



Genome-wide analysis of the role of copy-number variation in pancreatic cancer risk

Jason A. Willis^{1,2}, Semanti Mukherjee², Irene Orlow³, Agnes Viale⁴, Kenneth Offit^{1,2}, Robert C. Kurtz¹, Sara H. Olson³ and Robert J. Klein^{1,2*}

¹ Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

² Program in Cancer Biology and Genetics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

³ Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

⁴ Genomics Core Laboratory, Memorial Sloan-Kettering Cancer Center, New York, NY, USA

Edited by:

Hakon Hakonarson, University of Pennsylvania, USA

Reviewed by:

Peter A. Kanetsky, University of Pennsylvania, USA

Lynette Phillips, Kent State University, USA

*Correspondence:

Robert J. Klein, Department of Medicine, Memorial Sloan Kettering Cancer Center, 1275 York Ave, Box 337, New York, NY 10065, USA
e-mail: kleinr@mskcc.org

Although family history is a risk factor for pancreatic adenocarcinoma, much of the genetic etiology of this disease remains unknown. While genome-wide association studies have identified some common single nucleotide polymorphisms (SNPs) associated with pancreatic cancer risk, these SNPs do not explain all the heritability of this disease. We hypothesized that copy number variation (CNVs) in the genome may play a role in genetic predisposition to pancreatic adenocarcinoma. Here, we report a genome-wide analysis of CNVs in a small hospital-based, European ancestry cohort of pancreatic cancer cases and controls. Germline CNV discovery was performed using the Illumina Human CNV370 platform in 223 pancreatic cancer cases (both sporadic and familial) and 169 controls. Following stringent quality control, we asked if global CNV burden was a risk factor for pancreatic cancer. Finally, we performed *in silico* CNV genotyping and association testing to discover novel CNV risk loci. When we examined the global CNV burden, we found no strong evidence that CNV burden plays a role in pancreatic cancer risk either overall or specifically in individuals with a family history of the disease. Similarly, we saw no significant evidence that any particular CNV is associated with pancreatic cancer risk. Taken together, these data suggest that CNVs do not contribute substantially to the genetic etiology of pancreatic cancer, though the results are tempered by small sample size and large experimental variability inherent in array-based CNV studies.

Keywords: pancreatic cancer, copy number variation, cancer risk, SNP microarrays, CNVs

INTRODUCTION

Pancreatic adenocarcinoma is the fourth-leading cause of cancer mortality in the United States for both men and women (Siegel et al., 2012). Despite recent advances in screening methods and surgical treatment, it is a rapidly fatal disease with a poor 5-year survival rate of 5–6%. Thus, a challenge exists to develop new and more effective therapeutic interventions.

Inherited genetic predisposition to pancreatic cancer is hypothesized to play a role in both familial and non-familial forms of the disease. In large-scale genome-wide association studies, common single-nucleotide polymorphisms (SNPs) on chromosomes 9q34, 13q22, 1q32, and 5p15 were associated with pancreatic cancer risk (Amundadottir et al., 2009; Petersen et al., 2010); however, the true causal variants underlying these associations and their functional mechanisms remain unclear. Additional studies have focused their analyses on SNPs within candidate genes (Jiao et al., 2006, 2008; Li et al., 2007; McWilliams et al., 2008). Under this approach, SNPs within DNA damage response and repair genes—particularly *ATM*, *LIG3*, *XRCC1*, and *XRCC2* genes—were associated with increased risk, suggesting the involvement of inherited genetic variants within these pathways in pancreatic tumorigenesis.

Importantly, such efforts to identify inherited genetic variants that contribute to pancreatic cancer susceptibility may lead to novel biological insights about the disease and useful biomarkers for risk prediction. However, while these efforts have primarily focused on the analysis of SNPs, the additional contribution of germline copy number variations (CNVs) remains unclear.

CNVs are generally defined as inherited or *de novo* deletions or duplications of the genome ranging in size from 100 bp to 3 Mb (Zhang et al., 2009). Such variations may lead to changes in gene dosage and/or expression (Diskin et al., 2009). As of August 2012, approximately 179,450 human CNVs have been reported in the Database of Genomic Variants (DGV) (Iafate et al., 2004; Zhang et al., 2006). Although there are substantially fewer reported CNVs than SNPs, it is estimated that more than 30% of the human genome is covered by at least one CNV (compared to the <1% covered by SNPs). Thus, CNVs are hypothesized to be of functional significance.

The specific role of CNVs in familial forms of pancreatic cancer has been investigated previously. Lucito et al. used representational oligonucleotide microarray analysis (ROMA) to identify a total of 56 germline CNVs that were present in patients with a family history of pancreatic cancer and absent from a

cohort of healthy controls (Lucito et al., 2007). Al-Sukhni et al. followed a similar approach by analyzing the germline DNA of 91 familial pancreatic cancer patients and a 950 healthy controls using high-resolution Affymetrix 500 K and SNP 6.0 platforms (Al-Sukhni et al., 2012). There, investigators found a total of 93 germline CNVs that were unique to familial pancreatic cancer patients. Taken together, these studies nominate several CNVs as putative risk loci for familial pancreatic cancer. However, additional studies are needed to confirm these findings. More importantly, evidence for the broader role of CNVs outside of the familial pancreatic cancer setting is still unclear.

Here, we report a genome-wide analysis of CNVs in a hospital-based, European ancestry cohort of pancreatic cancer cases and controls. Germline CNV discovery was performed using the Illumina Human CNV370 platform in 223 pancreatic cancer cases (both sporadic and familial) and 169 controls. Following stringent quality control, we explored whether global CNV burden was a risk factor for pancreatic cancer. Finally, we performed *in silico* CNV genotyping and association testing to discover novel CNV risk loci.

MATERIALS AND METHODS

SAMPLE COLLECTION AND SNP ARRAY GENOTYPING

Participants were part of an ongoing hospital-based case-control project conducted in conjunction with the Familial Pancreatic Tumor Registry (FPTR) at Memorial Sloan-Kettering Cancer Center (MSKCC; New York, NY) as described previously (Mukherjee et al., 2011; Willis et al., 2012). Briefly, patients were eligible if they were 21 years or older, spoke English, and had pathologically or cytologically confirmed adenocarcinoma of the pancreas. Patients were recruited between June 2003 and July 2009 from the surgical and medical oncology clinics at MSKCC at the time of their initial diagnosis or during follow-up. Controls were spouses of patients or visitors accompanying patients with other diseases, had the same age and language eligibility requirements as the cases, had no personal history of cancer (except for non-melanoma skin cancer), and were not blood relatives of the cases. The participation rate among approached and eligible individuals was 76% among cases and 56% among controls. The study was approved by the MSKCC Institutional Review Board, and all enrolled participants signed informed consent.

Consented participants provided a blood or buccal (mouth-wash or saliva) sample to the MSKCC FPTR research study assistant and completed risk factor and family history questionnaires administered by the research study assistant in person or via telephone. Biospecimens were subsequently delivered for genomic DNA extraction and banking to the Molecular Epidemiology Laboratory. DNA was isolated from mouthwash specimens using the Puregene DNA Purification Kit (Qiagen, Inc.), from saliva samples with the Oragene saliva kits (DNA Genotek), and from whole blood using the GentraPuregene Blood Kit (Qiagen Inc.). DNA samples were hydrated in $1 \times$ TE buffer. The quality and quantity of the DNA was assessed by spectrophotometry and agarose gel electrophoresis.

A total of 464 individuals (263 cases and 201 controls) recruited at MSKCC were available for inclusion in downstream analyses. DNA samples were genotyped in 28 batches on the

Illumina Human CNV370 bead array (either the Illumina Human CNV370-Duo or Illumina Human CNV370-Quad format) at the Genomics Core Laboratory of MSKCC according to the manufacturer's protocol. Normalization and SNP genotype calling was performed using the Illumina BeadStudio software package. Ten individuals had their DNA analyzed twice for quality control during the course of the genotyping experiments, yielding a total of 474 samples. Normalized probe intensities were exported for downstream CNV discovery and genotyping.

SNP genotype calls for 474 individual samples (including 10 duplicate pairs) were exported to PLINK (version 1.07; Purcell et al., 2007) for processing. Identity-by-descent (IBD) analysis was performed to confirm that none of the genotyped individuals were blood relatives. For each known duplicate sample pair, priority was given to the sample that passed CNV-level quality-control (described below). We removed SNPs with call rates $<95\%$, minor allele frequency $<5\%$, or Hardy-Weinberg equilibrium (HWE) test p -value $< 1 \times 10^{-7}$ in controls, leaving a total of 315,136 SNPs.

Population structure was examined by principal component analysis (PCA) of the SNP genotype calls. As reference, we obtained whole-genome SNP data for Utah residents with northern and western European ancestry (abbreviated CEU) and individuals living in Toscani in Italia (abbreviated TSI) from the International HapMap project (phase 3, draft release 2) (International Hapmap 3 Consortium et al., 2010). The top four principal components of genetic structure identified by EIGENSOFT were used as covariates in downstream CNV association testing (Patterson et al., 2006).

