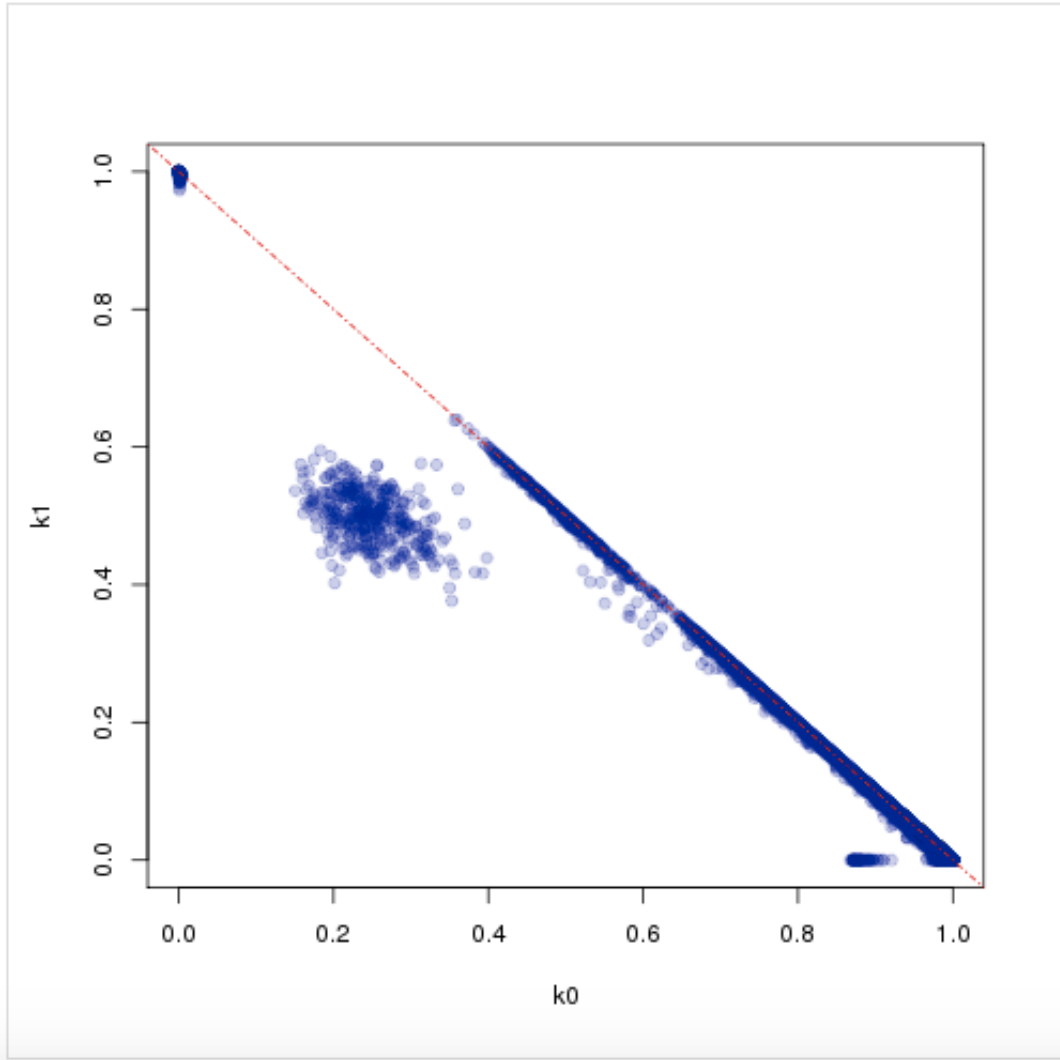
Z0	Z1	Z2	Kinship	Relationship
0.0	0.0	1.0	1.0	MZ twin or duplicate
0.0	1.0	0.0	0.50	Parent-offspring
0.25	0.50	0.25	0.50	Full siblings
0.50	0.50	0.0	0.25	Half siblings
0.75	0.25	0.0	0.125	Cousins
1.0	0.0	0.0	0.0	Unrelated


Full Sibs
Half Sibs
Other related
Parent-Offspring
Unrelated
(<<Identical twins)
Z1
Z0
0.0
0.2
0.4
0.6
0.8
1.0

