Published in final edited form as:

Pancreas. 2013 March ; 42(2): 209–215. doi:10.1097/MPA.0b013e318264cea5.

A re-sequence analysis of genomic loci on chromosomes 1q32.1, 5p15.33 and 13q22.1 associated with pancreatic cancer risk

Hemang Parikh, PhD¹, Jinping Jia, PhD¹, Xijun Zhang, MS^{2,3}, Charles C. Chung, PhD¹, Kevin B. Jacobs, PhD^{2,3}, Meredith Yeager, PhD^{2,3}, Joseph Boland, MS^{2,3}, Amy Hutchinson, MS^{2,3}, Laura Burdett, PhD^{2,3}, Jason Hoskins, PhD¹, Harvey A. Risch, MD, PhD⁴, Rachael Z. Stolzenberg-Solomon, MPH, PhD², Stephen J. Chanock, MD¹, Brian M. Wolpin, MD, MPH^{5,6}, Gloria M. Petersen, PhD⁷, Charles S. Fuchs, MD, MPH^{5,6}, Patricia Hartge, ScD², and Laufey Amundadottir, PhD¹

¹Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

²Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

³Core Genotyping Facility, SAIC-Frederick, Inc., NCI-Frederick, Frederick, MD, USA

⁴Yale University School of Public Health, School of Medicine, New Haven, CT, USA

⁵Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA

⁶Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

⁷Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA

Abstract

Objectives—To fine-map common pancreatic cancer susceptibility regions.

Methods—We conducted targeted Roche-454 re-sequencing across 428 kb in three genomic regions identified in genome-wide association studies (GWAS) of pancreatic cancer, on chromosomes 1q32.1, 5p15.33 and 13q22.1.

Correspondence to: Laufey Amundadottir, Ph.D., Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Institutes of Health, 8717 Grovemont Circle, Gaithersburg, MD 20877. Phone: 301-594-8131, Fax: 301-402-3134, amundadottirl@mail.nih.gov.

Conflict of interest statement

The authors report no financial interests or potential conflicts of interests.

Electronic database information

dbSNP database at the National Center for Biotechnology Information, NIH: <http://www.ncbi.nlm.nih.gov/SNP/>

The International HapMap database: <http://www.hapmap.org/>

1000 Genomes database: <http://www.1000genomes.org/>

GLU-genetics: <http://code.google.com/p/glu-genetics/>

NetPrimer: <http://www.premierbiosoft.com/netprimer/index.html>

Newbler: <http://my454.com/products/analysis-software/index.asp>

Primer3: <http://frodo.wi.mit.edu>

GATK bundle: <ftp://ftp.broadinstitute.org/bundle/1.2/hg19/>

R Project for Statistical Computing: <http://www.r-project.org/>

UCSC Genome Browser: <http://genome.ucsc.edu/cgi-bin/hgBlat>

454 GS FLX system: (<http://www.454.com/products-solutions/productlist.asp>)

Results—An analytical pipeline for calling genotypes was developed using HapMap samples sequenced on chr5p15.33. Concordance to 1000 Genomes data for chr5p15.33 was >96%. The concordance for chr1q32.1 and chr13q22.1 with pancreatic cancer GWAS data was >99%. Between 9.2–19.0% of variants detected were not present in 1000 Genomes for the respective continental population. The majority of completely novel SNPs were less common (MAF < 5%) or rare (MAF < 2%), illustrating the value of enlarging test sets for discovery of less common variants. Using the dataset, we examined haplotype blocks across each region using a tag SNP analysis ($r^2 > 0.8$ for MAF < 5%) and determined that at least 196, 243 and 63 SNPs are required for fine-mapping chr1q32.1, chr5p15.33, and chr13q22.1, respectively, in European populations.

Conclusions—We have characterized germline variation in three regions associated with pancreatic cancer risk and show that targeted re-sequencing leads to the discovery of novel variants and improves the completeness of germline sequence variants for fine-mapping GWAS susceptibility loci.

Keywords

pancreatic cancer; targeted re-sequencing; GWAS; susceptibility loci; SNP; 1000G

Introduction

Pancreatic cancer is the fourth leading cause of cancer mortality in the U.S. with a 5 year survival of less than 5% for all stages combined¹. Known risk factors include tobacco smoking, diabetes, obesity, chronic pancreatitis, heavy alcohol consumption and a family history of pancreatic cancer^{2–7}. Genome-wide association studies (GWAS) have been successful in identifying novel genomic regions associated with many complex diseases and traits, including pancreatic cancer⁸. We recently conducted a GWAS to identify common genetic markers of pancreatic cancer susceptibility within twelve prospective epidemiologic cohort studies (Pancreatic Cancer Cohort Consortium) and eight case-control studies (Pancreatic Cancer Case Control Consortium, PanC4) and identified four genomic regions associated with pancreatic cancer risk; chromosomes 1q32.1, 5p15.33, 9q34.2 and 13q22.1^{9,10}. Additional GWAS have identified susceptibility regions on chromosomes 6p25.3, 7q36.2 and 12p11.21 in a Japanese case-control study¹¹ and on chromosomes 5p13.1, 10q26.11, 21q21.3, 21q22.3 and 22q13.32 in a Chinese case-control study¹².

