

RESEARCH ARTICLE

Open Access

# Copy number polymorphisms near SLC2A9 are associated with serum uric acid concentrations

Robert B Scharpf<sup>1,5\*</sup>, Lynn Mireles<sup>2</sup>, Qiong Yang<sup>3</sup>, Anna Köttgen<sup>2,4</sup>, Ingo Ruczinski<sup>5</sup>, Katalin Susztak<sup>6</sup>, Eitan Halper-Stromberg<sup>7</sup>, Adrienne Tin<sup>2</sup>, Stephen Cristiano<sup>8</sup>, Aravinda Chakravarti<sup>9</sup>, Eric Boerwinkle<sup>10</sup>, Caroline S Fox<sup>11</sup>, Josef Coresh<sup>2</sup> and Wen Hong Linda Kao<sup>2^</sup>

## Abstract

**Background:** Hyperuricemia is associated with multiple diseases, including gout, cardiovascular disease, and renal disease. Serum urate is highly heritable, yet association studies of single nucleotide polymorphisms (SNPs) and serum uric acid explain a small fraction of the heritability. Whether copy number polymorphisms (CNPs) contribute to uric acid levels is unknown.

**Results:** We assessed copy number on a genome-wide scale among 8,411 individuals of European ancestry (EA) who participated in the Atherosclerosis Risk in Communities (ARIC) study. CNPs upstream of the urate transporter *SLC2A9* on chromosome 4p16.1 are associated with uric acid ( $\chi^2_{df} = 3545$ ,  $p = 3.19 \times 10^{-23}$ ). Effect sizes, expressed as the percentage change in uric acid per deleted copy, are most pronounced among women ( $_{3.974.935.87}$  [2.55097.5 denoting percentiles],  $p = 4.57 \times 10^{-23}$ ) and independent of previously reported SNPs in *SLC2A9* as assessed by SNP and CNP regression models and the phasing SNP and CNP haplotypes ( $\chi^2_{df} = 3190$ ,  $p = 7.23 \times 10^{-08}$ ). Our finding is replicated in the Framingham Heart Study (FHS), where the effect size estimated from 4,089 women is comparable to ARIC in direction and magnitude ( $_{1.414.707.88}$ ,  $p = 5.46 \times 10^{-03}$ ).

**Conclusions:** This is the first study to characterize CNPs in ARIC and the first genome-wide analysis of CNPs and uric acid. Our findings suggests a novel, non-coding regulatory mechanism for *SLC2A9*-mediated modulation of serum uric acid, and detail a bioinformatic approach for assessing the contribution of CNPs to heritable traits in large population-based studies where technical sources of variation are substantial.

**Keywords:** Copy number polymorphism, Hyperuricemia, Genomewide association study

## Background

Serum uric acid levels are highly heritable and associated with several diseases, including gout, hypertension, and cardiovascular disease [1-4]. Genome-wide association studies have identified several single nucleotide polymorphisms (SNPs) that are strongly associated with uric acid levels [5-10], but a large proportion of the heritability of uric acid is unexplained by common SNPs. While variation of DNA copy number has been implicated in many heritable diseases, there has been no association studies of

copy number polymorphisms (CNPs) and serum uric acid levels on a genome-wide level.

High-throughput platforms used to genotype SNPs are useful for copy number estimation, though additional steps are required to reduce technical artifacts that are prevalent in studies of copy number. Estimates of the relative copy number (log R ratios) and B allele frequencies measured at each marker on the array are mutually informative for the latent copy number [11]. Various hidden Markov model (HMM) implementations integrate the log R ratios and B allele frequencies to infer copy number [12-19]. Copy number estimation is challenging, in part, due to technical artifacts that contribute to false positives. Among the most common artifacts are *genomic waves* [20,21], an autocorrelation of the marker-level estimates when plotted against physical position, and *batch*

\*Correspondence: rscharpf@jhu.edu

<sup>^</sup>Deceased

<sup>1</sup>550 N. Broadway, Suite 1101, Department of Oncology, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, USA

<sup>2</sup>Department of Biostatistics, Johns Hopkins School of Public Health, Baltimore, Maryland, USA

Full list of author information is available at the end of the article

effects, differences between groups of samples arising from technical sources of variation such as sample preparation, reagents, and laboratory personnel [22-24]. Approaches to reduce wave and batch artifacts include models for adjusting log R ratios by the GC composition of the local sequence as in [21] and surrogates of batch such as chemistry plate in association models when confounding between batch and phenotype is incomplete.

Here, we implement a HMM to infer integer copy number from B allele frequencies and wave-corrected log R ratios obtained from 8,411 ARIC participants of European ancestry assayed on Affymetrix 6.0 arrays. We evaluate the association between CNPs and uric acid concentrations through mixed effects regression models that adjust for available clinical risk factors as well as technical covariates such as chemistry plate and study center. For loci reaching genome-wide significance, we replicate our findings in the Framingham Heart Study (FHS). In addition, we assess whether statistically significant associations among EA participants persist in a smaller cohort of 3,392 African Americans in ARIC. Finally, we establish the independence of the relationship between copy number and uric acid concentrations from genome-wide significant SNP associations among ARIC EA participants.

## Results and discussion

Among 8,411 ARIC samples of European ancestry passing SNP and copy number metrics for quality control (see Methods), 47 percent are male and the mean BMI, uric acid concentration, and age are 27 kg/m<sup>2</sup>, 5.9 mg/dL, and 54 years, respectively.

Copy number estimates 0-4 were obtained from a HMM [14]. In this population, the median number of deletions and duplications is 55, and the median cumulative number of bases spanned by copy number variants (CNVs) in autosomal chromosomes is 3,530 kb (Additional file 1: Figure S1 and Table S1). The number of CNVs estimated for an individual is dependent on array quality and is associated with batch (chemistry plate). In particular, the detection of small CNVs (< 25 kb) requires high quality arrays, whereas identification of large CNVs (> 200 kb) is robust to array quality and batch (Additional file 1: Figure S2). From the distribution of CNV breakpoints across all EA subjects, we identified 12,397 disjoint (non-overlapping) genomic intervals for which copy number is unambiguous and at least 1 percent of ARIC participants have a duplication or deletion (see Methods). These genomic intervals capture 317 non-contiguous loci constituting the CNPs ascertained by the HMM among EA ARIC participants, and nearly all span known regions of copy number variation reported in the Database of Genomic Variants [25].

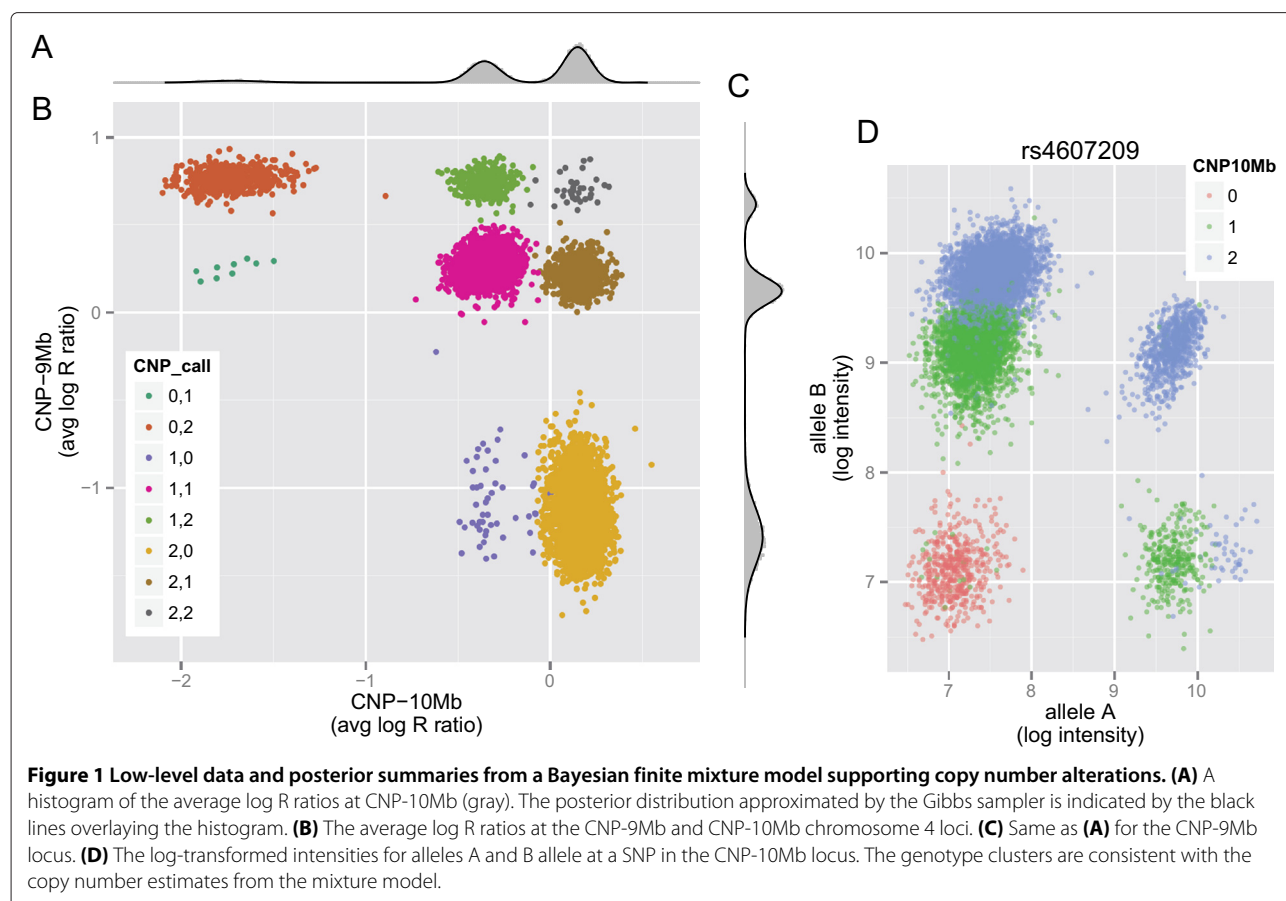
Prior to our assessment of CNPs as potential risk factors for hyperuricemia, we removed seasonal trends of uric

acid concentrations using a loess smoother with span  $\frac{1}{10}$  fit to women and men independently. Our baseline mixed effects model for seasonally adjusted log uric acid concentrations includes fixed effects for study center, age, log BMI, gender, and the interaction of age and log BMI with gender, as well as a random effect for chemistry plate.

