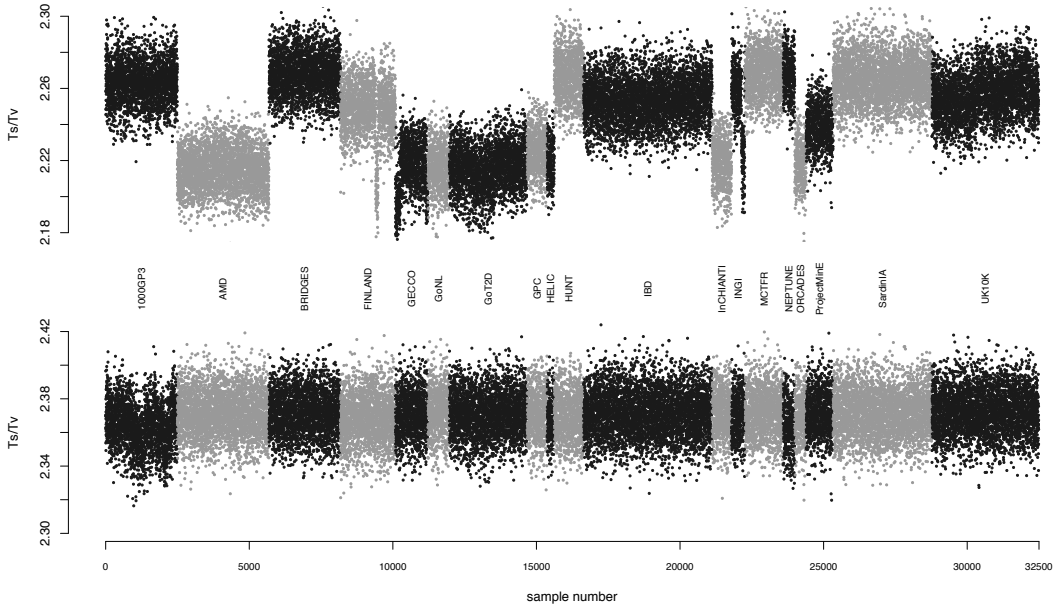
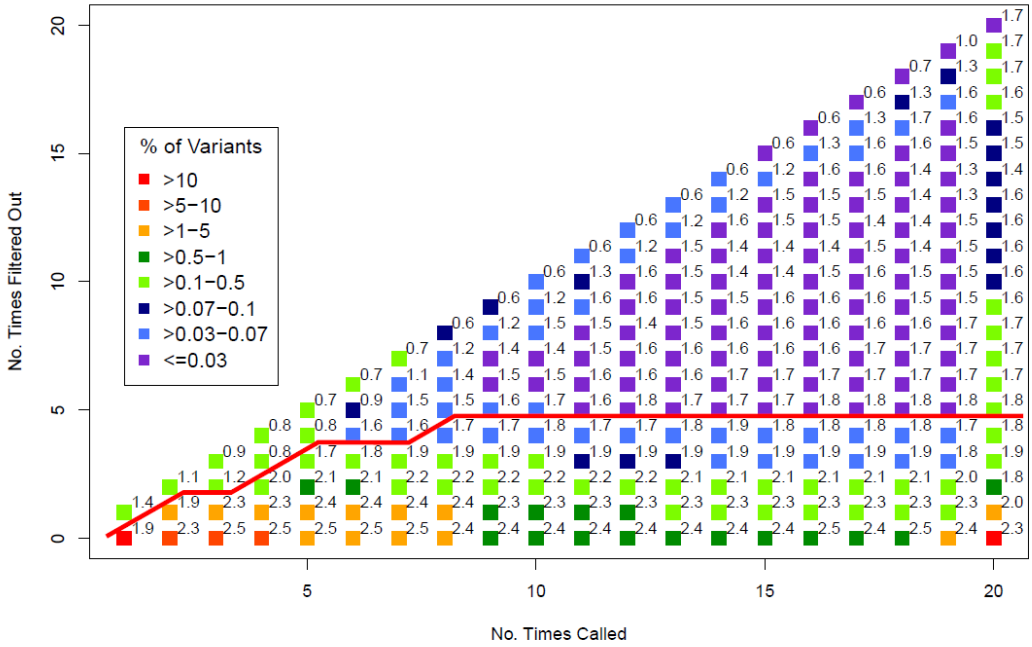
