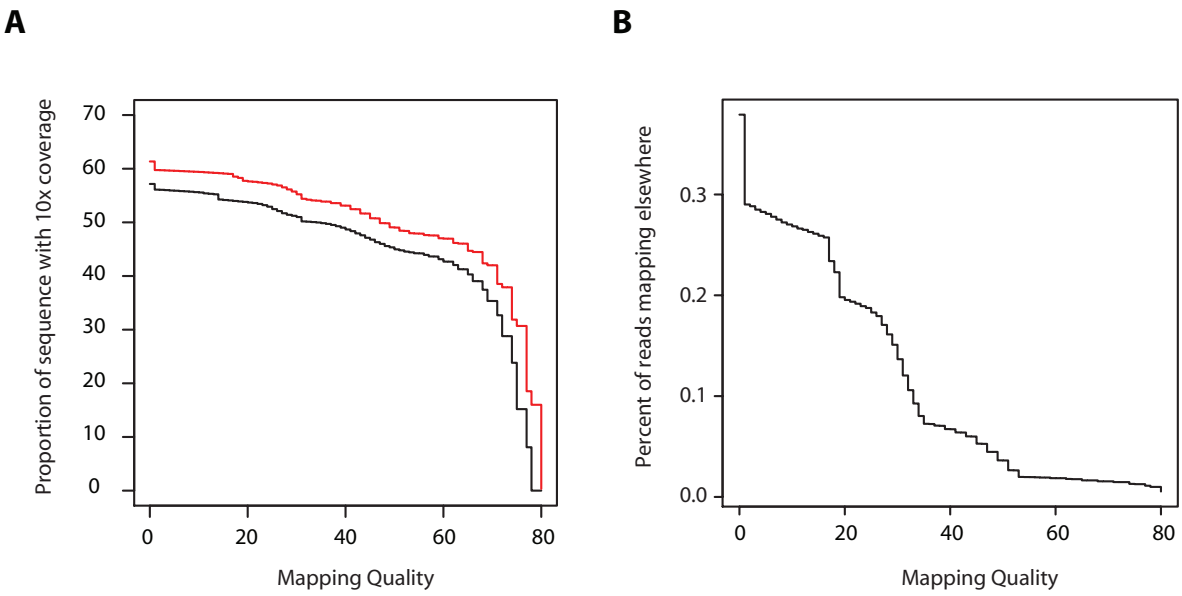
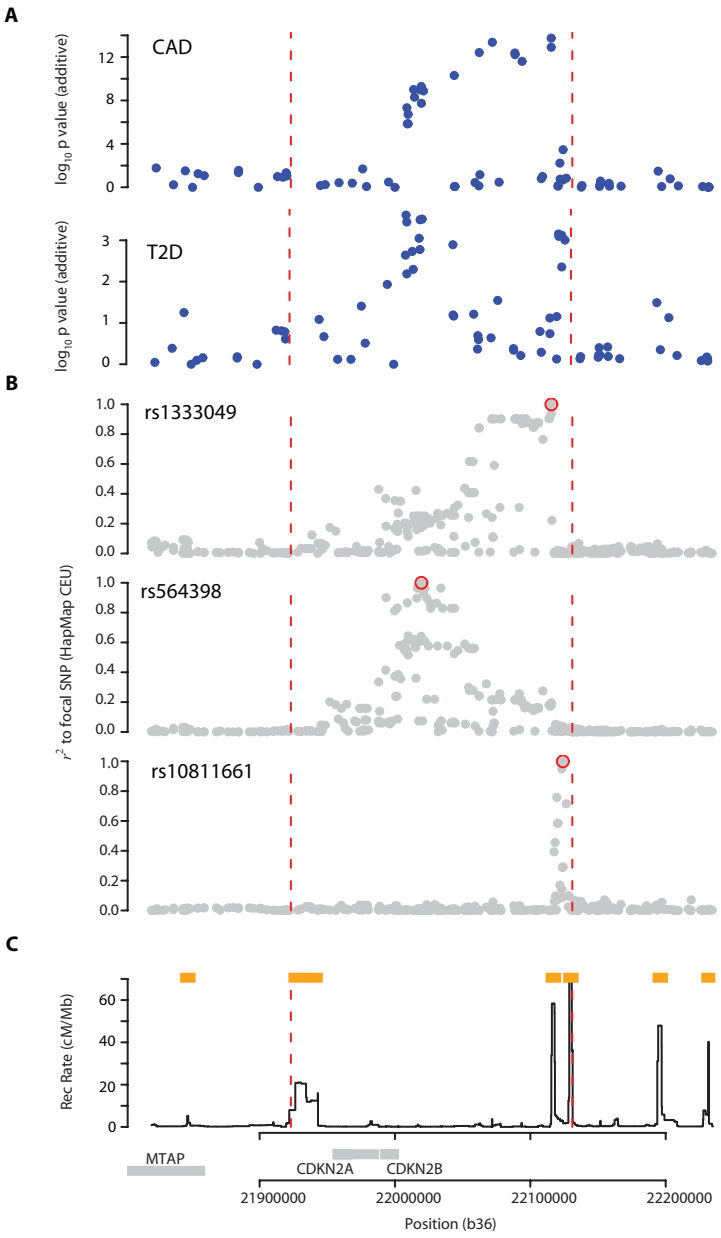
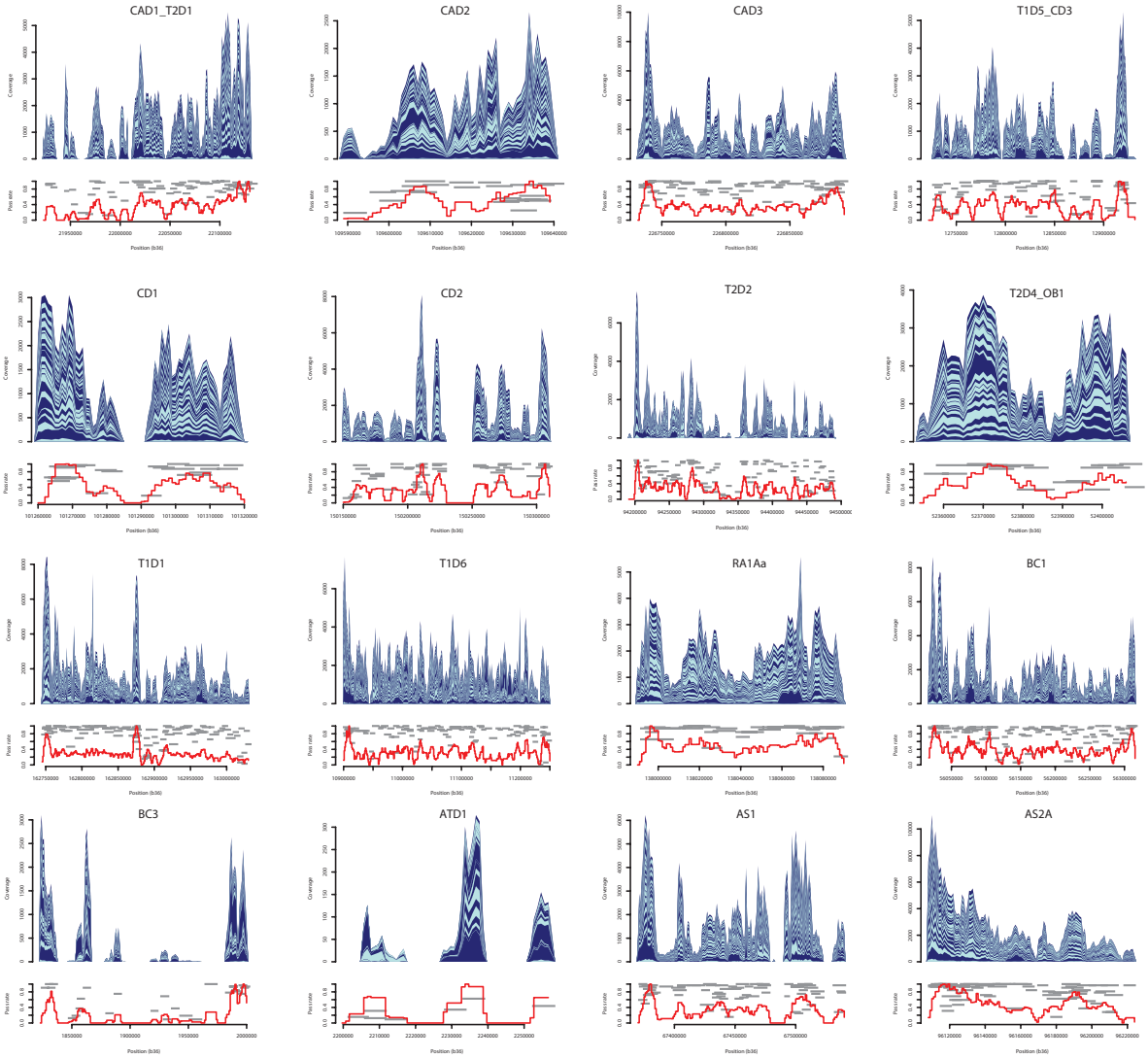


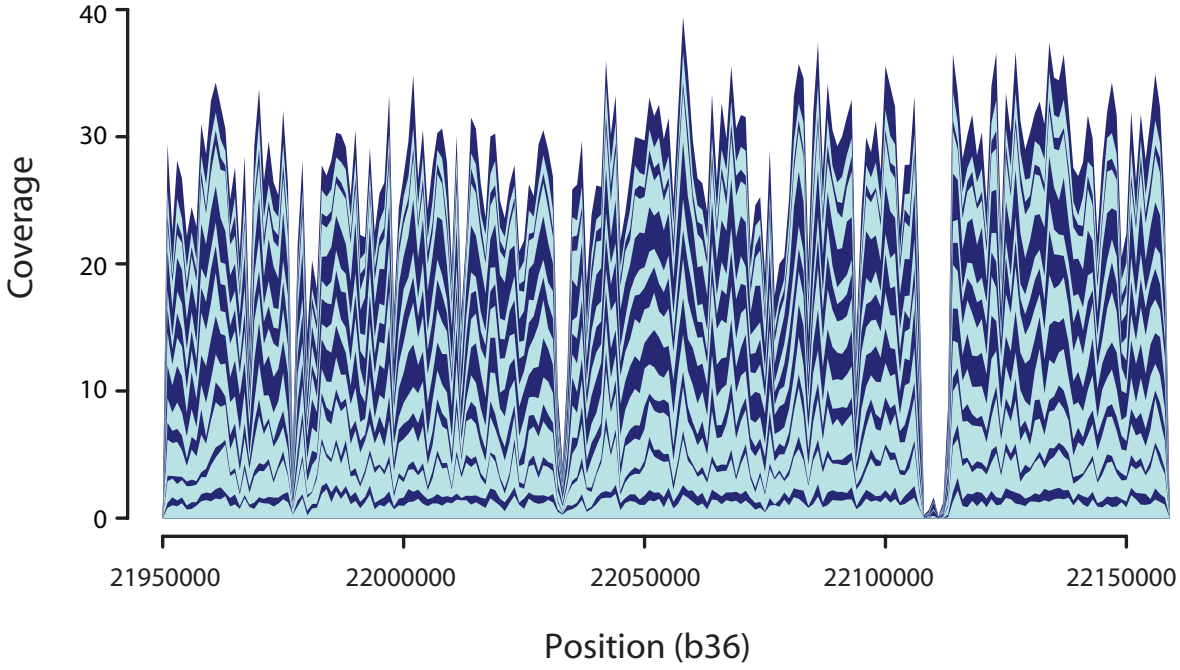
Supplementary Figure 1: Comparison of local versus global mapping. A) Mapping of reads to only those parts of the human genome contained within the amplicons leads to a roughly 10% increase in the fraction of the target regions that achieve at least 10x coverage across all categories of mapping quality (global mapping: black, local mapping: red). B) Chart showing the fraction of reads for which a position of higher mapping quality can be found outside the target regions than within, for different thresholds on local mapping quality. If the local mapping quality is 40 or higher, about 1 in 1500 reads map better elsewhere in the genome than within the targeted regions (only considering reads with non-zero mapping quality).



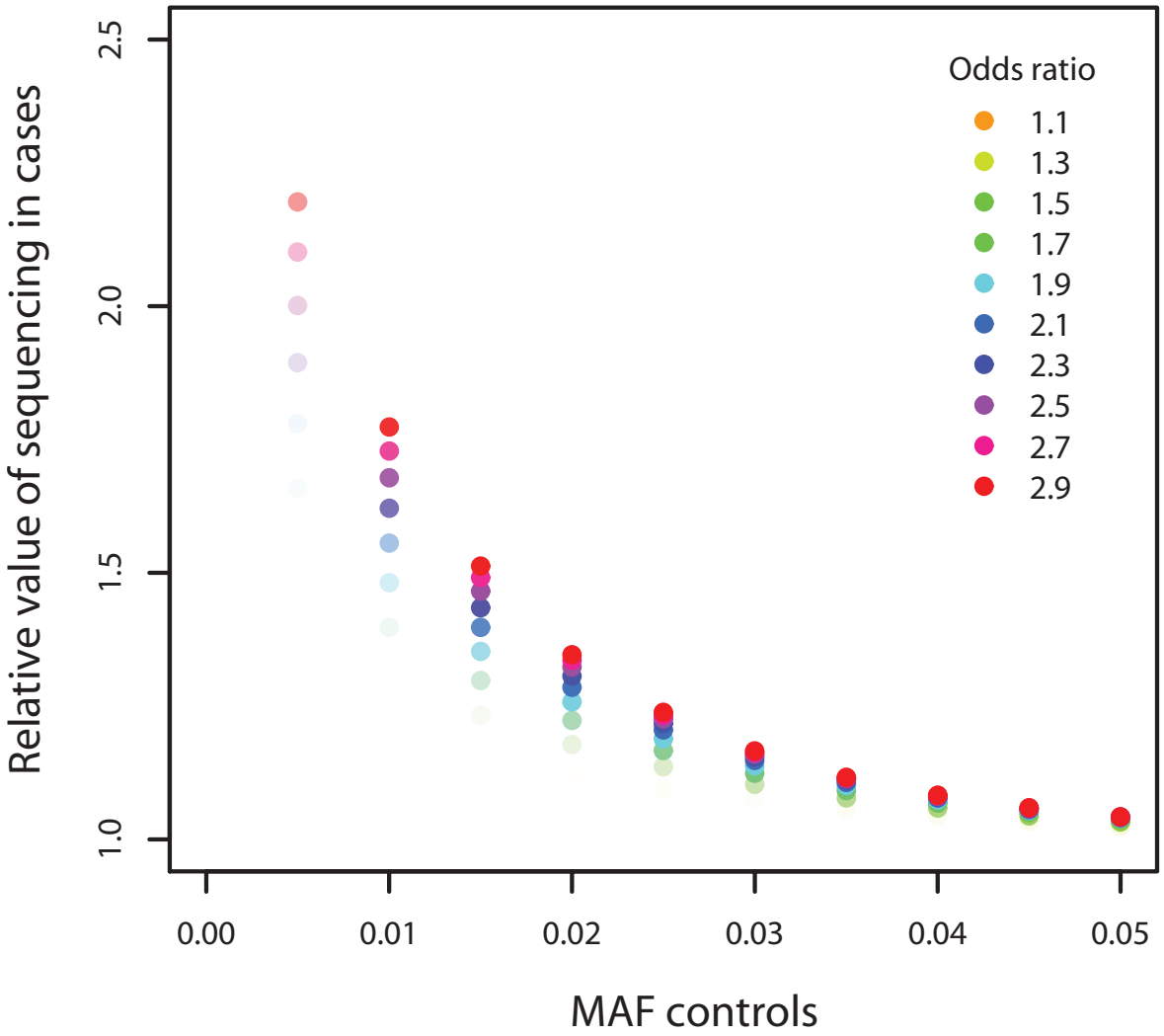
Supplementary Figure 2: A) Signal plots, B) LD to focal SNPs (as estimated from the CEU panel in HapMap) and C) fine-scale recombination rates used to define region boundaries for one of the 16 regions sequenced (the CDKN2A/CDKN2B gene region on chromosome 9 showing association to both CAD and T2D). The region chosen for sequencing is represented by the red vertical lines. In part B, the red circle indicates the focal SNP (rs1333049 for CAD and rs564398 and rs10811661 for T2D).



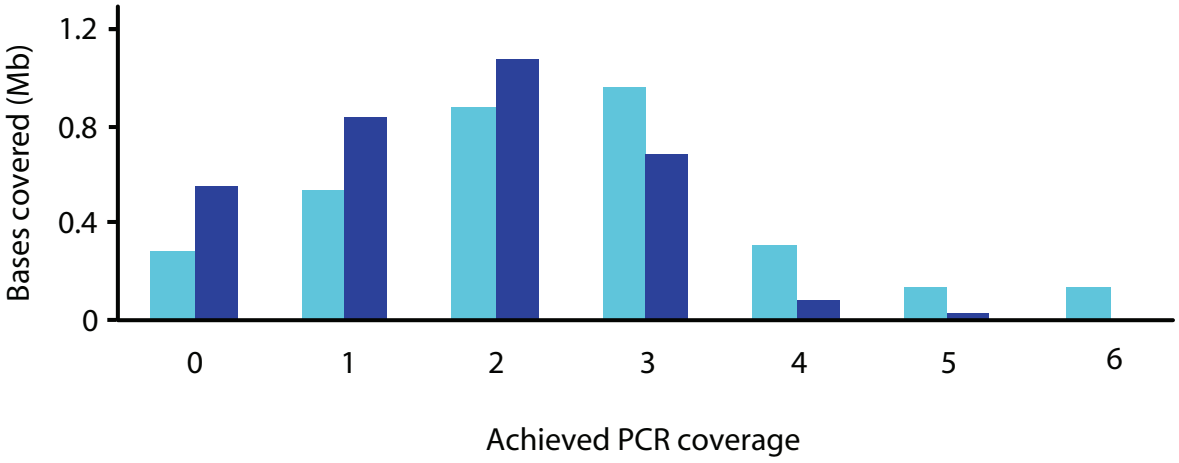
Supplementary Figure 3: Stacked charts showing the average per base coverage of confidently-mapped reads for each individual in 1kb windows across the 16 regions studied. The lower portion of each chart shows the successful PCR amplicons (indicated by the grey bars) and the proportion of each amplicons deemed successful across individuals (indicated by the vertical position). Combing the success rates for each amplicon generates a predicted relative coverage for each 1kb window (red line).



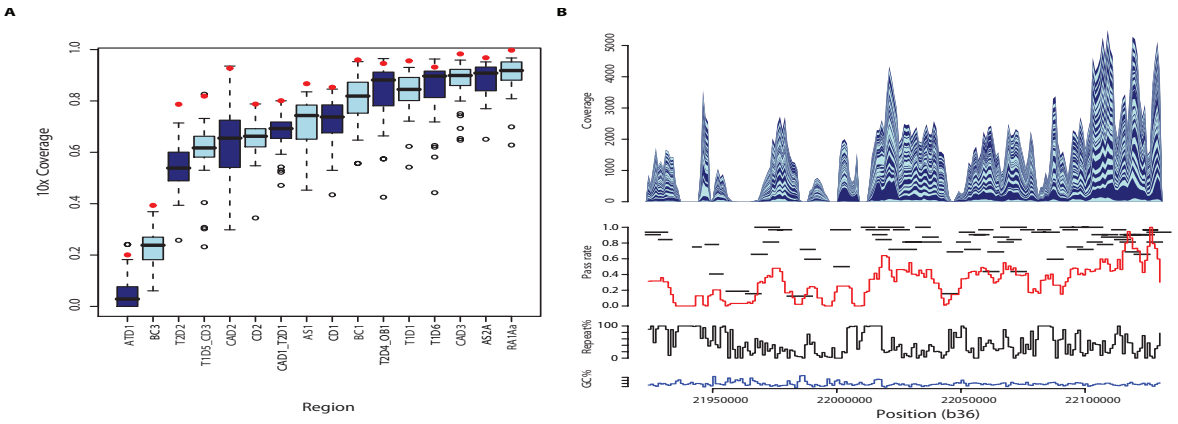
Supplementary Figure 4: Coverage of the chromosome 9 CAD/T2D hit region in 16 individuals from the CEU panel sequenced as part of the 1000 Genomes Project. The stacked area chart shows the coverage for each individual (indicated by the stripes) in 1kb bins. Only data from the Illumina technology (Freeze 3) are shown. Plots for all regions can be found in Figure 12.



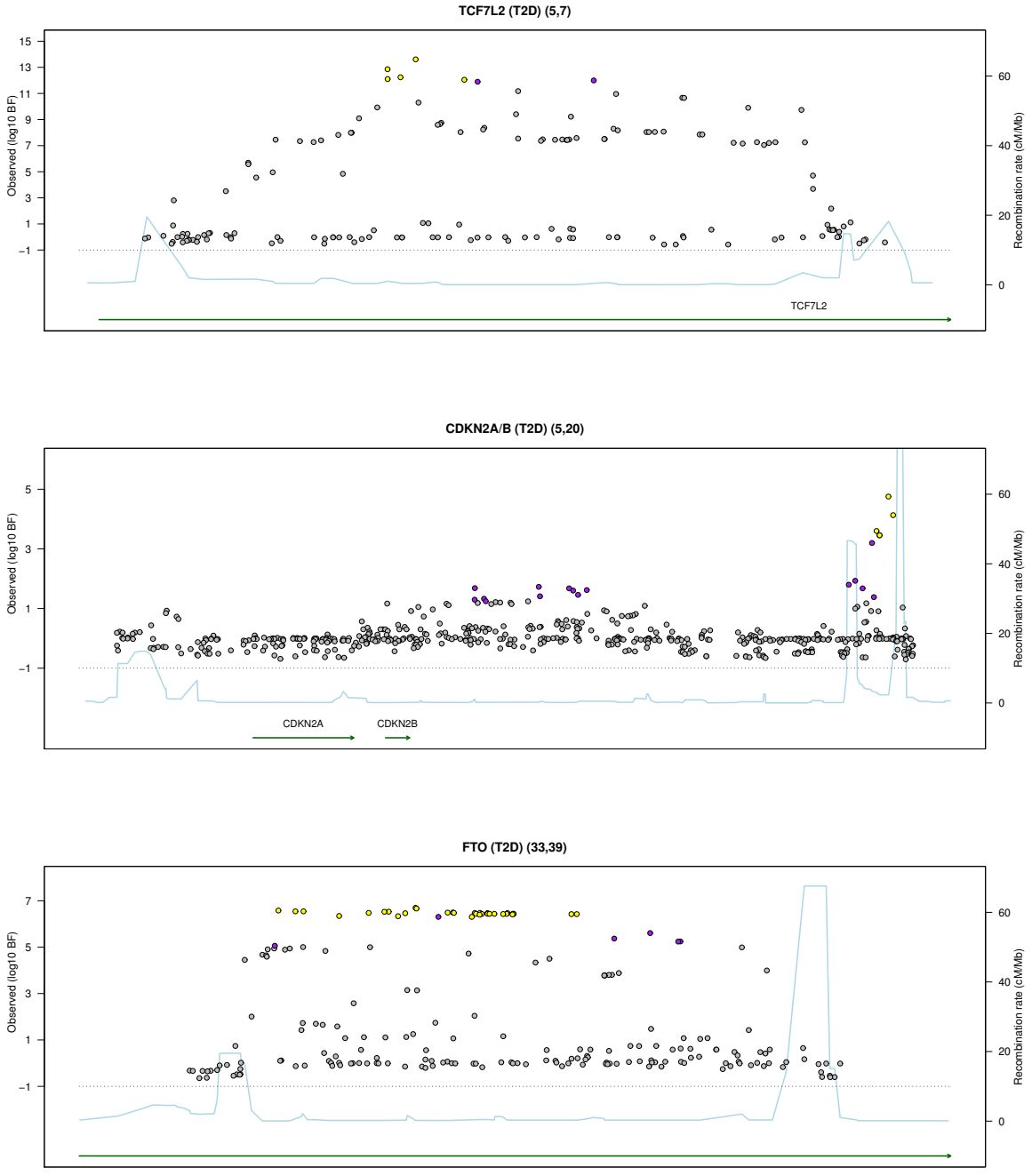
Supplementary Figure 5: The relative value of sequencing in cases compared to controls under disease models and allele frequencies of the risk variant. The relative value is defined by the ratio of the probability of at least one copy of the risk allele being present in a set of 32 randomly sampled cases to the same probability for 32 individuals sampled from the population (i.e. phenotype unknown). The intensity of points reflects the joint probability of identifying the variant in 32 cases and the variant being demonstrated to have a significant association (Cochran-Armitage trend test: $P<10^{-5}$) to disease status in a study of 2000 cases and 2000 controls. Consequently, there is a narrow range of disease models where there is both substantial power to detect an association and also value in sequencing controls as opposed to cases.

