PLOS ONE

RESEARCH ARTICLE

Genome-Wide Association of Copy Number Polymorphisms and Kidney Function

Man Li[1,2], Jacob Carey[1], Stephen Cristiano[3], Katalin Susztak[4], Josef Coresh[1,5], Eric Boerwinkle[6], Wen Hong L. Kao[1,5†], Terri H. Beaty[1], Anna Köttgen[1,7], Robert B. Scharpf[8]*

1 Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America, **2** Division of Nephrology and Hypertension, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, Utah, United States of America, **3** Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, United States of America, **4** Renal Electrolyte and Hypertension Division, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States of America, **5** Welch Center for Prevention, Epidemiology and Clinical Research, Baltimore, Maryland, United States of America, **6** Human Genetics Center, University of Texas Health Science Center at Houston, Houston, Texas, United States of America, **7** Division of Genetic Epidemiology, Medical Center–University of Freiburg, Faculty of Medicine, Freiburg, Germany, **8** Department of Oncology, Johns Hopkins School of Medicine, Baltimore, Maryland, United States of America

† Deceased.
* rscharpf@jhu.edu

Citation: Li M, Carey J, Cristiano S, Susztak K, Coresh J, Boerwinkle E, et al. (2017) Genome-Wide Association of Copy Number Polymorphisms and Kidney Function. PLoS ONE 12(1): e0170815. doi:10.1371/journal.pone.0170815

Editor: Giuseppe Remuzzi, Istituto Di Ricerche Farmacologiche Mario Negri, ITALY

Received: June 21, 2016

Accepted: January 11, 2017

Published: January 30, 2017

Data Availability Statement: All primary data (CEL files) are available from dbGaP (accession number phs000090.v1.p1): https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000090.v1.p1&phv=22859&phd=2066&pha=&pht=114&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1.

Funding: The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C,

Abstract

Genome-wide association studies (GWAS) using single nucleotide polymorphisms (SNPs) have identified more than 50 loci associated with estimated glomerular filtration rate (eGFR), a measure of kidney function. However, significant SNPs account for a small proportion of eGFR variability. Other forms of genetic variation have not been comprehensively evaluated for association with eGFR. In this study, we assess whether changes in germline DNA copy number are associated with GFR estimated from serum creatinine, eGFRcrea. We used hidden Markov models (HMMs) to identify copy number polymorphic regions (CNPs) from high-throughput SNP arrays for 2,514 African (AA) and 8,645 European ancestry (EA) participants in the Atherosclerosis Risk in Communities (ARIC) study. Separately for the EA and AA cohorts, we used Bayesian Gaussian mixture models to estimate copy number at regions identified by the HMM or previously reported in the HapMap Project. We identified 312 and 464 autosomal CNPs among individuals of EA and AA, respectively. Multivariate models adjusted for SNP-derived covariates of population structure identified one CNP in the EA cohort near genome-wide statistical significance (Bonferroni-adjusted p = 0.067) located on chromosome 5 (876–880kb). Overall, our findings suggest a limited role of CNPs in explaining eGFR variability.

Introduction

Chronic kidney disease (CKD) is defined by reduced kidney function or kidney damage and can progress over time. It is estimated that CKD effects about 26 million US adults[1] and its

HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research. Funding support for "Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium" was provided by the NIH through the American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419). ML was supported by a National Heart, Lung, and Blood Institute T32-HL0072024 Cardiovascular Epidemiology Training Grant. AK was supported by the Emmy Noether Program (KO 3598/2-1) of the German Research Foundation.

Competing Interests: The authors have declared that no competing interests exist.

prevalence is increasing both in the United States and globally[1, 2]. CKD has a genetic component, which may contribute to the development or progression of CKD in addition to or through established epidemiological CKD risk factors such as hypertension, diabetes, and proteinuria. Genome-wide association studies limited to single nucleotide polymorphisms (SNPs) have identified more than 50 loci associated with estimated glomerular filtration rate (eGFR), a measure of kidney function, in European ancestry (EA) populations[3–5]. These loci account for only 3% of the variation in eGFR[3], while the heritability of eGFR has been estimated at 33%–75% [6–9]. We hypothesized that heritable loss and gain of germline DNA copy number may contribute to kidney function.

Glomerular filtration rate (GFR) is considered the best quantitative measure of kidney function. However, gold standard estimates of GFR by urinary or plasma clearance of exogenous filtration markers are too cumbersome and expensive for use in clinical and research settings. Serum creatinine has emerged as a reliable indicator of GFR[10] and is currently the most commonly used biomarker of kidney function.

Copy number variants (CNVs) have been reported to encompass genes involved in the regulation of cell growth and metabolism, implicating vital roles in the variability of human traits and disease risk[11–13]. Copy number polymorphisms (CNPs), regions of the genome where CNVs segregate in the germline at a population frequency of at least 2 percent, have been implicated in a broad range of human diseases, including mental health (bipolar disease[14], schizophrenia[15–17], and autism spectrum disorder[18, 19]), metabolic disease (type I diabetes[20] and obesity[21–23]), congenital anomalies (kidney and urinary tract defects[24–26], oral clefts[27–29]), and cancer (breast cancer[30], melanoma[31], and colorectal cancer[32]). While previous SNP association studies have identified several loci strongly associated with kidney disease, the causal variants are generally not known. CNPs occurring at known risk loci may help establish a genetic basis for this disease. For example, a deletion of any part of a gene or its promoter can disrupt transcription to mRNA; amplification of a gene can lead to over expression of the mRNA product. In Scharpf *et al.*, 2014[33], we showed the association between copy number and uric acid levels at *SLC2A9* was independent of nearby SNP-association signals. As many array platforms include probes targeting regions of the genome that are monomorphic at the single nucleotide level (i.e. there is only one allele at the probe), the discovery of risk loci not well tagged by SNPs is also possible. The array platform used in ARIC has approximately 1 million monomorphic probes in addition to 1 million polymorphic (SNP) markers, enabling identification of CNPs in regions not well tagged by SNPs.

Here, we use Bayesian Gaussian Mixture Models (GMMs) to estimate copy number at polymorphic ($>$ 2% of subjects) regions identified by a Hidden Markov Model (HMM) and regions previously reported as polymorphic in the HapMap Project[34]. We evaluated models for eGFR by serum creatinine (eGFRcrea) levels that include copy number at CNP regions as a covariate. Findings presented here are the most comprehensive analyses to date of CNPs and quantitative measures of kidney function in an adult population. Further, this is the first study to present genomic analyses of copy number for individuals of AA in the ARIC study.

