Supplementary Information for:

Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction

Jessica Shea[1–3], Vineeta Agarwala[1,3–5], Anthony A Philippakis[1,3–7], Jared Maguire[1], Eric Banks[1], Mark DePristo[1], Brian Thomson[1], Candace Guiducci[1], Robert C Onofrio[8], The Myocardial Infarction Genetics Consortium[9], Sekar Kathiresan[1,6,10-12], Stacey Gabriel[1], Noël P Burtt[1], Mark J Daly[1,6,10,12], Leif Groop[13] & David Altshuler[1,3,6,10,12,14]

[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]Genetic Analysis Platform, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [9]A list of members is provided in this Supplementary Note. [10]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA. [11]Cardiovascular Research Center, Massachusetts General Hospital, Boston, Massachusetts, USA. [12]Department of Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA. [13]Department of Clinical Sciences, Diabetes and Endocrinology Research Unit, University Hospital Malmö, Lund University, Malmö, Sweden. [14]Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts, USA. Correspondence should be addressed to D.A. (altshuler@molbio.mgh.harvard.edu).

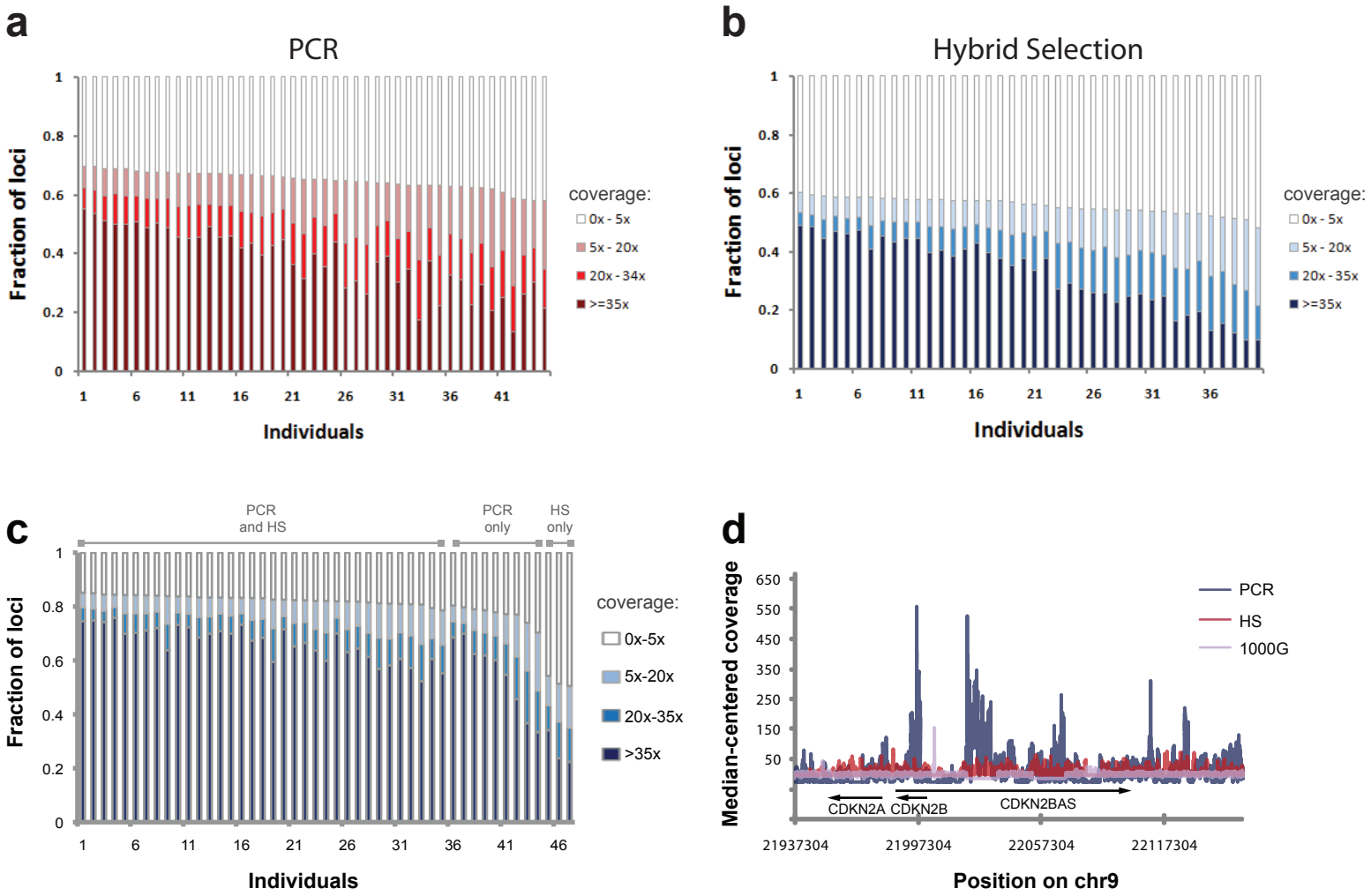
Information contained in this supplement:

Supplementary information in separate files:

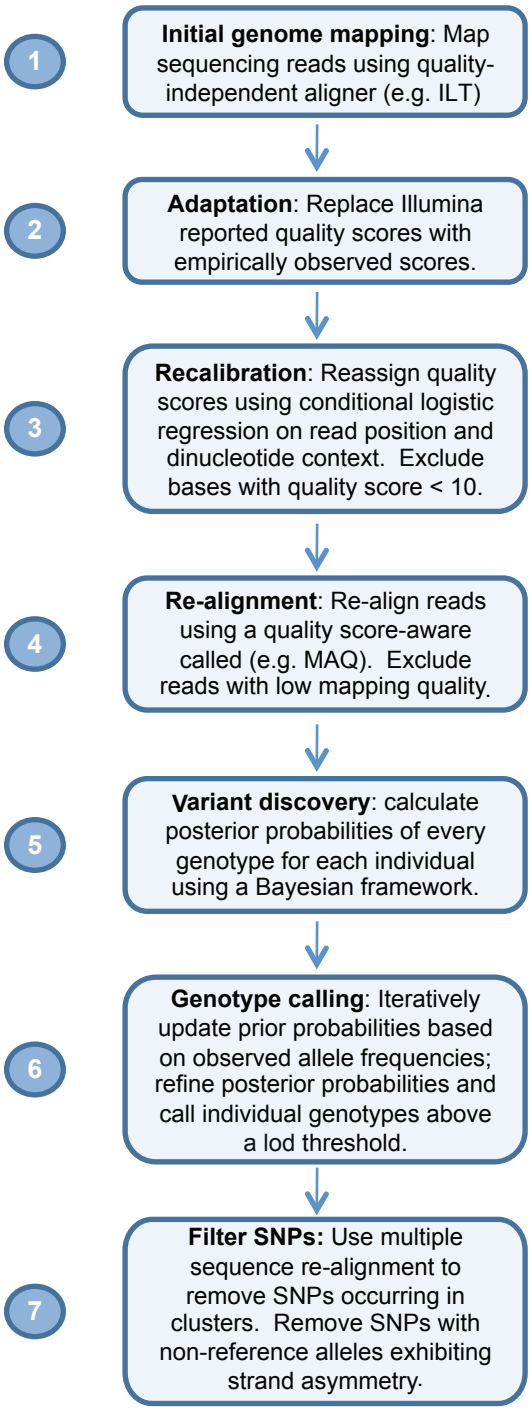
Supplementary Tables 2-6 (Excel files)

Supplementary Table 1: Regions sequenced and variants identified. Six T2D-associated regions were sequenced in 47 individuals from the HapMap CEU population. For each region, boundaries were defined to encompass all SNPs having an $r^2 \geq 0.2$ to the SNP with the lowest association p-value. The region on chromosome 9p21 near *CDKN2A* and *CDKN2B* was extended to cover the MI association. These regions comprise a total of 1.8Mb of sequence.