GWAS identify markers within susceptibility regions but do not provide functional explanations for the association signals. Fine-mapping of each region is required to survey the possible variants for further follow-up studies designed to understand the biological basis. In this study, we have focused on three regions, chromosomes 1q32.1, 5p15.33 and 13q22.1 and have generated detailed maps of common single nucleotide polymorphisms (SNPs) based on next generation sequence analysis and public data bases (HapMap and 1000 Genomes Project). In the process we have established an analytical pipeline for genotype calling using the Roche-454¹³ platform for targeted re-sequencing analysis.

Methods

Samples

DNA samples used for re-sequence analysis were drawn from the International HapMap CEPH (CEU), Han Chinese (CHB), Japanese (JPT) and Yoruba African (YRI) populations as well as from two cohort studies included in the PanScan GWAS^{9,10}, the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study (ATBC)¹⁴ and the National Cancer Institute's (NCI) Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial¹⁵. For the region on chromosome 5p15.33, DNA samples from 59 HapMap CEU, 60 HapMap

CHB and JPT and 60 HapMap YRI individuals were analyzed. For chromosome 1q32.1, DNA samples were from 48 pancreatic cancer cases and 47 control subjects from ATBC. For chromosome 13q22.1, DNA samples were from 49 pancreatic cancer cases and 48 controls subjects from PLCO and ATBC. Samples for chr1q32.1 and 13q22.1 were selected to give a uniform distribution of pancreatic cancer susceptibility alleles at these loci.

Regions sequenced

We sequenced the following regions: 140 kb genomic region on chromosome 5p15.33 (1,165,215–1,305,678 bps, NCBI Build 37), 209 kilobase (kb) on chromosome 1q32.1 (199,864,160–200,072,966 bps, National Center for Biotechnology Information (NCBI) Build 37) and 79 kb on chromosome 13q22.1 (73,866,409–73,945,872 bps, NCBI Build 37.3). Linkage disequilibrium (LD) patterns were visualized using Haploview¹⁶.

Primers, sequence capture and sequencing

For chromosome 5p15.33, the RainDance Rainstorm microdroplet PCR technology was used for target-specific capture. Supplemental Table 1A provides a list of primer sets (n=469) designed by RainDance. For chromosome 1q32.1, we used the Nimblegen solution based target capture technology using biotinylated solution capture probe pools (n=240) targeting the region of interest (Supplemental Table 1B) designed by Nimblegen (Roche NimbleGen, Madison, WI)¹⁷. Capture was performed in solution in 0.2 ml PCR strip tubes on a thermal cycler. After ~72 h, the capture probe/sample duplexes were bound to streptavidin magnetic beads. Captured samples were then amplified directly off the beads, emulsified and sequenced according to approved Roche 454 GS FLX protocols (<http://www.454.com/products-solutions/productlist.asp>). For chromosome 13q22.1, sets of long-range PCR primers (n=19) were designed to cover the region targeted using Primer3 (<http://frodo.wi.mit.edu>)¹⁸ as previously described (Parikh PMID: 19823874)¹⁹ (Supplemental Table 1C). Amplicon size ranged from 4,515 to 5,095 bps with an average overlap of 100 bps. Primers were ordered from Integrated DNA Technologies (Coralville, IA). After long-range PCR, all sequencing protocols were followed in accordance with standard protocols for the 454 GS FLX system (<http://www.454.com/products-solutions/productlist.asp>).

Alignment and detection of genetic variation

We developed a computational pipeline to process sequence reads generated by 454 FLX Genome Sequencers using the data from 158 HapMap samples for the 5p15.33 region. Sequence reads were pooled based on barcodes provided by Roche-454. Pre-alignment quality control (QC) was performed using 454-supplied software. Keypass, dot, mixed, signal intensity, primer and trimback valley filters were applied. Sequence reads that passed QC were mapped to the entire genome using Newbler (version 2.3, <http://my454.com/products/analysis-software/index.asp>)²⁰. Post-alignment quality control was done by an in-house pipeline written in Python (<http://python.org/>) and R (<http://www.r-project.org/>), which use the GLU package (<http://code.google.com/p/glu-genetics/>) as the core library. The median depth and sequence coverage over all samples is shown in Supplementary Figure 1 for the target regions (Panels A through E). A cutoff for sample exclusion due to low coverage was set at median coverage < 10x, leading to the exclusion of 21 samples (10 HapMap CEU, 6 CHB and JPT and 5 YRI) sequenced on chromosome 5p15.33, 7 samples (2 cases and 5 control subjects) on chromosome 1q32.1, and 3 samples (1 case and 2 controls) on chromosome 13q22.1.