For each disjoint interval, we extended the baseline model for uric acid with copy number (0-4) modeled as a continuous covariate. A Manhattan plot of the  $-\log_{10}$  *p*-value revealed a cluster of statistically significant associations on chromosome 4 (Additional file 1: Figure S3, A). The statistically significant coefficients are derived from two non-overlapping CNPs with NCBI36 build coordinates 9,832,502–9,844,354 bp (CNP-9Mb) and 10,002,240–10,009,754 bp (CNP-10Mb; Additional file 1: Figure S3, B). Together, the two CNPs span 19.368 kb, are interrogated by 49 nonpolymorphic markers and 1 SNP, overlap common deletions previously identified in HapMap Phase 1 [26], and are upstream of the *SLC2A9* gene that is transcribed in the reverse direction. With the exception of the chromosome 4 locus, the distribution of *p*-values is approximately uniform (Additional file 1: Figure S4).

The marginal distribution of the average log R ratios at CNP-10Mb and CNP-9Mb can be approximated by a mixture of normal distributions, where the components of the mixture are induced by differences in the latent copy number (Figure 1A and 1C). Our approximation to the posterior is derived from a Gibbs' sampler [27,28], an approach conceptually similar to the Bayesian mixture model described in [29] and extending some of the originally proposed heuristics using mixture models for CNPs [30]. A scatterplot of the average log R ratios at CNP-9Mb and CNP-10Mb provides a non-discrete visualization of their joint distribution (Figure 1B). Assuming the mixture components correspond to latent copy numbers 0, 1, and 2, the integer copy number for each sample is inferred from the component with highest posterior probability. The copy number estimates from the mixture model are further corroborated by the genotype clusters for SNP rs4607209 in the CNP-10 Mb locus (Figure 1D). For example, samples belonging to the second mixture component (copy number 1) populate the 'A' and 'B' genotype clusters at SNP rs4607209 (green). Hereafter, regression models for uric acid utilize the maximum a posteriori copy number estimates from the Bayesian mixture model.

Copy number estimates at the CNP-9Mb and CNP-10Mb loci have a Spearman correlation coefficient of -0.82. Homozygous deletions are common at each locus (46% of subjects at the CNP-9Mb locus and 6% of subjects at the CNP-10Mb locus), yet none of the subjects have a homozygous deletion at both loci (233 expected by chance). Evaluated in separate regression models, each deleted copy at CNP-9Mb and CNP-10Mb is associated

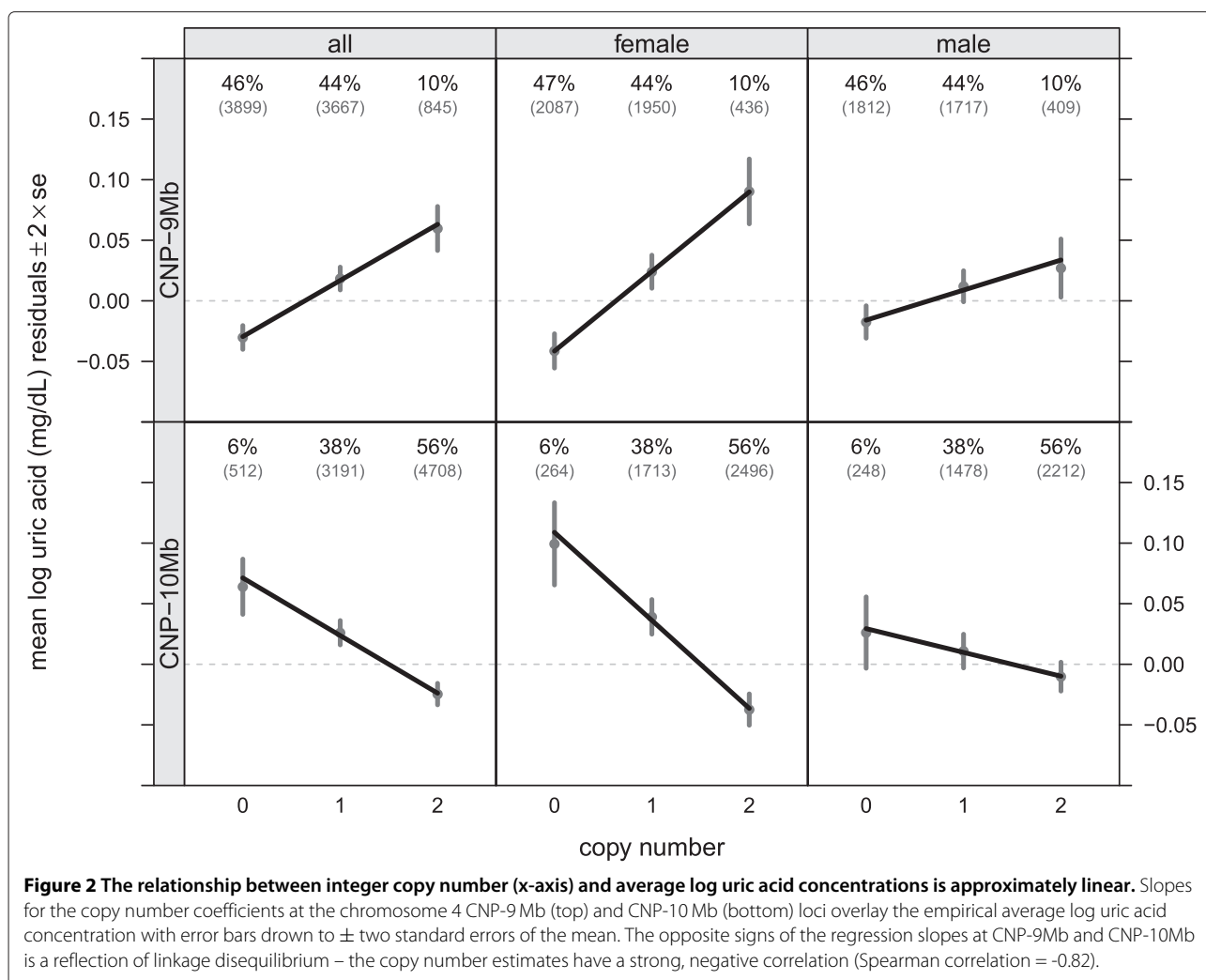


with a  $1.171.501.82$  percentage decrease ( $p = 5.43 \times 10^{-20}$ ) and a  $1.832.633.42$  percentage increase ( $p = 1.54 \times 10^{-10}$ ) in uric acid concentrations, respectively (Figure 2). While the regression coefficients at CNP-9Mb and CNP-10Mb are opposite in sign, the data is consistent with a dose response to copy number at only one CNP and an opposing sign for the tagging CNP attributable to its strong linkage disequilibrium. At each locus, the interaction of copy number and gender is statistically significant with more pronounced slopes observed among women. For example, each deleted copy at the CNP-10 Mb CNP among women is associated with a  $3.974.935.87$  ( $p = 4.57 \times 10^{-23}$ ) percentage increase of uric acid concentrations, whereas among men each deleted copy is associated with a  $0.311.362.39$  ( $p = 0.001$ ) percentage increase in uric acid concentrations.

To evaluate whether CNPs at the chromosome 4 loci are associated with uric acid in an independently sampled EA population for which uric acid measurements are available, we pursued replication in FHS. Because access to the intensity-level data in FHS was not available, we used missing genotype calls for SNP rs4607209 in the CNP-10 Mb CNP as a surrogate for the deletion polymorphism (justification in Methods). With the missing genotype indicator as a surrogate for homozygous

deletions, we fit a mixed effects model implemented in the R package kinship [31] with log uric acid concentrations as the dependent variable and clinical covariates age, gender, and log-transformed BMI as explanatory variables. The gender-specific slopes for the surrogate copy number variable in FHS are comparable to the copy number slopes in ARIC with respect to magnitude, direction, and statistical significance (Figure 3). In particular, missing genotypes are associated with a  $1.414.707.88$  percentage increase of uric acid concentrations among FHS women ( $p = 5.46 \times 10^{-03}$ ) compared to a  $3.974.935.87$  percentage increase among ARIC women ( $p = 4.57 \times 10^{-23}$ ). As in ARIC, the  $-3.120.173.36$  percentage change in uric acid concentrations among men is small and not statistically significant in FHS ( $p = 0.92$ ). Replication at the CNP-9 Mb CNP is not possible as the array platform used in FHS does not contain markers in this region.