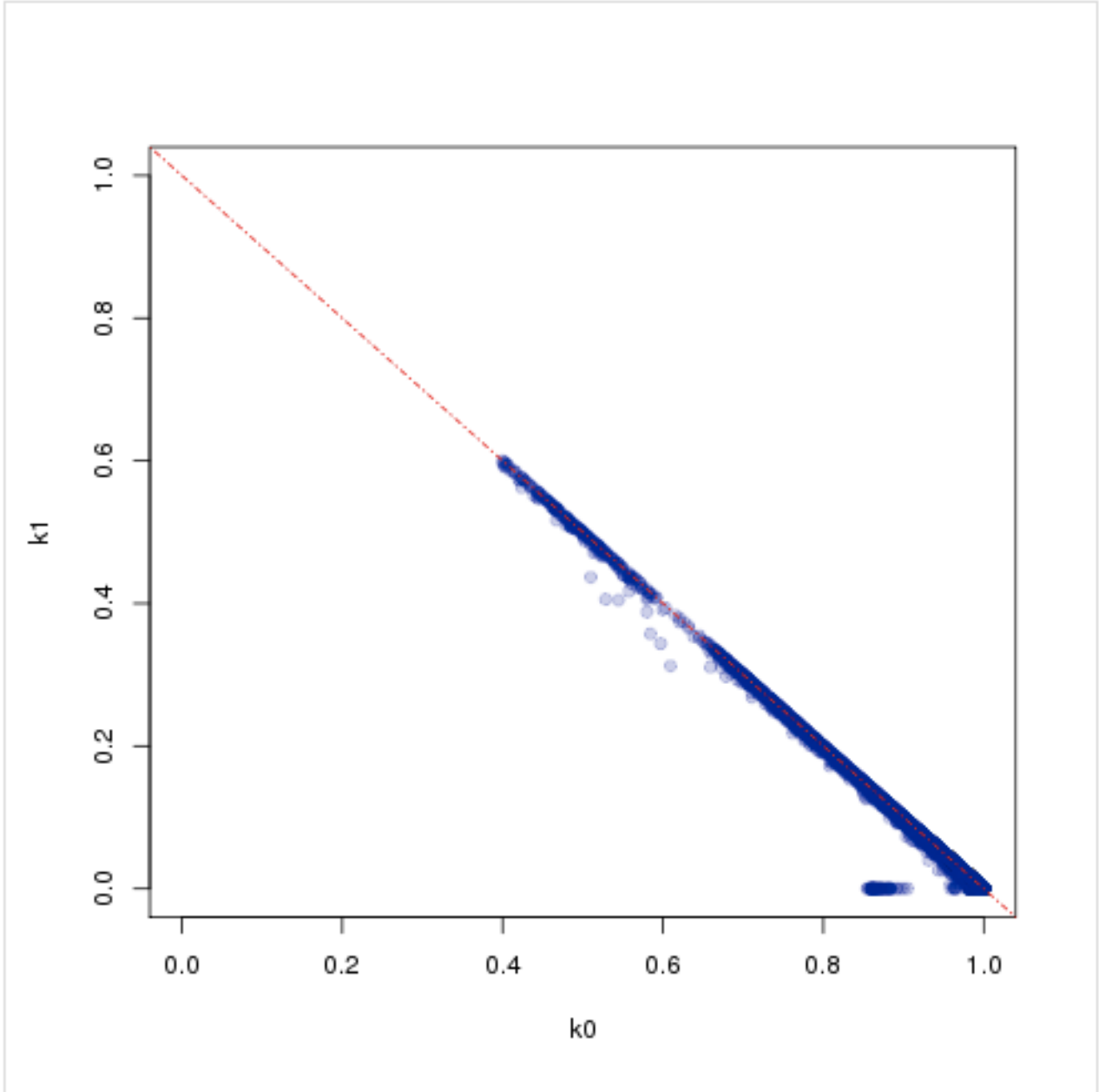
IBD:

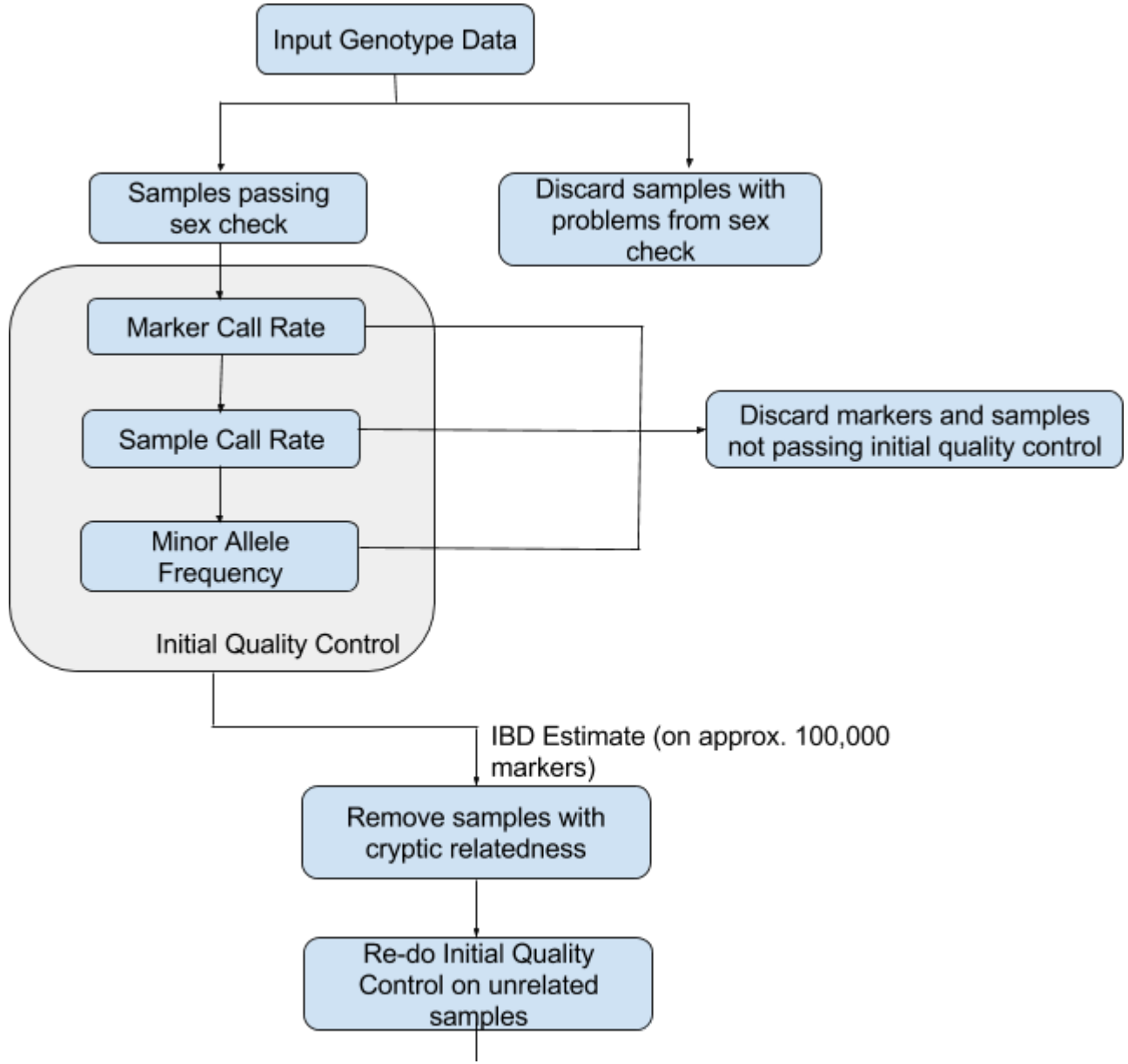
```
runQC -p 4 -f T2D_sex-check_geno99_mind99_maf05 -l 0.3 -C -P8
```

```
runQC -p 5 -f T2D_sex-check_geno99_mind99_maf05_prune03 -C -P8
```