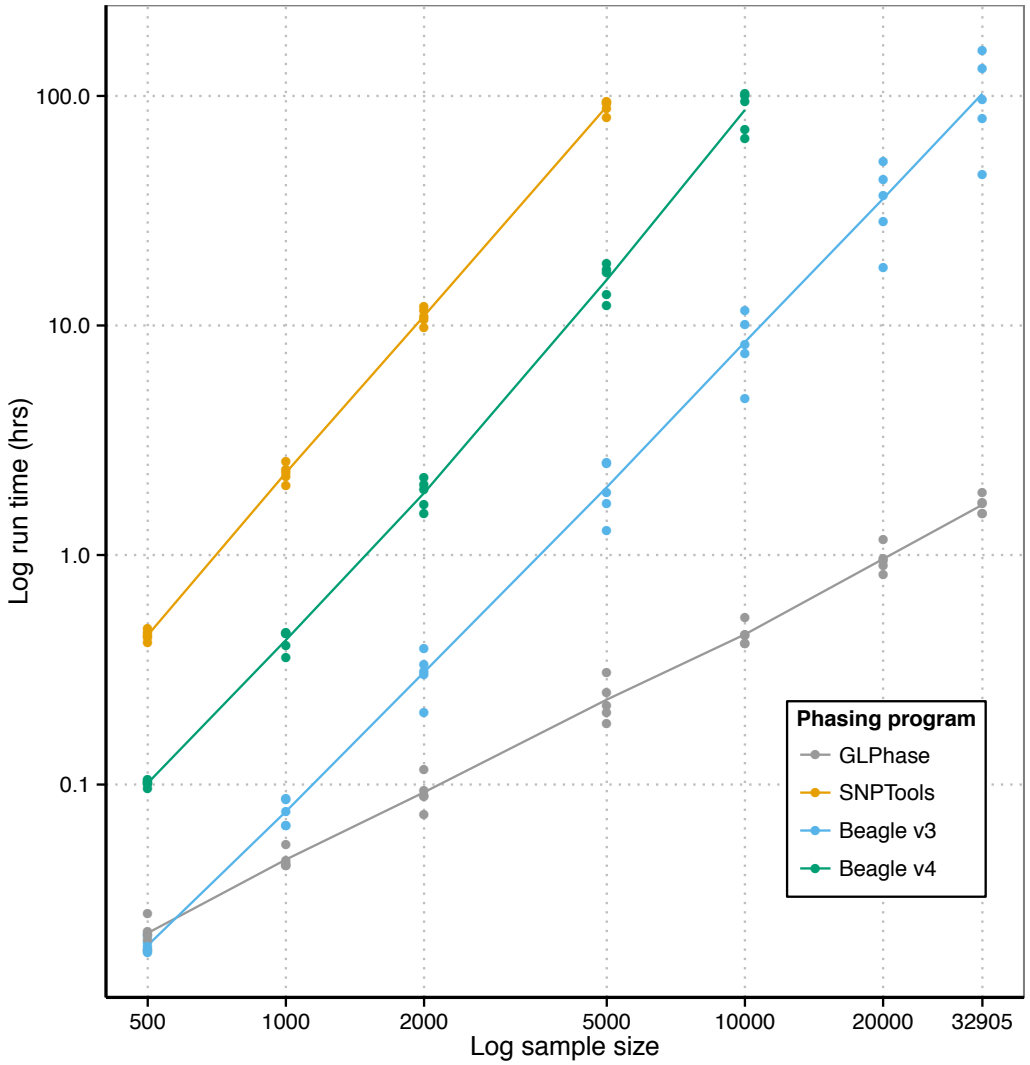
Supplementary Figures and Tables