To investigate whether the association between copy number and uric acid concentrations is present in non-EA populations, we estimated the copy number at both chromosome 4 CNPs for 3,392 African American (AA) participants in ARIC using the Bayesian mixture model described previously for the EA cohort. Homozygous deletions occur in approximately 46 and 6% of

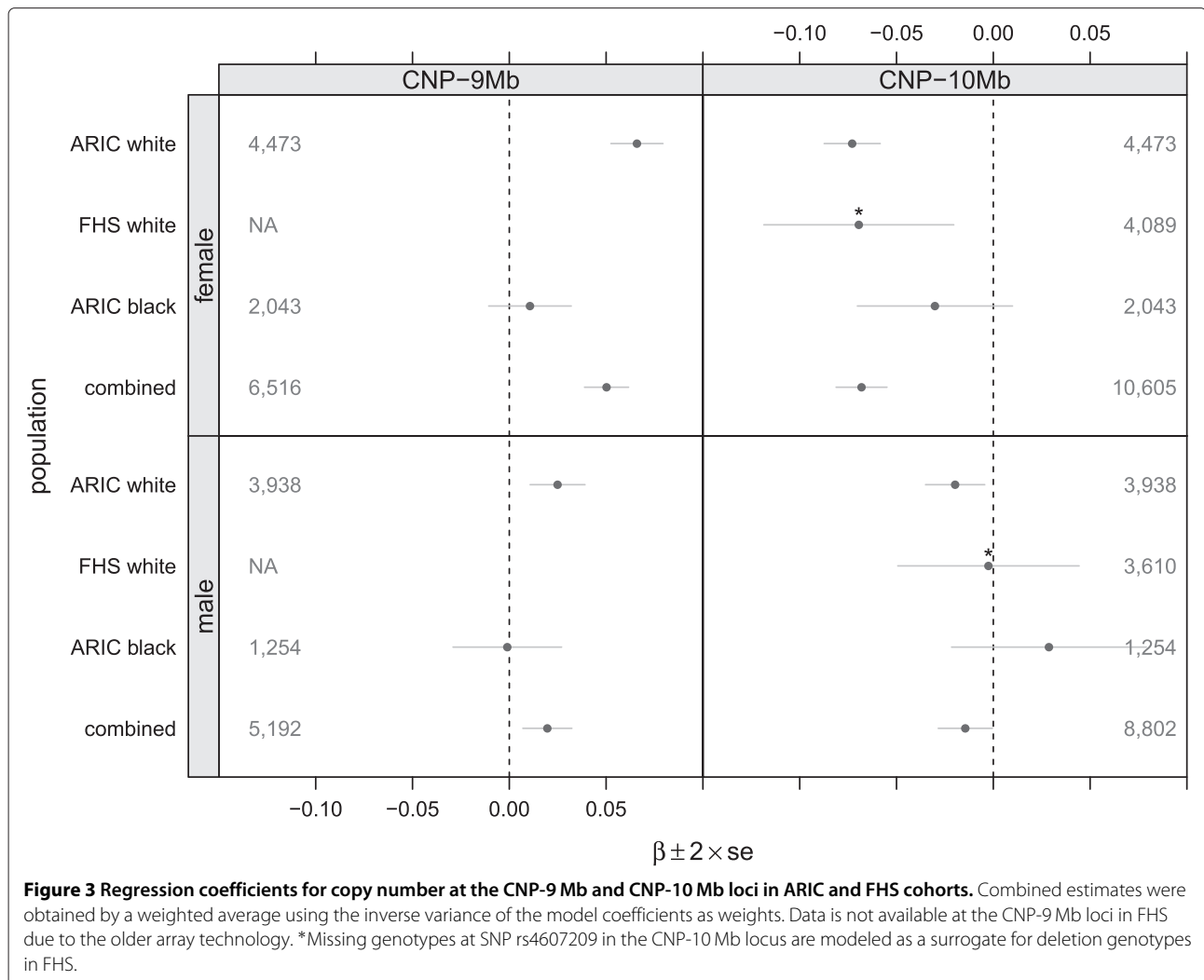


EA participants at the CNP-9Mb and CNP-10 Mb loci, respectively, but only 33 and 0.6% of AA participants have homozygous deletions at these loci. The percentage decrease of uric acid concentrations associated with each deleted copy at CNP-9 Mb is  $-0.75$  $0.73$  $_{2.22}$  among women ( $p = 0.335$ ) and  $-1.90$  $0.05$  $_{1.97}$  among men ( $p = 0.957$ ). Similarly, copy number is not associated with uric acid levels among AA women or men at CNP-10 Mb ( $\chi^2_{2df} = 3.45, p = 0.179$ ) (Figure 3).

To assess whether the CNP associations are independent of *SLC2A9* SNPs among EA participants, we evaluated a series of models for uric acid concentrations that include SNPs and/or the gender-specific CNP slopes. Marginally, the association between SNPs and CNPs with uric acid concentrations is the strongest for SNPs directly in the *SLC2A9* transcript, and the associations 200 kb upstream of *SLC2A9* are comparable for SNPs and CNPs (Figure 4, top). Adjusting for the SNP with the strongest marginal association (rs7675964), effect sizes for other SNPs near *SLC2A9* decrease. The CNP effect sizes are

also attenuated but remain genome-wide significant (minimum  $\chi^2_{2df} = 3190, p = 7.23 \times 10^{-08}$ ) (Figure 4, bottom). Adjusted for the CNP with the strongest marginal association (CNP-9 Mb), the effect size for SNP rs7675964 is comparable to the marginal model (data not shown).

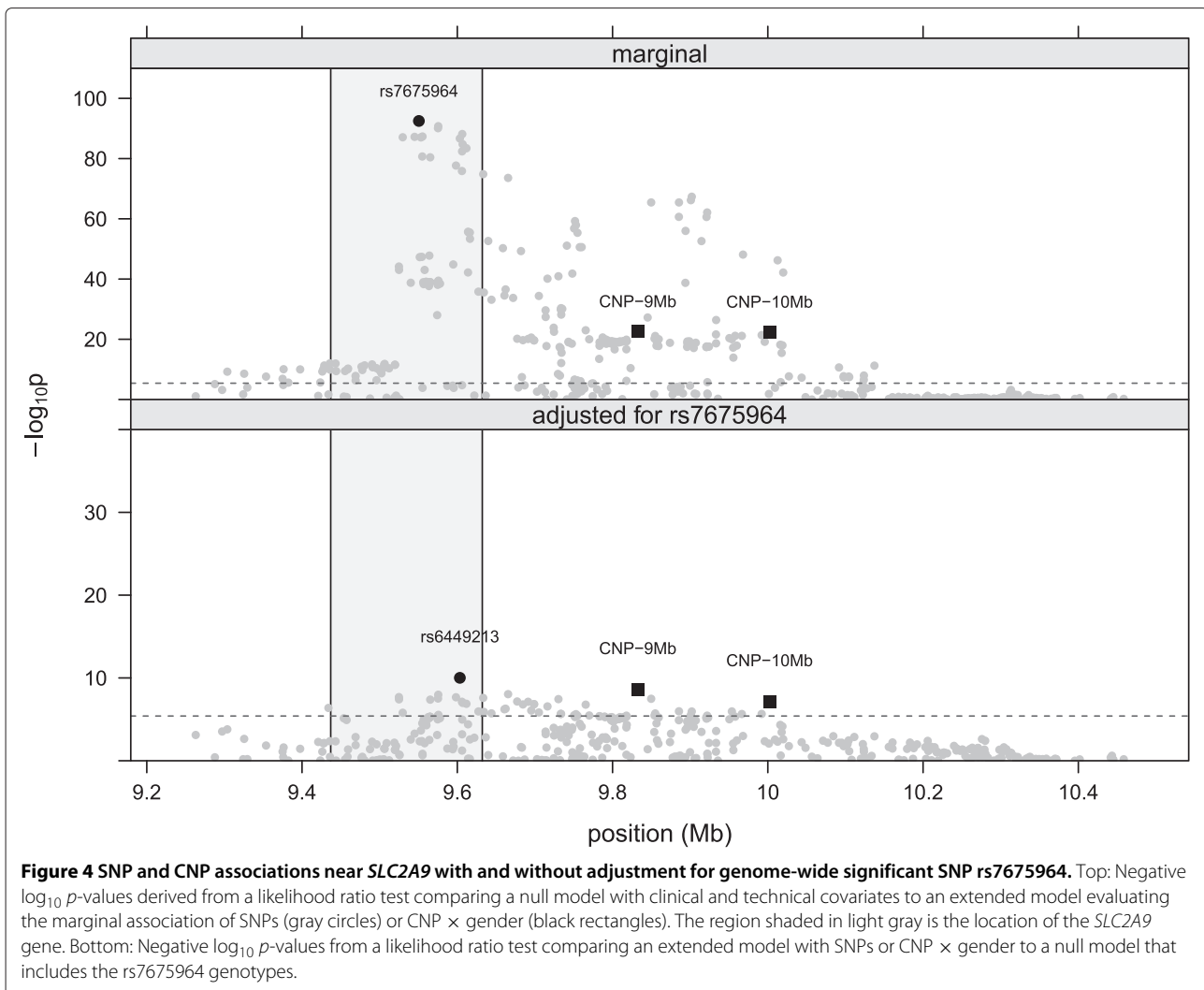
While regression coefficients for SNPs near *SLC2A9* are attenuated in the rs7675964-adjusted models, SNP rs6449213 (and others) remain genome-wide significant ( $p = 9.46 \times 10^{-11}$ ). To assess the independence of the CNP association with uric acid after adjusting for the rs6449213 and rs7675964 genotypes, we compared the baseline mixed effects model with rs6449213 and rs7675964 genotypes to an extended model with gender-specific slopes for copy number. A 2 degree of freedom likelihood ratio test comparing the baseline and extended models is statistically significant at both CNP loci (CNP-9 Mb:  $\chi^2_{2df} = 31, p = 2.01 \times 10^{-07}$ ; CNP-10 Mb:  $\chi^2_{2df} = 33, p = 8.72 \times 10^{-08}$ ). To further evaluate whether CNPs contribute to inter-individual variation of uric acid concentrations independently of SNPs in *SLC2A9*, we phased the genotypes at