Base quality score recalibration, variant discovery and genotype calling were performed using the Genome Analysis Toolkit (GATK) with the QUAL filter (threshold 10)²¹. We also performed variant calling using Newbler²⁰ as per standard AllDiffs criteria, and considered only variants identified by both GATK and Newbler. We filtered out SNPs with a

contiguous homopolymer run (HRun) in either direction with a size of >3 bp, SNPs within clusters (3 SNPs within 10 bps of each other) and SNPs close to indels (using the 1000 Genomes indel mask file from the GATK version 1.2 bundle, <ftp://ftp.broadinstitute.org/bundle/1.2/hg19/>). In addition, we removed singletons (alleles called only once) and SNPs with completion rates < 50%. The remaining, potentially novel SNPs, were manually inspected using the Integrative Genomics Viewer (IGV)²² for sequence or alignment artifacts. Supplemental Figure 2 outlines the SNP calling pipeline. For exonic regions, we manually inspected all variants (including singletons) using IGV for sequence or alignment artifacts. Finally, we reinserted all SNPs reported in the 1000 Genomes Project (October 2011 release) and/or dbSNP (in any population), if they were called by GATK, even if removed by the pipeline.

The potential functional effects of SNPs were predicted using SIFT (Sorting Intolerant From Tolerant) with outcomes of T for “tolerated” and D for “deleterious”; and Polyphen-2 (Polymorphism Phenotyping) with prediction outcomes: “neutral” or “deleterious” and probabilities listed^{23,24}. SNP effects on splicing were assessed using NetGene2 (<http://www.cbs.dtu.dk/services/NetGene2/>).

Concordance analysis

For chromosome 5p15.33, genotype data were downloaded from the GATK resource bundle v1.2 (Omni 2.5M genotyping array, <ftp://ftp.broadinstitute.org/bundle/1.2/hg19/>), the 1000 Genomes project (<http://www.1000genomes.org/>; October, 2011) and the International HapMap project (release 28), to evaluate genotype concordance between the re-sequencing data and the two different publicly available data sources. The Omni genotyping array contains genotype data from the 1000 Genomes samples genotyped on the Omni 2.5M chip. Twenty six HapMap CEU, 47 HapMap CHB and JPT and 34 YRI samples included in the 1000 Genomes dataset were used to calculate concordance between re-sequencing data and 1000 Genomes data using the GLU software package (<http://code.google.com/p/glu-genetics/>). Concordance analysis between the re-sequencing data and pancreatic cancer GWAS data^{9,10} was assessed for 88 individuals (pancreatic cancer cases and control subjects) for chromosome 1q32.1 and 94 individuals (cases and controls) for chromosome 13q22.1.

Descriptive statistics

Genotype completion, MAF estimations, deviations from fitness for Hardy-Weinberg proportion (HWP), pair-wise linkage disequilibrium (LD) and tag SNP selection were computed using the GLU software package. Data from samples that passed QC (as described above) were utilized for SNP tagging using the GLU software package. Only SNPs with HWP > 10⁻⁴ were used for tagging. In total, 44 SNPs were removed due to lack of with fitness for Hardy-Weinberg proportion including 12 SNPs on 5p15.33 for HapMap CEU, 10 SNPs on 5p15.33 for HapMap HCB/JPT, 8 SNPs on 5p15.33 for HapMap YRI, 10 SNPs on 1q32.1 for PanScan and 4 SNPs on 13q22.1 for PanScanLD was visualized in Haploview¹⁶.

Results

Variant discovery and quality control assessment

To characterize common and uncommon genetic variants in three genomic regions associated with pancreatic cancer risk, we developed a semi-automated analytical approach (Supplemental Figure 2) for the analysis of targeted next-generation sequencing data generated by the Roche-454 platform. For the development of the approach, we examined 158 HapMap samples sequenced across a 140 kb genomic region on chromosome 5p15.33

