AD_________________

Award Number: W81XWH09-1-0694

TITLE: Assessing the role of copy number variants in prostate cancer risk and progression using a novel genomewide screening method

PRINCIPAL INVESTIGATOR: Donna Lehman, PhD

CONTRACTING ORGANIZATION:
University of Texas Health Science Center San Antonio
San Antonio TX 78229-3901

REPORT DATE: December 2014

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

X Approved for public release; distribution unlimited

□

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE December 2014	**2. REPORT TYPE** Final	**3. DATES COVERED** 15Sep2009 to 14Sep2014

4. TITLE AND SUBTITLE
Assessing the role of copy number variants in prostate cancer risk and progression using a novel genomewide screening method.

5a. CONTRACT NUMBER
W81XWH-09-1-0694

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Donna Lehman
Robin Leach

E-Mail: lehman@uthscsa.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Texas Health Science Center
Office of Grants Management
7703 Floyd Curl Drive
San Antonio TX 78229-3901

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Individual copy number variations in the genome may play a substantial role in influencing trait variation, yet due to technical limitations they have been understudied. We have performed the first genome-wide association of copy number variants and risk for prostate cancer in Mexican Americans. Our results from direct testing of CNVs are consistent with the growing consensus that there are not common genetic variants of large additive effects on prostate cancer predisposition, regardless of variant type. We found a highly protective rare deletion on 8q24 which is present in Mexican Americans but extremely rare in Caucasians. Due to the strong effect of this deletion, this discovery has implications for prostate cancer risk assessment and for understanding the etiology of prostate cancer. This variant warrants further study. We have also identified a rare 900 bp deletion in the PTEN gene to be associated with increased risk for prostate cancer and have provided confirmatory data that a rare heritable deletion on 2p24.3 is associated with prostate cancer risk in non-Hispanic Caucasians. These data support our hypothesis that heritable structural variation may affect risk for prostate cancer and/or its progression, but these variants are likely rare. Moreover, these variants may be unique to ethnic population and underscores the need to investigate genetic risk in multiple populations. As genes are identified from these studies, they may prove to be both useful biomarkers for early diagnosis and/or novel therapeutic targets for both prevention and treatment of prostate cancer.

15. SUBJECT TERMS
Heritable copy number variation, prostate cancer, Hispanic

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	28	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Page

INTRODUCTION

Prostate cancer is known to have a strong genetic component. Thus, the identification of the heritable genetic alteration(s) that precedes or increases susceptibility to somatic cancerous changes in the prostate could likely lead to improved identification of high risk individuals for early screening and possibly to new treatment strategies. Standard methodologies, including linkage analysis in familial prostate cancer patients and genome-wide single nucleotide polymorphism (SNP) screening have not identified sufficient genetic alterations to account for the hereditary component of prostate cancer. Recently, it has become apparent that structural variation comprises similar diversity of human genomes as SNPs and may play a significant role in disease susceptibility and resistance. Since CNV regions often contain genes, parts of genes, or regulatory regions, they could result in different levels of gene expression. In addition, through deletion between genes or insertion of duplicated sequences into a gene, CNVs may also contribute to creation of new genes. Thus, they may play a substantial role in influencing trait variation, yet due to technical limitations they have been understudied, and at the outset of this project little was known about this new class of variant, including their distribution in most human populations and impact on common diseases. The goal of this research project was to screen the entire autosomal genome for copy number variation (CNVs) in constitutional DNA to assess their role in risk of development of prostate cancer and then evaluate any direct effect on the prostate. Differences in these genetic risk factors may partially explain the ethnic disparities in incidence of prostate cancer. This study was one of few examining genetic risk factors among the Mexican American population.

KEYWORDS: Heritable copy number variation, prostate cancer, Hispanic

OVERALL PROJECT SUMMARY

The study participants were part of the San Antonio Center for Biomarkers of Risk of Prostate Cancer (SABOR) cohort. SABOR is one of three prostate cancer centers in the country funded by the National Cancer Institute and has been prospectively enrolling healthy male volunteers in San Antonio since 2001. As of 2008, over 3,500 men had been enrolled and screened annually. Age criteria are not exclusionary; however, patients are counseled that they will best benefit from enrollment if over the age of 40 years and are high risk (African American or family history), or if over the age of 50 with low risk. The cohort is approximately 50% non-Hispanic Caucasian, 36% Hispanic Caucasians and 14% African American. Data collected on exam include a medical and surgical history, anthropometric measures such as body mass index (bmi), information on use of dietary supplements, medication records, and family cancer history. In addition, a digital rectal examination (DRE) is performed and prostate specific antigen (PSA) level is determined annually

We had previously screened the entire genome of 100 Hispanic prostate cancer subjects and 67 Hispanic controls of the SABOR cohort for copy number variants using an Infinium-based array by Illumina that covered all published CNVs as well as an additional ~13,000 regions not previously covered on SNP arrays. These regions include segmental duplications, megasatellites, and regions lacking SNPs. Coverage with this tool was provided by 44,220 SNPs or non-polymorphic probes representing ~29,000 segments, 15,559 of which are non-redundant segments. In collaboration with DeCode Genetics, the microarray genotyping data underwent preprocessing to remove noise and artifacts using DeCode´s unique protocol based on in-house data models and analytical methods developed using a large body of proprietary data for the CNV chips. Next, the number of copies of each "allele" was estimated using information from the intensity values for each probe. The association between prostate cancer and each polymorphic marker was tested using logistic regression analysis (using a logit link function). A likelihood ratio test was performed comparing the null hypothesis of no association to the two-sided alternative hypothesis of association. In order to minimize the effects of confounding, relevant covariates such as age were included in the model. The model was also adjusted for potential confounding by admixture (genetic population substructure) using principal components methodology. We used an additive genetic model to model the effect of the CNV, such that each additional copy of a variant would increase (or decrease) the trait by the same amount and used a Bonferroni correction in interpreting statistical significance. Using a conservative Bonferroni corrected significance threshold of $p \leq 10^{-6}$, which correlates with an experiment-wise p-value of 0.05, we observed 13 unique probes to be significantly associated with prostate cancer. We observed 25 associated CNV loci at a significance threshold of $p \leq 10^{-5}$.

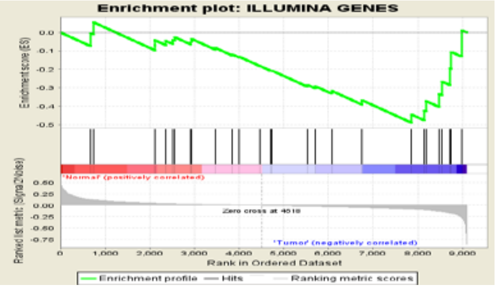
During the first year of this project, we attempted to test the associated CNVs in all SABOR participants of Mexican American descent. Although 3 CNVs remained significantly associated in the entire dataset, we quickly observed the difficulty of designing assays for these complex regions. Therefore, we adjusted our aims by 1) supplementing with in silico gene set enrichment assays of existing datasets, and 2) conducting additional discovery assays in a second set of subjects using a dense whole-genome SNP array followed by computational analyses to identify and genotype CNVs.

GSEA: Since a number of the associated CNV regions could not be validated using the method of quantitative PCR, we sought to find a means to help predict which gene or genes in those CNV regions may be affected by the variant and influence prostate cancer risk, thereby prioritizing genes for further analysis. Task 2 of our SOW is to identify candidate genes from CNV regions and assess any effects on expression of these genes. For this task we proposed to select candidate genes by comparison of current genomic database information with the current literature. We augmented this approach by conducting a gene set enrichment analysis (GSEA) using published expression data from prostate tissues. Briefly, the closest genes to the 25 most significantly associated variants were identified as a gene set. Next, we used Robust Probe-level Linear Model to normalize Affymetrix HG-U95Av2 expression data from Gene Expression Omnibus for 16 disease-free prostate tissue samples, and 20 prostate tumor samples and their adjacent normal prostate tissue samples, using affylmGUI in R. The expression data used is a subset of the dataset record GDS2545 available at Gene Expression Omnibus5. GDS2545 is a dataset of normal prostate, prostate cancer, adjacent normal, and metastatic prostate cancer tissues analyzed on the HG-U95Av2 array from Affymetrix. affylmGUI is an open source software available through Bioconductor (http://www.bioconductor.org/index.html) that normalizes Affymetrix expression data using limma, another program available through Bioconductor. Of the 94 genes associated with prostate cancer, 27 genes were annotated on the HG-U95Av2 chip. These 27 genes were used in the subsequent Gene Set Enrichment Analysis using the Java version at http://www.broadinstitute.org/gsea/. We also performed simulations using random sets of genes taken from the Database of Genomic Variants (http://projects.tcag.ca/variation) in order to account for potential Type 1 error due to possible association of hypervariable regions that may preferentially undergo genomic alterations in cancer. Using GSEA it was determined that this CNVR gene set showed enrichment in tumor tissue and in adjacent normal tissue compared to disease free normal tissue. *This work was presented at the 59th Annual meeting of the American Society of Human Genetics, Honolulu HI.* The genes that contribute to the core enrichment in both GSEA comparisons are shown in the column labeled “Overlap” in table 1.

Table 1. Results of GSEA with Genes in CNV regions.

Gene Set for GSEA		Leading Edge 'cancer vs. normal'	Leading Edge 'adjacent vs. normal'	Overlap	# Markers on Illumina array
UGT8	ARIH2	UGT8	UGT8	UGT8	3
HINT1	CYFIP1	HINT1	HINT1	HINT1	4
PTPRK	UQCRB	PTPRK	PTPRK	PTPRK	1
IGHM	TYRP1	IGHM	IGHM	IGHM	12
PPEF2	SETBP1	PPEF2	PPEF2	PPEF2	7
PTPRN2	TIPRL	PTPRN2	PTPRN2	PTPRN2	1
PSPC1	ADAM3A	PSPC1	PSPC1	PSPC1	4
AUTS2	ULK1	TYRP1	ULK1		
OTUD4	HINT1		HINT1		
COX10	BTG1		BTG1		
PPIE	ZBTB20		ZBTB20		
SLITRK3	GRIA1		AUTS2		
PDE1C	MAGI1				
LY6E					

	Normal vs. Adjacent	Normal vs. Tumor
Normalized Enrichment Score	-1.372	-1.497
Nominal p-value	0.091	0.036



CNV detection through SNP array: We selected cases with the earliest age-at-onset of prostate cancer and controls that were among the oldest controls. The samples were run on the Illumina OmniExpress array and the data analyzed using the programs PennCNV[1] and QuantiSNP[2]. We genotyped 192 samples total: 96 cases, 92 controls, and 4 replicate samples. All 4 replicate samples had SNP genotyping reproducibility rates >0.9999. All samples had SNP genotyping call rates >0.99. Three samples (2 cases and 1 control) did not meet criteria for CNV calling due to LogR ratios having standard deviation >0.3 (a standard QC setting in the program PennCNV). Therefore, for association testing in CNVtools there were 94 cases and 91 controls. The average age of cases genotyped on the OmniExpress was 60.43 ± 6.33 years and the average age of controls was 70.88 ± 5.89 years. After QC criteria, the average age of cases used for association testing with CNVtools was 60.36 ± 6.36, and the average age for controls was 70.89 ± 5.93. The groups did not differ in admixture estimates based upon the individual measures that were previously calculated for this cohort[3] as shown in Table 2.