rs7675964 and rs6449213 with copy number at CNP-9 Mb and CNP-10 Mb (see Methods). Notationally, we denote the CNP portion of the haplotypes by  $H1: -c_{1,1} -c_{1,2} -$ ,  $H2: -c_{2,1} -c_{2,2} -$ , where  $c_{i,j}$  is the copy number at the  $j^{th}$  CNP locus ( $c_{i,j} \in \{0, 1\}$ ) for haplotype  $H_i$  ( $i \in \{1, 2\}$ ). Similarly, the portion of the haplotypes for rs7675964 and rs6449213 are denoted by  $H1: -g_{1,1} -g_{1,2} -$ ,  $H2: -g_{2,1} -g_{2,2} -$ , where  $g_{i,j}$  is the allele at the  $j^{th}$  SNP ( $g_{i,j} \in \{a, b\}$ ). Of the  $2^4$  possible allelic haplotypes, 14 were observed in the 8,411 EA participants and only 3 SNP haplotypes had variation in the corresponding CNP haplotype. Specifically, the 3 SNP haplotypes for we observed variation in the phased copy number estimates are  $H1: -a -a -a -$ ,  $H2: -a -a -a -$ ,  $H1: -b -b -b -$ ,  $H2: -a -a -a -$ , and  $H1: -b -a -a -$ ,  $H2: -a -a -a -$ . For 2,195 subjects with the allelic haplotype  $H1: -b -b -b -$ ,  $H2: -a -a -a -$ , CNP haplotypes  $H1: -0 -0 -1 -$ ,  $H2: -1 -1 -0 -$  and  $H1: -0 -1 -1 -$ ,  $H2: -1 -1 -1 -$  are weakly associated with uric acid concentrations with similar effect sizes observed

in men and women ( $\chi^2_{4df} = 9.05$ ,  $p = 0.0599$ ). For 4,313  $H1: -a -a -a -$ ,  $H2: -a -a -a -$  subjects, CNP haplotypes are associated with uric acid concentrations in women ( $\chi^2_{4df} = 14.3$ ,  $p = 6.3 \times 10^{-03}$ ) but not men ( $\chi^2_{4df} = 0.757$ ,  $p = 0.944$ ). CNP haplotypes are not associated with uric acid concentrations for  $H1: -b -a -a -$ ,  $H2: -a -a -a -$  subjects ( $\chi^2_{2df} = 2.06$ ,  $p = 0.357$ ), though the sample size for this population is small and the effect size among the 66 women in this subgroup is comparable to the effect size in the much larger  $H1: -b -b -b -$ ,  $H2: -a -a -a -$  and  $H1: -a -a -a -$ ,  $H2: -a -a -a -$  subgroups for which the CNP haplotype association is statistically significant (Figure 5).

As the CNP association appears independent of *SLC2A9* SNPs and the CNP loci are located in an intergenic region approximately 200 kb upstream of the *SLC2A9* gene (*SLC2A9* is transcribed in the reverse orientation), we examined publicly available regulatory data for human kidney tissue where *SLC2A9* is known to function in the



transport of uric acid from urine to blood [32]. Examination of DNase hypersensitivity for human fetal kidney tissue and adult kidney cell line HKC8 revealed a peak adjacent to CNP-10 Mb, suggesting that CNP-10 Mb abuts a regulatory element. We did not observe DNase hypersensitivity peaks near CNP-9 Mb, but nearly half of EA participants have a homozygous deletion at CNP-9 Mb. It is unclear whether the absence of peaks at CNP-9 Mb reflect the absence of a regulatory element in the fetal kidney or whether the fetal kidney has a deletion at this locus (i.e., loss of a regulatory element by deletion).

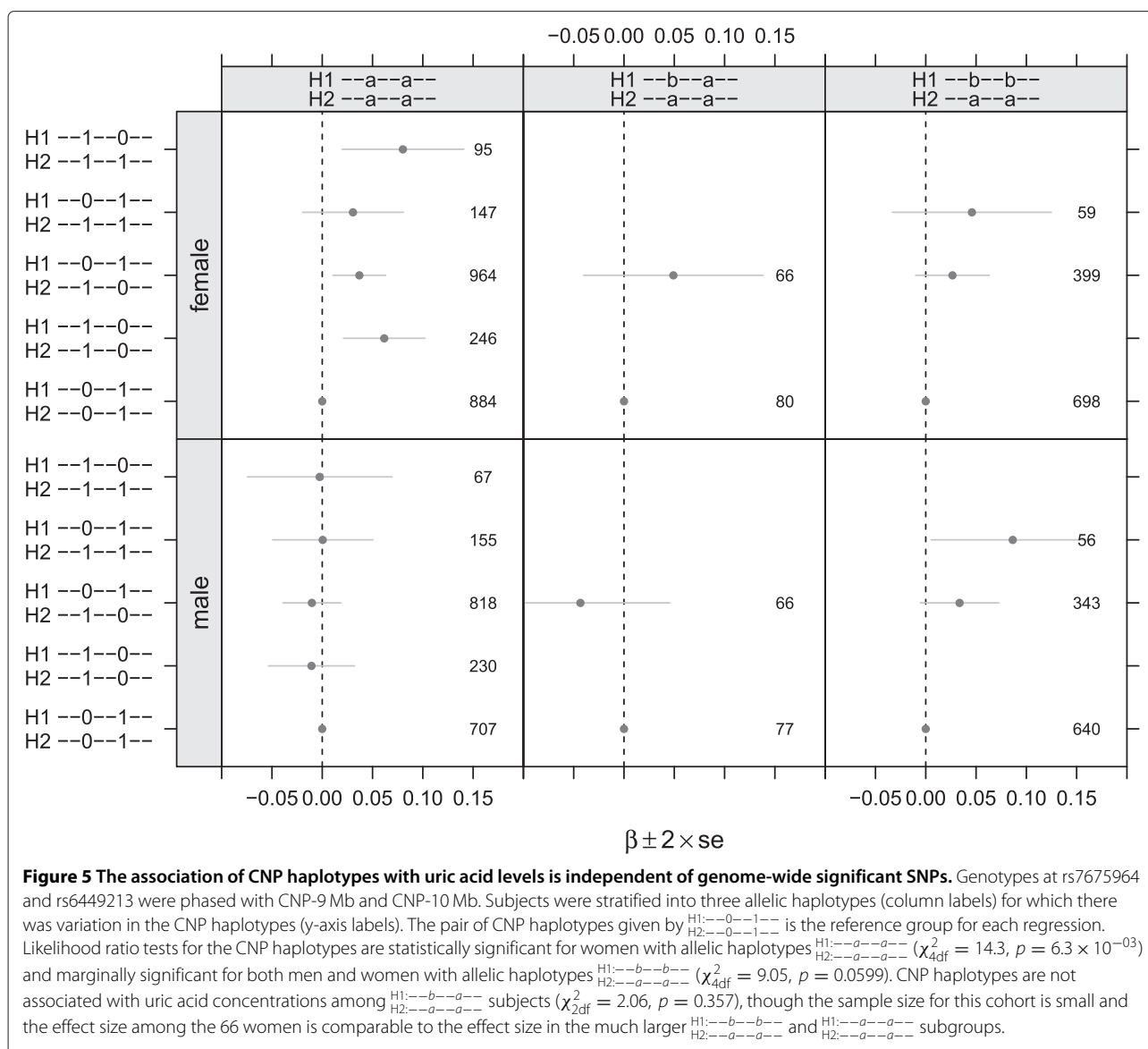
Given the strong association between CNPs and uric acid, we modeled the relationship between CNPs and gout. Of the 8,411 ARIC EA participants, 609 had gout at some point during the study's follow-up. In a logistic regression model including technical and clinical covariates described previously, the odds of gout is 1.21 times higher comparing subjects who differ by one copy of CNP-9 Mb ( $p = 0.003$ ). As expected, this association is largely

mediated through the CNP's association with serum uric acid. After including uric acid in the model, the association between copy number at CNP-9 Mb and gout is attenuated (1.11 odds ratio;  $p=0.12$ ). Results are qualitatively similar at the CNP-10 Mb locus with a statistically significant gout association in the marginal model that is attenuated after adjusting for uric acid concentrations (data not shown).

