Published in final edited form as:
Nat Genet. ; 43(8): 801–805. doi:10.1038/ng.871.

Comparing strategies to fine map the association of common SNPs on chromosome 9p21 to Type 2 Diabetes and Myocardial Infarction

Jessica Shea[1,2,3], **Vineeta Agarwala**[1,3,4,5], **Anthony A. Philippakis**[1,3,4,5,6,7], **Jared Maguire**[1], **Eric Banks**[1], **Mark DePristo**[1], **Brian Thomson**[1], **Candace Guiducci**[1], **The Myocardial Infarction Genetics Consortium**, **Sekar Kathiresan**[1,6,8,9,10], **Stacey Gabriel**[1], **Noël P Burtt**[1], **Mark J. Daly**[1,6,8,10], **Leif Groop**[11], and **David Altshuler**[1,3,6,9,10,12]

[1]Program in Medical and Population Genetics, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

[2]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts, USA

[3]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA

[4]Program in Biophysics, Harvard University, Cambridge, Massachusetts, USA

[5]Harvard-Massachusetts Institute of Technology Division of Health Sciences and Technology, Harvard Medical School, Boston, Massachusetts, USA

[6]Department of Medicine, Harvard Medical School, Boston, Massachusetts, USA

[7]Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA

[8]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA

[9]Cardiovascular Research Center, Massachusetts General Hospital, Massachusetts, USA

[10]Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA

[11]Department of Clinical Sciences, Diabetes and Endocrinology Research Unit, University Hospital Malmö, Lund University, Malmö, Sweden

[12]Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts, USA

Abstract

Correspondence should be addressed to David Altshuler (altshuler@molbio.mgh.harvard.edu)..

Author Contributions Manuscript writing: J.S., V.A., A. A. P., D.A.
Clinical samples: S.K., L.G., D.A., The Myocardial Infarction Genetics Consortium
Next-generation sequencing data generation: C.G., N.P.B., S.G., Broad Institute Sequencing Platform
Sequencing analysis and variant calling: A.A.P, J.M., E.B., M.D., S.G., M.J.D., D.A.
Imputation and association analysis: J.S., V.A., M.J.D., D.A.
Genotyping and analysis: J.S., V.A, B.T., C.G., N.P.B., Broad Institute Genetic Analysis Platform

Competing Interests Statement No competing interests identified.

Non-coding variants at human chromosome 9p21 near *CDKN2A* and *CDKN2B* are associated with type 2 diabetes (T2D)[1-4], myocardial infarction (MI)[5-7], aneurysm[8], vertical cup disc ratio[9], and at least five cancers[10-16]. We compared approaches to more comprehensively assess genetic variation in the region. We performed targeted sequencing at high coverage in 47 individuals and compared the results to pilot data from the 1000 Genomes Project. We imputed variants into T2D and MI cohorts directly from targeted sequencing, from a genotyped reference panel derived from sequencing, and from 1000 Genomes low-coverage data. Common polymorphisms were captured similarly by all strategies. Imputation of intermediate frequency polymorphisms required a higher density of tag SNPs in disease samples than available on first generation Genome Wide Association Study (GWAS) arrays. Association analyses identified more comprehensive sets of variants demonstrating equivalent statistical association to T2D or MI, but did not identify stronger associations the original GWAS signals.

Following the identification of a disease-associated region by GWAS, comprehensive study of sequence variation in the region is required to identify the full set of variants that might explain the association signal. Since GWAS arrays incompletely capture DNA variation in each region, it has been hypothesized that causal variants partially captured by linkage disequilibrium (LD) – due to location near recombination hotspots or lower minor allele frequency – might, if directly tested, display stronger association to phenotype than the tag SNPs used in GWAS. In particular, because HapMap and GWAS arrays contain primarily variants with minor allele frequency (MAF) >5%, first generation GWAS studies failed to test polymorphisms of somewhat lower frequency that might have larger effects on disease risk. Finally, even in regions where the true association signal is well captured by LD to array SNPs, enumeration of all associated variants is a necessary prerequisite to functional experiments that will identify causal mutation(s). Thus, an important next step following GWAS is to assemble a more complete catalog of variation present in an associated region, and to test it for association to the phenotype of interest.

