OPEN ACCESS Freely available online

PLOS ONE

Best Practices and Joint Calling of the HumanExome BeadChip: The CHARGE Consortium

Megan L. Grove[1]*, Bing Yu[1], Barbara J. Cochran[1], Talin Haritunians[2], Joshua C. Bis[3], Kent D. Taylor[2], Mark Hansen[4], Ingrid B. Borecki[5], L. Adrienne Cupples[6,7], Myriam Fornage[8], Vilmundur Gudnason[9,10], Tamara B. Harris[11], Sekar Kathiresan[12,13,14], Robert Kraaij[15], Lenore J. Launer[11], Daniel Levy[7], Yongmei Liu[16], Thomas Mosley[17], Gina M. Peloso[12,14], Bruce M. Psaty[3,18,19,20,21], Stephen S. Rich[22], Fernando Rivadeneira[15,23,24], David S. Siscovick[3], Albert V. Smith[9,10], Andre Uitterlinden[23,24], Cornelia M. van Duijn[23,24], James G. Wilson[25], Christopher J. O'Donnell[7,26], Jerome I. Rotter[2], Eric Boerwinkle[1,27]

1 School of Public Health, Human Genetics Center, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America, **2** Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, California, United States of America, **3** Cardiovascular Health Research Unit, University of Washington, Seattle, Washington, United States of America, **4** Illumina, Inc., San Diego, California, United States of America, **5** Division of Statistical Genomics, Department of Genetics, Washington University School of Medicine, St. Louis, Missouri, United States of America, **6** Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, United States of America, **7** Framingham Heart Study of the National, Heart, Lung, and Blood Institute, Framingham, Massachusetts, United States of America, **8** Institute of Molecular Medicine, Center for Human Genetics, The University of Texas Health Science Center at Houston, Houston, Texas, United States of America, **9** Icelandic Heart Association, Research Institute, Kopavogur, Iceland, **10** Faculty of Medicine, University of Iceland, Reykjavík, Iceland, **11** Laboratory of Population Science, National Institute on Aging, Bethesda, Maryland, United States of America, **12** Center for Human Genetic Research and Cardiovascular Research Center, Massachusetts General Hospital, Boston, Massachusetts, United States of America, **13** Harvard Medical School, Boston, Massachusetts, United States of America, **14** Broad Institute of Harvard and MIT, Cambridge, Massachusetts, United States of America, **15** Department of Internal Medicine, Erasmus University Medical Center, Rotterdam, The Netherlands, **16** Department of Epidemiology and Prevention, Wake Forest University School of Medicine, Winston-Salem, North Carolina, United States of America, **17** Department of Medicine, University of Mississippi Medical Center, Jackson, Mississippi, United States of America, **18** Department of Epidemiology, University of Washington, Seattle, Washington, United States of America, **19** Department of Medicine, University of Washington, Seattle, Washington, United States of America, **20** Department of Health Services, University of Washington, Seattle, Washington, United States of America, **21** Group Health Research Institute, Seattle, Washington, United States of America, **22** Center for Public Health Genomics, University of Virginia, Charlottesville, Virginia, United States of America, **23** ErasmusAGE and Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands, **24** Netherlands Consortium for Healthy Aging, Netherlands Genomics Initiative, Leiden, The Netherlands, **25** Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi, United States of America, **26** Cardiology Division, Massachusetts General Hospital, Boston, Massachusetts, United States of America, **27** Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, United States of America

Abstract

Genotyping arrays are a cost effective approach when typing previously-identified genetic polymorphisms in large numbers of samples. One limitation of genotyping arrays with rare variants (e.g., minor allele frequency [MAF] <0.01) is the difficulty that automated clustering algorithms have to accurately detect and assign genotype calls. Combining intensity data from large numbers of samples may increase the ability to accurately call the genotypes of rare variants. Approximately 62,000 ethnically diverse samples from eleven Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium cohorts were genotyped with the Illumina HumanExome BeadChip across seven genotyping centers. The raw data files for the samples were assembled into a single project for joint calling. To assess the quality of the joint calling, concordance of genotypes in a subset of individuals having both exome chip and exome sequence data was analyzed. After exclusion of low performing SNPs on the exome chip and non-overlap of SNPs derived from sequence data, genotypes of 185,119 variants (11,356 were monomorphic) were compared in 530 individuals that had whole exome sequence data. A total of 98,113,070 pairs of genotypes were tested and 99.77% were concordant, 0.14% had missing data, and 0.09% were discordant. We report that joint calling allows the ability to accurately genotype rare variation using array technology when large sample sizes are available and best practices are followed. The cluster file from this experiment is available at www.chargeconsortium.com/main/exomechip.