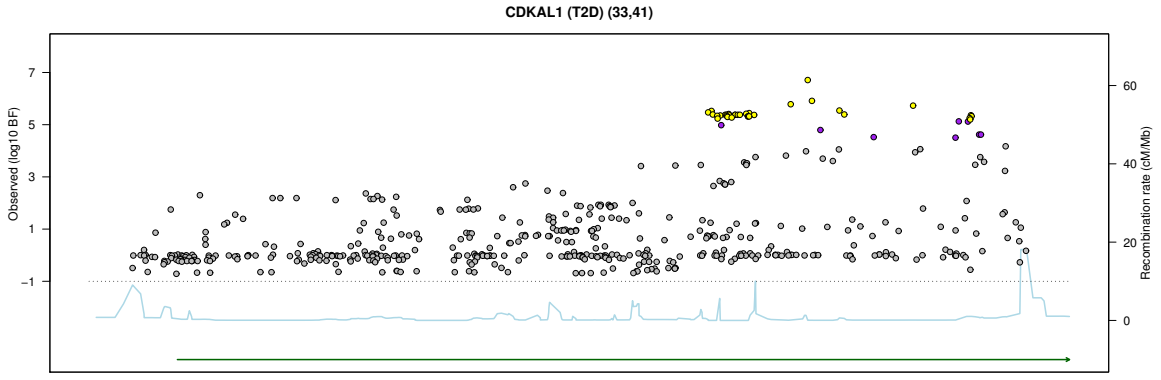


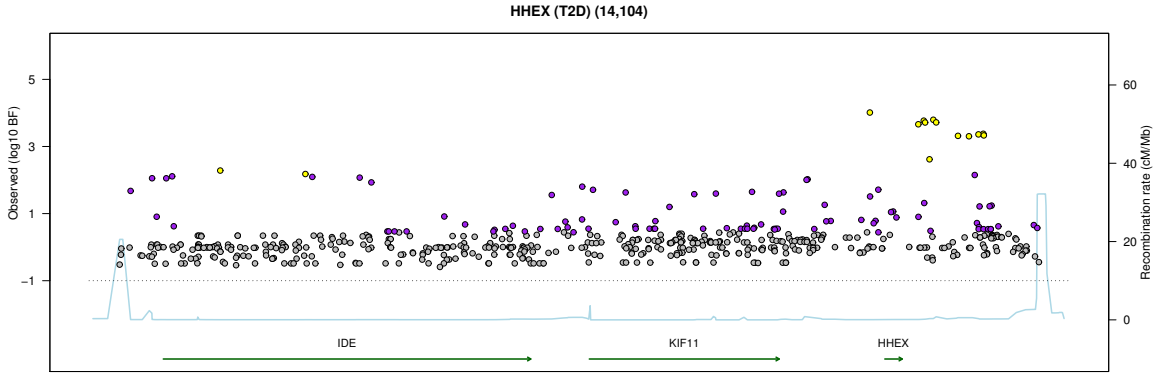
Supplementary Figure 6: The effect of overdesigning PCR amplicons on achieved coverage. The chart shows that initially starting with 5-fold PCR coverage (light blue bars) leads to an approximately 50% reduction in the number of bases that have zero coverage compared to initially starting with 3-fold PCR coverage (dark blue bars).

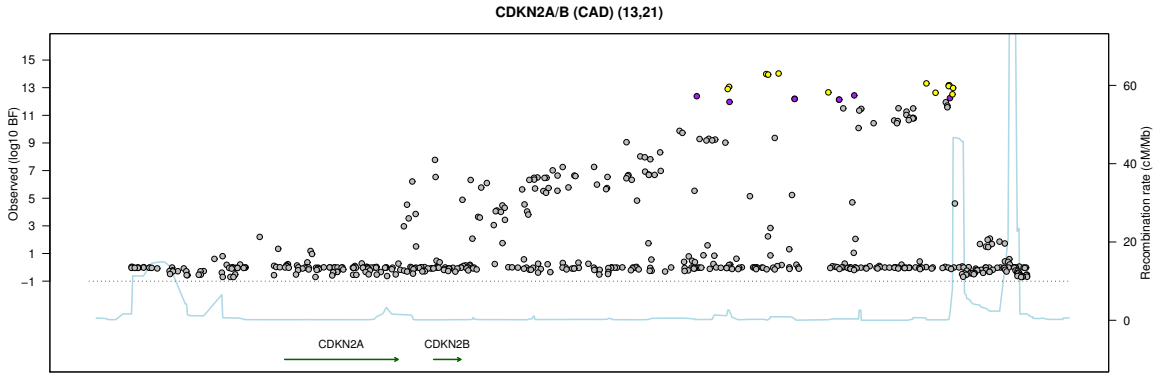


Supplementary Figure 7: A) Boxplots summarising the distribution of sequencing coverage across individuals for each region. Coverage is measured as the fraction of nucleotides within a given region that achieve at least 10x coverage of confidently-mapped reads (Q40). Red points indicate the maximum possible coverage achieved from the PCRs deemed successful. Note that in a few cases achieved coverage exceeds this maximum, presumably because some apparently failed PCRs were amplifying at a low level. B) Stacked chart showing the average per base per coverage of confidently-mapped reads for each individual in 1kb windows across the CDKN2A/CDKN2B region on chromosome 9. The lower portion of the chart shows the successful PCR amplicons (indicated by the black bars) and the proportion of each amplicons deemed successful across individuals (indicated by the vertical position). Combing the success rates for each amplicon generates a predicted relative coverage for each 1kb window (red line).

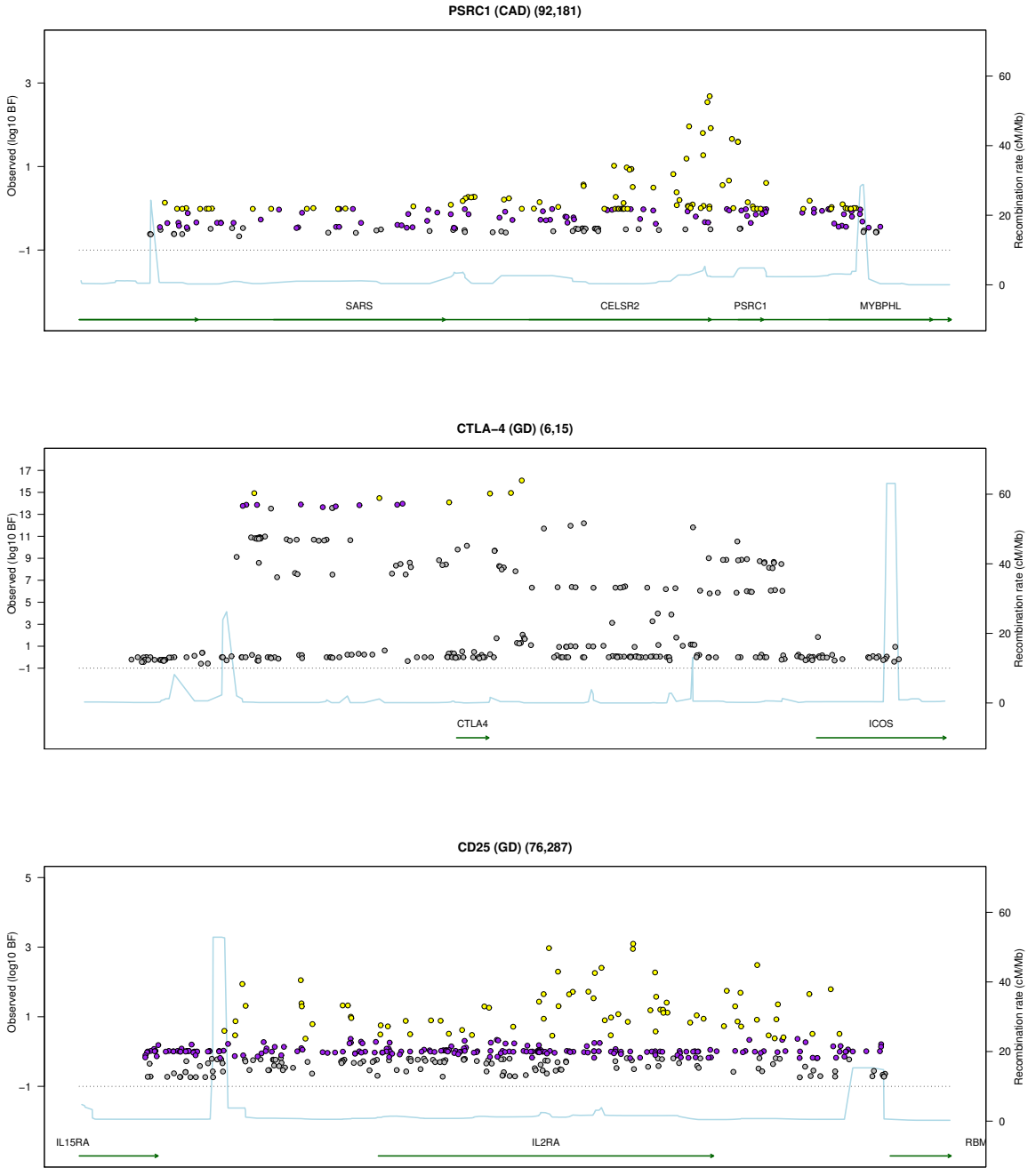

TCF7L2 (T2D) (5,7)
Observed (log10 BF)
Recombination rate (cM/Mb)
TCF7L2
CDKN2A/B (T2D) (5,20)
Observed (log10 BF)
Recombination rate (cM/Mb)
CDKN2A
CDKN2B
FTO (T2D) (33,39)
Observed (log10 BF)
Recombination rate (cM/Mb)

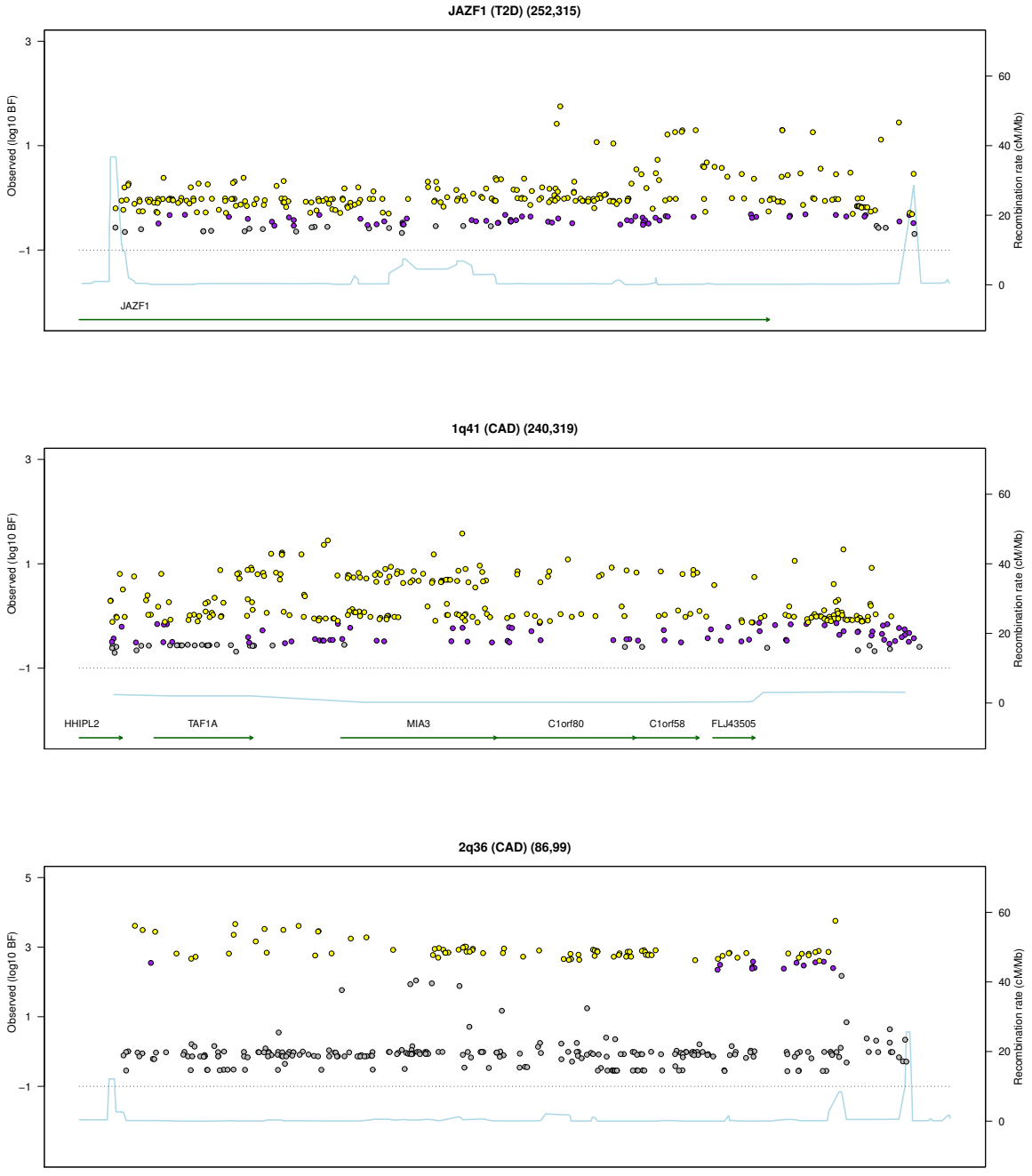


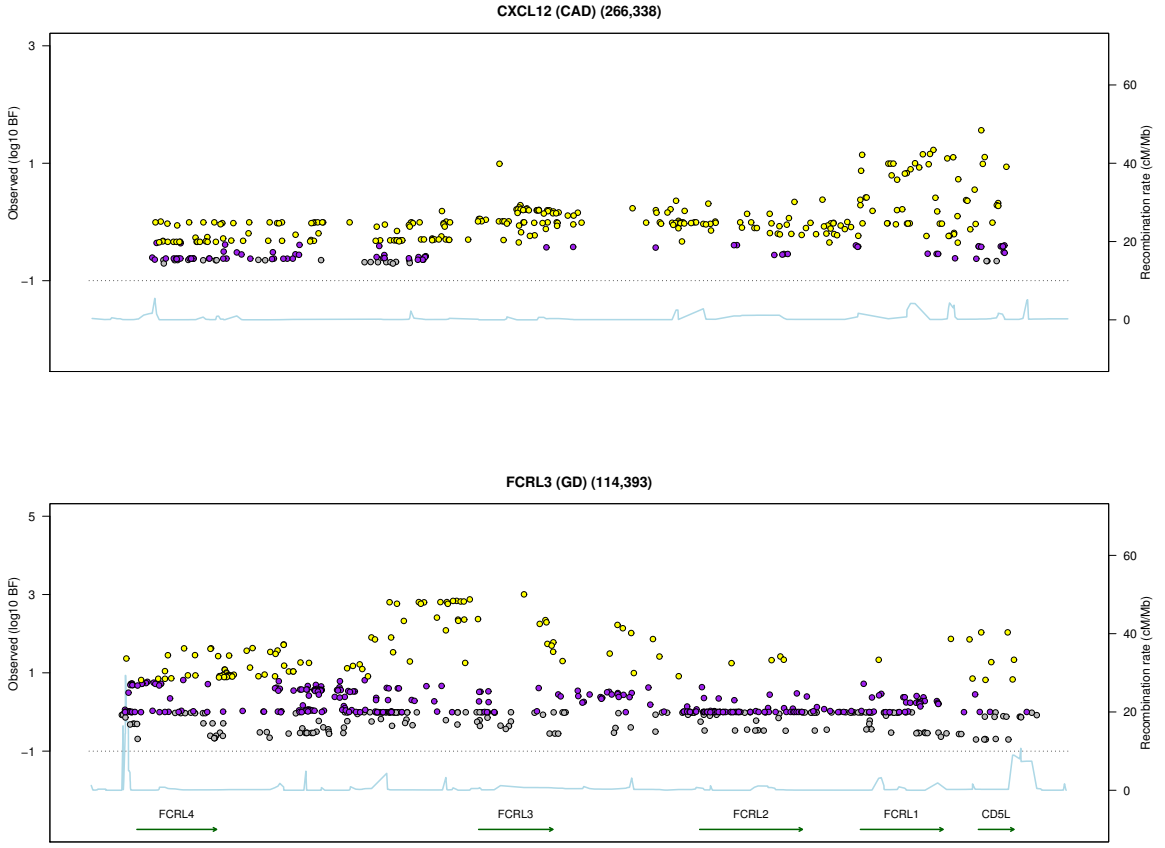




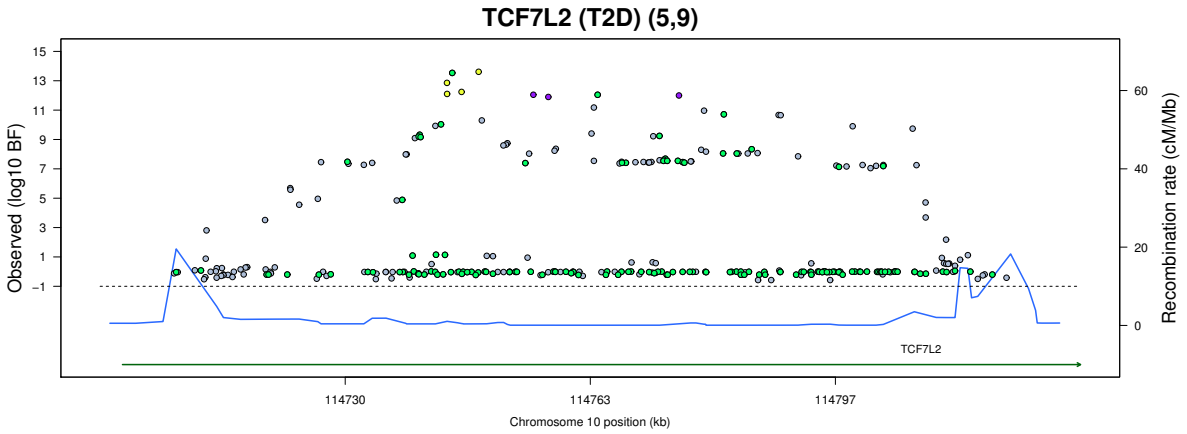
Nature Genetics: doi:10.1038/ng.2435

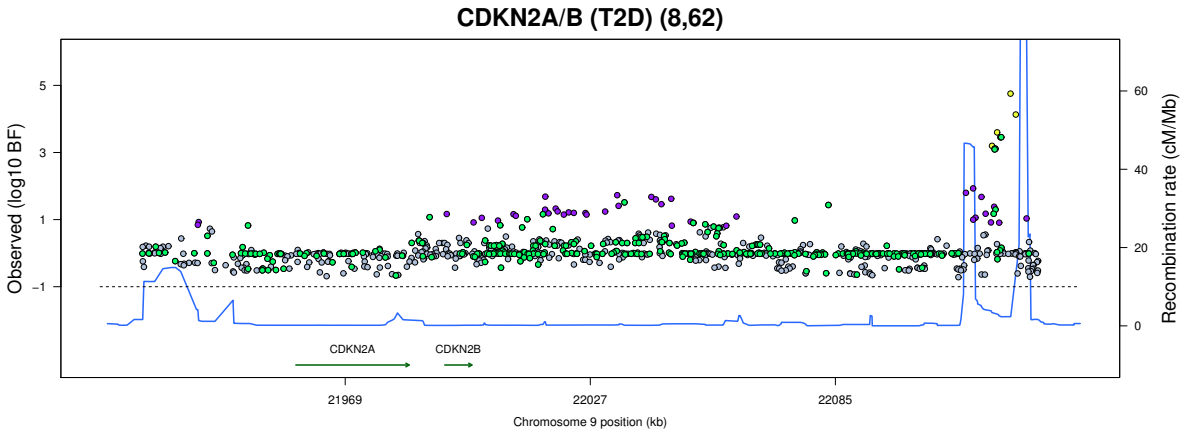

PSRC1 (CAD) (92,181)
Observed (log10 BF)
Recombination rate (cM/Mb)
SARS
CELSR2
PSRC1
MYBPHL
CTLA−4 (GD) (6,15)
CTLA4
ICOS
CD25 (GD) (76,287)
IL15RA
IL2RA
RBM