Results

The ARIC study includes 9,483 EA and 2,822 AA participants with both baseline eGFRcrea and Affymetrix 6.0 genotype data. The EA and AA cohorts differ in known clinical risk factors for kidney disease and eGFRcrea levels (Table 1). In particular, the percentage of AA participants with hypertension was 56.5, more than twice the percentage among EA participants (26.6 percent). Similarly, the percentage of AA with diabetes was 19.3 percent compared to

Table 1. Study sample characteristics. Descriptive statistics are shown as mean and (standard deviation) unless otherwise indicated.

	European ancestry	African ancestry
Sample size eGFRcrea/eGFRcys	8645/6843	2514/1673
Women, N (%)	4592 (53.1)	1576 (62.7)
Age (years)	54.2 (5.7)	53.5 (5.8)
Center N (%)	F 2606 (30.1)	F 288 (11.5)
	J 0 (0)	J 2226 (88.5)
	M 3226 (37.3)	M 0 (0)
	W 2813 (32.5)	W 0 (0)
eGFRcrea (ml/min/1.73m^2)	89.8 (18.0)	103.2 (25.0)
eGFRcys (ml/min/1.73m^2)	84.3 (19.6)	91.7 (24.9)
HTN, N (%)	2288 (26.6)	1413 (56.5)
DM, N (%)	745 (8.6)	484 (19.3)

doi:10.1371/journal.pone.0170815.t001

only 8.6 percent in the EA cohort. The average eGFRcrea among AA participants was 103.2 ml/min/1.73m^2, compared to an average of 89.8 ml/min/1.73m^2 among EA subjects.

To identify CNVs, we fit a 6-state HMM for all participants passing previously established quality control steps[35]. Additional statistics for quality control in this study include the median absolute deviation (MAD) and lag-10 autocorrelation of autosomal $\log_2$ R ratios (LRRs) (Figures A and B in S1 File). On average, the HMM identifies more CNVs in participants of AA than EA with median frequencies of 68 and 57, respectively. Among EA participants, approximately 10% of the CNVs span fewer than 10 SNPs or monomorphic markers and were excluded from further analysis. Of the remaining CNVs, approximately 64% (429,162) occur at regions that are copy number altered in 2 percent or more of the EA or AA participants (e.g., Fig 1A). Hereafter, we refer to these regions as CNPs.

A major challenge for identifying CNVs is the substantial intra-subject variance of the LRRs. As we and others have demonstrated[34–36], the signal to noise ratio can be improved by modeling the distribution of only the markers involved in the polymorphism across many subjects. To comprehensively identify CNPs, including those too small to be estimated by the HMM, we evaluated 785 additional regions reported as polymorphic in HapMap[34] and spanning 3 or more Affymetrix 6.0 markers. For HapMap regions that overlap the HMM-derived CNPs, we used the genomic coordinates from the HMM.

For each candidate CNP, maximum *a posteriori* estimates of relative copy number were obtained from a GMM implemented in the R package cnvCall[37]. Excluding monomorphic regions, we identified 312 and 464 autosomal polymorphic regions in the EA and AA cohorts, respectively (Fig 1C and Figure C in S1 File). After translating mixture component indices to copy number and manually recalling rare homozygous deletions (see Methods), we found roughly 85 percent of the deletion CNPs in the EA and AA cohorts occur at frequencies consistent with Hardy Weinberg equilibrium ($p > 0.01$; Figure D in S1 File).

To contrast the CNP regions by methodology (HapMap or HMM), we assessed the extent to which the CNP regions overlap. Interestingly, 15% of the EA CNPs (n = 46) and 13% of the AA CNPs (n = 60) did not overlap with published regions in HapMap (Fig 2A and 2B). To evaluate whether the CNPs identified by only the HMM (not reported in HapMap) were common in other studies, we examined 17 studies each having at least 100 subjects deposited in the Database of Genomic Variants as of May 15, 2016 (http://dgv.tcag.ca, NCBI build 36). With few exceptions, nearly all of these CNPs were in one or more of these studies at a frequency of 2 percent or more (Figure E in S1 File).

