Quality Control Report for Exome Chip Data

University of Michigan

April, 2015

Project: Health and Retirement Study

Support: U01AG009740

NIH Institute: NIA

1. Summary and recommendations for users

A total of 15,603 study subjects were genotyped on the HumanExome-12v1-1 array. The median call rate is 99.95% and the error rate estimated from 339 pairs of sample duplicates is $1x10^{-6}$. Genotypic data are provided for all subjects and SNPs. We recommend selective filtering of genotypic data prior to analysis to remove whole samples with missing call rate >2% or inbreeding coefficient greater than 6SD from the mean (see Section 15). We also recommend that SNPs be filtered according to the criteria we suggest in Section 16. A composite SNP filter is provided, along with each of the component criteria so that the user may vary thresholds.

2. Project overview

Since 1992, the Health and Retirement Study (HRS, a cooperative agreement between the National Institute on Aging (NIA) and the University of Michigan) has been the largest, most representative longitudinal study of Americans over age 50. Built on a national probability sample with oversamples of minorities, it is the model for a network of harmonized international longitudinal studies that monitors work, health, social, psychological, family and economic status, and assesses critical life transitions and trajectories related to retirement, economic security, health and function, social and behavioral function and support systems.

The HRS is a nationally representative sample with a 2:1 oversample of African-American and Hispanic populations. The target population for the original HRS cohort includes all adults in the contiguous United States born during the years 1931–1941 who reside in households. HRS was subsequently augmented with additional cohorts in 1993 and 1998 to represent the entire population 51 and older in 1998 (b. 1947 and earlier). Since then, the steady-state design calls for refreshment every six years with a new six-year birth cohort of 51–56 year olds. This was done in 2004 with the Early Baby Boomers (b. 1948-53) and in 2010 with the Mid Boomers (b. 1954–59). The current sample in 2010 included over 22,000 persons in 15,000 households.

Core interview data are collected every two years using a mixed mode design, combining in-person and telephone interviews. In 2006, a random one-half of the sample was pre-selected to complete an enhanced face-to-face (EFTF) interview, which included a set of physical performance tests, anthropometric measurements, blood and saliva samples, and a psychosocial self-administered questionnaire in addition to the core HRS interview. The sample was selected at the household-level. In 2008, an EFTF interview was conducted on the remaining half of the sample. In 2010, a new cohort was enrolled and a random half of this cohort was selected to

complete an EFTF interview, including saliva collection. Respondents who don't consent to saliva collection at one wave are reasked at their next EFTF interview to provide a sample. Respondents who consented to the saliva collection in 2006, 2008 or 2010 are included in this data release.

3. HRS phenotypic data

Phenotypic data are available on a variety of dimensions. Health measures include self-reported doctor-diagnosed disease and some aspects of treatment, including medications, health insurance and utilization, smoking, drinking, height, weight, physical function, family characteristics and interactions, income, wealth and financial management, and job conditions. In addition, cognitive ability in several domains as well as depression are assessed at every wave. Innovative measures of risk tolerance and time preference have been used, as well as probabilistic expectations. The study is supplemented with administrative linkages to Medicare claims files providing diagnostic and utilization information, to the National Death Index, and to Social Security.

Beginning in 2006 the study added direct measures of physical function (grip strength, gait speed, balance, lung function), biomarkers of cardiovascular risk (blood pressure, total and HDL cholesterol, HbA1c, C-reactive protein and cystatin-C, height, weight, and waist circumference), and greatly expanded measurement of psychological traits (e.g., big 5 personality measures, affect, sense of control) and social networks.

4. Genotyping process

The DNA samples were genotyped at the Center for Inherited Disease Research (CIDR) with the Illumina HumanExome-12v1 array and using the calling algorithms from GenomeStudio version 2011.1, Genotyping Module 1.9.4 and GenTrain version 1.0. The SNP annotation provided by CIDR is "HumanExome-12v1-1_A," using genome build 37/hg19. Genotypic data that passed initial quality control at CIDR were released to the Quality Assurance/Quality Control (QA/QC) analysis team at the University of Michigan.