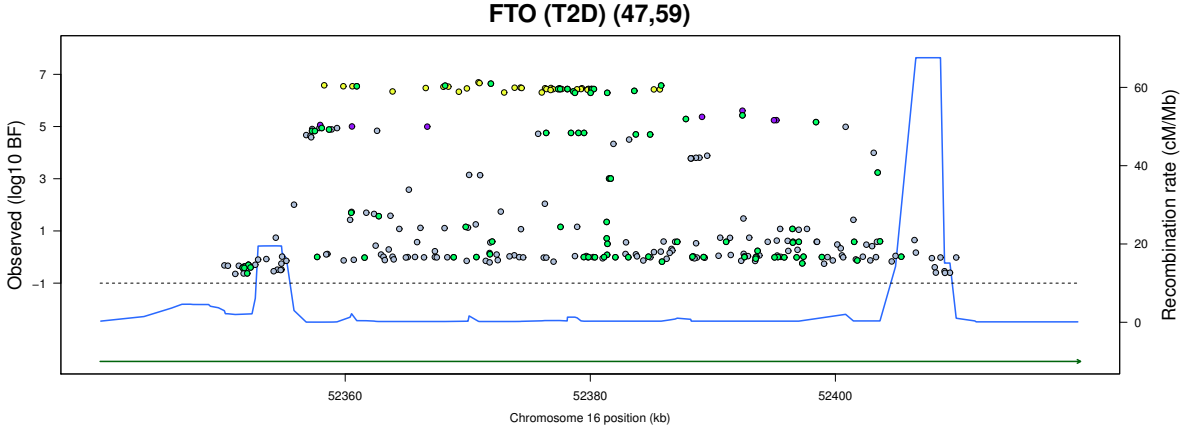

JAZF1 (T2D) (252,315)
Observed (log10 BF)
Recombination rate (cM/Mb)
JAZF1
1q41 (CAD) (240,319)
HHIPL2
TAF1A
MIA3
C1orf80
C1orf58
FLJ43505
2q36 (CAD) (86,99)

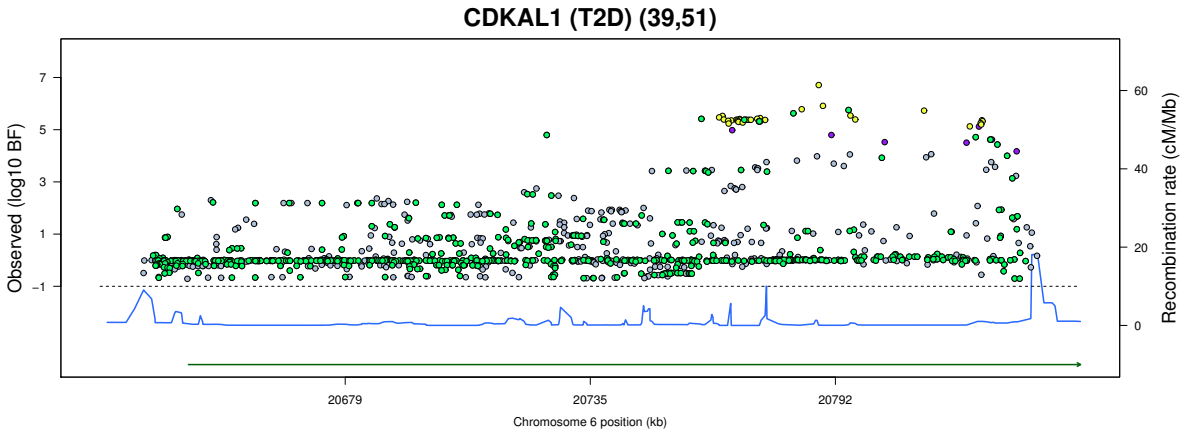


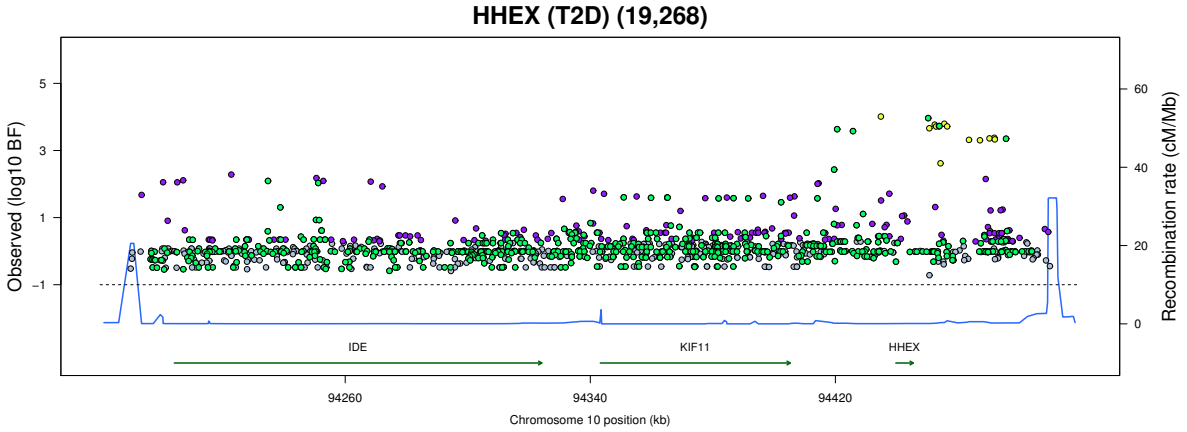
Supplementary Figure 8: Signal plots showing the strength of signal for each region, shown as $\log_{10}$ Bayes Factors for SNPs passing quality control. Estimated recombination rate is shown in blue, with scale on the right vertical axis. SNPs are coloured according to membership in credible sets: yellow for 95% credible set, purple for 99%, and grey otherwise. Genes in the region are shown towards the bottom in green. The title for each plot shows region name, follow by phenotype in parentheses, followed by credible set size (95,99) in parentheses. Genomic positions are from NCBI build 36.

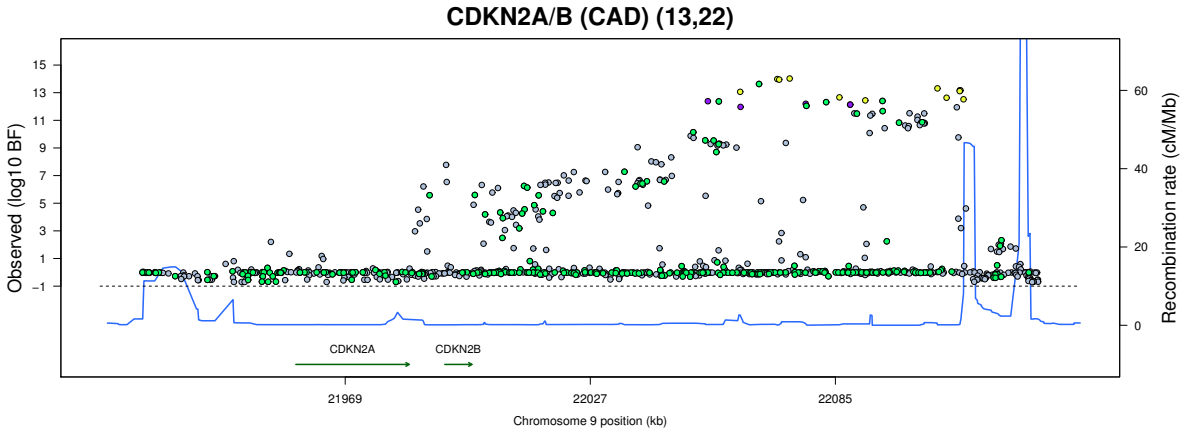

TCF7L2 (T2D) (5,9)
Observed (log10 BF)
Recombination rate (cM/Mb)
TCF7L2
114730
114763
114797
Chromosome 10 position (kb)


CDKN2A/B (T2D) (8,62)
Observed (log10 BF)
Recombination rate (cM/Mb)
CDKN2A
CDKN2B
21969
22027
22085
Chromosome 9 position (kb)

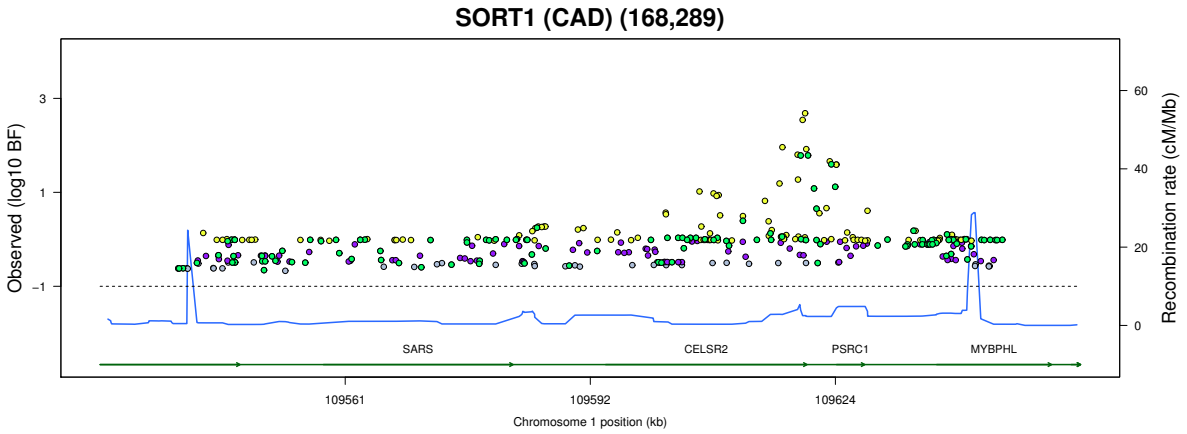

FTO (T2D) (47,59)
Observed (log10 BF)
Recombination rate (cM/Mb)
52360
52380
52400
Chromosome 16 position (kb)

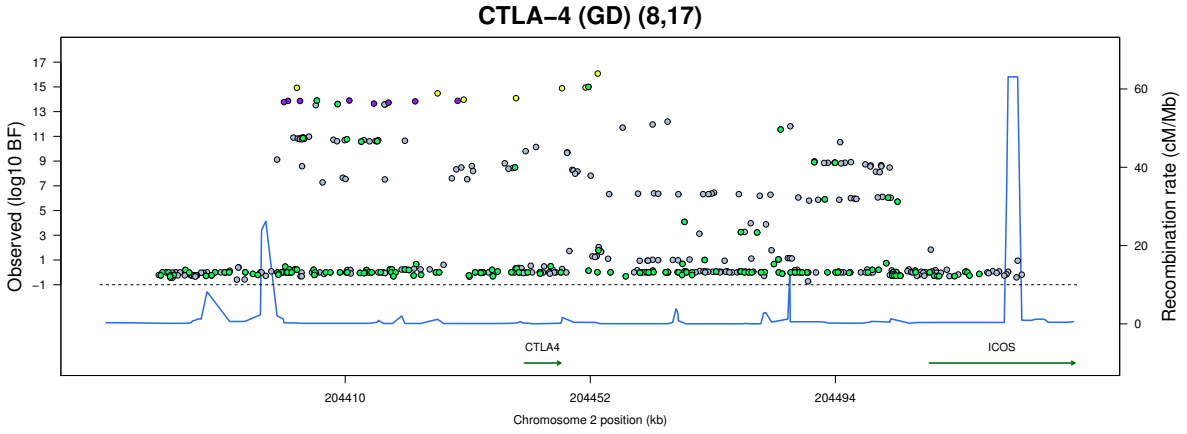


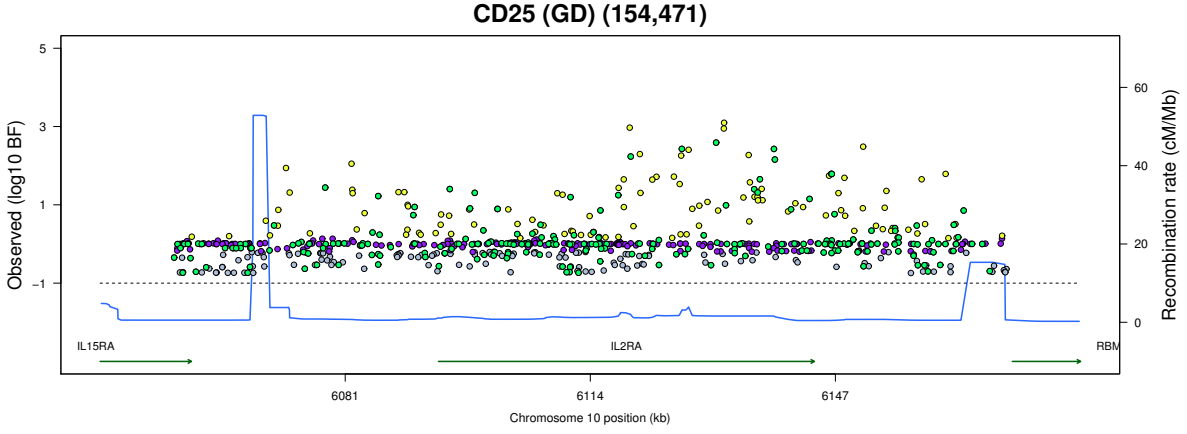


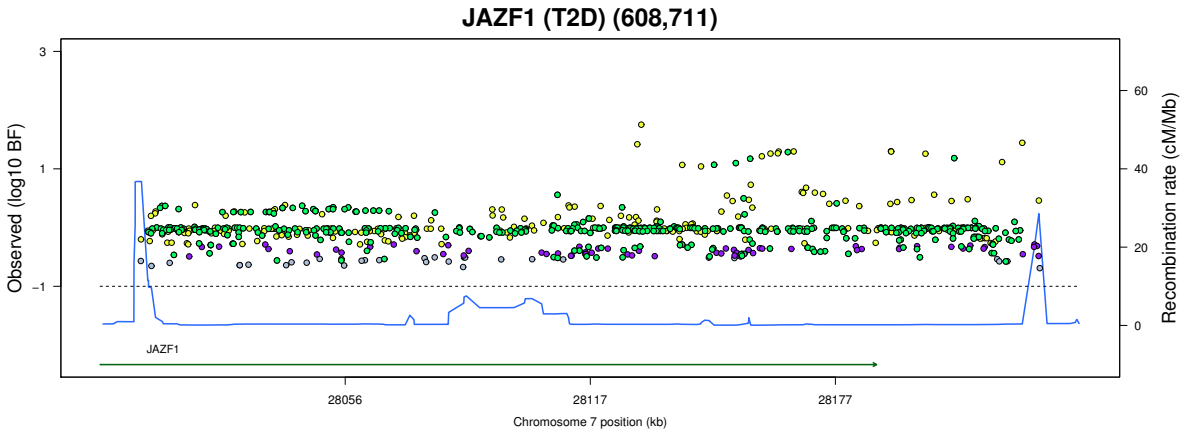


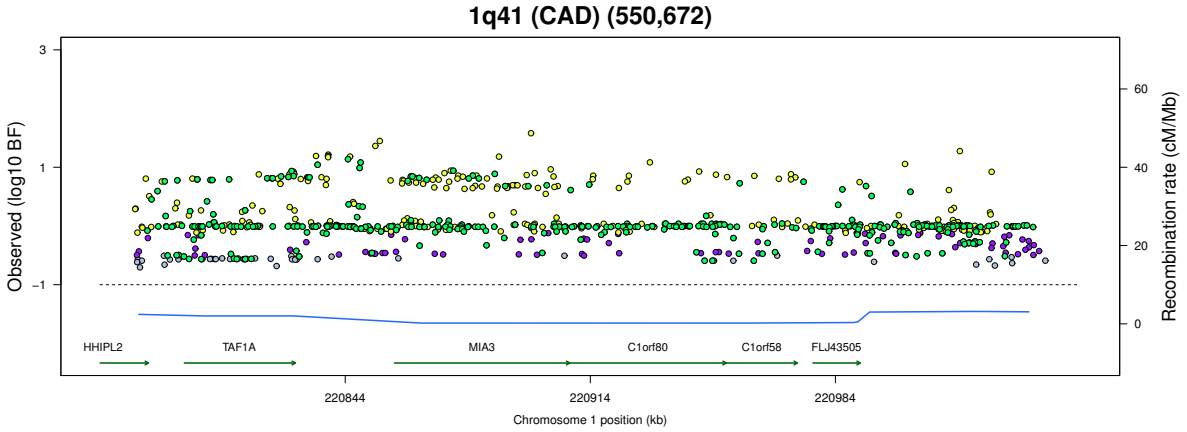
Nature Genetics: doi:10.1038/ng.2435

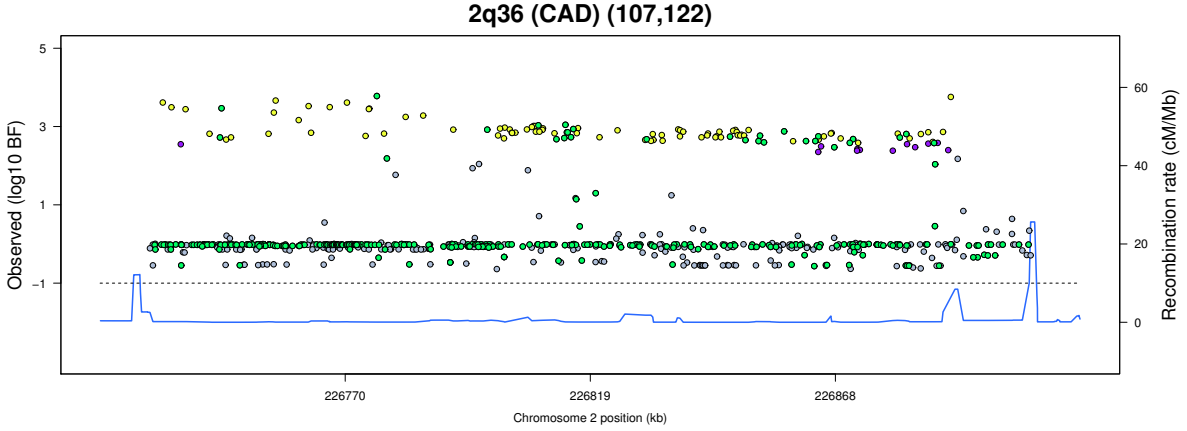

SORT1 (CAD) (168,289)
Observed (log10 BF)
Recombination rate (cM/Mb)
SARS
CELSR2
PSRC1
MYBPHL
109561
109592
109624
Chromosome 1 position (kb)

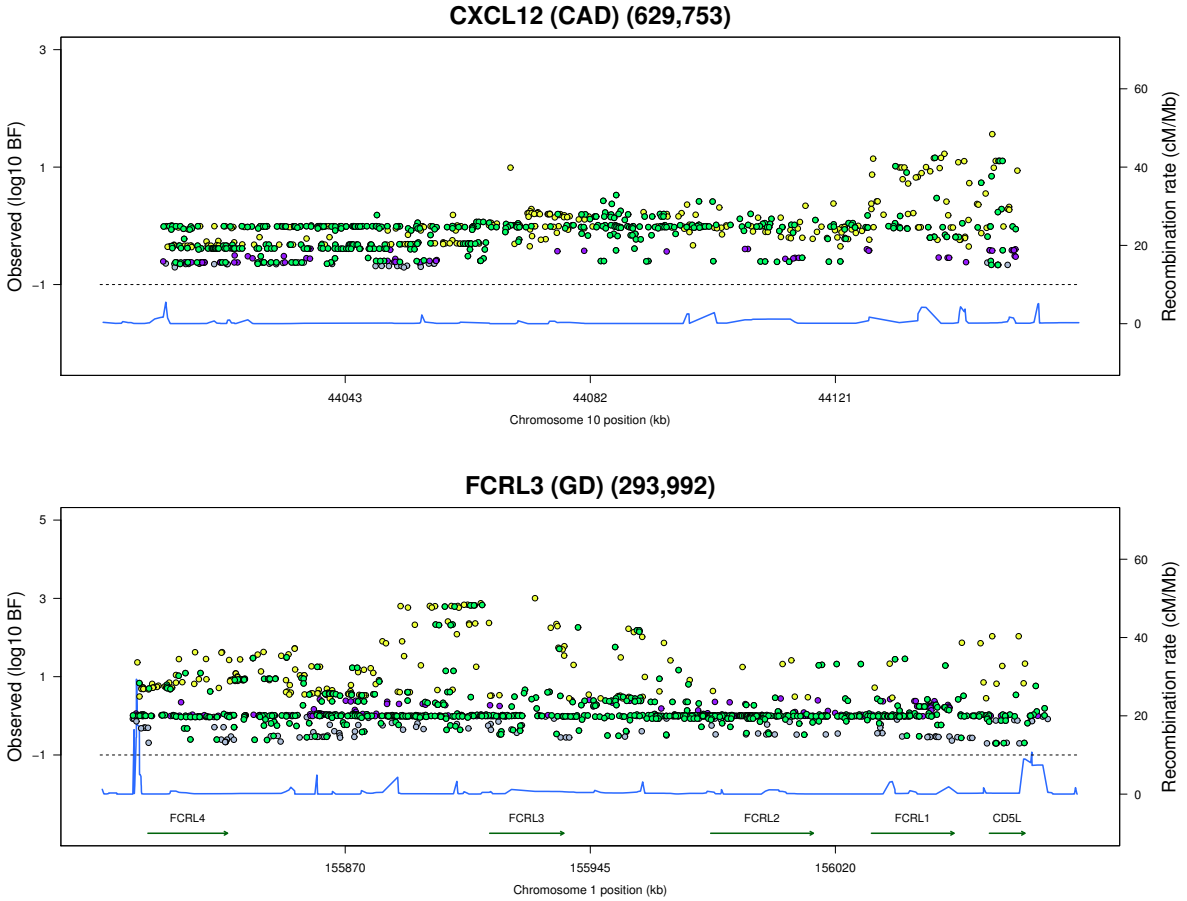

CTLA–4 (GD) (8,17)
Observed (log10 BF)
Recombination rate (cM/Mb)
CTLA4
ICOS
204410
204452
204494
Chromosome 2 position (kb)


CD25 (GD) (154,471)
Observed (log10 BF)
Recombination rate (cM/Mb)
IL15RA
IL2RA
6081
6114
6147
Chromosome 10 position (kb)

