Genome-wide Association of Copy-Number Variation Reveals an Association between Short Stature and the Presence of Low-Frequency Genomic Deletions

Andrew Dauber,[1,2,4,19] Yongguo Yu,[5,6,19] Michael C. Turchin,[2,3,8,9] Charleston W. Chiang,[2,3,8,9,10] Yan A. Meng,[2,3] Ellen W. Demerath,[11] Sanjay R. Patel,[12] Stephen S. Rich,[13] Jerome I. Rotter,[14] Pamela J. Schreiner,[15] James G. Wilson,[16] Yiping Shen,[5,6,7,19,*] Bai-Lin Wu,[6,7,17,18,19] and Joel N. Hirschhorn[1,2,3,8,9,10,19,*]

Height is a model polygenic trait that is highly heritable. Genome-wide association studies have identified hundreds of single-nucleotide polymorphisms associated with stature, but the role of structural variation in determining height is largely unknown. We performed a genome-wide association study of copy-number variation and stature in a clinical cohort of children who had undergone comparative genomic hybridization (CGH) microarray analysis for clinical indications. We found that subjects with short stature had a greater global burden of copy-number variants (CNVs) and a greater average CNV length than did controls ($p < 0.002$). These associations were present for lower-frequency (<5%) and rare (<1%) deletions, but there were no significant associations seen for duplications. Known gene-deletion syndromes did not account for our findings, and we saw no significant associations with tall stature. We then extended our findings into a population-based cohort and found that, in agreement with the clinical cohort study, an increased burden of lower-frequency deletions was associated with shorter stature ($p = 0.015$). Our results suggest that in individuals undergoing copy-number analysis for clinical indications, short stature increases the odds that a low-frequency deletion will be found. Additionally, copy-number variation might contribute to genetic variation in stature in the general population.

Height is a highly heritable complex trait, and up to 90% of the variation in height is due to genetic factors. It is a classic polygenic trait and has been used as a model for understanding the genetic architecture of complex traits.[1] Most association studies of polygenic traits—height in particular—have examined common single-nucleotide polymorphisms.[2] A recent genome-wide association study identified 180 independent loci associated with height,[3] demonstrating the highly polygenic nature of stature. Despite the tremendous progress made by these studies, the loci identified only explain ~10% of the variation in adult height.[3] It is possible that other types of genetic variation, such as low-frequency variation and/or copy-number variation, might contribute to the genetic variation in stature.

Although there are a few examples of common copy-number variants (CNVs) that are associated with complex traits, such as obesity[4] and Crohn disease,[5] the data thus far suggest that the majority of the heritability of complex traits is not due to common copy-number variation.[6] Most studies showing association with copy-number variation in polygenic traits have found associations with low-frequency CNVs that are not well tagged by SNPs.[7] Studies of schizophrenia[8] and autism[9] have identified an increased global burden of rare CNVs in affected subjects. These studies of rare copy-number variation have typically been performed in cohorts ascertained for the presence of disease. We sought to examine the role of both rare and common CNVs in the stature of cohorts that were not specifically ascertained on the basis of height.

To investigate whether CNVs play a role in short or tall stature, we conducted a genome-wide association study of copy-number burden in a cohort of children who had undergone comparative genomic hybridization (CGH) analysis for clinical indications, and we observed an excess of rare deletions in children with short stature. We then extended our findings to a large population-based cohort and again observed an excess of low-frequency deletions in shorter individuals. We also explored whether individual regions in the genome have CNVs associated with stature, and we preliminarily identified three candidate regions in our clinical cohort.

[1]Division of Endocrinology, Children's Hospital Boston, Boston, MA 02115, USA; [2]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02141, USA; [3]Metabolism Program, Broad Institute, Cambridge, MA 02141, USA; [4]Clinical Investigator Training Program, Beth Israel Deaconess Medical Center, Harvard Medical School, in collaboration with Pfizer Inc. and Merck & Co., Boston, MA 02115, USA; [5]Shanghai Children's Medical Center, Jiaotong University, Shanghai 200127, China; [6]Department of Laboratory Medicine, Children's Hospital Boston and Harvard Medical School, Boston, MA 02115, USA; [7]Department of Pathology, Children's Hospital Boston and Harvard Medical School, Boston, MA 02115, USA; [8]Division of Genetics, Children's Hospital Boston, Boston, MA 02115, USA; [9]Center for Basic and Translational Obesity Research, Children's Hospital Boston, Boston, MA 02115, USA; [10]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA; [11]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA; [12]Division of Sleep Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA; [13]University of Virginia, Charlottesville, VA 22908, USA; [14]Medical Genetics Institute, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA; [15]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA; [16]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS 39216, USA; [17]Children's Hospital, Fudan University, Shanghai 200032, China; [18]Institutes of Biomedical Science, Fudan University, Shanghai 200032, China
[19]These authors contributed equally to this work
*Correspondence: yiping.shen@childrens.harvard.edu (Y.S.), joelh@broadinstitute.org (J.N.H.)
DOI 10.1016/j.ajhg.2011.10.014.