## Conclusions

This study is the first genome-wide scan of CNPs and uric acid. We identified an association between serum uric acid concentrations and two common, intergenic deletions that are 200 kb and 350 kb, respectively, upstream of the urate transporter *SLC2A9*. Loss of DNA copy number in these regions is associated with  $\approx 5$  percent change of uric acid concentrations among women and a one percent change among men with the direction of the effect depending on the CNP locus ( $\chi^2_{df} =$





3545,  $p = 3.19 \times 10^{-23}$ ). Gender-specific associations between *SLC2A9* polymorphisms and uric acid concentrations have been reported by others and are consistent with our observations with CNPs near *SLC2A9* [7,33-36]. Independent replication of the association between copy number and uric acid concentrations in FHS provides further support for our finding. Among ARIC AA participants, CNP-10 Mb is weakly associated with uric acid concentrations and there was no association at CNP-9 Mb in men or women. The CNP association in ARIC EA is independent of previously reported SNP associations in *SLC2A9*, as assessed by joint CNP and SNP regression models as well as regression models with phased SNP and CNP haplotypes.

The physiological role of *SLC2A9* in the kidney is the reabsorption of urate from urine into blood, leading to

increased levels of serum uric acid concentrations when *SLC2A9* expression is up-regulated and decreased levels with loss of function mutations such as deletions. When phased with genome-wide significant SNPs in *SLC2A9*, the haplotypes with homozygous deletions at CNP-9 Mb had lower uric acid concentrations as we would hypothesize if CNP-9 Mb spans an enhancer for *SLC2A9*. DNase hypersensitivity assays suggest that CNP-10 Mb abuts a regulatory element, but we did not find DNase hypersensitivity or ChIP-seq peaks at CNP-9 Mb. Assays from other cell lines in ENCODE are consistent with our findings in the kidney. For example, CNP-10 Mb spans DNase hypersensitivity peaks in normal esophageal epithelial cells (HEEpiC cell line), airway epithelial cells (SAEC cell line), epidermal keratinocytes (cell line NHEK), and mammary epithelial cells (HMEC cell line), as well as a H3KMe1

histone mark in HMEC cells [37]. As nearly 50 percent of EA participants in ARIC have homozygous deletions at CNP-9 Mb, it is possible that the fetal kidney cell line harbors a homozygous deletion at this locus and that the absence of ChIP-seq binding and DNase hypersensitivity reflect absence of regulatory elements due to loss of DNA copy number. Gene expression data for kidney or liver tissues and germline copy number for the same samples is not currently available in ARIC or FHS.

Our CNP GWAS has low sensitivity for deletions less than 50 kb in size and/or having fewer than 10 Affymetrix 6.0 markers. For amplifications, the inability to discriminate high copy amplifications from single- and two- copy duplications because of the limited dynamic range of the array platform will attenuate the regression coefficients for copy number. The attenuation of the copy number coefficients for amplifications occurs irrespective of the size of the amplicon, but will be worse for small, focal amplifications due to the limited resolution of the platform. Our analyses do not rule out the contribution of small insertions and deletions as well as high copy repeats that are beyond the dynamic range of high-throughput arrays. Sequencing platforms will be useful for elucidating whether additional structural and mutational variants near *SLC2A9* contribute to inter-individual heterogeneity of uric acid concentrations. In addition, our association analysis only included CNPs. Rare duplications and deletions such as those directly spanning the *SLC2A9* transcript (5 deletions and 9 duplications in ARIC) were not evaluated in our analysis of CNPs and may have a larger effect on uric acid concentrations than the CNPs studied here. While these limitations impact sensitivity, our results indicate that CNP genome-wide association studies can achieve a high degree of specificity. As in any high-throughput setting, the specificity of a genome-wide screen depends on the extent to which technical factors influencing estimation can be modeled and the degree to which they are independent of the outcome of interest. Participants in ARIC were neither enrolled nor processed on the basis of their uric acid concentrations. Due to the merits of the experimental design and mixed models for uric acid that adjust for study center and chemistry plate, we feel the major sources of artefactual associations in ARIC have been addressed.

In summary, the loss of several kilobases of DNA in close proximity to *SLC2A9*, a known uric acid transporter and a candidate gene for gout [38-40], presents a biologically plausible mechanism for regulation of *SLC2A9* expression and modulation of serum uric acid concentrations. Gene expression data on the same set of individuals in target kidney and liver tissues is needed to evaluate whether loss of DNA copy number effects transcription of *SLC2A9* as hypothesized, and to evaluate gender differences in *SLC2A9* expression.

## Methods

This paper follows the guidelines for communicating confidence intervals as suggested in [41]. Institutional Review Board (IRB) approval was obtained by the Johns Hopkins University ARIC study center, and the research was conducted in accordance with the principles described in the Declaration of Helsinki.

### ARIC study

The ARIC study is an ongoing, prospective community-based cohort of 15,792 persons (27% black) aged 45-64 years at baseline (1987-89) [42]. Participants were selected by probability sampling from four U.S. communities (Forsyth County, North Carolina; Jackson, Mississippi; Minneapolis, Minnesota; and Washington County, Maryland). Participants took part in examinations starting with a baseline visit between 1987 and 1989 and three follow-up visits, thereafter, administered three years apart (visit 2: 1990-1992; visit 3: 1993-1995; visit 4: 1996-1998). At baseline, a home interview assessed participants' sociodemographic characteristics, smoking, and alcohol-drinking habits, medication use, and medical history. A clinical examination included measurement of various risk factors. All participants self-reported race as Asian, black, American Indian, or white. Body-mass index (BMI) was measured according to published methods [43]. Central laboratories performed analyses on baseline fasting specimens using conventional assays to obtain uric acid values [44]. Uric acid was measured by the uricase method [45]. The reliability coefficient of uric acid was 0.91, and within-person variability was 7.2 [46].

### CNV estimation

Raw CEL files from scanned Affymetrix 6.0 arrays were processed using Affymetrix power tools (APT, version 1.14.3) and PennCNV to derive estimates of log R ratios and B allele frequencies at each marker. While the log R ratio estimates were wave-adjusted [21], genomic waves persisted in many of the ARIC samples. We further processed the log R ratios using the R package ArrayTV [47] – an approach adapted from software for removing waves in high-throughput sequencing data [48]. A 6-state HMM comprising 5 distinct copy number states (0-4) implemented in the R package VanillaICE (VI) and the stand-alone tool PennCNV were applied independently to each sample [13,14,49]. CNVs with fewer than 10 markers were excluded due to the level of noise of the log R ratios and the difficulty in assessing the validity of low-coverage CNVs without experimental validation. As inference from association models using the PennCNV- and VI- derived copy number estimates were found to be qualitatively similar, only the VI copy number associations were reported.



### Quality control measures

Among 9,779 samples of EA for whom uric acid concentrations were measured at visit 1, we excluded 743 samples that did not meet criteria for SNP genome-wide association analyses in ARIC as described in Köttgen *et al.* [50]. For the estimation of germline CNVs, high CNV call frequencies often indicate problems with the normalization such as genomic waves that were incompletely removed by the wave correction methods. We excluded 625 participants with autosomal log R ratios having high autocorrelation or variance (lag 10 autocorrelation > 0.03 or median absolute deviation > 0.32), or if the number of CNVs called by the VI algorithm exceeded 100. We used the signal to noise ratio (SNR) implemented in the R package *crlmm* as a sample-specific measure of array quality as assessed by the overall separation of the canonical genotype clusters at SNPs [51,52], but we did not exclude samples on the basis of this statistic. Following the above quality control filters, 8,411 EA participants were evaluated in the subsequent association models.

### Genome-wide scan of copy number and uric acid levels

From the set of genomic intervals defining CNVs derived by the VI HMM fit to 8,411 EA subjects, we constructed rectangular matrices of the inferred integer copy number. Element  $[i, j]$  of the matrix is the copy number at genomic interval  $i$  for sample  $j$ . The genomic intervals were obtained from the union of the start and end coordinates across all CNVs detected for each of the autosomal chromosomes with the requirement that each non-overlapping (disjoint) interval contain at least one marker. For each disjoint interval, we calculated the number of samples harboring a CNV, excluding intervals for which fewer than one percent of the samples had a CNV. Across samples, the CNVs are partially overlapping and any given CNV may span one or many disjoint intervals. As a consequence, adjacent disjoint intervals often convey similar information with comparable frequencies of deletions and duplications. As the test statistics are correlated, Bonferonni correction is conservative. Because none of the loci were of borderline statistical significance (Additional file 1: Figure S3), more sophisticated simulation-based approaches for multiple testing correction with dependent test statistics were not assessed.

Mixed effects regression models for ARIC cohorts were implemented using the R package *lme4* [53]. Specifically, we modeled seasonally adjusted serum log uric acid concentrations (continuous) in a regression model with fixed effects for copy number (modeled as continuous with scale 0-4), age (continuous), log-transformed BMI (continuous), gender, and study center (categorical). As the heavy-tailed uric acid concentrations were log-transformed, we report

the percentage change of uric acid concentrations per integer increase in copy number. To take into account the heterogeneity of CNV call frequencies between chemistry plates, we include chemistry plate as a random effect. For regression models with canonical genotypes as covariates, we treated the frequency of the B-allele (an integer in the set 0, 1, or 2) as continuous. For FHS, we implemented mixed effects regression models using the R package *kinship* (<http://cran.uvigo.es/src/contrib/Archive/kinship/>) [31].