CNV DISCOVERY AND QUALITY CONTROL

CNV discovery was performed for each MSKCC sample using two parallel methods. First, we applied a hidden Markov model (HMM)-based algorithm implemented in the PennCNV package (2009 Aug 27 release, Wang et al., 2007). PennCNV makes use of normalized probe intensity (R) and allelic intensity ratio (BAF) values measured across different probes on the bead array to detect regions of copy number variation in the sample. For each probe, a ratio of the observed R value to the expected R value is calculated (here, the expected value is pre-defined as the average intensity observed at the locus in a pool of healthy HapMap individuals from CEU, YRI, and CHB-JPT populations). Positive or negative deviations of the $\log R_{\text{observed}}/R_{\text{expected}}$ ratio (LRR) from zero indicate that the locus may be either duplicated or deleted, respectively. The algorithm incorporates spatial information as well, such that the probability of transitioning between different copy number states is dependent upon the physical map distance between two adjacent loci.

PennCNV was applied to 474 individual samples (including 10 duplicate pairs) using default parameters and GC-wave correction (Diskin et al., 2008). From each duplicate pair, one sample was kept for downstream analysis on the basis of having the lowest LRR value. Quality-control (QC) was then applied at the sample-level to PennCNV output by excluding samples on the basis of: (1) LRR standard deviation >0.28 (mean LRR plus 3 times the interquartile range); (2) BAF standard deviation >0.13 (approximately the mean BAF plus three times the interquartile range); or

(3) a total number of CNV calls >124 (approximately the mean call rate plus 1.5 times the interquartile range). Additional QC was applied at the CNV-level by excluding CNV calls <5 kb in length and spanning <5 probes. After QC, we derived a total of 11,635 CNV calls from 417 unique samples using PennCNV.

Secondly, we applied an Objective Bayes HMM-based algorithm implemented in QuantiSNP (version 2.3 beta, Colella et al., 2007). QuantiSNP was run using default parameters and GC-wave correction on 474 individual samples (including 10 duplicate pairs). One sample from each duplicate pair was kept for downstream analysis on the basis of having the lowest LRR value. QC was then applied at the sample-level by excluding samples on the basis of: (1) LRR standard deviation >0.21; (mean plus 3 times the interquartile range); (2) BAF standard deviation >0.102 (approximately the mean plus three times the interquartile range); (3) a total number of CNV calls >160 (approximately the mean call rate plus two times the interquartile range). Additional QC was applied at the CNV-level by excluding CNV calls with logBF confidence scores <15. After QC, we derived a total of 5422 CNV calls from 414 unique samples using QuantiSNP.

Lastly, as an additional QC procedure, we retained only those CNV calls that were made in the same individual by both PennCNV and QuantiSNP. Any sample or CNV call that was present in just one of the result-sets was excluded. The boundaries of each merged CNV call were defined using the smallest starting coordinate and largest end coordinate from either algorithm. This procedure yielded 3520 merged CNV calls from 392 unique individuals.

CNV ANNOTATION

The start and end coordinates of each CNV in our dataset were based on the March 2006 human genome build (NCBI36/hg18). For comparison to previously-reported CNV loci, we obtained the 2012-03-29 release of the Database of Genomic Variants (DGV).

DEFINITION OF CNV REGIONS (CNVRs)

CNVRs were defined as any contiguous segment of the genome spanned by a CNV call from any sample. To identify CNVRs, we applied an iterative clustering procedure to the QC-filtered CNV calls, whereby CNV calls with a mutual overlap of $\geq 40\%$ were considered to be members of the same CNVR cluster. The boundaries of the CNVR clusters were “relaxed,” such that the starting and ending coordinates were based by the smallest and largest coordinates of any CNV that was a member of the cluster, respectively.

IN SILICO CNVR GENOTYPING

In silico CNVR genotyping was performed using the CNVtools package (Barnes et al., 2008). For each CNVR of interest, we systematically evaluated the parameter space of data summarization methods, number of copy-number components, and signal/variance model specifications:

1. Data summarization. As starting input for CNVtools, we extracted the normalized signal intensities of probes on the

array that mapped within the boundaries of a given CNVR of interest. The probe intensities for a given region and sample were summarized using either one of two methods: principal component analysis or simply taking the mean. For each CNVR under investigation, we selected the method which gave the best separation between different copy-number clusters by visual inspection.

Table 1 | Association of demographic and experimental factors with CNV burden.

Variable	Beta	P-value
DNA SOURCE		
Blood	1.0 (ref)	
Buccal	0.72	0.29
Saliva	0.69	0.38
Age	0.02	0.30
GENDER		
Female	1.0 (ref)	
Male	0.38	0.30
EXPERIMENTAL BATCH		
Batch1	1.0 (ref)	
Batch2	3.07	0.02
Batch3	2.26	0.08
Batch4	-0.14	0.91
Batch5	1.26	0.37
Batch6	1.41	0.33
Batch7	0.36	0.81
Batch8	0.41	0.78
Batch9	1.26	0.32
Batch10	1.17	0.35
Batch11	3.64	0.005
Batch12	1.14	0.38
Batch13	2.86	0.03
Batch14	1.21	0.35
Batch15	0.46	0.75
Batch16	4.41	0.003
Batch17	2.09	0.11
Batch18	1.71	0.18
Batch19	0.66	0.60
Batch20	0.63	0.63
Batch21	0.44	0.74
Batch22	2.36	0.22
Batch23	-1.39	0.47
Batch24	-0.14	0.91
Batch25	0.00	1.00
Batch26	3.58	0.001
Batch27	3.24	0.002
Batch28	3.57	0.02
PLATFORM		
Illumina CNV370-duo	1.0 (ref)	
Illumina CNV370-quad	2.10	3.66×10^{-7}

Differences in CNV burden between different levels of a factor were evaluated using linear regression.

- Copy-number components. For each CNVR under investigation, we used the Bayesian Information Criterion (BIC) and our subjective visual assessment of clustering quality to select the optimal number of copy-number classes used for genotyping.
- Signal and variance model selection. For each CNVR under investigation, we explored different combinations of linear models to describe the signal mean and variance of each copy-number class (*model.mean* and *model.var* parameters, respectively). We considered models for signal mean that were “free” (stratified by copy-number class) or proportional to copy-number. Similarly, we considered models for signal variance that were either free (stratified by copy-number class), proportional to copy-number, or constant for each copy-number. Selection was based on successful convergence of the model fitting procedure and visual assessment of the clustering quality.

Individual CNVR loci were excluded from downstream risk association testing on the basis of being too problematic for *in silico* CNVR genotyping—e.g., rare or singleton (detected in only one sample) events, and noisy or insufficient separation between different copy-number clusters.

STATISTICAL METHODS TO EVALUATE RISK ASSOCIATION

Comparison of CNV burden

In this study, “CNV burden” was estimated on a per-individual basis by counting the number of CNV calls made in a given individual. We compared the estimated CNV burden between

Table 2 | Characteristics of the CNV discovery samples.

	Cases <i>n</i> = 223	Controls <i>n</i> = 169	<i>P</i> -value ^a
	<i>n</i> (%)	<i>n</i> (%)	
GENDER			
Male	129 (57.8)	71 (42.0)	0.003
Female	94 (42.2)	98 (58.0)	
AGE			
≤50	58 (26.0)	33 (19.5)	0.45
51–60	60 (26.9)	54 (32.0)	
61–70	78 (35.0)	62 (36.7)	
>70	27 (12.1)	20 (11.8)	
FAMILY HISTORY			
Yes	24 (10.8)	0 (0.0)	<0.001
No	197 (88.3)	169 (100.0)	
Missing	2 (0.9)	0 (0.0)	
DNA SOURCE			
Blood	31 (13.9)	0 (0.0)	<0.001
Buccal	156 (70.0)	143 (84.6)	
Saliva	36 (16.1)	26 (15.4)	
SNP array platform			
Illumina human CNV370-duo	144 (64.6)	151 (89.3)	<0.001
Illumina human CNV370-quad	79 (35.4)	18 (10.7)	

^a*P*-value based on Fisher's exact test.

cases and controls using either univariate or multivariate logistic regression models in R. Notably, as reported by others, estimates of CNV burden are susceptible to non-specific sources of bias, including batch effects, DNA source effects, and age (International Hapmap 3 Consortium et al., 2010). Therefore, we analyzed the effects of DNA source, SNP array platform (Illumina Human CNV370-duo vs. CNV370-quad), and experimental batch on CNV burden using univariate linear regression models (Table 1). Variables that were associated with CNV burden ($p < 0.1$) were used as covariates in a multivariable logistic regression model to test for the association of CNV burden with risk. Our final multivariate model adjusted for age, gender, the top four principal components of genetic ancestry, experimental batch, and SNP array platform. The case-control comparison was made for (1) both deletion and duplication calls together, (2) deletion calls alone, or (3) duplication calls alone. In either univariate or multivariate models, we estimated the odds ratio per-unit-of-CNV burden, and significance was determined using the 1-degree-of-freedom (df) Wald test.

