ADGC 1000 Genomes combined data workflow

This data set was prepared by Kevin L. Boehme (Brigham Young University, kevinlboehme@gmail.com), Shubhabrata Mukherjee "Joey" (University of Washington, smukherj@uw.edu), Paul K. Crane (University of Washington, pcrane@uw.edu) and John Kauwe (Brigham Young University, kauwe@byu.edu). As the above listed individuals have provided significant intellectual effort in developing these data they should be named as authors in manuscripts that use these data. Please contact these authors early in the manuscript preparation process to resolve authorship decisions.

College of Life Sciences

Brigham Young University

Provo, UT

Department of Medicine

Division of General Internal Medicine

University of Washington

Seattle, WA

December 3, 2014

Citation: Boehme KL, Mukherjee S, Crane PK, Kauwe JSK. ADGC 1000 genomes combined workflow (electronic document). September 2014. http://kauwelab.byu.edu/Portals/22/adgc_combined_1000G_12032014.pdf

Contents

1 Introduction **2**

2 ADGC 1000 Genomes combined data workflow **3**

2.1 Converting data sets from SNPTEST format to PLINK allele calls format . . 3

2.2 Further processing of PLINK binary files 4

2.3 Merging data sets together . . . 5
2.4 Identifying common genotyped (not imputed) SNPs and QC steps . . . 5
2.5 Addressing known and cryptic relatedness . . . 6
2.6 Principal components calculation . . . 7
2.7 Final files . . . 7
2.7.1 Genotype and Covariate Data Files . . . 7
2.7.2 Auxiliary Data Files . . . 7
2.7.3 README Files . . . 8
2.7.4 Final Notes . . . 8

3 Summary **10**

4 Caveats **11**

5 References **12**

1 Introduction

This document details the analyses we performed to generate the combined 1000 Genomes-imputed data sets from the imputed ADGC data sets. These data sets may prove useful to other investigators for a variety of purposes. Of course, for other purposes, the parent data sets may be a better choice. We provide a "Caveats" section 4 where we describe points to consider before using this data. In either case, some of the steps we took to ensure that the structure of all of the data sets were the same may prove useful.

In broad strokes, this document details steps we took to merge ADGC Stage 1 and Stage 2 imputed data to generate the final files ADGC_full.[bim|bam|fam|covar] and ADGC_unrelated.[bim|bam|fam|covar].

Both the ADGC Stage 1 and Stage 2 data can be accessed via ftps from the UPENN server (alois.med.upenn.edu).

The resulting ADGC_full.[bim|bam|fam|covar] and ADGC_unrelated.[bim|bam|fam|covar] files, along with all other auxiliary data generated by this project (and referenced in the text), can be found bundled together on the UPENN server (alois.med.upenn.edu) in the ADGC_Combined directory in the ADGC Phase 2 data location.

To accomplish these steps we used the following software packages:

- PLINK 1.9 ([2] [1]; https://www.cog-genomics.org/plink2)
- GTOOL ([3]; http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html)
- KING-Robust ([4]; http://people.virginia.edu/~wc9c/KING/index.html)
- EIGENSTRAT ([5]; http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html)

- R ([6]; http://www.r-project.org/)

These analyses were accomplished using the computational resources of the Fulton Supercomputing Lab (FSL) at Brigham Young University. FSL maintains 896 compute nodes (servers) comprising 12,100 processor cores with a compute capacity of over 120 Teraflops. All resources are supported by approximately one petabyte of high performance storage (https://marylou.byu.edu/about).

Section 2 of this document details specific elements of the steps taken to accomplish this work. Section 3 provides boilerplate methods text that may be used in publications that use these data. Section 5 provides references cited.

2 ADGC 1000 Genomes combined data workflow

Note: Code used in this project can be found in the README directory in the file ADGC_create_combined_dataset_codes_09222014.txt.

2.1 Converting data sets from SNPTEST format to PLINK allele calls format

- Initially, all of the data sets were in SNPTEST format (GEN/SAMPLE) files. For a description of the file format see http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html. There were 32 studies with each study divided by autosomal chromosome (1-22) for a total of 704 GEN files.

- The first step was to filter SNPs with low info (info <0.5). The info metric is used as a measure of imputation quality with values near 1 meaning the SNP was imputed with high certainty (see http://mathgen.stats.ox.ac.uk/impute/output_file_options.html#info_metric_details). We used *awk* to gather SNPs for exclusion in each of the data sets then GTOOL [3] to write the information quality-filtered SNP data sets.