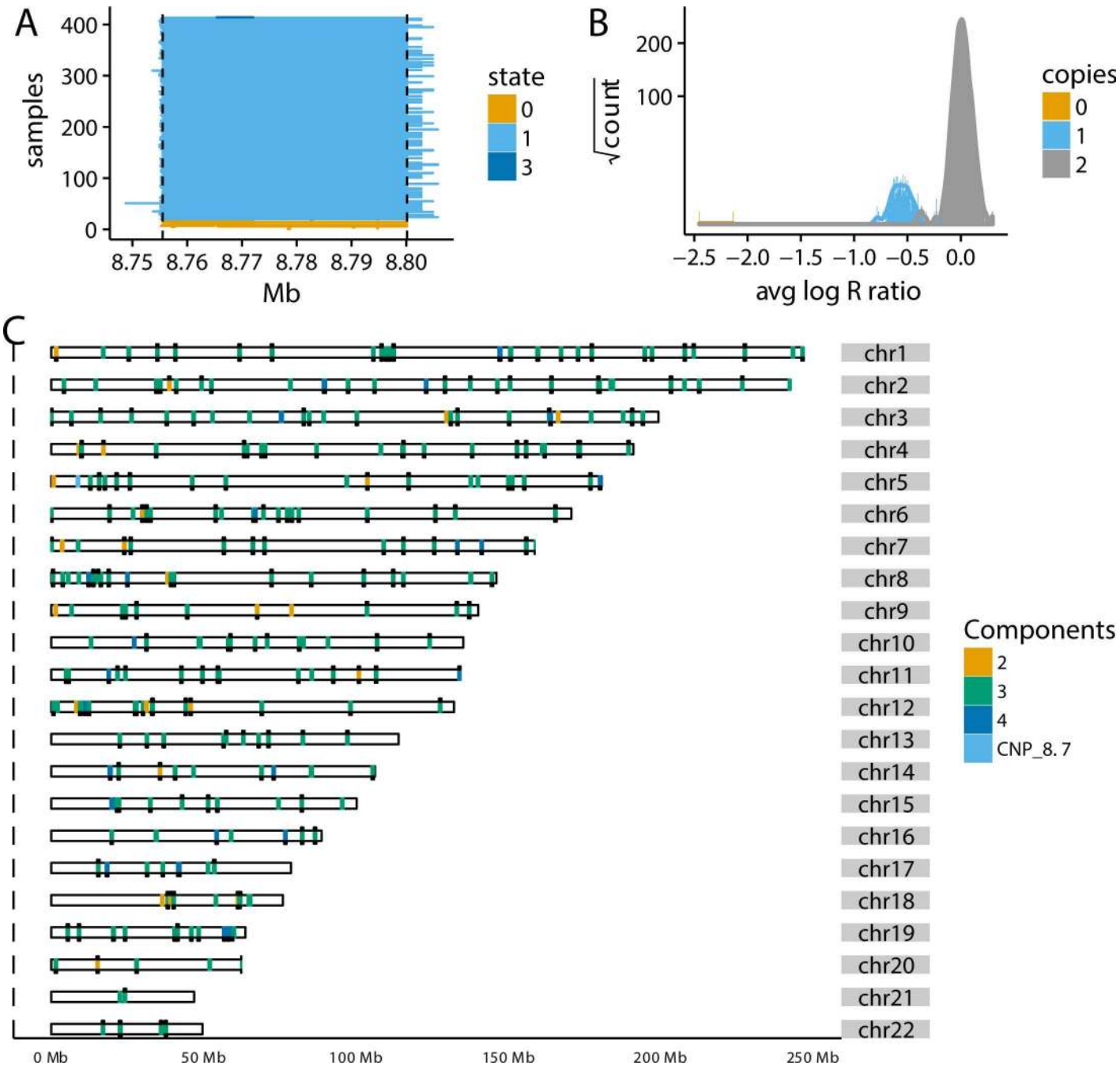

Fig 1. Developing a profile of autosomal CNP regions in ARIC. (A) CNVs identified from the HMM often have similar genomic endpoints across samples, shown here as colored rectangles for ~450 EA participants at a region on chromosome 5 (top signal in EA analysis). (B) The distribution of the average for 8,645 EA participants at the region on chromosome 5 approaching genome-wide significance. Copy number is called by the maximum *a posteriori* estimates from a normal mixture model. (C) All autosomal CNP regions identified either by HapMap or from the HMM among EA participants color-coded by the number of copy number states. Black ticks above the ideograms are additional regions from HapMap identified as polymorphic by the GMM in ARIC. Black ticks below the ideograms are CNPs that are also present in the AA cohort (see also Figure C in S1 File). The region on chromosome 5p is highlighted.

doi:10.1371/journal.pone.0170815.g001

Comparing CNP coordinates by ancestry, we find that only 215 of the 464 regions identified in the AA cohort are also polymorphic in the EA cohort (Fig 2C). As the ratio of EA to AA

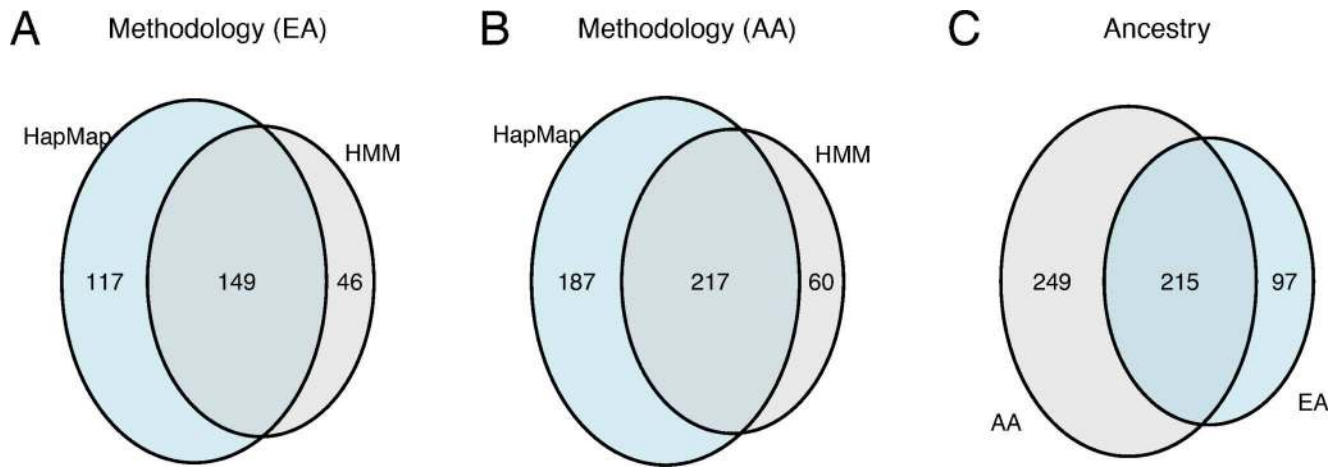

Fig 2. Overlap of CNP regions by methodology of identification and ancestry. (A, B) The number of CNPs identified by methodology for EA (left) and AA ancestry (middle). (C) The overlap of polymorphic regions by ancestry. The EA and AA cohorts shared 215 CNP regions.

doi:10.1371/journal.pone.0170815.g002

participants in ARIC was more than 3:1, an unstratified analysis would have failed to identify regions absent among EA participants and occurring in less than 10 percent of AA participants.

To assess whether copy number at identified CNP regions is associated with eGFRcrea, we evaluated linear models with log-transformed eGFRcrea as the dependent variable with copy number and SNP-derived covariates of population structure as covariates (Methods). Because global differences in allele frequencies may exist between subpopulations particularly within the AA cohort, we estimated the percentage of African ancestry by ANCESTRYMAP[38] using genotypes from 1401 ancestry-informative markers in 2,152 AA participants. For the AA cohort, none of the AA CNPs are statistically significant with or without global ancestry adjustment (Fig 3A). For the EA cohort, none of the models are statistically significant after Bonferroni-correction (Fig 3B), though CNP_8.7 (chr5: 8,755,522–8,800,142 bp) is suggestive (adjusted p = 0.067).

CNP_8.7 is a deletion polymorphism located in a gene desert[39, 40] and is associated with a 0.04 increase in log eGFRcrea (95% CI: 0.02–0.06 ml/min/1.73 m^2). The deletion allele is interrogated by 40 monomorphic markers and 7 SNPs on the Affymetrix platform and overlaps deletions previously identified in HapMap[34]. The deletion allele segregates in the EA population at Hardy Weinberg equilibrium (p = 0.4). In particular, 393 (4.5%) hemizygous deletions, 2 (0.02%) homozygous deletions, and 8,267 (95.6%) with diploid copy number were identified.