Supplementary Figure 1 : The top figure shows the per-sample transition-transversion ratio (Ts/Tv) for chromosome 20 after running the GLPhase genotype calling method on the full MAC5 site list. In the bottom figure, GLPhase was run after the site filtering described in the text.



Supplementary Figure 2 : This figure shows sites stratified by their called and filtered status. On the x-axis we show the number of studies a variant was called in (out of 20) and on the y-axis we show the number of times it was filtered out by the cohort-specific internal QC pipelines. The color shows the percentage of variants in each such cell (red means more than 10% of variants lie in that cell while blue means less than 0.1%). The number to the top right of each cell denotes the Ts/Tv ratio for all sites in that cell. Cells higher in the plot have been filtered out relatively often and usually represent poor variants, as is also seen from the low Ts/Tv ratio. All variants above the red line were filtered out (which excludes all cells which had been filtered independently by more than 4 studies or have Ts/Tv ratio less than 1.7)



Supplementary Figure 3 : Comparison of methods for genotype calling as sample size increases. The figure shows a log-log plot of run time vs sample size for four different methods of genotype calling from GL data. For each sample size 5 random 1024 site chunks from chromosome 20 were used. Each dot represents the run time of a single dataset. Lines are drawn between successive means of run times for each value of sample size.

Supplementary Table 1 : Table detailing studies that contributed datasets to the HRC. Details of study size, sequencing depth, webpages of studies and primary references are given. **(see SuppTable1.xlsx**)

Sites	**Samples**	REF/REF	REF/ALT	ALT/ALT	NRD
1000GP3	**1000GP3 (N=2,525)**	0.10	0.61	0.43	0.70
	HRC Pilot (N=13,309)	0.07	0.36	0.27	0.43
	HRC full (N=32,905)	0.06	0.34	0.25	0.41
HRC MAC5	**1000GP3 (N=2,525)**	0.10	0.59	0.40	0.67
	HRC Pilot (N=13,309)	0.06	0.38	0.26	0.43
	HRC full (N=32,905)	0.06	0.36	0.24	0.41
HRC MAC5 + site filters	**1000GP3 (N=2,525)**	0.10	0.59	0.40	0.67
	HRC Pilot (N=13,309)	0.06	0.36	0.25	0.42
	HRC full (N=32,905)	0.06	0.34	0.23	0.39

Supplementary Table 2: Evaluation of genotype calling process. The table reports discordance rates of genotypes called using different sites lists (Sites column) and sample sets (Samples column). NRD = non reference allele discordance percentage.

Variants present in HRC panel so HRC MAF can be determined (affects 1000G counts relative to all 1000G variants passing QC)					
Imputation Panel	HRC Minor Allele Frequency (minus 1000G Samples)				Total
	MAC >= 0 & <1%	MAC >0 & <1%	MAF ≥1% to <5%	MAF ≥5%	
1000 Genomes Phase 3 v5	3,488,324	3,461,032	1,930,001	5,309,997	10,728,322
HRC v1	7,833,143	7,817,460	2,243,365	5,425,008	15,501,516

The total number of SNPs in the InCHIANTI study passing imputation quality (≥0.5) broken down by allele frequency bins defined by the HRC (after excluding 1000 Genomes Phase 3 (v5) samples).

Variants present in 1000G Phase 3 (v5) AND HRC Panels (Michigan server) (affects HRC counts) : intersection N = 28,106,873					
Imputation Panel	HRC Minor Allele Frequency (minus 1000G Samples)				Total
	MAC >= 0 & <1%	MAC >0 & <1%	MAF ≥1% to <5%	MAF ≥5%	
1000 Genomes Phase 3 v5	3,488,324	3,461,032	1,930,001	5,309,997	10,728,322
HRC v1	5,734,139	5,718,467	2,231,559	5,399,097	13,364,795

The total number of SNPs in the InCHIANTI study imputed from both imputation panels passing imputation quality (≥0.5) broken down by allele frequency bins defined by the HRC (after excluding 1000 Genomes Phase 3 (v5) samples).

Supplementary Table 3 : Summary of imputed variants in the InCHIANTI study.