- The GEN and SAMPLE files were modified according to PLINK 1.9 [1] specifications in order to convert them to best guess genotype/allele calls format.

  - We recoded the SAMPLE files, where the original coding was (Control = 1, Case = 2, Missing = -9) changed to (Control = 0, Case = 1, Missing = NA). This is the format required by PLINK 1.9 for *.sample* files.
  - We updated the GEN files to fill in the chromosome position.

- We took these modified files and used PLINK 1.9 to convert dosage information to best guess PLINK allele call format files: bed/bim/fam. We used the default PLINK 1.9 uncertainty cutoff of .1, meaning any imputed call with uncertainty greater than .1 was treated as missing and anything less, as a hard call.

2.2 Further processing of PLINK binary files

We completed several processing steps to prepare to merge the PLINK binary files together.

- We used a perl script to identify duplicate SNPs within each data set and used PLINK 1.9 to exclude all instances of these SNPs. Note that it is default behavior for PLINK 1.9 to remove both instances of duplicate SNPs. Below we provide a summary table of the number of duplicate SNPs removed from each study (Table 1). We provide a file with the rs numbers of the SNPs excluded from each study in the file ADGC_duplicate_SNPs.txt located in the auxiliary_data subdirectory.

Study (Phase 1)	Duplicates	Study (Phase 2)	Duplicates
ACT 1	10	ACT 2	10
ADC 1	10	ADC 4	10
ADC 2	10	ADC 5	10
ADC 3	10	ADC 6	9
ADNI	10	BIOCARD	10
GSK	8	CHAP2	10
NIA-LOAD	12	EAS	10
MAYO	12	UMVUTARC2	10
MIRAGE	12	NBB	33
OHSU	12	RMAYO	10
ROSMAP	10	ROSMAP2	10
TGEN2	12	TARCC1	10
UMVUMSSM_A	10	UKS	10
UMVUMSSM_B	10	WASHU2	10
UMVUMSSM_C	10	WHICAP	10
UPITT	12		
WASHU	12		

Table 1: Duplicates Removed

- We combined the 22 PLINK-formatted files (one per chromosome) from each study into a single data set per study.

- We checked the consistency of the genomic physical location data across all of the data sets, choosing ADC1 as our reference data set because it contained the most SNPs at this point. We identified 4 SNPs within the UM/VU/MSSM A and UM/VU/MSSM B files that had different genomic physical locations specified than in all of the other data sets. Those SNPs were rs4433978, rs4664277, rs4256345, and rs3819263. We changed the genomic physical location data in the files for the two UM/VU/MSSM data sets to match the genomic physical location data in all of the rest of the data sets.

- After confirming the genomic physical location was identical across studies, we split each study back into 22 chromosome files. This step facilitated merging the data quicker and using less computational resources.

2.3 Merging data sets together

- For each chromosome, using ADC1 as our reference data set, we systematically merged the other data sets.
- In the process of merging, any strand flip errors will cause PLINK to stop and print out the offending SNPs. We found only 1 variant on chromosome 6 (rs9453295) which was flipped in 13 of the data sets (ADC2, ADC4, ADC5, ADC6, BIOCARD, CHAP, EAS, MTV, NBB, RMAYO, ROSMAP2, WASHU2, and WHICAP). We recoded this variant in those data sets so that all strands matched ADC1.
- We used a MAF of 0.01 to filter the resulting merged data sets.
- We then combined all of the chromosomes together to generate a single combined data set. Table 2 shows basic summary statistics for each combined data set.

	ADGC_full	ADGC_unrelated
# SNPs	8,631,242	8,631,242
# Samples	37,635	28,730
Genotyping Rate	93.44%	93.28%

Table 2: Summary Statistics

2.4 Identifying common genotyped (not imputed) SNPs and QC steps

In order to a) evaluate relatedness across studies and b) calculate principal components to account for population-specific variations in allele distributions, we created a data set with observed/raw SNPs which were common across the 32 studies.

- We extracted a common list of genotyped SNPs (no. of SNPs=17,146) across all ADGC studies based on quality controlled GWAS data. Steps we took to do this are as follows:
  - Downloaded the *.bim* files from each genotyped (not imputed) data set.
  - Using R [6] we found the intersection of genotyped SNPs across each study.