Single-locus risk association testing

We evaluated individual CNVRs for association with risk using two approaches. First, we used the approach implemented in CNVtools and described previously (Barnes et al., 2008). Briefly,

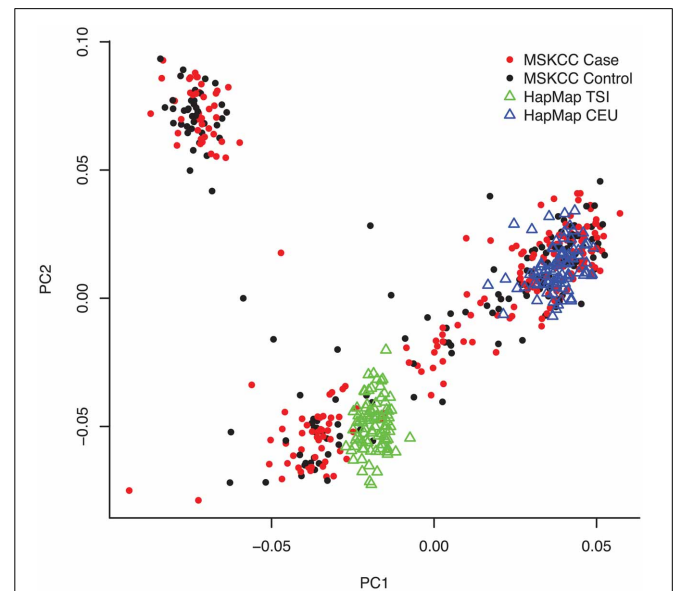


FIGURE 1 | Population structure of the study samples revealed by principal component analysis. Following SNP array genotyping, we applied the EIGENSTRAT package to 43,909 pairwise independent ($r^2 < 0.1$) SNPs with minor allele frequency (MAF) > 0.05 and call rates $> 95\%$ among the 223 pancreatic cancer cases, 169 controls, and 253 individuals from reference HapMap CEU and TSI populations. A plot of the top two principal components of genetic variation (PC1 and PC2) is shown with cases as red dots, controls as black dots, CEU reference samples as blue triangles, and TSI reference samples as green triangles. As expected for our New York-based population study, the majority of cases and controls clustered with either the CEU reference samples (i.e., central European ancestry) or TSI (southern Italian ancestry). A subset of cases and controls (representing those with Ashkenazi Jewish ancestry) clustered separately.

for each CNVR locus, the approach is to jointly fit two models: (1) a Gaussian mixture model describing the relationship of signal intensity to copy-number genotype and (2) a generalized logit linear model describing the log-linear relationship of case-control status to copy-number phenotype. The models are fit twice, once under the null hypothesis of no risk association and again under the alternative hypothesis. A likelihood ratio test is then performed comparing the likelihood of the two fits with 1 df.

Our second approach was a multivariate logistic regression model that adjusted for age, gender, the top four principal components of genetic ancestry, and DNA source. For each locus of interest, copy-number genotypes were obtained based on *in silico* genotyping (described above); we estimated the per-copy odds ratio (OR) and significance was determined using the 1-df Wald test.

RESULTS

CHARACTERISTICS OF THE STUDY PARTICIPANTS

Table 2 describes the characteristics of 223 pancreatic cancer cases and 169 healthy controls included in our analyses. The majority of samples were processed on the Illumina Human CNV370-duo platform using DNA collected from buccal sources (mouthwash or saliva). We observed a significant association of case-control status with gender (p -value = 0.003), family history (p -value < 0.001), DNA source (p -value < 0.001), and array

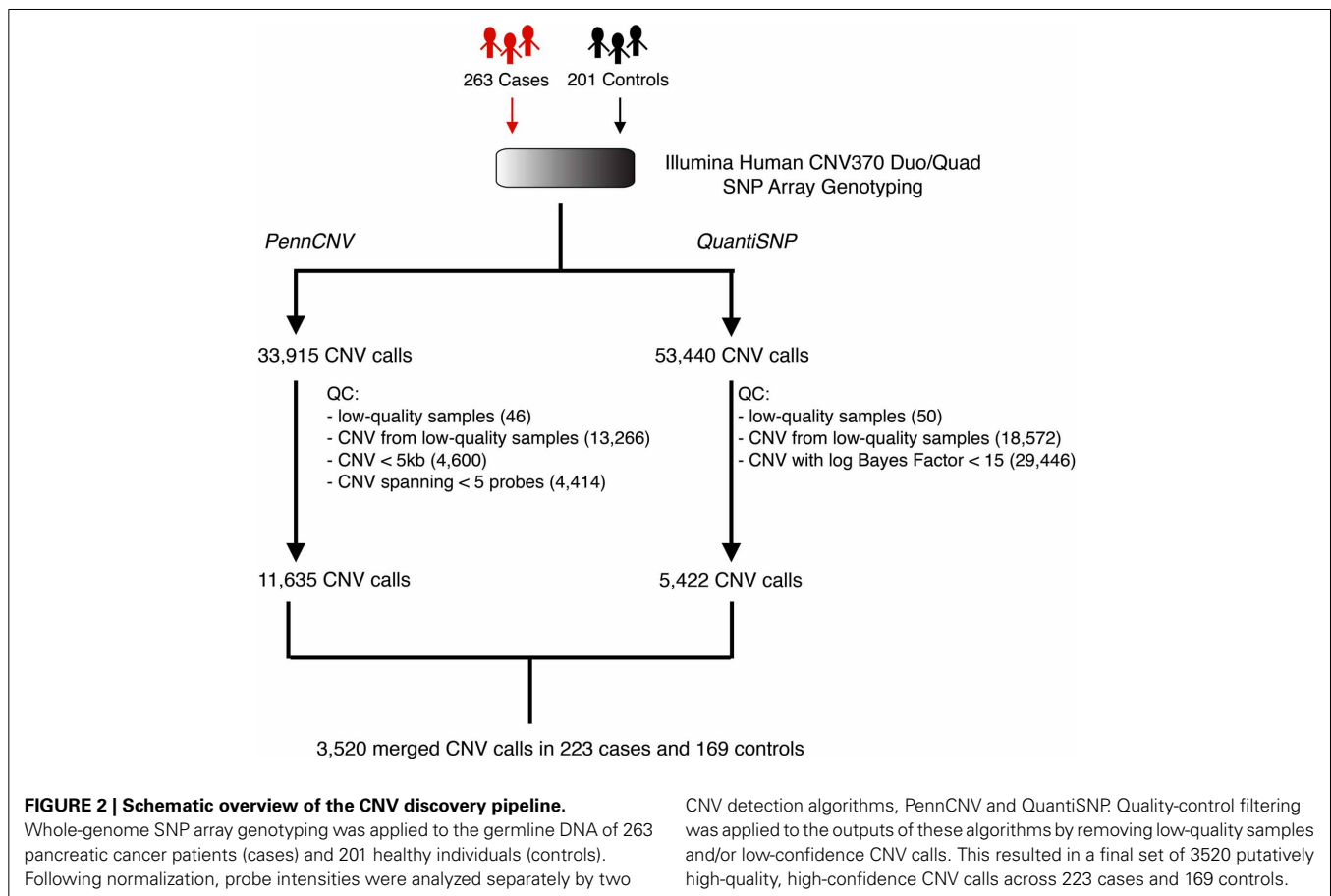
platform (p < 0.001). Following SNP array genotyping, principal component analysis revealed that the majority of our cases and controls clustered into northern and southern European genetic ancestry groups (**Figure 1**). We also observed a smaller subset of individuals that clustered separately into a group identified as Ashkenazi Jewish.

CNV DISCOVERY

Our approach to CNV discovery is summarized in **Figure 2**. After sample-level and CNV-level quality control filtering, we derived a total of 3520 high-confidence CNV calls from 223 cases and 169 controls. Of the total 3520 CNV calls, 1912 (54.3%) were deletions and 1608 (45.7%) were duplications. The median CNV length was 50.3 kb and 135.2 kb for deletions and duplications, respectively.

Notably, using a minimum overlap threshold of 40%, we found that 3407 (96%) of the CNVs discovered in our study were overlapped by a CNV previously reported in the DGV. Of the remaining 113 putatively novel CNVs, 17 (1.5%) were observed among study participants with Ashkenazi Jewish ancestry. Furthermore, 377 CNVs (194 deletions, 183 duplications) were found to be “singletons” in our study (i.e., detected in only one study sample).

To determine whether non-specific technical factors influenced our CNV discovery results, we first compared the distributions of CNV call rates across different genotyping batches



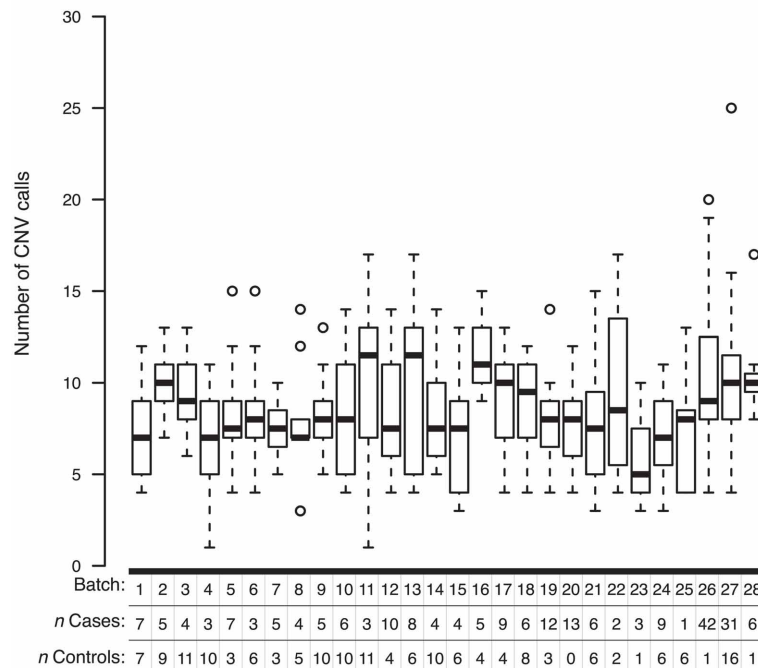


FIGURE 3 | Box-and-whisker plots of the number of CNV calls made within different experimental batches. Whole-genome SNP array genotyping was performed in 28 batches. Based on the final derived set of 3520 QC-filtered CNV calls, the interquartile range and median number of

calls in a given genotyping batch are represented by a white box and black bar, respectively. The whiskers are drawn to 1.5 times the interquartile range; circles are drawn to represent individuals with a total number of CNV calls beyond that range.