To confirm copy number estimates at CNP_8.7 with an alternative technology, we obtained next generation sequencing data from dbGaP for five subjects with a putative hemizygous deletion (accession number phs000090.v1.p1). Preprocessing the read depth in 10kb bins to adjust for GC-content and excluding regions of low mappability (see Methods), we find that all five samples have two or more 10kb bins in the region with log ratios of relative copy number less than -1 (Figure F in S1 File). Further, the three samples with lowest technical variation in the sequencing platform (F159225, F264060, and W156974) have approximately the same boundaries as identified by the array platform.

In the absence of any stronger candidate than CNP_8.7 to pursue for replication in an independent study, we evaluated an alternative GFR surrogate. In particular, we hypothesized bona fide modulation of latent GFR by copy number dosage would be captured by multiple GFR surrogates. As cystatin C is also available in ARIC and well regarded as a surrogate for calculating

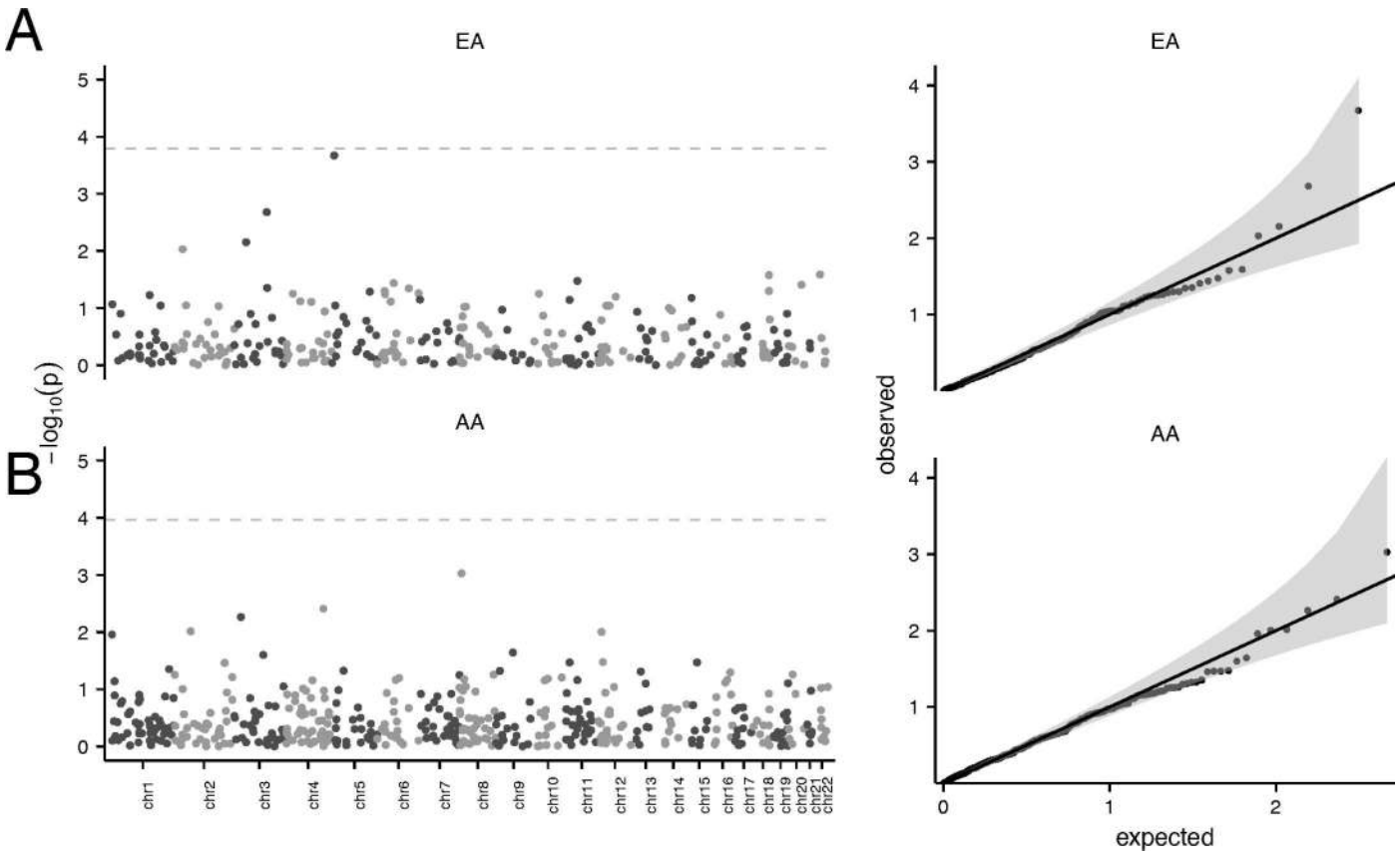

Fig 3. Statistical significance of copy number in linear regression models for eGFRcrea. (A) Manhattan plot for CNP association analysis in eGFRcrea among 8,645 European ancestry and 2,514 AA participants in the ARIC study. The gray line indicates genome-wide statistical significance. (B) Quantile-quantile plots of the expected–log 10 p-values under the null hypothesis of no association versus the observed–log 10 p-values. The lower and upper bounds of the shaded region indicate 0.025 and 0.975 quantiles, respectively, of the null.

doi:10.1371/journal.pone.0170815.g003

eGFR[41], we evaluated the same regression model as described previously with eGFR by cystatin C (eGFRcys) as the dependent variable. For the 6,830 subjects with available eGFRcys, subjects with one less copy of CNP_8.7 had a very modest 0.01 increase of log eGFRcys (95% CI: -0.02–0.04 ml/min/1.73 m^2; p = 0.54; n = 6,854). To assess empirically whether the qualitative difference in interpretation of the eGFRcrea and eGFRcys models could be attributable to lack of statistical power in the latter, we re-evaluated the eGFRcrea model using only the 6,854 participants with eGFRcys measurements available. Our findings in the restricted data set are qualitatively similar to the full dataset (effect size = 0.04, p = 0.098).

Discussion

We implemented a genome-wide association study of CNPs and eGFRcrea in two large EA and AA cohorts represented in ARIC. We identified 312 and 464 CNPs among EA and AA participants, respectively (S1 and S2 Tables). For each CNP, we evaluated copy number in the context of multivariate models for eGFRcrea including known risk factors of kidney disease and principal component-derived surrogates for subpopulation strata in ARIC. While our findings revealed no genome-wide statistically significant associations between copy number and eGFRcrea in either the EA or AA sub-population, we identified one region in the EA cohort close to genome-wide statistical significance (Bonferroni-adjusted p = 0.053). However,

we found no evidence to support our expectation that bona fide modulation of latent GFR by copy number dosage would be captured by the alternative GFR surrogate, eGFRcys (p = 0.54). We caution that our secondary analysis using eGFRcys merely provides additional context for the interpretation of the borderline association observed at CNP_8.7 using existing data in ARIC. A definitive analysis of the biological significance of CNP_8.7 (or lack thereof) would require replication in an independent study.