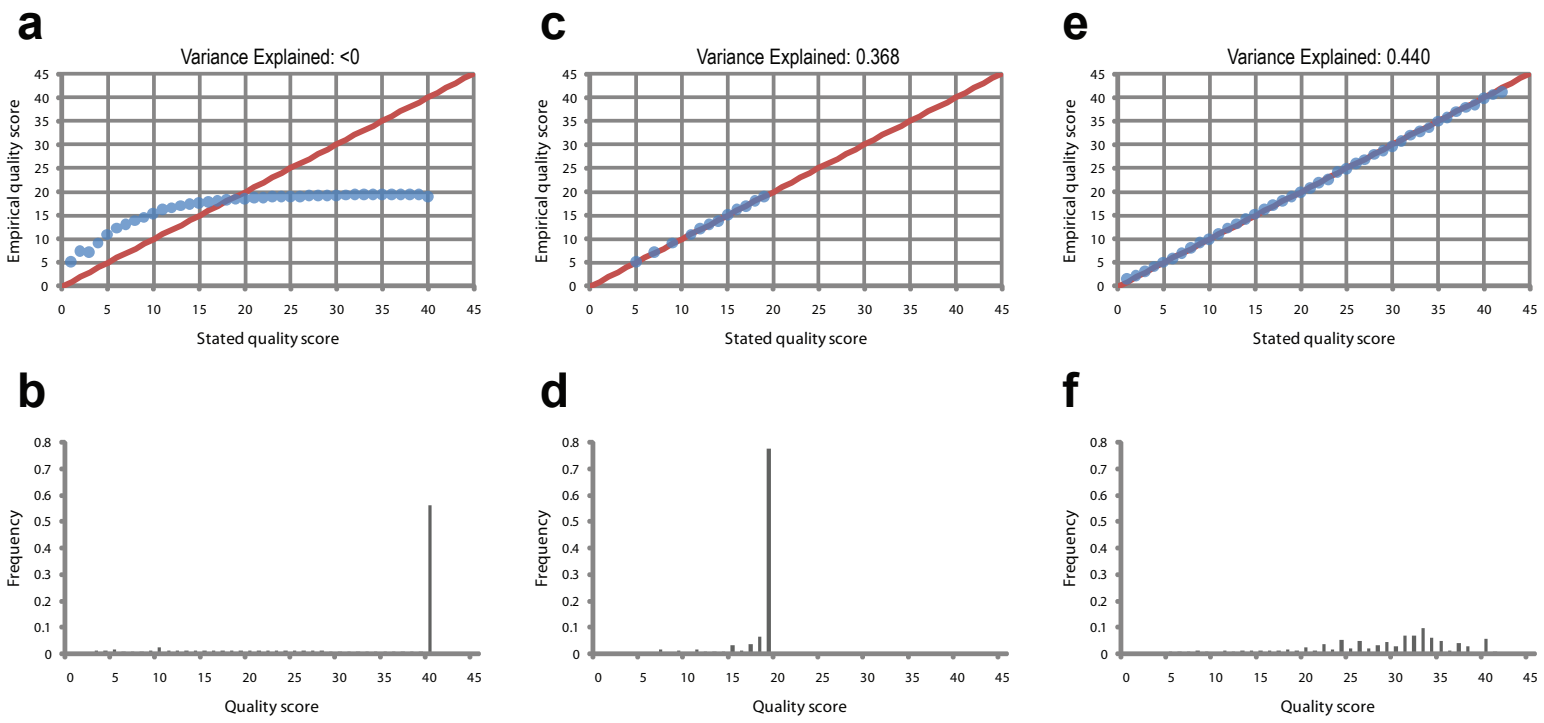
Chr	Coordinates (hg18)	Region Size (bp)	Genes	Percent of Bases Called	Variants Identified	Heterozygosity
3	186812966 - 187078478	265512	*SENP2, IGF2BP2*	62%	442	1/2,501
6	20617611 - 21368293	750682	*CDKAL1*	73%	2394	1/1,202
8	118194972 - 118290242	95270	*SLC30A8*	65%	192	1/2,030
9	21936711 - 22176221	239510	*CDKN2A, CDKN2B*	76%	635	1/1,266
10	94084878 - 94484911	400033	*MARCH5, IDE, KIF11, HHEX*	38%	611	1/2,016
16	52357187 - 52404063	46876	*FTO*	77%	189	1/621



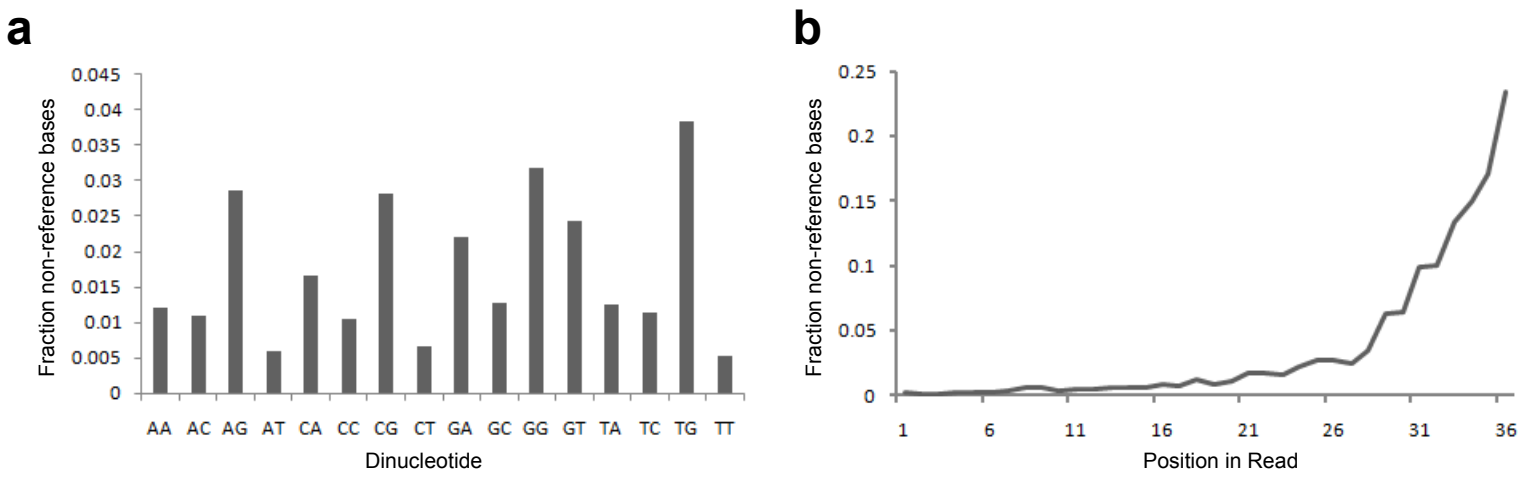
Supplementary Figure 1: Coverage of loci for (**a**) PCR and (**b**) Hybrid Selection DNA capture methods. Each bar represents one individual, and the coverage of loci across the six (for PCR) or three (hybrid selection) target regions for ranges 0x - 5x, 5x - 20x, 20x - 35x, >= 35x coverage is indicated by different shadings. Overall, approximately 70% of loci have 5x or greater coverage across all individuals; for hybrid selection only 50% of loci have 5x or greater coverage. Coverage for the the 9p21 locus across individuals (**c**) and genomic position (**d**) is also shown.



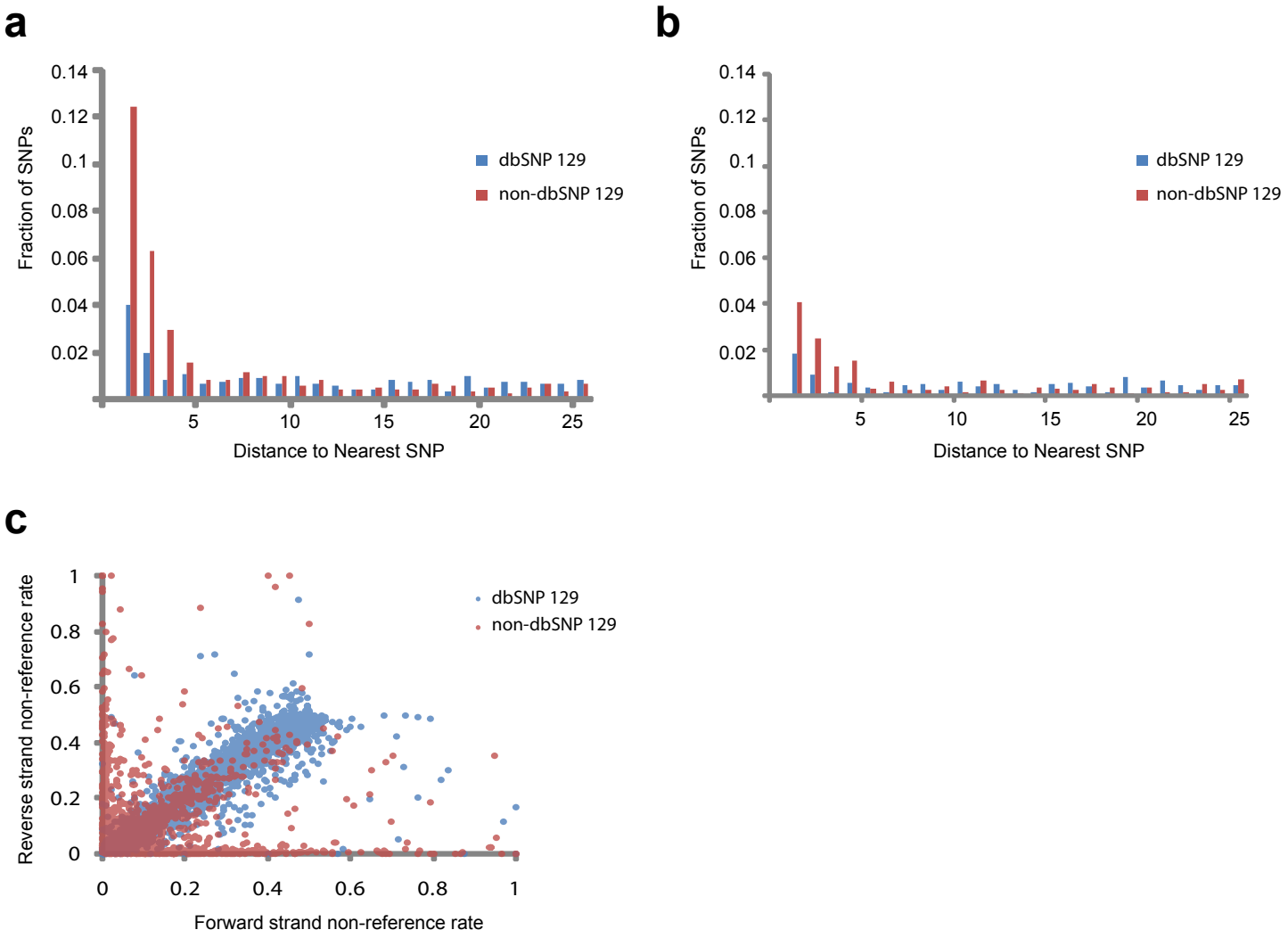
Supplementary Figure 2: Approach for identifying variants in next-generation sequencing.