(Figure 3). Indeed, significant batch-to-batch variation was observed, suggesting that experimental “batch effects” may have played a role. Similarly, we observed significant differences in the distributions of CNV call rates for samples genotyped on the Illumina Human CNV370-duo vs.-quad platform (Figure 4). In contrast, no significant differences were observed when comparing the CNV call rate across different sample DNA sources (Figure 5).

Finally, using an iterative clustering procedure (described in Materials and Methods), we collapsed the 3520 individual CNVs into 809 unique CNV regions (CNVRs), i.e., continuous segments of the genome spanned by one or more CNVs (Figure 6).

COMPARISON OF CNV BURDEN BETWEEN CASES AND CONTROLS

Global CNV burden

Under the hypothesis that CNV burden is a risk factor for pancreatic cancer, we first sought to compare CNV burden between cases and controls. Here, considering all 3520 CNVs discovered in our study regardless of frequency, we defined CNV burden as the number of CNV calls made in an individual. This measure was derived on a per-individual basis by counting (1) deletions and duplications together, (2) deletions only, or (3) duplications only and then averaged across case and control groups (Table 3).

The average case/control CNV burden ratio was observed to be 1.07 counting all CNV types together, 0.98 counting deletions only, and 1.17 counting duplications only. To assess whether these differences in CNV burden were significant, we employed a logistic regression model and estimated the odds ratio (OR) per unit

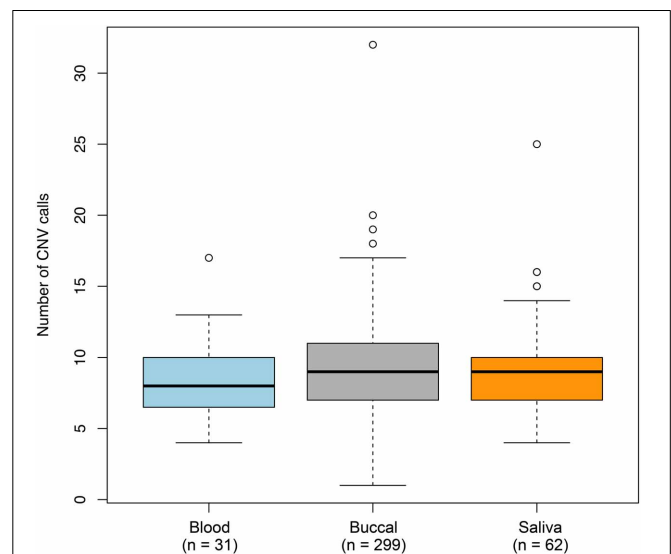


FIGURE 4 | Box-and-whisker plots of the number of CNV calls made across different DNA sources. Germline DNA was extracted from either a blood, buccal, or saliva biospecimen offered by each individual in our study. Based on the final set of 3520 QC-filtered CNV calls, the interquartile range and median number of calls derived from a given DNA source are represented by a shaded box and black bar, respectively. The whiskers are drawn to 1.5 times the interquartile range; circles are drawn to represent individuals with a total number of CNV calls beyond that range. We observed moderate (but non-statistically significant) variation in the number of CNVs detected between the DNA sources.

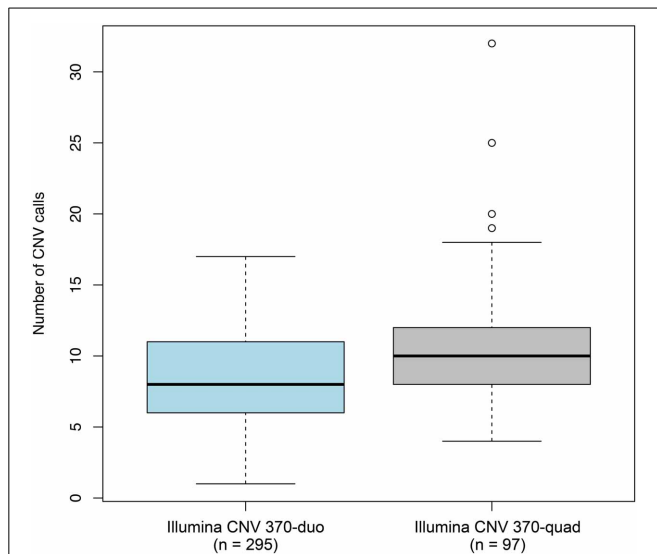


FIGURE 5 | Box-and-whisker plots of the number of CNV calls made between different configurations of the genotyping platform.

Whole-genome SNP array genotyping was performed on two different configurations of the Illumina HumanCNV370 bead array: the duo and quad. Based on the final set of 3520 QC-filtered CNV calls, the interquartile range and median number of calls derived from each configuration are represented by a shaded box and black bar, respectively. The whiskers are drawn to 1.5 times the interquartile range; circles are drawn to represent individuals with a total number of CNV calls beyond that range. We observed statistically significant differences in the number of CNVs detected between the two configurations.

of CNV burden. Under a univariate model, we observed no significant association between pancreatic cancer risk and CNV burden when counting all CNV types together ($OR = 1.05$, $p = 0.12$) or deletions only ($OR = 0.99$, $p = 0.73$). A nominally significant (p -value < 0.05) association was observed when counting duplications only ($OR = 1.10$, $p = 0.02$). However, under a multivariate model controlling for age, gender, genetic ancestry, and the non-specific effects of experimental batch and SNP array platform, we observed no statistically significant associations.

We further hypothesized that CNV burden would be enriched in patients with a family history of pancreatic cancer or early-onset disease. To evaluate this, we compared the CNV burden between controls and cases ($n = 24$) who reported a history of pancreatic cancer in at least one first-degree relative. Similarly, we compared CNV burden between controls ($n = 33$) aged 50 or younger and cases ($n = 58$) diagnosed at or prior to age 50. Again, although we observed minor differences in case/control CNV burden, these differences were not statistically significant in either univariate or multivariate analysis.

Putative rare or de novo CNV burden

To explore whether putative rare or *de novo* CNV burden is associated with pancreatic cancer risk, we extended the above analysis by considering only the subset of 377 CNV calls detected in a single individual in our study (Table 4). In both univariate and multivariate analyses, no statistically significant case-control differences in CNV burden were detected.

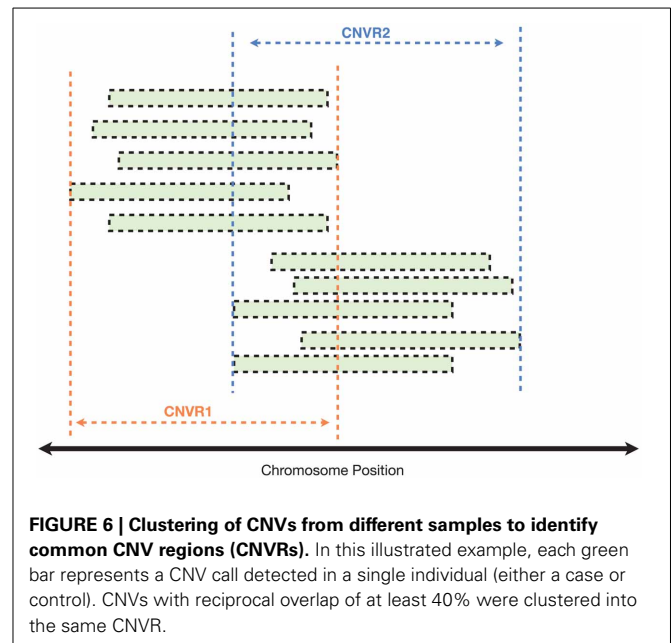


FIGURE 6 | Clustering of CNVs from different samples to identify common CNV regions (CNVRs). In this illustrated example, each green bar represents a CNV call detected in a single individual (either a case or control). CNVs with reciprocal overlap of at least 40% were clustered into the same CNVR.

ANALYSIS OF CNV LOCI PREVIOUSLY DISCOVERED IN FAMILIAL PANCREATIC CANCER

Next, we compared our CNV discovery results to loci that have been previously implicated in familial pancreatic cancer by scanning for CNVs that overlapped with 28 duplications and 25 deletions identified by Lucito et al. and 40 duplications and 53 deletions identified by Al-Sukhni et al. (Table 5). Seven overlapping CNV loci of the same type were identified, including 4 deletions and 3 duplications. At five of these loci, CNV events were observed in both our cases and controls together, or our controls alone. However, for the remaining two CNV loci (a duplication at chr18:6838462-7291170 and deletion at chr9:2235919-2351848) we observed a single CNV event exclusively in one of our cases.

ASSOCIATION OF INDIVIDUAL CNVR LOCI WITH PANCREATIC CANCER RISK

Lastly, we evaluated whether specific CNVR loci were associated with pancreatic cancer risk. To facilitate robust *in silico* CNVR genotyping and to avoid potential biases in signal characteristics between the Illumina CNV370-duo and Illumina CNV370-quad platforms, we focused this analysis on the subset of 295 samples (144 cases, 151 controls) genotyped using Illumina Human CNV370-duo. *In silico* copy-number genotyping was attempted across 706 CNVR loci that were derived from samples in this subset (Figure 7, described in Materials and Methods). Of those loci, only 176 were successfully genotyped with high quality.