- There were 17,146 directly genotyped SNPs in common across all of the 32 studies. A file with those rs numbers is named ADGC_common_genotyped_SNPs.txt and can be found in the auxiliary_data subdirectory.

- Symmetrical or strand-ambiguous SNPs can create problems in some settings, especially when considering data across multiple studies [7]. We observed no symmetrical SNPs in the common, directly genotyped data set.

- Some of the directly genotyped SNPs are in LD with each other. From the previous step, we created an LD thinned/pruned by invoking the following thresholds in PLINK 1.9 (–maf 0.01 –geno 0.02 –indep-pairwise 1500 150 0.2). The LD pruned data set contains 14,675 SNPs is named ADGC_LD_pruned_common_genotyped_SNPs.txt and can be found in the auxiliary_data subdirectory. These SNPs were used for the cryptic relatedness analysis and the generation of principle components (See following sections).

2.5 Addressing known and cryptic relatedness

The ADGC family of data sets includes two that are family-based (NIA-LOAD and MIRAGE). In addition, study participants may be related to other study participants within a study and also across studies. We used KING-Robust to purge study participants more closely related than 3rd degree relatives from our unrelated data set.

- We used the ADGC_LD_pruned_common_genotyped_SNPs.txt file for this step.

- We used a kinship coefficient cut-off of 0.0442, which indicates 3rd degree relatives, in KING-Robust [4].

  - We compared the number of people we would exclude if we set the threshold at 2nd degree relatives, and found that extending to 3rd degree relatives came at a cost of only n=84 study participants.

- The ADGC_full.[bim|bam|fam|covar] data set includes data from 37,635 individuals, while the ADGC_unrelated.[bim|bam|fam|covar] data set includes data from 28,730 individuals who are no more closely related than 3rd degree relatives.

Note: Previously, for our similar work with the Hapmap2 data set, we selected one individual from each family in NIA-LOAD and MIRAGE to be included (using an R script) in the combined unrelated data set. In the current analysis, we used KING-Robust to check for unrelated individuals (see Figure 1). We ran a relatedness analysis in NIA-LOAD and MIRAGE using KING-Robust and with the common 14,675 directly genotyped SNPs across 32 studies and found more unrelated samples. As a sensitivity analysis, we re-ran the analysis using LD-pruned common SNPs (∼100K) between NIA-LOAD and MIRAGE and the results were similar. On digging further into the KING-Robust results, we found that not all individuals in some families were related (>= 4th degree).

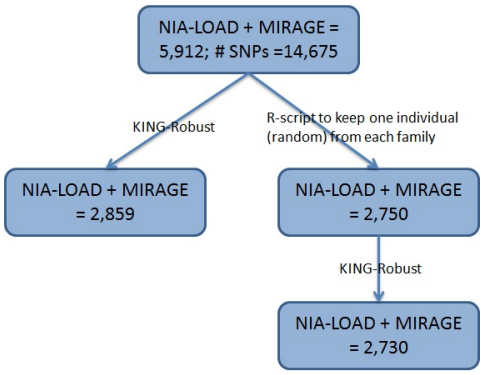

Figure 1

2.6 Principal components calculation

We used the ADGC_LD_pruned_common_genotyped_SNPs.txt file for this step.

- We used EIGENSTRAT [5] to calculate the first 10 PCs for the 28,730 unrelated individuals. The file with the first 10 PCs for the 28,730 unrelated people is called ADGC_unrelated_PCs.pca.evec, and can be found in the auxiliary_data subdirectory.
- These PCs were added to the covariate file, which already contains demographic variables such as sex, age, APOE genotype, etc.

2.7 Final files

All data resulting from this project can be accessed via ftps from the UPENN server (alois.med.upenn.edu). The final deliverables consisting of data previously referenced within the text as well as some additional README files not yet mentioned are as follows:

2.7.1 Genotype and Covariate Data Files

- **ADGC_full.[bim|bam|fam|covar]** - ADGC combined data set with all individuals (n=37,635).
- **ADGC_unrelated.[bim|bam|fam|covar]** - ADGC combined data set with only unrelated individuals (n=28,730).

2.7.2 Auxiliary Data Files

- **ADGC_duplicate_SNPs.txt** - List of duplicate SNPs removed from each study.
- **ADGC_common_genotyped_SNPs.txt** - List of genotyped SNPs that are present in each study.