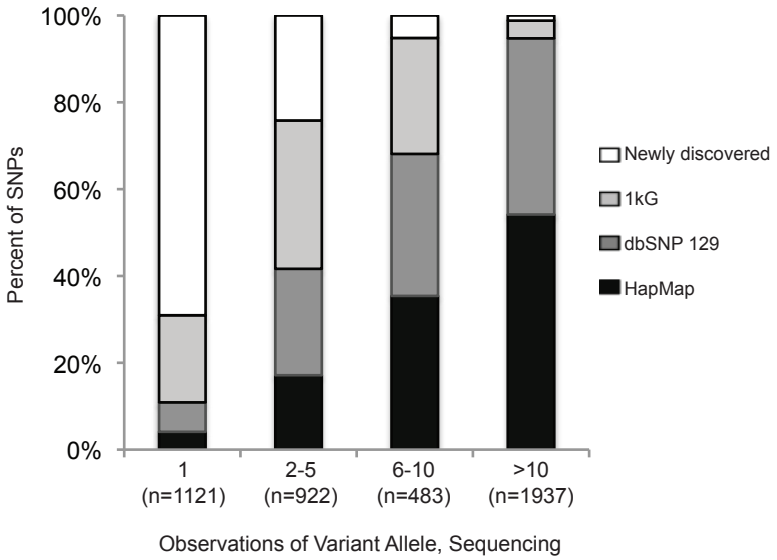
Supplementary Figure 3: Variance explained under various stages of adjusting quality scores. Each base comes equipped with a phred-style quality score q=-10log(ε), where ε is the probability that the base was miscalled. Stated quality scores diverged from empirical estimates (**a**), and were predominantly the same quality (**b**). We developed a method of adapting quality scores to make them empirically accurate (**c**), but noted that the bases were still predominantly a single quality score after applying this method (**d**). We therefore developed an additional method of recalibrating bases according to covariates such as position in read and dinculeotide context; this gave a markedly more dispersed distribution of quality scores (**e**) that were still empirically accurate (**f**).



Supplementary Figure 4: Properties of sequencing data used in recalibrating quality scores. For each base not aligning to a locus in dbSNP, we determined whether or not the base was non-reference. We then tallied the fraction of non-reference bases after stratifying on their dinucleotide context or position in the read. Plotted above are representative lanes showing that the non-referecnce rate depends on (**a**) the dinucleotide context and (**b**) the position in the read. The dinucleotides most predictive of high and low non-reference rates and the profile of non-reference reads as function of read position varied from lane to lane.

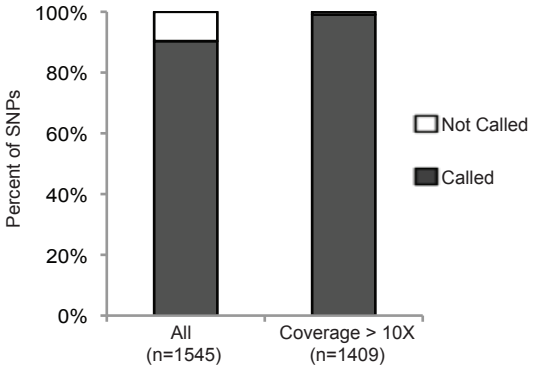


Supplementary Figure 5: Removal of additional artifacts to improve SNP calls. We observed an over-representation of SNPs in close proximity at non-dbSNP sites (**a**), which were removed by multiple requence re-alignment (**b**). We also observed non-reference calls for which the evidence was confined to one strand (**c**), which occured predominantly at non-dbSNP sites; these sites were filtered out of the final call set.



Supplementary Figure 6: Properties of variants identified across all six regions.

a



b

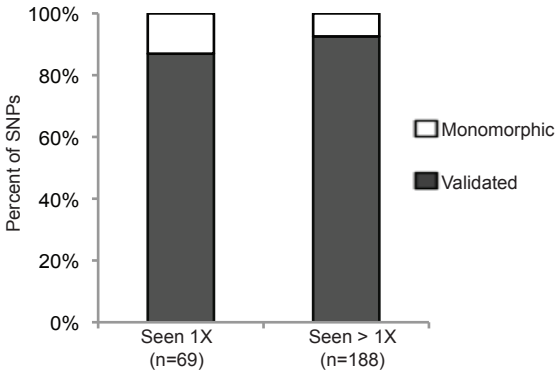
HapMap Genotype \ Sequencing Call	Hom Ref	Het	Hom Var
Hom Ref	99.74%	0.07%	0.19%
Het	1.21%	97.90%	0.88%
Hom Var	1.78%	2.24%	95.98%

c

NA12891	
1593	total snps called in 1kG Pilot 2
1122	sites at which coverage was >=10 in T2D
1074	sites at which our project called a variant allele
1059	sites at which both projects' genotype calls match
96%	**Sensitivity**
99%	**Genotype concordance**

NA12892	
1769	total snps called in 1kG Pilot 2
1232	sites at which coverage was >=10 in T2D
1172	sites at which our project called a variant allele
1158	sites at which both projects' genotype calls match
95%	**Sensitivity**
99%	**Genotype concordance**