Table 2. Admixture estimates of cases and controls

Admixture estimate	Cases	Controls	P-value
% European American	0.587 ± 0.180	0.615 ± 0.182	0.38
% Native American	0.385 ± 0.185	0.358 ± 0.189	0.41
% African American	0.028 ± 0.040	0.027 ± 0.039	0.82

We identified 462 copy number variants (CNVs) which were polymorphic in at least 2 individuals. We also applied the newly released program IMPUTE2.2, the 1000 Genomes project reference panel, and our SNP genotyping data, to impute known CNVs in our samples. We used allelic association in the program PLINK to test for association with cancer status. In total, 3725 CNVs with minor allele frequency (MAF) >0.01 were identified and tested. None of the 760 common CNVs with MAF >0.1 were significantly associated. Our results from direct testing of CNVs are consistent with the growing consensus that there are not common genetic variants of large additive effects on prostate cancer predisposition, regardless of variant type.

We next focused on analyzing rare CNVs using multiple methods. We undertook replication efforts of those novel CNVs reported in the literature[4,5] to be associated with prostate cancer and published our results in the Journal of Urologic Oncology (attached in Appendix). Our findings are consistent with the reported observation that a heritable deletion on 2p24.3 is associated with prostate cancer risk in non-Hispanic Caucasians. Additionally, our observations indicate that the 2p24.3 variant is associated with risk for high grade prostate cancer in a recessive manner. In our own San Antonio subjects, we had observed 3 other rare CNVs which were biologically interesting and nominally significant in the entire Mexican American SABOR cohort: deletions at 2p32.3, 8q24 and 10q23.31. The deletion at 2p32.3 was the most strongly associated CNV in this study. Chromosomal region 8q24 is of interest due to its consistent implication in GWA studies for prostate cancer. A rare non-recurrent 8486 base pair deletion on 8q24 (distinct from the MYC locus) was associated with decreased prostate cancer risk in 989 Mexican American men (Odds ratio 0.20, p=0.02). Only 3 of 1530 Caucasians carried the deletion, indicating that this deletion is not likely to affect risk in the Caucasian population. The deleted sequence contains a putative conserved transcription factor binding site for NKX3.1. NKX3.1 is an androgen regulated homeobox gene involved in prostate cancer development, is required for stem cell maintenance, and marks the luminal epithelial cell that is the cell of origin for prostate cancer. The deletion at 10q23.31 is in the gene PTEN, a known prostate cancer tumor suppressor genes. Sequencing data from our lab last year revealed that this deletion is 896 base pairs long with the closest breakpoint residing 57 base pairs from exon 2. We noted that the controls bearing the PTEN deletion were significantly younger than controls not harboring the variant (53.3 yr vs. 60.2, respectively; p=0.002) which leaves the possibility of conversion to prostate cancer early for these men. Therefore, we modeled time to prostate cancer diagnosis as a function of CNV genotype using Cox proportional hazards model and observed significant increase in risk for bearers of this deletion (OR 2.1 95%CI: 1.14-3.87, p=0.017). A manuscript describing these results is currently under review.

We conducted extensive burden testing and found no strong evidence to support any form of enrichment of CNVs, either rare or common, despite investigating the issue using multiple approaches; including a gene-centric approach, a pathway driven approach, gene set enrichment based approaches, and investigating the characteristics of the CNVs themselves. We have not tested interactions between these variants, SNPs and diet, however.

For Task 3, we planned to examine prostate tumors of bearers of common CNVs. However, no common CNVs appear to be involved in prostate cancer development. Analyses of rare CNVs, such as the PTEN deletion (and 2 others) that appear to increase risk for prostate cancer, require tumors from the specific SABOR subjects, but these were not available. We moved forward to maximize the extensive CNV data from this project by combining it with that being generated under the auspices of other projects on which Dr. Robin Leach is an investigator. Whole genome expression and methylation profiles were generated on a second set of prostate cancer samples from the local biorepository. Preliminary data show distinct expression profiles for genes involved in the immune system, among others, to discriminate tissue that is adjacent to tumors that progress vs. tumors that do not progress. Additional biological pathways have been identified that have distinct methylation patterns between tumors which later metastasize vs those that do not. This work is ongoing. We are currently examining these data in the regions of the CNVs that we and others have identified to be associated with prostate cancer in order to investigate their potential function through changes in methylation and/or expression.

KEY RESEARCH ACCOMPLISHMENTS

1. Developed an efficient and comprehensive workflow for identification of copy number variation using dense SNP arrays.
2. Identified and cataloged copy number variation in 2 large Mexican American cohorts, the SABOR and the San Antonio Family Studies consisting of over 2000 subjects. (Blackburn A, Göring HHH, Dean A, Carless MA, Dyer T, Kumar S, Fowler S, Curran JE, Almasy L, Mahaney M, Comuzzie A, Duggirala R, Blangero J, Lehman DM (2013). Utilization of extended pedigree information for discovery and confirmation of copy number variable regions among Mexican Americans. Eur J Hum Genet. 21(4):404-9. PMID:22909773)
3. Identified novel rare CNVs that may play an ethnic specific role in prostate cancer and provided confirmatory evidence for others that have been identified by other groups. (Blackburn A, Wilson D, Gelfond J, Yao L, Hernandez J, Thompson IM, Leach RJ, Lehman DM (2013). Validation of copy number variants associated with prostate cancer risk and prognosis. Urol Oncol. 2014 Jan;32(1):44.e15-20 PMID:24054869)

CONCLUSION

We have performed the first genome-wide association of copy number variants and risk for prostate cancer in Mexican Americans. Our results from direct testing of CNVs are consistent with the growing consensus that there are not common genetic variants of large additive effects on prostate cancer predisposition, regardless of variant type. We found a highly protective rare deletion on 8q24 which is present in Mexican Americans but extremely rare in Caucasians. Due to the strong effect of this deletion, this discovery has implications for prostate cancer risk assessment and for understanding the etiology of prostate cancer. This variant warrants further study. We have also identified a rare 900 bp deletion in the PTEN gene to be associated with increased risk for prostate cancer and have provided confirmatory data that a rare heritable deletion on 2p24.3 is associated with prostate cancer risk in non-Hispanic Caucasians. These data support our hypothesis that heritable structural variation may affect risk for prostate cancer and/or its progression, but these variants are likely rare. Moreover, these variants may be unique to ethnic population and underscores the need to investigate genetic risk in multiple populations. As genes are identified from these studies, they may prove to be both useful biomarkers for early diagnosis and/or novel therapeutic targets for both prevention and treatment of prostate cancer.

PUBLICATIONS, ABSTRACTS, PRESENTATIONS

1. Blackburn A., Gelfond J., Goring H.H., Beuten Y., Thompson I., Leach RJ, and Lehman DM. (2009) Identification of Copy Number Variable Regions (CNVRs) Associated with Risk of Prostate Cancer in Mexican-Americans. Abstract presented at 59th Annual meeting of the American Society of Human Genetics, Honolulu HI, October 2009
2. Blackburn A, Gelfond J, Yao L, Thompson IA, Leach RJ, Lehman DM (2011). A heritable deletion on 8q24 lowers risk for prostate cancer in Mexican Americans. Abstract presented at the Cancer Therapy and Research Center Annual Symposium, UT Health Science Center, San Antonio TX
3. Lehman DM. Identification of Copy Number Variable Regions (CNVRs) Associated with Risk of Prostate Cancer in Mexican-Americans, DOD IMPaCT conference, Health Disparities, Department of Defense, Orlando, FL March 2011 (Invited Speaker)
4. August Blackburn, Jonathan Gelfond, Yao Li, Iriscilla Ayala, Ian Thompson, Robin J. Leach, Donna M. Lehman (2012). A Heritable Deletion on 8q24 Lowers Risk for Prostate Cancer in Mexican Americans. Platform presentation at Texas Genetic Society annual meeting, March 22-24, San Antonio TX
5. Blackburn A, Wilson D, Gelfond J, Yao L, Hernandez J, Thompson IM, Leach RJ, Lehman DM (2013). Validation of copy number variants associated with prostate cancer risk and prognosis. Urol Oncol. 2014 Jan;32(1):44.e15-20 PMID:24054869
6. Blackburn A, Göring HHH, Dean A, Carless MA, Dyer T, Kumar S, Fowler S, Curran JE, Almasy L, Mahaney M, Comuzzie A, Duggirala R, Blangero J, Lehman DM (2013). Utilization of extended pedigree information for discovery and confirmation of copy number variable regions among Mexican Americans. Eur J Hum Genet. 21(4):404-9. PMID:22909773
7. Lehman DM. Rare germline copy number variants and prostate cancer risk among Mexican American men. Prostate Cancer Foundation Annual Retreat, Carlsbad CA 2014 (invited speaker)

INVENTIONS, PATENTS AND LICENSES

None

REPORTABLE OUTCOMES

1. Blackburn A., Gelfond J., Goring H.H., Beuten Y., Thompson I., Leach RJ, and Lehman DM. (2009) Identification of Copy Number Variable Regions (CNVRs) Associated with Risk of Prostate Cancer in Mexican-Americans. Abstract presented at 59th Annual meeting of the American Society of Human Genetics, Honolulu HI, October 2009
2. Blackburn A, Gelfond J, Yao L, Thompson IA, Leach RJ, Lehman DM (2011). A heritable deletion on 8q24 lowers risk for prostate cancer in Mexican Americans. Abstract presented at the Cancer Therapy and Research Center Annual Symposium, UT Health Science Center, San Antonio TX
3. August Blackburn, Jonathan Gelfond, Yao Li, Iriscilla Ayala, Ian Thompson, Robin J. Leach, Donna M. Lehman (2012). A Heritable Deletion on 8q24 Lowers Risk for Prostate Cancer in

Mexican Americans. Platform presentation at Texas Genetic Society annual meeting, March 22-24, San Antonio TX

4. Blackburn A, Wilson D, Gelfond J, Yao L, Hernandez J, Thompson IM, Leach RJ, Lehman DM (2013). Validation of copy number variants associated with prostate cancer risk and prognosis. Urol Oncol. 2014 Jan;32(1):44.e15-20 PMID:24054869
5. Blackburn A, Göring HHH, Dean A, Carless MA, Dyer T, Kumar S, Fowler S, Curran JE, Almasy L, Mahaney M, Comuzzie A, Duggirala R, Blangero J, Lehman DM (2013). Utilization of extended pedigree information for discovery and confirmation of copy number variable regions among Mexican Americans. Eur J Hum Genet. 21(4):404-9. PMID:22909773

OTHER ACHIEVEMENTS

1. August Blackburn, PhD, graduate of Integrated Multidisciplinary Graduate Program, August 2013, UT Health Science Center San Antonio. A major portion of Dr. Blackburn's dissertation project focused on and was funded by this Idea Development Award. This project provided him with necessary training to lead to his PhD degree and to his authorship of a book chapter in press on the topic of CNVs and complex diseases.
2. August Blackburn, PhD, is now a Postdoctoral Scientist at the Texas Biomedical Research Institute, San Antonio TX, based upon experience and training supported by this award.