(1,165,215–1,305,678 bps, NCBI Build 37.3). The median sequence depth (Supplemental Figure 1) was 56.0, 96.5 and 54.0 fold for the three HapMap sets (HapMap CEU, CHB/JPT and YRI samples). We used two algorithms to enhance confidence in variant calls: we first identified 1,291 variants in our targeted region on chromosome 5p15.33 in the 49 HapMap CEU samples using GATK; of these variants, 1,070 were also identified by Newbler (see corresponding numbers for the CHB/JPT and YRI populations in Table 1). In order to determine the most appropriate filters and thresholds for this set of SNPs, we used HapMap SNPs in our dataset as a set of high quality SNPs. Standard GATK filters and thresholds did not reliably group HapMap SNPs in our dataset apart from filters relating to SNP calling in or close to indels and homopolymer tracts. By applying these filters on the 1,070 SNPs called in the HapMap CEU samples on chromosome 5p15.33, we removed ~20% of SNPs called by both GATK and Newbler but retained all HapMap SNPs. We excluded singletons and SNPs with low completion rates (<50%) and manually inspected the remaining novel SNPs using IGV. This further refined the number of SNPs by excluding variants called due to alignment artifacts. By using this approach we observed a total of 590 SNPs in 49 HapMap CEU samples on chromosome 5p15.33 (Table 1). Of these SNPs, 509 (86.3%) were already listed in the 1000 Genomes database (October 2011 release) for the HapMap CEU population and 528 (89.5%) in dbSNP (build 132) (Table 2). We used the same approach to identify 510 SNPs in 54 HapMap CHB and JPT samples as well as 854 SNPs in 55 HapMap YRI samples on chromosome 5p15.33. A similar proportion of SNPs was listed in 1000 Genomes for the HapMap CHB/JPT and YRI populations (81.0% and 86.5%) and in dbSNP (85.3% and 87.1%). Note that our filters excluded 87 (CEU), 78 (CHB/JPT) and 89 (YRI) SNPs listed by the 1000 Genomes and/or dbSNP that were pulled back into the pool of variants if they were called by GATK, reducing the loss of known SNPs in our pipeline. Most of these SNPs were removed due to the “SNP cluster” filter indicating alignment artifacts that are probably due to falsely called indels.

We applied this approach to the other two regions (Table 1). For chromosome 1q32.1 (209 kb; chr1:199,864,160–200,072,966) we called 791 SNPs based on 88 individuals of European ancestry; for chromosome 13q22.1 (79 kb; chr13:73,866,409–73,945,872), we called 277 SNPs in 94 individuals. The median sequence depth (Supplemental Figure 1) was 20.0 and 48.5 fold for 1q32.1 and 13q22.1, respectively. A high percentage of SNPs observed in these two regions was previously listed in 1000 Genomes (October 2011 release) for the HapMap CEU population (90.8% for 1q32.1 and 85.2% for 13q22.1) or in dbSNP (build 132) (84.6% for 1q32.1 and 88.8% for 13q22.1). Table 2 compares the number of variants identified in the three genomic regions in the current study to HapMap, 1000 Genomes and NCBI’s dbSNP. Most SNPs discovered by our approach were common (MAF ≥ 5%); 80% of SNPs identified for HapMap CEU samples on 5p15.33; 79% for 5p15.33 CHB/JPT; 69% for 5p15.33 YRI; 75% for 1q32.1; and finally 63% for 13q22.1. In total, we identified 65 unique novel SNPs on chr5p15.33 not previously identified by 1000G or dbSNP in any population (21 in the HapMap CEU, 21 in HapMap CHB and JPT, 33 in the HapMap YRI samples). Similarly, we identified 25 novel SNPs on chromosome 1q32.1 and 5 novel SNP on chromosome 13q22.1. Of the 95 novel SNPs identified in all three regions, 26 were common (MAF ≥ 5%) in one or more of the populations.

Concordance analysis

To assess the accuracy of our analytical pipeline, we compared genotype calls from our targeted sequencing to those identified by the 1000 Genomes Project and PanScan separately (Table 3). For chromosome 5p15.33, the comparison was limited to the 26 CEU, 47 CHB/JPT and 34 YRI HapMap samples included in both our project and 1000 Genomes. A concordance rate of 96.2% was seen for HapMap CEU samples, 97.9% for HapMap CHB/JPT samples and 97.1% for HapMap YRI samples. A similar number of SNPs was

identified by both projects but not all of them overlapped: 421, 383 and 667 SNPs were seen in both projects for the HapMap CEU, CHB/JPT and YRI samples, respectively (Supplemental Figure 3). Our targeted re-sequencing identified 87, 114 and 111 SNPs not reported in the 1000 Genomes Project for these samples (median MAF 0.13, 0.13 and 0.09). Furthermore, our approach missed 100, 94 and 136 SNPs seen in 1000 Genomes data for the three populations (median MAF 0.15, 0.15 and 0.10). The majority of SNPs missed by either approach on 5p15.33 were common: 75.9% (HapMap CEU), 70.2% (HapMap CHB/JPT) and 64.0% (HapMap YRI) of SNPs identified by sequencing but missed by 1000 Genomes had MAF $\geq 5\%$. Similarly, 83.0% (HapMap CEU), 76.6% (HapMap CHB/JPT) and 73.5% (HapMap YRI) of SNPs seen in 1000 Genomes but missed by our approach had MAF $\geq 5\%$ (Supplemental Figure 3). The most likely reason for not observing these SNPs by either approach may be low sequence depth; the median coverage at sites missed by re-sequencing was 1.2, 2.0 and 8.0x for the CEU, CHB/JPT and YRI samples.