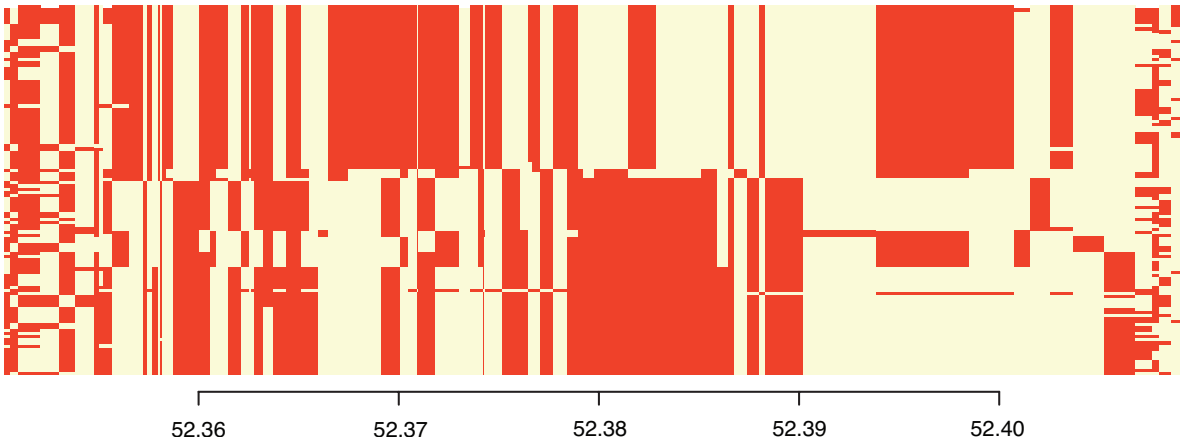

JAZF1 (T2D) (608,711)
Observed (log10 BF)
Recombination rate (cM/Mb)
JAZF1
28056
28117
28177
Chromosome 7 position (kb)


1q41 (CAD) (550,672)
Observed (log10 BF)
Recombination rate (cM/Mb)
HHIPL2
TAF1A
MIA3
C1orf80
C1orf58
FLJ43505
220844
220914
220984
Chromosome 1 position (kb)

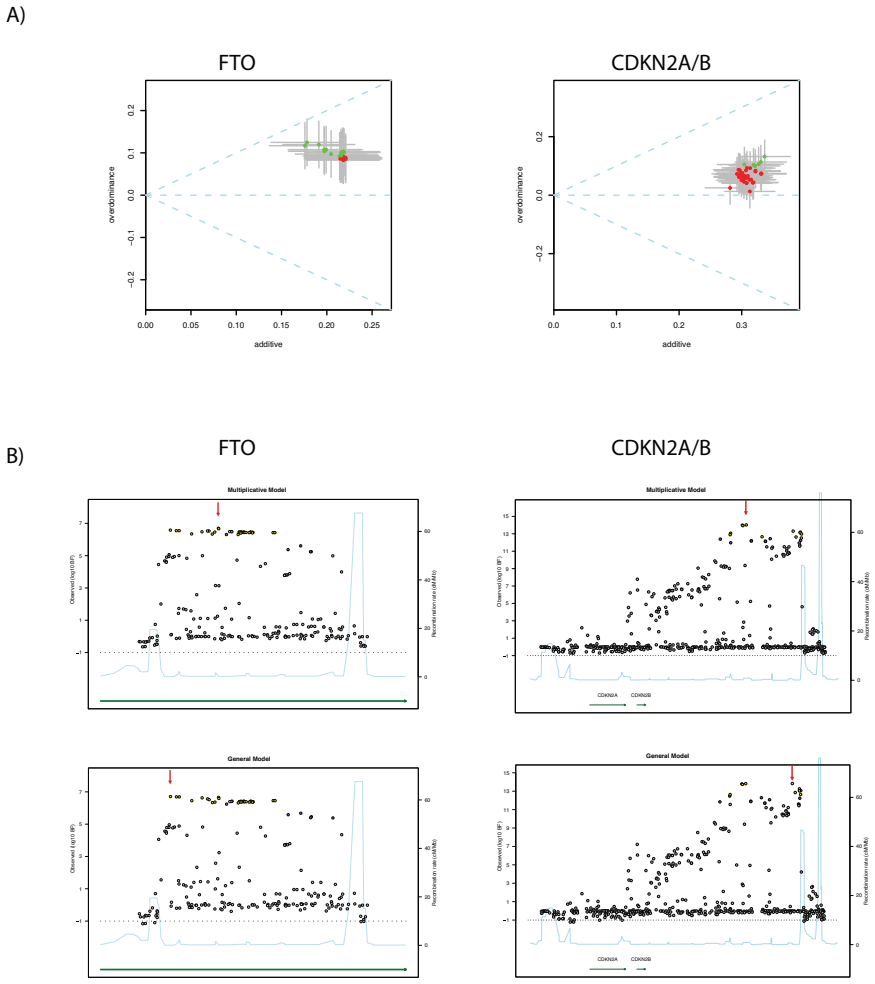

2q36 (CAD) (107,122)
Observed (log10 BF)
Recombination rate (cM/Mb)
226770
226819
226868
Chromosome 2 position (kb)



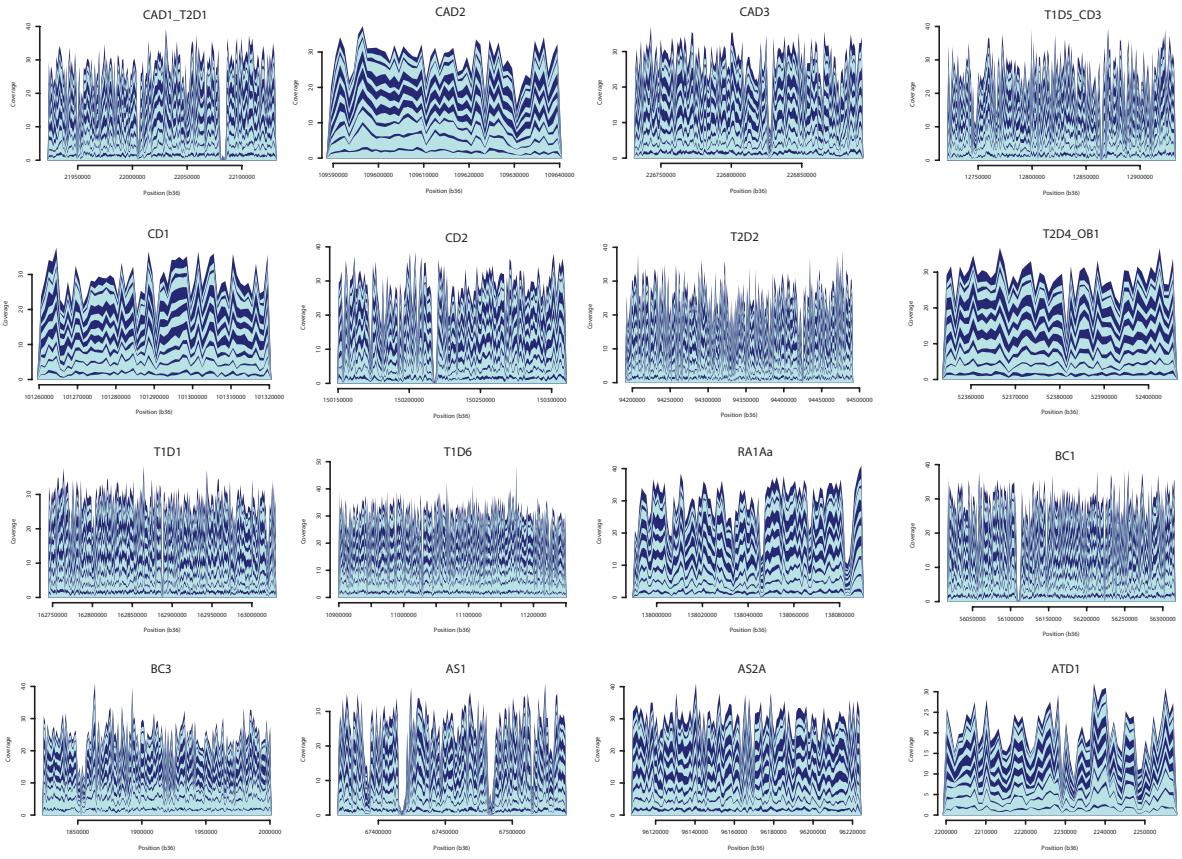
Supplementary Figure 9: Signal plots showing the strength of signal for each region, shown as $\log_{10}$ Bayes Factors for genotyped SNPs and for imputed SNPs (see imputation section of SOM for details). Estimated recombination rate is shown in blue, with scale on the right vertical axis. SNPs are coloured according to membership in credible sets: yellow for 95% credible set, purple for 99%, and grey otherwise. All imputed SNPs are coloured green, regardless of credible set membership. Genes in the region are shown towards the bottom in green. The title for each plot shows region name, follow by phenotype in parentheses, followed by credible set size (95,99) in parentheses. Credible set sizes reflect potential additions of imputed SNPs. Genomic positions are from NCBI build 36.



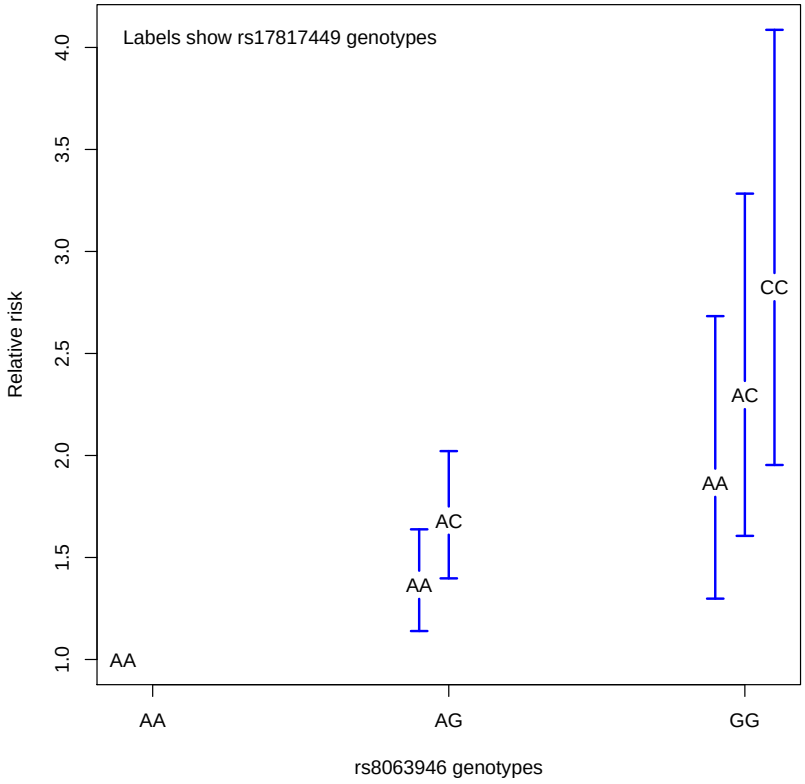
Supplementary Figure 10: The haplotype structure of the 120 CEU HapMap haplotypes across the FTO region. Each row represents an individual haplotype and each column a SNP (minor allele in red and major in white). The physical location, in Mb, are displayed on the x-axis. Two major haplotypes dominate the picture across the entire region.



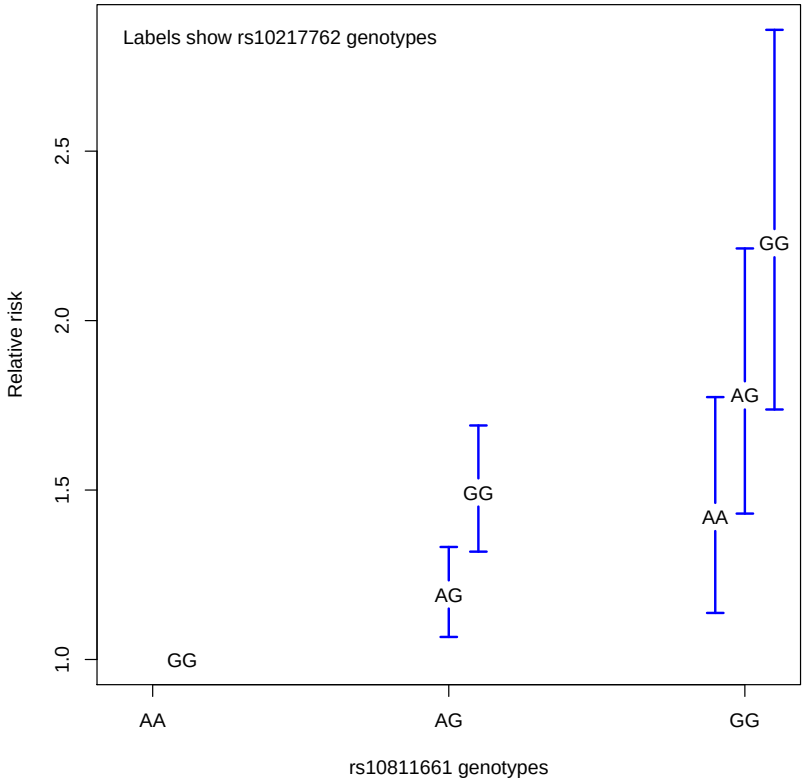
Supplementary Figure 11: Disease model comparison for FTO in T2D and CDKN2A/B in CAD. A) Scatter plot of parameter estimates for the general disease model for a selection of SNPs showing the strongest association signal in the region. The additive (multiplicative) and dominance parameters, on the log-odds (odds) scale, are plotted on the x- and y-axes respectively. Each point shows the maximum likelihood estimates for one SNP; the grey cross-hairs extend by one unit of standard error in each direction. Genotypes were coded with respect to the risk allele so that the additive parameter estimates were always positive. The SNPs plotted are those with $\log_{10}$ BF > 3 for the general model and are coloured depending on their $\log_{10}$ BF for the additive/general model comparison: those with a positive value (i.e. evidence in favour of the general model) are coloured green, those with a negative value are coloured red. The dashed cyan lines show specific models of interest: the horizontal line shows the space of additive models (i.e. the dominance parameter is zero), the top diagonal line shows the dominant models and the bottom diagonal line shows the recessive models. B) Signal plots showing the $\log_{10}$ BFs for all SNPs in the two regions for the additive (multiplicative) model (top panels) and the general model (bottom panels). In each plot, the arrow highlights the most associated SNP in each case, demonstrating that the signal peak moves depending on the model chosen.



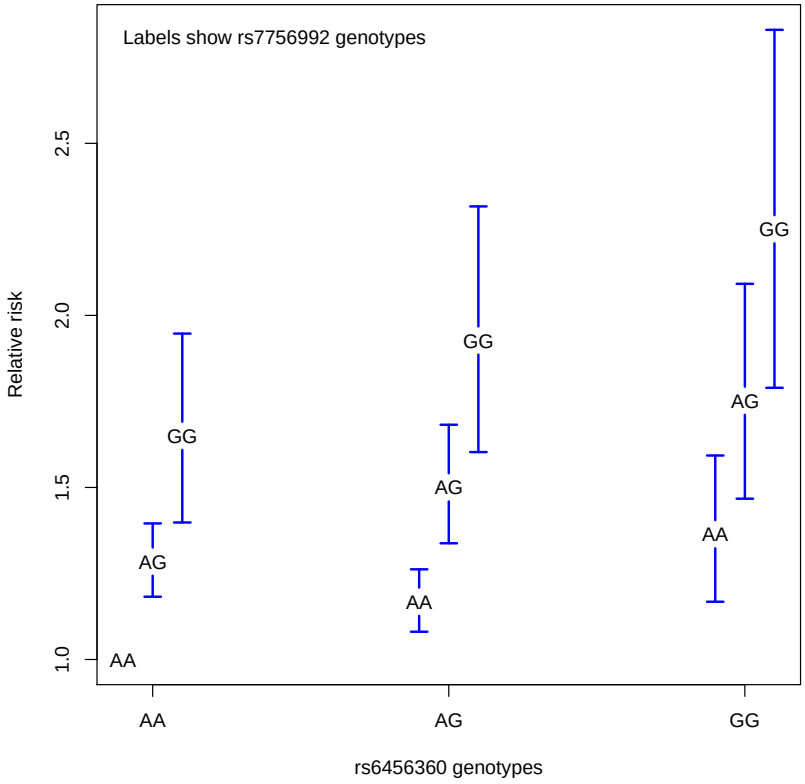
Supplementary Figure 12: Stacked charts showing the average per base coverage of confidently-mapped reads (in 1kb windows) from the 1000 Genomes Project for 16 individuals that overlap individual in 1kb windows across the 16 regions studied. Only data from the Illumina technology (Freeze 3) are shown.



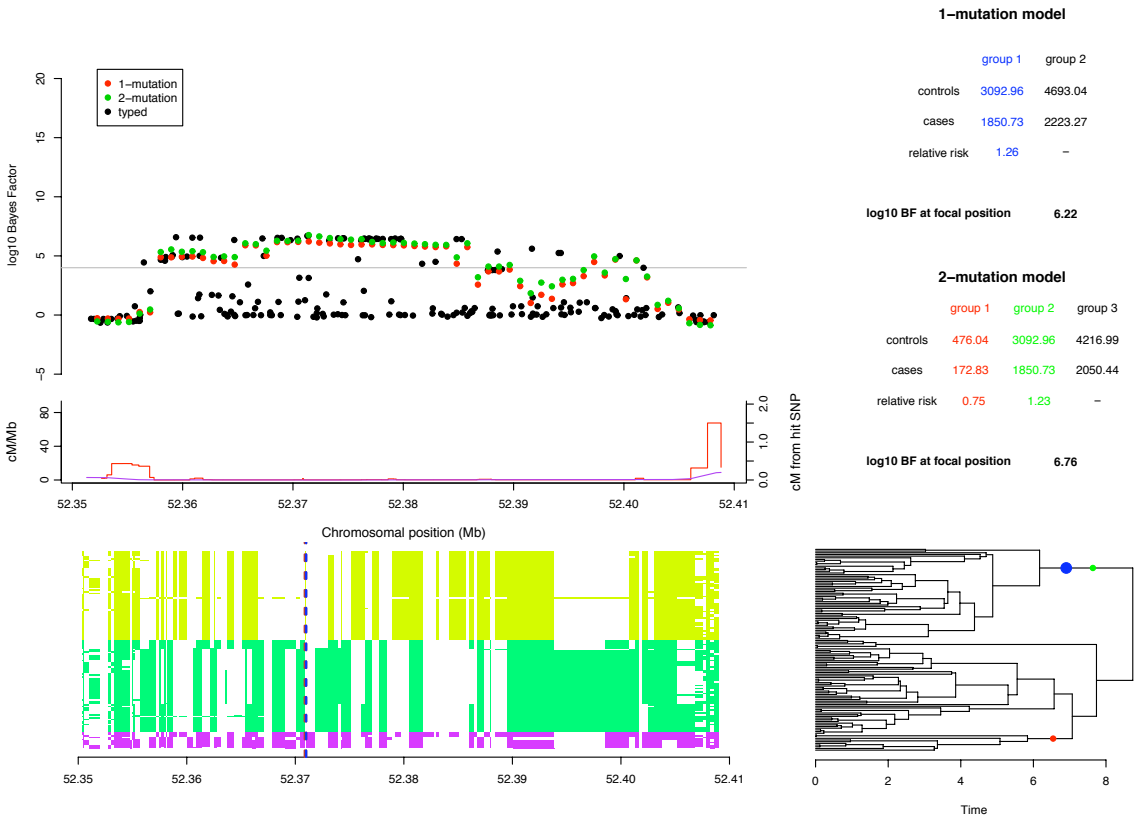
Supplementary Figure 13: **Two-SNP additive disease model for T2D, FTO.** RRs and 95% confidence intervals for each genotype combination, relative to the combination with lowest risk (which is shown with an RR of 1.0).



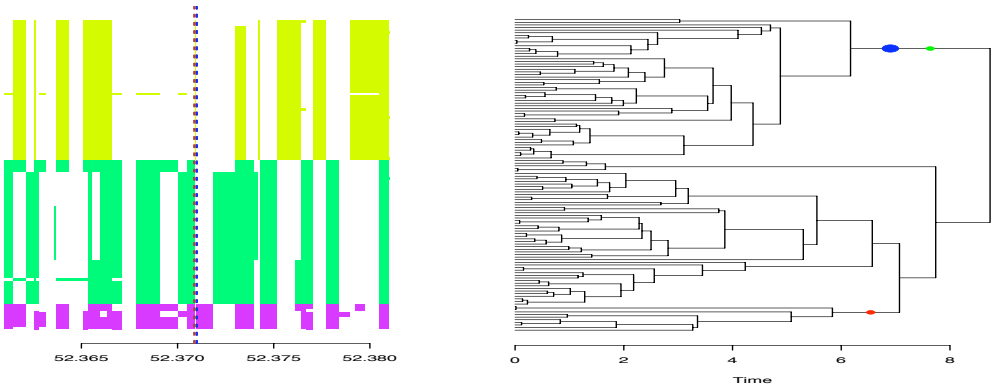
Supplementary Figure 14: **Two-SNP additive disease model for T2D, CDKN2A.** RRs and confidence intervals for each genotype combination, relative to the combination with lowest risk (which is shown with an RR of 1.0).



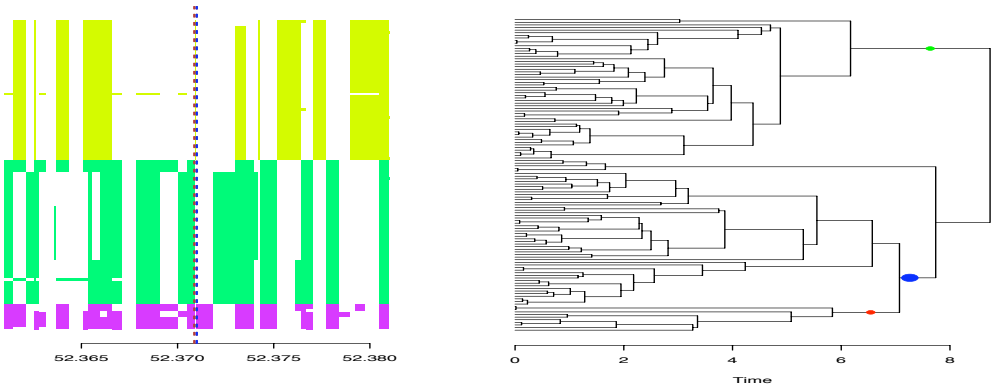
Supplementary Figure 15: **Two-SNP additive disease model for T2D, CDKAL1.** RRs and confidence intervals for each genotype combination, relative to the combination with lowest risk (which is shown with an RR of 1.0).