In the clinical cohort, subjects were eligible if they had a height measurement recorded between the ages of 2 and 20 years and had had a chromosomal microarray performed as part of their clinical evaluation. All microarrays were performed on the Agilent 244K platform (Agilent Technologies, Santa Clara, CA). Subjects with aneuploidy and poor microarray quality were removed, leaving a final sample size of 4,411 individuals. All CNV data were called with NEXUS software (BioDiscovery, El Segundo, California). Copy-number polymorphisms in candidate regions were validated in a subgroup of individuals via multiplex ligation-dependent probe amplification (Table S1, available online).

The height measurements from the three visits closest to the date of the microarray testing were abstracted from the electronic medical record. Age- and height-adjusted Z scores were calculated on the basis of the United States CDC growth charts.[10] The mean of the three height Z scores was used as the final Z score. Tall and short cases were defined as subjects with Z scores greater than +2 and less than −2 standard deviations, respectively. Individuals with Z scores between −2 and +2 were used as controls. Ethnicity information was not available for subjects in the clinical cohort and thus could not be controlled for in the analysis.

All genome-wide association analyses of CNV burden were performed with PLINK v1.07.[11] We compared, in both cases and controls, the distributions of the total number of CNV segments, the total span of CNVs, and the average CNV length. We then stratified the analysis by deletions (copy number <2) and duplications (copy number >2), as well as by CNV frequency. We defined common CNVs as being present in greater than 5% of our cohort and lower-frequency and rare CNVs as being present in less than 5% and less than 1% of the cohort, respectively. Statistical significance is based on one-tailed comparisons of the observed metrics in cases versus in controls (cases were tested for an excess of copy-number variation) to the distribution of the same metrics in 10,000 permutations of case-control status. Using the hg18 gene list of all Refseq genes provided on the PLINK resources page, we were able to identify all CNVs that fell within 20 kb of a gene. We then performed a genic CNV analysis by including all such CNVs.

We performed a regional association analysis to look for candidate genomic regions that contained CNVs associated with height. All genomic coordinates provided are in the NCBI36/hg18 genome build. Separate analyses were performed for deletions and duplications. For this analysis, the genome was divided into 10 kb segments. For each segment with two or more CNVs, we first identified each individual with a CNV in that segment. Then, we calculated the sum of all the individuals' Z scores. For example, if three individuals with deletions in a particular 10 kb segment had height Z scores of −2, −1, and +1, the sum of the Z scores for that segment would equal −2. Consecutive segments with identical CNVs present in the same sets of individuals were collapsed into one larger segment. We then performed 100,000 permutations in which the Z scores of the 4,403 individual subjects with CNVs were randomly shuffled. Using this permuted data, we calculated summed Z scores for each segment. We then counted the number of permuted summed Z scores that were more extreme (i.e., more negative for a negative sum or more positive for a positive sum) than the observed summed Z score for each segment. For one-tailed tests, we calculated a p value by dividing the number of more extreme summed Z scores by 100,000. Two-sided p values were calculated as twice the one-sided p value (or 1 minus the one-sided p value—whichever was greater). Regions were considered to have experiment-wide significance if they had a p-value of $\leq 1\times10^{-5}$, i.e., if there were no permutations with a more extreme summed Z score. This threshold is below a p value of 0.05 after a Bonferroni correction for the number of segments with two or more CNVs (3,225 segments for duplications and 2,720 segments for deletions). The genomic inflation factor was calculated from the observed p values as previously described.[12,13] All statistical analyses and permutations for the regional association analyses were performed with custom Perl (v5.8) and R (v2.11) scripts.

To extend the findings in the initial clinical cohort, we used CNV data from the National Heart, Lung, and Blood Institute (NHLBI)'s Candidate Gene Association Resource (CARe), which is a consortium that performs genetic analyses across nine NHLBI cohorts and encompasses cohorts with population-, community-, family-, and hospital-based designs.[14] A total of 7,363 African-American individuals with height data available from the Atherosclerosis Risk in Communities (ARIC), Coronary Artery Risk Development in Young Adults (CARDIA), Cleveland Family Study (CFS), Jackson Heart Study (JHS), and Multi-Ethnic Study of Atherosclerosis (MESA) cohorts were genotyped on the Affymetrix 6.0 platform as part of CARe.[15] Individuals who were removed during the genome-wide analysis of SNP association in this cohort[15] were also excluded from this analysis. In total, 6,892 individuals were analyzed: 2,285 from ARIC, 668 from CARDIA, 373 from CFS, 1,960 from JHS, and 1,606 from MESA. To define adult height in the CARe cohorts, we excluded men <23 years of age and women <21 years of age, as well as individuals >85 years of age. Stratified by cohort and gender, height was regressed on the basis of age and study site, when available. Residuals were normalized to a standard normal distribution. To account for the family structure in CFS, we used the linear mixed effects (LME) routine in genome-wide association analyses with family data.[16] Outliers (>4 SD or <−4 SD) were excluded from the analyses.