5. Sample and participant number and composition

In the following, the term "sample" refers to a DNA sample and, for brevity, "scan" refers to a genotyping instance (including genotyping chemistry, array scanning, genotype calls, etc.). A total of 15,888 samples (including duplicates) from study subjects were put into genotyping production, of which 15,773 were successfully genotyped and passed CIDR's QC process (**Table 1**). The subsequent QA process identified 21 samples with gender mismatch issues due to incorrect coding, and these errors were corrected. There were 258 samples with sex chromosome anomalies including apparent XX/XO, XXY, XY/XO, XYY and XXY with LOH on X. These samples remain in the release. However, this issue is annotated in the "Sample_analysis.csv" file. We suggest that users be cautious with these samples when analyzing the sex chromosome. We also identified an additional 32 samples (31 subjects) with unresolved identity issues including gender mismatch and unexpected duplicates, and these subjects were dropped. Therefore, the set of scans that was originally to be posted includes 15,741 study samples and 340 HapMap

controls. The 15,741 study samples derive from 15,572 subjects and include 169 duplicate scans. The 340 HapMap control samples derive from 146 subjects. All of the QA/QC processes that are described in this document are based on this set of scans.

However, after the initial QA/QC process was completed, it was determined that an additional 11 participants needed to be dropped due to identity issues discovered through genome-wide genotyping on the Illumina Infinium HumanOmni2.5 Beadchip (performed separately). Therefore, the final set of scans that are now posted includes 15,730 study samples and 340 HapMap controls, which derive from 15,561 subjects and 146 HapMap controls (we did not redo the QA/QC process after we dropped these 11 participants).

Table 1. Summary of DNA samples and genotyping instances

	Study	**HapMap**	**Both**
DNA samples into genotyping production	15888	340	16228
Failed samples	115	0	115
Scans released by genotyping center	15773	340	16113
Scans failing post_release QC	0	0	0
Scans with unresolved identity issues	32	0	32
Scans after initial QA/QC Subjects Replicated Scans	15741 15572 169	340 146 194	16081 15718 363
Scans dropped due to identity issues discovered through separate GWAS genotyping	11	0	11
Scans to be posted Subjects Replicated Scans	15730 15561 169	340 146 194	16070 15707 363

6. Relatedness

The relatedness between each pair of participants was evaluated by estimating genome-wide identity-by-descent (IBD) using PLINK. IBD coefficients were estimated using SNPs filtered for proximate linkage equilibrium. The filtering was done using PLINKs linkage disequilibrium (LD) based SNP procedure which

a) considered a window of 100 SNPs,

b) estimated the LD between each pair using pairwise genotypic correlation,

c) removed any pair of SNPs if the LD estimate was greater than 0.3,

d) shifted the window forward 25 SNPs, and repeated the process from (a)

After the SNPs were filtered, SNPs with minor allele frequency (MAF) less than 5% were removed, and SNPs with more than 1% missingness were removed. There were 18,929 SNPs remaining.

PLINK estimated the IBD for each pair of individuals using the formula

$$\text{pi-hat} = P(IBD=2) + 0.5 * P(IBD=1)$$

where the IBD coefficients were estimated using a Method of Moments procedure. A list of pairs of individuals with pi-hat greater than 0.125 was generated using the filtered SNP list of 18,929 SNPs.

We combined these results with existing knowledge about the relationship among the participants from other HRS resources. *We identified 104 families in total, and we suggest keeping one individual per family for an analysis that assumes all participants are unrelated (removal of 109 subjects).* Recommended unrelated individuals to include in an analysis sample are designated as *TRUE* in the "unrelated" column of the "Sample_analysis.csv" file. For each family, we preferentially kept samples that were included in the recommended analysis sample for the genome-wide genotyping performed on the Illumina Infinium HumanOmni2.5 Beadchip (a separate genotyping project that was composed of participants genotyped in 2006 and 2008), or the sample with the highest call rate.