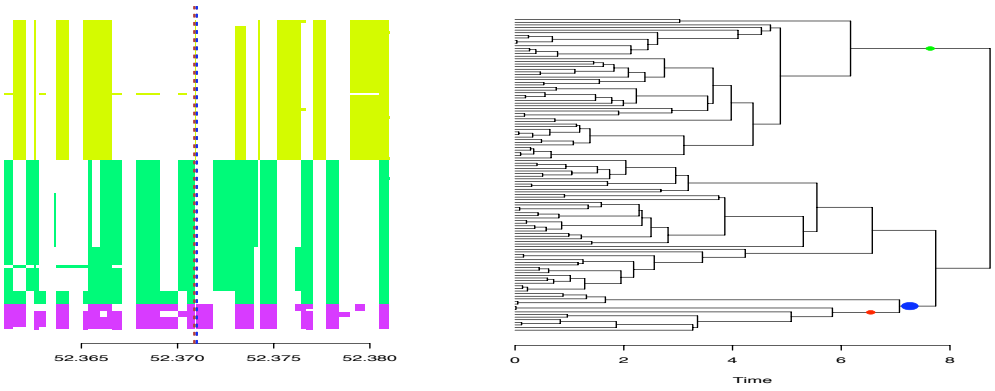
Supplementary Figure 16: GENECLUSTER analysis of T2D in the FTO region using the HapMap CEU haplotypes as the reference panel. The top left panel of the plot shows the $\log_{10}$ Bayes factors for the 1-mutation model (red), 2-mutation model (green), and the additive single SNP test (black). The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The panel haplotypes are coloured to indicate the 3 haplotypes that occur at the SNPs rs17817449 and rs8063946 (yellow=GC, purple=TT, green = TC). The dashed vertical blue line indicate the position of the largest $\log_{10}$ Bayes factor for the 2-mutation model (the focal position). The bottom right panel shows the estimated TREESIM tree at the focal position. The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes factors are shown with blue and red/green dots respectively. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel.



(a)

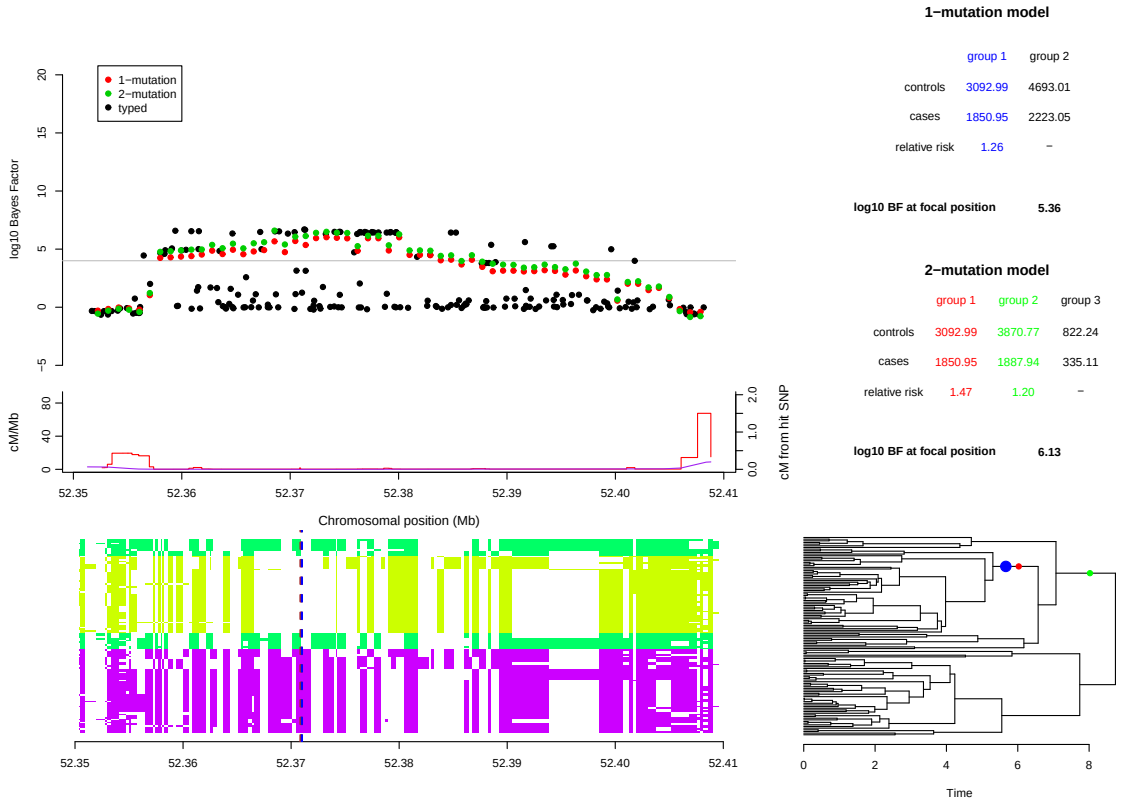


(b)

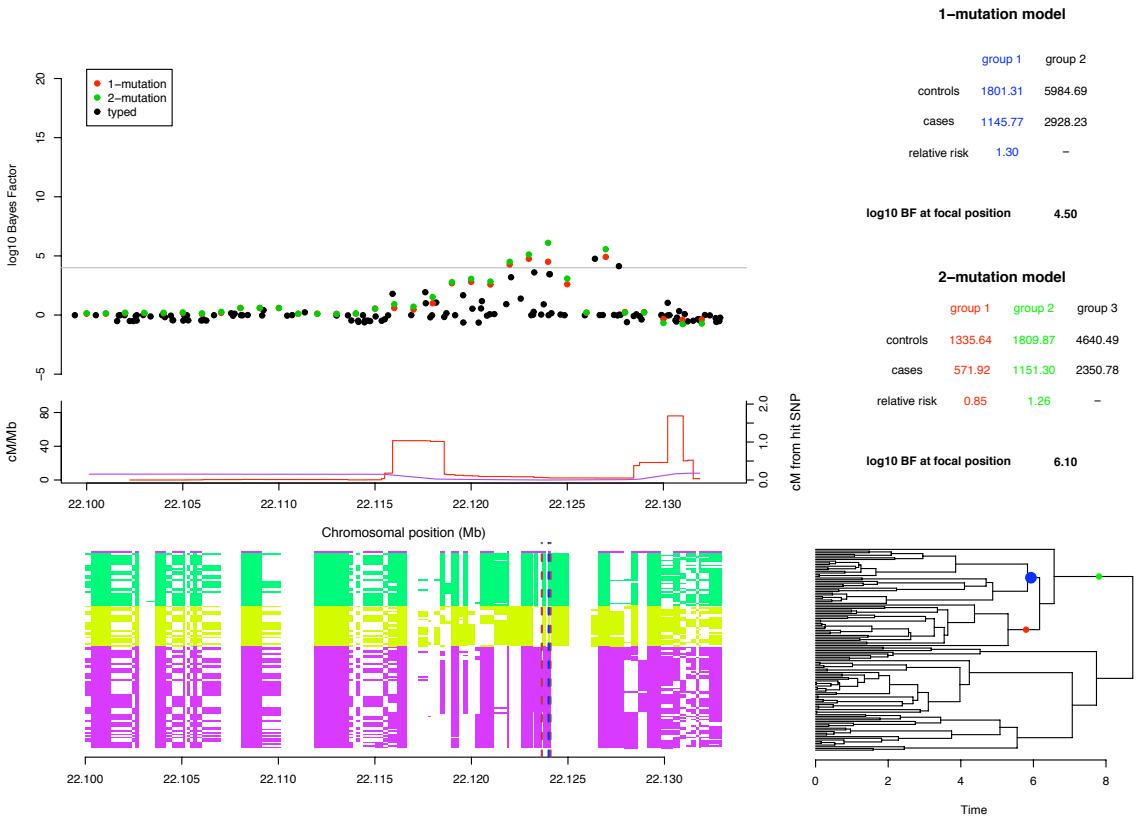


(c)

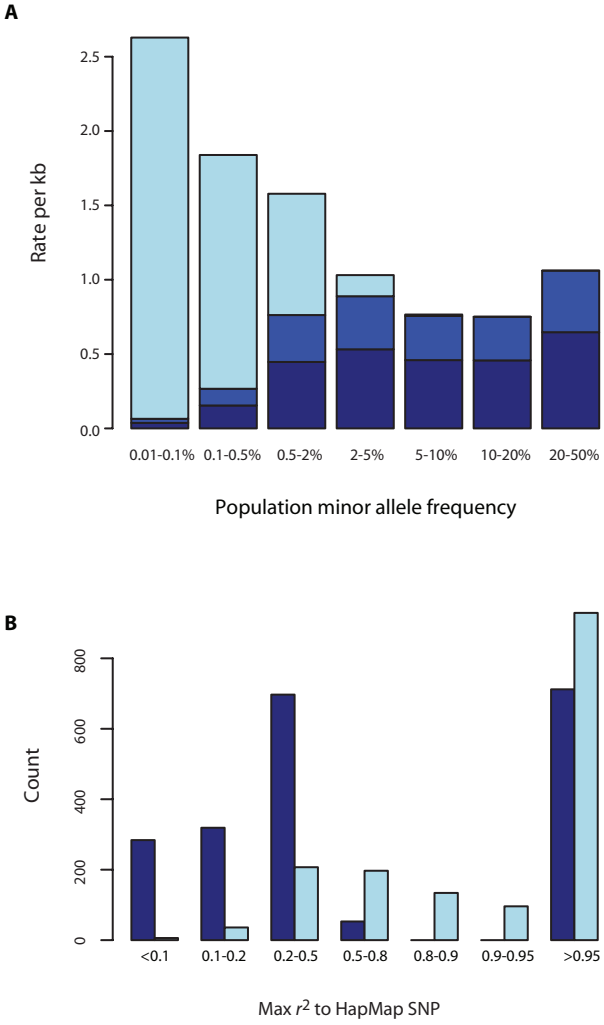
Supplementary Figure 17: **(a)** The estimated genealogy for the HapMap CEU reference panel at the focal position in Figure16, with the most likely single mutation (blue dot) and two mutations (red and green) under the 1-mutation and 2-mutation models respectively. Adjacent are the HapMap haplotypes, which are coloured to indicate the 3 haplotypes that occur at the SNPs rs17817449 and rs8063946 (yellow=GC, purple=TT, green = TC). **(b)** and **(c)** show the other possible mutations, indicated by the blue dot, that falls above all of the low risk TT haplotypes and some of the intermedite risk TC haplotypes. In **(c)**, the tree is slightly altered by changing the order of the coalescent events near the root of the tree in (a).



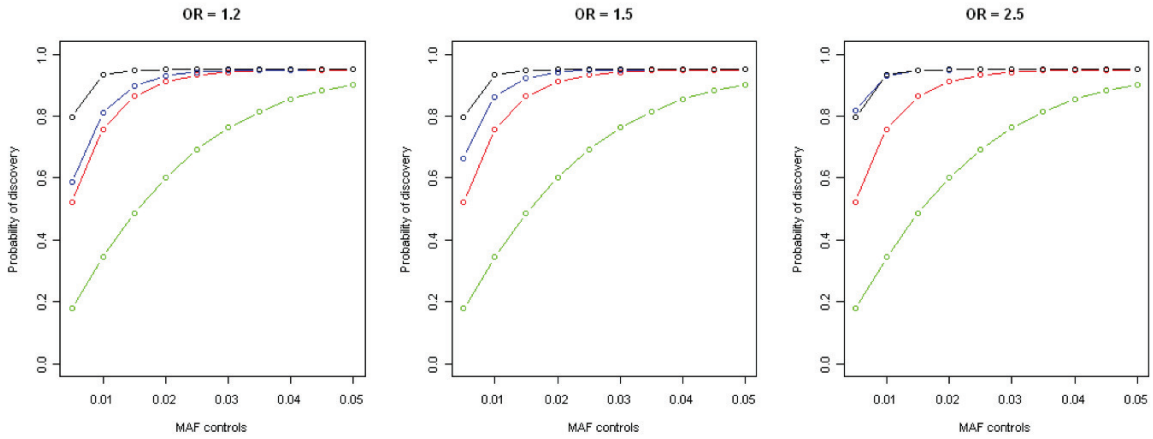
Supplementary Figure 18: GENECLUSTER analysis of T2D in the FTO region using the HapMap YRI haplotypes as the reference panel. The top left panel of the plot shows the $\log_{10}$ Bayes factors for the 1-mutation model (red), 2-mutation model (green), and the additive single SNP test (black). The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom left panel shows the 120 YRI HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The panel haplotypes are coloured to indicate the 3 haplotypes that occur at the SNPs rs17817449 and rs8063946 (yellow=GC, purple=TT, green=TC). The dashed vertical blue line indicate the focal position in Figure 16. The bottom right panel shows the estimated TREESIM tree at the focal position. The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes factors are shown with blue and red/green dots respectively. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel.



Supplementary Figure 19: GENECLUSTER analysis of T2D in the CDKN2A region using the HapMap CEU haplotypes as the reference panel. The top left panel of the plot shows the $\log_{10}$ Bayes factors for the 1-mutation model (red), 2-mutation model (green), and the additive single SNP test (black). The recombination map (red line) and the cumulative recombination map (purple line) are shown below this. The bottom left panel shows the 120 CEU HapMap haplotypes across the region. Each row of this panel is a haplotype and each column is a SNP. The panel haplotypes are coloured to indicate the 3 haplotypes that occur at the SNPs rs10811661 and rs10217762 (green=TC, yellow=CC, purple=TT). The dashed vertical blue line indicate the position of the largest $\log_{10}$ Bayes factor for the 2-mutation model (the focal position). The bottom right panel shows the estimated TREESIM tree at the focal position. The x-axis of the plot was chosen to provide a clear view of all the branches in the tree. The branches associated with the best 1-mutation and 2-mutation models that make the largest contributions to the Bayes factors are shown with blue and red/green dots respectively. The top right panel shows the tables of expected allele counts for the best 1-mutation and 2-mutation models together with a summary of the Bayes factors that occur at the focal position. The columns of the tables are colour matched to the mutations on the tree in the bottom right panel.



Supplementary Figure 20: A) Estimated rates of polymorphism in the human genome for variants in different population minor allele frequency categories. Colours indicate the rates across all genomes in the population (light blue), the set of 64 genomes sampled (blue) and the rates of variants discovered through the targeted sequencing strategy employed here (dark blue). B) Tagging of non-HapMap SNPs (both those newly discovered and those recorded in dbSNP but not genotyped in HapMap) by HapMap SNPs. Colours indicate rare SNPs ($MAF \leq 0.05$: dark blue, $MAF > 0.05$: light blue).



Supplementary Figure 21: The probability of discovery of a causal variant as a function of MAF in the control sample, effect size (odds-ratio) and experimental design - a random sample of 80 controls sequenced to 50x (red), a random sample of 80 cases sequenced to 50x (blue), 60 samples sequenced to 4x (green) and 400 samples sequenced to 4x (black). For all scenarios we assume that 95% of genomic sequence is accessible. For the low-coverage model we assume that read depth varies systematically across the genome such that the variance in read depth is twice that expected from a model of Poisson coverage and that a minimum of three reads from the minor allele are required for detection. In practice, variability in coverage for targeted sequencing considerably reduces the discovery rate.

Supplementary Table 1. Regions sequenced

Region	Associated diseases[a]	Chromosome	Start (b36)	End (b36)	focal SNP	MAF - CEU (HapMap)	Length (kb)	Recombination rate (cM/Mb)	STS coverage[b]	Mapped bases[c]	10x coverage[d]
IL23R	AS	1	67370000	67540000	rs11209026	0.067	170	1.12	87.6	89.4	71.3
SORT1	CAD	1	109589016	109640000	rs4970834	0.280	51	3.14	96.7	95.6	62.3
IFIH1	T1D	2	162745000	163030000	rs3788964	0.142	285	0.56	98.1	97.8	83.3
2q36	CAD	2	226731390	226892875	rs2943634	0.358	161	0.31	100.0	100.0	87.1
MAP3K1	BC	5	56017346	56315415	rs889312	0.308	298	0.77	99.6	99.3	79.8
ARTS1	AS	5	96108000	96224000	rs30187	0.300	116	0.52	97.8	97.9	88.5
MST150	CD	5	150150000	150310000	rs1000113	0.033	160	0.31	85.5	85.4	65.3
6q23.3	RA	6	137990000	138090000	rs6920220	0.175	100	1.00	100.0	100.0	90.0
CDKN2A/B	CAD/T2D	9	21923100	22131000	rs1333049	0.492	208	3.27	89.2	85.5	67.2
HHEX	T2D	10	94191000	94491000	rs5015480	0.432	300	0.57	90.9	91.7	54.3
NKX23	CD	10	101260000	101320000	rs10883365	0.500	60	3.67	89.2	87.3	71.6
LSP1	BC	11	1822948	2000443	rs3817198	0.342	177	2.66	61.3	59.4	22.3
KIAA0350	T1D	16	10900000	11250000	rs2542151	0.192	350	1.29	99.2	98.7	84.2
FTO	T2D/OB	16	52354000	52406000	rs9939609	0.450	52	3.65	100.0	100.0	82.8
PTPN2	T1D/CD	18	12721854	12932218	rs12708716	0.288	210	1.52	90.3	88.4	59.7
JSRP1	ATD	19	2199683	2257575	rs7250822	0.025	58	4.31	55.3	55.2	5.0
Average							172.3	1.79	90.0	89.5	67.2

[a]AS: Ankylosing spondylitis, CAD: coronary artery disease, T1D: type I diabetes, BC: breast cancer, CD: Crohn's disease, RA: rheumatoid arthritis, T2D: type II diabetes, OB: obesity, ATD: autoimmune thyroid disease

[c]The percent of bases with at least one successful PCR amplicon at design stage.

[c]The percent of bases with at least one read mapping in one individual

[d]The average percent of bases with at least 10x coverage

Supplementary Table 2. Properties of newly-discovered SNPs

	Illumina Resequencing				Capillary Resequencing				Percent SNPs novel	
Region	Mean size covered (kb)	Polymorphic known SNPs[a]	Novel SNPs	Double-hit novel SNPs	Mean size covered (kb)[b]	Polymorphic known SNPs	Novel SNPs	Double-hit novel SNPs	MAF ≤ 0.05	MAF > 0.05
IL23R	121.2	257	165	71	9.4	11	4	3	94	20
SORT1	31.8	104	57	13	6.2	9	6	4	59	20
IFIH1	237.4	63	265	75	5.5	1	4	1	94	27
2q36	140.2	198	200	64	5.1	2	6	4	83	35
MAP3K1	237.8	471	437	249	22	34	16	11	85	18
ARTS1	102.7	327	240	165	6.8	13	6	5	69	32
MST150	104.5	173	173	83	11.7	18	21	15	84	36
6q23.3	90.0	281	171	84	3.2	9	4	3	91	28
CDKN2A/B	139.8	237	191	65	37.2	31	18	16	90	64
HHEX	162.9	287	338	177	52.3	65	27	16	79	22
NKX23	43.0	127	69	29	4.1	9	0	0	80	19
LSP1	39.5	133	96	48	6	6	6	6	87	30
KIAA0350	294.7	703	477	230	49.1	77	36	26	84	29
FTO	43.1	107	83	38	2.1	5	0	0	82	23
PTPN2	125.4	288	232	98	8.2	22	5	4	78	32
JSRP1	2.9	19	5	1	2.7	11	4	0	100	14
Total	1916.7	3775	3199	1490	231.6	313	163	114		

[a]from dbSNP release 127.

[b]On average 8% of regions sequenced by capillary overlap with regions sequenced by Illumina.

[c]Percent of SNPs called from resequencing which are not in dbSNP release 127.

Supplementary Table 3. Average numbers and proportions of excluded (ie not in 95% set) SNPs in each region, stratified by "success" of fine-mapping in region.

Region	Total SNPs[a]	SNPs excluded [b]	Proportion Excluded[c]
TCF7L2	157	152	0.97
CDKN2A/B (T2D)	519	514	0.99
FTO	207	174	0.84
CDKAL1	497	464	0.93
HHEX	560	546	0.98
CDKN2A/B (CAD)	515	502	0.97
SORT1	231	139	0.60
CTLA4	293	287	0.98
CD25	426	350	0.82
JAZF1	339	87	0.26
1q41	354	114	0.32
2q36	343	257	0.75
CXCL12	363	97	0.27
FCRL3	607	493	0.81

[a]Number of genotyped SNPs in each region which pass QC and are polymorphic.
[b]Number of SNPs in each region not included in the credible set accounting for 95% of the posterior probability
[c]Proportion of SNPs in each region not included in the credible set accounting for 95% of the posterior probability

Supplementary Table 4.

	Region	SNP Set	rs ID	Risk allele frequency (controls)	Risk Allele	Effect Size (RR)	log BF	BF Ratio	λ_s
CAD	*SORT1*	Genotyped in FM	rs3832016	0.79	A	1.22 (1.11-1.35)	2.68	1.00	1.006
		Imputation	rs3832016	0.79	A	1.22 (1.11-1.35)	2.68	1.00	1.006
	CDKN2A/B	Genotyped in FM	rs1537370	0.47	A	1.40 (1.29-1.51)	14.03	1.00	1.028
		Imputation	rs1537370	0.47	A	1.40 (1.29-1.51)	14.03	1.00	1.028
	CXCL12	Genotyped in FM	rs34161818	0.84	A	1.21 (1.08-1.35)	1.56	1.00	1.004
		Imputation	rs34161818	0.84	A	1.21 (1.08-1.35)	1.56	1.00	1.004
	1Q41	Genotyped in FM	rs2936023	0.85	T	1.21 (1.08-1.36)	1.58	1.00	1.004
		Imputation	rs2936023	0.85	T	1.21 (1.08-1.36)	1.58	1.00	1.004
	2Q36	Genotyped in FM	rs2673145	0.41	A	1.20 (1.11-1.30)	3.76	1.00	1.008
		Imputation	rs2943634	0.34	A	1.22 (1.12-1.33)	3.78	1.01	1.009
T2D	*FTO*	Genotyped in FM	rs17817449	0.40	C	1.26 (1.17-1.36)	6.69	1.00	1.013
		Imputation	rs17817449	0.40	C	1.26 (1.17-1.36)	6.69	1.00	1.013
	CDKN2A/B	Genotyped in FM	rs12555274	0.23	C	1.26 (1.15-1.37)	4.75	1.00	1.011
		Imputation	rs12555274	0.23	C	1.26 (1.15-1.37)	4.75	1.00	1.011
	HHEX	Genotyped in FM	rs10882098	0.59	G	1.21 (1.12-1.31)	4.01	1.00	1.009
		Imputation	rs10882098	0.59	G	1.21 (1.12-1.31)	4.01	1.00	1.009
	CDKAL1	Genotyped in FM	rs7756992	0.27	G	1.29 (1.19-1.40)	6.71	1.00	1.014
		Imputation	rs7756992	0.27	G	1.29 (1.19-1.40)	6.71	1.00	1.014
	TCF7L2	Genotyped in FM	rs7903146	0.30	A	1.40 (1.29-1.52)	13.61	1.00	1.027
		Imputation	rs7903146	0.30	A	1.40 (1.29-1.52)	13.61	1.00	1.027
	JAZF1	Genotyped in FM	rs12531540	0.51	G	1.14 (1.06-1.23)	1.75	1.00	1.004
		Imputation	rs12531540	0.51	G	1.14 (1.06-1.23)	1.75	1.00	1.004
GD	*CTLA-4*	Genotyped in FM	rs11571297	0.51	A	1.39 (1.29-1.50)	16.08	1.00	1.027
		Imputation	rs11571297	0.51	A	1.39 (1.29-1.50)	16.08	1.00	1.027
	CD25/IL2RA	Genotyped in FM	rs10905669	0.23	A	1.20 (1.10-1.30)	3.10	1.00	1.007
		Imputation	rs10905669	0.23	A	1.20 (1.10-1.30)	3.10	1.00	1.007
	FCRL3	Genotyped in FM	rs11264798	0.52	C	1.17 (1.09-1.26)	3.00	1.00	1.006
		Imputation	rs11264798	0.52	C	1.17 (1.09-1.26)	3.00	1.00	1.006

Supplementary Table 5. Contribution of SNPs on Affymetrix 500k chip to fine mapping results.