CNVs were called by two different methods—one designed for known (more common) CNVs and one for rarer CNVs. In brief, known CNVs (copy-number polymorphisms, or CNPs) were genotyped with Birdsuite[17]

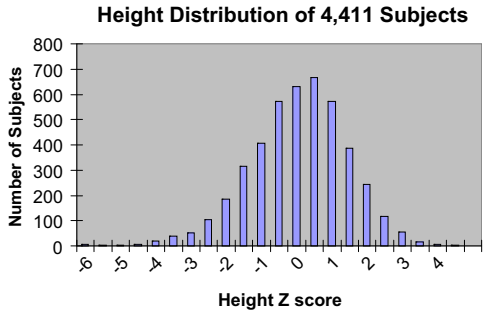

Figure 1. Height Distribution of 4,411 Subjects in the Clinical Cohort

version 1.6, which is based on two recently published CNV maps.[6,18] Rare CNVs were called with a hidden Markov model implemented in Birdsuite, which, for each sample, assigns a discrete copy number at each segment of the genome, along with a LOD score that reflects the confidence of the assignment. Individuals with a total number of non-copy-number-2 CNV calls more than 3 standard deviations from the mean were removed. CNPs with low call rate (<0.9), low frequency (<0.01), and bad calls due to plate effects were removed. Rare CNV segments with LOD scores <3 were removed. A repeat analysis including only those CNVs with LOD scores >5 was performed but did not substantially change the results. For the analysis of common CNVs, only CNPs genotyped by Canary with a frequency of >5% in the cohort were included. For the analysis of lower-frequency and rare CNVs, the common and rare CNVs were merged to create a final dataset of CNV calls. CNV frequency was defined as in the clinical cohort.

Stratified by deletions and duplications, the total number of CNV segments, the total span of CNVs, and the average CNV length for each individual were tabulated. We then regressed the standardized height residuals against each of these measures of CNV burden and corrected for both genotyping-plate membership and the top ten principal components. Analysis was independently conducted for each CNV class (deletions and duplications) and frequency threshold in each CARe cohort, and data were then meta-analyzed for the final evidence of association between global low-frequency CNV burden and stature. CNV association analysis and frequency calculation were done with PLINK v.1.07.[11] Linear regression was performed with R v.2.8.1. Meta-analysis was done with Metal (April 2010 release).[19]

This study complied with all necessary ethical procedures concerning human subjects and was approved by the Children's Hospital Boston institutional review board. A waiver of consent was also granted for the chart review of the clinical cohort. The component CARe studies were approved by local institutional review boards, and the analysis was approved by the Committee on the Use of Humans as Experimental Subjects at the Massachusetts Institute of Technology.

After compiling the CNV and height data in the clinical cohort, we studied the 4,411 subjects who had height data and for whom a clinician had obtained a microarray to test for copy-number variation. Figure 1 shows the height distribution of the 4,411 subjects, of whom 1,463 (33%) were female. The mean and median height Z scores were −0.19 and −0.12, respectively, indicating that this clinically ascertained population was slightly shorter than the overall population. There were 415 short cases with Z scores below −2, 196 tall cases with Z scores above +2, and 3,800 control subjects. We identified a total of 45,644 CNVs, including 23,931 deletions and 21,713 duplications, in 4,403 individuals. Eight subjects did not have any CNVs identified. The respective median values for deletion and duplication sizes were 115 kb and 143 kb. The respective interquartile ranges for deletion and duplication sizes were 70–294 kb and 71–341 kb. Histograms of deletion and duplication lengths are shown in Figures S1 and S2. The mean number, total CNV length, and average CNV length per individual were 10.3 kb, 2,944 kb, and 281 kb, respectively.

We first tested the general hypothesis that there was a greater burden of copy-number variation in the cases (tall and short) than in the normal-height controls from the same cohort. The initial global burden analysis comparing the CNV burden in all 611 cases (tall and short) to that of the 3,800 controls showed a 1.17-fold increase in total CNV length and a 1.15-fold increase in average CNV size in cases ($p = 0.008$ for total length and $p = 0.01$ for average size; Table 1). When we stratified the cases into short cases and tall cases, the associations with copy-number burden were present for short cases but not for tall cases; the statistical significance increased after we limited the cases to short individuals (Table 1). Thus, short stature is associated with increased copy-number burden in this cohort.

We next stratified the association analysis of the short cases both by type of CNV (deletion or duplication) and by CNV frequency (Table 2). Significant associations for total CNV burden and average CNV size were found for deletions but not for duplications. Furthermore, we observed these associations only with lower-frequency (frequency <5%) and rare (frequency <1%) deletions and not with common (frequency >5%) deletions. Compared to those in the controls, rare deletions in short cases showed a 2.30-fold increase in total CNV length and a 1.98-fold increase in average CNV size ($p = 0.005$ and $p = 0.002$, respectively). Additionally, short cases showed higher rates of CNVs in lower-frequency and rare deletions than did the controls. Thus, the association between copy-number burden and short stature is driven by lower-frequency and rare deletions. These associations persisted when we removed individuals with known gene-deletion or -duplication syndromes associated with short stature (Table S2). We attempted to further characterize this association on the basis of CNV size but were not able to identify a specific range of CNV sizes associated with short stature. However,