Trait	SNP chr:bp (build 37)	Effect Size	SE	P-value	InCHIANTI MAF	Nearest Gene	1000G P-value	HRC Imputation	1000G imputation

		(SD)						Quality	Quality
Lactic Dehydrogenase	3:52551566	2.37	0.29	$6x10^{-16}$	0.006	*STAB1*	$3x10^{-05}$	0.79	0.74
Magnesium	17:9689132	2.26	0.38	$4.5x10^{-09}$	0.004	*DHRS7C*	NA	0.67	0.01
Resistin	16:17685354	1.36	0.24	$1.9x10^{-08}$	0.009	*XYLT1*	$1.2x10^{-06}$	0.9	0.79
Free Thyroxine (FT4)	9:7092298	3.26	0.58	$2x10^{-08}$	0.002	*KDM4C*	NA	0.84	NA
Vitamin D	19:11599979	1.47	0.19	$2.8x10^{-08}$	0.013	*ZNF653*	$1.4x10^{-05}$	0.9	0.59
Retinol	11:19870163	1.79	0.32	$3x10^{-08}$	0.004	*NAV2*	NA	0.94	NA
	5:129871638	2.8	0.5	$3.6x10^{-08}$	0.002	*CHSY3*	NA	0.79	0.41
β-Globulins	15:100810864	1.71	0.31	$3.6x10^{-08}$	0.005	*ADAMTS17*	$2x10^{-07}$	0.81	0.53
Potassium	11:102811514	1.43	0.26	$3.7x10^{-08}$	0.006	*MMP13*	$1.2x10^{-06}$	0.94	0.77

Supplementary Table 4 : Potentially novel associations in the InCHIANTI study found by imputing SNPs using HRC release 1. The table lists the trait being tested and summary information about the SNP and the association signal.
(see SuppTable4.xlsx)

Imputation Panel	Effect size (SDs)	p-value	Imputation quality
HRC release 1	**-2.07**	**$2x10^{-13}$**	**0.98**
1000GP3	**-1.86**	**$3x10^{-11}$**	**0.87**
HapMap2	**-**	**-**	**-**

Supplementary Table 5 : Summary of association signal at SNP rs28929474 using imputed data from 3 different reference panels. The SNP does not exist in the HapMap2 panel.

Site List	Source/Type
Human CNV370-Quad	Chip (Illumina)
Human 660W-Quad	Chip (Illumina)
Human OmniExpress	Chip (Illumina)
Human Omni1-Quad	Chip (Illumina)
Human Omni5	Chip (Illumina)
Human CoreExome	Chip (Illumina)
Metabochip	Chip (Illumina)
Affy_Genome-Wide Human SNP Nsp/Sty	Chip (Affy)
Affy Genome-Wide Human SNP Array 6.0	Chip (Affy)
UK Biobank	Biobank Study
GIANT Consortium	GWAS Study
Global Lipids Genetics Consortium	GWAS Study

Supplementary Table 6 : List of commercial genotyping arrays and study lists used to add SNPs back in after site filtering.

Number of samples	Studies		Removal choice
5	AMD	AMD	Removed the duplicates randomly.
36	IBD	IBD	These were duplicates between Crohns and UC studies. Removed the duplicates randomly.
5	FINLAND	FINLAND	Removed lower coverage (4x) Kuusamo samples in preference to the higher coverage (6x) SiSu samples.
14	GECCO	GECCO	Removed the duplicates randomly.
17	GoT2D	FINLAND	Removed the FINLAND samples.
79	GoT2D	UK10K	Removed the UK10K/UK10K duplicates, then removed randomly otherwise.
34	GPC	BRIDGES	Removed the GPC samples.
32	GPC	GPC	Removed the duplicates randomly.
14	MCTFR	MCTFR	Removed the duplicates randomly.
1	NEPTUNE	NEPTUNE	Removed the duplicates randomly.
1	ORCADES	1000GP3	Removed the ORCADES sample.
1	ProjectMinE	GoNL	Removed the ProjectMinE sample.
1	ProjectMinE	ProjectMinE	Removed the duplicates randomly.
3	SardinIA	SardinIA	Removed the duplicates randomly.
26	UK10K	UK10K	These were monozygotic twins already identified by UK10K. Removed based on a list from UK10K of samples they had already excluded from downstream analysis.

Supplementary Table 7 : Details of duplicate removal. Each row of the table details the number of duplicate pairs found within and between studies together with the method by which duplicates were removed.