REFERENCES

1. Wang K, Li M, Hadley D et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 2007 November;17(11):1665-74.
2. Colella S, Yau C, Taylor JM et al. 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):2013-25
3. Beuten J., et al. *Wide disparity in genetic admixture among Mexican Americans from San Antonio, TX.* Ann Hum Genet 2011 July;75(4):529-38.
4. Demichelis, F., et al., *Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk.* Proc.Natl.Acad.Sci.U.S.A, 2012. 109(17): p. 6686-6691.
5. Liu, W., et al., *Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer.* Cancer Res., 2009. 69(6): p. 2176-2179.

APPENDIX

1. Blackburn A, Wilson D, Gelfond J, Yao L, Hernandez J, Thompson IM, Leach RJ, Lehman DM (2013). Validation of copy number variants associated with prostate cancer risk and prognosis. Urol Oncol. 2014 Jan;32(1):44.e15-20 PMID:24054869
2. Blackburn A, Göring HHH, Dean A, Carless MA, Dyer T, Kumar S, Fowler S, Curran JE, Almasy L, Mahaney M, Comuzzie A, Duggirala R, Blangero J, Lehman DM (2013). Utilization of extended pedigree information for discovery and confirmation of copy number variable regions among Mexican Americans. Eur J Hum Genet. 21(4):404-9. PMID:22909773

Personnel: Donna Lehman, Robin Leach, Jon Gelfond, August Blackburn

Published in final edited form as:
Urol Oncol. 2014 January ; 32(1): 44.e15–44.e20. doi:10.1016/j.urolonc.2013.06.004.

Validation of copy number variants associated with prostate cancer risk and prognosis

August Blackburn[#a], **Desiree Wilson**[#a], **Jonathan Gelfond, MD, PhD**[b,e], **Li Yao, MD, PhD**[c], **Javier Hernandez, MD**[d,e], **Ian M Thompson, MD**[d,e], **Robin J Leach, PhD**[a,d,e], and **Donna M Lehman, PhD**[c,e]

[a] University of Texas Health Science Center at San Antonio, Department of Cellular and Structural Biology

[b] University of Texas Health Science Center at San Antonio, Department of Epidemiology and Biostatistics

[c] University of Texas Health Science Center at San Antonio, Department of Medicine – Division of Clinical Epidemiology

[d] University of Texas Health Science Center at San Antonio, Department of Urology

[e] University of Texas Health Science Center at San Antonio, Cancer Therapy and Research Center

[#] These authors contributed equally to this work.

Abstract

Objective—Two recent studies have reported novel heritable copy number variants (CNVs) on chromosomes 2p, 15q, and 12q to be associated with prostate cancer risk in non Hispanic Caucasians. The goal of the current study was to determine whether these findings could be independently confirmed in the Caucasian population from the South Texas area.

Methods and Materials—The study subjects consisted of participants of the San Antonio Biomarkers Of Risk for prostate cancer (SABOR) cohort and additional cases ascertained in the same metropolitan area. We genotyped all 7 of the reported copy number variants using real time quantitative PCR in 1536 (317 cases, 1219 controls) non Hispanic Caucasian men, and additionally genotyped 632 (191 cases, 441 controls) Hispanic Caucasian men for one of these variants, a deletion on 2p24.3.

Results—Association of the deletion on 2p24.3 with overall prostate cancer risk did not meet our significance criteria but was consistent with previous reports [odds ratio (OR), 1.40; 95%

Corresponding author: Donna Lehman, Ph.D. Department of Medicine – Clinical Epidemiology UTHSCSA 7702 Floyd Curl Drive San Antonio, TX 78229 Office phone: (210) 567 6714 lehman@uthscsa.edu.

All other authors can be reached through correspondence with the corresponding author.

There are no personal or other conflicts of interest that need to be disclosed.

confidence interval (95% CI), 0.99 2.00; P = 0.06]. Among Hispanic Caucasians, this deletion is much less prevalent (MAF of 0.059 and 0.024 in non Hispanic and Hispanic Caucasians respectively) and did not show evidence of association with risk for prostate cancer. Interestingly, among non Hispanic Caucasians, carrying a homozygous deletion of 2p24.3 was significantly associated with high grade prostate cancer as defined by a Gleason sum ≥8 [OR, 27.99; 95% CI, 1.99 392.6; P = 0.007 (Fischer's Exact)]. The remaining six copy number variable regions were either not polymorphic in our cohort of non Hispanic Caucasians, or showed no evidence of association.

Conclusions—Our findings are consistent with the reported observation that a heritable deletion on 2p24.3 is associated with prostate cancer risk in non Hispanic Caucasians. Additionally, our observations indicate that the 2p24.3 variant is associated with risk for high grade prostate cancer in a recessive manner. We were unable to replicate any association with prostate cancer for the variants on chromosomes 15q and 12q which may be explained by regional population differences in low frequency variants and disease heterogeneity.

Keywords

Prostate; cancer; risk; deletion; prognosis

Introduction

Prostate cancer is the most common non skin cancer in American men. The majority of prostate cancer cases, however, is indolent and may not require treatments that are associated with significant rates of voiding and sexual function complications. There is great impetus to identify markers that distinguish indolent from aggressive disease. Given that risk for prostate cancer may be attributed to a heritable component in as many as 42% of cases (1), it may be possible to identify genetic polymorphisms that can be used to better predict risk and prognosis. Multiple studies have attempted to identify genetic variants that are associated with risk for prostate cancer.

The contribution of copy number variants (CNV) to risk for complex diseases have not yet been fully elucidated, partially due to the difficulty of identification and accurate genotyping (2). Efforts to characterize common and rare copy number variation in various human populations are still underway. Common CNVs have been shown to be in linkage disequilibrium with adjacent single nucleotide polymorphisms (SNPs) (3 5), which indicates that some common CNVs have already been indirectly assessed for association with traits in SNP based genome wide association studies (GWAS). However, recurrent variants and risk bearing alleles with low minor allele frequencies may not be well tagged by SNPs in GWAS, and these variants will likely require direct assessment.

Few studies have investigated risk for prostate cancer attributed to CNVs directly. Liu *et al.* published the first genome wide investigation of germ line CNVs and risk for prostate cancer, in which they report a germ line deletion spanning 5,947 base pairs at 2p24.3 associated with risk for prostate cancer and aggressive prostate cancer among men of European descent (6). Demichelis *et al*. reported 6 deletions associated with risk for prostate cancer with a false discovery rate <0.2 (7). Two of these variants, deletions on 15q and 12q, were further supported in this same publication with evidence from bioinformatics analysis and functional assays (7).

Genome wide association studies for any genetic variant type are subject to a range of errors and biases, and therefore replication of association signals in independent samples are necessary to confirm true associations(8). One of the primary purposes of the San Antonio Center for Biomarkers of Risk of Prostate Cancer (SABOR), a Clinical and Validation Center of the Early Detection Research Network of the National Cancer Institute, is to independently confirm prostate cancer biomarkers, including genetic variants that are predictive of risk. With this motivation, we report the investigation of these reported variants in non Hispanic and Hispanic Caucasians from San Antonio, Texas.

Materials and Methods

Study subjects

Study subjects consisted of 1372 (153 cases, 1219 controls) non Hispanic Caucasian and 516 (75 cases, 441 controls) Hispanic Caucasian participants of SABOR. SABOR is a prospective longitudinal study that examines behavioral, genetic, and other markers of risk of prostate cancer. All study participants are screened annually for prostate cancer using prostate specific antigen (PSA) serum measurements, with exception to those with PSA <1, who are screened every other year. An additional 164 non Hispanic Caucasian and 116 Hispanic Caucasian men were recruited from a parallel cohort study of prevalent cases (PREF) in the same metropolitan area. This study consists of men who were diagnosed with prostate cancer (PCa) prior to enrollment in the study and were recruited using the same methods as SABOR (local advertisement). Age for this study was calculated in the following manner: for prevalent cases, self reported age of diagnosis was used; for incident cases, age of diagnosis was used; non cancer (control) participants were censored at their most recent SABOR examination age. Self reported ages of diagnosis for most of the prevalent cases were confirmed through medical records. Institutional Review Board approval was obtained from the UT Health Science Center at San Antonio. Informed consent was obtained from all participants of both cohorts. We refer to the total population as the SABOR cohort.

Liu *et al.* defined cancer aggressiveness as meeting any of the following criteria: $T_{3/4}$, N+, M+, Gleason score sum ≥8, or PSA >50 ng/mL(6). Although it would be ideal to use identical criterion to define aggressiveness, PSA serum levels at time of diagnosis as well as staging were not available for all participants of PREF, and thus these variables were not utilized in defining aggressiveness. Gleason scores were more broadly available, and thus, in this study aggressive cases in this study were defined by a Gleason score of ≥8.

CNV genotyping

DNA was isolated from whole blood using QIAampfi DNA Blood Maxi Kit (Qiagen, Valencia, CA). For real time quantitative PCR, all primers and probes were purchased from Applied Biosystems (Valencia CA). The primer/probe pairs were either pre designed or were designed for the regions of interest using Primer Express (Applied Biosystems, Valencia CA).Quantitative PCR primer/probe information is summarized in Table 1. .