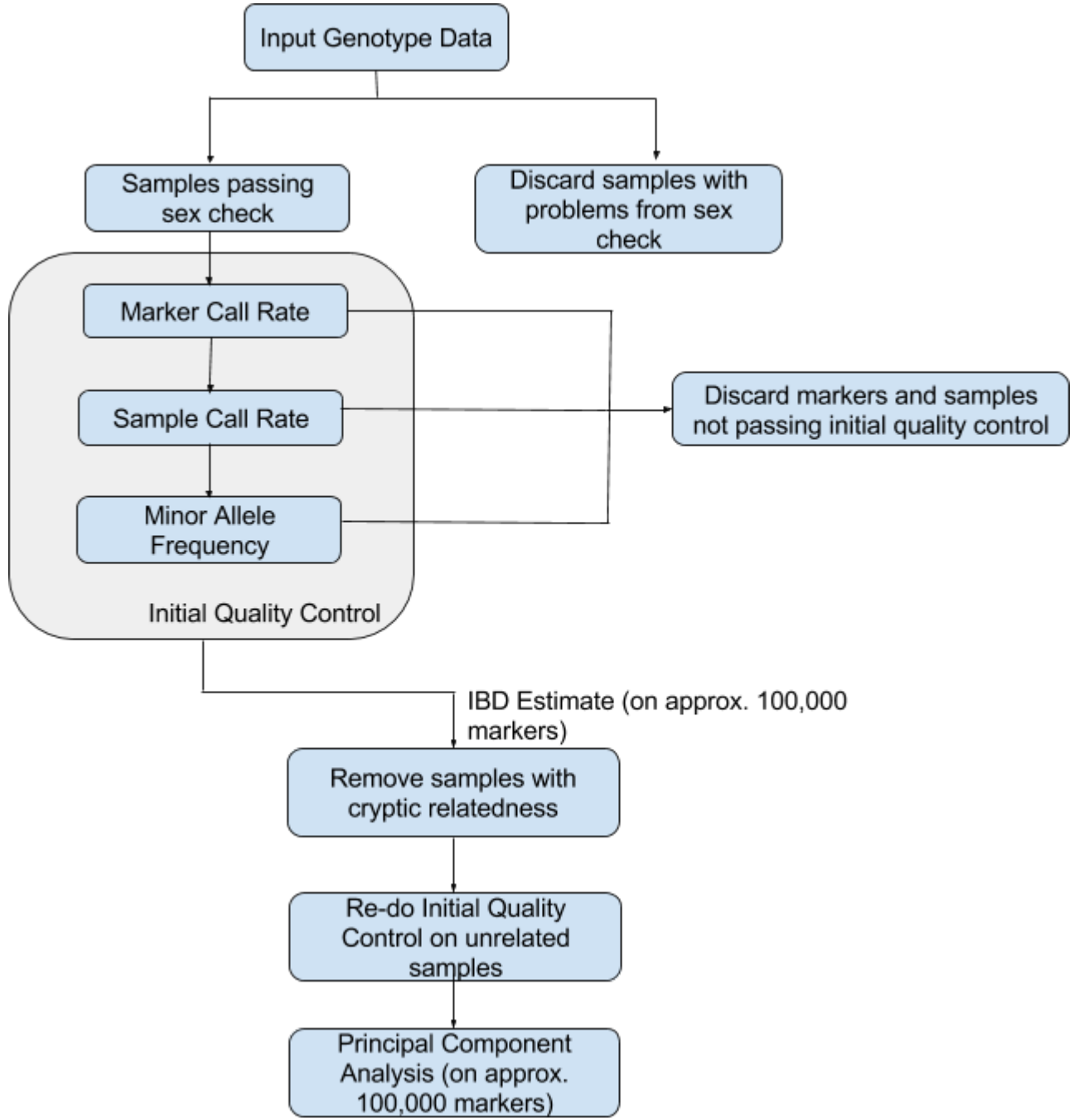


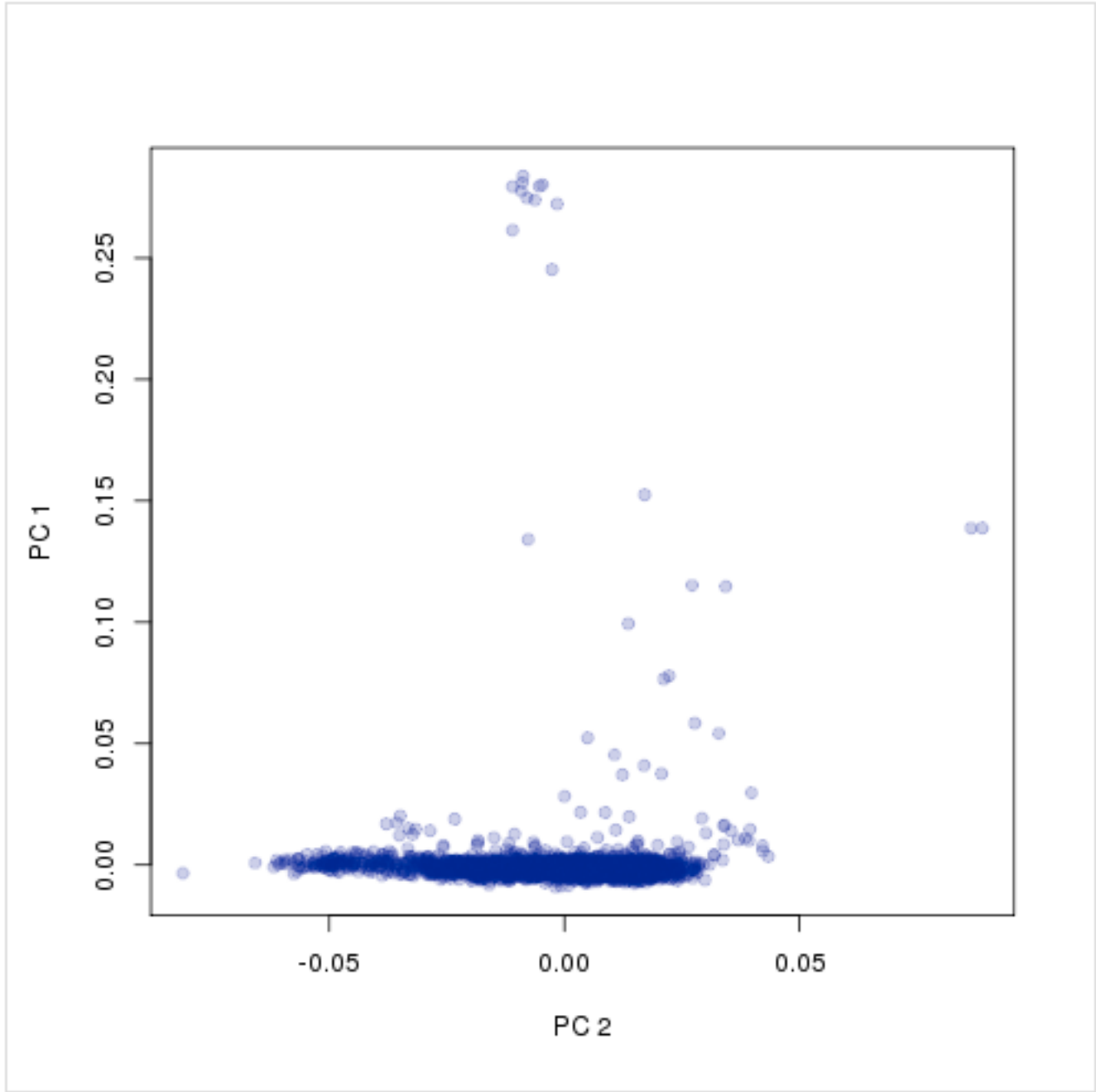
Dropped 513 samples.