This study extends previous work characterizing copy number variation in ARIC. First, the profile of CNPs in the EA cohort now includes published CNP regions too small for detection by HMMs but estimable by GMMs. Secondly, we provide the first genome-wide profile of CNPs among AA participants in ARIC. In both the EA and AA CNP profiles, the HMM- and HapMap-derived regions were confirmed by GMMs that explicitly model between subject variation of one-dimensional LRR summaries. Finally, we show that 13–16% of the CNPs identified in the EA and AA cohorts would not have been identified by using HapMap information alone. For populations less well-characterized by consortium efforts such as 1000 Genomes and HapMap, these percentages are likely to increase.

In summary, our study does not support a link between CNPs and kidney function as measured by estimated GFR. Nearly 30% of the CNVs identified in this study occur outside of CNPs. The statistical power to detect effect sizes of rare CNVs is limited, particularly in the AA cohort (Figure G in S1 File). Pathway-based analyses and/or meta-analysis of multiple cohorts to study the contribution of rare CNVs in EA and AA subpopulations to CKD require further investigation.

Methods

Study population

The ARIC Study is a prospective observational cohort study with participants aged between 45 and 64 at the baseline visit (visit 1) occurring between 1987 and 1989. The participants were recruited from 4 US communities: Forsyth County, North Carolina (F); Jackson, Mississippi (J); suburban Minneapolis, Minnesota (M); and Washington County, Maryland (W). After enrollment, there were three follow-up visits approximately every three years (1990–92, 1993–95, 1996–98). A fifth visit was completed in 2011–2013. Details of the study design have been reported previously[42]. All study participants provided written informed consent, and the study protocol was approved by the Johns Hopkins Bloomberg School of Public Health Institutional Review Board.

Measurements

In the ARIC study, serum creatinine is available at visits 1, 2, and 4. We used serum creatinine measurements for the visit with the largest participation, visit 1 (n = 15,792). Serum creatinine levels were measured using the modified kinetic Jaffe method and calibrated to the age-, sex-, and race-specific means in the Third National Health and Nutrition Examination Survey (NHANES III). We estimate GFR based on serum creatinine (eGFRcrea) using the Modification of Diet in Renal Disease (MDRD) Study 4-variable equation[10].

Cystatin C, an alternative surrogate quanitative measure of eGFR, was measured by a particle enhanced immunonephelometric assay (N Latex Cystatin C, Dade Behring). The eGFRcys levels were estimated as eGFRcys = 76.7 x (serum cystatin C)$^{-1.19}$ [41]. Both eGFRcrea and eGFRcys were approximately log-normally distributed.

Diabetes was defined as fasting glucose $\geq$126 mg/dL, non-fasting glucose $\geq$200 mg/dL, self-reported physician diagnosis of diabetes mellitus or the use of oral hypoglycemic medication or insulin. Hypertension was defined as systolic blood pressure $\geq$ 140 mmHg, diastolic blood pressure $\geq$ 90 mmHg or the use hypertension treatment medication.

Genotyping and quality controls

Genomic DNA was extracted from peripheral whole blood and SNPs were genotyped on the Affymetrix 6.0 chip as described previously[43]. Genetic outliers, first-degree relatives, gender mismatches, and participants who did not consent for use of DNA information were excluded. To control for population stratification, we computed the first 10 principal components using EIGENSTRAT[44] using high quality, independent SNPs. Details of principal component generation have been previously described[45]. All 10 principal components were included as covariates in our statistical model. To summarize, 11,827 participants (9,038 EA and 2,822 AA) attended visit 1, had valid data on serum creatinine, and had genotype data meeting the above quality control criteria (Figure A in S1 File).

Overview of CNP estimation

To comprehensively identify all CNP regions among EA and AA participants, we pursued a two-step approach. First, we fit a HMM to each individual sample. The HMM allows identification of both CNVs and CNPs. While helpful for identifying CNPs in populations not well represented in public repositories, HMMs have a limited resolution that depends on the inherent marker-to-marker variation of the LRR estimates. As the signal to noise ratio is increased by examining only the markers involved in a CNP across a large collection of samples (e.g., McCarroll et al., 2008[34] and Cardin et al., 2011[37]), we examined the distribution of region-level copy number summaries across all ARIC samples in a second step. In particular, we evaluated all CNPs previously reported in HapMap spanning at least 3 Affymetrix 6.0 markers using a previously described Gaussian mixture model for CNPs[35].

Hidden Markov model. B-allele frequencies and wave-adjusted LRRs were computed as described previously[33]. Estimates for copy number states 0–4 for all autosomes were derived as previously described for the EA participants using the VanillaICE HMM[46, 47]. We required at least 10 markers in a CNP region identified by the HMM to reduce false positive identifications. We excluded 393 EA and 275 AA subjects for one or more of the following reasons: median absolute deviation of the LRRs greater than 0.35, autosomal lag 10 autocorrelation of the LRRs greater than 0.05, or more than 150 CNVs (Figure B in S1 File).

CNP regions. We refer to CNVs occurring in the population at a frequency of at least 2 percent as CNPs. CNPs tend to have the same or very similar breakpoints across individuals. As CNP regions are known to differ by ancestry, we defined consensus start and stop genomic positions for CNP regions independently for the EA and AA cohorts. Specifically, the consensus start (end) was defined as the minimum (maximum) base-pair spanned by at least half of all CNVs identified by the HMM at a particular polymorphic region. In addition to HMM-derived CNPs, we included 785 candidate CNP regions available from HapMap and reported by McCarroll et al (2008)[34]. For partially overlapping regions, we kept only one region (copy number estimates were nearly identical in each case), yielding 312 non-overlapping regions in EA and 464 non-overlapping regions in AA.