With the advent of next generation sequencing and the emergence of data from the 1000 Genomes (1000G) Project, investigators must choose between (or combine) multiple strategies for creating and testing a reference panel of polymorphic sites. We re-sequenced ~240kb on chromosome 9p21 (chr9:21936711-22176221, hg18) spanning the T2D and MI associations in 47 unrelated individuals of European ancestry from the HapMap CEU population[17] as part of a sequencing project spanning six T2D-associated regions (Supplementary Table 1). Sequencing was performed at the Broad Institute on Illumina Genome Analyzers (Supplementary Note, all data available in the NCBI Short Read Archive). An analytical framework (Supplementary Note, Supplementary Table 2, Supplementary Figs. 1-5), since extended and incorporated in the Genome Analysis Tool Kit[18,19], was developed and includes methods to empirically recalibrate Illumina base quality scores, a Bayesian framework to call SNPs, local re-alignment to identify insertions/deletions (and remove clusters of false positive SNPs), and filters to remove false positive SNP calls based on discrepancy between forward and reverse strands.

This targeted sequencing identified 635 high-confidence SNPs on chromosome 9p21 (4,463 across the six regions) (Supplementary Table 3, Supplementary Fig. 6, SNPs available in

dbSNP). We evaluated sensitivity against HapMap II[17] and the high coverage Pilot 2 data from the 1000 Genomes Project[20] (Supplementary Note): at sites in overlapping samples with 10x or greater read coverage (70% of the region), sensitivity was 99% for HapMap variants and 97% for variants found in 1000G Pilot 2 (Supplementary Fig. 7a-c). To evaluate specificity, we genotyped 257 sites found on chromosome 9p21 but not previously genotyped in HapMap (Supplementary Fig. 7d, e). Overall, 96% of variants seen more than once in sequencing validated in the genotyping data (Supplementary Table 4).

We compared these variants to those discovered in the low-coverage Pilot 1 of the 1000 Genomes Project [20], limiting comparison to 32 CEU individuals studied in both projects. Across the six regions, both projects identified similar numbers of variants: 3,897 SNPs in the high coverage targeted sequencing as compared to 4,043 in 1000G Pilot 1. However, the variants found were in fact only partially overlapping. Of variants seen in the high coverage targeted sequencing, 22% were missed by 1000G Pilot 1 (Fig. 1), nearly all of which were rare: 72% of these sites were singletons and 12% were seen twice (Fig. 1, Table 1). (Pilot 1 successfully identified 97% of SNPs seen more than 5 times in high coverage sequencing (Table 1)). Of variants identified in Pilot 1 but not in targeted sequencing (n=998), nearly all were sites at which target capture failed to achieve high coverage: 65% of these sites had zero coverage. Thus, targeted capture and low-pass whole genome had distinct and non-overlapping failure modes.

We evaluated methods for testing these variants for association to disease via linkage disequilibrium and haplotype-based imputation. First, we genotyped SNPs found in targeted re-sequencing on chromosome 9p21 in 168 individuals (56 parent offspring trios) from the HapMap extended CEU population[21] (Supplementary Note). We used MACH[22,23] to impute variants from this reference panel into 1,000 T2D patients and 1,048 controls from the Diabetes Genetics Initiative (DGI) cohort[1] and 1,274 MI cases and 1,407 controls from the Myocardial Infarction Genetics (MIGen) Consortium cohort[6], each previously genotyped on Affymetrix GWAS arrays (Supplementary Note).