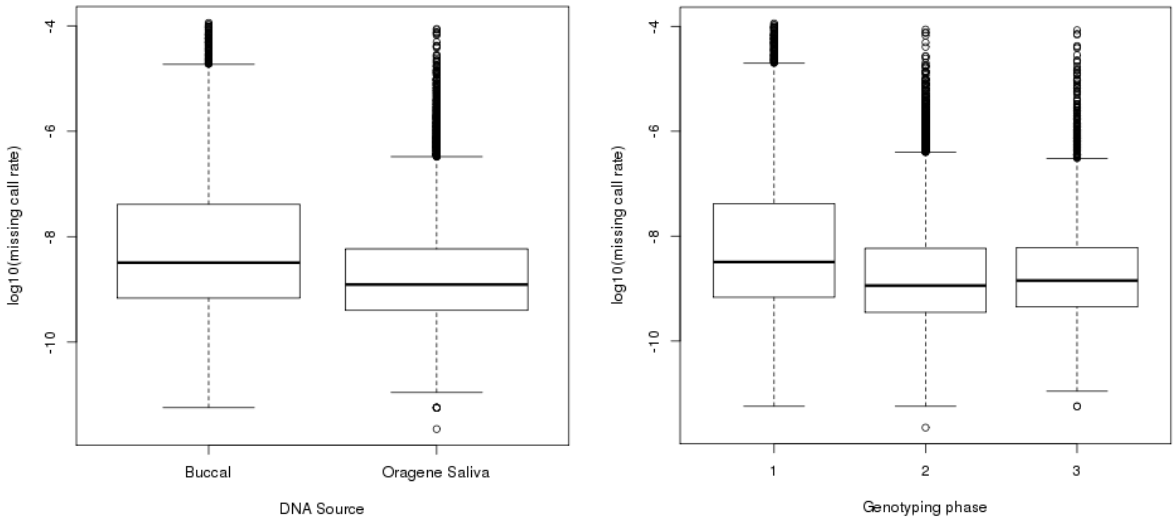
7. Missing call rates

Missing call rates were calculated for each sample and for each SNP in the following way (and are provided in files "SNP_analysis.csv" and "Sample_analysis.csv"). (1) missing.n1 is the missing call rate per SNP over all samples (including HapMap controls). (2) missing.e1 is the missing call rate per sample for all SNPs with missing.n1< 100%. (3) missing.n2 is the missing call rate per SNP over all samples with missing.e1< 5%. In this project, all samples have missing.e1< 5%, so missing.n1=missing.n2. (4) missing.e2 is the missing call rate per sample over all SNPs with missing.n2< 5%. In this study, the two missing rates by sample are very similar, with median values of 0.0001569 (missing.e1) and 0.0001395 (missing.e2). All samples have a missing rate less than 2%. For SNPs that passed the genotyping center QC, the median value of missing.n1 is 0 and 99.5% of SNPs have a missing call rate < 2%. *We recommend filtering out samples with a missing call rate > 2% and SNPs with a missing call rate > 2%.*

8. DNA source and genotype phase effects

Since the study was genotyped in three phases and there were both buccal samples and Oragene samples, we looked at missing call rate differences by genotyping phase and DNA source. There was a significant missing call rate difference by genotyping phase ($p<2.2x10^{-16}$, **Figure 1A**) as well as DNA source ($p<2x10^{-16}$, **Figure 1B**). Since Phase 1 is composed mostly of the buccal samples and Phase 2 and 3 are mostly Oragene samples, it appears that the major differences in missing call rate across genotyping phases originated from the DNA source. After adjusting for DNA source, the effect of the genotyping phase was greatly reduced (p=0.002). We suggest that investigators check the association between analysis phenotypes and DNA source prior to analysis.