- **ADGC_LD_pruned_common_genotyped_SNPs.txt** - List of LD-pruned genotyped SNPs that are present across all studies.
- **ADGC_unrelated_PCs.pca.evec** - EIGENSTRAT output containing first 10 PCs for all 28,730 unrelated samples.
- **ADGC_rectangular_snps.txt** - List of SNPs which pass the info filter (info >.5) and are common to each study and the final dataset. Useful for creating a rectangular data set with SNPs common across data sets. Created using the following steps:
  1. Collected all SNPs that pass info filter (info >.5) for each study.
  2. Took intersect of those SNPs. Gives 8,220,168 SNPs
  3. Took intersect of the previous results and the SNPs in our final data set. Gives 7,637,305 SNPs.

2.7.3 README Files

- **ADGC_combined_1000G_09222014.pdf** - A copy of this pdf.
- **ADGC_create_combined_dataset_codes_09222014.txt** - Contains codes used in this project.
- **ADGC_covar_DataDictionary.xlsx** - Describes the covariate file variables.

2.7.4 Final Notes

- Please note that we did not filter due to missing covariate data. Some participants have missing covariate data, including missing case/control status.
- The covariate file may be updated; investigators should check with ADGC for future updates.
- Table 3 on page 9 includes sex, case/control status, and overall sample size of the ADGC_full.[bim|bam|fam|covar] data set for each of the studies in ADGC wave 1 and ADGC wave 2.
- Table 4 on page 10 includes the same variables for the ADGC_unrelated.[bim|bam|fam|covar] data set.

Study	Sex (M/F)	Cases/Controls/Missing	Sample Size
ACT 1	1,090/1,458	567/1,701/280	2,548
ADC 1	1,247/1,494	1,961/731/49	2,741
ADC 2	429/499	738/160/30	928
ADC 3	768/995	942/638/183	1,763
ADNI	411/281	291/189/212	692
GSK/GenADA	615/957	799/773/0	1,572
LOAD/NIA-LOAD	1,675/2,735	1,852/1,991/567	4,410
YOUNKIN/MAYO	929/1,074	798/1,205/0	2,003
MIRAGE	603/897	603/885/14	1,502
KRAMER/OHSU	261/346	138/184/285	607
ROSMAP	522/1,163	388/878/419	1,685
TGEN2	639/872	952/559/0	1,511
MIAMI/UMVUMSSM	926/1,544	1,240/1,230/0	2,470
KAMBOH2/UPITT	825/1,394	1,283/849/87	2,219
WASHU/GOATE	289/381	403/225/42	670
ACT 2	182/205	23/8/356	387
ADC 4	451/603	385/464/205	1,054
ADC 5	507/717	354/701/169	1,224
ADC 6	575/760	485/399/451	1,335
BIOCARD	80/122	8/135/59	202
CHAP	299/450	32/198/519	749
EAS	134/152	12/240/34	286
MTV	212/330	303/235/4	542
NBB	96/204	215/85/0	300
RMAYO	261/171	24/320/88	432
ROSMAP 2	141/403	85/304/155	544
TARC1	242/382	399/225/1	625
UKS	849/897	770/976/0	1,746
WASHU 2	109/125	50/114/71	235
WHICAP	250/403	75/574/4	653
Total (Combined Phase 1 and 2)	15,617/22,014	16,175/17,176/4,284	37,635

Table 3: ADGC_full Sample Size

Study	Sex(M/F)	Cases/Controls/Missing	Sample Size
ACT 1	886/1,161	479/1,348/220	2,047
ADC 1	947/1,137	1,503/543/38	2,084
ADC 2	328/364	546/121/25	692
ADC 3	593/721	711/464/139	1,314
ADNI	317/207	215/140/169	524
GSK/GenADA	608/952	796/764/0	1,560
LOAD/NIA-LOAD	628/1,069	745/801/151	1,697
YOUNKIN/MAYO	706/835	616/925/0	1,541
MIRAGE	274/429	398/294/13	705
KRAMER/OHSU	142/188	59/109/162	330
ROSMAP	502/1,119	364/853/404	1,621
TGEN2	560/698	770/488/0	1,258
MIAMI/UMVUMSSM	817/1,380	1,085/1,112/0	2,197
KAMBOH2/UPITT	810/1,377	1,267/834/86	2,187
WASHU/GOATE	217/295	312/166/34	512
ACT 2	144/158	18/5/279	302
ADC 4	332/443	287/340/148	775
ADC 5	370/526	273/496/127	896
ADC 6	432/571	363/304/336	1,003
BIOCARD	75/113	8/123/57	188
CHAP	236/348	20/164/400	584
EAS	116/132	10/209/29	248
MTV	177/261	241/194/3	438
NBB	96/204	215/85/0	300
RMAYO	220/133	12/271/70	353
ROSMAP 2	105/323	62/237/129	428
TARC1	170/260	286/144/1	431
UKS	845/895	767/973/0	1,740
WASHU 2	68/67	30/65/40	135
WHICAP	246/394	74/562/4	640
Total (Combined Phase 1 and 2)	11,967/16,760	12,532/13,134/3,064	28,730