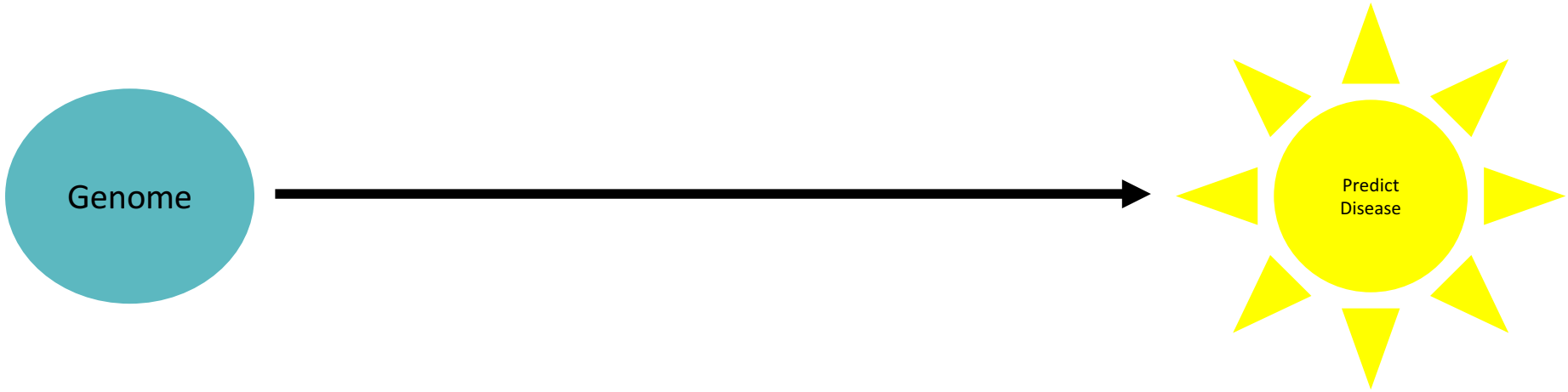



Input Genotype Data
Samples passing sex check
Discard samples with problems from sex check
Marker Call Rate
Sample Call Rate
Minor Allele Frequency
Initial Quality Control
Discard markers and samples not passing initial quality control
IBD Estimate (on approx. 100,000 markers)
Remove samples with cryptic relatedness
Re-do Initial Quality Control on unrelated samples