Each CNVR that could be successfully genotyped was then analyzed for association with pancreatic cancer risk by use of a likelihood ratio test (Table 6). We observed a total of seven loci associated with p -values < 0.05 . Considering the number of loci tested, only one association (PA-CNVR46.1, likelihood ratio test $p = 6.41 \times 10^{-5}$) met the Bonferroni threshold of significance.

Table 3 | Comparison of global CNV burden in cases and controls based on the final derived set of 3520 QC-filtered CNV calls.

Subjects	CNV type	Mean CNV burden ^a			Logistic regression			
		Cases	Controls	Case/control ratio	Univariate		Multivariate ^b	
					OR	P-value	OR	P-value
All cases (<i>n</i> = 223) vs. all controls (<i>n</i> = 169)	All	9.22	8.66	1.07	1.05	0.12	1.01	0.80
	Deletions	4.84	4.92	0.98	0.99	0.73	1.01	0.78
	Duplications	4.38	3.73	1.17	1.10	0.02	1.00	0.93
Cases with family history (<i>n</i> = 24) vs. all controls (<i>n</i> = 169)	All	8.83	8.66	1.02	1.02	0.79	1.02	0.78
	Deletions	5.29	4.92	1.07	1.07	0.48	1.07	0.52
	Duplications	3.54	3.73	0.95	0.95	0.67	0.96	0.73
Cases diagnosed age ≤50 (<i>n</i> = 58) vs. controls age ≤50 (<i>n</i> = 33)	All	8.21	8.52	0.96	0.96	0.62	0.93	0.41
	Deletions	4.17	4.97	0.84	0.80	0.07	0.87	0.30
	Duplications	4.03	3.55	1.14	1.16	0.22	0.96	0.79

OR, per-unit CNV burden odds ratio.

^aMean CNV burden is defined as the average number of CNVs detected in each group of samples, and is derived by counting deletions and duplications together, deletions only, or duplications only.

^bMultivariate model adjusted for age, gender, genetic ancestry, experimental batch, and SNP array platform.

Table 4 | Comparison of putative rare or *de novo* CNV burden in cases and controls based on a subset of 377 CNV calls observed in a single individual.

Subjects	CNV type	Mean CNV burden ^a			Logistic regression			
		Cases	Controls	Case/control ratio	Univariate		Multivariate ^b	
					OR	P-value	OR	P-value
All cases (<i>n</i> = 223) vs. all controls (<i>n</i> = 169)	All	0.96	0.97	0.98	0.99	0.89	1.00	0.97
	Deletions	0.52	0.46	1.13	1.09	0.50	1.12	0.37
	Duplications	0.43	0.51	0.85	0.87	0.32	0.85	0.27
Cases with family history (<i>n</i> = 24) vs. all controls (<i>n</i> = 169)	All	1.04	0.97	1.07	1.05	0.78	1.03	0.87
	Deletions	0.42	0.46	0.90	0.94	0.82	0.95	0.85
	Duplications	0.63	0.51	1.23	1.20	0.49	1.14	0.65
Cases diagnosed age ≤50 (<i>n</i> = 58) vs. controls age ≤50 (<i>n</i> = 33)	All	0.79	0.97	0.82	0.78	0.34	0.83	0.55
	Deletions	0.31	0.39	0.79	0.80	0.52	0.73	0.49
	Duplications	0.48	0.58	0.84	0.84	0.53	0.94	0.88

OR, per-unit CNV burden odds ratio.

^aMean CNV burden is defined as the average number of putative rare or *de novo* CNVs detected in each group of samples, and is derived by counting deletions and duplications together, deletions only, or duplications only.

^bMultivariate model adjusted for age, gender, genetic ancestry, experimental batch, and SNP array platform.

However, in a multivariate logistic regression model adjusted for age, gender, experimental batch and the top four principal components of genetic ancestry, this region did not remain statistically significantly associated with risk (per-copy OR = 0.86, 95% CI = 0.58–1.26, *p* = 0.44).

DISCUSSION

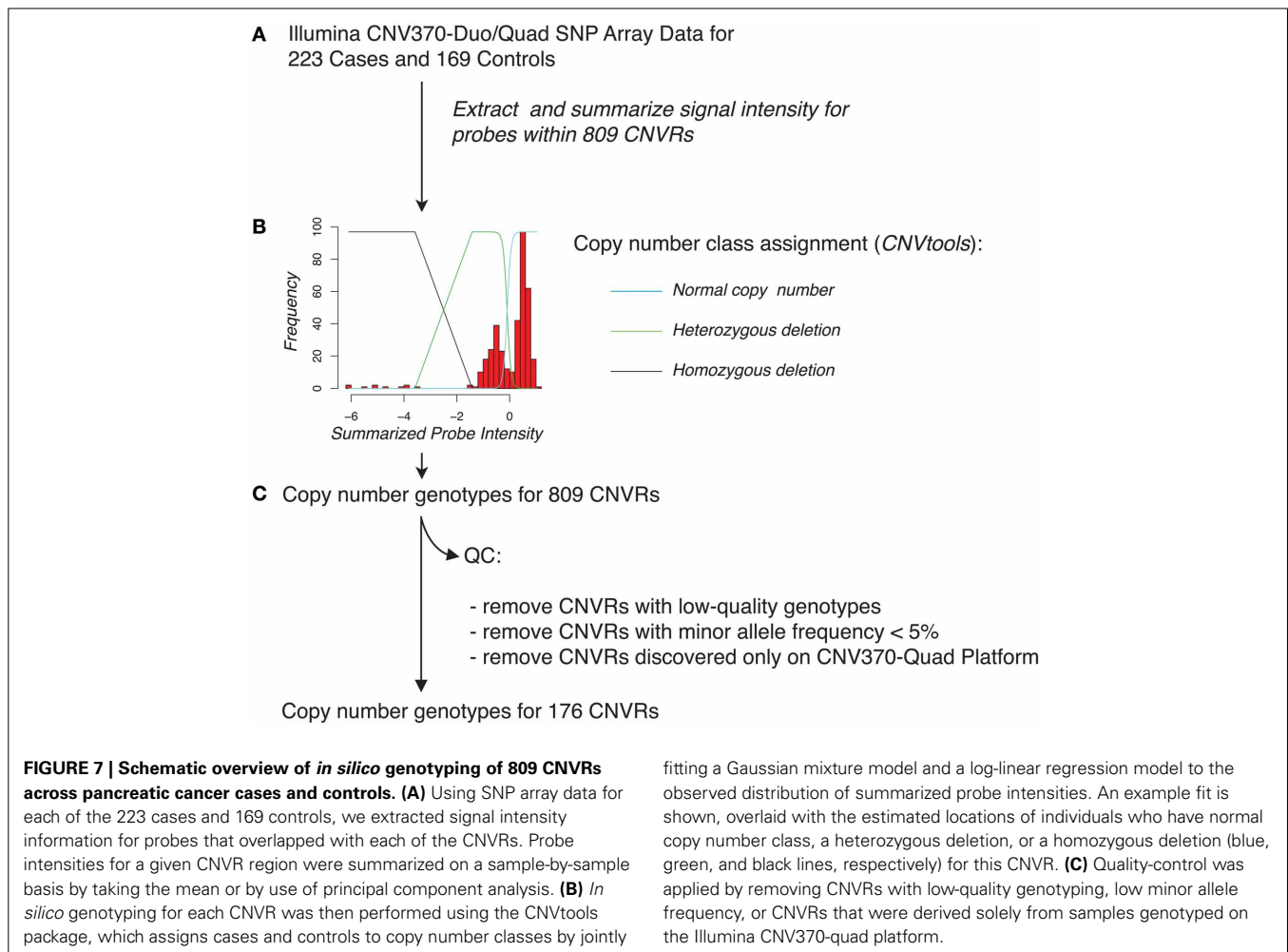
In this study, we sought to investigate the roles of CNV burden and individual CNV loci in pancreatic cancer susceptibility. Toward this end, we first performed genome-wide CNV discovery within a hospital-based cohort of 223 pancreatic cancer cases

and 169 healthy controls using a SNP array platform. To help minimize the proportion of false-positive CNVs in our data set, we took the approach of analyzing whole-genome SNP array data using two separate CNV discovery algorithms (PennCNV and QuantiSNP) followed by stringent QC filtering.

A small proportion (*n* = 113) of the CNV loci detected in our study were not overlapped by CNVs previously reported in the DGV. One likely explanation is that these regions are platform-specific artifacts. Indeed, because we did not experimentally validate the CNV loci discovered in our study, we cannot exclude the presence of artifacts in our downstream

Table 5 | Overlap between discovery CNVs (this study) and CNVs previously implicated in familial pancreatic cancer.

CNV locus (hg18)	Type		Number of samples (this study) with an overlapping CNV of same type		
			All cases (n = 223)	Cases with family history (n = 24)	Controls (n = 169)
chr12:130382166-130686668	Deletion	Al-Sukhni et al., 2012	3	1	2
chr3:60219748-60263116	Deletion	Lucito et al., 2007	3	0	6
chr18:6838462-7291170	Duplication	Al-Sukhni et al., 2012	1	0	0
chr9:2235919-2351848	Deletion	Al-Sukhni et al., 2012	1	0	0
chr11:39882017-40010124	Deletion	Al-Sukhni et al., 2012	0	0	1
chr19:2984601-5201290	Duplication	Lucito et al., 2007	0	0	1
chr7:133223330-133393933	Duplication	Al-Sukhni et al., 2012	0	0	1



analyses despite the application of rigorous filtering methods. In addition, 17 of the previously unreported CNV loci were observed among subjects with Ashkenazi Jewish ancestry. Therefore, we further speculate that at least some of the CNV calls unique to our study may be true population-specific CNVs from populations (i.e., Ashkenazi Jewish) that are underrepresented in the DGV.