Table 1. Association Analysis of CNV Burden

	Control	Tall and Short Cases	p Value	Short Cases	p Value	Tall Cases	p Value
Global CNV Burden							
Total Number of CNVs	39,241	6,403		4,307		2,096	
Number of CNVs per individual	10.3	10.5	0.21	10.4	0.41	10.7	0.12
Total CNV burden per individual (kb)	2,879	3,381	**0.008**	3,697	**0.001**	2,711	0.7
Average CNV size per individual (kb)	276	317	**0.01**	348	**0.002**	253	0.85
Genic CNV Burden							
Total Number of CNVs	28,785	4,710		3,178		1,532	
Number of CNVs per individual	7.6	7.7	0.19	7.7	0.31	7.8	0.17
Total CNV burden per individual (kb)	2,259	2,752	**0.009**	3,090	**0.001**	2,036	0.82
Average CNV size per individual (kb)	298	348	**0.01**	390	**0.002**	257	0.96

p values showing significant associations are in bold. Tall and short cases were defined as subjects with Z scores greater than +2 and less than −2 standard deviations respectively. Individuals with Z scores between −2 and +2 were used as controls. "Global CNV" refers to all CNVs in the genome. "Genic CNV" refers to all CNVs within 20 kb of a known gene.

short cases showed higher rates of lower-frequency deletions with length <100 kb and >500 kb ($p = 0.01$ and $p = 0.004$, respectively) than did the controls.

In addition to assessing total CNV burden, we repeated our analyses by using only genic CNVs (deletions or duplications within 20 kb of known genes; Table 1 and Table S3). Compared to the control subjects, all cases (short and tall) had a 1.22-fold increase in genic CNV burden ($p = 0.009$) and a 1.17-fold increase in average CNV length ($p = 0.01$). As with the overall burden analyses, an association was present for short cases (1.37-fold increased burden and 1.31-fold increased length, $p = 0.001$ and 0.002, respectively) but not for tall cases. Within the short cases, we observed the association with deletions ($p = 0.003$ and $p = 0.0002$ for total burden and average length, respectively) but not with duplications. Further stratification by CNV frequency in this smaller set of CNVs did not yield any convincingly significant results (Table S3).

Additionally, we performed linear-regression analyses of CNV burden versus height as a quantitative trait rather than as a case-control analysis. We observed all of the associations seen in our case-control analyses. Specifically, the analyses showed an inverse correlation between height and both the total CNV burden ($p = 0.013$) and average CNV size ($p = 3.1 \times 10^{-5}$) (Table S4 and Figure S3). These findings were stronger for genic CNVs and were driven by low-frequency and rare deletions (Table S5 and Figure S4). To explore whether the effects of CNV burden in the quantitative-trait analysis differ between shorter and taller individuals, we divided the cohort into subgroups with height Z scores greater than or less than zero, and the associations were only present for those subjects with height Z scores below zero (Table S4). This result is also consistent with the case-control analysis. In this quantitative-trait analysis, we saw an association between common deletion burden and increased height, but this result was not seen in the case-control analysis and did not replicate in the population-based cohort (see below).

Because our clinical cohort represents a highly ascertained population that underwent CNV analysis as a result of clinical concerns, the subjects are probably enriched for rare CNVs. Thus, it is difficult to draw any more general conclusions regarding the etiological role of CNVs in short stature. We asked whether our findings that short individuals show an increased burden of deletions would extend into a population-based cohort. To this end, we used CNV data from NHLBI's CARe.[14,15] In total, we analyzed the CNV burden of 6,892 African-American individuals from five population-based cohorts. We assessed the association—stratified by deletions and duplications—between three CNV burden measures (total number of rare CNVs, total CNV burden, and average CNV length) and height as a quantitative phenotype. In this analysis, we adjusted for covariates, including the top ten principal components of ancestry on the basis of genome-wide SNP genotypes from CARe (Table 3).

In general, we observed that shorter stature is associated with a significantly increased total burden of lower-frequency CNV deletions ($p = 0.015$) (Table 3). Each of the three measures of CNV burden (total burden, average CNV size, and total number of CNV segments) was significantly associated with height when we confined our analysis to CNVs within 20 kb of genes (Table 3), suggesting that deletions disrupting gene function might have a role in shorter stature. In the CARe cohort this translates into an average 0.3 cm decrement in height for every 1 Mb increase in lower-frequency deletion burden and a decrement of 0.4 cm for every 1 Mb increase in lower-frequency genic deletion burden. We also examined associations between total burdens of lower-frequency duplications and stature and did not observe any significant trends, although our power to detect associations is lower as

Table 2. Association Analysis of CNV Burden in 415 Short Cases versus 3,800 Controls

	All Frequencies			Common (>5%)			Lower Frequency (<5%)			Rare (<1%)		
	Case	Control	p Value	Case	Control	p Value	Case	Control	p Value	Case	Control	p Value
Deletions and Duplications												
Total number of CNVs	4,307	39,241		2,985	27,498		1,373	12,236		819	6,985	
Number of CNVs per individual	10.4	10.3	0.41	7.2	7.2	0.6	3.3	3.2	0.25	2	1.8	0.12
Total CNV burden per individual (kb)	3,697	2,879	**0.001**	1,725	1,768	0.77	2,107	1,226	**0.0006**	2,052	1,068	**0.001**
Average CNV size per individual (kb)	348	276	**0.002**	233	237	0.74	534	359	**0.003**	774	449	**0.0004**
Deletions Only												
Total number of CNVs	2,253	20,589		1,434	13,812		826	6,840		446	3,549	
Number of CNVs per individual	5.4	5.4	0.47	3.6	3.6	0.92	2	1.8	**0.02**	1.1	0.9	**0.02**
Total CNV burden per individual (kb)	2,073	1,421	**0.003**	814	849	0.78	1,567	743	**0.002**	1,786	776	**0.005**
Average CNV size per individual (kb)	415	250	**0.0002**	205	202	0.35	595	314	**0.0005**	888	449	**0.002**
Duplications Only												
Total number of CNVs	2,054	18,652		1,269	11,368		809	7,462		494	4,504	
Number of CNVs per individual	4.9	4.9	0.39	3.1	3	0.26	1.9	2	0.51	1.2	1.2	0.4
Total CNV burden per individual (kb)	1,673	1,503	0.08	865	866	0.5	1,086	867	0.05	993	743	0.07
Average CNV size per individual (kb)	310	295	0.22	248	260	0.84	417	372	0.20	486	418	0.20