High concordance rates were also observed between our dataset on 5p15.33 and other publicly available genotype datasets including the HapMap project (94.7%, 98.1% and 98.1%) and 1000 Genomes samples genotyped on the Omni 2.5M genotyping array (94.6%, 100.0% and 97.7%) for overlapping HapMap CEU, CHB/JPT and YRI samples, respectively (Table 3). For chromosomes 1q32.1 and 13q22.1, genotype concordance was compared to our previously reported pancreatic cancer GWAS with a 98.9% concordance rate for 88 samples on chr1q32.1 and 100% concordance rate for 94 samples and on chr13q22.1 (Table 3).

Loci within exonic regions on chromosomes 1q32.1 and 5p15.33

Twenty-three exonic SNPs were observed in the *TERT* gene on chromosome 5p15.33, of which ten resulted in non-synonymous amino acid changes (Supplemental Table 3). Four SNPs were predicted to have damaging effects on TERT protein function using SIFT and six SNPs by PolyPhen-2. Functional prediction using these two programs concurred for only two SNPs, both of whom were novel (at position 1,280,292 in exon 4; and at position 1,253,919 in exon 16). Both SNPs were only seen in the HapMap YRI samples and are singletons. In addition, two SNPs were observed in the 3' UTR and three potential splice sites were affected by SNPs. Four coding SNPs were observed in the *NR5A2* gene on chromosome 1q32.1 (Supplemental Table 3). Three of the four resulted in non-synonymous changes. A damaging effect on protein function was predicted by SIFT for one non-synonymous SNP (at position 200,017,398 in exon 5) but PolyPhen-2 predicted a neutral substitution at this site. No other deleterious changes were predicted for *NR5A2*. Moreover, one SNP located within the 2nd intron of *NR5A2* is predicted to create an alternative splicing site.

No coding SNPs were missed in the *NR5A2* gene by our data, as compared to 1000 Genomes CEU samples, and four potential new ones were identified. For *TERT*, two coding SNPs were missed (rs33956095, MAF=0.015 and rs33959226, MAF=0.002) and two new SNPs identified. All novel SNPs identified in the two genes were rare (MAF < 2%) and thus require additional validation.

Linkage disequilibrium (LD) and tag SNP selection

The maps of LD for common variants (MAF $\geq 5\%$) based on our sequence data in the three genomic regions are shown in Figure 1 for populations of European ancestry and in Supplemental Figure 4 for HapMap CHB/JPT and YRI populations. The pattern of LD for chromosome 5p15.33 shows that there has been extensive recombination across distinct populations resulting in a region with little LD. Chromosome 1q32.1 has an overall stronger

pattern of LD with more defined haplotype blocks. The region on chromosome 13q22.1 contained evidence of low LD, with a paucity of well-defined blocks.

Tagging analysis of SNPs in the 140 kb region on 5p15.33 in HapMap CEU samples at an r^2 threshold of 0.8 yielded at least 243 tag SNPs that are required to interrogate the 460 SNPs with a MAF $\geq 5\%$ (Table 4). At the same threshold, 207 tag SNPs are needed to cover the 393 SNPs identified in the CHB/JPT samples and 369 tag SNPs are needed to cover the 578 SNPs in the YRI samples for 5p15.33. For chromosome 1q32.1, 196 tag SNPs are required to cover the 586 SNPs; and 63 SNPs were needed to cover the 171 SNPs identified on chromosome 13q22.1. The former region was 209 kb in length and the latter 79 kb. Supplemental Table 4 lists bin and tag SNPs information using the same threshold.

Discussion

In this study, we have characterized common genetic variants for three genomic regions associated with pancreatic cancer risk in European populations. We sequenced 209 kb on chromosome 1q32.1, 140 kb on chromosome 5p15.33 and 79 kb on chromosome 13q22.1 and developed a systematic approach to call SNPs from sequence data generated by targeted Roche-454 next generation technology. Chromosome 5p15.33 is a bona fide multi-cancer risk locus noted in GWAS of pancreatic cancer, lung cancer, bladder cancer, prostate cancer, breast cancer, glioma, testicular cancer, basal cell carcinoma and melanoma^{10,25–35}. The locus harbors the *TERT* gene, encoding the catalytic subunit of telomerase frequently reactivated in human cancers³⁶. It also contains a less known gene, *CLPTMIL*, with a possible role in apoptosis³⁷. The region of chr1q32.1 harbors a plausible candidate gene, *NR5A2*, whereas chr13q22.1 is an intergenic region with less obvious candidate genes. It is notable that the association with pancreatic cancer has been confirmed for the latter in populations of European, Chinese and Japanese ancestry^{10–12}.