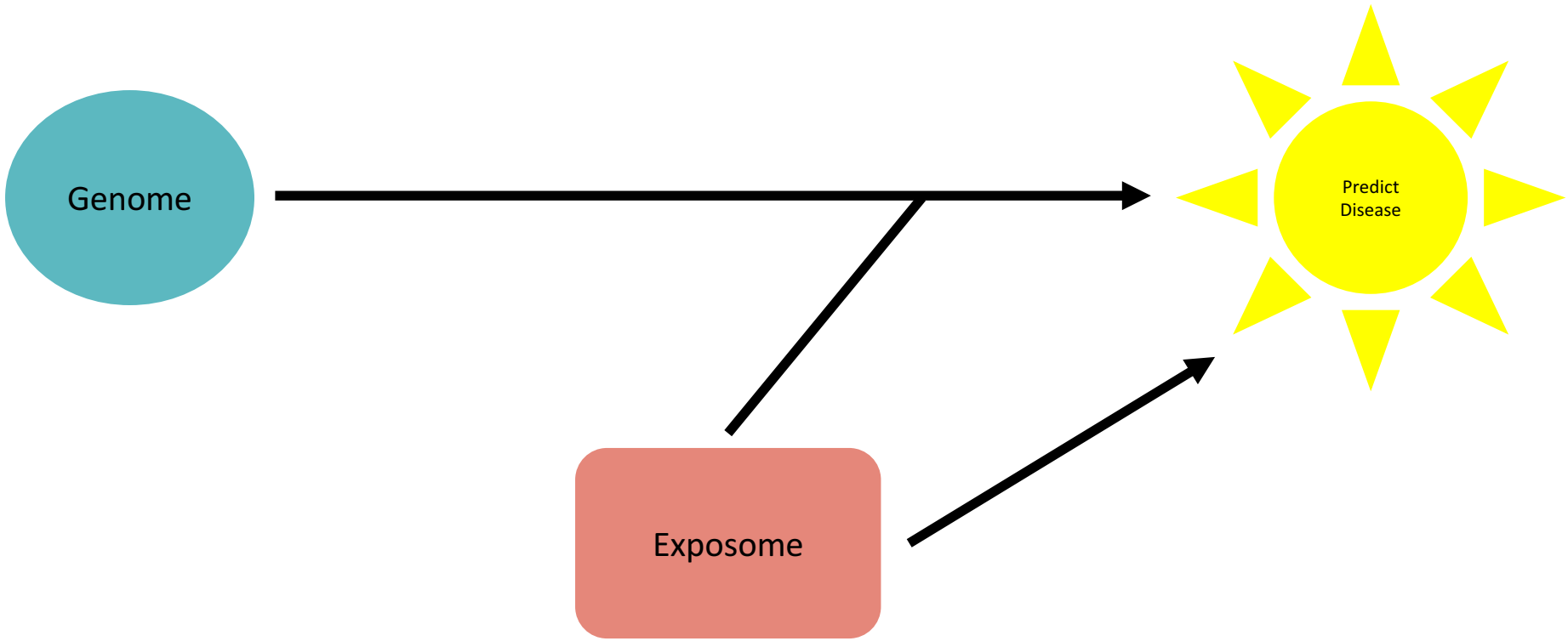
Any time samples are dropped - REPEAT


Input Genotype Data
Samples passing sex check
Discard samples with problems from sex check
Marker Call Rate
Sample Call Rate
Minor Allele Frequency
Initial Quality Control
Discard markers and samples not passing initial quality control
IBD Estimate (on approx. 100,000 markers)
Remove samples with cryptic relatedness
Re-do Initial Quality Control on unrelated samples
Principal Component Analysis (on approx. 100,000 markers)