For qPCR, primers and probes for a target sequence and reference sequence (RNAse P) were multiplexed in 384 well plates and all samples were run in duplicate. Fluorescence was detected using the 7900HT Real time PCR System (Applied Biosystems, Valencia CA). Real time PCR data was analyzed using the reference free ΔΔct method implemented in CopyCaller software (Applied Biosystems, Valencia CA). This approach estimates the mean Δct for a copy number of 1, and subsequently uses this value to calculate ΔΔct. We have found this method broadly consistent with the ΔΔct method using a reference sample. Discrete copy number calls were determined by plotting histograms of the raw calculated copy number values, which for polymorphic regions reveals non overlapping Gaussian distributions representing integer copy number states(9).

To confirm that CNV calls from qPCR accurately detected the 2p24.3 deletion identified by Liu *et al.* (6), we conducted PCR genotyping as reported by Liu *et al.* (6) for all homozygous deletions, 12 heterozygous deletions, and 14 wild type individuals, including 4 individuals on the Gaussian tails where the distributions approach each other in the non Hispanic Caucasian samples. We observed 100% concordance with the real time PCR results. An example of the genotyping analysis is shown in Figure 1.

Statistical Analysis

Power analyses for this study were conducted using Power for Genetic Association Analysis (PGA) (10) with the estimated disease frequency of 0.16, the minor allele frequency, estimated relative risk reported for each variant, and $\alpha = 0.05$.

Tests for deviation from Hardy Weinberg equilibrium were conducted separately for the non Hispanic Caucasian and Hispanic Caucasian samples. Frequency differences between cases and controls were tested using logistic regression with adjustment for age. Due to low counts, tests for association with aggressive prostate cancer were conducted using Fisher's exact test for a 2 by 2 contingency table under both dominant and recessive models.

Results

We genotyped 7 deletion variants previously reported to be associated with risk for prostate cancer using real time quantitative PCR in 1536 (317 cases, 1219 controls) non Hispanic Caucasian men, with an average of 1468 men genotyped per assay after applying strict genotyping criteria. The mean age of diagnosis was 66.7 ± 8.3 and the mean age of controls was 66.5 ± 9.5. Using this dataset, we estimated power to detect association with 6 variants reported by Demichelis *et al*.(7) at >0.9, and ≈0.37 for the 2p24.3 variant reported by Liu *et* al (6) with an α of 0.05. Four of 6 CNVRs reported to be associated by Demichelis *et al*.(7) were not observed to be polymorphic. Variants on 12q21.31 and 19q13.12 were polymorphic (MAF ≈ 0.10, 0.21 respectively), but showed no evidence for association with risk for prostate cancer (12q21.31: OR, 0.89; 95% CI, 0.28 2.77; P=0.84) (19q13.12: OR, 1.08; 95% CI, 0.87 1.34; P=0.49).

The deletion on 2p24.3 reported by Liu *et al.* (6) had a minor allele frequency (MAF) of 0.059 in non Hispanic Caucasians, and was not associated with risk of prostate cancer given our predetermined of 0.05 (OR, 1.40; 95% CI, 0.99 2.00; P = 0.06). However, these results are consistent with the reported MAF, effect size, and confidence intervals reported by Liu *et al* using a dominant model. Among SABOR participants, carrier frequency was not associated with high grade prostate cancer when compared to controls under a dominant model [OR, 1.28; 95% CI, 0.43 3.12; P = 0.62 (Fischer's Exact)], but was significantly associated with high grade prostate cancer under a recessive model (OR, 27.99; 95% CI, 1.99 392.6; P = 0.007). Two (4.44%) of 45 men with high grade prostate cancer carried homozygous deletions, compared to 2 (0.16%) men carrying homozygous deletions in 1219 controls. The 2 control carriers for the deletion were aged 43.4 and 55.35 years, which is relatively young for developing prostate cancer. Additionally, carrying a homozygous deletion is also significantly associated with prostate cancer grade with 2 out of 45 high grade cases carrying a homozygous deletion compared to 0 out of 272 low grade cases (OR, INF; 95% CI, 1.15 INF, P=0.02). The 2 case carriers had observed Gleason scores of 9 and 10. When considering controls and non aggressive cases together the enrichment for homozygous deletion in aggressive cases is strong (OR, 34.3; 95% CI, 2.43 477; P=0.005).

We genotyped the 2p24.3 deletion in 632 (191 cases, 441 controls) Hispanic Caucasian men, for whom the average age of diagnosis was 63.0 ± 8.0 and the average age of controls was 60.0 ± 8.3. Among Hispanic Caucasians, the deletion had a minor allele frequency of 0.023, and was not associated with prostate cancer risk (OR, 0.91; 95% CI, 0.36 2.31; P = 0.84). Out of 37 Hispanic Caucasians with high grade prostate cancer, we observed zero cases carrying the deletion. This would be expected based on the low minor allele frequency of this deletion in the Hispanic Caucasian population.

Discussion

In this study we have independently tested reported associations of germ line copy number variants with risk for prostate cancer. Of 6 variants reported by Demichelis *et* al. (7), 4 were not polymorphic in the non Hispanic Caucasian population from the SABOR cohort. This observation may be explained by regional population differences in low minor allele frequency variants. Alternatively, it is possible that although the primers used in this study fall within the reported boundaries of the variants, that they do not fall within the variant's true boundaries. To avoid this scenario, primers and probes were designed to fall within variants identified by the 1000 genomes project if applicable (11). Additionally, 3 separate primer probe sets, including those reported by Demichelis *et al.*, were used to genotype the deletion at 15q21.3. Taken together our data provides strong evidence that the non Hispanic Caucasian population from the SABOR cohort do not harbor these deletion variants.

For the 2p24.3 variant our observations are highly consistent with those reported by Liu *et al.*(6). Although we obtained a p value that is not strictly significant, the probability that these results would be observed by chance (P=0.06) is suggestive of an association with overall prostate cancer risk. It is worth noting that the estimated power to detect this association at $\alpha = 0.05$ was not strong ($\approx$0.37). A comparison of results under dominant (as they reported it) and recessive models is illustrated in Figure 2. Interestingly, in Caucasians we observed statistically significant enrichment of homozygous deletions in high grade prostate cancer cases. Although all three cohorts display overlapping confidence intervals, this observation appears to be somewhat discordant with the observations reported by Liu *et al(6)*. However, the criteria that defined aggressiveness varied among the studies, which may explain different associations with aggressive disease. The definition of aggressiveness in this study did not include staging or PSA measurements because data was not available for all prevalent cases. However, the measurement used in this study, Gleason sum, is a robust predictor of prostate cancer prognosis (12). Follow up of this variant in larger cohorts with various definitions of aggressiveness is merited.

Conclusions

The initial discovery, and now independent supporting evidence, of the association between the 2p24.3 deletion reported by Liu *et al*. and prostate cancer support the involvement of copy number variation in the etiology of heritable disease. Further studies are merited to confirm the relationship between this deletion and risk for aggressive prostate cancer, and to investigate the relationship between this deletion, other linked variants, and nearby genes. The evidence from this study provides additional support that this copy number variant is a true risk allele for prostate cancer in non Hispanic Caucasians.

Acknowledgments

The participation and cooperation of all study subjects is gratefully acknowledged. The study could not have been accomplished without the skilled assistance of the SABOR/PREF clinical staff. We thank the expert technical assistance of Yasmin Ench.

Funding Support

This work was funded in part by DOD grant PC081025, by grant U01CA086402 from the Early Detection Research Network of the National Cancer Institute, and by the CTRC at UTHSCSA grant P30CA054174. Data was generated in the Genomics Core Shared Resource, which is supported by UTHSCSA, and NIH NCI P30CA054174 (CTRC of UTHSCSA).

References

1. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med. 2000; 343:78–85. [PubMed: 10891514]
2. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011; 12:363–76. [PubMed: 21358748]
3. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010; 464:704–12. [PubMed: 19812545]
4. Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, et al. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–8. [PubMed: 20811451]
5. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010; 464:713–20. [PubMed: 20360734]
6. Liu W, Sun J, Li G, Zhu Y, Zhang S, Kim ST, et al. Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer. Cancer Res. 2009; 69:2176–9. [PubMed: 19258504]

7. Demichelis F, Setlur SR, Banerjee S, Chakravarty D, Chen JY, Chen CX, et al. Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. Proc Natl Acad Sci U S A. 2012; 109:6686–91. [PubMed: 22496589]

8. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008; 9:356–69. [PubMed: 18398418]

9. Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, et al. A robust statistical method for case-control association testing with copy number variation. Nat Genet. 2008; 40:1245–52. [PubMed: 18776912]

10. Menashe I, Rosenberg PS, Chen BE. PGA: power calculator for case-control genetic association analyses. BMC Genet. 2008; 9:36. [PubMed: 18477402]

11. Consortium TGP. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–73. [PubMed: 20981092]

12. Egevad L. Recent trends in gleason grading of prostate cancer. II. Prognosis, reproducibility and reporting. Anal Quant Cytol Histol. 2008; 30:254–60. [PubMed: 18980156]