Citation: Grove ML, Yu B, Cochran BJ, Haritunians T, Bis JC, et al. (2013) Best Practices and Joint Calling of the HumanExome BeadChip: The CHARGE Consortium. PLoS ONE 8(7): e68095. doi:10.1371/journal.pone.0068095

Editor: Yurii S. Aulchenko, Institute of Cytology & Genetics SD RAS, Russian Federation

Received March 6, 2013; **Accepted** May 25, 2013; **Published** July 12, 2013

Funding: Support for the centralized work was provided by Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium through the National Institutes of Health (NIH) American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419) (PI: E. Boerwinkle). Funding support of the individual cohorts is detailed below. The Age, Gene/Environment, Susceptibility-Reykjavik (AGES) study is funded by NIH contract N01-AG-12100, the NIA Intramural Research Program, the Icelandic Heart Association, and the Icelandic Parliament. The Atherosclerosis Risk in Communities (ARIC) Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute (NHLBI) contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN2682011000010C, HHSN2682011000011C, and HHSN2682011000012C). The authors thank the staff and participants of the ARIC study for their important contributions. The research reported in this manuscript on behalf of the Cardiac Arrest Blood Study (CABS) was supported by grants from the National Heart, Lung, and Blood Institute (HL41993 and HL091244); the University of Washington Clinical Nutrition Research Unit (DK-35816); and the Medic One Foundation, Seattle, WA. The authors gratefully acknowledge the Coronary Artery Risk Development in Young Adults (CARDIA) study participants and staff for their valuable contributions. The CARDIA study is funded by contracts N01-HC-95095, N01-HC-48047, N01-HC-48048, N01-HC-48049, N01-HC-48050, N01-HC-45134, N01-HC-05187, N01-HC-45205, and N01-HC-45204 from the National Heart, Lung, and Blood Institute to the CARDIA investigators. Genotyping of the CARDIA participants was supported by grants U01-HG-004729 and R01-HL-084099. This manuscript has been reviewed by CARDIA for scientific content and consistency of data interpretation with previous CARDIA publications. This CHS research was supported by NHLBI contracts N01-HC-85239, N01-HC-85079 through N01-HC-85086; N01-HC-35129, N01 HC-15103, N01 HC-55222, N01-HC-75150, N01-HC-45133, HHSN268201200036C and NHLBI grants HL080295, HL087652, HL105756 with additional contribution from NINDS. Additional support was provided through AG-023629, AG-15928, AG-20098, and AG-027058 from the NIA. See also http://www.chs-nhlbi.org/pi. DNA handling and genotyping was supported in part by National Center of Advancing Translational Technologies CTSI grant UL1TR000124, the National Institute of Diabetes and Digestive and Kidney Diseases grant DK063491 to the Southern California Diabetes Endocrinology Research Center, and Cedars-Sinai Board of Governors' Chair in Medical Genetics (JIR). The Rotterdam Study is supported by Erasmus Medical Center and Erasmus University, Rotterdam, Netherlands Organization for the Health Research and Development (ZonMw), the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture and Science, the Ministry for Health, Welfare and Sports, the European Commission (DG XII), and the Municipality of Rotterdam. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists. The NHLBI's Framingham Heart Study is a joint project of the National Institutes of Health and Boston University School of Medicine and was supported by contract N01-HC-25195 and its contract with Illumina, Inc. for genotyping services. The Family Heart Study (FamHS) was supported by grants from the National Heart, Lung, & Blood Institute (U01 HL56563, U01 HL56564, U01 HL56565, U01 HL56566, U01 HL56567, U01 HL56568, U01 HL56569, and K01-HL70444). The Health, Aging, and Body Composition (HABC) Study is supported by NIA contracts N01AG62101, N01AG62103, and N01AG62106. The genome-wide association study was funded by NIA grant 1R01AG032098-01A1 to Wake Forest University Health Sciences. This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging. The Jackson Heart Study (JHS) is supported by the National Heart, Lung, and Blood Institute (N01 HC-95170, N01 HC-95171 and N01 HC-95172) and the National Center on Minority Health and Health Disparities. The authors thank the staff, interns and participants in JHS for their long-term commitment and important contributions to the study. MESA is supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts N01-HC-95159 through N01-HC-95169 and UL1-RR-024156. Funding for MESA Family is provided by grants R01- HL-071051, R01-HL-071205, R01-HL-071250, R01-HL-071251, R01-HL-071252, R01-HL-071258, R01-HL-071259, UL1-RR-025005. The authors thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. A full list of participating MESA investigators and institutions can be found at http://www.mesa-nhlbi.org. Gina Peloso was supported by Award Number T32HL007208 from the National Heart, Lung, And Blood Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Heart, Lung, and Blood Institute or the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: Mark Hansen is an employee of Illumina, Inc. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials. All authors have declared that no competing interests exist.

* E-mail: Megan.L.Grove@uth.tmc.edu

Introduction

Exome- and whole-genome sequencing is becoming increasingly affordable and allows for detection and genotyping of rare variants in the human genome. Yet, genotyping arrays remain a cost-effective approach when investigating genetic polymorphisms previously identified in large populations. A limitation of using arrays to genotype rare variants is the difficulty that automated clustering algorithms have to accurately detect and assign accurate genotype calls [1,2]. Large sample sizes increase the number of occurrences of rare variants and, therefore, should facilitate automated clustering and genotyping.