Gaussian mixture model. For each HapMap- or HMM-derived CNP candidate, we fit the GMM implemented in the R package cnvCall (Cardin et al., 2011[37]). Briefly, a one-dimensional summary for each sample was derived from the first principal component of the LRR matrix at a CNP (rows are samples and columns are the marker-level LRR). The marginal distribution of the one-dimensional summary (marginal across subjects) was modeled as a mixture of normal distributions. Since the number of mixture components, k, was not known *a priori*, models k = 1 to k = 5 were evaluated at each CNP. The model with the lowest Bayesian Information Criterion (BIC) was then selected. To assign a mixture component index to each sample, we used the maximum *a posteriori* estimate. If the maximum *a posteriori* probability

was less than max (0.2, 1/k), an NA, indicating missing, was recorded. A post-post-hoc merging procedure implemented in cnvCall was used to reduce overfitting skewed-normal distributions.

To translate cnvCall component indices into absolute copy number, we implemented a simple relabeling heuristic. Let I denote the mixture component index, $I \in \{1,\ldots,k\}$ and CN denote the copy number, $CN \in \{0,\ldots,4\}$. If the average LRR of the first component was less than -1.5 (consistent with a homozygous deletion), we set $CN = I\text{-}1$. For common deletions and $k \geq 3$, homozygous deletions often have mean LRR greater than -1.5. Therefore, if the mean of the first component was less than -0.5 and the distance between the first and second mode was 1.5-fold the distance between the second and third mode, we also set $CN = I\text{-}1$. If the first component is not homozygous (i.e. neither of the above criteria were met), we set $CN = 2$ for the modal component index. The remaining indices were set to $CN = 1$ or $CN = 3$ depending on whether the component index is less or greater than the modal component index, respectively.

From a statistical point of view, two challenging aspects of developing mixture models to estimate copy number are model selection (i.e., choosing the right k) and model robustness to assumptions of approximate normality. While cnvCall selects the model with the lowest BIC and has an outlier component to capture outliers, we found that for less common deletions, a model having only hemizygous and diploid components ($k = 2$) was selected over a model that includes a homozygous deletion component ($k = 3$). While the more parsimonious model may be preferable in some situations, here the more parsimonious $k = 2$ model is biologically implausible if the deletion allele is segregating in the population at Hardy Weinberg equilibrium (HWE), as expected. Having implemented the above relabeling heuristic, we identified all CNPs in which the first component was hemizygous deletions. For these CNPs, we set $CN = 0$ for any sample with an average LRR consistent with homozygous deletion ($<$ -1.5) but assigned NA to indicate missing by cnvCall.

Analysis of whole genome sequencing platform. We downloaded and preprocessed low-pass whole genome sequencing data for 5 ARIC samples (dbGaP accession number phs000090.v1.p1). Briefly, we realigned each BAM file to the NCBI build 36 reference genome using ELAND [36]. Next, we tiled the genome into 10kb non-overlapping bins and counted the number of reads aligning to each bin. We transformed the bin-counts to the $\log_2$ scale, and GC-corrected the bin counts using a loess scatterplot smoother with a span of 1/3.

Association analysis

Copy number was obtained from the relabeled cnvCall component indices and manually identified rare homozygous deletions (as described previously). Other covariates included age, sex, study site, and principal components derived from the SNP genotypes (described previously). Since eGFRcrea is approximately log-normally distributed and previous studies have used the log-transformed response (instead of a log-link), we evaluated linear models with the log-transformed measurements. All CNP analyses were stratified by ancestry. The genome-wide statistical significance level by Bonferroni was $p < 1.61\text{x}10^{-4}$ for the EA participants (0.05/312) and $p < 1.08\text{x}10^{-4}$ (0.05/464) for the AA participants.

Statistical power

We used simulation to estimate the statistical power for identifying copy number alterations associated with eGFRcrea levels. Briefly, we randomly sampled the copy number status for 8,645 EA participants assuming prevalence of a deletion allele ranging from rare (0.02, top-left of Figure G in S1) to common (0.2, bottom-right of Figure G in S1). Conditional on the copy number assignment, we added a value β and $2\times\beta$ to the empirical log(eGFRcrea) estimates for

individuals with 1 copy and 2 copies, respectively. We evaluated a range of values for β that include estimates from previously replicated SNP-association studies, as well as values observed in this study. For example, the average slope in a GWAS based on SNPs in the ARIC EA cohort using the same log(eGFRcrea) response was 0.02 [5]. For each simulated dataset, we fit a generalized linear model with log-link and calculated the Bonferroni-adjusted p-value of the copy number regression coefficient. Repeating the simulation 100 times for each combination of deletion prevalence and estimated β, the statistical power is the fraction of simulated datasets with adjusted $p < 0.05$. We repeated this simulation with a sample size of 2,514 for the AA cohort.

Genomic annotation and software versions

Genomic annotation in this paper is based on UCSC build hg18 (NCBI36). The version of R and of the R packages used in this analysis is included in S1 File.

Supporting Information

S1 File. Includes Figures A-G and software versions.
(DOCX)

S1 Table. Genomic coordinates and model summary statistics for EA CNPs.
(CSV)

S2 Table. Genomic coordinates and model summary statistics for AA CNPs.
(CSV)

Acknowledgments

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research. Funding support for "Building on GWAS for NHLBI-diseases: the U.S. CHARGE consortium" was provided by the NIH through the American Recovery and Reinvestment Act of 2009 (ARRA) (5RC2HL102419). ML was supported by a National Heart, Lung, and Blood Institute T32-HL0072024 Cardiovascular Epidemiology Training Grant. AK was supported by the Emmy Noether Program (KO 3598/2-1) of the German Research Foundation.

Author Contributions

Conceptualization: ML WHLK RBS.

Data curation: ML RBS.

Formal analysis: ML RBS.

Funding acquisition: WHLK RBS.

Methodology: ML RBS.

Project administration: WHLK RBS.

Resources: RBS.

Software: ML SC J. Carey RBS.

Supervision: EB J. Coresh.

Validation: ML RBS.

Visualization: ML RBS.

Writing – original draft: ML.

Writing – review & editing: ML RBS KS J. Carey J. Coresh AK THB.

