PNC Data Cleaning

Matthew Parker
Mostafavi Lab

June 26, 2019

Introduction

This document details each of the data preparation steps taken in preparing the PNC genotype data for analysis. The data processing pipeline is detailed in the accompanying document, or online at Genotype-Imputation.pdf.

Data Cleaning Steps

We looked at the data from chips: Omni, Quad, v1, and v3.

- Omni: HumanOmniExpress-12v1-Multi_B
- Quad: Human610-Quadv1_B
- v1: BDCHP-1X10-HUMANHAP550_11218540_C
- v3: HumanHap550v3_A

For each chip:

1. extract all subjects with race = European American, subjects:
    - Omni: 1657 → 1345
    - Quad: 3807 → 2076
    - v1: 556 → 335
    - v3: 1914 → 1084
2. quality control to exclude individual subjects:

Check (PLINK2)	Subjects Removed			
	Omni	**Quad**	**v1**	**v3**
ambiguous sex check	0	0	0	1
heterozygosity cutoff	14	35	4	25
Subjects Remaining	1331	2037	329	1043

3. quality control to exclude variants:

Check (PLINK2)	Variants Removed			
	Omni	**Quad**	**v1**	**v3**
MAF (0.01)	70,988	29,326	21,111	21,156
HWE (0.000005)	1756	1626	220	769
MIND, Missing genotype data (0.05)	0	4	2	15
GENO, Genotyping call rate (0.05)	3824	4809	8498	10462
Variants Remaining	636,158	541,238	511,497	515,071

4. data was aligned to HRC reference panel

	Omni	**Quad**	**v1**	**v3**
Variants Removed	5620	5117	7523	3140
Variants Remaining	630,538	536,121	503,974	511,931
Subjects Remaining	1331	2037	329	1043

5. the genotype data was then imputed using Michigan Imputation Server, and merged into a single dataset:

Total SNPs	39,018,346
Total Subjects	4740

6. Finally, only the assayed SNPs were used in calculating SNP principal components:

	SNPs	**Subjects**
Assayed SNPs	860,219	4740
After LD Pruning	121,902	4740
After IBD Pruning	121,902	4481