An array focused on rare and low frequency coding variation, hereafter referred to as the exome chip, has been developed by querying the exomes sequenced in ~12,000 individuals and aggregating the variation that is seen in more than two individuals in more than two sequencing efforts (http://genome.sph.umich.edu/wiki/Exome_Chip_Design). Participating studies in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium [3] consented to have their Illumina Infinium HumanExome BeadChip intensity data analyzed collectively (n = 62,266) in order to increase the accuracy of rare variant genotype calls. The resulting cluster file (.egt) is publically available and we show that its use, along with best practices, increase genotype accuracy compared to other methods alone.

Results

Genotypes were obtained for 238,876 successful variants in accordance with our best practices (96.4% SNP pass rate) which were converted to PLINK format [4] by cohort and combined into a single aggregate file for further analyses. Of the 62,266 samples genotyped, 1,380 (2.2%) had a GenCall quality score in the lower 10^{th} percentile of the distribution across all variants genotyped (p10GC) <0.38 or call rate <0.97 and were excluded from allele frequency calculations. Because founder effects and unique population structure have been previously observed in Icelandic samples [5,6], the Age, Gene/Environment, Susceptibility-Reykjavik study was excluded from subsequent steps. Known duplicated samples, individuals without self-reported race, and the HapMap controls were also removed. After excluding duplicate variants (n = 811), the minor allele frequencies (MAF) for 238,065 successful SNPs and 56,407 samples by self-reported race are described in Table 1. There were 10,693 monomorphic SNPs (4.5%), and 78.6% of the variants on the exome chip have a MAF <0.005. Allele frequencies for each variant by race group are reported in the SNP information file (see Methods and Data Access sections). Ethnicity specific HapMap allele frequencies for the 96 controls (48 CEU and 48 YRI) and genotypes are also available.

Table 1. Exome chip minor allele frequency distribution by race.

MAF Interval	African Americans	Caucasians	Hispanics	Asians	All
	(n = 13,375) (%)	(n = 40,102) (%)	(n = 2,128) (%)	(n = 776) (%)	(n = 56,407) (%)
0	23.6	16.5	43.6	77.5	4.5
(0, 0.001]	36.8	58.1	22.3	3.9	58.8
(0.001, 0.005]	14.6	8.6	13.9	3.9	15.3
(0.005, 0.01]	4.4	2.3	3.7	1.6	4.3
(0.01, 0.05]	7.9	3.8	5.1	3.0	5.7
(0.05, 0.1]	2.7	1.7	1.7	1.4	1.8
(0.1, 0.2]	3.0	2.4	2.4	2.1	2.4
(0.2, 0.5]	7.1	6.7	7.3	6.6	7.3

The following samples were excluded: all AGES individuals, race unknown or not reported, known replicates, HapMap controls, individuals with p10GC <0.38, and individuals with call rate <0.97. Individuals with race designated as other were included in the overall MAF calculation, but data is not shown separately (n = 26). A total of 238,065 variants were used for calculating minor allele frequencies after excluding those that failed laboratory quality control (n = 8,994) and duplicates (n = 811).
doi:10.1371/journal.pone.0068095.t001

To evaluate the performance of the rare variant calling approach (see Methods), we compared exome chip genotypes derived from three calling methods to available exome sequencing data in 530 ARIC individuals. First, exome chip genotypes were called with the Illumina issued cluster file HumanExome-12v1.egt (see Data Access section for file location) (Dataset I). Second, we used zCall (threshold set to 7) [7] to determine genotypes for the missing variant calls in Dataset I to create Dataset Z. Third, we used the CHARGE best practices (see Data Access) and joint calling approach described to ascertain exome chip genotypes (Dataset C). A total of 185,119 variants that were present in the exome sequence dataset and passed our best practices were compared using genotype concordance and uncertainty coefficient tests. Results are presented in Table 2. The uncertainty coefficients indicate that we can predict 86.4% of the information (entropy) in the exome sequence data when using the Illumina cluster file, 91.2% when using the zCall algorithm, and 93.4% when the CHARGE clustering method was utilized.

These data demonstrate the importance of implementing stringent laboratory quality control measures in addition to the clustering algorithms and rare variant calling approaches tested. The complete list of 8,994 failing SNPs identified in the jointly called exome chip project are available for download on the CHARGE public website. Genotypes ascertained with the CHARGE jointly called exome chip cluster file (Dataset C) were 99.77% concordant with sequence data, 0.14% were missing in exome chip data, in the exome sequence data, or both, and 0.09% were discordant (Figure 1). Heterozygotes in Dataset C were most often misclassified when compared to the common allele homozygote, and mismatches were attributed equally to both sequencing and genotyping (Table 2).