Figure 1

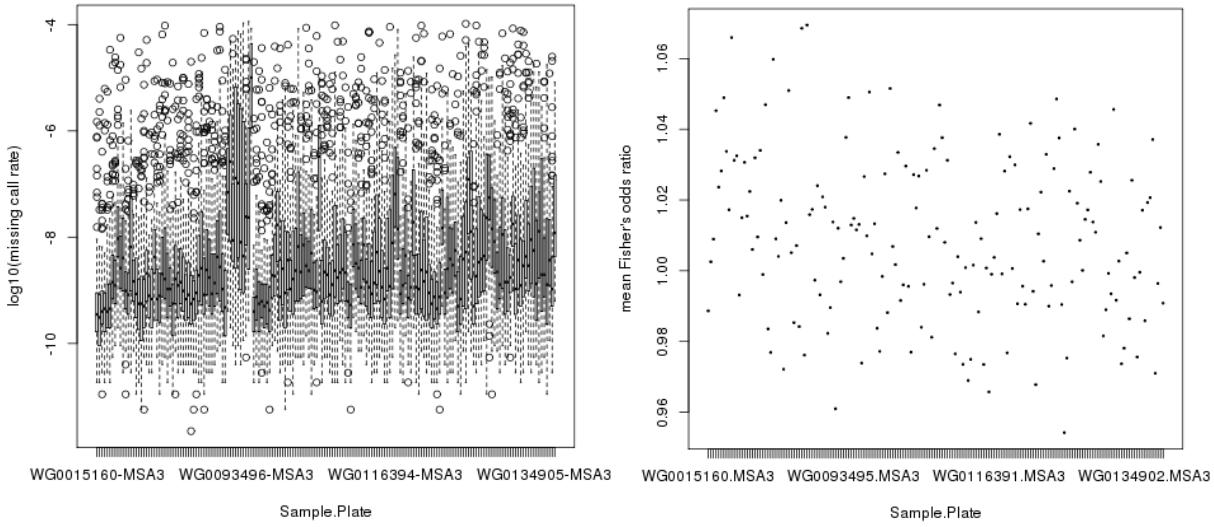


9. Plate effects

The samples were processed together in 176 complete or partial 96-well plates. There is highly significant variation among plates in log10 of the missing call rate ($p<2 \times 10^{-16}$), but all plates have a low missing call rate (**Figure 2A**).

We also assessed the difference in allelic frequencies between each plate and a pool of the other plates. To eliminate potential bias that could be introduced by many rare variants, we used only common variants for this purpose (MAF>0.05). We calculated the odds ratio (OR) from Fisher's exact test for each SNP and each plate and then averaged these statistics over SNPs, using only study samples. This statistic is a measure of how different each plate is from the other plates. **Figure 2B** shows the mean odds ratio for all plates (176 plates), and there are no outliers. We concluded that there are no problematic plate effects.

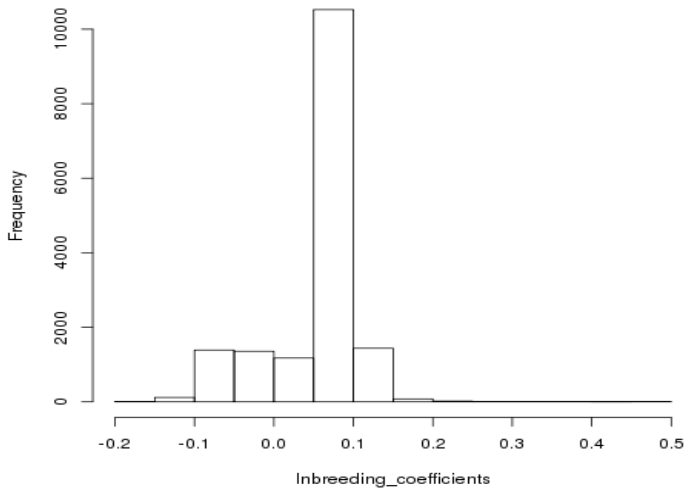
Figure 2



10. Inbreeding coefficient as a measure of sample quality

The inbreeding coefficient for a given sample can be used as a quality metric to detect unusual allelic distributions across the genome that may be due to low sample quality or sample contamination. We calculated the inbreeding coefficient for each sample using all of the SNPs. The distribution of the inbreeding coefficient is shown in Figure 3. *There is one sample that is outside of 6 times the standard deviation (SD) from the mean, and we recommend excluding this sample from the analysis.*