Table 4: ADGC_unrelated Sample Size

3 Summary

We converted IMPUTE2/SNPTEST format files to PLINK allele calls/best guess genotype (binary) format. We used the default PLINK 1.9 uncertainty cutoff of .1, meaning any imputed call with uncertainty greater than .1 was treated as missing. We filtered SNPs

imputed with low information (info<.5) from each dataset. We removed duplicate SNPs from each dataset. We identified 4 SNPs in 2 datasets with different genome physical locations and modified those so all physical locations were the same across all datasets. We identified one SNP with a flipped strand in 13 datasets, and flipped it so all SNPs had the same strand orientation in all datasets. We then merged all of the datasets together. We used a minor allele frequency of 0.01 to retain common SNPs.

We used directly genotyped (not imputed) SNPs for identifying cryptic relatedness and for calculating PCs to account for population structure. There were 17,146 directly genotyped SNPs in common across all 32 studies, none of which were symmetrical. We used PLINK to LD-prune these SNPs using the following settings: maf 0.01, geno 0.02, indep-pairwise 1500 150 0.2. These steps resulted in an LD-pruned directly observed non-ambiguous dataset with 14,675 SNPs.

We used KING-Robust to identify the 28,730 participants who were no more related than 3rd degree relatives (kinship coefficient 0.0442).

We used EIGENSTRAT to calculate the first 10 principal components for the 28,730 unrelated participants using the QC'd, LD-pruned directly observed set of SNPs common to all 32 studies.

4 Caveats

- The files described by this document were derived from ADGC files available September 22, 2014. Any errors in those files that we did not catch in our QC process will by necessity be incorporated in the final files. Any changes made to ADGC files after September 22, 2014 are not incorporated in the final files. As far as we know as of October 7, 2014, there have been no changes to those data, but the user should bear in mind that these files are static.

- Our workflow used called genotypes rather than dosage (see section 2.1)

- Many factors may impact genetic analysis results, including DNA quality, platform, and batch effects, not all of which are captured by study-specific indicator variables as we have used. Unfortunately data on platform for studies that used multiple platforms are not readily available, and DNA quality and batch data are not readily available. The user should be alert to these concerns, and should additional data become available on factors such as DNA quality, platform, and batch, these should be incorporated in analyses.

- The data set described by this document did not exist prior to our creating it. This is not the data set that has been used in any previous publication. For example, the Lambert et al. paper used a meta-analysis approach with these data sets, and did not perform all of the QC steps exactly as we have described. Thus GWAS results from this data set should not be expected to exactly match previously published results that analyzed different data sets.

- We would be most interested in any communications regarding additional QC issues, and will keep a published log of any questions that arise, along with responses.

5 References

References

[1] S. Purcell and C. Chang, "Plink 1.9: version v1.90b2a." https://www.cog-genomics.org/plink2.

[2] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, *et al.*, "Plink: a tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[3] C. Freeman and J. Marchini, "Gtool: A program for transforming sets of genotype data for use with the programs snptest and impute." http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html, 2007–2012.

[4] A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W.-M. Chen, "Robust relationship inference in genome-wide association studies," *Bioinformatics*, vol. 26, no. 22, pp. 2867–2873, 2010.

[5] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[6] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[7] C. T O'Dushlaine, C. Dolan, M. E. Weale, A. Stanton, D. T. Croke, R. Kalviainen, K. Eriksson, A.-M. Kantanen, R. A. Gibson, D. Hosford, *et al.*, "An assessment of the irish population for large-scale genetic mapping studies involving epilepsy and other complex diseases," *European Journal of Human Genetics*, vol. 16, no. 2, pp. 176–183, 2007.