We compared the results of imputation with this augmented reference panel (n=464 variants, Supplementary Table 5) to those obtained when imputing from HapMap II alone (n=238 variants). The addition of genotype data for a more complete collection of common variants provided imputation data for a much larger number of SNPs than was possible with HapMap II, which contains only 50-60% of common variants (Fig. 2a, b and Supplementary Fig. 8a, b). However, even with the augmented reference panel, the tag SNP density characteristic of the first generation GWAS arrays on which our disease samples were typed allowed only 80% of common (MAF > 5%) variants to be captured (either directly typed or imputed with a MACH-estimated $r^2 \geq 0.8$). Moreover, only a small fraction of intermediate frequency variation (MAF 2-5%) was imputed with an estimated r^2 above this stringent threshold (Fig. 2c, d and Supplementary Fig. 8c, d).

To evaluate the impact of tag SNP density on imputation performance, we increased the number of tags across the region to approximately 1 SNP per 1.5kb (the previous density was ~1SNP/5kb in T2D samples and ~1SNP/3kb in MI samples) in the T2D and MI cohorts (Supplementary Note). With this increased density of tag SNPs, nearly all common variants

(~98%) were captured with $r^2 \geq 0.8$ in disease samples. Moreover, performance for intermediate frequency variants was dramatically improved, rising from 2% to 75% with $r^2 \geq 0.8$ (Fig. 2e, f and Supplementary Fig. 8e, f). This result was not specific to the Affymetrix GWAS arrays, as we observed a similar improvement in imputation ability upon addition of tag SNPs using multiple other GWAS arrays (Supplementary Fig. 9).

We next compared different reference panels, imputing in each case into disease samples with the higher tag SNP density. The reference panels were: (a) the genotyped reference panel of 168 individuals above (112 unrelated individuals), (b) the targeted sequencing data (47 individuals, without genotyping and expansion into a larger sample set), and (c) 1000 Genomes Pilot 1 (55 individuals). We considered both the fraction of variants in each reference panel successfully imputed (which is related to the quality and completeness of SNP genotypes and to the size of the reference panel) and the fraction of all variation captured (which, in addition, depends on the proportion of all known SNPs represented in the reference panel).

The union of the three reference panels contained 582 variants (Fig. 3a). Each panel was partially incomplete, due to genotyping assay failure in the genotyped panel (14% of SNPs missing), sample size and low coverage in 1000 genomes (16% of SNPs missing), and sample size and gaps in coverage in the targeted sequencing (19% of SNPs missing). For common variants, there is little difference in bulk performance between the reference panels. Considering only SNPs contained in each reference panel (Fig. 3b) the genotyped panel has the highest proportion of variants imputed well. However, when all variation is considered (Fig. 3c), a *lower* proportion of common variation is captured by imputing from the genotyped reference panel, owing to the fact that some SNPs were missing in this panel because they failed assay design or quality control. Notably, the 1000G (freely available) and sequencing (costly) strategies performed equivalently for these common variants.

For intermediate frequency variants, there are more pronounced differences between the panels (Fig. 3b, c). These variants were best imputed from the genotyped reference panel (Fig. 3b), which was the largest and also contained trio information. This was true even when all variation was considered (Fig. 3c), suggesting that the improved imputation quality from genotype data and increased sample size offset the loss of variants in this panel due to genotyping failure. Comparing the high coverage re-sequencing and 1000G reference panels, lower frequency variants were better imputed from the high coverage re-sequencing data both when considering only the SNPs within each reference panel (Fig. 3b) and when considering the overall proportion of low frequency variants captured by imputation from each reference panel (Fig. 3c). This is consistent with the low coverage 1000G pilot 1 data being less complete and accurate for lower frequency variants[20].

We tested variants for association to T2D and MI using imputation from all three reference panels to maximize the number of variants captured (Supplementary Note). Overall, we have captured 461 of the 582 polymorphic variants observed across all three reference panels in our T2D and MI samples with a MACH-estimated r^2 of at least 0.8: this represents ~ 92% of all known common variants and ~52% of intermediate frequency variants (at a MACH-estimated r^2 of 0.5, these figures are 98% of common variants and 83% of

intermediate frequency variants). In comparison, only 176 of the 582 variants were previously captured by imputation from HapMap. Despite testing many additional SNPs in partial LD with the index GWAS hits and at allele frequencies not well captured by first generation GWAS arrays and HapMap, we found no example of a SNP with stronger association to T2D or MI than the initial GWAS signals.