To reduce uncertainty associated with genotype calling from sequence data, we used two distinct genotype-calling algorithms and selected quality filters that retained HapMap SNPs while reducing the total set of called SNPs considerably. In our analysis of three regions, each captured by a distinct technology, we observed that the most effective filters removed SNPs in regions of potential alignment artifacts, in part due to false SNP calling around indels and in homopolymer runs, known stumbling blocks for analysis with 454 data. Still, a substantial fraction of novel SNPs had to be excluded by manual inspection of SNPs suggesting this genotype-calling approach for Roche 454 sequence data is achievable for targeted genomic regions but not practical for exome sequencing or whole genome sequencing.

When comparing the set of samples common to the current study and the 1000 Genomes project, we observed a fraction of SNPs missed by either approach. These tended to locate to regions of low coverage, as expected. This indicates that additional re-sequencing is valuable to improve the completeness of publicly available genotype datasets. However, the majority of SNPs missed by either approach were already listed in public databases for other sample sets or populations, and the number of completely novel SNPs was small and mostly represented less common (MAF $\geq 5\%$) or rare (MAF $\geq 2\%$) SNPs. Additional high throughput sequencing in large sample sets at high depth is needed to discover more of the less common and rare variants reliably.

Our dataset, in concert with publicly available data sources, provides a comprehensive catalog of common genetic variation in three genomic regions associated with pancreatic cancer risk. It permits the selection of common tag SNP to thoroughly examine the association of genetic polymorphisms to pancreatic cancer risk in follow-up association

studies, with the aim of identifying causal variants and, ultimately, explain the underlying biology behind these three pancreatic cancer risk loci.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported in part by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health (NIH) under Contract No. HHSN261200800001E. We acknowledge and thank the study participants for donating their time and biospecimens and thereby making this study possible. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. Government.

References

1. Siegel R, Naishadham D, Jemal A. Cancer statistics, 2012. *CA Cancer J Clin.* Jan; 2012 62(1):10–29. [PubMed: 22237781]
2. Lynch SM, Vrieling A, Lubin JH, et al. Cigarette smoking and pancreatic cancer: a pooled analysis from the pancreatic cancer cohort consortium. *Am J Epidemiol.* Aug 15; 2009 170(4):403–413. [PubMed: 19561064]
3. Huxley R, Ansary-Moghaddam A, Berrington de Gonzalez A, et al. Type-II diabetes and pancreatic cancer: a meta-analysis of 36 studies. *Br J Cancer.* Jun 6; 2005 92(11):2076–2083. [PubMed: 15886696]
4. Arslan AA, Helzlsouer KJ, Kooperberg C, et al. Anthropometric measures, body mass index, and pancreatic cancer: a pooled analysis from the Pancreatic Cancer Cohort Consortium (PanScan). *Archives of internal medicine.* May 10; 2010 170(9):791–802. [PubMed: 20458087]
5. Raimondi S, Lowenfels AB, Morselli-Labate AM, et al. Pancreatic cancer in chronic pancreatitis; aetiology, incidence, and early detection. *Best Pract Res Clin Gastroenterol.* Jun; 2010 24(3):349–358. [PubMed: 20510834]
6. Michaud DS, Vrieling A, Jiao L, et al. Alcohol intake and pancreatic cancer: a pooled analysis from the pancreatic cancer cohort consortium (PanScan). *Cancer causes & control : CCC.* Aug; 2010 21(8):1213–1225.
7. Jacobs EJ, Chanock SJ, Fuchs CS, et al. Family history of cancer and risk of pancreatic cancer: a pooled analysis from the Pancreatic Cancer Cohort Consortium (PanScan). *Int J Cancer.* Sep 1; 2010 127(6):1421–1428. [PubMed: 20049842]
8. Hindorff, LA.; MacArthur, J. Institute EB. A Catalog of Published Genome-Wide Association Studies. 2012. Available at: www.genome.gov/gwastudies
9. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet.* Sep; 2009 41(9):986–990. [PubMed: 19648918]
10. Petersen GM, Amundadottir L, Fuchs CS, et al. A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat Genet.* Mar; 2010 42(3):224–228. [PubMed: 20101243]
11. Low SK, Kuchiba A, Zembutsu H, et al. Genome-wide association study of pancreatic cancer in Japanese population. *PloS one.* 2010; 5(7):e11824. [PubMed: 20686608]
12. Wu C, Miao X, Huang L, et al. Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. *Nat Genet.* 2011; 44(1):62–66. [PubMed: 22158540]
13. Rothberg JM, Leamon JH. The development and impact of 454 sequencing. *Nat Biotechnol.* Oct; 2008 26(10):1117–1124. [PubMed: 18846085]