p values showing significant associations are in bold.

a result of fewer duplications (Table 3). For common (>5% frequency) CNVs, we did not see any associations with height except for a weak association between the number of common deletions and decreasing height ($p = 0.036$). This is the opposite of the effect seen for common deletions in the clinical cohort. All of our findings were consistent across the five cohorts within CARe and showed little evidence of heterogeneity.

As an additional exploratory analysis, we examined whether any regions of the genome showed associations between height and copy-number variation of any kind. We divided the genome into segments and tested each segment for an association between height and either deletions or duplications within that segment; significance was assessed by permutation. Height proved to be associated much more with deletions than with duplications: Genomic inflation factors for deletions and duplications were 1.75 and 1.17, respectively (Figure 2). Three preliminary candidate regions were identified as having significant associations with stature after we corrected for multiple testing (nominal $p < 1\times10^{-5}$) in our cohort: a duplication at 11q11 and deletions at 14q11.2 and 17q21.31 (Figures S5–S7, Table S6). Detection of these CNVs was confirmed by multiplex ligation-dependent probe amplification (Figure S8). For each of these three regions, deletions and duplications had opposite effects on height. These regions all display common copy-number variation,[20] and 22%–49% of the individuals in our sample had CNVs in each region. Therefore, these regions do not account for our earlier observation of an excess of rare deletions in individuals with short stature. In addition to these three areas with common CNVs, we observed a 30 kb region that was directly adjacent to the 11q11 polymorphism and that was associated with increased height when deleted. However, this result is only based on the observation of a regional deletion in two subjects with extremely tall stature (Z scores of +3.3 and +4.3) and thus requires additional replication if we are to be certain of its validity.

Our study demonstrates an increased global burden of CNVs in individuals with short stature. This increased burden is driven by an excess of lower-frequency deletions. Furthermore, deletions intersecting with known genes were more common among cases of short stature than among controls. Similar findings of increased genic CNV burden have been found in prior studies of complex traits; an increase in rare genic CNVs has been found with autism,[9] and rare large (>100 kb) CNVs have been found

Table 3. CNV Association Analysis in Population-Based Replication Cohort

	Common (>5%)			Lower Frequency (<5%)			Rare (<1%)		
	B	Standard Error	p	B	Standard Error	p	B	Standard Error	p
Global Deletions									
Number of CNVs per individual (×1000)	−2.94	1.40	**0.036**	−1.82	1.37	0.18	−1.33	1.71	0.44
Total CNV burden per individual (Mb)	−0.088	0.053	0.098	−0.048	0.020	**0.015**	−0.038	0.023	0.10
Average CNV size per individual (Mb)	−6.72	7.95	0.40	−0.91	0.59	0.13	−0.29	0.30	0.34
Global Duplications									
Number of CNVs per individual (×1000)	−2.12	6.08	0.73	−0.078	1.21	0.95	0.11	1.25	0.93
Total CNV burden per individual (Mb)	−0.030	0.074	0.68	−0.0016	0.0053	0.77	−0.0011	0.0055	0.84
Average CNV size per individual (Mb)	−0.24	1.03	0.82	−0.022	0.10	0.83	0.0063	0.061	0.92
Genic Deletions									
Number of CNVs per individual (×1000)	−2.82	1.70	0.097	−4.37	2.04	**0.033**	−4.19	2.71	0.12
Total CNV burden per individual (Mb)	−0.052	0.062	0.40	−0.064	0.022	**0.0033**	−0.051	0.025	**0.045**
Average CNV size per individual (Mb)	−0.42	6.17	0.95	−0.94	0.46	**0.042**	−0.38	0.23	0.089
Genic Duplications									
Number of CNVs per individual (×1000)	−2.43	6.39	0.70	−0.11	1.75	0.95	0.53	1.84	0.77
Total CNV burden per individual (Mb)	−0.037	0.075	0.62	−0.0017	0.0055	0.75	−0.0013	0.0056	0.82
Average CNV size per individual (Mb)	−0.25	0.89	0.78	−0.032	0.082	0.69	0.0038	0.055	0.95

p values showing nominally significant associations are in bold. The beta value indicates the magnitude of the increase in the height Z score given the number of CNVs per individual (in thousands) or the length of CNVs in Mb based on the linear regression.

with schizophrenia.[8] The global burdens of duplications and common deletions were not associated with short stature in our clinical cohort. Interestingly, an increased CNV burden was found in short cases but not in tall cases in our clinical cohort, suggesting that lower-frequency deletions are more likely to impair growth than to cause overgrowth. We extended these findings by performing a quantitative-trait association analysis in a population-based cohort and observed a correlation between lower-frequency genic deletions and decreasing height. This finding strongly supports our hypothesis that an increasing burden of lower-frequency deletions can impair growth and suggests that this phenomenon extends to the general population.