However, we did identify multiple additional variants in strong LD with the GWAS hits that might underlie each association. We observed the three-tiered haplotypic association to T2D first reported by the Wellcome Trust Case Control Consortium with protective, risk, and neutral haplotypes (Table 2). The protective alleles of the GWAS SNP (rs10811661) and nine other SNPs in strong LD with this variant tag the protective haplotype (Fig. 4a, Supplementary Table 6). Interestingly, no single SNP yet identified marks the risk haplotype. Association analyses for MI identified 7 SNPs in LD with each other and with equivalent evidence for association ($P < 10^{-4}$) as well as 54 additional SNPs with only slightly less evidence for association ($P < 10^{-3}$) (Fig. 4b, Supplementary Table 6). Knockout of the MI-associated region in mouse alters regulation of *CDKN2A* and *CDKN2B*[24], and two of the associated SNPs have recently been shown to disrupt a STAT1 binding site[25]. Interestingly, in addition to the SNPs disrupting the STAT1 site, there are other variants with equivalent MI association and with putative functional annotations, including variants overlapping exons of the non-coding transcript *CDKN2BAS*, highly conserved regions, and predicted, conserved transcription factor binding sites (Supplementary Table 6).

This study is limited by the investigation of a single region (albeit one with at least eight different disease associations), by the early nature of the sequencing data analyzed, by the small number of samples sequenced in SNP discovery, and by the sample size of our disease cohorts. Nonetheless, the observations on the strengths and weaknesses of different methods for fine mapping GWAS signals are likely general: targeted high coverage sequencing provides high sensitivity for lower frequency variants, but has gaps in coverage; the 1000G Pilot 1 resource offers more even coverage at lower depth, currently sufficient for capture of most common variation; creating a genotyped reference panel improves accuracy and sample size, but is limited by assay conversion failures; tag SNP density characteristic of first generation GWAS is inadequate to maximally extract information with current imputation algorithms. To some extent, these limitations are transient: the growing 1000 Genomes Project resource is sequencing over 2,000 diverse samples with both low-pass whole genome and high coverage targeted exon approaches, increasing the accuracy and completeness of the public reference panel. However, our results suggest that fully exploiting this resource for imputation may require increasing tag density in GWAS disease samples and / or improved algorithms for imputation.

Finally, our study did not find evidence for stronger association at 9p21 to SNPs in moderate LD with the initial tags. While the maximum achievable association signal for lower frequency variants was limited by our sample size, we did not observe lower frequency variants with effect sizes that could individually explain the common variant associations. We do, however, identify additional common variants in LD with the GWAS hits that might underlie each association. Enumeration of all variants on 9p21 that might explain each association signal will be needed as a foundation for systematic functional studies that aim

to understand how different non-coding variants in this single genomic interval can lead to such varied and clinically significant phenotypic associations.

Methods

Targeted Re-Sequencing

Six regions associated with T2D were selected for targeted re-sequencing (Supplementary Table 1). Because the goal of this study to was to identify additional SNPs that might explain the initial GWAS signal, region boundaries were selected to encompass all SNPs showing detectable linkage disequilibrium ($r^2 \geq 0.2$) to the T2D associated SNP with the most significant p-value. DNA was captured for sequencing by long-range PCR with 2-5kb amplicons or by hybrid selection (HS) using 170bp baits tiled across the region on an Agilent microarray[26]. All sequencing was performed at the Broad Institute in 2008 using Illumina Genome Analyzers. Runs from PCR-based capture generated 36bp reads and runs from HS-based capture generated 46-50bp reads. Methods for alignment, quality score adaptation and recalibration, and variant calling are described in detail in the Supplementary Note.