14. The ATBC Cancer Prevention Study Group. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods participant characteristics, and compliance. *Ann Epidemiol.* Jan; 1994 4(1):1–10. [PubMed: 8205268]
15. Gohagan JK, Prorok PC, Hayes RB, et al. The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial of the National Cancer Institute: history, organization, and status. *Controlled clinical trials.* Dec; 2000 21(6 Suppl):251S–272S. [PubMed: 11189683]
16. Barrett JC, Fry B, Maller J, et al. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics.* Jan 15; 2005 21(2):263–265. [PubMed: 15297300]
17. Albert TJ, Molla MN, Muzny DM, et al. Direct selection of human genomic loci by microarray hybridization. *Nat Methods.* Nov; 2007 4(11):903–905. [PubMed: 17934467]
18. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol.* 2000; 132:365–386. [PubMed: 10547847]
19. Parikh H, Deng Z, Yeager M, et al. A comprehensive resequence analysis of the KLK15-KLK3-KLK2 locus on chromosome 19q13.33. *Human genetics.* Jan; 2010 127(1):91–99. [PubMed: 19823874]
20. Droege M, Hill B. The Genome Sequencer FLX System--longer reads, more applications, straight forward bioinformatics and more complete data sets. *Journal of biotechnology.* Aug 31; 2008 136(1–2):3–10. [PubMed: 18616967]
21. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet.* May; 2011 43(5):491–498. [PubMed: 21478889]
22. Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol.* Jan; 2011 29(1):24–26. [PubMed: 21221095]
23. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols.* 2009; 4(7):1073–1081.
24. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res.* May; 2001 11(5):863–874. [PubMed: 11337480]
25. Kote-Jarai Z, Olama AA, Giles GG, et al. Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. *Nat Genet.* Aug; 2011 43(8):785–791. [PubMed: 21743467]
26. Rafnar T, Sulem P, Stacey SN, et al. Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nat Genet.* Feb; 2009 41(2):221–227. [PubMed: 19151717]
27. Shete S, Hosking FJ, Robertson LB, et al. Genome-wide association study identifies five susceptibility loci for glioma. *Nat Genet.* Aug; 2009 41(8):899–904. [PubMed: 19578367]
28. Stacey SN, Sulem P, Masson G, et al. New common variants affecting susceptibility to basal cell carcinoma. *Nat Genet.* Aug; 2009 41(8):909–914. [PubMed: 19578363]
29. Rothman N, Garcia-Closas M, Chatterjee N, et al. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nat Genet.* Nov; 2010 42(11):978–984. [PubMed: 20972438]
30. Turnbull C, Rapley EA, Seal S, et al. Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. *Nat Genet.* Jul; 2010 42(7):604–607. [PubMed: 20543847]
31. Landi MT, Chatterjee N, Yu K, et al. A genome-wide association study of lung cancer identifies a region of chromosome 5p15 associated with risk for adenocarcinoma. *Am J Hum Genet.* Nov; 2009 85(5):679–691. [PubMed: 19836008]
32. McKay JD, Hung RJ, Gaborieau V, et al. Lung cancer susceptibility locus at 5p15.33. *Nat Genet.* Dec; 2008 40(12):1404–1406. [PubMed: 18978790]
33. Wang Y, Broderick P, Webb E, et al. Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nat Genet.* Dec; 2008 40(12):1407–1409. [PubMed: 18978787]
34. Truong T, Hung RJ, Amos CI, et al. Replication of lung cancer susceptibility loci at chromosomes 15q25, 5p15, and 6p21: a pooled analysis from the International Lung Cancer Consortium. *J Natl Cancer Inst.* Jul 7; 2010 102(13):959–971. [PubMed: 20548021]
35. Haiman CA, Chen GK, Vachon CM, et al. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat Genet.* Dec; 2011 43(12):1210–1214. [PubMed: 22037553]

36. Kim NW, Piatyszek MA, Prowse KR, et al. Specific association of human telomerase activity with immortal cells and cancer. *Science*. Dec 23; 1994 266(5193):2011–2015. [PubMed: 7605428]
37. Yamamoto K, Okamoto A, Isonishi S, et al. A novel gene, CRR9, which was up-regulated in CDDP-resistant ovarian tumor cell line, was associated with apoptosis. *Biochem Biophys Res Commun*. Feb 2; 2001 280(4):1148–1154. [PubMed: 11162647]

Figure 1A **Figure 1B**

Figure 1C

Figure 1. Linkage disequilibrium (LD) plot in HapMap CEU samples for SNPs with MAF $\geq$ 5% as measured by r^2 across a 140 kb region of chromosome 5p15.33, 209 kb for 1q32.1 and 79 kb for 13q22.1

Coordinates are based on NCBI genome build 37.3. SNPs with MAF $\geq$ 5% and completion rates $\geq$ 50% are included. LD plots for the HapMap CHB & JPT as well as YRI samples on chromosome 5p15.33 are shown in Supplemental Figures 4A and 4B, respectively.