In addition to analyzing the increase in global copy-number burden, we performed an exploratory analysis of regional association of CNVs with stature in our clinical cohort. Overall, we observed far more significant p values than we had expected to find, indicating that there might be many regions with copy-number variants that affect stature. However, we cannot exclude population stratification as a contributing factor leading to this inflation of test statistics. We identified three candidate regions associated with stature in our clinical cohort; none of these have known associations with stature or contain obvious candidate genes. These preliminary regional associations all require confirmation in additional cohorts. It is interesting to note that two of these regions have multiple olfactory receptor genes present. Although these are highly polymorphic regions from the standpoint of copy-number variation, our data set contains other highly CNV-polymorphic regions that did not show any association with height (data not shown). Of note, 11q11 copy-number variation was recently found to be significantly associated with early-onset obesity at a genome-wide level.[4] In that study, deletion of this region was associated with an increased risk of obesity; our study showed duplication of the region to be associated with decreased stature. These results could be complementary because obesity is often associated with increased linear growth in childhood.

One prior study explicitly examined the role of CNVs in height in a population cohort of 618 elderly Chinese Han subjects,[21] but did not examine global copy-number burden. The researchers identified four possible regional associations, although none of them reached a genome-wide significance level. None of their four regions reached experiment-wide significance in our study, although two of the regions from their study (on 16p12.1 and 9p23) did reach nominal significance ($p < 0.05$) in the duplication analysis in our cohort. Additionally, Kang et al. had previously performed a genome-wide association analysis of anthropometric traits in ~2,000 individuals of African ancestry; in that analysis, they looked at the association between common CNVs and height.[22] They did not find any CNV that reached genome-wide significance for an