References

1. Coresh J, Selvin E, Stevens LA, Manzi J, Kusek JW, Eggers P, et al. Prevalence of chronic kidney disease in the United States. JAMA: the journal of the American Medical Association. 2007; 298 (17):2038–47. doi: 10.1001/jama.298.17.2038 PMID: 17986697
2. El Nahas M. The global challenge of chronic kidney disease. Kidney international. 2005; 68(6):2918–29. doi: 10.1111/j.1523-1755.2005.00774.x PMID: 16316385
3. Pattaro C, Teumer A, Gorski M, Chu AY, Li M, Mijatovic V, et al. Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. Nat Commun. 2016; 7.
4. Pattaro C, Kottgen A, Teumer A, Garnaas M, Boger CA, Fuchsberger C, et al. Genome-wide association and functional follow-up reveals new loci for kidney function. PLoS genetics. 2012; 8(3):e1002584. doi: 10.1371/journal.pgen.1002584 PMID: 22479191
5. Kottgen A, Pattaro C, Boger CA, Fuchsberger C, Olden M, Glazer NL, et al. New loci associated with kidney function and chronic kidney disease. Nature genetics. 2010; 42(5):376–84. doi: 10.1038/ng.568 PMID: 20383146
6. Fox CS, Yang Q, Cupples LA, Guo CY, Larson MG, Leip EP, et al. Genomewide linkage analysis to serum creatinine, GFR, and creatinine clearance in a community-based population: the Framingham Heart Study. Journal of the American Society of Nephrology: JASN. 2004; 15(9):2457–61. doi: 10.1097/01.ASN.0000135972.13396.6F PMID: 15339995
7. Bochud M, Elston RC, Maillard M, Bovet P, Schild L, Shamlaye C, et al. Heritability of renal function in hypertensive families of African descent in the Seychelles (Indian Ocean). Kidney Int. 2005; 67(1):61–9. Epub 2004/12/22. doi: 10.1111/j.1523-1755.2005.00055.x PMID: 15610228
8. Langefeld CD, Beck SR, Bowden DW, Rich SS, Wagenknecht LE, Freedman BI. Heritability of GFR and albuminuria in Caucasians with type 2 diabetes mellitus. Am J Kidney Dis. 2004; 43(5):796–800. Epub 2004/04/28. PMID: 15112169
9. Placha G, Poznik GD, Dunn J, Smiles A, Krolewski B, Glew T, et al. A genome-wide linkage scan for genes controlling variation in renal function estimated by serum cystatin C levels in extended families with type 2 diabetes. Diabetes. 2006; 55(12):3358–65. Epub 2006/11/30. doi: 10.2337/db06-0781 PMID: 17130480
10. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. Modification of Diet in Renal Disease Study Group. Annals of Internal Medicine. 1999; 130(6):461–70. PMID: 10075613
11. Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, et al. Detection of large-scale variation in the human genome. Nature genetics. 2004; 36(9):949–51. doi: 10.1038/ng1416 PMID: 15286789
12. Itsara A, Cooper GM, Baker C, Girirajan S, Li J, Absher D, et al. Population analysis of large copy number variants and hotspots of human genetic disease. American Journal of Human Genetics. 2009; 84 (2):148–61. doi: 10.1016/j.ajhg.2008.12.014 PMID: 19166990
13. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. Annual review of genomics and human genetics. 2009; 10:451–81. doi: 10.1146/annurev.genom.9.081307.164217 PMID: 19715442
14. Zhang D, Cheng L, Qian Y, Alliey-Rodriguez N, Kelsoe JR, Greenwood T, et al. Singleton deletions throughout the genome increase risk of bipolar disorder. Molecular psychiatry. 2009; 14(4):376–80. doi: 10.1038/mp.2008.144 PMID: 19114987

15. Derks EM, Ayub M, Chambert K, Del Favero J, Johnstone M, MacGregor S, et al. A genome wide survey supports the involvement of large copy number variants in schizophrenia with and without intellectual disability. American journal of medical geneticsPart B, Neuropsychiatric genetics: the official publication of the International Society of Psychiatric Genetics. 2013; 162B(8):847–54.

16. Need AC, Ge D, Weale ME, Maia J, Feng S, Heinzen EL, et al. A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet. 2009; 5(2):e1000373. Epub 2009/02/07. PubMed Central PMCID: PMCPMC2631150. doi: 10.1371/journal.pgen.1000373 PMID: 19197363

17. Sutrala SR, Goossens D, Williams NM, Heyrman L, Adolfsson R, Norton N, et al. Gene copy number variation in schizophrenia. Schizophrenia research. 2007; 96(1–3):93–9. Epub 2007/09/11. doi: 10.1016/j.schres.2007.07.029 PMID: 17826036

18. Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, Liu XQ, et al. Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nat Genet. 2007; 39(3):319–28. Epub 2007/02/27. PubMed Central PMCID: PMCPMC4867008. doi: 10.1038/ng1985 PMID: 17322880

19. Kaminsky EB, Kaul V Fau—Paschall J, Paschall J Fau—Church DM, Church Dm Fau—Bunke B, Bunke B Fau—Kunig D, Kunig D Fau—Moreno-De-Luca D, et al. An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. (1530–0366 (Electronic)). D—NLM: PMC3661946 EDAT- 2011/08/17 06:00 MHDA- 2012/03/28 06:00 CRDT- 2011/08/17 06:00 AID—PST—ppublish.

20. Grayson BL, Smith ME, Thomas JW, Wang L, Dexheimer P, Jeffrey J, et al. Genome-wide analysis of copy number variation in type 1 diabetes. PloS one. 2010; 5(11):e15393. doi: 10.1371/journal.pone.0015393 PMID: 21085585

21. Aerts E, Beckers S, Zegers D, Van Hoorenbeeck K, Massa G, Verrijken A, et al. CNV analysis and mutation screening indicate an important role for the NPY4R gene in human obesity. Obesity (Silver Spring, Md). 2016; 24(4):970–6. Epub 2016/02/28.

22. Hasstedt SJ, Xin Y, Mao R, Lewis T, Adams TD, Hunt SC. A Copy Number Variant on Chromosome 20q13.3 Implicated in Thinness and Severe Obesity. Journal of obesity. 2015; 2015:623431. Epub 2016/02/18. PubMed Central PMCID: PMCPMC4736014. doi: 10.1155/2015/623431 PMID: 26881067

23. Wheeler E, Huang N, Bochukova EG, Keogh JM, Lindsay S, Garg S, et al. Genome-wide SNP and CNV analysis identifies common and low-frequency variants associated with severe early-onset obesity. Nature genetics. 2013; 45(5):513–7. doi: 10.1038/ng.2607 PMID: 23563609

24. Caruana G, Wong MN, Walker A, Heloury Y, Webb N, Johnstone L, et al. Copy-number variation associated with congenital anomalies of the kidney and urinary tract. Pediatric nephrology (Berlin, Germany). 2015; 30(3):487–95. Epub 2014/10/02.