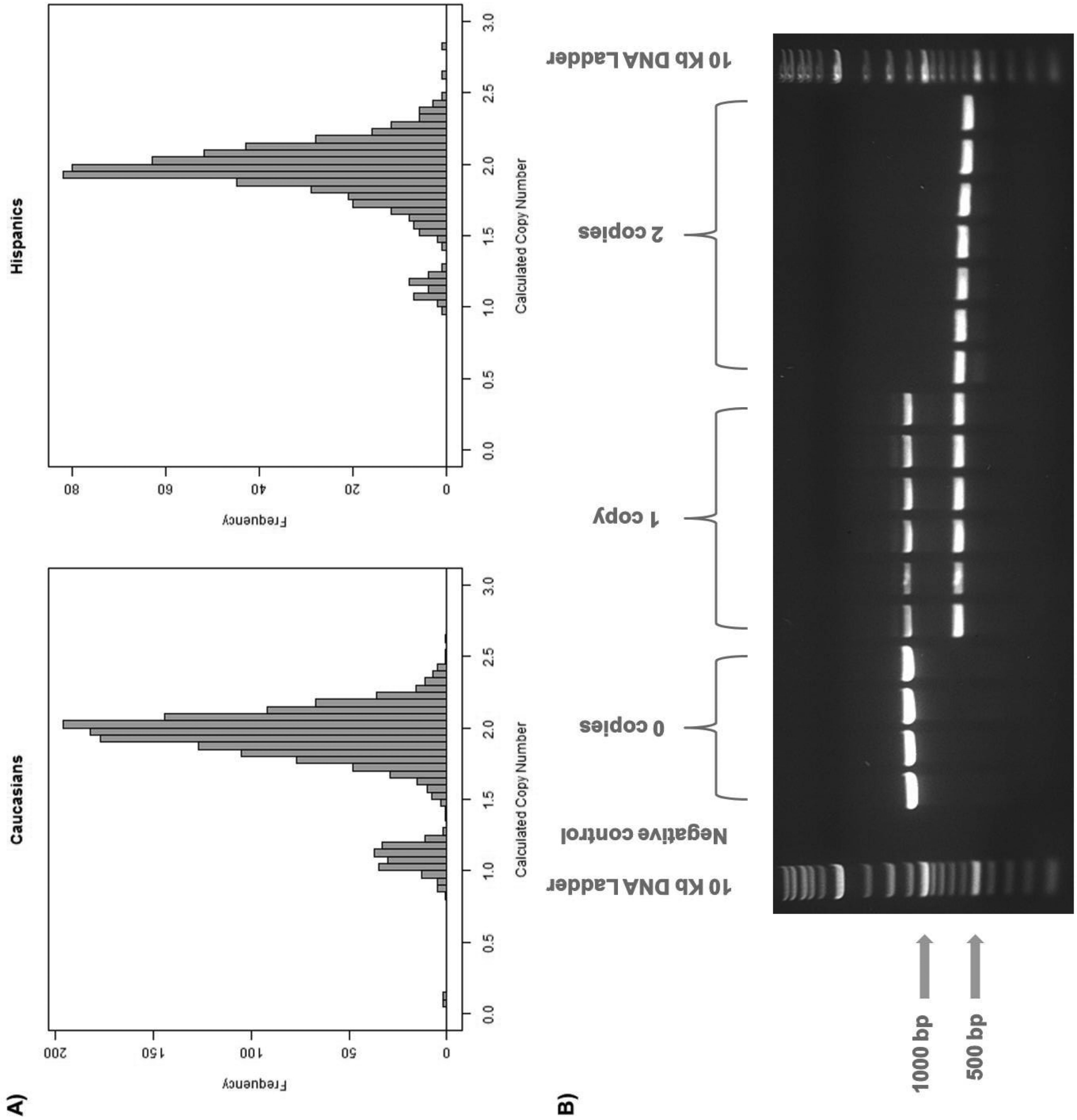


FIGURE 1. Genotyping of deletion in San Antonio Cohorts

A) Histograms of calculated copy number values in Caucasian and Hispanic individuals for a deletion on 2p24.3. Gaussian distributions are present representing measurement variation around integer copy number values of 0, 1, and 2. Discrete values were inferred based on these distributions. B) Gel electrophoresis of PCR products, as reported by Liu *et al*., was used to confirm the inferred genotypes in a subset of individuals, including the individuals with calculated copy number values at the edges of these distributions. This gel displays the genotyping results for a subset of the samples and displays all three genotypic states of this deletion. Negative control indicates no DNA.

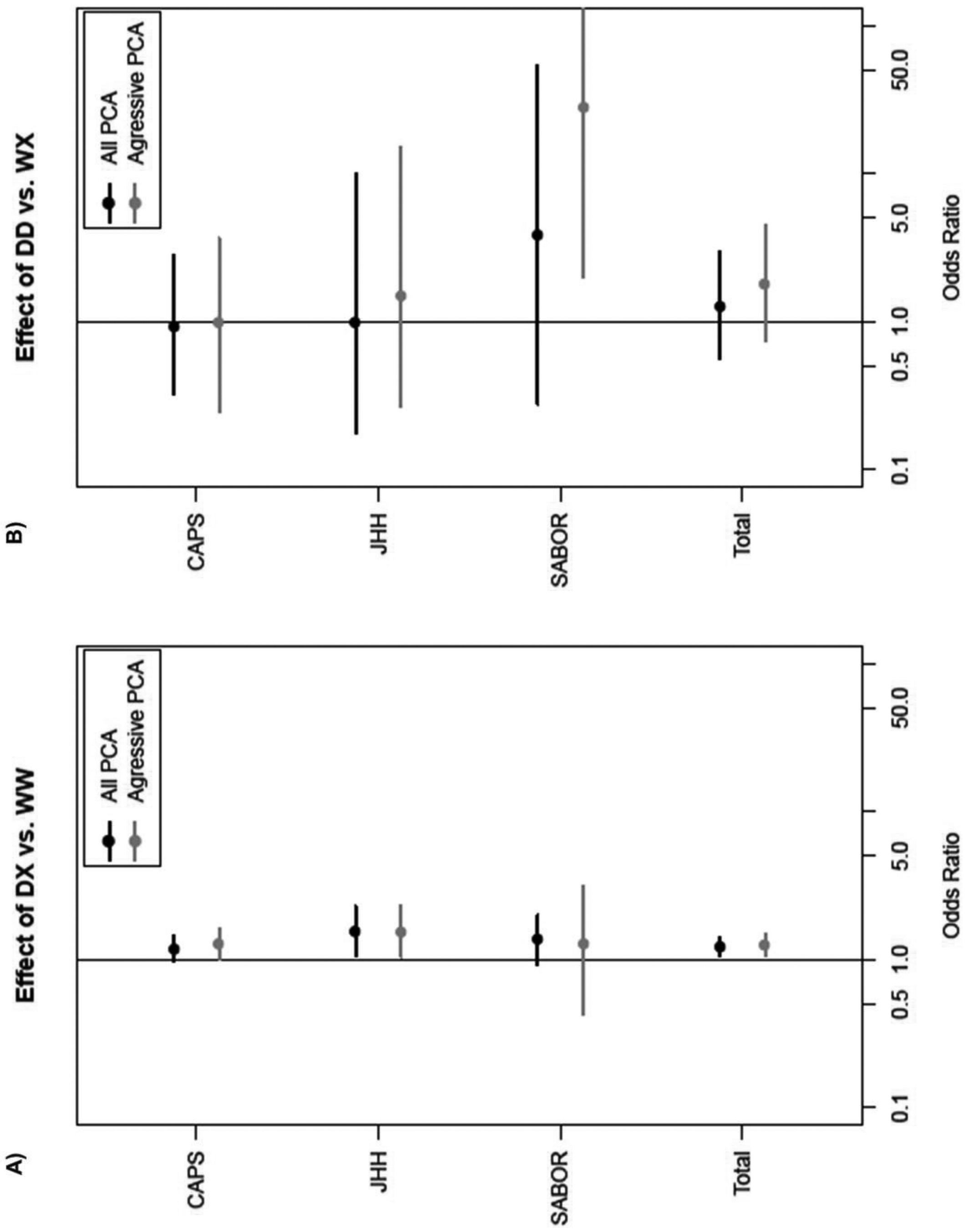


FIGURE 2. Comparison between studies of estimated effects for a deletion on 2p24.3

A comparison of odds ratios and 95% confidence intervals for Caucasians in this study (SABOR), and for cohorts previously reported on by Liu *et al.* Total indicates the results when counts are combined between studies.**A)** The results under a dominant model (DX vs WW) and **B)** a recessive model (DD vs WX).

Table 1

Summary of quantitative PCR primer/probe pairs

Deletion region	Primer design	Pre-designed Assay ID or Pubmed ID	Probe location (GRCh37/hg19)	Forward primer	Reverse primer	Probe
2p24.3	Custom		Chr2:14707671	TTGGGTCAGGCTGTCTTCT	TTTCAGCAGAGTGGTGAAGGAA	TCTTTCCCAACTCATTACC
15q21.3	Predesigned	Hs05369169_cn[1]	chr15:54199360			
15q21.3	Predesigned	Hs05345872_cn[1]	chr15:54198190			
19q13.12	Predesigned	Hs01617872_cn[1]	chr19:36829874			
8p23.2	Custom[1]		Chr8:3719510			
8q24.3	Predesigned	Hs03668742_cn[1]	chr8:145686582			
7p22.2	Predesigned	Hs01408227_cn[1]	chr7:4091360			
15q21.3	Previously reported	PMID:22496589				
12q21.31	Previously reported	PMID:22496589				

[1] Assays designed and ordered from Applied Biosystems and reported according to MIQE Guidelines.

Table 2

Genotyping counts in Caucasians

Deletion region	Pubmed ID	Cases				Controls			
		DD	DW	WW	MAF	DD	DW	WW	MAF
2p24.3	19258504	2	43	272	0.074	2	129	1088	0.055
15q21.3	22496589	0	0	310	0	0	0	1184	0
19q13.12	22496589	14	106	184	0.220	48	397	741	0.208
12q21.31	22496589	4	52	243	0.100	16	185	902	0.098
8p23.2	22496589	0	0	305	0	0	0	1173	0
8q24.3	22496589	0	0	297	0	0	0	1120	0
7p22.2	22496589	0	0	305	0	0	0	1154	0

Abbreviations: DD: homozygous deletion; DW: heterozygous deletion; WW: normal copy number; MAF: minor allele frequency

European Journal of Human Genetics (2013) 21, 404–409
© 2013 Macmillan Publishers Limited All rights reserved 1018-4813/13
www.nature.com/ejhg

ARTICLE

Utilizing extended pedigree information for discovery and confirmation of copy number variable regions among Mexican Americans

August Blackburn[1], Harald HH Göring[2], Angela Dean[3], Melanie A Carless[2], Thomas Dyer[2], Satish Kumar[2], Sharon Fowler[3], Joanne E Curran[2], Laura Almasy[2], Michael Mahaney[2], Anthony Comuzzie[2], Ravindranath Duggirala[2], John Blangero[2] and Donna M Lehman*,[3]

Copy number variation (CNV) remains poorly defined in many populations, including Mexican Americans. We report the discovery and genetic confirmation of copy number variable regions (CNVRs) in subjects of the San Antonio Family Heart and the San Antonio Family Diabetes Gallbladder Studies, both comprised of multigenerational pedigrees of Mexican American descent. In a discovery group of 1677 participants genotyped using Illumina Infinium Beadchips, we identified 2937 unique CNVRs, some with observation frequencies as low as 0.002, using a process that integrates pedigree information with CNV calls made by PennCNV and/or QuantiSNP. Quantitative copy number values had statistically significant ($P \leq 1.792e\text{-}5$) heritability estimates ranging from 0.139 to 0.863 for 2776 CNVRs. Additionally, 920 CNVRs showed evidence of linkage to their genomic location, providing strong genetic confirmation. Linked CNVRs were enriched in a set of independently identified CNVRs from a second group of 380 samples, confirming that these CNVRs can be used as predefined CNVRs of high confidence. Interestingly, we identified 765 putatively novel variants that do not overlap with the Database of Genomic Variants. This study is the first to use linkage and heritability in multigenerational pedigrees as a confirmation approach for the discovery of CNVRs, and the largest study to date investigating copy number variation on a genome-wide scale in individuals of Mexican American descent. These results provide insight to the structural variation present in Mexican Americans and show the strength of multigenerational pedigrees to elucidate structural variation in the human genome.