Region	Proportion SNPs on 500k assay (original study)[a]	Proportion of 500k SNPs in 95% credible set[b]	Contribution to 95% credible set posterior[c]
SORT1 (CAD)	0.07	0.04	0.04
CDKN2A/B (CAD)	0.07	0.31	0.28
CXCL12 (CAD)	0.11	0.12	0.14
1q41 (CAD)	0.07	0.05	0.06
2q36 (CAD)	0.07	0.13	0.08
FTO (T2D)	0.07	0.06	0.06
CDKN2A/B (T2D)	0.07	0.20	0.04
HHEX (T2D)	0.03	0.21	0.21
CDKAL1 (T2D)	0.06	0.12	0.07
TCF7L2 (T2D)	0.09	0.40	0.06
JAZF1 (T2D)	0.10	0.13	0.11
CTLA-4 (GD)	0.11	0.33	0.07
CD25 (GD)	0.08	0.17	0.50
FCRL3 (GD)	0.07	0.11	0.10

[a]Proportion of fine mapping SNPs that are on the 500k chip
[b]Proportion of 95% credible set SNPs that are on the 500k chip
[c]Contribution to the 95% posterior probability made by SNPs on the 500k chip

Supplementary Table 6. Complete biological annotations for the 109 and 247 SNPs making up the 95% and 99% credible sets across seven fine mapped regions.

Annotation type	Annotation[a]	proportion of SNPs in 95% (99%) credible set[b]	proportion of posterior probability in 95% (99%) credible set[c]
refGene exon/intron[d]	all exons	0 (0.03)	0 (0)
	coding exons	0 (0)	0 (0)
	largest intron	0.05 (0.09)	0.14 (0.14)
	first intron	0.3 (0.19)	0.14 (0.14)
ensembl introns[d]	largest intron	0.05 (0.11)	0.14 (0.14)
	first intron	0.3 (0.18)	0.14 (0.14)
dbSNP 130 functions[d]	nonsynonymous	0 (0)	0 (0)
	synonymous	0 (0)	0 (0)
	intron	0.67 (0.53)	0.43 (0.43)
	splicing	0 (0)	0 (0)
	5' utr	0 (0)	0 (0)
	3' utr	0 (0.02)	0 (0)
	near-gene-5' (<= 2kb)	0.01 (0.01)	0 (0)
	near-gene-3' (<= 2kb)	0.01 (0)	0.01 (0.01)
	unkown (intergenic)	0.31 (0.44)	0.56 (0.56)
other gene prediction methods[d]	alternative splicing events	0 (0)	0 (0)
	noncoding RNA	0 (0)	0 (0)
	AceScan	0 (0)	0 (0)
	EVOfold secondary structure	0 (0)	0 (0)
	miRNA (mirBase)	0 (0)	0 (0)
HGDP Fst and signals of selective sweep[d]	positive selection on human branch $p < 0.05$	0 (0)	0 (0)
	HGDP Fst $p < 0.05$	0.06 (0.07)	0.03 (0.03)
	HGDP Bantu iHS $p < 0.05$	0.12 (0.10)	0.14 (0.14)
	HGDP Mideast iHS $p < 0.05$	0 (0)	0 (0)
	HGDP Europe iHS $p < 0.05$	0 (0)	0 (0)
	HGDP S. Asia iHS $p < 0.05$	0 (0)	0 (0)
	HGDP E. Asia iHS	0.13 (0.42)	0.14 (0.14)

	p < 0.05		
	HGDP Oceania iHS p < 0.05	0 (0)	0 (0)
	HGDP Americas iHS p < 0.05	0.05 (0.02)	0.01 (0.01)
	HGDP Bantu XP-EHH p < 0.05	0.09 (0.08)	0.04 (0.04)
	HGDP Mideast XP-EHH p < 0.05	0.22 (0.10)	0.06 (0.06)
	HGDP Europe XP-EHH p < 0.05	0.15 (0.07)	0.04 (0.04)
	HGDP S. Asia XP-EHH p < 0.05	0.22 (0.10)	0.06 (0.06)
	HGDP E. Asia XP-EHH p < 0.05	0.12 (0.36)	0.24 (0.24)
	HGDP Oceania XP-EHH p < 0.05	0.02 (0.01)	0.12 (0.12)
	HGDP Americas XP-EHH p < 0.05	0.02 (0.01)	0.12 (0.12)
Broad/Uppsala/ GIS ChIP-PET histone methylation and acetylation[d]	Broad H3K4me1 any -logp > 5	0.71 (0.57)	0.67 (0.67)
	Broad H3K4me2 any -logp > 5	0.47 (0.38)	0.54 (0.54)
	Broad H3K4me3 any -logp > 5	0.17 (0.14)	0.15 (0.15)
	Broad H3K9ac any -logp > 5	0.33 (0.24)	0.21 (0.21)
	Broad H3K9me1 any -logp > 5	0.40 (0.36)	0.31 (0.31)
	Broad H3K27ac any -logp > 5	0.48 (0.34)	0.37 (0.37)
	Broad H3K27me3 any -logp > 5	0.38 (0.43)	0.71 (0.71)
	Broad H3K36me3 any -logp > 5	0.58 (0.71)	0.49 (0.50)
	Broad H4K20me1 any -logp > 5	0.72 (0.70)	0.48 (0.49)
	Uppsala H3ac	0 (0.02)	0 (0)
	GIS ChIP-PET H3K4me3	0 (0.01)	0 (0)
	GIS ChIP-PET H3K27me3	0.03 (0.03)	0.02 (0.02)
transcription factor binding sites (ChIP-Seq)[d]	Uppsala *USF1*	0 (0)	0 (0)
	Uppsala *USF2*	0 (0)	0 (0)
	HAIB *NRSF* signal > 5	0 (0)	0 (0)
	HAIB *Pol2* signal > 5	0 (0)	0 (0)
	Broad/Duke *Pol2* p < 0.05	0.03 (0.08)	0.06 (0.06)
	Yale *Pol2*	0 (0.03)	0 (0)

	q < 0.05		
	Broad/Duke *CTCF* p < 0.05	0.04 (0.05)	0.04 (0.04)
	Duke *c-Myc* p < 0.05	0 (0.01)	0 (0)
	GIS ChIP-PET *c-Myc* score >= 800	0.03 (0.02)	0.03 (0.03)
	Yale *JunD* q < 0.05	0 (0)	0 (0)
	Yale *Max* q < 0.05	0 (0)	0 (0)
	Yale *SREBP1* q < 0.05	0 (0)	0 (0)
	Yale *SREBP2* q < 0.05	0 (0)	0 (0)
	Yale *c-Fos* q < 0.05	0.04 (0.04)	0.02 (0.02)
	Yale *c-Jun* q < 0.05	0.01 (0)	0.01 (0)
	Yale *NF-E2* q < 0.05	0 (0)	0 (0)
	Yale *Rad21* q < 0.05	0 (0.01)	0 (0)
	Yale *ZNF263* q < 0.05	0 (0)	0 (0)
	Yale *GATA1* q < 0.05	0 (0)	0 (0)
	Yale *TCF7L2* q < 0.05	0.05 (0.05)	0.03 (0.03)
	Yale *SATA1* q < 0.05	0.06 (0.05)	0.05 (0.05)
	GIS ChIP-PET *p53* score >= 800	0 (0)	0 (0)
DNase I hypersensitivity and nucleosome occupancy/ accessibility[d]	Duke/UW DNaseHS p < 0.05	0.10 (0.11)	0.07 (0.08)
	EIO/JCV Nucleosome Accesibility CD34+	0.03 (0.02)	0.04 (0.04)
	EIO/JCV Nucleosome Accesibility CD34-	0.02 (0.02)	0.05 (0.05)
	UW Nucleosome Occupancy A375 > 1.0	0.01 (0.03)	0 (0)
	UW Nucleosome Occupancy Dennis > 1.0	0.03 (0.04)	0.02 (0.02)
	UW Nucleosome Occupancy Mec < -1.0	0.01 (0.01)	0 (0)
other regulatory	NHGRI NRE score >=800	0 (0)	0 (0)

predictions[d]	ORegAnno	0.01 (0)	0.01 (0.01)
	switchGear TSS	0 (0)	0 (0)
	TFBScons	0.01 (0.01)	0 (0)
	miRNA target site	0 (0)	0 (0)
	Vista Enhancer	0 (0)	0 (0)
	7Xreg score > 0.1	0.14 (0.1)	0.07 (0.07)
	FOX2 binding sites	0.14 (0.07)	0.15 (0.15)
	LI TAF1 sites	0 (0)	0 (0)
	NKI LADs	0.06 (0.06)	0.14 (0.14)
	eponine TSS	0 (0)	0 (0)
	firstEF	0 (0.01)	0 (0)
	NHGRI BiPro score >=800	0.30 (0.16)	0.14 (0.14)
sequence conservation[d]	17-way most conserved Vertebrate	0.05 (0.04)	0.02 (0.02)
	28-way most conserved Mammals	0.05 (0.03)	0.02 (0.02)

[a]annotation type.
[b]proportion of the SNPs in the 95% and 99% sets within each annotation class.
[c]proportion of the 95% and 99% posterior probability within each annotation class.
[d]see Methods for details on specific annotation classes.

Supplementary Table 7. Upper table: SNPs in 95% credible set for 7 regions

Lower table: Correlation matrix for SNPs in upper table, with same ordering.

CDKN2A/B (CAD)

SNP	Position	MAF (Control)	MAF (Case)	Bayes Factor	Posterior Probability	p-value
rs1537370	22074310	0.47	0.55	14.03	0.26	5.30E-17
rs10116277	22071397	0.47	0.55	13.99	0.23	5.80E-17
rs6475606	22071850	0.47	0.55	13.95	0.21	6.40E-17
rs1333045	22109195	0.50	0.42	13.31	0.05	3.10E-16
rs10757278	22114477	0.47	0.55	13.17	0.04	4.10E-16
rs10757279	22114630	0.47	0.55	13.17	0.04	4.20E-16
rs10757277	22114450	0.53	0.44	13.11	0.03	4.70E-16
rs9632885	22062638	0.46	0.54	13.06	0.03	5.30E-16
rs1333049	22115503	0.47	0.55	12.98	0.02	6.60E-16
rs9632884	22062301	0.47	0.55	12.90	0.02	7.90E-16
rs10757274	22086055	0.52	0.44	12.66	0.01	1.40E-15
rs10217586	22111349	0.48	0.40	12.63	0.01	1.60E-15
rs1333048	22115347	0.51	0.43	12.52	0.01	2.00E-15

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1.00	0.99	0.99	0.68	0.77	0.77	0.77	0.87	0.76	0.89	0.87	0.73	0.82
2	0.99	1.00	1.00	0.69	0.77	0.77	0.77	0.88	0.76	0.90	0.87	0.73	0.81
3	0.99	1.00	1.00	0.69	0.77	0.77	0.77	0.88	0.76	0.90	0.87	0.73	0.81
4	0.68	0.69	0.69	1.00	0.86	0.86	0.86	0.62	0.85	0.60	0.79	0.94	0.80
5	0.77	0.77	0.77	0.86	1.00	1.00	1.00	0.71	0.99	0.68	0.89	0.81	0.93
6	0.77	0.77	0.77	0.86	1.00	1.00	1.00	0.71	0.99	0.68	0.89	0.81	0.93
7	0.77	0.77	0.77	0.86	1.00	1.00	1.00	0.71	0.99	0.68	0.89	0.81	0.93
8	0.87	0.88	0.88	0.62	0.71	0.71	0.71	1.00	0.70	0.97	0.80	0.67	0.75
9	0.76	0.76	0.76	0.85	0.99	0.99	0.99	0.70	1.00	0.68	0.88	0.80	0.93
10	0.89	0.90	0.90	0.60	0.68	0.68	0.68	0.97	0.68	1.00	0.77	0.65	0.72
11	0.87	0.87	0.87	0.79	0.89	0.89	0.89	0.80	0.88	0.77	1.00	0.84	0.94
12	0.73	0.73	0.73	0.94	0.81	0.81	0.81	0.67	0.80	0.65	0.84	1.00	0.86
13	0.82	0.81	0.81	0.80	0.93	0.93	0.93	0.75	0.93	0.72	0.94	0.86	1.00

CDKN2A/B (T2D)

SNP	Position	MAF (Control)	MAF (Case)	Bayes Factor	Posterior Probability	p-value
rs12555274	22126440	0.23	.27	4.75	0.68	2.80E-07
rs7018475	22127685	0.24	.28	4.13	0.16	1.40E-06
rs10965250	22123284	0.17	.14	3.60	0.05	5.90E-06
rs10811660	22124068	0.17	.14	3.46	0.03	8.50E-06
rs10811661	22124094	0.17	.14	3.46	0.03	8.50E-06

	1	2	3	4	5
1	1.00	0.94	0.01	0.01	0.01
2	0.94	1.00	0.01	0.01	0.01
3	0.01	0.01	1.00	0.98	0.98
4	0.01	0.01	0.98	1.00	1.00
5	0.01	0.01	0.98	1.00	1.00

CDKAL1 (T2D)

SNP	Position	MAF (Control)	MAF (Case)	Bayes Factor	Posterior Probability	p-value
rs7756992	20787688	0.27	0.32	6.71	0.35	2.20E-09
rs9348441	20788657	0.26	0.31	5.91	0.06	1.60E-08
rs35261542	20783771	0.26	0.31	5.78	0.04	2.20E-08
rs6931514	20811931	0.26	0.31	5.73	0.04	2.50E-08
rs9368222	20794975	0.26	0.31	5.54	0.02	4.00E-08
rs4712523	20765543	0.31	0.36	5.53	0.02	4.30E-08
rs4712522	20764779	0.31	0.36	5.47	0.02	5.00E-08
rs2206738	20774225	0.31	0.36	5.44	0.02	5.30E-08
rs7748382	20773528	0.31	0.36	5.42	0.02	5.70E-08
rs9460544	20769508	0.31	0.36	5.41	0.02	5.80E-08
rs10440833	20796100	0.27	0.31	5.39	0.02	5.90E-08
rs4712524	20765844	0.31	0.36	5.39	0.02	6.10E-08
rs11759505	20768710	0.31	0.36	5.38	0.02	6.30E-08
rs10946398	20769013	0.31	0.36	5.38	0.02	6.30E-08
rs9460545	20769529	0.31	0.36	5.38	0.02	6.30E-08
rs4712525	20770945	0.31	0.36	5.38	0.02	6.30E-08
rs4712526	20771014	0.31	0.36	5.38	0.02	6.30E-08
rs9460546	20771611	0.31	0.36	5.38	0.02	6.30E-08
rs9465860	20772079	0.31	0.36	5.38	0.02	6.30E-08
rs9358356	20775361	0.31	0.36	5.37	0.02	6.40E-08
rs9465871	20825234	0.18	0.22	5.36	0.02	5.60E-08
rs6456367	20767566	0.31	0.36	5.35	0.02	6.70E-08
6-20766865	20766865	0.31	0.36	5.34	0.01	6.90E-08
rs7774594	20769122	0.31	0.36	5.33	0.01	7.10E-08
rs10946403	20825383	0.18	0.22	5.32	0.01	6.20E-08
rs7772603	20773925	0.31	0.36	5.32	0.01	7.30E-08
rs7752780	20774001	0.31	0.36	5.32	0.01	7.30E-08
rs2206739	20774223	0.31	0.36	5.32	0.01	7.30E-08
rs7754840	20769229	0.31	0.36	5.30	0.01	7.70E-08
rs35456723	20770196	0.31	0.36	5.27	0.01	8.10E-08
rs2328548	20824937	0.18	0.22	5.25	0.01	7.50E-08
6-20766957	20766957	0.31	0.36	5.23	0.01	9.00E-08
rs6935599	20825074	0.18	0.22	5.20	0.01	8.50E-08

(correlation matrix omitted due to size constraints)

FTO (T2D)

SNP	Position	MAF (Controls)	MAF (Case)	Bayes Factor	Posterior Probability	p-value
rs17817449	52370868	0.40	0.45	6.69	0.05	2.60E-09
rs8043757	52370951	0.40	0.45	6.66	0.05	2.80E-09
rs1421085	52358455	0.41	0.46	6.58	0.04	3.40E-09
rs11642015	52359995	0.41	0.46	6.54	0.03	3.80E-09
16-52360724	52360724	0.41	0.46	6.54	0.03	3.80E-09
16-52368444	52368444	0.40	0.45	6.53	0.03	3.90E-09
rs7187250	52368047	0.40	0.45	6.52	0.03	3.90E-09
rs8051591	52374253	0.40	0.45	6.49	0.03	4.20E-09
rs8050136	52373776	0.40	0.45	6.49	0.03	4.30E-09
rs9935401	52374339	0.40	0.45	6.48	0.03	4.40E-09
rs9923233	52376699	0.40	0.45	6.48	0.03	4.40E-09
16-52366624	52366624	0.41	0.46	6.48	0.03	4.40E-09
rs3751814	52376225	0.40	0.45	6.47	0.03	4.50E-09
rs11075989	52377378	0.40	0.45	6.47	0.03	4.50E-09
rs7202296	52379191	0.40	0.45	6.47	0.03	4.50E-09
rs8063057	52369934	0.40	0.45	6.46	0.03	4.60E-09
16-52376335	52376335	0.40	0.45	6.44	0.03	4.90E-09
rs11075990	52377394	0.40	0.45	6.44	0.03	4.90E-09
rs11075992	52377567	0.40	0.45	6.44	0.03	4.90E-09
rs9926289	52378004	0.40	0.45	6.44	0.03	4.90E-09
rs7202116	52379116	0.40	0.45	6.44	0.03	4.90E-09
16-52379684	52379684	0.40	0.45	6.44	0.03	4.90E-09
16-52379738	52379738	0.40	0.45	6.44	0.03	4.90E-09
16-52379740	52379740	0.40	0.45	6.44	0.03	4.90E-09
rs9923312	52376868	0.40	0.45	6.43	0.03	4.90E-09
rs7206410	52378798	0.40	0.45	6.43	0.03	5.00E-09
rs10468280	52384980	0.40	0.45	6.42	0.03	5.00E-09
16-52385463	52385463	0.40	0.45	6.42	0.03	5.00E-09
16-52379643	52379643	0.40	0.45	6.41	0.03	5.20E-09
rs9936385	52376670	0.40	0.45	6.40	0.02	5.30E-09
16-52363954	52363954	0.41	0.46	6.34	0.02	6.10E-09
16-52369289	52369289	0.40	0.45	6.33	0.02	6.30E-09
rs3751812	52375961	0.40	0.45	6.31	0.02	6.70E-09

(correlation matrix omitted due to size constraints)

HHEX (T2D)

SNP	Position	MAF (Control)	MAF (Case)	Bayes Factor	Posterior Probability	p-value
rs10882098	94434773	0.41	0.36	4.01	0.20	2.10E-06
rs5015480	94455539	0.41	0.36	3.80	0.12	3.50E-06
rs10882101	94452407	0.41	0.36	3.77	0.11	3.80E-06
rs10882102	94456475	0.41	0.36	3.72	0.10	4.30E-06
rs1111875	94452862	0.41	0.36	3.71	0.10	4.30E-06
rs10882099	94450630	0.41	0.36	3.66	0.09	4.90E-06
rs7923837	94471897	0.38	0.34	3.38	0.05	1.00E-05
rs10882106	94470314	0.38	0.34	3.36	0.04	1.00E-05
rs7923866	94472056	0.38	0.34	3.33	0.04	1.10E-05
rs7087591	94463609	0.38	0.34	3.32	0.04	1.20E-05
rs10748582	94467199	0.38	0.34	3.30	0.04	1.20E-05
rs12778642	94454287	0.43	0.39	2.62	0.01	6.60E-05
rs2149632	94222227	0.36	0.32	2.28	0.00	1.60E-04
rs10882071	94250085	0.36	0.32	2.18	0.00	2.00E-04

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	1.00	0.97	0.97	0.97	0.97	0.97	0.74	0.74	0.74	0.74	0.74	0.87	0.55	0.55
2	0.97	1.00	1.00	0.99	1.00	0.99	0.75	0.75	0.75	0.75	0.75	0.89	0.53	0.53
3	0.97	1.00	1.00	0.99	1.00	1.00	0.75	0.75	0.75	0.75	0.75	0.90	0.53	0.53
4	0.97	0.99	0.99	1.00	0.99	0.99	0.75	0.75	0.75	0.76	0.75	0.89	0.53	0.53
5	0.97	1.00	1.00	0.99	1.00	1.00	0.75	0.75	0.75	0.75	0.75	0.90	0.53	0.53
6	0.97	0.99	1.00	0.99	1.00	1.00	0.75	0.75	0.75	0.75	0.75	0.90	0.53	0.53
7	0.74	0.75	0.75	0.75	0.75	0.75	1.00	1.00	1.00	0.99	1.00	0.67	0.56	0.56
8	0.74	0.75	0.75	0.75	0.75	0.75	1.00	1.00	1.00	0.99	1.00	0.67	0.56	0.56
9	0.74	0.75	0.75	0.75	0.75	0.75	1.00	1.00	1.00	0.99	1.00	0.67	0.56	0.56
10	0.74	0.75	0.75	0.76	0.75	0.75	0.99	0.99	0.99	1.00	1.00	0.67	0.56	0.56
11	0.74	0.75	0.75	0.75	0.75	0.75	1.00	1.00	1.00	1.00	1.00	0.67	0.56	0.56
12	0.87	0.89	0.90	0.89	0.90	0.90	0.67	0.67	0.67	0.67	0.67	1.00	0.47	0.47
13	0.55	0.53	0.53	0.53	0.53	0.53	0.56	0.56	0.56	0.56	0.56	0.47	1.00	1.00
14	0.55	0.53	0.53	0.53	0.53	0.53	0.56	0.56	0.56	0.56	0.56	0.47	1.00	1.00

TCF7L2 (T2D)