SNP Genotyping and Quality Control

Genotyping was performed on the Sequenom MassARRAY iPLEX platform. Quality control filters included 1) > 95% genotyping rate, 2) Hardy Weinberg equilibrium (with $P > 0.001$) and 3) Mendel error rate < 5%.

Phasing and Imputation

We compared several strategies and publicly available tools for phasing and imputation directly from Illumina sequencing data (Supplementary Fig. 10-11). Phased haplotypes for all reference panels were created using the PHASE software package (Version 2.1)[27,28]. For the genotyped reference panel, trio information was used in phasing (-P1 option). For sequencing reference panels, known phase was specified at HapMap sites (-k option). All other PHASE parameters were default values. Imputation from reference haplotypes was performed using MACH[22,23] (Version 1.0.16). 100 rounds were used; all other MACH parameters were default values.

Association Analyses

Variants were tested for association using logistic regression on imputed genotype dosages and individual disease status. EIGENSTRAT[29] (DGI) or PLINK[30] (MIGen) was used to estimate principal components which track with the ancestry of the study samples[1,6]; the first ten components were used as covariates in logistic regression to account for population structure. For T2D analyses, additional covariates used were: age, gender, and body mass index. For MI analyses additional covariates used were age, gender, BMI, and smoking. Tests for haplotypic association to T2D were performed using the PLINK[30] (Version 1.05) software package.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Patient collections in the DGI study were funded by grants from the Sigrid Juselius and Folkhälsan foundations as well as the Swedish Research Council (LG). The DGI GWAS study was supported by a grant from Novartis.

The MIGen study was funded by the US National Institutes of Health (NIH) and National Heart, Lung, and Blood Institute's STAMPEED genomics research program through a grant to D.A (R01 HL087676). S.K. is supported by a Doris Duke Charitable Foundation Clinical Scientist Development Award, a charitable gift from the Fannie E. Rippel Foundation, the Donovan Family Foundation, a career development award from the NIH and the Department of Medicine and Cardiovascular Research Center at Massachusetts General Hospital. DA and JS are supported in part by a Distinguished Clinical Scholar Award from the Doris Duke Charitable Foundation (to D.A.)

Next-generation sequencing for this work was performed by the Broad Institute Sequencing Platform and genotyping was performed by the Broad Institute Genetic Analysis Platform. We acknowledge their excellence and collaboration on this study. Sequencing was supported in part by a grant from NHGRI and by the Broad Institute.

The authors thank Manny Rivas, Andrey Sivachenco, and Kiran Garimella for helpful discussions on sequencing and Benjamin Voight, Stephan Ripke, and Ron Do for helpful discussions on imputation.

References

1. Saxena R, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science. 2007; 316:1331–1336. [PubMed: 17463246]
2. Scott LJ, et al. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science. 2007; 316:1341–1345. [PubMed: 17463248]
3. Zeggini E, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nat Genet. 2008; 40:638–645. [PubMed: 18372903]
4. Zeggini E, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science. 2007; 316:1336–1341. [PubMed: 17463249]
5. Helgadottir A, et al. A common variant on chromosome 9p21 affects the risk of myocardial infarction. Science. 2007; 316:1491–1493. [PubMed: 17478679]
6. Kathiresan S, et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet. 2009; 41:334–341. [PubMed: 19198609]
7. McPherson R, et al. A common allele on chromosome 9 associated with coronary heart disease. Science. 2007; 316:1488–1491. [PubMed: 17478681]
8. Helgadottir A, et al. The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. Nat Genet. 2008; 40:217–224. [PubMed: 18176561]
9. Ramdas WD, et al. A genome-wide association study of optic disc parameters. PLoS Genet. 2010; 6:e1000978. [PubMed: 20548946]
10. Bishop DT, et al. Genome-wide association study identifies three loci associated with melanoma risk. Nat Genet. 2009; 41:920–925. [PubMed: 19578364]
11. Falchi M, et al. Genome-wide association study identifies variants at 9p21 and 22q13 associated with development of cutaneous nevi. Nat Genet. 2009; 41:915–919. [PubMed: 19578365]
12. Sherborne AL, et al. Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. Nat Genet. 42:492–494. [PubMed: 20453839]
13. Shete S, et al. Genome-wide association study identifies five susceptibility loci for glioma. Nat Genet. 2009; 41:899–904. [PubMed: 19578367]
14. Stacey SN, et al. New common variants affecting susceptibility to basal cell carcinoma. Nat Genet. 2009; 41:909–914. [PubMed: 19578363]