### Imputation of copy number in the Framingham heart study

To evaluate whether CNPs at the chromosome 4 loci are associated with uric acid in an independently sampled EA population, we explored replication in FHS. Challenges to replication in FHS include the older array architecture (Affymetrix 250k Nsp/Sty chips) and the unavailability of raw intensities needed for copy number estimation. While there were no markers for CNP-9 Mb on the 250k chips, SNP rs4607209 in CNP-10 Mb is present in the Affymetrix 250k Nsp chip. To verify that the expected non-diploid genotypes ('A', 'B', and NULL genotypes) can be observed from the normalized intensities for this SNP on the Affymetrix 250k Nsp chip, we genotyped the 270 phase 2 HapMap samples that were assayed on the the Affymetrix 250k platform using the BRLMM algorithm implemented in Affymetrix power tools. (The BRLMM algorithm was used to genotype FHS participants.) A scatterplot of the log intensities for the A and B alleles reveals three clusters corresponding to the deletion genotypes for rs4607209 in addition to the canonical biallelic clusters (Additional file 1: Figure S5), and is similar to the clusters observed on the Affymetrix 6.0 platform for ARIC EA participants (Figure 1D). Homozygous deletions occur in 8.9% of the HapMap CEPH samples and 6.1% of the ARIC EA participants. The canonical biallelic genotypes in HapMap have high genotype confidence scores (not shown) and no missing calls, while 6 out of 8 CEPH subjects with homozygous deletions have missing BRLMM genotype calls. These data demonstrate that the low level intensities for SNP rs4607209 in both the 250k Nsp and Affymetrix 6.0 platforms have distinct clusters corresponding to the latent copy number and that missing BRLMM genotypes occur in clusters that are consistent with homozygous deletions. The specificity of missing genotype calls as a surrogate for homozygous deletion genotypes at SNP rs4607209 in EA HapMap is 1 and the sensitivity is 0.75. We expect that missing genotype calls as a surrogate for homozygous deletions will lead to conservative parameter estimates of the copy number effect size in regression models as contamination of the diploid population with subjects harboring homozygous and hemizygous deletions will bias the regression slopes to zero.

### Estimation of copy number for ARIC AA participants

Log R ratios for markers in the CNP-9 Mb and CNP-10 Mb loci were averaged. The average log R ratios in AA participants are a mixture of 3 normal distributions as observed in the EA population, with the mixture components presumed to be induced by differences in the latent copy number. A Gibbs' sampler [27,28] was implemented in R to approximate the posterior distribution of the 3-component normal mixture. Each subject was assigned to the mixture component with the highest posterior probability. As in the EA cohort, the observed mixture components in the AA cohort are most consistent with homozygous deletion, hemizygous deletion, and diploid copy number on the basis of the expected log R ratios for these copy number states.

### Phasing SNPs and CNPs near *SLC2A9*

Genotypes from 8 SNPs having the largest marginal associations with uric acid (including rs7675964 and rs6449213) were phased with CNP-9 Mb and CNP-10 Mb using the fastPHASE software [54]. For diploid CNPs, we assumed that each haplotype had one copy. This assumption is supported empirically by the data—if haplotypes containing two copies were common, we would expect to see subjects with duplications. Haplotypes were modeled as categorical covariates in regression models for uric acid concentrations. Subjects with rare haplotypes and subjects with allelic haplotypes that had no variation in the corresponding CNP portion of the haplotypes were excluded (1,473 subjects).

### Genomic annotation and software versions

Genomic annotation in this paper is based on UCSC build hg18 (NCBI36) [55]. Gene *SLC2A9* has RefSeq accession numbers NM\_001001290.1 and NM\_020041.2. We used the May, 2010 version of PennCNV, version 1.14.3 of APT, and version 1.4.0 of fastPHASE [54]. All remaining analyses were performed in the statistical environment R [56]. Graphics were generated using the R packages lattice [57] or ggplot2 [58,59]. The analyses downstream of the VI algorithm relied on the infrastructure provided by the GenomicRanges package [60]. The complete listing of supporting R packages and their corresponding version numbers is provided below.

- R version 3.1.0 (2014-04-10), x86\_64-apple-darwin13.1.0
- Base packages: base, datasets, graphics, grDevices, grid, methods, parallel, stats, tools, utils
- Other packages: aricUricAcid 1.0.19, Biobase 2.24.0, BiocGenerics 0.10.0, Biostrings 2.32.0, DBI 0.2-7, devtools 1.5, foreach 1.4.2, GenomeInfoDb 1.0.2, GenomicRanges 1.16.3, ggplot2 1.0.0, gridExtra 0.9.1, gtable 0.1.2, IRanges 1.22.7, knitr 1.6, lattice 0.20-29,

lme4 1.1-6, Matrix 1.1-3, oligo 1.28.2, oligoClasses 1.26.0, pd.genomewidensp.6 1.10.0, RColorBrewer 1.0-5, Rcpp 0.11.1, RSQLite 0.11.4, XVector 0.4.0

- Loaded via a namespace (and not attached): affxparser 1.36.0, affyio 1.32.0, BiocInstaller 1.14.2, bit 1.1-12, codetools 0.2-8, colorspace 1.2-4, digest 0.6.4, evaluate 0.5.5, ff 2.2-13, formatR 0.10, gtools 3.4.0, httr 0.3, iterators 1.0.7, latticeExtra 0.6-26, MASS 7.3-33, memoise 0.2.1, minqa 1.2.3, munsell 0.4.2, nlme 3.1-117, plyr 1.8.1, preprocessCore 1.26.1, proto 0.3-10, RcppEigen 0.3.2.1.2, RCurl 1.95-4.1, reshape2 1.4, scales 0.2.4, splines 3.1.0, stats4 3.1.0, stringr 0.6.2, whisker 0.3-2, zlibbioc 1.10.0

### Availability of supporting data

The data set supporting the results of this article is available in the dbGaP repository, phs000090.v1.p1 ([http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000090.v1.p1](http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000090.v1.p1)). The ChIP-seq and DNAase hypersensitivity data for the kidney described in [32] is available from the GEO repository, accession: GSE49637 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE49637>).

### Additional file

**Additional file 1: Supplementary figures and tables. Figure S1:** Size, frequency and burden of CNVs among ARIC participants of European ancestry. **Figure S2:** Batch effects in processing arrays for copy number estimation. **Figure S3:** Manhattan plot of copy number associations. **Figure S4:** Quantile-quantile plot of the expected  $-\log_{10} p$ -values versus the observed  $-\log_{10} p$ -values. **Figure S5:** A scatterplot of the normalized intensities for the A and B alleles of SNP rs4607209 for 90 HapMap subjects of EA assayed on the Affymetrix 250k Nsp chip used in FHS. **Table S1:** Median and interquartile range (IQR) descriptive statistics of CNVs for 8,411 EA participants.

### Abbreviations

AA: African American; ARIC: Atherosclerosis risk in communities; BMI: Body mass index; ChIP: Chromatin immunoprecipitation; CNP: Copy number variant; CNP: Copy number polymorphism; EA: European ancestry; FHS: Framingham heart study; HMM: Hidden Markov model; MAD: Median absolute deviation; SNP: Single nucleotide polymorphism; SNR: Signal to noise ratio.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

RBS, JC, and WKHL conceived of the study. RBS, LM, AK, EB, CSF, AC, KS, and WKHL drafted the manuscript. KS and AK participated in the analysis and interpretation of DNAase hypersensitivity and ChIP-seq assays. RBS, LM, EHS, QY, IR, AT, and SC participated in the statistical analyses. All authors read and approved the final manuscript.

### Acknowledgements

This work was supported by National Institutes of Health grants R01HG005220 and R00HG005015 [R.B.S., L.M., A.C., W.H.L.K.]. 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 research was conducted in part using data and resources from the Framingham Heart Study of the National Heart Lung and Blood Institute of the National Institutes of Health and Boston University School of Medicine (Contract No. N01-HC-25195), its contract with Affymetrix, Inc for genotyping services (Contract No. N02-HL-6-4278) and National Institute of Health grants R01 NS017950-28 and R01-HL093328-01. The analyses reflect intellectual input and resource development from the Framingham Heart Study investigators participating in the SNP Health Association Resource (SHARe) project. Framingham Heart Study investigators were supported in part by the National Heart, Lung, and Blood Institute's Framingham Heart Study (Contract No. N01-HC-25195) and grant numbers R01HL093328, R01HL093029, R01NS017950 and R01HL093029 [C.F.S. and Q.Y.]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors thank the staff and participants of the ARIC and FHS studies for their important contributions. We thank Dan Arking for the suggestion of phasing the SNP and CNP haplotypes.