25. Westland R, Verbitsky M, Vukojevic K, Perry BJ, Fasel DA, Zwijnenburg PJ, et al. Copy number variation analysis identifies novel CAKUT candidate genes in children with a solitary functioning kidney. Kidney Int. 2015; 88(6):1402–10. Epub 2015/09/10. PubMed Central PMCID: PMCPMC4834924. doi: 10.1038/ki.2015.239 PMID: 26352300

26. Sanna-Cherchi S, Kiryluk K, Burgess KE, Bodria M, Sampson MG, Hadley D, et al. Copy-number disorders are a common cause of congenital kidney malformations. Am J Hum Genet. 2012; 91(6):987–97. Epub 2012/11/20. PubMed Central PMCID: PMCPMC3516596. doi: 10.1016/j.ajhg.2012.10.007 PMID: 23159250

27. Cao Y, Li Z, Rosenfeld JA, Pursley AN, Patel A, Huang J, et al. Contribution of genomic copy-number variations in prenatal oral clefts: a multicenter cohort study. Genetics in medicine: official journal of the American College of Medical Genetics. 2016; 18(10):1052–5. Epub 2016/02/26.

28. Klamt J, Hofmann A, Bohmer AC, Hoebel AK, Golz L, Becker J, et al. Further evidence for deletions in 7p14.1 contributing to nonsyndromic cleft lip with or without cleft palate. Birth defects research Part A, Clinical and molecular teratology. 2016; 106(9):767–72. Epub 2016/07/08. doi: 10.1002/bdra.23539 PMID: 27384521

29. Younkin SG, Scharpf RB, Schwender H, Parker MM, Scott AF, Marazita ML, et al. A genome-wide study of de novo deletions identifies a candidate locus for non-syndromic isolated cleft lip/palate risk. BMC genetics. 2014; 15:24-2156-15-24.

30. Sapkota Y, Narasimhan A, Kumaran M, Sehrawat BS, Damaraju S. A Genome-Wide Association Study to Identify Potential Germline Copy Number Variants for Sporadic Breast Cancer Susceptibility. Cytogenetic and genome research. 2016. Epub 2016/09/27.

31. Shi J, Zhou W, Zhu B, Hyland PL, Bennett H, Xiao Y, et al. Rare Germline Copy Number Variations and Disease Susceptibility in Familial Melanoma. The Journal of investigative dermatology. 2016. Epub 2016/08/02.

32. Fernandez-Rozadilla C, Cazier JB, Tomlinson I, Brea-Fernandez A, Lamas MJ, Baiget M, et al. A genome-wide association study on copy-number variation identifies a 11q11 loss as a candidate

susceptibility variant for colorectal cancer. Human genetics. 2014; 133(5):525–34. doi: 10.1007/s00439-013-1390-4 PMID: 24218287

33. Scharpf RB, Mireles L, Yang Q, Kottgen A, Ruczinski I, Susztak K, et al. Copy number polymorphisms near SLC2A9 are associated with serum uric acid concentrations. BMC genetics. 2014; 15:81-2156-15-81.

34. McCarroll SA, Kuruvilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nature genetics. 2008; 40(10):1166–74. doi: 10.1038/ng.238 PMID: 18776908

35. Scharpf RB, Parmigiani G, Pevsner J, Ruczinski I. Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays. The annals of applied statistics. 2008; 2 (2):687–713. doi: 10.1214/07-AOAS155 PMID: 19609370

36. Cox A. ELAND: Efficient Large-Scale Alignment of Nucleotide Databases. Illumina. 2007;San Diego, CA.

37. Cardin N, Holmes C, Wellcome Trust Case Control C, Donnelly P, Marchini J. Bayesian hierarchical mixture modeling to assign copy number from a targeted CNV array. Genetic epidemiology. 2011; 35 (6):536–48. doi: 10.1002/gepi.20604 PMID: 21769931

38. Patterson N, Hattangadi N, Lane B, Lohmueller KE, Hafler DA, Oksenberg JR, et al. Methods for high-density admixture mapping of disease genes. American Journal of Human Genetics. 2004; 74(5):979–1000. doi: 10.1086/420871 PMID: 15088269

39. Venter JC, Smith HO, Adams MD. The Sequence of the Human Genome. Clin Chem. 2015; 61 (9):1207–8. Epub 2015/07/18. doi: 10.1373/clinchem.2014.237016 PMID: 26185218

40. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet. 2007; 3(4):e58. Epub 2007/04/24. PubMed Central PMCID: PMCPMC1853118. doi: 10.1371/journal.pgen.0030058 PMID: 17447842

41. Stevens LA, Coresh J, Schmid CH, Feldman HI, Froissart M, Kusek J, et al. Estimating GFR using serum cystatin C alone and in combination with serum creatinine: a pooled analysis of 3,418 individuals with CKD. American Journal of Kidney Diseases: The Official Journal of the National Kidney Foundation. 2008; 51(3):395–406.

42. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. American Journal of Epidemiology. 1989; 129(4):687–702. PMID: 2646917

43. Kottgen A, Glazer NL, Dehghan A, Hwang SJ, Katz R, Li M, et al. Multiple loci associated with indices of renal function and chronic kidney disease. Nature genetics. 2009; 41(6):712–7. doi: 10.1038/ng.377 PMID: 19430482

44. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006; 38(8):904–9. doi: 10.1038/ng1847 PMID: 16862161

45. Tin A, Astor Bc Fau—Boerwinkle E, Boerwinkle E Fau—Hoogeveen RC, Hoogeveen Rc Fau—Coresh J, Coresh J Fau—Kao WHL, Kao WH. Genome-wide association study identified the human leukocyte antigen region as a novel locus for plasma beta-2 microglobulin. (1432–1203 (Electronic)). D—NLM: PMC3656139 EDAT- 2013/02/19 06:00 MHDA- 2013/07/10 06:00 CRDT- 2013/02/19 06:00 PHST- 2012/10/07 [received] PHST- 2013/02/06 [accepted] PHST- 2013/02/16 [aheadofprint] AID—PST—ppublish.

46. Scharpf RB, Irizarry RA, Ritchie ME, Carvalho B, Ruczinski I. Using the R Package crlmm for Genotyping and Copy Number Estimation. Journal of statistical software. 2011; 40(12):1–32. PMID: 22523482

47. Scharpf RB, Ruczinski I, Carvalho B, Doan B, Chakravarti A, Irizarry RA. A multilevel model to address batch effects in copy number estimation using SNP arrays. Biostatistics (Oxford, England). 2011; 12 (1):33–50.