Germline variants observed by targeted Roche 454 re-sequencing on chromosomes 5p15.33, 1q32.1 and 13q22.1

Table 1

Population	HapMap	HapMap	HapMap	PanScan	PanScan
Ethnicity	CEU	YRI	CEU	CEU	CEU
Locus	Chr5	Chr5	Chr1	Chr13	Chr13
SNPs called by GATK	1,291	1,350	1,783	1,883	499
SNPs also called by Newbler	1,070	1,130	1,517	1,778	483
SNPs removed if:					
Singletons	209	219	311	889	194
HRun > 3	5	6	4	4	3
SNPcluster	171	235	219	24	13
Presence of indels	3	3	5	4	2
Completion rate < 50%	32	27	34	15	0
Manual inspection in IGV	152	212	189	85	8
SNPs pulled back in if:					
SNPs in 1000 Genomes and/or dbSNP and called by GATK	87	78	89	30	14
Additional exonic SNPs	5	4	10	4	-
Total SNPs identified	590	510	854	791	277

Filtering criteria used for genotype calling, number of SNPs called and excluded and the final numbers of SNPs identified in each region are listed. Singletons, SNPs seen only once; HRun, homopolymer run, SNP clusters, 3 or more SNPs within 10 bps; presence of indels, SNPs in or close to indels (using GATK indel maske file). IGV: Integrated Genome Viewer. See methods for details on filtering and genotype calling.

Table 2
Comparison of SNPs observed on chromosomes 5p15.33, 1q32.1 and 13q22.1 to those found in public databases

Population	HapMap	HapMap	HapMap	PanScan	PanScan
Ethnicity	CEU	CHB	YRI	CEU	CEU
Locus	Chr5	Chr5	Chr5	Chr1	Chr13
Total SNPs identified	590	510	854	791	277
Overlap with known databases					
HapMap CEU/CHB/JPT/YRI populations	91	82	88	178	116
1000 Genomes CEU population	509			718	236
1000 Genomes CHB/JPT population		413			
1000 Genomes YRI population			739		
1000 Genomes all populations	520	435	761	751	269
dbSNP (with frequency)	474	387	668	619	234
dbSNP (total)	528	435	744	669	246
# of novel SNPs	21	21	33	25	5

The databases include HapMap (Release 28), 1000 Genomes (October 2011 release) and NCBI's dbSNP (build 132). The number of HapMap samples used for comparison was as follows: HapMap CEU: n=49 (this project) compared to n=85 in 1000 Genomes; CHB/JPT: n=54 (this project) compared to n=186 CHB and JPT in 1000 Genomes; YRI: n=55 (this project) compared to n=88 YRI in 1000 Genomes; 1000 Genomes all populations: n=1092, HapMap CEU/CHB/JPT/YRI: n=638.

Table 3

Genotype concordance rates for SNPs on chromosomes 5p15.33, 1q32.1 and 13q22.1 with publicly available datasets

Population	HapMap	HapMap	HapMap	PanScan	PanScan
Ethnicity	CEU	CHB	YRI	CEU	CEU
Locus	Chr5	Chr5	Chr5	Chr1	Chr13
The 1000 Genomes					
# of samples	26	47	34	-	-
# of SNPs	485	403	727	-	-
Concordance rate	0.962 (0.095–1)	0.979 (0.116–1)	0.971 (0.176–1)	-	-
Omni chip					
# of samples	37	52	44	-	-
# of SNPs	151	131	168	-	-
Concordance rate	0.946 (0.833–1)	1 (0.940–1)	0.977 (0.795–1)	-	-
HapMap data					
# of samples	49	54	55	-	-
# of SNPs	96	86	93	-	-
Concordance rate	0.947 (0.048–1)	0.981 (0.038–1)	0.981 (0.592–1)	-	-
PanScan GWAS					
# of samples	-	-	-	88	94
# of SNPs	-	-	-	42	26
Concordance rate	-	-	-	0.989 (0.784–1)	1 (0.915–1)

Genotype concordance rates between the current dataset and the 1000 Genomes (October 2011 release), GATK Omni 2.5M chip, HapMap (Release 28) and PanScan GWAS datasets for SNPs observed on chromosomes 5p.15.33, 1q32.1 and 13q22.1.

Table 4

Bin content and tag SNPs necessary to cover the three regions sequenced on chromosomes 5p15.33, 1q32.1 and 13q22.1

Population	HapMap	HapMap	HapMap	HapMap	PanScan	PanScan
Ethnicity	CEU	CHB & JPT	YRI	CEU	CEU	CEU
Locus	Chr5	Chr5	Chr5	Chr1	Chr1	Chr13
MAF 2% and r ² 0.8						
Bins monitored	314	266	485	245	83	
SNPs monitored	573	465	750	673	214	
MAF 5% and r ² 0.8						
Bins monitored	243	207	369	196	63	
SNPs monitored	460	393	578	586	171	

The number of SNPs required to cover the three regions at an r² threshold of 0.8 and MAF of either 2% or 5% is indicated. SNPs single nucleotide polymorphisms, MAF minor allele frequency.