15. Turnbull C, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet. 42:504–507. [PubMed: 20453838]
16. Wrensch M, et al. Variants in the CDKN2B and RTEL1 regions are associated with high-grade glioma susceptibility. Nat Genet. 2009; 41:905–908. [PubMed: 19578366]
17. Frazer KA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449:851–861. [PubMed: 17943122]
18. Depristo MA, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43:491–498. [PubMed: 21478889]
19. McKenna A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010; 20:1297–1303. [PubMed: 20644199]
20. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. Nature. 2010; 467:1061–1073. [PubMed: 20981092]
21. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010; 467:52–58. [PubMed: 20811451]
22. Li, Y.a.A.G. MACH 1.0: rapid haplotype reconstructoin and missing genotype inference. American Journal of Human Genetics. 2006; S70:2290.
23. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009; 10:387–406. [PubMed: 19715440]
24. Visel A, et al. Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. Nature. 464:409–412. [PubMed: 20173736]
25. Harismendy O, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-gamma signalling response. Nature. 2011; 470:264–268. [PubMed: 21307941]
26. Gnirke A, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009; 27:182–189. [PubMed: 19182786]
27. Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. Am J Hum Genet. 2005; 76:449–462. [PubMed: 15700229]
28. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet. 2001; 68:978–989. [PubMed: 11254454]
29. Price AL, et al. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006; 38:904–909. [PubMed: 16862161]
30. Purcell S, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–575. [PubMed: 17701901]

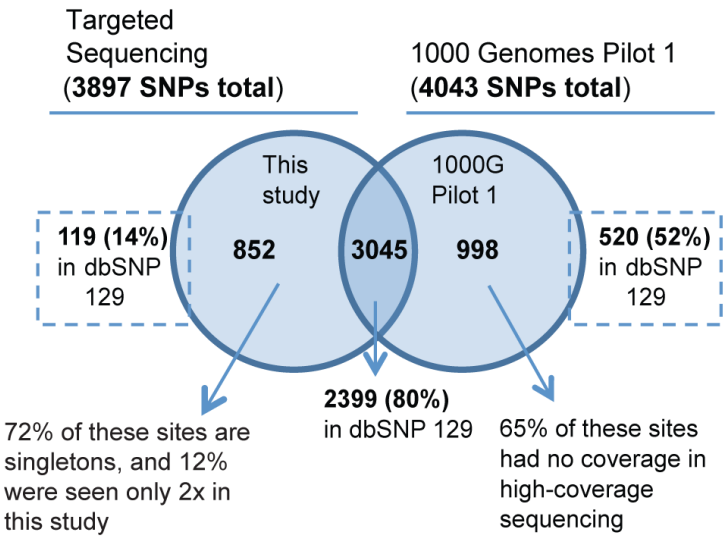

Figure 1. Comparison of targeted sequencing to 1000G Pilot 1 Data

Variant calls were made in all six regions of T2D association in the 32 individuals who were sequenced as part of both this targeted, high coverage sequencing effort (total 47 CEU HapMap individuals) as well as 1000G Pilot 1 (total 60 CEU HapMap individuals).