Figure 3



11. Duplicate sample discordance

Genotyping error rates can be estimated from duplicate discordance rates. The genotype at any SNP may be called correctly, or miscalled as either of the other two genotypes. If α and β are the two error rates, the probability that duplicate genotyping instances of the same participant will give a discordant genotype is $2[(1 - \alpha - \beta)(\alpha + \beta) + \alpha\beta]$. When α and β are very small, this is approximately $2(\alpha + \beta)$ or twice the total error rate. Potentially, each true genotype has different error rates (i.e. three α and three β parameters), but here we assume they are the same. In this case, since the discordance rate over all sample pairs is 2×10^{-6}, a rough estimate of the mean error rate is 1×10^{-6} per SNP per sample, indicating a high level of reproducibility. Duplicate discordance estimates for individual SNPs can be used as a SNP quality filter. The challenge here is to find a level of discordance that would eliminate a large fraction of SNPs with high error rates, while retaining a large fraction with low error rates. The probability of observing > x discordant genotypes in a total of n pairs of duplicates can be calculated using the binomial distribution. Table 2 shows these probabilities for x = 0–3 and n = 339. Here we chose n = 339 to correspond to the number of pairs of duplicate samples for both study and HapMap control samples. *We recommend a filter threshold of > 0 discordant calls because this removes > 49% of SNPs with an error rate > 10^{-3} and >99.8% of SNPs with error rate >10^{-2}. This threshold eliminates 171 SNPs.*

Table 2. Probability of observing more than the given number of discordant calls in 339 pairs of duplicate samples, given an assumed error rate. The number of SNPs with a given number of discordant calls is shown in the final column. The recommended threshold for SNP filtering (in bold) is >0 discordant calls.

# discordant calls	Assumed error rate				# SNPs
	1.0e-05	1.0e-04	1.0e-03	1.0e-02	
>0	**0.0068**	**0.0656**	**0.4927**	**0.998**	**171**
>1	2.281e-05	0.0022	0.1481	0.9916	14
>2	5.123e-08	4.896e-05	0.0314	0.9663	3
>3	8.603e-11	8.199e-07	0.0051	0.9083	1

12. Mendelian errors

Mendelian errors were analyzed in the 32 trios of HapMap control subjects. Only 220 SNPs have any Mendelian errors, and 24 SNPs have more than one error. We recommend filtering out SNPs with more than one Mendelian error.

Table 3. Number of SNPs with Mendelian errors

# Mendelian errors	# SNPs
1	196
2	**14**
3	4
4	2

5	2
6	1
15	1

13. Minor allele frequency

The minor allele frequency was calculated among the unrelated, non-duplicated study sample (see Section 16). Among the 229,456 SNPs with missing.n1< 1, there are 22,838 monomorphic SNPs. The distribution of all types of SNPs is listed in **Table 4**.

Table 4. Allele frequency distribution

Types of SNPs	N
Monomorphic SNPs	22,838
SNPs with 1 copy of the minor allele	15,609
SNPs with 2 copies of the minor allele	13,231
SNPs with 3 copies of the minor allele	11,260
SNPs with 4 copies of the minor allele	9,704
SNPs with 5 copies of the minor allele	8,204
SNPs with more than 5 copies of the minor allele and with MAF < 0.01	108,625
SNPs with $0.01 \leq$ MAF < 0.05	12,284
SNPs with MAF > 0.05	27,701
Total Number of SNPs	229,456

We also examined the sex differences in minor allele frequency for autosomal SNPs and the absolute differences range from 0 to 0.031.

14. Duplicate SNP probes

The HumanExome-12v1-1 array has 799 pairs of SNPs that occur in duplicate, as indicated by duplicated genomic position. Among the 799 pairs, there are 78 pairs of SNPs that had at least one SNP that did not pass quality control and was not released by CIDR (call rate=0). There are an additional 38 pairs of SNPs that have an annotation problem, in which the alleles in one probe are not the same as the alleles in the other probe. This problem can arise for two reasons: (1) the SNP is triallelic, in which the minor allele in one probe is not the same as the minor allele in the other probe, or (2) one or both of the probes have annotation errors. For the remaining 683 pairs, the concordance rate for each pair of probes was calculated. Most of these pairs have a very high level of concordance across the 16,081 study samples. The numbers of pairs of duplicate SNPs with various levels of discordance are given in Table 5.