We also tested the ability of the CHARGE exome chip cluster file to accurately assign genotypes in the three rarest variant bins: singletons (minor allele count = 1), doubletons (minor allele count = 2), and tripletons (minor allele count = 3). We observed high concordance between exome chip singletons (99.99%), doubletons (99.98%), and tripletons (99.97%) when compared to their respective sequence genotypes in the same 530 ARIC individuals previously described (data not shown). These results are consistent with the global concordance tests which suggest we are able to accurately call very rare variants.

Discussion

The results presented here demonstrate that rare variants on the exome chip can be accurately called when using a large, combined cluster file and best practices described when compared to existing clustering algorithms and rare variant calling methods. The joint calling protocol, accompanying cluster file, list of poor performing variants on the chip, and annotation data are a valuable resource for the scientific community and will be of great utility to those having smaller sample sets where the calling of rare variants is problematic. All new projects will require user decisions based on their own cohort data and the metrics and best practices presented here should be updated accordingly.

Materials and Methods

Subjects

Data from 62,266 participants from the following eleven studies in the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium [3] were included in this joint calling experiment and study descriptions were published previously: Age, Gene/Environment, Susceptibility-Reykjavik (AGES) Study [8], Atherosclerosis Risk in Communities (ARIC) Study [9], Cardiac Arrest Blood Study (CABS) [10], Cardiovascular Health Study (CHS) [11,12], Coronary Artery Risk Development in Young Adults (CARDIA) [13,14], Multi-Ethnic Study of Atherosclerosis (MESA) [15], Family Heart Study (FamHS) [16], Framingham Heart Study (FHS) [17], Health, Aging, and Body Composition (HABC) Study [18], Jackson Heart Study (JHS) [19], and the Rotterdam Study (RS) [20–23]. In addition, we genotyped 96 unrelated HapMap samples (48 CEU and 48 YRI) with each cohort and the list of sample IDs are available as a reference on the CHARGE exome chip public website.

Ethics Statement

All subjects provided written and informed consent to participate in genetic studies, and all study sites received approval to conduct this research from their local respective Institutional Review Boards (IRB) as follows: "The National Bioethics Committee" and "The Data Protection Authority" (AGES); University of Mississippi Medical Center IRB (ARIC – Jackson Field Center), Wake Forest University Health Sciences IRB (ARIC – Forsyth County Field Center), University of Minnesota IRB

Table 2. Results of missing data, genotype discordance, uncertainty coefficients and frequencies of exome chip data ascertained by three calling methods and compared to exome sequence genotypes.

Exome Sequence	Exome Chip					Missing	Discordance	Uncertainty
Genotypes	**Genotypes**					**(%)**	**(%)**	**Coefficient**
Dataset I	**AA**	**AB**	**BB**	**XX**	**Total**			
AA	94,878,501	33,679	4,467	183,952	95,100,599			
AB	41,395	2,350,644	4,777	15,967	2,412,783			
BB	3,658	4,642	495,626	1,809	505,735			
XX	89,104	2,905	711	1,233	93,953			
Total	95,012,658	2,391,870	505,581	202,961	98,113,070	0.30	0.09	0.864
Dataset Z	**AA**	**AB**	**BB**	**XX**	**Total**			
AA	94,964,611	115,849	4,394	15,745	95,100,599			
AB	41,864	2,365,462	5,137	320	2,412,783			
BB	3,557	5,635	496,351	192	505,735			
XX	89,480	2,996	714	763	93,953			
Total	95,099,512	2,489,942	506,596	17,020	98,113,070	0.11	0.18	0.912
Dataset C	**AA**	**AB**	**BB**	**XX**	**Total**			
AA	95,023,653	33,442	4,430	39,074	95,100,599			
AB	41,850	2,363,664	3,969	3,300	2,412,783			
BB	3,606	4,646	496,897	586	505,735			
XX	89,391	2,930	706	926	93,953			
Total	95,158,500	2,404,682	506,002	43,886	98,113,070	0.14	0.09	0.934

A total of 185,119 variants were used for these analyses, excluding duplicated variants, short insertion/deletions, XY chromosome SNPs, Y chromosome SNPs, mitochondrial SNPs, sites not identified in the exome sequencing dataset, and failing SNPs as identified by the CHARGE best practices guidelines. Genotype classes are represented as AA = common variant homozygote, AB = heterozygote, BB = rare variant homozygote, and XX = missing data. Dataset I: exome chip genotypes called with Illumina cluster file. Dataset Z: zCall assigned genotypes to missing data in Dataset I. Dataset C: exome chip genotypes called with the CHARGE cluster file.
doi:10.1371/journal.pone.0068095.t002