PC 1
0.00
0.05
0.10
0.15
0.20
0.25
-0.05
0.00
0.05
PC 2

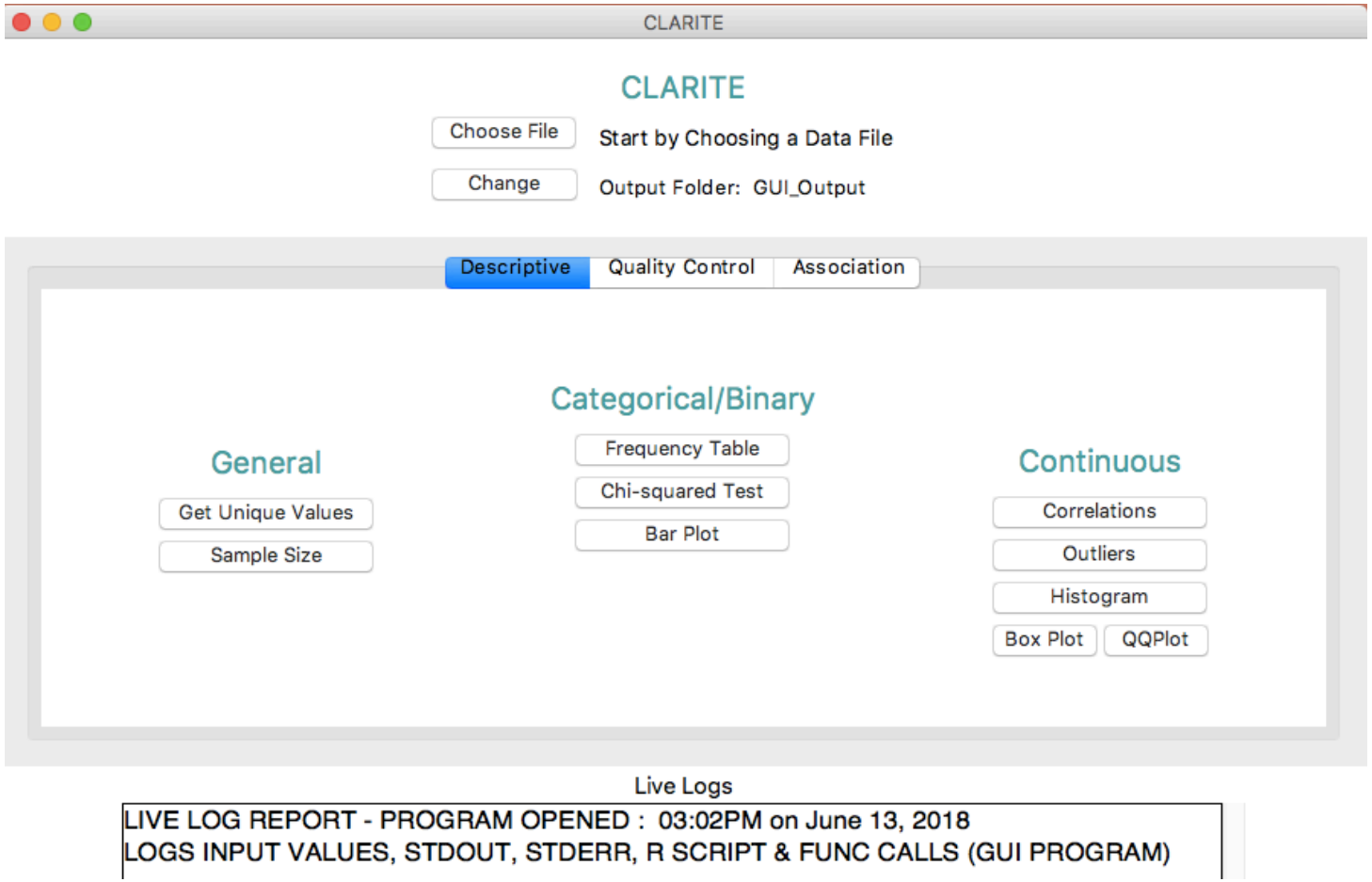

Genome
Predict
Disease

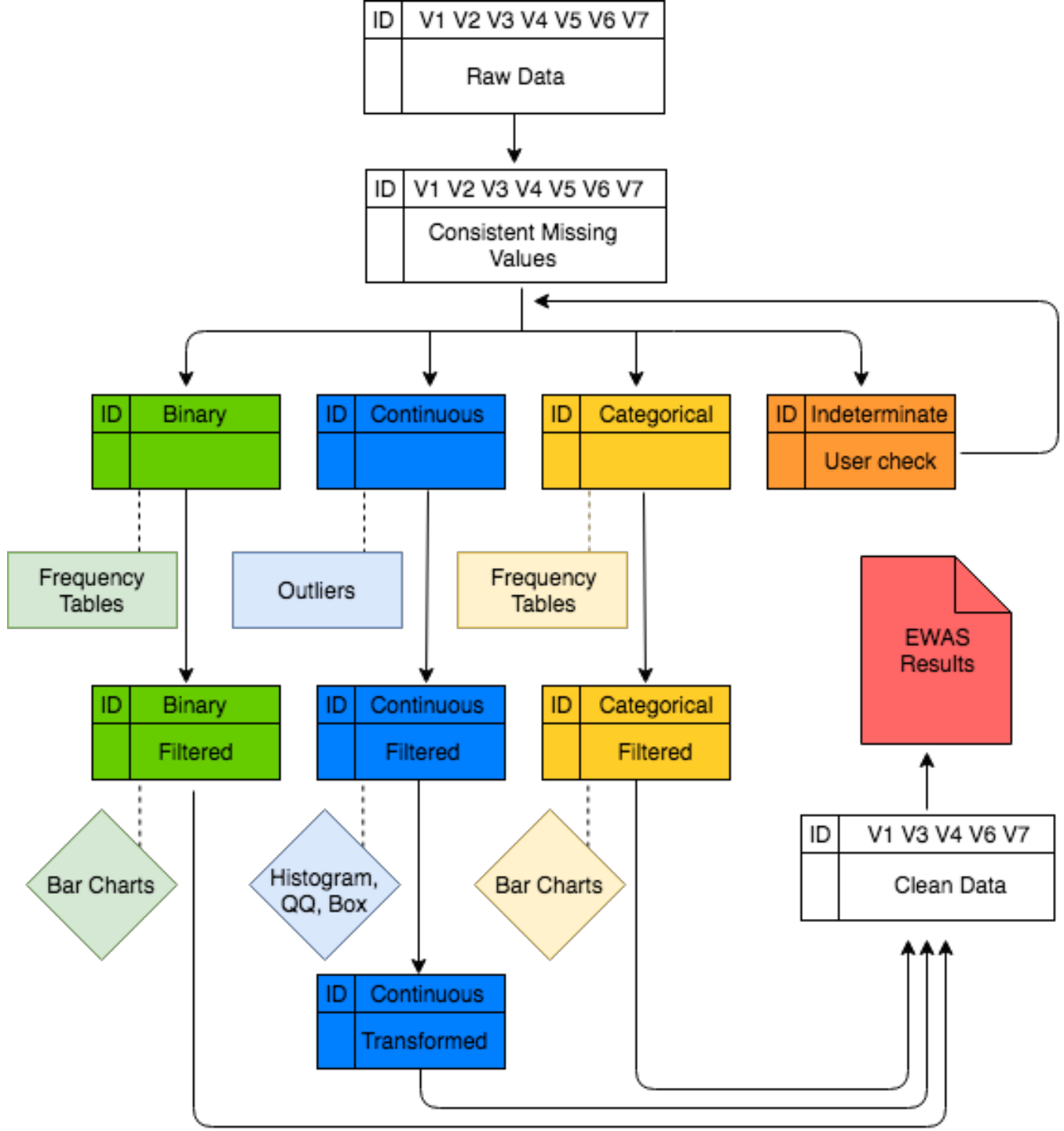
Future Directions



Issues to consider for exposure (and phenotype) data

- Outliers
- Survey data (bias, missingness, incorrect)
- Faulty lab measurement
- Small sample size
- Desperate data types (continuous, categorical, binary) combined
- Skew
- Can’t look by hand at big data!


CLARITE
CLARITE
Choose File
Start by Choosing a Data File
Change
Output Folder: GUI_Output
Descriptive
Quality Control
Association
Categorical/Binary
General
Frequency Table
Continuous
Get Unique Values
Chi-squared Test
Correlations
Sample Size
Bar Plot
Outliers
Histogram
Box Plot
QQPlot
Live Logs
LIVE LOG REPORT - PROGRAM OPENED : 03:02PM on June 13, 2018
LOGS INPUT VALUES, STDOUT, STDERR, R SCRIPT & FUNC CALLS (GUI PROGRAM)


ID V1 V2 V3 V4 V5 V6 V7
Raw Data
ID V1 V2 V3 V4 V5 V6 V7
Consistent Missing Values
ID Binary
ID Continuous
ID Categorical
ID Indeterminate
User check
Frequency Tables
Outliers
Frequency Tables
ID Binary
Filtered
ID Continuous
Filtered
ID Categorical
Filtered
EWAS Results
Bar Charts
Histogram, QQ, Box
Bar Charts
ID V1 V3 V4 V6 V7
Clean Data
ID Continuous
Transformed