Experimental batch effects are well known to the genomics field and require proper consideration when performing case-control analyses. In this study, we observed significant variation in distributions of CNV call rates across different sample batches, which is likely the result of technical variation during the course of batch processing. In support of this hypothesis, we also observed variations in CNV call rates at the individual

Table 6 | Likelihood ratio tests for the association of 176 CNVR loci with pancreatic cancer risk.

CNVR	Locus(hg18)	Type ^a	Likelihood ratio	P-value
PA_CNVR46.1	chr4:115387397-115401739	Multiallelic	15.98	6.41 × 10 ⁻⁵
PA_CNVR181.4	chr17:41792236-42143493	Multiallelic	10.79	0.001
PA_CNVR276.2	chr15:28304144-28591312	Multiallelic	7.09	0.01
PA_CNVR145.1	chr8:145589526-145701138	Deletion	6.51	0.01
PA_CNVR112.1	chr8:5586807-5592495	Deletion	6.31	0.01
PA_CNVR45.1	chr4:108285187-108293245	Deletion	4.87	0.03
PA_CNVR515.1	chr16:34325303-34618468	Duplication	4.76	0.03
PA_CNVR81.18	chr6:32563460-32577503	Multiallelic	3.87	0.05
PA_CNVR131.1	chr8:83443991-83456427	Deletion	3.84	0.05
PA_CNVR248.1	chr12:8533984-8665794	Deletion	3.83	0.05
PA_CNVR452.1	chr9:44683090-44844429	Duplication	3.80	0.05
PA_CNVR373.1	chr3:8801023-8832963	Duplication	3.59	0.06
PA_CNVR206.2	chr20:26155692-28255585	Duplication	3.35	0.07
PA_CNVR338.1	chr2:208742960-208760548	Deletion	3.29	0.07
PA_CNVR90.1	chr6:79029649-79104256	Deletion	3.26	0.07
PA_CNVR489.1	chr5:104461415-104631946	Deletion	3.08	0.08
PA_CNVR383.1	chr3:60043770-60251762	Deletion	2.79	0.09
PA_CNVR357.1	chr13:56654503-56698837	Deletion	2.78	0.10
PA_CNVR384.2	chr3:75502426-75552183	Deletion	2.76	0.10
PA_CNVR144.1	chr8:145046951-145275551	Deletion	2.68	0.10
PA_CNVR81.1	chr6:32560011-32648263	Multiallelic	2.64	0.10
PA_CNVR212.3	chr1:1191495-1278446	Multiallelic	2.56	0.11
PA_CNVR509.2	chr16:16225138-16708567	Multiallelic	2.48	0.12
PA_CNVR18.1	chr14:105630045-105837886	Multiallelic	2.42	0.12
PA_CNVR49.1	chr4:129993825-130159225	Deletion	2.41	0.12
PA_CNVR419.1	chr10:56877102-56916808	Deletion	2.37	0.12
PA_CNVR485.1	chr5:97073409-97127572	Deletion	2.35	0.13
PA_CNVR73.5	chr6:31479757-31502679	Deletion	2.31	0.13
PA_CNVR81.12	chr6:32608853-32635771	Multiallelic	2.19	0.14
PA_CNVR206.3	chr20:28043606-28255585	Multiallelic	2.12	0.14
PA_CNVR253.16	chr12:31248369-31298174	Duplication	2.12	0.15
PA_CNVR144.5	chr8:145064091-145118650	Deletion	2.08	0.15
PA_CNVR314.1	chr2:52577564-52637176	Deletion	2.04	0.15
PA_CNVR73.1	chr6:31465370-31562866	Deletion	2.03	0.15
PA_CNVR305.1	chr2:34556561-34580068	Multiallelic	2.01	0.16
PA_CNVR245.1	chr1:246802692-246852068	Deletion	1.92	0.17
PA_CNVR172.1	chr17:19446576-19475026	Deletion	1.90	0.17
PA_CNVR147.1	chr11:3179445-3351014	Deletion	1.89	0.17
PA_CNVR237.1	chr1:194992939-195168376	Multiallelic	1.88	0.17
PA_CNVR350.1	chr2:232958927-232976959	Deletion	1.79	0.18
PA_CNVR573.1	chr22:18693299-19048116	Duplication	1.78	0.18
PA_CNVR302.1	chr2:17086609-17095859	Deletion	1.76	0.18
PA_CNVR165.1	chr11:133851329-134227062	Duplication	1.75	0.19
PA_CNVR397.1	chr3:166524485-166560107	Deletion	1.74	0.19
PA_CNVR299.1	chr2:4191253-4201943	Deletion	1.68	0.19
PA_CNVR192.1	chr19:40613106-40636215	Deletion	1.68	0.20
PA_CNVR16.1	chr14:85528167-85560365	Duplication	1.53	0.22
PA_CNVR387.1	chr3:116125098-116150586	Deletion	1.46	0.23
PA_CNVR142.1	chr8:137757412-137926509	Deletion	1.42	0.23
PA_CNVR138.1	chr8:115704806-115710821	Deletion	1.41	0.23
PA_CNVR253.2	chr12:31157554-31298174	Duplication	1.41	0.23
PA_CNVR7.2	chr14:39308459-39982197	Deletion	1.35	0.24

(Continued)

Table 6 | Continued

CNVR	Locus(hg18)	Type ^a	Likelihood ratio	P-value
PA_CNVR364.1	chr13:108363498-108381356	Deletion	1.35	0.25
PA_CNVR509.1	chr16:16657423-16726778	Multiallelic	1.34	0.25
PA_CNVR423.4	chr10:67952976-68091312	Deletion	1.34	0.25
PA_CNVR353.1	chr13:18019741-18334782	Duplication	1.26	0.26
PA_CNVR472.1	chr5:31522297-31818133	Duplication	1.26	0.26
PA_CNVR234.1	chr1:187549425-187789366	Deletion	1.24	0.27
PA_CNVR347.1	chr2:230799467-230897291	Multiallelic	1.20	0.27
PA_CNVR156.1	chr11:55124465-55180783	Deletion	1.13	0.29
PA_CNVR438.2	chr9:181843-264641	Multiallelic	1.10	0.29
PA_CNVR59.2	chr4:173218118-173263440	Deletion	1.10	0.29
PA_CNVR59.1	chr4:173222335-173227450	Deletion	1.10	0.29
PA_CNVR114.3	chr8:7683445-7929107	Deletion	1.09	0.30
PA_CNVR212.2	chr1:1096336-1468043	Multiallelic	1.06	0.30
PA_CNVR156.8	chr11:55124465-55209499	Deletion	1.04	0.31
PA_CNVR81.127	chr6:32593190-32635057	Multiallelic	1.00	0.32
PA_CNVR247.1	chr12:7884583-8017012	Multiallelic	0.99	0.32
PA_CNVR444.1	chr9:11837376-12177104	Deletion	0.99	0.32
PA_CNVR292.1	chr15:85631534-85671028	Deletion	0.97	0.33
PA_CNVR193.1	chr19:48312997-48531928	Multiallelic	0.96	0.33
PA_CNVR28.1	chr4:34456032-34499424	Deletion	0.95	0.33
PA_CNVR264.1	chr12:78652724-78698368	Multiallelic	0.91	0.34
PA_CNVR114.18	chr8:7683445-7781955	Multiallelic	0.86	0.35
PA_CNVR190.1	chr19:20385941-20522325	Deletion	0.85	0.36
PA_CNVR396.1	chr3:163986784-164109279	Deletion	0.75	0.39
PA_CNVR102.1	chr6:168078929-168342182	Duplication	0.74	0.39
PA_CNVR4178	chr10:46611927-47218918	Multiallelic	0.73	0.39
PA_CNVR316.2	chr2:57256144-57299713	Multiallelic	0.73	0.39
PA_CNVR336.1	chr2:180123158-180129913	Deletion	0.72	0.39
PA_CNVR193.2	chr19:47935678-48430180	Multiallelic	0.65	0.42
PA_CNVR533.1	chr7:8791785-8833529	Deletion	0.64	0.42
PA_CNVR240.1	chr1:201109558-201523423	Multiallelic	0.63	0.43
PA_CNVR516.1	chr16:35041151-35074383	Deletion	0.59	0.44
PA_CNVR50.1	chr4:132165825-132547664	Deletion	0.56	0.45
PA_CNVR324.1	chr2:89682553-89911010	Multiallelic	0.55	0.46
PA_CNVR68.5	chr6:29959422-29969546	Deletion	0.51	0.47
PA_CNVR436.2	chr10:134890273-135000022	Deletion	0.50	0.48
PA_CNVR154.1	chr11:50578631-50687058	Duplication	0.49	0.48
PA_CNVR108.1	chr8:2232502-2570171	Duplication	0.49	0.49
PA_CNVR74.1	chr6:31566612-31580699	Deletion	0.46	0.50
PA_CNVR467.1	chr5:8756085-8797557	Deletion	0.46	0.50
PA_CNVR606.1	chr18:61352708-61358853	Multiallelic	0.45	0.50
PA_CNVR181.1	chr17:41780482-42092850	Multiallelic	0.42	0.51
PA_CNVR546.1	chr7:64407696-64593616	Deletion	0.39	0.53
PA_CNVR294.1	chr15:95622079-95632771	Deletion	0.39	0.53
PA_CNVR181.7	chr17:41636474-42143493	Multiallelic	0.38	0.54
PA_CNVR353.2	chr13:17922259-18120572	Duplication	0.38	0.54
PA_CNVR72.1	chr6:31382534-31406722	Deletion	0.37	0.54
PA_CNVR72.41	chr6:31382224-31419324	Deletion	0.37	0.54
PA_CNVR417.1	chr10:46830464-47218918	Multiallelic	0.37	0.54
PA_CNVR514.2	chr16:32467276-32498422	Deletion	0.37	0.54
PA_CNVR517.1	chr16:44943958-45048915	Duplication	0.37	0.54
PA_CNVR392.1	chr3:133185033-133195707	Multiallelic	0.36	0.55