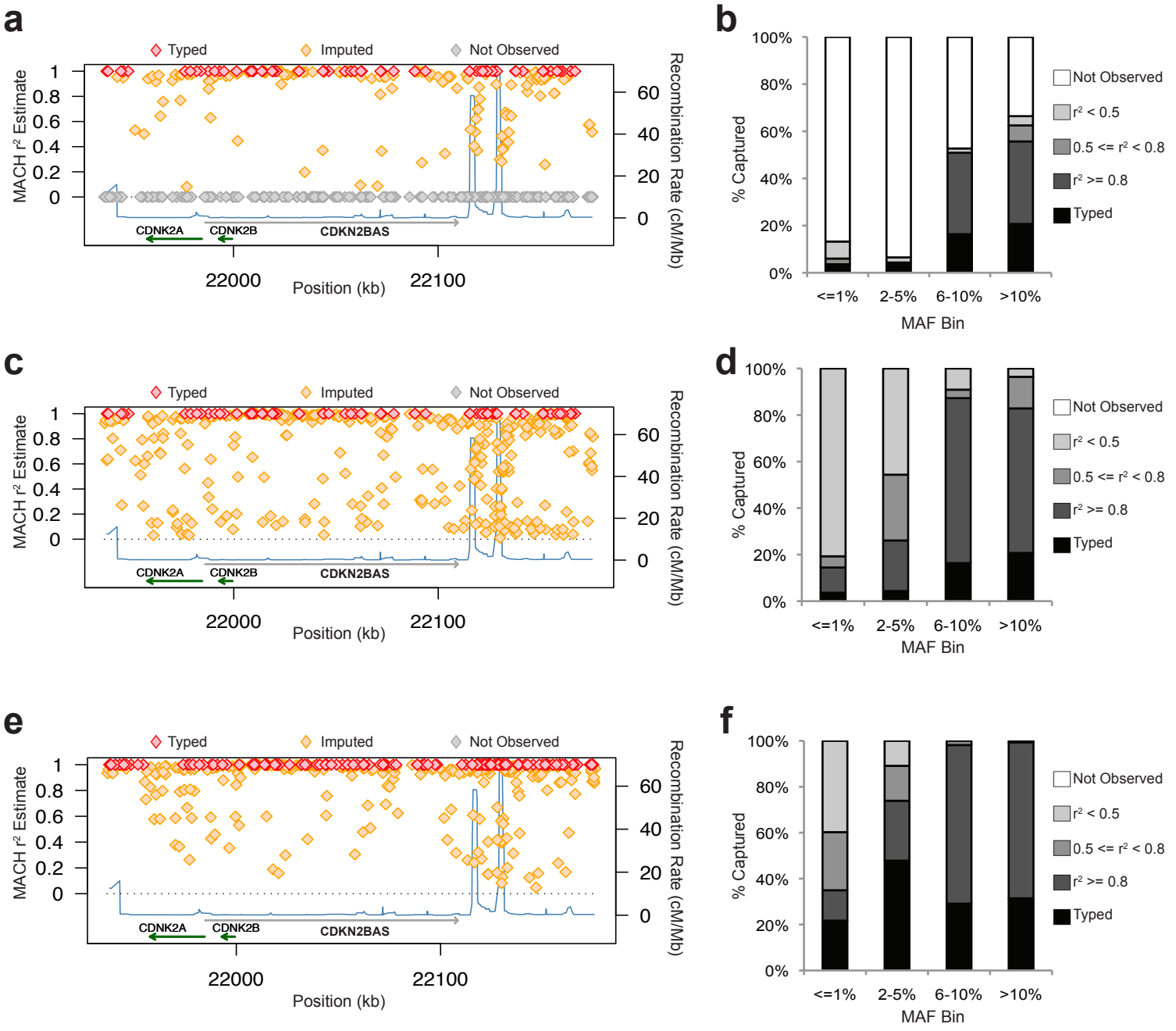
d



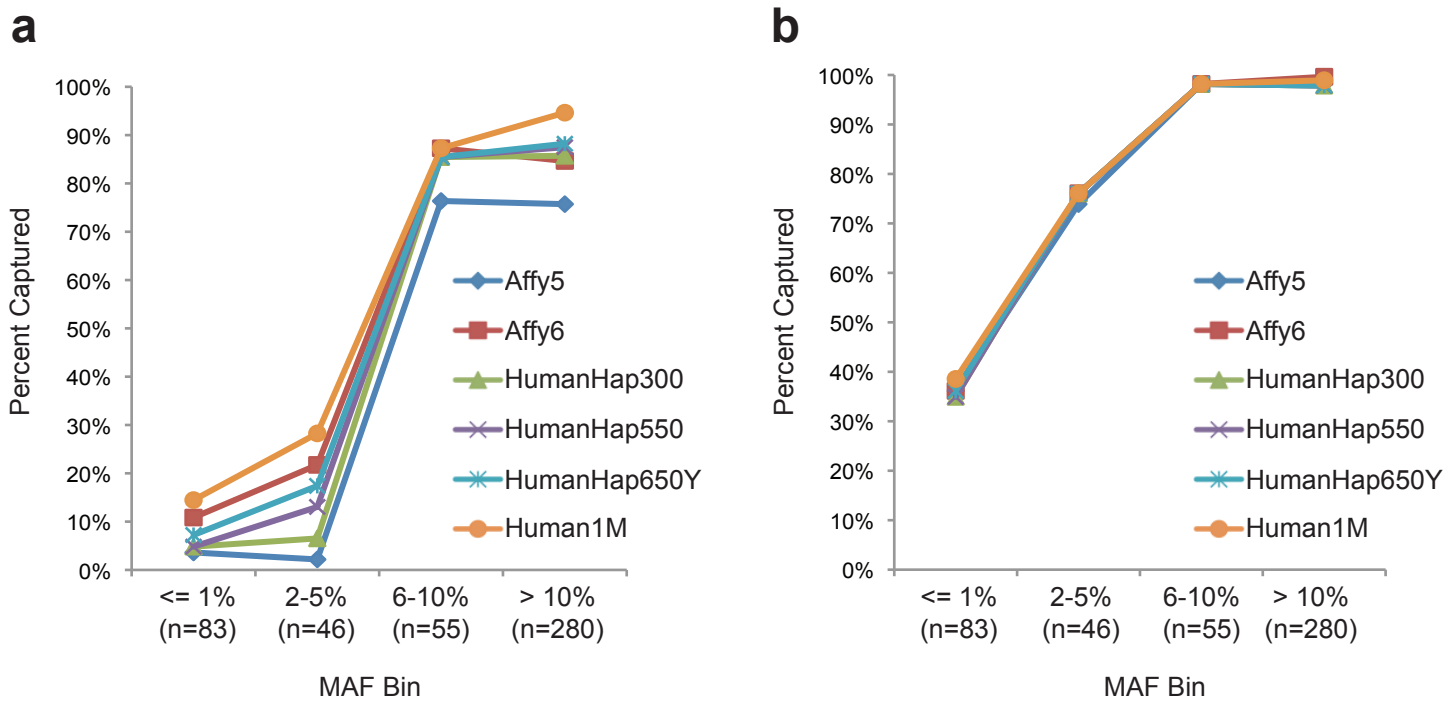
e

Validation Genotype \ Sequencing Call	Hom Ref	Het	Hom Var
Hom Ref	98.58%	1.27%	0.15%
Het	5.93%	93.54%	0.53%
Hom Var	0.80%	6.36%	92.84%

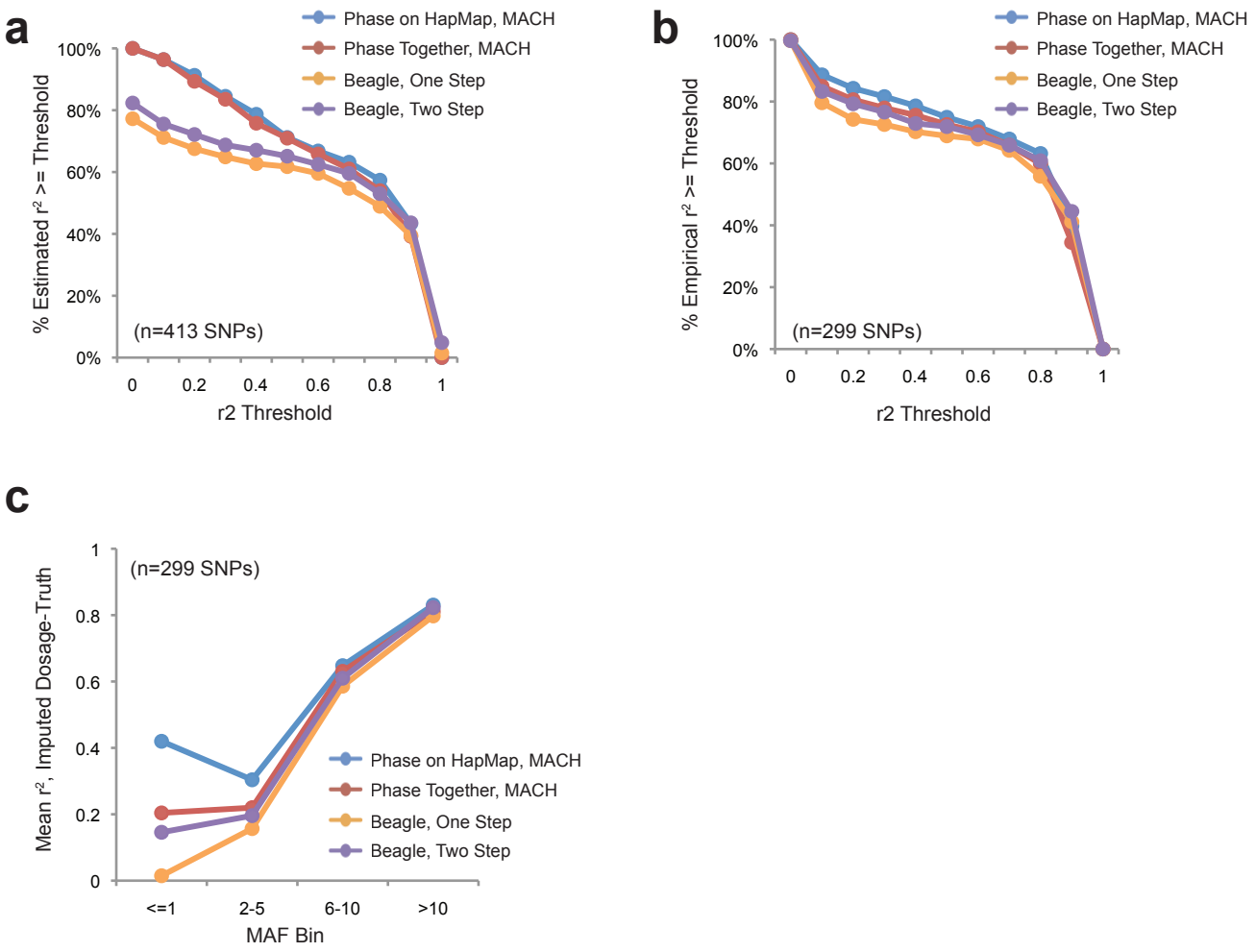
Supplementary Figure 7: Sensitivity, specificity and accuracy of variant calls in sequencing data. (**a**) Sensitivity for HapMap sites in high-coverage sequencing data across all six regions. We detected 90% of HapMap SNPs polymorphic in the 47 sequenced individuals overall, and 99% when we had at least 10-fold coverage in individuals with the variant allele. (**b**) Accuracy of individual genotype calls made from sequencing data at HapMap sites across all six regions sequenced (n=28,507 comparisons). (**c**) Comparison of our high-coverage sequencing to high-coverage 1kG Pilot 2 data across all six regions in the two CEU individuals common to both projects. (**d**) Specificity at sites not previously genotyped in HapMap based on validation genotyping on chromosome 9p21. (**e**) Accuracy of individual genotype calls made from sequencing data at sites not previously genotyped in HapMap based on validation genotyping on chromosome 9p21 (n=7,573 comparisons).