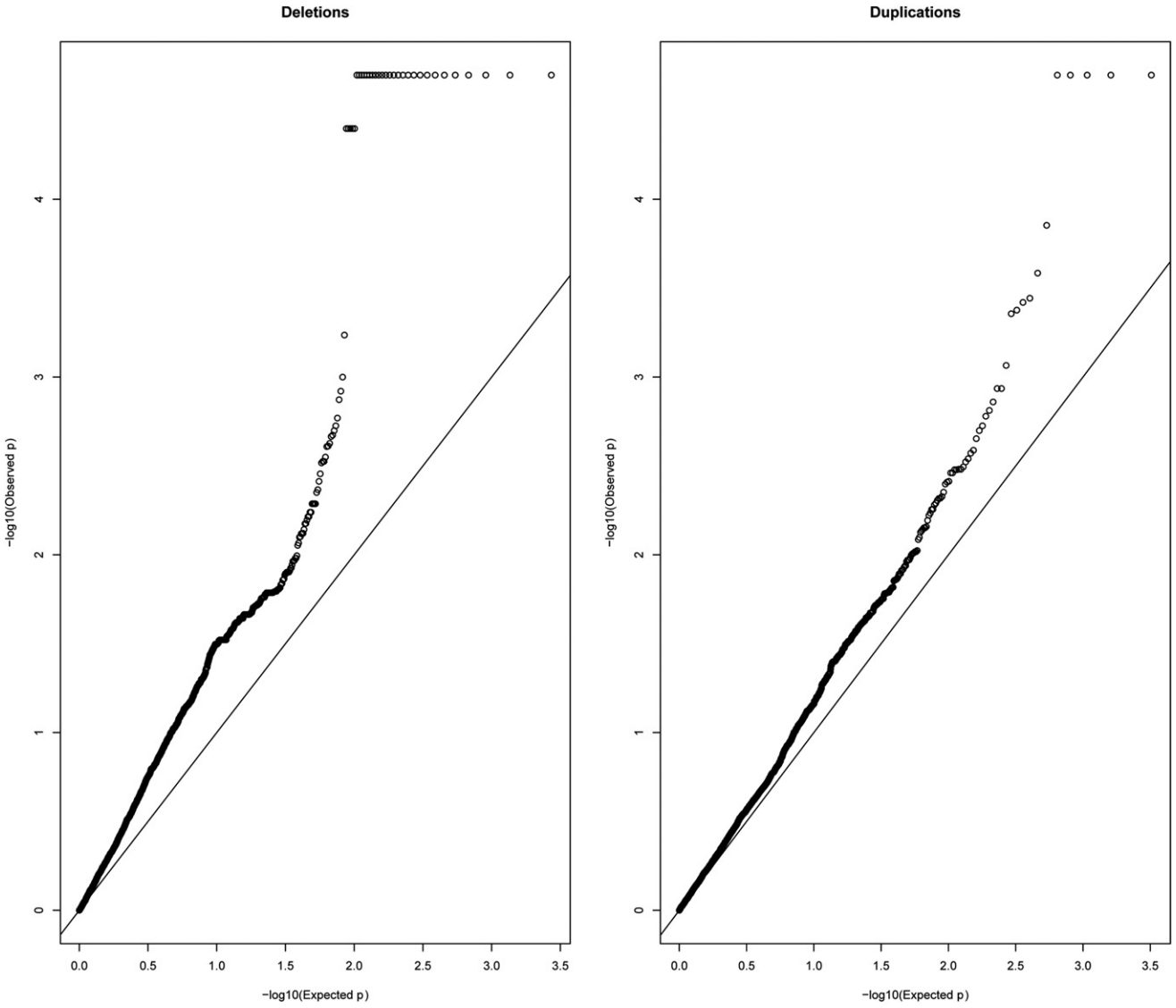

Figure 2. Quantile-Quantile Plot of p Values for Regional-CNV Association Analysis
Quantile-quantile plots for deletion and duplication CNVs are plotted separately. The *x* axis represents the negative log p value for the expected distribution of p values, and the *y* axis represents the negative log p value for the observed p values. The circles forming a horizontal line at a negative log p value of ~5 on the *y* axis represent regions where there are no simulations with summed Z scores more extreme than our observed data. These areas represent regions with CNVs whose associations with stature have experiment-wide significance.

association with height, and they did not perform a global burden analysis for height.

There are a number of important limitations to our study. First, our initial cohort is a clinically ascertained cohort consisting of subjects who had undergone CGH analysis for clinical reasons. Typical reasons for performing this analysis in our institution include developmental delay, autism spectrum disorders, and multiple congenital anomalies. Therefore, it is possible that individuals with an increased burden of rare CNVs are more likely to have increased severity of an underlying disease, leading to poor growth. However, our population-based analysis suggests that this finding might be generalizable to a more representative population. Further replication studies examining the role of copy-number variation in growth will be needed as additional cohorts with CNV data become available.

A second limitation of our study is that we do not have access to information about the genetic ancestry of the subjects in the clinical cohort. It is possible that some of the associations seen in our sample might be due to population stratification, although ancestry was controlled for in analyses of the population-based CARe cohort. Furthermore, our population-based study was performed exclusively in African-American individuals and thus might not be generalizable to other ancestries.

The third and final limitation is that the platform used for determining the CNVs in the clinical cohort does not allow for fine resolution of CNV length; it is possible that multiple small copy-number polymorphisms were merged

into a larger CNV. Given this limitation, our association analysis of individual regions was based on genomic location rather than on individual CNVs. It is possible, and perhaps even likely given the apparent size of the CNVs in these regions, that our candidate regions actually represented an overall local burden of multiple small common copy-number polymorphisms as opposed to an association with individual larger and common CNVs present in those regions. Given these concerns, we are not able to definitively highlight particular causal CNVs but rather can suggest regions of the genome in which copy-number variation, perhaps integrated over a local region, might affect height.

In conclusion, we have performed a genome-wide association study of copy-number burden and height in a clinical cohort. We found a significant increase in the global burden of lower-frequency deletions and genic deletions in a clinically ascertained population, demonstrating that individuals who meet clinical indications for CNV analysis and have short stature are more likely to have a lower-frequency deletion. This finding suggests that the presence of short stature might appropriately lower the clinical threshold for obtaining a microarray-based analysis of copy-number variation. We extended these findings into a population-based cohort of African-Americans and observed a similar outcome. Our results suggest that lower-frequency copy-number variants play a role in the genetic basis of height. Thus, height, as a model polygenic trait, is influenced by many rare sequence variants with large effect size and hundreds of common SNPs with small effect sizes, as well as low-frequency copy-number changes and, in particular, genic deletions. Additional studies of height and other polygenic traits will be needed if we are to further assess the contribution of copy-number variation to human phenotypic variation.

Supplemental Data

Supplemental Data include six tables, eight figures, and funding acknowledgments for the CARe cohorts and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

We would like to thank Guillaume Lettre and the CARe anthropometric working group for generating the height Z scores for the CARe cohort. We would like to thank Steve McCaroll, Joshua Korn, and James Nemesh for their assistance in calling the CNVs in the CARe cohorts. Funding for the work with the clinical cohort was supported by the March of Dimes 6-FY09-507 (J.N.H.), by the National Basic Research Program of China (973 Program) (2010CB529601) (B.L.W.), and by the Science and Technology Council of Shanghai (09JC1402400 and 09ZR1404500) (B.L.W.).

For the CARe cohort, the authors wish to acknowledge the support of the National Heart, Lung, and Blood Institute and the contributions of the research institutions, study investigators, field staff, and study participants in creating this resource for biomedical research. The ARIC, CARDIA, CFS, JHS, and MESA studies contributed parent study data, ancillary study data, and DNA samples through the Broad Institute (N01-HC-65226) to create this genotype/phenotype database for wide dissemination to the biomedical research community. Further acknowledgments of the individual CARe cohort funding can be found in the supplemental materials and include the following grants: N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, N01-HC-55022, R01HL087641, R01HL59367, R01HL086694, RC2 HL102419, U01HG004402, HHSN268200625226C, UL1RR025005, N01-HC-95095, N01-HC-48047, N01-HC-48048, N01-HC-48049, N01-HC-48050, N01-HC-45134, N01-HC-05187, N01-HC-45205, N01-HC-45204, HL46380-01-16, N01-HC-95170, N01-HC-95171, N01-HC-95172, N01 HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, RR-024156, 5RC1HL099911, and 1DP3DK085695.