(Continued)

Table 6 | Continued

CNVR	Locus(hg18)	Type ^a	Likelihood ratio	P-value
PA_CNVR593.1	chr21:29458916-29478852	Deletion	0.36	0.55
PA_CNVR68.19	chr6:29988619-29999402	Deletion	0.35	0.55
PA_CNVR68.6	chr6:29972182-29996478	Deletion	0.34	0.56
PA_CNVR208.1	chr20:52078573-52094148	Deletion	0.32	0.57
PA_CNVR607.1	chr18:64003719-64042401	Deletion	0.30	0.59
PA_CNVR146.1	chr8:146116506-146137021	Deletion	0.30	0.59
PA_CNVR133.1	chr8:85420095-85433884	Deletion	0.29	0.59
PA_CNVR143.1	chr8:144686338-144780417	Deletion	0.28	0.59
PA_CNVR231.1	chr1:173054347-173067547	Deletion	0.28	0.60
PA_CNVR68.1	chr6:29950151-30021706	Deletion	0.27	0.60
PA_CNVR421.1	chr10:66980652-66988454	Deletion	0.25	0.62
PA_CNVR253.1	chr12:31180151-31237140	Duplication	0.25	0.62
PA_CNVR68.7	chr6:30000415-30007126	Deletion	0.24	0.62
PA_CNVR63.1	chr6:243700-326918	Multiallelic	0.24	0.62
PA_CNVR88.1	chr6:67075448-67104015	Deletion	0.24	0.63
PA_CNVR246.1	chr12:6114170-6134080	Deletion	0.20	0.65
PA_CNVR36.1	chr4:68608212-68676295	Deletion	0.20	0.65
PA_CNVR468.1	chr5:9951962-9981862	Deletion	0.20	0.66
PA_CNVR516.3	chr16:35041151-35141900	Deletion	0.19	0.66
PA_CNVR78.7	chr6:32059186-32065343	Deletion	0.19	0.66
PA_CNVR125.1	chr8:39331592-39509376	Multiallelic	0.18	0.67
PA_CNVR254.1	chr12:31898373-31965665	Multiallelic	0.18	0.67
PA_CNVR482.13	chr5:69611483-69791981	Multiallelic	0.18	0.68
PA_CNVR572.9	chr22:17318367-17396663	Duplication	0.17	0.68
PA_CNVR102.6	chr6:168209041-168274300	Duplication	0.16	0.69
PA_CNVR255.1	chr12:33193705-33201064	Deletion	0.16	0.69
PA_CNVR273.1	chr15:21601351-21612590	Deletion	0.16	0.69
PA_CNVR81.35	chr6:32610165-32656281	Multiallelic	0.16	0.69
PA_CNVR256.3	chr12:34261193-34692538	Duplication	0.14	0.71
PA_CNVR436.1	chr10:134913018-134948335	Deletion	0.14	0.71
PA_CNVR422.1	chr10:67749354-67785209	Deletion	0.10	0.75
PA_CNVR34.1	chr4:63352170-63357704	Deletion	0.08	0.77
PA_CNVR37.1	chr4:69007217-69210001	Deletion	0.08	0.77
PA_CNVR73.2	chr6:31465923-31485621	Deletion	0.07	0.78
PA_CNVR78.1	chr6:32057331-32118241	Deletion	0.07	0.78
PA_CNVR451.1	chr9:43515795-43730292	Deletion	0.07	0.79
PA_CNVR70.1	chr6:30565183-30617261	Deletion	0.05	0.82
PA_CNVR387.2	chr3:116143746-116150586	Deletion	0.05	0.82
PA_CNVR352.1	chr2:242565979-242692820	Deletion	0.05	0.83
PA_CNVR352.9	chr2:242412215-242645262	Deletion	0.04	0.84
PA_CNVR229.1	chr1:167436480-167513579	Deletion	0.03	0.85
PA_CNVR147.2	chr11:3240658-3297012	Multiallelic	0.03	0.86
PA_CNVR228.1	chr1:147435422-147637598	Multiallelic	0.03	0.86
PA_CNVR578.1	chr22:22653131-22728586	Multiallelic	0.03	0.86
PA_CNVR556.5	chr7:110847122-110883322	Deletion	0.03	0.87
PA_CNVR473.1	chr5:32137157-32202977	Duplication	0.03	0.87
PA_CNVR158.1	chr11:70966737-71226822	Duplication	0.03	0.87
PA_CNVR227.1	chr1:146409913-146483416	Multiallelic	0.02	0.88
PA_CNVR102.4	chr6:168092530-168162650	Duplication	0.02	0.88
PA_CNVR382.1	chr3:46771354-46825614	Deletion	0.02	0.88
PA_CNVR482.19	chr5:69724106-69791981	Duplication	0.02	0.89
PA_CNVR278.1	chr15:29812822-30302218	Multiallelic	0.02	0.90

(Continued)

Table 6 | Continued

CNVR	Locus(hg18)	Type ^a	Likelihood ratio	P-value
PA_CNVR280.1	chr15:32459510-32625184	Deletion	0.01	0.91
PA_CNVR511.1	chr16:19857185-19862969	Deletion	0.01	0.92
PA_CNVR451.5	chr9:43599125-43616717	Deletion	0.01	0.94
PA_CNVR451.20	chr9:43515795-43616717	Deletion	0.01	0.94
PA_CNVR11.1	chr14:44251087-44294325	Deletion	0.00	0.95
PA_CNVR324.5	chr2:89029231-91046486	Multiallelic	0.00	0.95
PA_CNVR8.1	chr14:40660199-40763690	Deletion	0.00	0.95
PA_CNVR7.1	chr14:39506301-39650286	Deletion	0.00	0.95
PA_CNVR249.1	chr12:9496550-9616735	Multiallelic	0.00	0.97
PA_CNVR57.1	chr4:162239613-162310205	Deletion	0.00	0.97
PA_CNVR3.1	chr14:21014014-21038716	Deletion	0.00	0.97
PA_CNVR19.1	chr14:106185238-106297061	Multiallelic	0.00	0.97
PA_CNVR170.1	chr17:9937668-10337719	Duplication	0.00	0.97
PA_CNVR182.1	chr17:46849574-46910094	Duplication	0.00	0.98
PA_CNVR175.1	chr17:21967881-22013983	Deletion	0.00	0.99
PA_CNVR71.1	chr6:30854607-30864253	Deletion	0.00	0.99
PA_CNVR237.3	chr1:195089923-195168376	Multiallelic	0.00	0.99
PA_CNVR276.4	chr15:28714502-28881771	Multiallelic	0.00	0.99
PA_CNVR579.1	chr22:23983992-24254444	Multiallelic	0.00	1.00
PA_CNVR210.1	chr20:61880157-62011862	Deletion	-1.09	1.00

^a CNVR type is defined by the observation of either deletion genotypes only ("deletion"), duplication genotypes only ("duplication"), or both deletion and duplication genotypes ("multiallelic") among the analyzed individuals.

sample-level by comparing CNVs from 10 duplicate pairs of samples that were genotyped twice on the same array platform but in two different batches (data not shown). Furthermore, the distribution of CNV call rates for samples genotyped on the Illumina Human CNV370-duo vs. -quad platforms were significantly different. While batch effects may have contributed to this observation, we speculate that it also reflects differences in probe-level characteristics between the duo and quad bead array configurations.

Following CNV discovery, we explored several hypotheses that CNV burden is a risk factor for pancreatic cancer. Different formulations of "CNV burden" have been employed in the literature, including: a simple CNV count, the total CNV length, and the total number of genes overlapped by a CNV. Here, we regarded CNV burden as the number of CNV calls made per individual—a metric that was evaluated across a spectrum of different CNV types and frequency, and different case-control subgroups.

When we considered all 3520 CNVs discovered in our study across all 223 cases and 169 controls, we found no strong evidence of an association between CNV burden and pancreatic cancer risk. Similarly, no evidence was found when we restricted our analyses to cases with early-onset (age ≤ 50) disease or with a family history of pancreatic cancer. Notably, a recent study by Stadler et al. found that a significant proportion of men with early-onset testicular cancer harbored *de novo* CNVs (Stadler et al., 2012). Although the pathogenicity of these specific *de novo* CNVs has yet to be confirmed, this finding suggested a novel framework for understanding the genetic basis of sporadic cancers. We explored this hypothesis by restricting our analyses to

singleton CNVs (i.e., putative rare or *de novo* CNVs), but found no evidence of association.

Although our results do not support the role of CNV burden in pancreatic cancer risk, we emphasize that our conclusions are tempered by the presence of experimental variability in the CNV discovery scheme. It remains possible that true differences in case/control CNV burden might have been masked by the presence of varying DNA sources, genotyping platforms, and experimental batch performance.