SNP	Position	MAF (Control)	MAF (Case)	Bayes Factor	Posterior Probability	p-value
rs7903146	114748339	0.30	0.37	13.61	0.75	9.20E-17
rs34872471	114744061	0.29	0.36	12.86	0.13	5.50E-16
rs4506565	114746031	0.32	0.39	12.24	0.03	2.80E-15
rs7901695	114744078	0.32	0.39	12.10	0.02	3.90E-15
rs4575195	114755737	0.32	0.39	12.05	0.02	4.60E-15

	1	2	3	4	5
1	1.00	0.99	0.91	0.90	0.88
2	0.99	1.00	0.91	0.90	0.88
3	0.91	0.91	1.00	0.98	0.96
4	0.90	0.90	0.98	1.00	0.95
5	0.88	0.88	0.96	0.95	1.00

CTLA4 (GD)

SNP	Position	MAF (Control)	MAF (Case)	Bayes Factor	Posterior Probability	p-value
rs11571297	204453248	0.49	0.41	16.08	0.77	4.70E-19
rs11571302	204451179	0.46	0.38	14.95	0.06	7.50E-18
rs1968351	204401981	0.42	0.34	14.93	0.05	9.00E-18
rs3087243	204447164	0.45	0.37	14.89	0.05	9.20E-18
rs11571293	204425958	0.41	0.33	14.48	0.02	2.70E-17
rs11571316	204439334	0.42	0.34	14.09	0.01	6.70E-17

	1	2	3	4	5	6
1	1.00	0.93	0.74	0.84	0.70	0.75
2	0.93	1.00	0.80	0.90	0.75	0.81
3	0.74	0.80	1.00	0.88	0.90	0.91
4	0.84	0.90	0.88	1.00	0.83	0.89
5	0.70	0.75	0.90	0.83	1.00	0.93
6	0.75	0.81	0.91	0.89	0.93	1.00

Supplementary Table 8. Resequencing contribution to posterior

Region (phenotype)	% of SNPs unique to resequencing	Contribution to 95% posterior
CDKN2A/B (T2D)	33%	0%
FTO (T2D)	46%	32%
HHEX (T2D)	42%	0.5%
CDKN2A/B (CAD)	33%	0%
SORT1 (CAD)	18%	4.5%
2q36 (CAD)	30%	7.5%

Supplementary Table 9

Phenotype	Region	Reported Lead snp	Reported Effect size	Effect size in our study	Lead SNP(s) in our study	Effect size	r^2
T2D	FTO	rs9936385	1.13	1.26	rs17817449	1.26	0.995
	CDKN2A/B	rs10811661			rs10811661		
	HHEX	rs1111875	1.11	1.2	rs10882098	1.21	0.967
	CDKAL1	rs7756992			rs7756992		
	TCF7L2	rs7903146			rs7903146		
	JAZF1	rs849135	1.11	1.13	rs12531540	1.14	0.876
CAD	SORT1	rs599839	1.14	1.17	rs3832016	1.22	0.917
	CDKN2A/B	rs4977574	1.29	1.37	rs1537370	1.4	0.85
	CXCL12	rs1746048	1.09	1.12	rs34161818	1.21	0.178
	1q41	rs17465637	1.14	1.12	rs2936023	1.21	0.485
	2q36	not present			rs2673145	1.2	

Comparison between results of our study and recent publications. T2D results compared with meta-analysis results for 34,840 cases and 114,981 controls from Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet* (2012). CAD results compared with those listed from the review Peden, J.F. & Farrall, M. Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour. *Hum Mol Genet* 20, R198-205 (2011). Region headings follow those used in our paper. Subsequent columns give the rsid of the lead SNP in the relevant recent study followed by the effect size estimate from that study. Next, the table gives the effect size estimate for that SNP in our study, the rsid of our top SNP for the region, and its effect size estimate from our study. The final column gives the r2 value, calculated in our control data, between the two SNPs. Where the pair of SNPs in a row are identical, other columns are left blank.

Members of the WTCCC+ consortium (Dec 2011)

Jan Aerts[1], Tariq Ahmad[2], Hazel Arbury[1], Anthony Attwood[1,3,4], Adam Auton[5], Stephen G Ball[6], Anthony J Balmforth[6], Chris Barnes[1], Jeffrey C Barrett[1], Inês Barroso[1], Anne Barton[7], Amanda J Bennett[8], Sanjeev Bhaskar[1], Katarzyna Blaszczyk[9], John Bowes[7], Oliver J Brand[8,10], Peter S Braund[11], Francesca Bredin[12], Gerome Breen[13,14], Morris J Brown[15], Ian N Bruce[7], Jaswinder Bull[16], Oliver S Burren[17], John Burton[1], Jake Byrnes[18], Sian Caesar[19], Niall Cardin[5], Chris M Clee[1], Alison J Coffey[1], John MC Connell[20], Donald F Conrad[1], Jason D Cooper[17], Anna F Dominiczak[20], Kate Downes[17], Hazel E Drummond[21], Darshna Dudakia[16], Andrew Dunham[1], Bernadette Ebbs[16], Diana Eccles[22], Sarah Edkins[1], Cathryn Edwards[23], Anna Elliot[16], Paul Emery[24], David M Evans[25], Gareth Evans[26], Steve Eyre[7], Anne Farmer[14], I Nicol Ferrier[27], Edward Flynn[7], Alistair Forbes[28], Liz Forty[29], Jayne A Franklyn[10,30], Timothy M Frayling[2], Rachel M Freathy[2], Eleni Giannoulatou[5], Polly Gibbs[16], Paul Gilbert[7], Katherine Gordon-Smith[19,29], Emma Gray[1], Elaine Green[29], Chris J Groves[8], Detelina Grozeva[29], Rhian Gwilliam[1], Anita Hall[16], Naomi Hammond[1], Matt Hardy[17], Pile Harrison[31], Neelam Hassanali[8], Husam Hebaishi[1], Sarah Hines[16], Anne Hinks[7], Graham A Hitman[32], Lynne Hocking[33], Chris Holmes[5], Eleanor Howard[1], Philip Howard[34], Joanna MM Howson[17], Debbie Hughes[16], Sarah Hunt[1], John D Isaacs[35], Mahim Jain[18], Derek P Jewell[36], Toby Johnson[34], Jennifer D Jolley[3,4], Ian R Jones[29], Lisa A Jones[19], George Kirov[29], Cordelia F Langford[1], Hana Lango-Allen[2], G Mark Lathrop[37], James Lee[12], Kate L Lee[34], Charlie Lees[21], Kevin Lewis[1], Cecilia M Lindgren[8,18], Meeta Maisuria-Armer[17], Julian Maller[18], John Mansfield[38], Jonathan L Marchini[5], Paul Martin[7], Dunecan CO Massey[12], Wendy L McArdle[39], Peter McGuffin[14], Kirsten E McLay[1], Gil McVean[5,18], Alex Mentzer[40], Michael L Mimmack[1], Ann E Morgan[41], Andrew P Morris[18], Craig Mowat[42], Patricia B Munroe[34], Simon Myers[18], William Newman[26], Elaine R Nimmo[21], Michael C O'Donovan[29], Abiodun Onipinla[34], Nigel R Ovington[17], Michael J Owen[29], Kimmo Palin[1], Aarno Palotie[1], Kirstie Parnell[2], Richard Pearson[8], David Pernet[16], John RB Perry[2,18], Anne Phillips[42], Vincent Plagnol[17], Natalie J Prescott[9], Inga Prokopenko[8,18], Michael A Quail[1], Suzanne Rafelt[11], Nigel W Rayner[8,18], David M Reid[33], Anthony Renwick[16], Susan M Ring[39], Neil Robertson[8,18], Samuel Robson[1], Ellie Russell[29], David St Clair[13], Jennifer G Sambrook[3,4], Jeremy D Sanderson[40], Stephen J Sawcer[43], Helen Schuilenburg[17], Carol E Scott[1], Richard Scott[16], Sheila Seal[16], Sue Shaw-Hawkins[34], Beverley M Shields[2], Matthew J Simmonds[8,10], Debbie J Smyth[17], Elilan Somaskantharajah[1], Katarina Spanova[16], Sophia Steer[44], Jonathan Stephens[3,4], Helen E Stevens[17], Kathy Stirrups[1], Millicent A Stone[45,46], David P Strachan[47], Zhan Su[5], Deborah PM Symmons[7], John R Thompson[48], Wendy Thomson[7], Martin D Tobin[48], Mary E Travers[8], Clare Turnbull[16], Damjan Vukcevic[18], Louise V Wain[48], Mark Walker[49], Neil M Walker[17], Chris Wallace[17], Margaret Warren-Perry[16], Nicholas A Watkins[3,4], John Webster[50], Michael N Weedon[2], Anthony G Wilson[51], Matthew Woodburn[17], B Paul Wordsworth[52], Chris Yau[5], Allan H Young[27,53], Eleftheria Zeggini[1], Matthew A Brown[52,54], Paul R Burton[48], Mark J Caulfield[34], Alastair Compston[43], Martin Farrall[55], Stephen CL Gough[8,10,30], Alistair S Hall[6], Andrew T Hattersley[2,56], Adrian VS Hill[18], Christopher G Mathew[9], Marcus Pembrey[57], Jack Satsangi[21], Michael R Stratton[1,16], Jane Worthington[7], Matthew E Hurles[1], Audrey Duncanson[58], Willem H Ouwehand[1,3,4], Miles Parkes[12], Nazneen Rahman[16], John A Todd[17], Nilesh J Samani[11,59], Dominic P Kwiatkowski[1,18], Mark I McCarthy[8,18,60], Nick Craddock[29], Panos Deloukas[1], Peter Donnelly[5,18].

[1] The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA UK.
[2] Genetics of Complex Traits, Peninsula College of Medicine and Dentistry University of Exeter, EX1 2LU, UK.
[3] Department of Haematology, University of Cambridge, Long Road, Cambridge, CB2 0PT, UK.
[4] National Health Service Blood and Transplant, Cambridge Centre, Long Road, Cambridge CB2 0PT, UK.
[5] Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG, UK.
[6] Multidisciplinary Cardiovascular Research Centre (MCRC), Leeds Institute of Genetics, Health and Therapeutics (LIGHT), University of Leeds, Leeds, LS2 9JT, UK.
[7] arc Epidemiology Unit, Stopford Building, University of Manchester, Oxford Road, Manchester, M13 9PT, UK.
[8] Oxford Centre for Diabetes, Endocrinology and Medicine, University of Oxford, Churchill Hospital, Oxford OX3 7LJ, UK.
[9] Department of Medical and Molecular Genetics, King's College London School of Medicine, 8th Floor Guy's Tower, Guy's Hospital, London, SE1 9RT, UK.
[10] Centre for Endocrinology, Diabetes and Metabolism, Institute of Biomedical Research, University of Birmingham, Birmingham, B15 2TT, UK.
[11] Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Groby Road, Leicester LE3 9QP, UK.
[12] IBD Genetics Research Group, Addenbrooke's Hospital, Cambridge, CB2 0QQ, UK.
[13] University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen AB25 2ZD, UK.
[14] SGDP, The Institute of Psychiatry, King's College London, De Crespigny Park, Denmark Hill, London SE5 8AF, UK.
[15] Clinical Pharmacology Unit, University of Cambridge, Addenbrookes Hospital, Hills Road, Cambridge CB2 2QQ, UK.
[16] Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton SM2 5NG, UK.
[17] Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge CB2 0XY, UK.
[18] The Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK.
[19] Department of Psychiatry, University of Birmingham, National Centre for Mental Health, 25 Vincent Drive, Birmingham, B15 2FG, UK.
[20] BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow, G12 8TA, UK.
[21] Gastrointestinal Unit, Division of Medical Sciences, School of Molecular and Clinical Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU, UK.
[22] Academic Unit of Genetic Medicine, University of Southampton, Southampton, UK.
[23] Endoscopy Regional Training Unit, Torbay Hospital, Torbay TQ2 7AA, UK.
[24] Academic Unit of Musculoskeletal Disease, University of Leeds, Chapel Allerton Hospital, Leeds, West Yorkshire LS7 4SA, UK.
[25] MRC Centre for Causal Analyses in Translational Epidemiology, Department of Social Medicine, University of Bristol, Bristol, BS8 2BN, UK.
[26] Department of Medical Genetics, Manchester Academic Health Science Centre (MAHSC), University of Manchester, Manchester M13 0JH, UK.
[27] School of Neurology, Neurobiology and Psychiatry, Royal Victoria Infirmary, Queen Victoria Road, Newcastle upon Tyne, NE1 4LP, UK.
[28] Institute for Digestive Diseases, University College London Hospitals Trust, London NW1 2BU, UK.
[29] MRC Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Heath Park, Cardiff, CF14 4XN, UK.
[30] University Hospital Birmingham NHS Foundation Trust, Birmingham, B15 2TT, UK.
[31] University of Oxford, Institute of Musculoskeletal Sciences, Botnar Research Centre, Oxford, OX3 7LD, UK.
[32] Centre for Diabetes and Metabolic Medicine, Barts and The London, Royal London Hospital, Whitechapel, London, E1 1BB, UK.
[33] Bone Research Group, Department of Medicine and Therapeutics, University of Aberdeen, Aberdeen, AB25 2ZD, UK.
[34] Clinical Pharmacology and Barts and The London Genome Centre, William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK.
[35] Institute of Cellular Medicine, Musculoskeletal Research Group, 4th Floor, Catherine Cookson Building, The Medical School, Framlington Place, Newcastle upon Tyne, NE2 4HH, UK.
[36] Gastroenterology Unit, Radcliffe Infirmary, University of Oxford, Oxford, OX2 6HE, UK.
[37] Centre National de Genotypage, 2, Rue Gaston Cremieux, Evry, Paris 91057, France.
[38] Department of Gastroenterology & Hepatology, University of Newcastle upon Tyne, Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, UK.
[39] ALSPAC Laboratory, Department of Social Medicine, University of Bristol, BS8 2BN, UK.
[40] Division of Nutritional Sciences, King's College London School of Biomedical and Health Sciences, London SE1 9NH, UK.
[41] NIHR-Leeds Musculoskeletal Biomedical Research Unit, University of Leeds, Chapel Allerton Hospital, Leeds, West Yorkshire LS7 4SA, UK.
[42] Department of General Internal Medicine, Ninewells Hospital and Medical School, Ninewells Avenue, Dundee DD1 9SY, UK.
[43] Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge, CB2 2QQ, UK.
[44] Clinical and Academic Rheumatology, Kings College Hosptal National Health Service Foundation Trust, Denmark Hill, London SE5 9RS, UK.
[45] University of Toronto, St. Michael's Hospital, 30 Bond Street, Toronto, Ontario M5B 1W8, Canada.
[46] University of Bath, Claverdon, Norwood House, Room 5.11a Bath Somerset BA2 7AY, UK.
[47] Division of Community Health Sciences, St George's, University of London, London SW17 0RE, UK.
[48] Departments of Health Sciences and Genetics, University of Leicester, 217 Adrian Building, University Road, Leicester, LE1 7RH, UK.
[49] Diabetes Research Group, School of Clinical Medical Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK.
[50] Medicine and Therapeutics, Aberdeen Royal Infirmary, Foresterhill, Aberdeen, Grampian AB9 2ZB, UK.
[51] School of Medicine and Biomedical Sciences, University of Sheffield, Sheffield, S10 2JF, UK.
[52] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Nuffield Orthopaedic Centre, University of Oxford, Windmill Road, Headington, Oxford, OX3 7LD, UK.
[53] UBC Institute of Mental Health, 430-5950 University Boulevard Vancouver, British Columbia, V6T 1Z3, Canada.

[54] Diamantina Institute of Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Ipswich Road, Woolloongabba, Brisbane, Queensland, 4102, Australia.
[55] Cardiovascular Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK.
[56] Genetics of Diabetes, Peninsula College of Medicine and Dentistry, University of Exeter, Barrack Road, Exeter, EX2 5DW, UK.
[57] Clinical and Molecular Genetics Unit, Institute of Child Health, University College London, 30 Guilford Street, London WC1N 1EH, UK.
[58] The Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK.
[59] Leicester NIHR Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester, LE3 9QP, UK.
[60] Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, OX3 7LJ, UK.

Fine mapping of genome-wide association loci. Supplementary Note

August 30, 2012

1 Resequencing Experiment

To assess the performance of a high-throughput sequencing platform for discovery of variants in GWAS association regions, we undertook a resequencing experiment in 32 unrelated CEU HapMap control individuals across 16 regions showing association in WTCCC diseases (in our own data or other published data). These choices reflect a diversity of diseases, signal strength, frequency of the estimated risk allele, region size, and recombination rate (Supplementary Table 1). We used long-range PCR to target a total of 2.75Mb, and pooled the PCR products from the same individual for sequencing in a single lane of an Illumina GI sequencer with single-end sequencing reactions.

On average, 90% of the targeted sequence was covered by at least one successful amplicon and had at least one mapped read. However, this figure differs between regions (range 55% to 100%) and individuals (Supplementary Table 1 and Supplementary Figure 3). Among samples, variation in coverage is highly repeatable and the primary predictor is the success of the PCR tiling path (Supplementary Table 1; PCR success explains 68% of the variance in coverage at the 1kb scale). The addition of repeat content and GC content explains 73% of the variance in coverage, however SNP density, conservation, segmental duplications, structural variation and the presence of exons have no strong or systematic effect on residual coverage. To increase sequence coverage we also designed short amplicons across remaining gaps that were sequenced by conventional capillary sequencing.

Putative SNPs were identified independently within individuals through comparison of the mapped reads to the reference genome using MAQ2. Across the regions analysed, 7450 distinct and polymorphic SNPs were called with a frequency distribution characterised by a strong and predictable bias towards rare variants (Supplementary Figure 20). Of the SNPs identified, 45% were novel, 37% were previously known and had been genotyped in HapMap and 18% were previously known (from dbSNP), but had not been typed in HapMap. These rates differ considerably between regions and also between allele-frequency categories (Supplementary Table 2). Again as expected, we observed a bias towards rare variants in the newly-discovered SNPs: on average 84% of SNPs with MAF$\leq$ 0.05 in the sample are novel, compared to 28% of SNPs with MAF$>$0.05.

To validate the newly-discovered variants, 732 SNPs from across five regions were successfully genotyped on the same individuals as part of the fine-mapping project described below. Of

Nature Genetics: doi:10.1038/ng.2435

these, 32.5% were not polymorphic in the genotyping (35% for singletons and 30% for other SNPs). In contrast, the fine-mapping project had a high validation rate for SNPs in the dbSNP repository (99.5%), suggesting that most of the discordance between sequencing and genotyping reflects false positives in the resequencing. This is to be expected because the filters for SNP detection were deliberately relaxed so as to improve detection rates. False positive rates are dramatically improved at sites with higher coverage and higher minor allele frequency (estimated false positive rate at non-singletons in regions where all samples have at least 10x coverage is 1.2%), and the use of additional filters can substantially decrease the false positive rate further.

False negative rates were estimated in two ways. First, through comparison to HapMap Phase II genotypes we estimate that at sites where 80% of individuals have at least 10x coverage the false negative rate is 1.7% (10% for singletons). However, this figure for false negatives is a substantial underestimate of the genome-wide position because of the variation in coverage, and the location of HapMap SNPs in regions of high sequence complexity to which it is relatively easy to map short reads. Secondly, to try to quantify the effect of inadequate coverage of our regions, if we consider only sites where all individuals had coverage of 10x, we detected 2381 polymorphic SNPs, a polymorphism rate of approximately 1 SNP every 250 bases. Extrapolating this rate to the entire 2.75Mb we targeted for resequencing suggests that we should have discovered 11,000 SNPs, whereas we have identified only 68% of these. However, it is important to note that this figure varies considerably between regions and between variants of different allele frequency classes. For example, in the 2q36, 6q23.3, and *FTO* regions we expected to have detected at least 95% of variants with a minor allele frequency of 0.05 or greater (Supplementary Table 2).

In addition, we would like to know the fraction of all variants present in the population within a given allele frequency range that we have captured. This calculation requires knowledge of the population allele frequency spectrum, which we have estimated from the resequencing data at those sites where all individuals have 10x coverage using a modelling approach (see Methods). We can therefore predict the number of SNPs per kb within a given minor allele-frequency range in a very large sample, the fraction of these we would expect to have sampled in 32 individuals and the fraction of those detected (Supplementary Figure 20). For variants with a frequency range of 0.5-2% within the population we estimate that roughly half would be present in the sample of size 32, but due to heterogeneity in coverage the fraction discovered is less than a third. In contrast, for variants of 10% or greater frequency in the population, we expect to have detected roughly 80%. Another important factor is the success of genotyping. In the fine-mapping experiment approximately one third of all variants could not be successfully genotyped on the chosen platform (through design or assay failures), so that despite the extensive sequencing, only a small fraction (<20%) of the rare (MAF<0.05) variants not yet typed in the case-control study were discovered and characterised using the route of resequencing followed by genotyping. Again, it is worth noting that this figure varies considerably between regions.

The accuracy of genotype calls from the resequencing data was assessed through comparison with those from HapMap. After correcting for strand flips, allele-labelling problems, and false monomorphism in the Phase II HapMap data, we estimate a discrepancy rate of 0.62% at sites where coverage was at least 10x. Discrepancies are enriched at sites identified as heterozygous through genotyping (2.4%).

Another natural question is the extent to which we gained from the substantive resequencing efforts we had undertaken. Supplementary Table 8 shows, for each of the five regions where we had undertaken resequencing, the proportion of the posterior 95% credible sets accounted for by SNPs which we only knew about because of our resequencing experiment. The first four regions in the table correspond to those where fine mapping appears to have added to our understanding, and in these, most or all of the top SNPs after fine mapping were already known before our resequencing efforts. Our experiment only resequenced control individuals, but as we note below, there is less of a disadvantage to this approach than might be first thought.

2 Resequencing Methods

2.1 Primer design

Primers were designed automatically using Primer3 to obtain a 5-fold depth of coverage across each region of both 5 kb and 10 kb product sizes. All primer pairs were pre-screened and a non-redundant set selected at 3-fold coverage depth using a combination of the successful 5 and 10 kb products. Any regions refractory to automatic design, or remaining uncovered after prescreening had primers designed to them manually.

2.2 DNA preparation

CEPH DNA samples were obtained from Corriell Cell Repositories, quantitated using picogreen (protocol available on request) and diluted to the appropriate working concentration (see PCR method) using T0.1E.