Received: July 27, 2011
Revised: October 26, 2011
Accepted: October 28, 2011
Published online: November 23, 2011

Web Resources

The URLs for data presented herein are as follows:

Metal, http://www.sph.umich.edu/csg/abecasis/Metal/index.html
PLINK, http://pngu.mgh.harvard.edu/~purcell/plink/

References

1. Hirschhorn, J.N., and Lettre, G. (2009). Progress in genome-wide association studies of human height. Horm. Res. *71* (*Suppl 2*), 5–13.
2. Lettre, G. (2011). Recent progress in the study of the genetics of height. Hum. Genet. *129*, 465–472.
3. Lango Allen, H., Estrada, K., Lettre, G., Berndt, S.I., Weedon, M.N., Rivadeneira, F., Willer, C.J., Jackson, A.U., Vedantam, S., Raychaudhuri, S., et al. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature *467*, 832–838.
4. Jarick, I., Vogel, C.I., Scherag, S., Schäfer, H., Hebebrand, J., Hinney, A., and Scherag, A. (2011). Novel common copy number variation for early onset extreme obesity on chromosome 11q11 identified by a genome-wide analysis. Hum. Mol. Genet. *20*, 840–852.
5. McCarroll, S.A., Huett, A., Kuballa, P., Chilewski, S.D., Landry, A., Goyette, P., Zody, M.C., Hall, J.L., Brant, S.R., Cho, J.H., et al. (2008). Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. Nat. Genet. *40*, 1107–1112.
6. Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., et al; Wellcome Trust Case Control Consortium. (2010). Origins and functional impact of copy number variation in the human genome. Nature *464*, 704–712.
7. Bochukova, E.G., Huang, N., Keogh, J., Henning, E., Purmann, C., Blaszczyk, K., Saeed, S., Hamilton-Shield, J., Clayton-Smith, J., O'Rahilly, S., et al. (2010). Large, rare chromosomal deletions associated with severe early-onset obesity. Nature *463*, 666–670.

8. Consortium, T.I.S.; International Schizophrenia Consortium. (2008). Rare chromosomal deletions and duplications increase risk of schizophrenia. Nature *455*, 237–241.
9. Pinto, D., Pagnamenta, A.T., Klei, L., Anney, R., Merico, D., Regan, R., Conroy, J., Magalhaes, T.R., Correia, C., Abrahams, B.S., et al. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. Nature *466*, 368–372.
10. Kuczmarski, R., Ogden, C., and Guo, S. (2002). 2000 CDC growth charts for the United States: Methods and development. Vital Health Stat. 11 *246*, 1–190.
11. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J., and Sham, P.C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.
12. Devlin, B., and Roeder, K. (1999). Genomic control for association studies. Biometrics *55*, 997–1004.
13. Reich, D.E., and Goldstein, D.B. (2001). Detecting association in a case-control study while correcting for population stratification. Genet. Epidemiol. *20*, 4–16.
14. Musunuru, K., Lettre, G., Young, T., Farlow, D.N., Pirruccello, J.P., Ejebe, K.G., Keating, B.J., Yang, Q., Chen, M.H., Lapchyk, N., et al; NHLBI Candidate Gene Association Resource. (2010). Candidate gene association resource (CARe): Design, methods, and proof of concept. Circ Cardiovasc Genet *3*, 267–275.
15. Lettre, G., Palmer, C.D., Young, T., Ejebe, K.G., Allayee, H., Benjamin, E.J., Bennett, F., Bowden, D.W., Chakravarti, A., Dreisbach, A., et al. (2011). Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project. PLoS Genet. *7*, e1001300.
16. Chen, M.H., and Yang, Q. (2010). GWAF: an R package for genome-wide association analyses with family data. Bioinformatics *26*, 580–581.
17. Korn, J.M., Kuruvilla, F.G., McCarroll, S.A., Wysoker, A., Nemesh, J., Cawley, S., Hubbell, E., Veitch, J., Collins, P.J., Darvishi, K., et al. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat. Genet. *40*, 1253–1260.
18. Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., Peltonen, L., et al; International HapMap 3 Consortium. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.
19. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: Fast and efficient meta-analysis of genomewide association scans. Bioinformatics *26*, 2190–2191.
20. Kollars, J., Zarroug, A.E., van Heerden, J., Lteif, A., Stavlo, P., Suarez, L., Moir, C., Ishitani, M., and Rodeberg, D. (2005). Primary hyperparathyroidism in pediatric patients. Pediatrics *115*, 974–980.
21. Li, X., Tan, L., Liu, X., Lei, S., Yang, T., Chen, X., Zhang, F., Fang, Y., Guo, Y., Zhang, L., et al. (2010). A genome wide association study between copy number variation (CNV) and human height in Chinese population. J. Genet. Genomics *37*, 779–785.
22. Kang, S.J., Chiang, C.W., Palmer, C.D., Tayo, B.O., Lettre, G., Butler, J.L., Hackett, R., Adeyemo, A.A., Guiducci, C., Berzins, I., et al. (2010). Genome-wide association of anthropometric traits in African- and African-derived populations. Hum. Mol. Genet. *19*, 2725–2738.