#### Author details

<sup>1</sup>550 N. Broadway, Suite 1101, Department of Oncology, Johns Hopkins School of Medicine, Baltimore, Maryland 21205, USA. <sup>2</sup>Department of Epidemiology, Johns Hopkins School of Public Health, Baltimore, Maryland, USA. <sup>3</sup>Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, USA. <sup>4</sup>Department of Medicine IV, University Hospital Freiburg, Freiburg im Breisgau, Germany. <sup>5</sup>Department of Biostatistics, Johns Hopkins School of Public Health, Baltimore, Maryland, USA. <sup>6</sup>Renal Electrolyte and Hypertension Division, Perelman School of Medicine, University of Pennsylvania, Philadelphia PA, USA. <sup>7</sup>Computational Biosciences Program, University of Colorado, Denver, Aurora, Colorado, USA. <sup>8</sup>Department of Biostatistics, Johns Hopkins School of Public Health, Baltimore, Maryland, USA. <sup>9</sup>Department of Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland, USA. <sup>10</sup>IMM Center for Human Genetics, University of Texas School of Public Health, Houston, Texas, USA. <sup>11</sup>Laboratory for Metabolic and Population Health, National Heart Lung and Blood Institute, National Institutes of Health, Framingham, Massachusetts, USA.

Received: 31 March 2014 Accepted: 30 June 2014

Published: 9 July 2014

#### References

- Liese AD, Hense HW, Löwel H, Döring A, Tietze M, Keil U: **Association of serum uric acid with all-cause and cardiovascular disease mortality and incident myocardial infarction in the MONICA Augsburg cohort. World Health Organization monitoring trends and determinants in cardiovascular diseases.** *Epidemiology* 1999, **10**(4):391–397.
- Feig DI, Kang D-H, Johnson RJ: **Uric acid and cardiovascular risk.** *N Engl J Med* 2008, **359**(17):1811–1821. doi:10.1056/NEJMra0800885.
- Rao DC, Laskarzewski PM, Morrison JA, Khoury P, Kelly K, Glueck CJ: **The clinical lipid research clinic family study: familial determinants of plasma uric acid.** *Hum Genet* 1982, **60**(3):257–261.
- Rice T, Vogler GP, Perry TS, Laskarzewski PM, Province MA, Rao DC: **Heterogeneity in the familial aggregation of fasting serum uric acid level in five North American populations: the lipid research clinics family study.** *Am J Med Genet* 1990, **36**(2):219–225. doi:10.1002/ajmg.1320360216.
- Charles BA, Shriner D, Doumatey A, Chen G, Zhou J, Huang H, Herbert A, Gerry NP, Christman MF, Adeyemo A, Rotimi CN: **A genome-wide association study of serum uric acid in African Americans.** *BMC Med Genomics* 2011, **4**:17. doi:10.1186/1755-8794-4-17.
- Wallace C, Newhouse SJ, Braund P, Zhang F, Tobin M, Falchi M, Ahmadi K, Dobson RJ, Marécano ACB, Hajat C, Burton P, Deloukas P, Brown M, Connell JM, Dominiczak A, Lathrop GM, Webster J, Farrall M, Spector T, Samani NJ, Caulfield MJ, Munroe PB: **Genome-wide association study identifies genes for biomarkers of cardiovascular disease: serum urate and dyslipidemia.** *Am J Hum Genet* 2008, **82**(1):139–149. doi:10.1016/j.ajhg.2007.11.001.

- Dehghan A, Köttgen A, Yang Q, Hwang S-J, Kao WL, Rivadeneira F, Boerwinkle E, Levy D, Hofman A, Astor BC, Benjamin EJ, van Duijn CM, Witteman JC, Coresh J, Fox CS: **Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study.** *Lancet* 2008, **372**(9654):1953–1961. doi:10.1016/S0140-6736(08)61343-4.
- Karns R, Zhang G, Sun G, Rao Indugula S, Cheng H, Havas-Augustin D, Novokmet N, Rudan D, Durakovic Z, Missoni S, Chakraborty R, Rudan P, Deka R: **Genome-wide association of serum uric acid concentration: replication of sequence variants in an island population of the Adriatic coast of Croatia.** *Ann Hum Genet* 2012, **76**(2):121–127. doi:10.1111/j.1469-1809.2011.00698.x.
- Tin A, Woodward OM, Kao WHL, Liu C-T, Lu X, Nalls MA, Shriner D, Semmo M, Akyzbekova EL, Wyatt SB, Hwang S-J, Yang Q, Zonderman AB, Adeyemo AA, Palmer C, Meng Y, Reilly M, Shlipak MG, Siscovick D, Evans MK, Rotimi CN, Flessner MF, Köttgen M, Cupples LA, Fox CS, Köttgen A, Candidate Gene Associated Resource (C. A. Re), Cohorts for Heart and Aging Research in Genetic Epidemiology (C. H. A. R. G. E.): **Genome-wide association study for serum urate concentrations and gout among African Americans identifies genomic risk loci and a novel URAT1 loss-of-function allele.** *Hum Mol Genet* 2011, **20**(20):4056–4068.
- Kolz M, Johnson T, Sanna S, Teumer A, Vitart V, Perola M, Mangino M, Albrecht E, Wallace C, Farrall M, Johansson A, Nyholt DR, Aulchenko Y, Beckmann JS, Bergmann S, Bochud M, Brown M, Campbell H, European Special Population Research Network (EUROSPAN), Connell J, Dominiczak A, Homuth G, Lamina C, McCarthy MI, European Network for Genetic and Genomic Epidemiology (ENGAGE), Meitinger T, Mooser V, Munroe P, Nauck M, Peden J, et al: **Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations.** *PLoS Genet* 2009, **5**(6):1000504. doi:10.1371/journal.pgen.1000504.
- Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL: **High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping.** *Genome Res* 2006, **16**(9):1136–1148. doi:10.1101/gr.5402306.
- Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J: **QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data.** *Nucleic Acids Res* 2007, **35**(6):2013–2025. doi:10.1093/nar/gkm076.
- Wang K, Li M, Hadley D, Liu R, Glennier J, Grant SFA, Hakonarson H, Bucan M: **PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data.** *Genome Res* 2007, **17**(11):1665–1674. doi:10.1101/gr.6861907.
- Scharpf RB, Parmigiani G, Pevsner J, Ruczinski I: **Hidden Markov models for the assessment of chromosomal alterations using high-throughput SNP arrays.** *Ann Appl Stat* 2008, **2**(2):687–713.
- Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, Altshuler D: **Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs.** *Nat Genet* 2008, **40**(10):1253–1260. doi:10.1038/ng.237.
- Greenman CD, Bignell G, Butler A, Edkins S, Hinton J, Beare D, Swamy S, Santarius T, Chen L, Widada S, Futreal PA, Stratton MR: **PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data.** *Biostatistics* 2010, **11**(1):164–175. doi:10.1093/biostatistics/kxp045.
- Su S-Y, Asher JE, Jarvelin M-R, Froguel P, Blakemore AIF, Balding DJ, Coin LJM: **Inferring combined CNV/SNP haplotypes from genotype data.** *Bioinformatics* 2010, **26**(11):1437–1445. doi:10.1093/bioinformatics/btq157.
- Yau C, Mouradov D, Jorissen RN, Colella S, Mirza G, Steers G, Harris A, Ragoussis J, Sieber O, Holmes CC: **A statistical approach for detecting genomic aberrations in heterogeneous tumor samples from single nucleotide polymorphism genotyping data.** *Genome Biol* 2010, **11**(9):92. doi:10.1186/gb-2010-11-9-r92.
- Yau C, Papaspiliopoulos O, Roberts GO, Holmes C: **Bayesian nonparametric hidden Markov models with application to the**