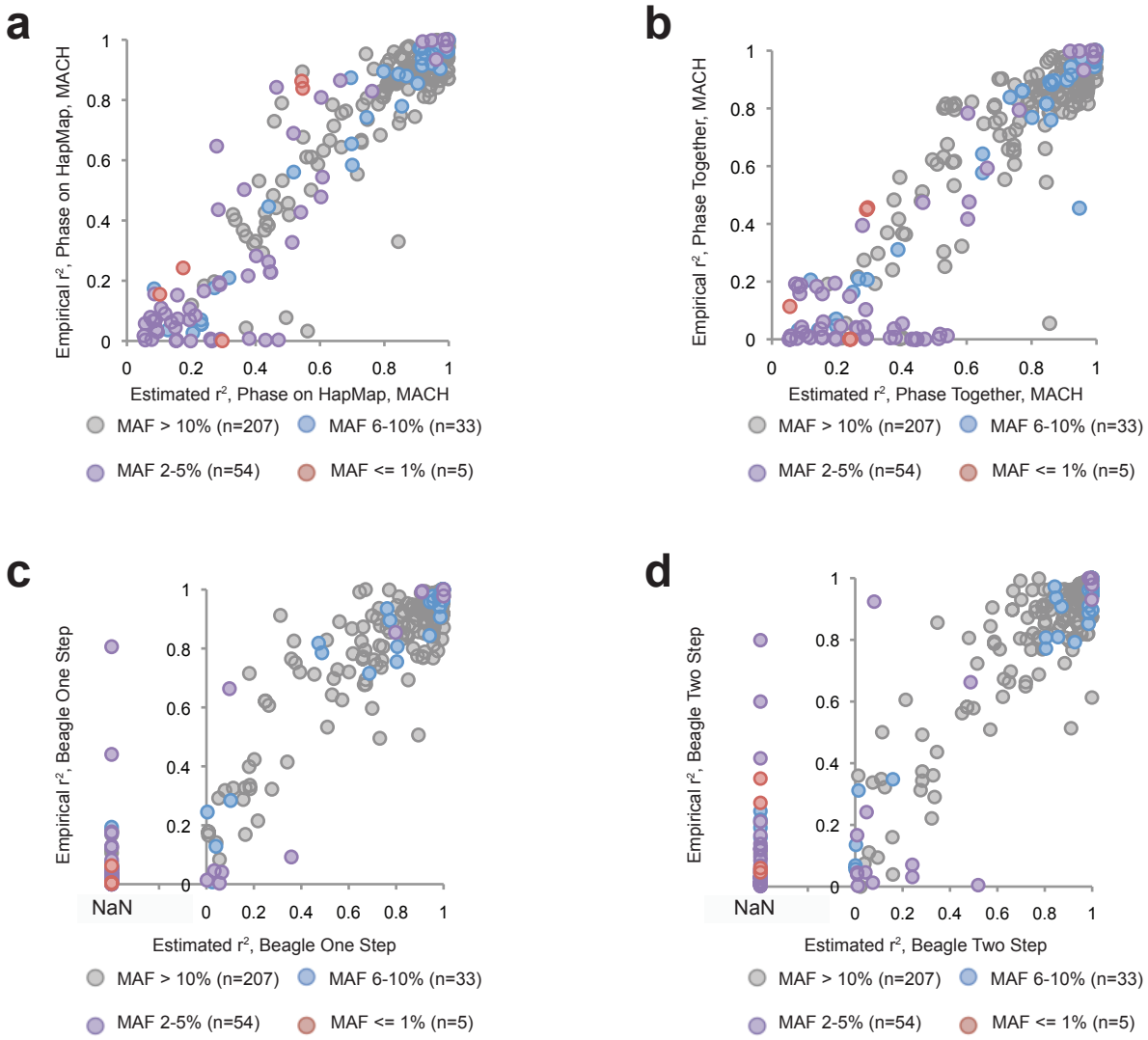
Supplementary Figure 8: Percentage of variation on chromosome 9p21 captured in the MI disease cohort by different imputation scenarios. (**a-f**) MACH imputation quality estimates (**a, c, e**) and overall percentage of variation captured in MI samples (**b, d, f**) for different imputation scenarios. (**a, c, e**) MACH-estimated r^2 for each SNP versus genomic position. SNPs not observed in the reference panel are assigned $r^2 = 0$. Recombination rate was estimated from HapMap II. (**b, d, f**) The fraction of variation captured in MI case-control samples versus MAF and MACH-estimated r^2. Imputation scenarios include imputing from HapMap II into the SNPs genotyped on the Affymetrix 6.0 array (**a, b**), imputing from 112 CEU individuals genotyped at HapMap II sites and validated sequencing sites into the SNPs genotyped on the Affymetrix 6.0 array (**c, d**) and imputing from the same reference panel as **c** and **d** into SNPs genotyped on the Affymetrix 6.0 array plus additional tag SNPs genotyped in the MI cohort.



Supplementary Figure 9: Adding tag SNPs improves imputation from multiple chips. We performed imputation from a genotyped reference panel (n=464 SNPs in 112 HapMap CEU individuals from 56 trios) into 90 HapMap TSI individuals genotyped at the same set of SNPs. (**a**) Imputation results when genotypes in TSI individuals were downsampled to the SNPs present on the indicated commercial GWAS arrays. Y-axis shows the overall percent of SNPs in the reference panel captured in TSI samples. (**b**) Imputation results when genotypes in TSI individuals were downsampled to SNPs present on the indicated commercial GWAS arrays plus the set of tag SNPs we genotyped in our entire T2D and MI cohorts to improve imputation (eg, the set of tag SNPs used in Fig. 2 and Supplementary Fig. 8). Affy5 and Affy6 indicated the Affymetrix 500K and 6.0 arrays, respectively. The HumanHap and Human1M arrays are Illumina arrays.



Supplementary Figure 10: Imputation directly from Illumina data. We compared four methods for imputing directly from Illumina data into our T2D samples on chromosome 9p21 and used genotypes for a sampling of the imputed SNPs (n=299 out of 413 imputed SNPs) in a subset of our T2D samples (n = 319 individuals) to evaluate imputation performance. All methods performed comparably in terms of the fraction of imputed SNPs captured as a function of the MACH- or Beagle-estimated (**a**) and empirical (**b**) r^2 between imputed dosages and true genotypes. Phasing sequencing genotypes on top of HapMap haplotypes and imputing with MACH resulted in slightly higher accuracy at rare SNPs (**c**).



Supplementary Figure 11: Comparison of MACH- and Beagle estimated r^2 to empirical values when imputing directly from Illumina data into disease samples. We compared four methods of imputing directly from Illumina data. In each case, the imputation engine (MACH or Beagle) provides an estimate of how well each SNP is imputed. To evaluate how accurate these estimates are when imputing directly from Illumina data, we used genotypes at a sampling of imputed SNPs (n=299 out of 413 imputed SNPs) in a subset of our T2D samples (n=319 individuals) to calculate the r^2 between imputed dosages and observed genotypes and compared these values to the estimates provided by the imputation engines for each of the four methods: (**a**) phasing sequencing genotypes on top of HapMap haplotypes and imputing with MACH, (**b**) phasing sequencing genotypes along with HapMap genotypes and imputing with MACH, (**c**) imputing from sequencing likelihoods and HapMap genotypes directly into study samples using Beagle, and (**d**) creating phased haplotypes from sequencing likelihoods and HapMap genotypes using Beagle, and then imputing from those haplotypes into study samples using Beagle.

SUPPLEMENTARY NOTE

Sequencing

I. Sequencing protocols and data pre-processing

Six regions associated with Type 2 Diabetes (T2D) were selected for targeted re-sequencing. Region boundaries were selected to contain all SNPs in linkage disequilibrium ($r^2 \geq 0.2$) with the T2D-associated SNP of lowest p-value. For chromosome 9p21, we also sequenced the region associated with myocardial infarction (MI). In total, these 6 regions comprise nearly 1.8Mb of sequence, with individual regions ranging from ~47kb to ~750kb in size. **Supplementary Table 1** lists the exact coordinates for each region, as well as a list of all genes contained within them.