(ARIC – Minnesota Field Center), and Johns Hopkins University (Bloomberg School of Public Health) IRB (ARIC – Washington County Field Center); University of Washington IRB (CABS); Wake Forest University Health Sciences IRB (CHS – Forsyth County Field Center), University of California, Davis IRB (CHS – Sacramento County Field Center), Johns Hopkins University (Bloomberg School of Public Health) IRB (CHS – Washington County Field Center), and University of Pittsburgh IRB (CHS – Pittsburgh Field Center); University of Alabama at Birmingham (CARDIA – Birmingham Field Center), Northwestern University IRB (CARDIA – Chicago Field Center), University of Minnesota IRB (CARDIA – Minneapolis Field Center), and Kaiser Permanente IRB (CARDIA – Oakland Field Center); Washington University IRB (FamHS); Boston University IRB (FHS); Wake Forest University Health Sciences IRB (HABC); University of Mississippi Medical Center IRB (JHS); Columbia University IRB (MESA – New York Field Center), Johns Hopkins University IRB (MESA – Baltimore Field Center), Northwestern University IRB (MESA – Chicago Field Center), University of California IRB (MESA – Los Angeles Field Center), University of Minnesota IRB (MESA – Twin Cities Field Center), Wake Forest University Health Sciences IRB (MESA – Winston-Salem Field Center) and the National Heart, Lung, and Blood Institute; Medisch Ethische Toetsings Commissie (METC) at the Erasmus Medical Center, and the Netherlands Ministry of Health, Welfare and Sport (VWS) (RS). Joint calling of the array data was approved by the Committee for the Protection of Human Subjects (CPHS) which serves as the IRB for the University of Texas Health Science Center at Houston.

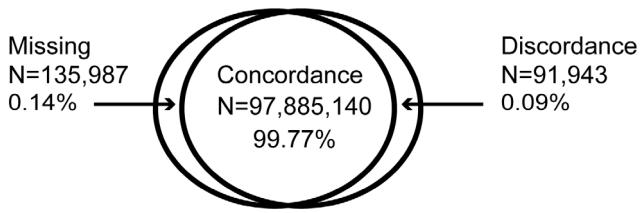

Figure 1. Results of CHARGE exome chip genotype calls compared to exome sequence data in 530 individuals.
doi:10.1371/journal.pone.0068095.g001

Genotyping

Study samples were processed on the HumanExome BeadChip v1.0 (Illumina, Inc., San Diego, CA) querying 247,870 variable sites described elsewhere (see Data Access) using standard protocols suggested by the manufacturer at the following seven genotyping centers: Broad Institute (JHS), Cedars-Sinai Medical Center (CHS, FamHS and MESA), Erasmus Medical Center (RS), Illumina Fast Track Services (FHS), University of Texas Health Science Center at Houston (AGES, ARIC and CARDIA), University of Washington (CABS), and Wake Forest University (HABC). Each center genotyped a common set of 96 HapMap samples to be utilized for quality control and determination of batch effects. The two channel raw data files (.idat) for all samples were transferred to a central location and assembled into a single project for joint calling. A summary of the samples genotyped within each cohort by race and gender is described in Table 3.

Table 3. Sample sizes of cohorts participating in joint calling effort by gender and self-reported race.

Cohort	African Americans		Caucasians		Hispanics		Asians		Other		HapMaps		Replicates			Total
	M	F	M	F	M	F	M	F	M	F	M	F	M	F	U[1]	
AGES	0	0	1,305	1,767	0	0	0	0	0	0	24	24	6	9	0	3,135
ARIC	1,121	1,832	5,198	5,873	0	0	0	0	0	0	77	76	62	90	200	14,529
CABS	283	172	3,701	1,174	0	0	0	0	0	0	57	59	93	29	0	5,568
CHS	318	526	2,008	2,603	0	0	1	3	14	13	48	46	25	34	0	5,639
CARDIA	900	1,185	1,063	1,189	0	0	0	0	0	0	48	48	18	30	0	4,481
FamHS	213	409	933	1,191	0	0	0	0	0	0	14	23	1	0	0	2,784
FHS	0	0	3,702	4,475	0	0	0	0	0	0	75	69	47	76	0	8,444
HABC	515	680	930	839	0	0	0	0	0	0	48	47	7	5	0	3,071
JHS	1,063	1,795	0	0	0	0	0	0	0	0	64	64	0	0	0	2,986
MESA	1,129	1,464	1,282	1,397	978	1,151	387	386	0	0	44	44	51	39	0	8,352
RS	0	0	1,459	1,720	0	0	0	0	0	0	47	47	0	0	4	3,277
Total	**5,542**	**8,063**	**21,581**	**22,288**	**978**	**1,151**	**388**	**389**	**14**	**13**	**546**	**547**	**310**	**312**	**204**	**62,266**

[1]Gender is unavailable for blinded replicates in the ARIC study and four RS samples.

doi:10.1371/journal.pone.0068095.t003

The following variables were provided for each sample included in the project: study specific sample ID, cohort name, sample type (DNA or WGA), race (self-reported), gender, sample plate, sample well, chip barcode, chip position, replicate ID, father and mother IDs (if applicable).