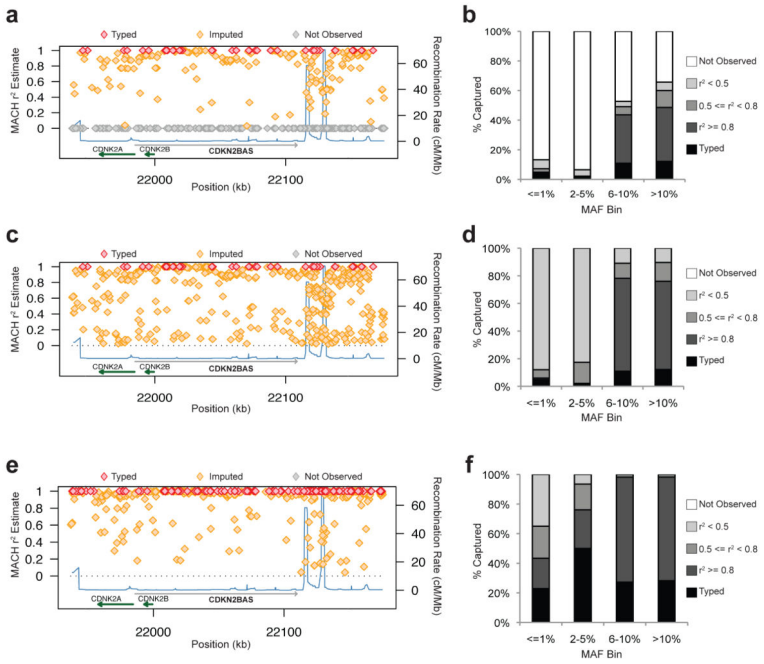

Figure 2. Fraction of variation on chromosome 9p21 captured in T2D disease cohort by different imputation scenarios

MACH imputation quality estimates **(a, c, e)** and overall fraction of variation captured in T2D samples **(b, d, f)** for different imputation scenarios. **(a, c, e)** The MACH-estimated r^2 for each SNP is plotted as a function of genomic position. SNPs not observed in the reference panel are assigned an r^2 of zero. Recombination rate (estimated from HapMap) is plotted to reflect local LD structure. Gene annotations were taken from the University of California-Santa Cruz Genome Browser. **(b, d, f)** The fraction of variants captured in T2D samples is shown as a function of MAF and MACH-estimated r^2. Imputation scenarios are: **(a, b)** Imputing from HapMap II (n=238 SNPs in 60 individuals) into the SNPs genotyped on the Affymetrix 500K array; **(c, d)** Imputing from 112 individuals genotyped at HapMap II sites and validated sequencing sites (total n=464 SNPs) into the SNPs genotyped on the Affymetrix 500K array; **(e, f)** Imputing from the same reference panel as c, d into SNPs genotyped on the Affymetrix 500K array plus additional tag SNPs genotyped in the T2D cohort (genotyped marker density in T2D samples ~1 SNP/1.5kb).

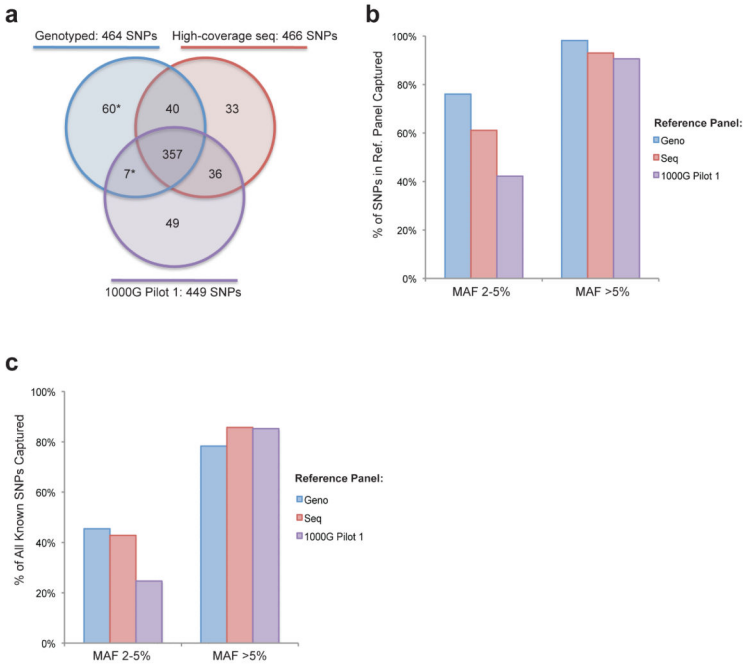

Figure 3. Comparison of imputation from a genotyped reference panel, directly from high coverage re-sequencing data, and directly from 1000G Pilot 1 data