DNA was captured for sequencing by long-range PCR with 2-5kb amplicons (**Supplementary Table 2.1**) or by hybrid selection (HS) using 170bp baits (**Supplementary Table 2.2**) tiled across the region on an Agilent microarray[1]. All re-sequencing was performed at the Broad Institute in 2008 using Illumina Genome Analyzers (GA). Runs from PCR-based capture generated 36bp reads and runs from HS-based generated 46-50bp reads. We sequenced these regions in 47 individuals of European ancestry from the HapMap CEU population[2].

We noted that HS resulted in more uniform, but lower, coverage than PCR-based capture (**Supplementary Fig. 1 a-d**). Overall, 70% of the targeted regions had at least 5x coverage in all individuals (**Supplemental Fig. 1c**).

II. Alignment and data pre-processing

An overview of our analysis framework for variant calling from Illumina sequencing data is shown in **Supplementary Figure 2**. Reads were first aligned to the reference genome using Imperfect Lookup Table (ILT), an aligner developed at the Broad Institute that does not utilize base quality scores when aligning. This is important given that quality scores from the Illumina GA are not reflective of true error rates (see below).

After quality score adaptation (Section III below), we re-aligned the reads using MAQ[3], which provides greater accuracy than ILT but relies on having accurate base quality scores. In aligning with MAQ, we used the following settings: -D, -e 100, -s 0. We discarded all reads with a mapping quality score < 20.

III. Quality score adaptation

The Illumina GA provides each sequenced base (b_i) with a phred-style quality score (q_i)

$$q_i = -10 log_{10} \varepsilon_i \quad (3.1)$$

where ε_i represents the probability that base b_i has been erroneously called. Inspection of output quality scores reveals that stated error rates diverge from empirical values and are predominantly the same value (**Supplementary Fig. 3a, b**). Specifically, let B_q be the collection of all bases with quality score q, and let $M_q \subset B_q$ be the subset of bases with quality score q that are non-reference (i.e., aligned bases that do not agree with the reference). Then for each q, we compute:

$$\hat{q} = -10 log_{10} \left(\frac{|M_q|}{|B_q|} \right) \quad (3.2)$$

Thus, $\hat{q}$ represents the empirical quality score for all bases with stated quality score q. Note that this calculation implicitly assumes that all non-reference bases are errors, an overly conservative estimate. To address this difficulty, in computing $\hat{q}$ we ignore all dbSNP positions; since the frequency of polymorphisms outside dbSNP sites is ~10^{-5}, and since the error rate of the Illumina GA is much higher (~1-5%), this provides a reasonable estimate of the empirical error rate $\hat{q}$. As shown in **Supplementary Figure 3a,** $\hat{q}$ is not, in general, equal to q; in an "adaptation" procedure, we replace stated quality scores with empirical values (**Supplementary Fig. 3c**).

IV. Quality score recalibration

After adaptation, the majority of bases were still assigned identical quality scores (**Supplementary Fig. 3d**), resulting in low information content. To address this, we identified several covariates, including dinucleotide context and base position within a read, that were predictive of base error rates even after stratifying on quality score. Specifically, we associate each sequenced base b_i to a dinucleotide context d_i as well as its relative position p_i in a read. We observed that error rates (as measured by empirical non-reference rates) were non-uniform across both these metrics (**Supplementary Fig. 4a, b)**. Moreover, the values of d and p causing the highest and lowest error rates varied from lane to lane, depending on the experimental chemistries used.

We developed a "recalibration" method to repartition observed quality scores based on these metrics using conditional logistic regression. Each base emitted from the sequencer gives a 4-tuple (b_i, q_i, d_i, p_i), and the following regression across all bases yields values for the coefficients $\beta_{j,k,d}$

$$log\left(\frac{\mathbb{P}(b\neq ref|d)}{\mathbb{P}(b=ref|d)}\right) = \sum_{j=0}^{4}\sum_{k=0}^{4}\beta_{j,k,d}q^j p^k \qquad (4.1)$$

The coefficients $\beta_{j,k,d}$ can then be used to compute a new estimate $\widetilde{q_i}$ of the quality score for base b_i

$$\tilde{q}_i = \frac{exp\left(\sum_{j=0}^{4}\sum_{k=0}^{4}\beta_{j,k,d_i}q_i{}^j p_i{}^k\right)}{1+exp\left(\sum_{j=0}^{4}\sum_{k=0}^{4}\beta_{j,k,d_i}q_i{}^j p_i{}^k\right)} \qquad (4.2)$$

After performing re-calibration, the distribution of quality scores was markedly more uniform, and empirical quality scores matched their stated values when conditioned on both position in read and dinucleotide (**Supplementary Fig. 3e, f**). Because there were frequently large jumps in error rates at the first and last bases of a read, these bases were dropped in all analyses.

We also noted that false-positive calls often occurred at positions where aligned bases were of lower quality or fewer in number than bases immediately adjacent to them. To quantify this, we used a "neighborhood quality standard" (NQS). At a given genomic locus *l,* let $b_{l,j}$ be all bases across all individuals that map to position *l*. We define the NQS at *l* by

$$NQS_l = \frac{\sum_j b_{l,j}}{\frac{1}{20}\sum_{\substack{k=l-10 \\ k\neq l}}^{k=l+10}\sum_j b_{k,j}} \qquad (4.3)$$

NQS is a weak but consistent predictor of base error rate; we removed its effects by discretizing the NQS score at each locus and adapting the quality scores.

We evaluated the fraction of variance explained 1) in the raw quality scores, 2) after adaptation, 3) after re-calibration over the position and dinucleotide covariates, and 4) after recalibration over NQS. The variance explained was computed using the method of Nagelkerke[4]. As shown in **Supplementary Figure 3**, the raw quality scores explain no variance and assigning each base the average error rate of the lane explains more variance than the raw quality scores. After all recalibrations were made to the quality scores, approximately 40% of the variance is explained. Thus, adaptation and recalibration act to explain a substantial amount of variance, yet there are also likely other covariates that might also be utilized to further improve quality scores.

We also examined the effect of recalibration on SNP calling. Adaptation and recalibration lead to a dramatic reduction in the total number of SNPs while causing almost no change to the number of dbSNP sites discovered, suggesting that the majority of removed SNPs were false positives.

V. SNP calling algorithm

We developed a Bayesian framework for detecting SNPs which has since been incorporated into the Genome Analysis Toolkit (GATK) [5,6]. Specifically, let $G=\{AA, AC, AG, AT, CC, ..., TT\}$ represent the ten possible diploid genotypes at a given locus, and let D represent the collection of bases and their quality scores observed at that locus. Following the notation of the previous section, $D=\{(b_1,q_1), ...(b_N,q_N)\}$, where b_i is the i'th base, q_i is its corresponding quality score after recalibration, and q_i=-10logε_i where ε_i is the probability that base b_i has been miscalled.

We seek to compute the posterior probability of each genotype G given the data D

$$\mathbb{P}(G|D) = \frac{\mathbb{P}(D|G)\mathbb{P}(G)}{\mathbb{P}(D)} = \frac{(\prod_{b_i \in D}\mathbb{P}(b_i|G))\mathbb{P}(G)}{\sum_G(\prod_{b_i \in D}\mathbb{P}(b_i|G))\mathbb{P}(G)} \qquad (5.1)$$

Let N_i denote any base not equal to b_i. Then G$\in$ {b_ib_i, b_iN_i, N_iN_i} at the i'th base, and $\mathbb{P}(b_i|G)$ will vary with the value of G and b_i

$$\mathbb{P}(b_i|G) = \begin{cases} (1-\varepsilon_i) & if\ G = b_ib_i \\ \varepsilon_i & if\ G = N_iN_i \\ (0.5 - \varepsilon_i/2) & if\ G = N_ib_i \end{cases} \qquad (5.2)$$

The term $\mathbb{P}(G)$ represents our prior probability of observing genotype G at the locus. For most applications, we use a variation-aware prior. Specifically, if the locus has been genotyped in the HapMap project, we let $\mathbb{P}(G)$ be the expected frequency of G under the assumption of Hardy-Weinberg equilibrium based on the reported CEU frequencies; for sites that are in dbSNP but not genotyped in HapMap, we assume a minor allele frequency of 0.05 and compute the expected frequency of G under the assumption of Hardy-Weinberg equilibrium; finally, for sites that are not in dbSNP, we assume a prior probability of $2*10^{-4}$ for heterozygous non-reference genotypes and 10^{-5} for homozygous non-reference genotypes. Note, however, that in several applications (e.g., discovering the strand asymmetry and clustering artifacts shown in **Supplementary Fig. 5a-c**), it is useful to evaluate SNP-calling performance by stratifying on membership in dbSNP. In these cases, the preceding priors would clearly be problematic as they are biased toward re-discovering known sites. Therefore, in these applications we place a prior of 10^{-3} for any heterozygous non-reference genotype, and 10^{-5} for any homozygous non-reference genotype.