European Journal of Human Genetics (2013) **21,** 404–409; doi:10.1038/ejhg.2012.188; published online 22 August 2012

Keywords: copy number variation; Mexican Americans; MODY5; pedigree CNVRs; pedigree

INTRODUCTION

Copy number variants (CNVs), gains or losses of DNA sequence larger than 1 kb, were first recognized as widespread in the human genome in 2004.[1] Since this initial discovery multiple studies have further characterized copy number variation in the human genome.[2–4] Recent reports suggest that CNVs have a role in multiple complex disorders, such as schizophrenia,[5] autism,[6] autoimmune disorders,[7] and diabetes syndromes.[7,8]

However, despite this progress, copy number variation remains poorly defined in many populations. Understanding genetic variation in human populations besides Caucasians may reveal important biological insights not observable in the Caucasian population and is important for extending the benefits of understanding genetic risk to these underrepresented populations.

Despite methodological advancements for identifying, genotyping, and characterizing CNVs,[9] there is currently no comprehensive cost-effective method that has reached universal adoption. Several studies, including HapMap3,[4] have recently used the Illumina Infinium technology to characterize CNVs.

Given the current limitations of CNV analysis, most studies have taken the approach of limiting type I error by either requiring CNVs to be identified by more than one algorithm, or by limiting the size or number of probes identifying a CNV so that an acceptable portion are confirmed by orthogonal approaches. However, investigating copy number variation in multigenerational pedigrees allows for additional quality control metrics such as observation of transmission, and linkage with adjacent markers to confirm the genomic location of CNVs. Additionally, some rare CNVs will be present in the founders and may be inherited by younger generations in the pedigree, thus allowing for higher confidence in rare CNV identification. Despite these strengths, there are currently no large-scale studies reporting CNVs identified in multigenerational cohorts that use this information as a form of quality control.

In this study, we investigate copy number variation in 2057 participants of the San Antonio Family Heart Study (SAFHS) and San Antonio Family Diabetes/Gallbladder Study (SAFDGS), both comprised of multigenerational pedigrees of Mexican American descent. We present the identification, genotyping, and confirmation of copy number variable regions (CNVRs) using heritability and linkage. We report their genomic distribution, potential disease relevance, and discovery of novel variants. Most importantly, this

[1]Department of Cellular and Structural Biology, UT Health Science Center, San Antonio, TX, USA; [2]Department of Genetics, Texas Biomedical Research Institute, San Antonio, TX, USA; [3]Department of Medicine, Division of Clinical Epidemiology, UT Health Science Center, San Antonio, TX, USA
*Correspondence: Dr DM Lehman, Department of Medicine, Division of Clinical Epidemiology, UT Health Science Center San Antonio, 7702 Floyd Curl Drive, San Antonio, TX 78229, USA. Tel: +210 567 6714; Fax: +210 567 1990; E-mail: lehman@uthscsa.edu

Received 3 April 2012; revised 18 July 2012; accepted 19 July 2012; published online 22 August 2012

study provides novel insight into the structural variation specific to Mexican Americans.

MATERIALS AND METHODS

Study design

Participants in this study are members of extended, multigenerational families of Mexican American descent who have taken part in the SAFHS or the SAFDGS. Study-related clinical exams were conducted in San Antonio, Texas. SAFHS is a family study where the subjects were not ascertained on disease status. SAFDGS probands were ascertained on type 2 diabetes status. The current study was approved by the Institutional Review Board at the University of Texas Health Science Center San Antonio, and informed consent was obtained from all participants. Both cohorts have previously been described in detail.[10,11]

Study group 1. DNA isolated from primary blood mononuclear cells for 1677 participants was previously genotyped using four versions of the Illumina (San Diego, CA, USA) Infinium Beadchips: 767 participants were genotyped on the 1M duo beadchip, 327 individuals were genotyped on the 1M beadchip, and 583 individuals were genotyped on both the 510 and 550 beadchips. The SNP markers on the 510 and 550 beadchips are unique and together represent the content on the 1M beadchip.

Study group 2. Participants of this group are members of the SAFDGS for which DNA from lymphoblastoid cell lines were genotyped on the Illumina 660W beadchip. Data from a total of 380 participants were available for analyses.

CNVR identification

Approach. We first applied a 'wide' method of identification of CNVs to these data by applying CNV-calling algorithms to identify CNVs in individual samples. As these methods are known to have high rates of type I and type II errors, we also reassigned copy number values to individual participants based on comparison of the samples to each other, described in the section 'CNVR genotyping' referred to here as a 'deep' method.

To minimize type II errors that may be present using one algorithm, we chose to employ two algorithms, PennCNV[12] and QuantiSNP.[13] Standard quality control measures implemented within PennCNV were utilized for sample exclusion. The Log ratio SD, b-allele ratio drift, and genomic waviness were set at maximums of 0.3, 0.01, and 0.05, respectively. QuantiSNP calls were limited to CNV calls with a Maximum Bayesian Factor ≥ 10.

Recent CNV studies have taken the approach of reporting only those CNVs that are similarly identified by at least two algorithms in the same individual.[14] Given that our downstream analyses could be used to identify type I errors, we chose to take a more liberal approach to CNV inclusion so as to identify as much variation as possible. As pedigree data gives us additional information by which we could validate CNVs, we took the approach of including CNVs even if they are only identified by one algorithm.

Using the individual CNV calls, CNVRs, which are regions of overlap of CNVs, were identified. To reduce the rate of type I errors we limited our downstream analysis to regions harboring a CNV call in at least two individuals of the same pedigree on the same beadchip. To determine breakpoints for this set of CNVRs, we expanded the CNV breakpoints to the largest overlapping region identified in each pedigree. This set of CNVRs we have termed Pedigree CNVRs. To further summarize the observed CNV calls, we consolidated CNVRs across all pedigrees by using the most common breakpoints observed for overlapping Pedigree CNVRs. For those consolidated CNVRs that were observed to be overlapping on multiple platforms, we used the breakpoints identified by the higher density beadchips. Priority was given to breakpoints identified using PennCNV, when the CNV was identified by both algorithms. Each CNVR was then manually inspected to ensure the individual CNV calls indicated the identified CNVR breakpoints, which resulted in removing five CNVRs and redefining the breakpoints for CNVR887. All other CNVRs remained unchanged.

CNVR genotyping

Considering each beadchip independently, for each final CNVR we used CNVtools[15] to identify the first principal component of log R ratios of markers falling within the CNVR breakpoints. Using this value, we used CNVtools to cluster individuals into groups harboring the same copy number genotype. To improve our power for downstream analyses, we combined the Log R ratio data from the 1M duo, 1 M, 550, and 510 beadchips into a single matrix and performed the same procedures using CNVtools.

CNVR characterization

PennCNV calls and histograms of the first principal component were used to categorize each CNVR as a deletion, duplication, complex, overlapping, or unknown. CNVRs were considered overlapping if there were two clear variants of different lengths, either duplication or deletion, which were both overlapping and present in multiple individuals. Complex regions were labeled based on their location in either centromeric, telomeric, or immunoglobulin regions. Unknown regions were those that we were unable to clearly classify based on the available data, but may fit into any of the other four categories. Tables from the UCSC genome browser summarizing OMIM genes, RefSeq genes, segmental duplications, microRNAs, and disease association SNPs from the National Human Genome Research Institute were downloaded on 1 March 2011 and used as reference datasets. A table from the UCSC genome browser summarizing the Database of Genomic Variants (DGVs) was downloaded on 4 February 2012 and used as a reference dataset. Copy Number Polymorphisms were downloaded from HapMap3 on 23 March 2011 for comparison of lengths.

Heritability and linkage analysis

For each CNVR, the first principal component value identified by CNVtools was rank normalized. Subsequently, heritability was calculated and each CNVR was tested for linkage to its own genomic location. Both tests were conducted using a variance components approach using the statistical models implemented within the software package SOLAR.[16,17] Heritability and linkage was also calculated for binned copy number values by treating the values as a categorical trait, either harboring or not harboring a deletion or duplication.

Statistical analyses of results

CNVR lengths and minor allele frequencies were log transformed to provide a normal distribution. Correlation of lengths between this study and HapMap3 was tested using linear regression. Correlations between the observed minor allele frequency and heritability or linkage LOD scores were tested using linear regression. The difference in lengths between deletions and duplications was tested using a one-sided *t*-test. The difference in ratios of deletions and duplications which overlap genes was tested using a χ^2 test. The correlation between length and observed minor allele frequency of deletions was tested using a linear model. The difference in size between novel and known CNVRs was tested using a two sided *t*-test. CNVRs that were linked to their genomic location in group 1were tested for enrichment in group 2 using a χ^2 test.

RESULTS

Study group 1

We identified 2937 unique CNVRs, representing 120 959 and 75 932 autosomal CNV calls by either PennCNV or QuantiSNP, respectively, detected across all individuals genotyped on any SNP microarray type. In all, 1201 CNVRs are pedigree specific, 399 of which have not been previously reported in the DGVs and potentially represent private variants enriched within the pedigrees through transmission. Summary information for CNV calls is presented in Table 1.

When applying the Gaussian mixture model implemented within CNVtools,[15] we were able to confidently fit 186 CNVRs into defined classes. When coded as a dichotomous trait, 169 (90.9%) of these CNVRs had a statistically significant ($P \leq 2.7e\text{-}4$) heritability of 1.00, as would be expected for a correctly genotyped copy number variant.