(a) Variants present in the three reference panels and their overlap. The 67 variants present in the genotyped reference panel but not in the high coverage sequencing reference panel (denoted by asterisk) were called in high coverage sequencing as singletons and so were excluded from the sequencing reference panel. 40% of these variants are not singletons in the larger genotyped reference panel. **(b)** The fraction of sites within each reference panel captured with a MACH-estimated r^2 of at least 0.8. **(c)** The overall fraction of known variants captured with a MACH-estimated r^2 of at least 0.8 by imputation from each reference panel.

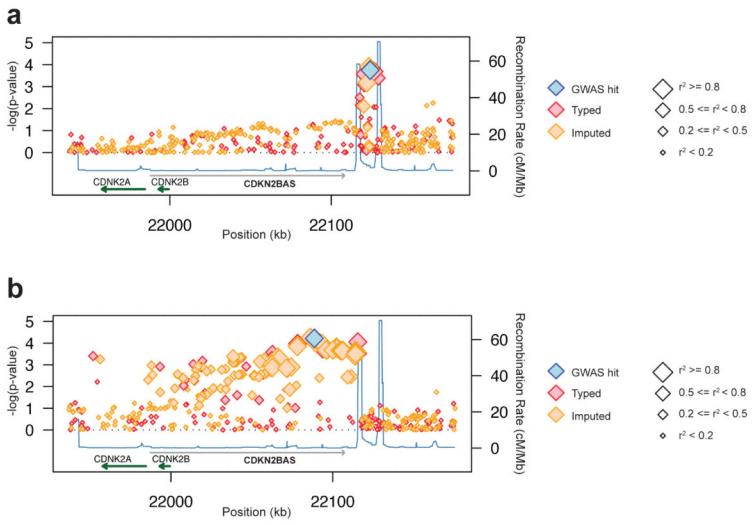

Figure 4. Association results for T2D and MI on chromosome 9p21

Regional plots showing association signal for **(a)** T2D and **(b)** MI. The signal for each SNP (represented as $-\log_{10}$ p-value) is plotted as a function of genomic position. The size of the diamond for each SNP represents the LD (measured as r^2) between that SNP and the original GWAS SNP (rs10811661 for T2D and rs4977574 for MI). Recombination rate (estimated from HapMap) is plotted to reflect the local LD structure in the region. Gene annotations were taken from the University of California-Santa Cruz Genome Browser.

Table 1

Sensitivity of 1000G Pilot 1 for variants detected in targeted, high coverage sequencing of samples common to both projects

Pilot 1 of the 1000 Genomes Project contains 97% of variants seen more than 5 times in high coverage sequencing, and 35% of variants seen once.

Number of times non-reference allele observed in this study	Number of SNPs called, this study	% Contained in 1000G Pilot 1	% in dbSNP, build 129	% Validated on chr9p21
1X	941	35%	13%	91%
2X	300	68%	42%	88%
3X	239	82%	55%	100%
4X	154	87%	66%	86%
5X	186	91%	67%	70%
>5X	2077	97%	92%	98%

Table 2
Haplotypic association to T2D on chromosome 9p21

rs10757282 and the reported SNP from GWAS, rs10811661, define haplotypes with three levels of risk (risk, protective, and neutral) for T2D.

Haplotypes defined by rs10757282, rs10811661			
Haplotype	**Frequency**	**OR**	***P*-value**
Overall Evidence	--	--	4.40×10^{-5}
CT	0.30	1.29	3.99×10^{-4}
TT	0.54	0.96	5.24×10^{-1}
CC	0.16	0.72	2.71×10^{-4}