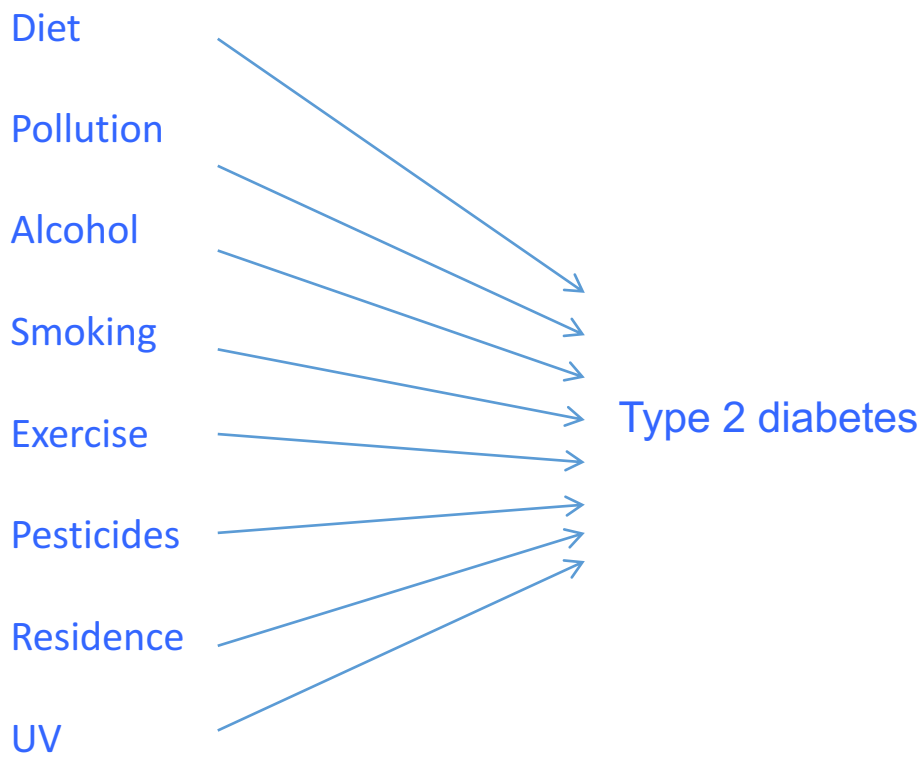

Click to take a virtual tour of NHANES
nhanes
Visit our website:
www.cdc.gov/nhanes
National Health and Nutrition Examination Survey
...monitoring the nation's health

Example of QC Protocol (NHANES HDL-C)

- Drop any sample that's missing a covariate value or phenotype
- Split by variable type:
  - Split into 4 tables by variable type: binary, continuous (min values = 15), categorical (3-6), and ambiguous (6-15)
  - By hand, determine ambiguous variable type and merge into appropriate file
  - Drop any variables that are indeterminant according to the NHANES data dictionary
- Sample Size Filter:
  - Drop variables < 200 samples ("Min # Samples") and < 200 samples in a category ("Min Category Size")
- Remove any variable with > 90% of the samples with a 0 value.
- Log(x+1) transformation all exposures and phenotypes

Environment-Wide Association Studies (EWAS)

- Test a variety of environmental variables in a high-throughput manner for association with phenotype(s)
- Analogous to GWAS method of testing loci across the genome
- Enable agnostic exposure assessment



Patel CJ, Bhattacharya J, Butte AJ (2010) An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. PLoS One 5: e10746

EWAS Discovery and Replication

- Following QC…
- Linear regression (HDL-C)
- Covariates: Sex, Age, BMI, SES, Race, Series

- Repeat QC and EWAS in Replication dataset (2 later surveys in NHANES)

MANHATTAN PLOT OF REPLICATING EWAS RESULTS TO BE ADDED

Summary

- The first step in data reproducibility is ensuring high quality data.
- To do this, rigorous and well-documented (so it can be reproduced!) QC is essential.
- Standardized QC protocols and tools are well-established and utilized in genomics.
- Few standardized protocols and tools are established for environment and phenotype data but are needed.

Key QC Papers:

- Stephen Turners GWAS QC paper: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3066182/
- Other useful papers: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3025522
- http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3061487/
- http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3592376/
- http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2896766/
- PLINK paper: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1950838/

Tools:

- PLINK: Most commonly used in Hall Lab for genotype quality control (http://pngu.mgh.harvard.edu/~purcell/plink/)
- PLATO: Most commonly used in Hall Lab for complex association studies (http://ritchielab.psu.edu/software/plato-download)
- Eigensoft/smartpca: For principal component analysis (PCA) (http://www.hsph.harvard.edu/alkes-price/software/)
- SNPrelate: For relatedness and PCA (https://www.bioconductor.org/packages/release/bioc/html/SNPRelate.html)

Hall-lab.org
mah546@psu.edu

Acknowledgements:

Molly Hall, PhD
Anastasia Lucas *
Nicole Palmiero
Bryan Almonte
Tuyen Pham