Table 1 Summary of CNV calls made by either PennCNV or QuantiSNP

Beadchip	*1mduo*		*1M*		*550*		*510*		*660W*	
Samples passing QC	*688 (89.7%)*		*289 (88.4%)*		*568 (97.4%)*		*564 (96.7%)*		*324 (85.3%)*	
Algorithm	*PennCNV*	*QuantiSNP*	*PennCNV*	*QuantiSNP*	*PennCNV*	*QuantiSNP*	*PennCNV*	*QuantiSNP*	*PennCNV*	*QuantiSNP*
Total autosomal calls	46 440	34 841	25 924	16 574	16 054	9713	32 541	14 804	106 219	116 505
Median size (all)	24 314	31 540	22 271	35 998	29 730	35 407	22 134	32 215	2821	2848
Pedigree CNVRs	6552	5294	3583	2383	2210	1591	4762	2342	20 243	20 500
Deletions										
Median Size	20 398	25 659	14 935	18 064	18 344	15 800	12 779	14 823	2683	2324
CNV calls/individual (median/mean)	29/41.12	23/34.41	42/57.01	28/32.77	15/18.1	11/11.1	33/36.48	16/16.38	232.5/240.5	278/287.1
Markers/CNV call (median/mean)	10/18.82	11/20.6	8/15.19	8/17.56	6/12.41	6/13.12	6/10.01	6/10.26	17/16.34	16/15.3
Duplications										
Median size	31 214	47 731	40 285	74 701	69 100	105 804	41 523	181 879	3493	8934
CNV calls/individual (median/mean)	20/26.38	13/19.11	23/32.99	15/24.98	5/10.16	4/6.33	17/21.21	7/10.29	77.5/87.34	67.5/72.52
Markers/CNV call (median/mean)	12/25.8	19/43.6	12/36.31	22/67.38	13/39.92	16/51.43	9/25.54	15/32.47	16/19.26	18/25.76
Deletion:duplication ratio	1.56	1.80	1.73	1.31	1.78	1.75	1.72	1.59	2.75	3.96

Abbreviations: CNV, copy number variable; CNVR, copy number variable region; QC, quality control.

Table 2 Summary information for CNVRs identified in group 1

		Heritable	*Linked*		
	All CNVRs	*($P\leq$1.702E-5)*	*($P\leq$0.05)*	*($P\leq$1.702E-5)*	*Novel*
Total	2937	2776 (94.5%)	920 (31.3%)	431 (14.7%)	727 (24.8%)
Deletion	1453 (49.5%)	1373	505	235	403
Duplication	464 (15.8%)	437	178	68	110
Complex	131 (4.5%)	128	36	21	1
Overlapping Variants	48 (1.6%)	45	16	7	4
Unknown	841 (28.6%)	793	185	100	209
Within 50 kb of disease SNP from NHGRI	409 (13.9%)	390	109	44	95
Novel	727 (24.8%)	670	146	32	–
1 Pedigree	1201 (40.9%)	1115	273	61	399
>1 Pedigree	1736 (59.1%)	1661	647	370	328

Abbreviation: CNVR, copy number variable region.

Ten additional CNVRs had statistically significant ($P\leq$2.7e-4) heritabilities ranging from 0.723 to 0.977. In all, 151 (81.2%) CNVRs showed evidence of linkage to their genomic location ($P\leq$0.05), 131 (70.4%) of which are linked after Bonferroni correction for the number of CNVRs investigated ($P\leq$2.7e-4).

Given the high ratio of observed posterior probability errors when fitting these measurements into defined bins, we chose to work with the first principal component as a measurement of copy number as it was available for all 2937 CNVRs, an approach that has been used previously.[18] For 2776 (94.5%) CNVRs the first principal component had statistically significant ($P\leq$1.702e-5) heritabilities ranging from 0.139 to 0.863. For 920 (31.3%) CNVRs the first principal component showed evidence of linkage to its genomic location ($P\leq$0.05). Of 727 novel CNVRs, 670 (92.2%) are significantly heritable, and 146 (20.1%) show evidence of linkage ($P\leq$0.05), providing very strong evidence of the validity of these novel variants. Linkage and heritability information for different classes of variants is presented in Table 2.

We hypothesized that rare variants may not be sufficiently measured by the first principal component value for significant linkage to be observed. We investigated the relationship between the observation frequency of 186 CNVRs that were binned into defined classes and their respective heritability and linkage LOD values using the first principal component. These CNVRs had observation frequencies ranging from 0.002 to 0.477. Observation frequency was positively correlated with heritability ($P\leq$2e-16), as shown in Figure 1. Observation frequency was also associated with linkage LOD values ($P\leq$2e-16), supporting our hypothesis that common variants were more likely to be linked in our analysis. On the basis of this observation and the observation that dispersed duplications may insert elsewhere in the genome, we conclude that lack of linkage to their genomic location does not indicate that a CNVR is the result of a type 1 error. Rather we consider those that have statistically significant heritability estimates to be confirmed, and those that are linked to be of the highest confidence, having evidence of their genomic location. The location, class, heritability estimates, and

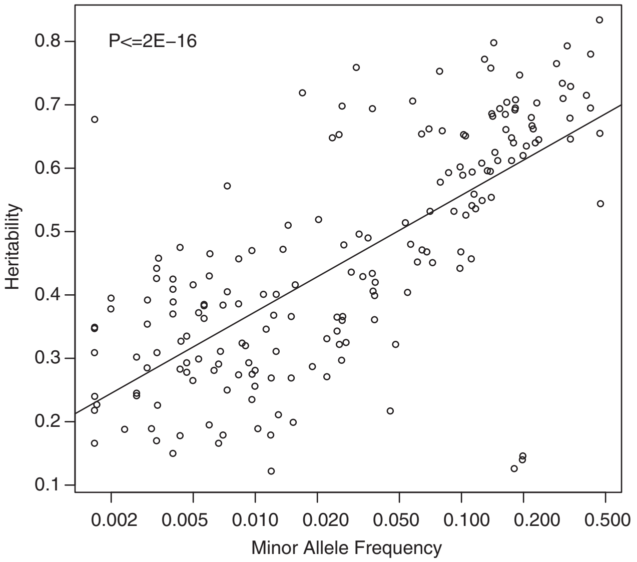


Figure 1 Relationship of minor allele frequency and heritability of the first principal component. Minor allele frequency is significantly associated with the heritability of the first principal component measurement of each CNVR.

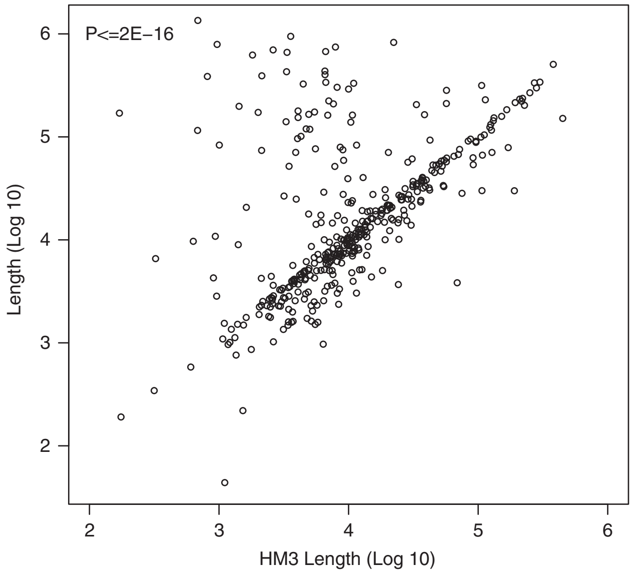


Figure 2 Comparison of lengths between this study and HapMap3. The length of CNVRs that overlap a single CNP from HapMap3 are plotted on the y axis. The length of corresponding CNPs from HapMap3 are plotted on the x axis.

linkage LOD values for all CNVRs are presented in the Supplementary Materials.

The lengths of the CNVRs identified in this study correlated significantly ($P\leq2.2e\text{-}16$) with lengths of overlapping CNVRs from HapMap3 (392 CNVRs).[4] Visual inspection, presented in Figure 2, indicates a generally good agreement of lengths between studies. When considering those CNVRs that show evidence of linkage ($P<0.05$), the average length of deletions and duplications are 21.8 and 45.5 kb, respectively ($P=4.62E\text{-}5$), as shown in Figure 3. Twenty-nine of 178 (16.3%) duplications encompass at least 1 gene compared with 25 of 505 (5.0%) deletions ($P=1.42E\text{-}6$), but this difference is not statistically significant when corrected for the length of the CNVRs. Interestingly, within 186 CNVRs, which were binned into defined states, large deletions had lower observation frequencies than smaller deletions ($P=1.2E\text{-}8$).

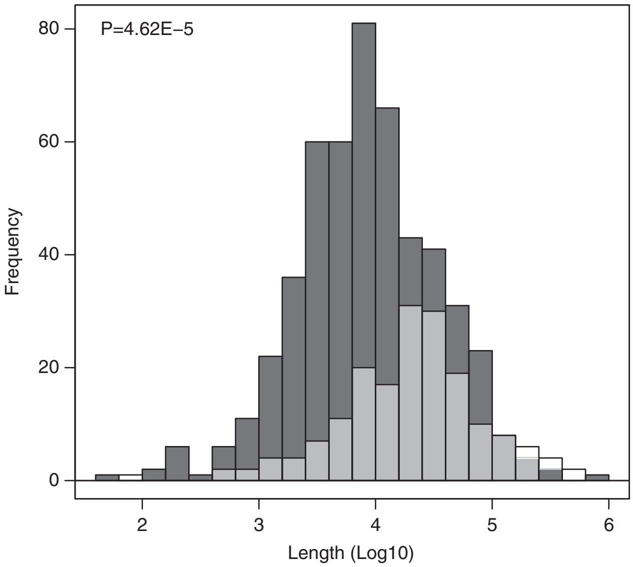


Figure 3 Size distribution of deletions and duplications linked to their genomic location. The size distribution of duplications is presented in transparent white and is superimposed over the size distribution of deletions, which is presented in dark gray. There is a statistically significant difference in size distribution likely representing an improved ability to detect small deletions.

Four-hundred and nine CNVRs identified in this study are within 50 kb of a disease associated SNP from the NHGRI GWAS catalog. Among these, was a CNVR harboring two overlapping deletions, a common ~1-kb deletion within the *ACACA* gene and a ~1.44-Mb deletion of 16 genes, including *HNF1B*, which is responsible for Renal Cyst and Diabetes Syndrome, also referred to as Maturity Onset Diabetes of the Young 5 (MODY5[MIM 137920]). This is a recurrent deletion that has been associated with multiple phenotypes including MODY5 and psychiatric disorders.[19–21] The deletion was apparent in three individuals, a woman and her two daughters. We hypothesized that these individuals had MODY5. Retrospective investigation of clinical data showed that the mother and one daughter were diagnosed with diabetes at ages 17 and 22.4 years, respectively. One daughter remained diabetes free at her last visit at age 31, indicating incomplete penetrance.

We discovered 727 putatively novel CNVRs that do not overlap with variants reported in the DGVs.[1] Of these, 328 were detected in multiple pedigrees. These CNVRs may have been missed in previous studies or may be unique to the Mexican American population in this study. Considering CNVRs, which showed evidence of linkage to their genomic location ($P\leq0.05$), novel CNVRs identified in this study were smaller on average than previously known CNVRs ($P=0.0004$), as shown in Figure 4.