VI. Removing SNP clusters

As shown in **Supplementary Figure 5a,** SNPs often occurred in clusters. This clustering of SNPs tended to occur more frequently at non-dbSNP sites, suggestive of false positives. In many cases, we observed that the cause of SNP clustering was either an insertion/deletion or a locally repetitive region. To address this, we developed a method to perform multiple sequence realignment (MSR) at specified regions. At each interval where SNP clustering is observed, the MSR allows at most one insertion or deletion in each read. Let $b_{i,j}$ denote a base aligning to locus i in read j before MSR, and let $\widetilde{b_{i,j,D}}$ be a base aligning to locus i in read j after MSR when allowing an insertion or deletion D. For each possible insertion or deletion D at the locus, a measure of the goodness of fit is computed by

$$\varphi_D = log\left(\frac{\mathbb{P}(\widetilde{b_{i,j,D}})}{\mathbb{P}(b_{i,j})}\right)$$

where

$$\mathbb{P}(b_{i,j}) = \begin{cases} 1-\varepsilon_{i,j}\ if\ b_{i,j}\ is\ reference \\ \varepsilon_{i,j}\ if\ b_{i,j}\ is\ non-reference \end{cases}$$

and $\varepsilon_{i,j}$ is the estimated error rate of $b_{i,j}$ from the previously described recalibration procedure. We then take the insertion or deletion D that maximizes φ_D, and retain those insertions or deletions for which $\max(\varphi_D) > 5.0$. The effect of MSR is shown in **Supplementary Figure 5b**; this markedly reduced the number of clustered SNPs.

VII. Improving SNP calls through iterative updating of priors

We developed a two-step method of iteratively updating the prior probability of genotypes at each SNP to incorporate information from all individuals in the study. Intuitively, the goal of this analysis is to demand less evidence in order to call a non-reference genotype at loci that are more frequently observed to be non-reference.

Step I: Estimate the population allele frequencies f_A, f_C, f_G, f_T using the sequenced individuals according to the formula

$$f_j = \frac{\sum_i \left(2\mathbb{P}_i(N_jN_j|D) + \sum_{k \neq j}\left(\mathbb{P}_i(N_jN_k|D)\right)\right)}{2I} \qquad (7.1)$$

where $\mathbb{P}_i(N_jN_k|D)$ is the posterior probability of the genotype N_jN_k in sequenced individual i, and I is the number of sequenced individuals.

Step II: Update the prior probability of the genotype N_jN_k according to the expected frequencies under Hardy-Weinberg equilibrium

$$\hat{\mathbb{P}}(N_jN_k) = \begin{cases} f_j^2 \; if \; j = k \\ 2f_jf_k \; if \; j \neq k \end{cases} \qquad (7.2)$$

and use this value to re-compute the posterior probability of genotype N_jN_k.

$$\mathbb{P}(N_jN_k|D) = \frac{\mathbb{P}(D|N_jN_k)\hat{\mathbb{P}}(N_jN_k)}{\mathbb{P}(D)} \qquad (7.3)$$

We iterate Steps I and II until the new frequency of each nucleotide computed in Step I is < 0.0001 of its previous value. The output of this procedure is both an updated posterior probability for each genotype, as well as an estimate of the allele frequencies f_j. We have found that this procedure substantially increases the sensitivity of SNP calling, especially at lower coverage sites.

VI. Removal of strand asymmetries

We noted that for many SNPs, particularly at non-dbSNP sites, the evidence for the non-reference allele was confined to either the forward or reverse strand (**Supplementary Fig. 5c**). We therefore developed a method for testing whether any observed strand asymmetries were statistically significant.

Let D be the collection of bases at a given locus, and let D^+ and D^- be those bases that occur on the forward and reverse strands, respectively. We apply the iterative procedure of the previous section in order to estimate the allele frequencies f_j^- and f_j^+ on each strand separately, as well as the allele frequencies f_j using both strands simultaneously. Let $\mathbb{P}_{f^+}(D^+)$, $\mathbb{P}_{f^-}(D^-)$ and $\mathbb{P}_f(D)$ be the probabilities of the data using these estimated allele frequencies as in equation 7.3 and 5.1. Then we can compute the statistical significance of any observed strand asymmetries with a standard likelihood ratio test

$$2log\left(\frac{\mathbb{P}_{f^+}(D^+)\mathbb{P}_{f^-}(D^-)}{\mathbb{P}_f(D^+)\mathbb{P}_f(D^-)}\right) \sim \chi_1^2$$

where χ_1^2 is a chi-squared random variable with one degree of freedom.

Sensitivity, Specificity, and Accuracy Analysis for Sequencing Calls

I. Sensitivity for HapMap SNPs

We used HapMap II data[2] (Release 23a) for all six regions sequenced. This dataset contained 1545 SNPs in the sequenced regions that were polymorphic in the 47 individuals sequenced, of which we called 90%. The vast majority of HapMap SNPs that we did not call were missed due to having low or no coverage in the sequencing data. When we had greater than 10-fold coverage in individuals with the variant allele, we called 99% of HapMap SNPs (**Supplementary Fig. 7a**).

The 1% of well-covered SNPs that we missed fell into two categories: (1) SNPs that were initially called by our SNP calling algorithm but that were subsequently filtered out due to strand asymmetry or an overlapping indel and (2) SNPs represented by a single heterozygous individual for whom we had high (30-300X) coverage but saw no evidence for the variant allele, suggesting these are possible genotype errors in HapMap.

II. Comparison to 1000G Pilot 2 Data

1000G Pilot 2[7] included high coverage (>25x) data for one CEU trio; the parents in this trio were also sequenced in our project. We compared our sequencing calls for these individuals to 1000G Pilot 2 data. Across the six regions we sequenced, 1000G Pilot 2 called 1593 SNPs in the CEU father (NA12891) and 1769 SNPs in the CEU mother (NA12892). The 1000G Pilot 2 project had much more uniform coverage than our project; we had greater than 10-fold coverage at only 70% of these sites of variation. At this subset of well-covered sites, we detected 97% of the 1000G Pilot 2 SNPs in our sequencing across all 47 individuals, and made a variant call within NA12891 and NA12892 in 95-96% of cases; furthermore, at variant sites called by both projects, we called exactly the same genotype in 99% of cases (**Supplementary Fig. 7c**).

Over half of the SNPs that targeted re-sequencing did not detect (20 out of 37 variants in NA12891, for example) were initially called and then subsequently filtered out due to the presumed presence of insertion-deletion variants or due to strand asymmetry. However, 15 of these 20 sites are in fact annotated in dbSNP as potential insertion-deletion sites, suggesting that our MSR filtering procedure is relatively specific. Thus, after correcting for likely false positive sites, the true sensitivity of our project as compared to the 1000G Pilot 2 data is 98% for overall variant detection and 97% for detection of the variant allele within NA12891 and NA12892. These sensitivities are slightly lower than the 99% sensitivity achieved for sites of common variation in HapMap, but this was expected given that less evidence is available for the detection of rare alleles.

III. Validation genotyping of variant calls in sequencing data

To evaluate the specificity of variant calls from targeted re-sequencing, we performed validation genotyping of non-HapMap sites on chromosome 9p21 (**Supplementary Fig. 7d**). In this validation, we attempted to type all variants seen more than once in our sequencing data, as well as a random sampling of variants seen only once (singletons). We did not attempt all singletons because the vast majority of these were novel (**Supplementary Fig. 6**) and possibly private mutations. A detailed summary of validation analysis on chromosome 9p21 is provided in **Supplementary Table 4**.

IV. Accuracy of individual genotype calls in sequencing data

As described above, the Bayesian SNP caller generates a posterior probability of each possible genotype for each individual at each site. We made individual genotype calls at sites where the lod score ($\log_{10}$ of the ratio of the two highest genotype posterior probabilities) exceeded 3.0. We compared these high-confidence genotypes to HapMap genotypes across the six regions sequenced and to validation genotypes on chromosome 9p21 (**Supplementary Fig. 7b, e)**.

Study Samples and Clinical Characteristics

The study samples used here have been previously described[8,9]. For T2D, the samples were from the Diabetes Genetics Initiative cohort. This cohort has 2931 individuals previously analyzed in our T2D GWAS: 2097 unrelated individuals and 834 individuals in discordant sibships[9]. We used only individuals from the unrelated case-control group. We excluded 49 individuals for whom we found genotype discrepancies in multiple rounds of genotyping and used the remaining 2048 individuals (1000 cases and 1048 controls). These individuals had been previously genotyped on the Affymetrix GeneChip Human Mapping 500K SNP Array Set[9].

For MI, we used early-onset MI cases and matched controls from five studies that are part of the Myocardial Infarction Genetics Consortium[8]. The five studies were (i) Heart Attack Risk in Puget Sound; (ii) Registre Gironi del Cor; (iii) Massachusetts General Hospital Premature Coronary Artery Disease; iv) FINRISK; and v) Malmö Diet and Cancer Study. The collection and clinical characteristics of these samples have been previously described[8]. These studies included a total of 1274 cases and 1407 controls. These samples had been previously genotyped on the Affymetrix Genome-Wide Human SNP Array 6.0[8].