We next attempted *in silico* validation of CNV loci that were previously implicated in familial pancreatic cancer risk. By comparing our CNV discovery results against those reported by Lucito et al. and Al-Sukhni et al. we identified two loci in which the CNV events were present exclusively in our cases—and, thus, consistent with the original hypotheses that each locus confers a strong risk of pancreatic cancer. The first such locus, on chromosome 9p24.2, was reported by Al-Sukhni et al. as a deletion event in a single case, but harbors no known RefSeq genes. The second locus, on chromosome 18p11.31, was reported by Al-Sukhni as a duplication event in a single case and harbors four known genes including *ARHGAP28*, *LINC00668*, *LAMA1*, *LRRC30*. Notably, *LAMA1* (laminin subunit alpha-1 precursor) encodes a subunit of the extracellular protein *laminin*. While the specific role of *laminin* in pancreatic tumorigenesis remains poorly understood, a recent study by Vincent et al. found that *LAMA1* was among several genes that were hypermethylated and underexpressed in pancreatic tumor samples compared to normal pancreas (Vincent et al., 2011). Our results add further support to the hypothesis that inherited duplications at the *LAMA1* locus may be involved

in pancreatic cancer risk. However, as we did not experimentally validate the duplication involving *LAMA1* in our data set, we cannot exclude the possibility of either false positives or negatives.

Lastly, to determine whether specific CNV loci play a role in pancreatic cancer risk, we performed *in silico* copy-number genotyping and association testing across 176 CNVR loci identified in our discovery experiment. Using a likelihood ratio testing approach, we identified seven regions putatively associated with pancreatic cancer risk at the nominal threshold of $p < 0.05$. Only one locus (PA-CNVR46.1) reached the Bonferroni level of significance. This locus maps to a non-genetic region of chromosome 4q26 and was found to be multiallelic in our samples (i.e., both deletion and duplication genotypes were detected). However, its effect on risk was not significant in a multivariate logistic regression model adjusted for gender, age, ancestry, and DNA source. Follow-up studies of these seven putative loci would be necessary to validate their associations with pancreatic cancer risk.

Yet, prior to the conclusion of our study, work performed by Conrad et al. suggested that nearly 77% of all common CNVs in the human genome are tagged ($r^2 > 0.8$) by SNPs through linkage disequilibrium (Conrad et al., 2009). Hence, one could speculate that any common CNV locus with weak-to-moderate effects on pancreatic cancer risk would have been detected already through large-scale GWAS involving several thousand cases and controls (Amundadottir et al., 2009; Petersen et al., 2010). Thus, in this context, we strongly emphasize that our inability to detect such CNV loci is not unexpected given the relatively small sample size of our study.

As a corollary, one could further hypothesize that CNVs with large contributions to pancreatic cancer risk are likely to be individually rare and poorly tagged by SNPs on commercially available genotyping arrays. Hence, we also emphasize that such CNV loci were likely to have been missed in our study due to not only small sample size, but also the resolution of the Illumina Human CNV370 bead array.

In conclusion, based on the results of genome-wide CNV discovery in a hospital-based case-control cohort, our study found no evidence that CNVs contribute substantially to the genetic etiology of pancreatic cancer. However, in light of recent population-wide CNV data and the challenges faced by our study, future efforts to address the role of CNVs in pancreatic cancer will require larger case-control groups and high-resolution discovery platforms.

ACKNOWLEDGMENTS

We are grateful to the Geoffrey Beene Cancer Research Center at MSKCC and the Emerald Foundation for supporting this work.

REFERENCES

- Al-Sukhni, W., Joe, S., Lionel, A. C., Zwingerman, N., Zogopoulos, G., Marshall, C. R., et al. (2012). Identification of germline genomic copy number variation in familial pancreatic cancer. *Hum. Genet.* 131, 1481–1494. doi: 10.1007/s00439-012-1183-1
- Amundadottir, L., Kraft, P., Stolzenberg-Solomon, R., Fuchs, C., Petersen, G., Arslan, A., et al. (2009). Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat. Genet.* 41, 986–990. doi: 10.1038/ng.429
- Barnes, C., Plagnol, V., Fitzgerald, T., Redon, R., Marchini, J., Clayton, D., et al. (2008). A robust statistical method for case-control association testing with copy number variation. *Nat. Genet.* 40, 1245–1252. doi: 10.1038/ng.206
- Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., et al. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* 35, 2013–2025. doi: 10.1093/nar/gkm076
- Conrad, D., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., et al. (2009). Origins and functional impact of copy number variation in the human genome. *Nature* 464, 704–712. doi: 10.1038/nature08516
- Diskin, S. J., Hou, C., Glessner, J. T., Attiyeh, E. F., Laudenslager, M., Bosse, K., et al. (2009). Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* 459, 987–991. doi: 10.1038/nature08035
- Diskin, S. J., Li, M., Hou, C., Yang, S., Glessner, J., Hakonarson, H., et al. (2008). Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* 36, e126. doi: 10.1093/nar/gkn556
- Iafate, A., Feuk, L., Rivera, M., Listewnik, M., Donahoe, P., Qi, Y., et al. (2004). Detection of large-scale variation in the human genome. *Nat. Genet.* 36, 949–951. doi: 10.1038/ng1416
- International Hapmap 3 Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., et al. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature* 467, 52–58. doi: 10.1038/nature09298
- Jiao, L., Bondy, M., Hassan, M., Wolff, R., Evans, D., Abbruzzese, J., et al. (2006). Selected polymorphisms of DNA repair genes and risk of pancreatic cancer. *Cancer Detect. Prev.* 30, 284–291. doi: 10.1016/j.cdp.2006.05.002
- Jiao, L., Hassan, M., Bondy, M., Wolff, R., Evans, D., Abbruzzese, J., et al. (2008). XRCC2 and XRCC3 gene polymorphism and risk of pancreatic cancer. *Am. J. Gastroenterol.* 103, 360–367. doi: 10.1111/j.1572-0241.2007.01615.x
- Li, D., Li, Y., Jiao, L., Chang, D., Beinart, G., Wolff, R., et al. (2007). Effects of base excision repair gene polymorphisms on pancreatic cancer survival. *Int. J. Cancer* 120, 1748–1754. doi: 10.1002/ijc.22301
- Lucito, R., Suresh, S., Walter, K., Pandey, A., Lakshmi, B., Krasnitz, A., et al. (2007). Copy-number variants in patients with a strong family history of pancreatic cancer. *Cancer Biol. Ther.* 6, 1592–1599. doi: 10.4161/cbt.6.10.4725
- McWilliams, R., Bamlet, W., Cunningham, J., Goode, E., De Andrade, M., Boardman, L., et al. (2008). Polymorphisms in DNA repair genes, smoking, and pancreatic adenocarcinoma risk. *Cancer Res.* 68, 4928–4935. doi: 10.1158/0008-5472.CAN-07-5539
- Mukherjee, S., Simon, J., Bayuga, S., Ludwig, E., Yoo, S., Orlow, I., et al. (2011). Including additional controls from public databases improves the power of a genome-wide association study. *Hum. Hered.* 72, 21–34. doi: 10.1159/000330149
- Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet* 2:e190. doi: 10.1371/journal.pgen.0020190
- Petersen, G., Amundadottir, L., Fuchs, C., Kraft, P., Stolzenberg-Solomon, R., Jacobs, K., et al. (2010). A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat. Genet.* 42, 224–228. doi: 10.1038/ng.522
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795
- Siegel, R., Naishadham, D., and Jemal, A. (2012). Cancer statistics, 2012. *CA Cancer J. Clin.* 62, 10–29. doi: 10.3322/caac.20138
- Stadler, Z. K., Esposito, D., Shah, S., Vijai, J., Yamrom, B., Levy, D., et al. (2012). Rare *de novo* germline copy-number variation in testicular cancer. *Am. J. Hum. Genet.* 91, 379–383. doi: 10.1016/j.ajhg.2012.06.019
- Vincent, A., Omura, N., Hong, S.-M., Jaffe, A., Eshleman, J., and Goggins, M. (2011). Genome-wide analysis of promoter methylation associated with gene expression profile in pancreatic adenocarcinoma. *Clin. Cancer Res.* 17, 4341–4354. doi: 10.1158/1078-0432.CCR-10-3431
- Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S., et al. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* 17, 1665–1674. doi: 10.1101/gr.6861907
- Willis, J. A., Olson, S. H., Orlow, I., Mukherjee, S., McWilliams, R. R., Kurtz, R. C., et al. (2012). A replication study and genome-wide scan of single-nucleotide

polymorphisms associated with pancreatic cancer risk and overall survival. *Clin. Cancer Res.* 18, 3942–3951. doi: 10.1158/1078-0432.CCR-11-2856

Zhang, F., Gu, W., Hurles, M., and Lupski, J. (2009). Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* 10, 451–481. doi: 10.1146/annurev.genom.9.081307.164217

Zhang, J., Feuk, L., Duggan, G. E., Khaja, R., and Scherer, S. W. (2006). Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. *Cytogenet. Genome Res.* 115, 205–214. doi: 10.1159/000095916

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 25 September 2013; paper pending published: 28 October 2013; accepted: 26 January 2014; published online: 13 February 2014.

Citation: Willis JA, Mukherjee S, Orlov I, Viale A, Offit K, Kurtz RC, Olson SH and Klein RJ (2014) Genome-wide analysis of the role of copy-number variation in pancreatic cancer risk. *Front. Genet.* 5:29. doi: 10.3389/fgene.2014.00029

This article was submitted to *Applied Genetic Epidemiology*, a section of the journal *Frontiers in Genetics*.

Copyright © 2014 Willis, Mukherjee, Orlov, Viale, Offit, Kurtz, Olson and Klein. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.