Of 146 novel and linked CNVRs, 21 are within 50 kb of a disease associated SNP from the NHGRI. Of these 21, 10 overlap with gene exons, providing novel testable hypotheses, which are summarized in the Supplementary Materials. One of these 10 is a ~5.2-kb deletion of the *IL2* gene. This CNVR is significantly heritable ($P=6.35E\text{-}12$), linked ($P=0.045$), and is flanked by SNPs associated with immune-related functions.[22,23] A second of these 10 is a 370-bp CNVR within

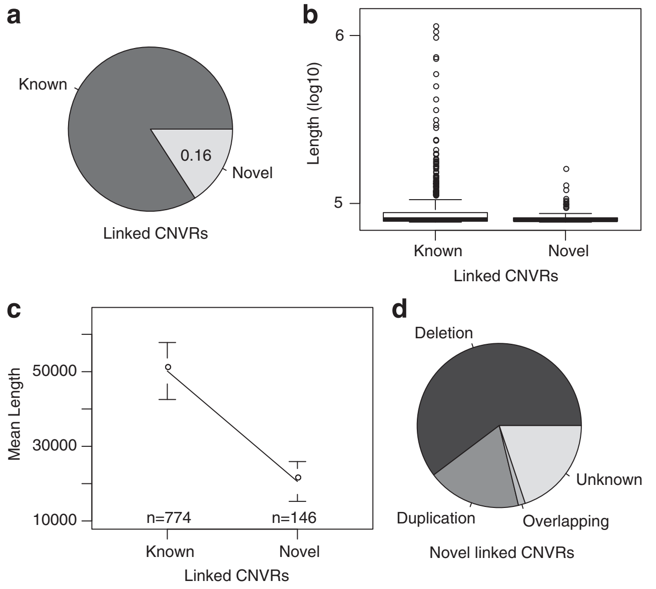


Figure 4 Novel linked CNVR information. (**a**) The ratio of novel CNVRs within CNVRs linked to their own location. (**b**) Boxplot showing that there were few very large (>100KB) novel, linked CNVRs. (**c**) The mean and 95% CI of lengths of novel and known CNVRs. (**d**) The ratio of various classes of novel linked CNVR.

the first exon of the *UGT1A7* gene. This CNVR is highly linked to its own genomic location ($P = 5.44\text{E-}60$), and ~6.6 kb from rs2602381, which was previously associated with attention-deficit hyperactivity disorder.[24]

Study group 2

We identified 2555 CNVRs representing 106 219 and 116 505 autosomal CNV calls using PennCNV and QuantiSNP, respectively. Of 72 putatively novel CNVRs discovered in this group, 34 were also discovered in group 1. Despite being identified using separate platforms, CNVRs in group 2 overlap 745 CNVRs identified in Study group 1, of which 420 (56.4%) showed evidence of linkage ($P \leq 0.05$) in group 1. This enrichment is statistically significant ($P \leq 1.0\text{e-}10$), showing that those which were linked in the first study were more likely to be observed in a second study group, as well as validating linkage as a useful confirmation approach.

DISCUSSION

Copy number variation makes up a significant portion of genetic variation in humans. The current limitations regarding CNVs are largely due to a lack of an affordable comprehensive identification and genotyping strategy, although methods have been proposed to address this issue.[9,25] We have applied a strategy that takes advantage of the benefit of pedigree information to identify, confirm, and localize CNVs in the largely understudied Mexican American population. In an effort to limit type 1 errors, previous reports have limited their analysis to CNVs identified using at least two algorithms in the same individual. However, in this study we were able to rely on pedigree information as a form of quality control, so we did not restrict CNVRs based on this convention. Of 431 CNVRs which we have confirmed beyond doubt using linkage analysis, 144 CNVRs were identified by PennCNV only and 21 CNVRs were identified by QuantiSNP only in at least two individuals in the same pedigree, indicating that restricting CNVRs based on algorithm overlap can be overly conservative.

Similar to previous reports,[7,18] we observed poor cluster separation for many regions and were unable to confidently bin individuals into defined classes. However, we are able to show that for ~95% of the CNVRs identified here, representative values for these regions had statistically significant heritability estimates. Additionally, 920 CNVRs showed evidence of linkage to their genomic location, providing exceptionally strong genetic confirmation. This nicely highlights the continued difficulty of genotyping CNVs, and supports the use of representative values in the absence of high-confidence binning.

Through linkage we obtained evidence that 178 duplications have inserted near their genomic location. This does not mean that these are tandem duplications, because a dispersed duplication could potentially be close enough to its original genomic location to be linked. The extent to which dispersed duplications are responsible for associations on other chromosomes is currently unknown. Future studies aimed at using linkage analysis in multigenerational pedigrees may help to identify the insertion locations of common duplications.

The observed difference in size between duplications and deletions in this study could be the result of a methodological bias toward detecting large CNVs. Similarly, the correlation between deletion size and frequency may indicate an increased ability to detect large deletions. Alternatively, these observations could indicate that large deletions are under stronger selective pressure than duplications or small deletions.

In summary, we have identified and genotyped CNVRs that are polymorphic in Mexican Americans from San Antonio, Texas. The majority of CNVRs identified have been previously reported, indicating Mexican Americans share much of their genetic diversity with other populations. However, about 25% of copy number variation in this population may be specific to this ethnic group and has not been previously characterized due to the limited number of studies in Mexican American populations.

Importantly, we show that CNVRs that were confirmed using linkage analysis in this study are likely to be identified again in a separate study, and therefore can be used as predefined CNVRs of high confidence in future studies investigating CNVs in Mexican Americans. Our application of heritability and linkage analysis to confirm CNVR genotype measurements shows the promise of using multigenerational pedigrees to improve the power and accuracy with which we can characterize structural variation in the human genome, and should be considered orthogonal to other quality metrics for CNV calling. We suggest that future studies investigating copy number variation in multigenerational pedigrees should incorporate similar approaches to for CNVR confirmation.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This study was supported in part by grants from the National Institutes of Health (DK47482, DK70746, DK053889, HLO45222, RR013556, MH059490) and the Department of Defense (DOD PC081025). We thank the participants of the SAFHS and SAFDGS for their generous cooperation.

1 Iafrate AJ, Feuk L, Rivera MN *et al*: Detection of large-scale variation in the human genome. *Nat Genet* 2004; **36**: 949–951.
2 Redon R, Ishikawa S, Fitch KR *et al*: Global variation in copy number in the human genome. *Nature* 2006; **444**: 444–454.
3 Conrad DF, Pinto D, Redon R *et al*: Origins and functional impact of copy number variation in the human genome. *Nature* 2010; **464**: 704–712.

4 Altshuler DM, Gibbs RA, Peltonen L *et al*: Integrating common and rare genetic variation in diverse human populations. *Nature* 2010; **467**: 52–58.

5 Xu B, Roos JL, Levy S, van Rensburg EJ, Gogos JA, Karayiorgou M: Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet* 2008; **40**: 880–885.

6 Sebat J, Lakshmi B, Malhotra D *et al*: Strong association of de novo copy number mutations with autism. *Science* 2007; **316**: 445–449.

7 Craddock N, Hurles ME, Cardin N *et al*: Genome-wide association study of CNVs in 16 000 cases of eight common diseases and 3000 shared controls. *Nature* 2010; **464**: 713–720.

8 Jarick I, Vogel CI, Scherag S *et al*: Novel common copy number variation for early onset extreme obesity on chromosome 11q11 identified by a genome-wide analysis. *Hum Mol Genet* 2011; **20**: 840–852.

9 Alkan C, Coe BP, Eichler EE: Genome structural variation discovery and genotyping. *Nat Rev Genet* 2011; **12**: 363–376.

10 Hunt KJ, Lehman DM, Arya R *et al*: Genome-wide linkage analyses of type 2 diabetes in Mexican Americans: the San Antonio Family Diabetes/Gallbladder Study. *Diabetes* 2005; **54**: 2655–2662.

11 Mitchell BD, Kammerer CM, Blangero J *et al*: Genetic and environmental contributions to cardiovascular risk factors in Mexican Americans. The San Antonio Family Heart Study. *Circulation* 1996; **94**: 2159–2170.

12 Wang K, Li M, Hadley D *et al*: PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 2007; **17**: 1665–1674.

13 Colella S, Yau C, Taylor JM *et al*: QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res* 2007; **35**: 2013–2025.

14 Wineinger NE, Pajewski NM, Kennedy RE *et al*: Characterization of autosomal copy-number variation in African Americans: the HyperGEN Study. *Eur J Hum Genet* 2011.

15 Barnes C, Plagnol V, Fitzgerald T *et al*: A robust statistical method for case-control association testing with copy number variation. *Nat Genet* 2008; **40**: 1245–1252.

16 Boerwinkle E, Chakraborty R, Sing CF: The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. *Ann Hum Genet* 1986; **50**: 181–194.

17 Almasy L, Blangero J: Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 1998; **62**: 1198–1211.

18 Surakka I, Kristiansson K, Anttila V *et al*: Founder population-specific HapMap panel increases power in GWA studies through improved imputation accuracy and CNV tagging. *Genome Res* 2010; **20**: 1344–1351.

19 Nagamani SC, Erez A, Shen J *et al*: Clinical spectrum associated with recurrent genomic rearrangements in chromosome 17q12. *Eur J Hum Genet* 2010; **18**: 278–284.

20 Mefford HC, Clauin S, Sharp AJ *et al*: Recurrent reciprocal genomic rearrangements of 17q12 are associated with renal disease, diabetes, and epilepsy. *Am J Hum Genet* 2007; **81**: 1057–1069.

21 Moreno-De-Luca D, Mulle JG, Kaminsky EB *et al*: Deletion 17q12 is a recurrent copy number variant that confers high risk of autism and schizophrenia. *Am J Hum Genet* 2010; **87**: 618–630.

22 Plagnol V, Howson JM, Smyth DJ *et al*: Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLoS Genet* 2011; **7**: e1002216.

23 Ramasamy A, Curjuric I, Coin LJ *et al*: A genome-wide meta-analysis of genetic variants associated with allergic rhinitis and grass sensitization and their interaction with birth order. *J Allergy Clin Immunol* 2011; **128**: 996–1005.

24 Mick E, Todorov A, Smalley S *et al*: Family-based genome-wide association scan of attention-deficit/hyperactivity disorder. *J Am Acad Child Adolesc Psychiatry* 2010; **49**: e893.

25 Park H, Kim JI, Ju YS *et al*: Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat Genet* 2010; **42**: 400–405.

Supplementary Information accompanies the paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)