Table 5. Probability of observing more than the given number of discordant calls in 683 pairs of duplicate SNPs, given an assumed error rate. The number of SNP pairs with a given number of discordant calls is shown in the final column. The recommended threshold for SNP filtering (in bold) is >1 discordant calls.

# discordant calls	Assumed error rate				# SNPs
	1.0e-05	1.0e-04	1.0e-03	1.0e-02	
>0	0.2750	0.9599	1	1	67
>1	**0.0419**	**0.8309**	**1**	**1**	**39**
>2	0.0044	0.6235	1	1	29
>3	3.451E-04	0.4011	1	1	25
>4	2.195E-05	0.2223	1	1	18
>5	1.167E-06	0.1072	1	1	17
>6	5.327E-08	0.0456	1	1	15
>7	2.131E-09	0.0173	1	1	13
>8	7.584E-11	0.0059	1	1	11
>9	2.431E-12	0.0018	1	1	9
>10	7.083E-14	0.0005	1	1	6

A high level of discordance may indicate that the SNP has a high error rate or that the two members of the pair may not be assaying the same SNP. *We recommend filtering out both members of each SNP pair with > 1 discordances because this is expected to eliminate only 4% of the SNPs with an error rate of 0.0001, > 83% of SNPs with error rate 0.001 and essentially all of the SNPs with error rate of 0.01 (this removes 39 pairs of SNPs). We also recommend filtering out SNPs that have the aforementioned annotation problems (38 pairs). Finally, we recommend filtering out one member of each pair with <= 1 discordant calls—the one with lowest missing call rate, since these provide redundant information (this removes one member of each of the remaining 644 pairs).* Filtering information for the duplicated SNPs and the annotation problems is provided in the file "SNP_analysis.csv".

15. Sample exclusion and filtering summary

As discussed in Section 5, genotyping was attempted for a total of 15,888 study samples, of which 15,773 passed CIDR's QC process (Section 2). The subsequent data cleaning QA process identified 32 samples with unresolved identity issues. After all the QA/QC processes were completed, it was determined that an additional 11 participants needed to be dropped due to identity issues discovered through genome-wide genotyping using the Illumina Infinium HumanOmni2.5 Beadchip (performed separately). Therefore, 15,730 study scans are posted.

We recommend filtering out whole samples with missing call rate > 2% and samples with inbreeding coefficient out of 6SD from the mean. *Subjects that we recommend for inclusion in an analysis sample are designated as TRUE in the "Analysis" column of the file "Sample_analysis.csv".* This list includes the recommended filters, has just one scan per subject (unduplicated), and has one subject per family (unrelated).

16. SNP filter summary

Table 6 summarizes SNP failures applied by CIDR prior to data release and a set of additional filters suggested for removing assays of low quality. There are 22,838 monomorphic SNPs that

were not removed because they could be informative when meta-analysis is conducted across cohorts.

Table 6: SNP Filters

Filter	SNPs lost	SNPs kept
SNP probes		242,901
Intensity-only SNPs	0	242,901
CIDR technical filters	13,445	229,456
Duplicate SNPs	798	228,658
>0 discordant calls in 339 study duplicates	171	228,487
>1 Mendelian error	24	228,463
Missing call rate >=2%	1098	227,365
Percent of SNPs lost due to quality filters	6.4%	

The suggested composite quality filter is provided as a TRUE/FALSE vector in the "SNP_analysis.csv" file, which also has the individual quality metrics so that the user can apply alternative thresholds. The recommended filters remove 6.4% of the 242,901 SNP assays attempted.

17. Project Participants

University of Michigan

David R. Weir, Jessica D. Faul, Sharon L.R. Kardia, Jennifer A. Smith, and Wei Zhao

Center for Inherited Disease Research, Johns Hopkins University

Kim Doheny, Jane Romm, Michelle Zilka, Tameka Shelford, Hua Ling, Elizabeth Pugh, and Marcia Adams

National Institute on Aging

Jonathan King, Georgeanne Patmios, and John Phillips