Clustering, Genotype Calling and Laboratory Quality Control

The Illumina GenomeStudio v2011.1 software was utilized with the GenTrain 2.0 clustering algorithm. Genomic DNA study samples and HapMap controls with call rates >99% (n = 55,142) were used to define genotype clusters with races combined and reruns excluded. The no-call threshold was set to 0.15 and we excluded female Y SNPs when calculating SNP statistics. The genotype quality score, representing the 10th percentile of the distribution of GenCall scores across all SNPs genotyped (p10GC), was visually examined in a scatter plot across all samples (Index vs. p10GC). Samples with an empirically determined p10GC <0.38 were identified as outliers and flagged for exclusion. The SNP parameters "Expected Number of Clusters of Y SNPs" and "Expected Number of Clusters of mtSNPs" were set to 2. Following automated clustering, all variants meeting the criteria provided in Table 4 (n = 107,175) were visually inspected and manually clustered, if possible, by two independent laboratory technicians. AA and BB theta deviation cutoffs were determined empirically. Variants removed from the HumanExome BeadChip v1.1 (n = 4,969) and cautious sites, as defined by the exome chip design committee (n = 333) (ftp://share.sph.umich.edu/exomeChip/IlluminaDesigns/cautiousSites/cautiousSite.sorted.sites), were also inspected. Samples with a call rate between 0.95 and 0.99 that had been previously excluded were brought back in to the project and re-inspected based on the criteria listed in Table 4. This additional review was necessary as the CHARGE exome chip project contains samples from multiple DNA sources and ethnicities that were genotyped at several centers. SNPs exhibiting obvious batch effects were excluded. After joint calling, reproducibility and heritability statistics, SNP statistics and sample statistics were updated and the SNP-level quality control criteria described in Table 5 were implemented. SNPs with reproducibility (rep) errors >2, parent-parent-child (PPC) error >1, or parent-child (PC) error >1 were not excluded, but were flagged and reported back to the participating studies for further investigation. A list of the 8,994 excluded variants is provided on the CHARGE exome chip website as cluster positions for these sites are zeroed out in the.egt file (note: all SNP statistics for these sites will be converted to zero when the cluster file is imported into the Genome Studio project). A portion of excluded SNPs may be recoverable in projects with a homogenous population substructure, and we recommend clustering and reviewing the subset of variants with the user's high quality samples. Table 6 describes the exome chip content and number of variants excluded by functional category (see Annotation). Importantly, the v1.0 cluster file should not be used for calling the Illumina v1.1 exome chip as the two versions were manufactured with different bead pools.

Exome Chip Performance

Genotypes derived from available exome sequencing of 540 ARIC participants were used as the comparison dataset to test the performance of the exome chip. We excluded 10 individuals from the sequencing dataset due to a high missing data rate <0.90, or non-overlap of individuals with existing exome chip data. Exome sequencing data is accessible via dbGaP as part of the National Heart Lung and Blood Institute (NHLBI) GO-ESP: Heart Cohorts Component of the Exome Sequencing Project (ARIC) (Study Accession: phs000398.v1.p1).

The following variants were excluded from the exome chip dataset as they were not available in the genotype data derived from exome sequencing results: replicate sites that were determined as triallelic or duplicates on opposite strands, short insertion/deletions, XY chromosome SNPs, Y chromosome SNPs, mitochondrial SNPs, or sites not identified in the exome sequencing dataset (n = 56,042). Poor performing variants identified by our best practices criteria were removed if not previously excluded (n = 6,709), thus a total of 185,119 variants were available for concordance analyses in 530 individuals.

Since concordance results are potentially high due to rare variation on the exome chip, we also calculated uncertainty coefficients [24] to determine the degree of association between each of the exome chip calling methods and exome sequence data. The uncertainty coefficient is a measure of association that is based

Table 4. Best practices criteria used to identify SNPs for visual inspection and manual reclustering.

Best Practices Criteria
All X, Y, XY and MT variants
Call frequency between 0.95 and 0.99
Cluster separation <0.4
AB frequency >0.6
AB R mean <0.2
Het excess >0.1
Het excess<−0.9
AA theta mean between 0.2 and 0.3
BB theta mean between 0.7 and 0.8
AB theta mean between 0.2 and 0.3
AB theta mean between 0.7 and 0.8
AA theta deviation >0.025
AB theta deviation ≥0.07
BB theta deviation >0.025
AB frequency = 0 and minor allele frequency >0
AA frequency = 1 and call rate <1
BB frequency = 1 and call rate <1
MAF <0.0001 and call rate ≠ 1
Rep error >2
PPC error >1
PC error >1
Variants removed from v1.1 exome chip
Cautious sites

AA: allele A homozygote; AB: heterozygote; BB: allele B homozygote; Het: heterozygote; MAF: minor allele frequency; MT: mitochondrial; PC: parent-child; PPC: parent-parent-child; R: normalized intensity; Rep, reproducibility.
doi:10.1371/journal.pone.0068095.t004

on information entropy [25], or the uncertainty in a random variable, that is, a variable subject to chance variations. Uncertainty coefficients are useful when evaluating results obtained from clustering algorithms since genotype classification is usually random (all minor alleles are not classified as either AA or BB), thus the algorithm is not susceptible to rare variation bias in which the more common genotype could have been called by chance alone. See Press et al. (1992), pp. 758–762, for further clarification of the uncertainty coefficient metric [26].