Imputation From Genotyped Reference Panel on Chromosome 9p21

I. Reference Panel

To create a genotyped reference panel for chromosome 9p21, we genotyped all validated sequencing SNPs in 168 individuals (56 parent-offspring trios) of European ancestry from the HapMap III CEU population[10]. HapMap III (Release 2) contained 137 SNPs polymorphic in the 168 individuals; we used HapMap III data for these sites. The final reference panel contained 464 SNPs across ~240kb on chromosome 9p21. A detailed list of SNPs in this reference panel is given in **Supplementary Table 5**. For **Figure 2** and **Supplementary Figure 8**, we evaluated the proportion of the 464 SNPs represented in the genotyped reference panel that were captured in the T2D and MI cohorts.

II. Phasing and Imputation

When imputing from the genotyped reference panel or HapMap II (**Fig. 2, Supplementary Fig. 8**), we used PHASE[11,12] (Version 2.1) to create phased haplotypes for the reference panel. We used trio information to inform phasing (-P1 option); all other PHASE parameters were default values. Imputation from reference haplotypes was performed using MACH[13,14] (Version 1.0.16). We used 100 rounds and the default values for other MACH parameters.

III. Identification of tag SNPs

Tagger[15] (pairwise tagging, r^2 threshold 0.8) was used to identify a set of tag SNPs to capture poorly imputed (MACH-estimated $r^2 < 0.8$) variants. These tags were genotyped in the entire T2D and MI cohorts. After quality control filtering, this added 94 SNPs to the T2D cohort (final number of SNPs = 145) and 73 SNPs to the MI cohort (final number of SNPs = 167). This corresponds to a marker density of ~1SNP/1.5kb across the ~240kb region. We note that we performed this tagging on an early version of our dataset that was subsequently revised as we improved our SNP caller (for calling variants in Illumina data). While these tags were designed with an earlier dataset, we find that they effectively tag the common variation in our final dataset (**Fig. 2e, f and Supplementary Fig. 8e, f**).

IV. Evaluation of tagging for additional GWAS arrays

To evaluate whether the improvement in imputation observed upon addition of tag SNPs was specific to Affymetrix arrays, we tested the effect of adding the same set of tag SNPs to additional commercial arrays. We considered, in addition to the Affymetrix 500K and 6.0 arrays, the Illumina HumanHap300, HumanHap500, HumanHap650Y, and Human1M arrays.

We imputed from our genotyped reference panel (n=464 SNPs in 112 CEU individuals from 56 trios) into 90 additional HapMap samples of European ancestry (from the HapMap III TSI population) which we had previously genotyped at the same set of 464 SNPs present in the genotyped reference panel. For these imputations, we downsampled the markers in the 90 TSI individuals to the set of SNPs present on each of the commercial GWAS arrays and evaluated the overall fraction of markers in the reference panel captured with a MACH-estimated r^2 of at least 0.8 (**Supplementary Fig. 9a**). Consistent with our observations for the Affymetrix arrays, intermediate frequency (MAF 2-5%) variants were largely not

imputed well, although – as expected – they were imputed better from the more dense GWAS chips (e.g. the Human IM). Adding the tag SNPs to the array SNPs (**Supplementary Fig. 9b**) resulted in improved imputation across all minor allele frequencies for all chips tested.

Imputation Directly from Illumina Sequencing Data

I. Methods tested

We tested several methods for imputing directly from high coverage re-sequencing data. In all cases, the reference panel consisted of only the 47 individuals sequenced. Imputation was performed into our T2D cohort, using GWAS genotypes in the T2D samples.

Haplotype creation with PHASE, imputation with MACH

As described above, we made individual genotype calls from the sequencing data at sites where the lod score ($\log_{10}$ of the ratio of the two highest genotype posterior probabilities) exceeded 3.0; we used these genotypes for imputation (singletons in the sequencing data were excluded).

Individual genotypes from sequencing data (n=220 non-singleton SNPs) were either phased on top of HapMap II haplotypes (Release 22, n=246 SNPs) (-k used to specify known phase at HapMap sites) or phased along with HapMap genotypes for the 47 sequenced individuals using PHASE. The HapMap haplotypes were created using trio information; phasing on top of these haplotypes has the advantage of preserving this information. Imputation from the reference haplotypes was performed using MACH (100 rounds, all other parameters were default values).

Likelihoods-based imputation with Beagle

Of course, information is lost in converting the genotype probabilities from the sequencing data into hard genotype calls. Most notably, there are many instances in which we do not make a genotype call due to an insufficient lod score. We therefore also tried likelihoods-based imputation with a newly available version of Beagle[16-18] (Version 3.1).

As described above, our SNP caller calculates the posterior probability of each genotype, given the observed sequencing data (P(Genotype|Data)). The genotype likelihood used in Beagle imputation is the probability of observing the sequencing data given a particular genotype at the site (P(Data|Genotype)), which we calculate with a re-arrangement of equation 5.1 above.

For these methods, we expressed HapMap genotypes (n=246) as likelihoods (P = 0.999 for the observed genotype and P = 0.001 for other possible genotypes) and integrated these with genotype likelihoods calculated from sequencing data at non-HapMap, non-singleton sites (n=220). We then used Beagle to impute directly from this likelihoods reference panel into T2D samples (single step procedure). We also tried using Beagle to first create phased reference haplotypes from the likelihoods reference panel and then to impute from the phased haplotypes into T2D samples (two step procedure).

II. "Truth" dataset and evaluation of method performance

Of the 466 SNPs in the reference panel, 53 were genotyped in the T2D samples, leaving 413 SNPs to be imputed. We used several metrics to evaluate imputation performance. First, both MACH and Beagle provide, for each SNP, an measure of imputation quality related to the estimated correlation between the imputed allele dosage and the true allele dosage; we refer to these values as the "estimated r^2". To enable empirical evaluation of imputation performance, we genotyped a sampling of the imputed SNPs (n=299) in a subset of our T2D samples (n=319 individuals). For each variant we calculated the squared correlation between the imputed dosage from imputation and the observed genotype. We refer to these values as the "empirical r^2".

We found that all methods performed comparably in terms of the fraction of variants captured by both estimated and empirical r^2 over a range of r^2 thresholds (**Supplementary Fig. 10a, b**), but noted that phasing the sequencing calls on top of known HapMap haplotypes gave slightly higher imputation accuracy at low-frequency sites (**Supplementary Fig. 10c**). We also found that the estimated r^2 values provided by MACH or Beagle for each imputation analysis were well calibrated to the empirical values (**Supplementary Fig. 11**). Because of the improved accuracy at low-frequency sites, we used the method of phasing sequencing genotypes on top of HapMap haplotypes and imputing with MACH for all

subsequent imputations from sequencing data (this method was used for **Fig. 3**).

III. Imputation from 1000 Genomes
Out of the 58 CEU founders sequenced in 1000G Pilot 1, 55 individuals had phased haplotypes available from HapMap III[10]. Non-singleton 1000G genotype calls (n=325 SNPs) for these 55 individuals were phased on top of HapMap haplotypes (n=124 SNPs) using PHASE and imputation was performed using MACH as described above.

List of Members of the Myocardial Infarction Genetics Consortium

Heart Attack Risk in Puget Sound (Seattle, Washington, USA): Stephen M Schwartz, David S Siscovick

Registre Gironi del Cor (Spain): Roberto Elosua

Massachusetts General Hospital Premature Coronary Artery Disease Study (Boston, Massachusetts, USA): Sekar Kathiresan

FINRISK (Finland): Veikko Salomaa

Malmö Diet and Cancer Study (Malmö, Sweden): Olle Melander

Broad Institute of Harvard and Massachusetts Institute of Technology (Cambridge, Massachusetts, USA): Benjamin F Voight, Sekar Kathiresan, David Altshuler

Supplementary References

1. Gnirke, A.*, et al.* Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol* **27**, 182-189 (2009).
2. Frazer, K.A.*, et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851-861 (2007).
3. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851-1858 (2008).
4. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691-692 (1991).
5. Depristo, M.A.*, et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498 (2011).
6. McKenna, A.*, et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).
7. Durbin, R.M.*, et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073 (2010).
8. Kathiresan, S.*, et al.* Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet* **41**, 334-341 (2009).
9. Saxena, R.*, et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331-1336 (2007).
10. Altshuler, D.M.*, et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52-58 (2010).
11. Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am J Hum Genet* **76**, 449-462 (2005).
12. Stephens, M., Smith, N.J. & Donnelly, P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**, 978-989 (2001).
13. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387-406 (2009).
14. Li, Y.a.A.G. MACH 1.0: rapid haplotype reconstructoin and missing genotype inference. *American Journal of Human Genetics* **S70**, 2290 (2006).

15. de Bakker, P.I.*, et al.* Efficiency and power in genetic association studies. *Nat Genet* **37**, 1217-1223 (2005).
16. Browning, B.L. & Browning, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**, 210-223 (2009).
17. Browning, B.L. & Yu, Z. Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet* **85**, 847-861 (2009).
18. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-1097 (2007).