- analysis of copy number variation in mammalian genomes. *J R Stat Soc Series B Stat Methodol* 2011, **73**(1):37–57. doi:10.1111/j.1467-9868.2010.00756.x.**
20. Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H, Andrews TD, Stranger BE, Lynch AG, Dermitzakis ET, Carter NP, Tavaré S, Hurles ME: **Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization.** *Genome Biol* 2007, **8**(10):r228. doi:10.1186/gb-2007-8-10-r228.
  21. Diskin SJ, Li M, Hou C, Yang S, Gleessner J, Hakonarson H, Bucan M, Maris JM, Wang K: **Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms.** *Nucleic Acids Res* 2008, **36**(19):126. doi:10.1093/nar/gkn556.
  22. Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, Hurles ME: **A robust statistical method for case-control association testing with copy number variation.** *Nat Genet* 2008, **40**(10):1245–1252. doi:10.1038/ng.206.
  23. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA: **Tackling the widespread and critical impact of batch effects in high-throughput data.** *Nat Rev Genet* 2010, **11**(10):733–739. doi:10.1038/nrg2825.
  24. 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* 2011, **12**(1):33–50. doi:10.1093/biostatistics/kxq043.
  25. lafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C: **Detection of large-scale variation in the human genome.** *Nat Genet* 2004, **36**(9):949–951. doi:10.1038/ng1416.
  26. McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, Dallaire S, Gabriel SB, Lee C, Daly MJ, Altshuler DM, Consortium IH: **Common deletion polymorphisms in the human genome.** *Nat Genet* 2006, **38**(1):86–92.
  27. Geman S, Geman D: **Stochastic relaxation, gibbs distributions, and the bayesian restoration of images.** *IEEE Trans Pattern Anal Mach Intell* 1984, **6**(6):721–741.
  28. Diebolt J, Robert CP: **Estimation of finite mixture distributions through Bayesian sampling.** *J R Stat Soc Series B Methodol* 1994, **56**:363–375.
  29. Cardin N, Holmes C, WTCCC, Donnelly P, Marchini J: **Bayesian hierarchical mixture modeling to assign copy number from a targeted cnv array.** *Genet Epidemiol* 2011, **35**(6):536–548. doi:10.1002/gepi.20604.
  30. McCarroll SA, Kuruwilla FG, Korn JM, Cawley S, Nemesh J, Wysoker A, Shapero MH, de Bakker PIW, Maller JB, Kirby A, Elliott AL, Parkin M, Hubbell E, Webster T, Mei R, Veitch J, Collins PJ, Handsaker R, Lincoln S, Nizzari M, Blume J, Jones KW, Rava R, Daly MJ, Gabriel SB, Altshuler D: **Integrated detection and population-genetic analysis of SNPs and copy number variation.** *Nat Genet* 2008, **40**(10):1166–1174. doi:10.1038/ng.238.
  31. Atkinson B, Therneau T: **Kinship: mixed-effects cox models, sparse matrices, and modeling data from large pedigrees.** 2013. R package version 1.1.0-23. [http://cran.uvigo.es/src/contrib/Archive/kinship/]
  32. Ko Y-A, Mohtat D, Suzuki M, Park ASD, Izquierdo MC, Han SY, Kang HM, Si H, Hostetter T, Pullman JM, Fazzari M, Verma A, Zheng D, Grealley JM, Susztak K: **Cytosine methylation changes in enhancer regions of core pro-fibrotic genes characterize kidney fibrosis development.** *Genome Biol* 2013, **14**(10):108. doi:10.1186/gb-2013-14-10-r108.
  33. Döering A, Gieger C, Mehta D, Gohlke H, Prokisch H, Coassin S, Fischer G, Henke K, Klopp N, Kronenberg F, Paulweber B, Pfeufer A, Rosskopf D, Völzke H, Illig T, Meitinger T, Wichmann H-E, Meisinger C: **SLC2A9 influences uric acid concentrations with pronounced sex-specific effects.** *Nat Genet* 2008, **40**(4):430–436. doi:10.1038/ng.107.
  34. McArdle PF, Parsa A, Chang Y-PC, Weir MR, O'Connell JR, Mitchell BD, Shuldiner AR: **Association of a common nonsynonymous variant in GLUT9 with serum uric acid levels in old order amish.** *Arthritis Rheum* 2008, **58**(9):2874–2881. doi:10.1002/art.23752.
  35. Hu M, Tomlinson B: **Gender-dependent associations of uric acid levels with a polymorphism in SLC2A9 in Han Chinese patients.** *Scand J Rheumatol* 2012, **41**(2):161–163. doi:10.3109/030097422011.637952.
  36. Köttgen A, Albrecht E, Teumer A, Vitart V, Krumsiek J, Hundertmark C, Pistis G, Ruggiero D, O'Seaghdha CM, Haller T, Yang Q, Tanaka T, Johnson AD, Kutalik Z, Smith AV, Shi J, Struchalin M, Middelberg RPS, Brown MJ, Gaffo AL, Pirastu N, Li G, Hayward C, Zemunik T, Huffman J, Yengo L, Zhao JH, Demirkan A, Feitosa MF, Liu X, et al: **Genome-wide association analyses identify 18 new loci associated with serum urate concentrations.** *Nat Genet* 2013, **45**(2):145–154. doi:10.1038/ng.2500.
  37. ENCODE Project Consortium, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M: **An integrated encyclopedia of DNA elements in the human genome.** *Nature* 2012, **489**(7414):57–74. doi:10.1038/nature11247.
  38. Li S, Sanna S, Maschio A, Busonero F, Usala G, Mulas A, Lai S, Dei M, Orrù M, Albai G, Bandinelli S, Schlessinger D, Lakatta E, Scuteri A, Najjar S, S, Guralnik J, Naitza S, Crisponi L, Cao A, Abecasis G, Ferrucci L, Uda M, Chen W-M, Nagaraja R: **The GLUT9 gene is associated with serum uric acid levels in Sardinia and Chianti cohorts.** *PLoS Genet* 2007, **3**(11):194. doi:10.1371/journal.pgen.0030194.
  39. Matsuo H, Chiba T, Nagamori S, Nakayama A, Domoto H, Phetdee K, Wiriyasermkul P, Kikuchi Y, Oda T, Nishiyama J, Nakamura T, Morimoto Y, Kamakura K, Sakurai Y, Nonoyama S, Kanai Y, Shinomiya N: **Mutations in glucose transporter 9 gene SLC2A9 cause renal hypouricemia.** *Am J Hum Genet* 2008, **83**(6):744–751. doi:10.1016/j.ajhg.2008.11.001.
  40. Doblado M, Moley KH: **Facilitative glucose transporter 9, a unique hexose and urate transporter.** *Am J Physiol Endocrinol Metab* 2009, **297**(4):831–835. doi:10.1152/ajpendo.00296.2009.
  41. Louis TA, Zeger SL: **Effective communication of standard errors and confidence intervals.** *Biostatistics* 2009, **10**(1):1–2. doi:10.1093/biostatistics/kxn014.
  42. ARIC investigators: **The atherosclerosis risk in communities (aric) study: design and objectives. The ARIC investigators.** *Am J Epidemiol* 1989, **129**(4):687–702.
  43. Center ARIC: *Operations Manual No. 2: Cohort Component Procedures.* Chapel Hill: University of North Carolina School of Public Health; 1987.
  44. Center ARIC: *Operations Manual No. 10: Clinical Chemistry Determinations.* Chapel Hill: University of North Carolina School of Public Health; 1987.
  45. Iribarren C, Folsom AR, Eckfeldt JH, McGovern PG, Nieto FJ: **Correlates of uric acid and its association with asymptomatic carotid atherosclerosis: the ARIC study. Atherosclerosis Risk in Communities.** *Ann Epidemiol* 1996, **6**(4):331–340.
  46. Eckfeldt JH, Chambless LE, Shen YL: **Short-term, within-person variability in clinical chemistry test results. Experience from the atherosclerosis risk in communities study.** *Arch Pathol Lab Med* 1994, **118**(5):496–500.
  47. Halper-Stromberg E, Scharpf RB, ArrayTV: **Wave Correction for Arrays.** 2013. R package version 1.0.0. [http://www.bioconductor.org/packages/release/bioc/html/ArrayTV.html]
  48. Benjamini Y, Speed TP: **Summarizing and correcting the GC content bias in high-throughput sequencing.** *Nucleic Acids Res* 2012, **40**(10):72. doi:10.1093/nar/gks001.
  49. Scharpf RB, Beaty TH, Schwender H, Younkin SG, Scott AF, Ruczinski I: **Fast detection of de novo copy number variants from SNP arrays for case-parent trios.** *BMC Bioinformatics* 2012, **13**(1):330. doi:10.1186/1471-2105-13-330.
  50. Köttgen A, Glazer NL, Dehghan A, Hwang S-J, Katz R, Li M, Yang Q, Gudnason V, Launer LJ, Harris TB, Smith AV, Arking DE, Astor BC, Boerwinkle E, Ehret GB, Ruczinski I, Scharpf RB, Chen Y-DI, de Boer IH, Haritunians T, Lumley T, Sarnak M, Siscovick D, Benjamin EJ, Levy D, Upadhyay A, Aulchenko YS, Hofman A, Rivadeneira F, Uitterlinden AG, et al: **Multiple loci associated with indices of renal function and chronic kidney disease.** *Nat Genet* 2009. doi:10.1038/ng.377.
  51. Carvalho B, Bengtsson H, Speed TP, Irizarry RA: **Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data.** *Biostatistics* 2007, **8**(2):485–499. doi:10.1093/biostatistics/kxl042.
  52. Lin S, Carvalho B, Cutler D, Arking D, Chakravarti A, Irizarry R: **Validation and extension of an empirical Bayes method for SNP calling on Affymetrix microarrays.** *Genome Biol* 2008, **9**(4):63. doi:10.1186/gb-2008-9-4-r63.
  53. Bates D, Maechler M, Bolker B: **Lme4: Linear mixed-effects models using S4 classes.** 2012. R package version 0.999999-0. [http://CRAN.R-project.org/package=lme4]
  54. Scheet P, Stephens M: **A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase.** *Am J Hum Genet* 2006, **78**(4):629–644. doi:10.1086/502802.

55. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, Kent WJ: **The UCSC genome browser database: update 2011**. *Nucleic Acids Res* 2011, **39**(Database issue):876–882. doi:10.1093/nar/gkq963.
56. R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2012. [<http://www.R-project.org/>].
57. Sarkar D: *Lattice: Multivariate Data Visualization With R*. New York: Springer; 2008. ISBN 978-0-387-75968-5.
58. Wickham H: *ggplot2: Elegant Graphics for Data Analysis. Use R!*. 233 Spring Street, New York, NY 10013, USA: Springer; 2009.
59. Yin T, Cook D, Lawrence M: **ggbio: an R package for extending the grammar of graphics for genomic data**. *Genome Biol* 2012, **13**(8):77.
60. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, Gentleman R, Morgan MT, Carey VJ: **Software for computing and annotating genomic ranges**. *PLoS Comput Biol* 2013, **9**(8):1003118. doi:10.1371/journal.pcbi.1003118.

doi:10.1186/1471-2156-15-81

**Cite this article as:** Scharpf et al.: Copy number polymorphisms near SLC2A9 are associated with serum uric acid concentrations. *BMC Genetics* 2014 **15**:81.

Submit your next manuscript to BioMed Central  
and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