Annotation

Annotation of the v1.0 exome chip variants was performed with dbNSFP [27]. The dbNSFP v2.0 annotations are available on the CHARGE exome chip public website in the SNP information file. dbSNP rs information has been curated and a look up table with the associated Illumina SNP name is also available. The reason for inclusion of the variant on the exome chip by the design team is also provided in the SNP info file (ftp://share.sph.umich.edu/exomeChip/IlluminaDesigns/annotatedList.txt).

Data Access

The following CHARGE supporting documents are located at chargeconsortium.com/main/exomechip: CHARGE_ExomeChip_Best_Practices.pdf, CHARGE_ExomeChip_v1.0_Cluster_File.egt (cluster file for v1.0 chip), CHARGE_ExomeChip_v1.0_Excluded_Variants.txt (list of 8,994 zeroed out variants in

Table 5. Exome chip SNP exclusion criteria.

Exclusion Criteria
Call frequency <0.95 (except Y chr)
Cluster separation <0.4
AB frequency >0.6
AB R mean <0.2
Het excess >0.1
Het excess<−0.9
AA theta mean >0.3
BB theta mean <0.7
AB theta mean <0.2 or >0.8
AA theta deviation >0.06
AB theta deviation ≥0.07
BB theta deviation >0.06
Obvious batch effects

AA: allele A homozygote; AB: heterozygote; BB: allele B homozygote; Het: heterozygote; R: normalized intensity.
doi:10.1371/journal.pone.0068095.t005

Table 6. Exome chip content and CHARGE excluded variants by functional category.

Category[1]	Total Variants	Variants Excluded
exonic;stopgain	5,193	145
exonic;splicing;stopgain	90	1
exonic;stoploss	239	2
exonic;splicing;stoploss	5	0
splicing	2,263	60
exonic;splicing;synonymous	3,363	74
exonic;splicing	70	1
exonic;splicing;nonsynonymous	5,237	105
exonic;nonsynonymous	208,779	7,369
exonic;synonymous	6,415	281
UTR3	518	46
UTR5	77	6
ncRNA_splicing	1	1
ncRNA_exonic	111	8
ncRNA_UTR3	8	0
ncRNA_UTR5	1	0
intronic	5,762	254
ncRNA_intronic	447	23
downstream	187	19
upstream	181	7
upstream;downstream	8	0
intergenic	8,549	528
indel	137	10
mitochrondrial	226	54
no annotation	3	0
Total	247,870	8,994

[1]dbNSFP was used for annotating variants [27] (see Methods).
doi:10.1371/journal.pone.0068095.t006

cluster file), CHARGE_ExomeChip_SNP_Info_File.tsv.txt and Read Me file includes Illumina annotation, dbNSFP annotation, dbSNP rs numbers, overlapping sites between the HumanExome BeadChip v1.0 and v1.1, reason for inclusion, and race specific allele frequencies for each variant, including HapMap controls. Sample identifiers (CHARGE_ExomeChip_HapMap96_Control_List.csv) and genotypes for the 96 unrelated HapMap controls (CHARGE_ExomeChip_HapMap96_Genotype_Data.csv) are also available.

The Illumina genotyping protocol (Infinium_Best_Practices_370-2009-010.pdf) and cluster file (HumanExome-12v1.egt) are available with a MyIllumina login at https://icom.illumina.com/. The exome chip content data sheet is publicly available at http://www.illumina.com/documents/products/datasheets/datasheet_humanexome_beadchips.pdf.

zCall is a rare variant caller for array-based genotyping provided by Goldstein et al. and available for download at github.com/jigold/zCall [7]. PLINK is a freely available analysis toolset at http://pngu.mgh.harvard.edu/purcell/plink/ [4].

Acknowledgments

We thank all participating cohorts and institutions for their collaboration in this large-scale effort, and acknowledge the important role of the CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) consortium in the development and support of this manuscript. We would also like to recognize the following individuals at UT for their expertise and participation in the genotype calling, data management and analyses, respectively: Irina Strelets, Genesis Williams, and Kim Lawson. Also, we are thankful to Jennifer Brody at the University of Washington for her contributions to the exome chip supporting documentation.

The CHARGE investigators request that publications resulting from these exome chip data also cite their original publication: Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, Rotter JI, Uitterlinden AG, Harris TB, Witteman JC, Boerwinkle E; CHARGE Consortium. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from five cohorts. Circ Cardiovasc Genet 2:73-80, 2009.

Author Contributions

Conceived and designed the experiments: JCB KDT IBB LAC MF VG SK DL YL BMP SSR DSS AVS AU CMVD JGW CJO JIR EB. Performed the experiments: MLG BJC TH MH YL RK FR. Analyzed the data: MLG BY EB. Contributed reagents/materials/analysis tools: KDT IBB MF VG SK YL GMP BMP SSR DSS AVS AU CMVD JGW CJO JIR EB. Wrote the paper: MLG BY BJC TH JCB KDT MH IBB LAC MF VG TBH SK RK LJL DL YL TM GMP BMP SSR FR DSS AVS AU CMVD JGW CJO JIR EB.

References

1. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, et al. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 40: 1253–1260.
2. Ritchie ME, Liu R, Carvalho BS, Irizarry RA (2011) Comparing genotyping algorithms for Illumina's Infinium whole-genome SNP BeadChips. BMC Bioinformatics 12: 68.
3. Psaty BM, O'Donnell CJ, Gudnason V, Lunetta KL, Folsom AR, et al. (2009) Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium: design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet 2: 73–80.
4. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559–575.
5. Gudnason V, Sigurdsson G, Nissen H, Humphries SE (1997) Common founder mutation in the LDL receptor gene causing familial hypercholesterolaemia in the Icelandic population. Hum Mutat 10: 36–44.
6. Thorlacius S, Olafsdottir G, Tryggvadottir L, Neuhausen S, Jonasson JG, et al. (1996) A single BRCA2 mutation in male and female breast cancer families from Iceland with varied cancer phenotypes. Nat Genet 13: 117–119.
7. Goldstein JI, Crenshaw A, Carey J, Grant G, Maguire J, et al. (2012) zCall: a rare variant caller for array-based genotyping. Bioinformatics 28: 2543–2545.
8. Harris TB, Launer LJ, Eiriksdottir G, Kjartansson O, Jonsson PV, et al. (2007) Age, Gene/Environment Susceptibility-Reykjavik Study: multidisciplinary applied phenomics. Am J Epidemiol 165: 1076–1087.
9. The ARIC Investigators (1989) The Atherosclerosis Risk in Communities (ARIC) study: design and objectives. Am J Epidemiol 129: 687–702.
10. Siscovick DS, Raghunathan TE, King I, Weinmann S, Wicklund KG, et al. (1995) Dietary intake and cell membrane levels of long-chain n-3 polyunsaturated fatty acids and the risk of primary cardiac arrest. JAMA 274: 1363–1367.
11. Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, et al. (1991) The Cardiovascular Health Study: design and rationale. Ann Epidemiol 1: 263–276.
12. Tell GS, Fried LP, Hermanson B, Manolio TA, Newman AB, et al. (1993) Recruitment of adults 65 years and older as participants in the Cardiovascular Health Study. Ann Epidemiol 3: 358–366.
13. Friedman GD, Cutter GR, Donahue RP, Hughes GH, Hulley SB, et al. (1988) CARDIA: study design, recruitment, and some characteristics of the examined subjects. J Clin Epidemiol 41: 1105–1116.
14. Cutter GR, Burke GL, Dyer AR, Friedman GD, Hilner JE, et al. (1991) Cardiovascular risk factors in young adults. The CARDIA baseline monograph. Control Clin Trials 12: 1S–77S.
15. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, et al. (2002) Multi-ethnic study of atherosclerosis: objectives and design. Am J Epidemiol 156: 871–881.
16. Higgins M, Province M, Heiss G, Eckfeldt J, Ellison RC, et al. (1996) NHLBI Family Heart Study: objectives and design. Am J Epidemiol 143: 1219–1228.
17. Dawber TR, Meadors GF, Moore FE, Jr. (1951) Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health Nations Health 41: 279–281.
18. Park SW, Goodpaster BH, Strotmeyer ES, de Rekeneire N, Harris TB, et al. (2006) Decreased muscle strength and quality in older adults with type 2 diabetes: the Health, Aging, and Body Composition Study. Diabetes 55: 1813–1818.
19. Taylor HA, Jr., Wilson JG, Jones DW, Sarpong DF, Srinivasan A, et al. (2005) Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. Ethnic Dis 15: S6-4-17.
20. Hofman A, Grobbee DE, de Jong PTVM, van den Ouweland FA (1991) Determinants of disease and disability in the elderly: the Rotterdam Study. Eur J Epidemiol 1991: 403–422.
21. Hofman A, Breteler MMB, van Duijn CM, Krestin GP, Pols HA, et al. (2007) The Rotterdam Study: objectives and design update. Eur J Epidemiol 22: 819–829.
22. Hofman A, Breteler MMB, van Duijn CM, Janssen HL, Krestin GP, et al. (2009) The Rotterdam Study: 2010 objectives and design update. Eur J Epidemiol 24: 553–572.
23. Hofman A, van Duijn CM, Franco OH, Ikram MA, Janssen HL, et al. (2011) The Rotterdam Study: 2012 objectives and design update. Eur J Epidemiol 26: 657–686.
24. Mills P (2011) Efficient statistical classification of satellite measurements. Int J Remote Sens 32: 6109–6132.
25. Cover TM, Thomas JA (2006) Elements of Information Theory, 2nd Edition. Hoboken, NJ: John Wiley & Sons, Inc. 748 p.
26. Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1992) Numerical Recipes: the Art of Scientific Computing, 3rd Edition. New York, NY: Cambridge University Press. 1235 p.
27. Liu X, Jian X, Boerwinkle E (2011) dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32: 894–899.