2.3 PCR amplification

For 10 kb PCR: Primers were aliquoted at a concentration of 10 ng/μl, 4μl per well into 384-well microtitre plates and stored frozen until required. DNA samples were diluted to 60 ng/μl. A premix was made consisting of 2 μl of 10X Buffer (as supplied with the enzyme), 0.4 μl 10 mM dNTPs, 0.8 μl 50 mM $MgSO^4$ (as supplied with the enzyme), 0.16 μl Platinum Hi-Fi Taq, 11.14 μl DDW per PCR. 1.5 μl of diluted DNA were added per PCR to the premix and 16 μl of this mix added to each well of the microtitre plate containing the primers. Plates were heat sealed and subjected to PCR under the following conditions: 98 °C for 3 minutes; 15 cycles of 94 °C for 30 seconds, 68 °C for 30 seconds, −1 °C per cycle, 68 °C for 10 minutes; 19 cycles of 94 °C for 30 seconds, 58 °C for 30 seconds, 68 °C for ten minutes; 68 °C for 10 minutes. Reactions were held at 4 °C until required. For 5 kb PCR: The method was as for the 10 kb PCR except the DNA concentration was 20 ng/μl and 0.12 μl of Platinum Hi-Fi Taq were used per reaction. PCR was performed under the following conditions: 98 °C for 3 minutes; 15 cycles of 94 °C for 30 seconds, 68 °C for 30 seconds, −1 °C per cycle, 68 °C for 5 minutes; 19 cycles of 94 °C for 30 seconds, 58 °C for 30 seconds, 68 °C for 5 minutes; 68 °C for 10 minutes. Reactions were held at 4 °C until required.

2.4 QC of amplification products

Following PCR, products were separated by electrophoresis on 0.8 % agarose gels in 1 X TBE and visualised using ethidium bromide staining. Products were scored 1 — 5 according to the intensity of the bands on the gel with 1 representing a failed or messy product and 5 a very strong product. PCR products were pooled according to the scoring system with 2 μl added to the pool for a score of 5 and 10 μl for a score of 2. PCR was repeated for any failures which, if successful, were added to the pool to give maximum PCR product coverage across the regions.

2.5 Library making

Sequencing libraries for the Illumina GA platform were constructed by shearing 1μg of pooled PCR fragments by nebulisation (35psi, 6min), followed by end-repair with klenow polymerase, T4 DNA polymerase and T4 polynucleotide kinase (to blunt-end the DNA fragments). A single 3' adenosine base was then added to the fragments using klenow exo- and dATP, before ligation of SE Illumina adapters (containing primer sites for sequencing and flowcell surface annealing). Gel-electrophoresis was used to separate library DNA fragments from unligated adapters. Ligated DNA fragments were in the 100-200 bp size range were excised from the gel and DNA extracted then amplified by 18 cycles of PCR with Phusion polymerase. Sequencing libraries were denatured with sodium hydroxide and diluted to 3.5 pM in hybridisation buffer for loading onto a single lane of an Illumina GA flowcell. Cluster formation, primer hybridisation and single-end, 36 cycle sequencing were performed using proprietary reagents according to manufacturers recommended protocol .

2.6 Base calling, read mapping and SNP calling

Images generated during each sequencing cycle on the Genome Analyzer platform were processed using the Illumina's image analysis software to produce FASTQ sequence files (35bp reads). Poor quality reads *e.g.* greater than four uncalled bases, and potentially spurious reads e.g. poly A only, are filtered out prior to alignment. The MAQ software package v0.6.0 (Li et al. 2008) was used to generate ungapped alignments the remaining reads to NCBI build 36.1 of the human genome. Alignments involving more than two mismatches in the first 24bp of the read or more than four mismatches in total are screened out. The MAQ algorithm is used to generate consensus sequences from the reads alignments, from which potential heterozygous and homozygous single nucleotide polymorphism (SNP) sites are extracted. SNPs reported in non unique regions of the human genome, within 2bp of another SNP, having a consensus base quality score of less than 23, sequence depth less than 10x or greater than 600x are excluded from further analysis.

2.7 Estimating population rates of polymorphism

To estimate the rate at which variants of different minor allele frequency would be expected in a very large sample from the population we modeled the derived allele-frequency distribution

as a beta-distribution and allowed for some fraction of the genome to be immutable. Parameters were estimated by maximum likelihood using the beta-binomial model (Balding and Nichols 1995) using only those parts of the regions where all individuals had 10x coverage or more. Estimated values: alpha parameter = 0.0012 (compared to 0 for the standard neutral model), beta parameter = 1.70 (compared to 1 for the standard neutral model) and fraction immutable = 0.026.

3 Region Definition for Fine Mapping

Regions were defined based on the focal SNP (as reported in the literature) as well as HapMap recombination estimates, HapMap LD data and GWAS signal data. There was no genome-wide association data available for GD, and so it was difficult to assess how far the association signal stretched for the three GD regions. Association signal data from the original WTCCC was used for T2D and CAD regions. Our strategy was to first choose a set of boundaries using recombination information, and then refine those boundaries based first on correlation and then using association signal. Specifically, the boundaries were chosen to be a distance of at least 0.1 centimorgans both upstream and downstream from the focal SNP. We then checked if any SNPs outside these boundaries had an $r^2 > 0.2$ to the focal SNP and expanded the boundaries to include any such SNPs. The boundaries were then further adjusted to incorporate any SNPs with a p-value within 2 orders of magnitude of the p-value of the focal SNP. In most cases the initial boundaries based solely on recombination ended up as the final boundaries, though some minor adjustments were made.

4 Fine Mapping SNP Selection

We attempted to design assays for and genotype all polymorphic HapMap SNPs, and all SNPs from our resequencing pilot and the other (smaller) resequencing datasets we had access to for these regions. We did the same for any SNP in dbSNP (version 128) which had genotype or frequency data showing variation and any SNPs which had been reported by more than one group.

5 Genotyping and Quality Control

Genotype calling was performed in two stages. First, genotypes were called using Illuminus.[1] Only SNPs which Illuminus called with high confidence were taken forward directly; the remainder underwent manual cluster inspection and re-calling where appropriate. The first QC filter applied was to remove individual genotypes with an Illuminus "confidence score" less than 0.2, which is the recommended threshold.[1] Samples with call rates lower than 90% were filtered out. Sample heterozygosity was calculated, as it is a useful marker for sample failure or contamination, but the structure of the data (a small number of relatively small regions) made it unlikely that we would be able to clearly identify outliers. We filtered out SNPs with call rate less than 0.95, with Hardy-Weinberg p-value less than 0.001 or minor allele frequency less than 0.001.

6 Bayesian Fine Mapping

6.1 Introduction

A brief derivation of results for Bayesian analysis of fine mapping SNP data. Evidence of association is measured by the Bayes factor (BF). Under the assumptions that there is a single causal SNP in a given region, and that it is typed in the study, we derive two convenient results when using a uniform prior on SNPs being causal. Firstly, that the BF for the region is the mean of the BFs for the individual SNPs. Secondly, the posterior that a given SNP is the causal SNP is proportional to these 'single-SNP' BFs.

6.2 Disease model

Let $\boldsymbol{X} = (\boldsymbol{X}^{(R)}, \boldsymbol{X}^{(S)})$ be the genotype data on n_R controls and n_S cases at k loci in a genomic region where $\boldsymbol{X}_i = (\boldsymbol{X}_i^{(R)}, \boldsymbol{X}_i^{(S)})$ denotes the data at locus $i = 1, \ldots, k$. Consider models M_0 (null model, no genetic effects) and $M_1, \ldots, M_k$, where M_i says that SNP i is the (only) causal locus in the region, and denote $M = M_1 \cup \ldots \cup M_k$. In other words, M represents the model where exactly one SNP in the region is causal.

Under the null model M_0, both cases and controls are sampled directly from the general population. Under model $M_i, i > 0$, the controls are still sampled directly from the general population but the case haplotypes are sampled in two steps. First, the case genotypes at the causal locus are sampled from a distribution in which the risk allele/genotype is overrepresented compared to the general population according to an odds-ratio parameter θ. Second, for the non-causal loci, (whose data are denoted by $\boldsymbol{X}_{-i}$),

$$\boldsymbol{X}_{-i}^{(S)} | (\boldsymbol{X}_i^{(S)}, \theta, M_i) \sim \boldsymbol{X}_{-i}^{(S)} | (\boldsymbol{X}_i^{(S)}, M_0),$$

i.e., given the case genotypes at the causal SNP, the case haplotypes over the non-causal SNPs are sampled from the general population distribution after conditioning on the genotypes at the causal SNP.

6.3 Inference

We are interested in measuring the evidence that there is a causal SNP in the region and also determining which SNP is likely to be the causal SNP. We do the former using the Bayes factor (BF) and the later by calculating the posterior probability of each SNP being causal.

6.3.1 Single-SNP BFs

Let BF_i be the Bayes factor comparing models M_i and M_0.

$$\begin{aligned}
\mathrm{BF}_i &= \frac{\Pr(\boldsymbol{X}|M_i)}{\Pr(\boldsymbol{X}|M_0)} \\
&= \frac{\Pr(\boldsymbol{X}_i|M_i)\Pr(\boldsymbol{X}_{-i}|\boldsymbol{X}_i, M_i)}{\Pr(\boldsymbol{X}|M_0)} \\
&= \frac{\Pr(\boldsymbol{X}_i|M_i)\Pr(\boldsymbol{X}_{-i}|\boldsymbol{X}_i, M_0)}{\Pr(\boldsymbol{X}|M_0)} \\
&= \frac{\Pr(\boldsymbol{X}_i|M_i)\Pr(\boldsymbol{X}_{-i}|\boldsymbol{X}_i, M_0)\Pr(\boldsymbol{X}_i|M_0)}{\Pr(\boldsymbol{X}|M_0)\Pr(\boldsymbol{X}_i|M_0)} \\
&= \frac{\Pr(\boldsymbol{X}_i|M_i)\Pr(\boldsymbol{X}|M_0)}{\Pr(\boldsymbol{X}|M_0)\Pr(\boldsymbol{X}_i|M_0)} \\
&= \frac{\Pr(\boldsymbol{X}_i|M_i)}{\Pr(\boldsymbol{X}_i|M_0)}.
\end{aligned}$$

Thus, BF_i depends only on the genotype data at locus i, and is thus called a single-SNP BF. It follows that in order to compute BF_i, we need not specify explicitly the joint model for the genotypes in the region, and in practice, we compute the single-SNP BFs by using the conventional single-SNP prospective likelihood as described in the subsection 6.2 below.

6.3.2 Region BF

Let $\mathrm{BF}_{\mathrm{reg}}$ be the BF that compares M and M_0. We will refer to this as the region BF. It measures the evidence that there is exactly one causal SNP in the region. We can write it in terms of the single-SNP BFs,

$$\begin{aligned}
\mathrm{BF}_{\mathrm{reg}} &= \frac{\Pr(\boldsymbol{X} \mid M)}{\Pr(\boldsymbol{X} \mid M_0)} \\
&= \frac{\sum_{i=1}^{k} \Pr(\boldsymbol{X} \mid M_i)\Pr(M_i \mid M)}{\Pr(\boldsymbol{X} \mid M_0)} \\
&= \sum_{i=1}^{k} \mathrm{BF}_i \Pr(M_i \mid M)\,.
\end{aligned}$$

Assuming a uniform prior on any particular SNP in the region being the causal SNP,

$$\Pr(M_i \mid M) = \frac{1}{k}\,,$$

results in the region BF as being simply the mean of the single-SNP BFs,

$$\mathrm{BF}_{\mathrm{reg}} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{BF}_i\,.$$

6.3.3 Posteriors on SNPs

Under the assumption that there is exactly one causal SNP, we show that the posterior that a given SNP is causal is proportional to its BF. By Bayes' Theorem,

$$\begin{aligned}\Pr(M_i \mid \boldsymbol{X}, M) &= \frac{\Pr(\boldsymbol{X} \mid M_i, M)\Pr(M_i \mid M)}{\Pr(\boldsymbol{X} \mid M)} \\ &= \frac{1}{k}\frac{\Pr(\boldsymbol{X} \mid M_i)}{\Pr(\boldsymbol{X} \mid M)} \\ &= \frac{1}{k}\frac{\Pr(\boldsymbol{X} \mid M_i)/\Pr(\boldsymbol{X} \mid M_0)}{\Pr(\boldsymbol{X} \mid M)/\Pr(\boldsymbol{X} \mid M_0)} \\ &= \frac{\mathrm{BF}_i}{k\mathrm{BF}_{\mathrm{reg}}} \propto \mathrm{BF}_i\,.\end{aligned}$$

6.4 Discussion

Calculating region BFs from single-SNP BFs has been previously described in certain contexts.[2,3] Our formulation makes explicit the assumptions that have been made under a retrospective disease model.

These derivations apply irrespective of the correlation amongst SNPs. In the situation of significantly associated but also highly correlated SNPs, the correct conclusion is that any of these could be causal but without necessarily identifying which one.[3] This will manifest itself through high single-SNP and region BFs, but with the posterior distributed quite evenly across multiple SNPs.

We have assumed that the causal SNP is typed in the study, which might be reasonable for a sufficiently thorough investigation of the variation in a region. Where this does not hold, the above methods are still applicable if there is a good surrogate SNP for the true effect.

In the presence of multiple causal SNPs, these methods are no longer optimal. They will tend to pick out the SNP with the best marginal effect, which may or may not be one of the causal SNPs.

7 Non-additive and Secondary Effects

7.1 Introduction

Analyses were carried out to detect and characterise any disease effects additional to the additive effect at the main SNP in each region. Briefly, in each region we looked for evidence of: (i) a significant deviation from an additive model at significantly associated SNPs; or (ii) a significant effect at other SNPs after conditioning on the genotypes from the main SNP. Where either of these were discovered, we analysed the region in more detail to determine the nature of the effects.

7.2 Additive (Multiplicative) vs. General Model

To test for deviations from the additive model, we must compare the standard additive model, where genotype contributes additively (multiplicatively) to the log-odds (odds) of disease, to one with an additional parameter that accounts for non-additive effects of genotype on the log-odds. Briefly, evidence of association for a SNP is measured by the Bayes Factor comparing models M_{alt} and M_{null}. By using the conventional prospective likelihood, this is

$$\mathrm{BF}_i = \frac{\int \Pr(\boldsymbol{Y}|\boldsymbol{X}_i, \theta_i, M_{alt}) \Pr(\theta_i|M_{alt}) d\theta_i}{\int \Pr(\boldsymbol{Y}|\theta_{null}, M_{null}) \Pr(\theta_{null}|M_{null}) d\theta_{null}}, \tag{1}$$

where $\boldsymbol{Y}$ is a vector of individual phenotypes and $\boldsymbol{X}_i$ is a vector of individual genotypes at SNP i (0,1, or 2). M_{alt} is a model where allele 1 at SNP i contributes to the log-odds of disease, and M_{null} is a model of no association. Both models rely on a logistic regression model for the likelihood,

$$\Pr(\boldsymbol{Y}|\boldsymbol{X}_i, \theta, M) = \prod_{j=1}^{N} p_j^{Y_j} (1-p_j)^{Y_j} \tag{2}$$

where for model M_{null}, we have

$$\theta_{null} = (\mu) \qquad\qquad \log \tfrac{p_j}{1-p_j} = \mu,$$

and for model M_{alt}, we have either

$$\theta_{alt} = (\mu, \gamma) \qquad\qquad \log \tfrac{p_j}{1-p_j} = \mu + \gamma Z_{j,i}$$

in the additive model case or

$$\theta_{alt} = (\mu, \gamma, \delta) \qquad\qquad \log \tfrac{p_j}{1-p_j} = \mu + \gamma Z_{j,i} + \delta \mathbb{I}_{(Z_{j,i}=1)}$$

in the general model case. Y_j is the phenotype for individual j (0 if j is a control, 1 if j is a case) and $Z_{j,i}$ is the count of allele 1 at SNP i for individual j. Parameter μ is the baseline log-odds of disease, γ is the additive contribution of genotype to odds, and δ is the deviation from additivity. We use both the default priors and the Laplace approximation as developed in the original.[4]

To test for a non-additive effect at a SNP, we calculated a BF which compares the additive and general models. This turns out to be just a ratio of the additive and general BFs,

$$\begin{aligned}
\mathrm{BF}_{\text{non-additive}} &= \frac{\Pr(\text{data} \mid M_{\text{general}})}{\Pr(\text{data} \mid M_{\text{additive}})} \\
&= \frac{\Pr(\text{data} \mid M_{\text{general}}) / \Pr(\text{data} \mid M_{\text{null}})}{\Pr(\text{data} \mid M_{\text{additive}}) / \Pr(\text{data} \mid M_{\text{null}})} \\
&= \frac{\mathrm{BF}_{\text{general}}}{\mathrm{BF}_{\text{additive}}}\,.
\end{aligned}$$

We are only interested in comparing models at SNPs which actually show a substantial association signal. In particular, we considered all SNPs with $\log_{10}(\mathrm{BF}_{\text{general}}) > 3$ and then looked for those that also had $\log_{10}(\mathrm{BF}_{\text{non-additive}}) > 0.2$. Two regions had SNPs with this

property: CDKN2A in CAD, and FTO in T2D. Of these, CDKN2A had the strongest evidence of a departure from an additive model, but nevertheless the evidence is only suggestive with $\log_{10}(\mathrm{BF}_{\text{non-additive}}) \approx 0.5$ at the best SNP. The FTO region showed a somewhat similar picture, with $\log_{10}(\mathrm{BF}_{\text{non-additive}}) \approx 0.5$ at the SNP showing greatest deviation, however the deviation decays as the evidence of association increases, with the best SNPs having $\log_{10}(\mathrm{BF}_{\text{non-additive}})$ values in the range -0.1 to 0.2. Thus, the evidence of deviation can at best only be described as suggestive.

It is interesting to note that for the regions shown here, the evidence for a departure from additivity is often inconsistent with that observed in the original study.[4] In particular, neither of the two regions highlighted above showed significant departures from an additive model previously. Furthermore, the only region that did show significant departures previously, *CDKAL1* for T2D, here shows a definite additive effect.

7.3 Secondary effects

We tested for a secondary effect at a SNP using two approaches.

First, we fitted an additive model while using the genotype calls at the best SNP in the region as a covariate, and compared it to the null model with the same covariate using a maximum likelihood ratio test. These were carried out using the *expanded reference group*. Individuals that have missing data at either SNP were excluded. A p-value threshold of 1×10^{-3} was used to find regions of interest.

Second, we analysed the genealogy of the case-control sample using GENECLUSTER[5] to find evidence for multiple mutations. GENECLUSTER uses the genealogy of a reference haplotype panel to approximate the genealogy of the case-control sample at each position, by clustering the case-control haplotypes under the leaves of the reference genealogy. Placing a disease mutation on a branch of the reference panel genealogy defines a hypothetical disease SNP, where those case-control haplotypes that fall under the mutation carry the disease allele. For a given mutation, m, in the tree, GENECLUSTER carries out a Bayesian test of association, where haplotypes carrying the disease allele are assigned a common disease penetrance parameter and those that do not are assigned another, independent, penetrance parameter under the alternative model, and all case-control haplotypes are assigned a common penetrance parameter under the null model. The test statistic of this association test is a Bayes factor, $\mathrm{BF}_m = \frac{\Pr(\text{data}|m)}{\Pr(\text{data}|M_{\text{null}})}$. The Bayes factor of the *1-mutation* model, against the null model, is the average of the Bayes factors over all possible single mutations m that can occur in the genealogy, that is

$$\begin{aligned} \mathrm{BF}_{1-\text{mutation}} &= \frac{\Pr(\text{data} \mid M_{\text{1-mutation}})}{\Pr(\text{data} \mid M_{\text{null}})} \\ &= \sum_m \frac{\Pr(\text{data}|m)\Pr(m)}{\Pr(\text{data}|M_{\text{null}})} \\ &= \sum_m \mathrm{BF}_m \Pr(m), \end{aligned} \tag{3}$$

where $\Pr(m)$ is the prior probability on mutation m. The model can also be extended to two disease mutations, m_1 and m_2, in the genealogy, which defines two hypothetical disease SNPs,

and haplotypes that carry the same set of mutant alleles are assigned the same penetrance parameter. The Bayes factor of the *2-mutation* model, against the null model, is

$$\begin{aligned}
\mathrm{BF}_{2-\mathrm{mutation}} &= \frac{\Pr(\mathrm{data}|M_{2\text{-mutation}})}{\Pr(\mathrm{data}|M_{\mathrm{null}})} \\
&= \sum_{m_1,m_2} \frac{\Pr(\mathrm{data}|m_1,m_2)\Pr(m_1,m_2)}{\Pr(\mathrm{data}|M_{\mathrm{null}})} \\
&= \sum_{m_1,m_2} \mathrm{BF}_{m_1,m_2}\Pr(m_1,m_2).
\end{aligned}$$

We can quantify the evidence for the 2-mutation model over the 1-mutation model, that is the evidence for a secondary effect, in terms of a Bayes factor:

$$\begin{aligned}
\mathrm{BF}_{\mathrm{secondary}} &= \frac{\Pr(\mathrm{data} \mid M_{2\text{-mutation}})}{\Pr(\mathrm{data} \mid M_{1\text{-mutation}})} \\
&= \frac{\Pr(\mathrm{data} \mid M_{2\text{-mutation}})/\Pr(\mathrm{data} \mid M_{\mathrm{null}})}{\Pr(\mathrm{data} \mid M_{1\text{-mutation}})/\Pr(\mathrm{data} \mid M_{\mathrm{null}})} \\
&= \frac{\mathrm{BF}_{2\text{-mutation}}}{\mathrm{BF}_{1\text{-mutation}}}.
\end{aligned}$$

For example, if we assume even prior odds for the 1-mutation model versus 2-mutation model, then $\log_{10}(\mathrm{BF}_{\mathrm{secondary}}) = 0.5$ results in a posterior probability for the 2-mutation model, and hence secondary effects, of 0.76. In addition, the mostly likely pair of mutations, a posteriori, in the genealogical tree under the 2-mutation model induces three haplotype risk groups in the reference panel (each with a different penetrance parameter). It is possible to identify potential disease SNPs by searching for SNPs in the reference panel that are highly correlated with the haplotype risk groups, as we will demonstrate below.

We have identified three regions that showed evidence for secondary effects, all in T2D: *CDKAL1*, *CDKN2A*, *FTO*, which we discuss in detail below.

7.3.1 T2D, *FTO*

The best SNP in this region is rs17817449, and the best SNP on the conditional tests is rs8063946 with a p-value of 5.9×10^{-4}. Taken together, only 6 different genotype-pair combinations are observed in the data. This therefore corresponds to 3 underlying haplotypes, for which phase is unambiguous.

Comparing the two-SNP additive model to the two-SNP general model (here, 3 parameters vs 6 parameters) gives a p-value of 0.4, thus the additive model provides an adequate fit. Details of this model are shown in Supplementary Table 10 and Supplementary Figure 13.

A more extensive haplotype-based analysis was carried out in this region, taking all SNPs with MAF greater than 0.3 and from them aiming to determine a richer set of haplotypes in differing risk classes, but this did not improve on the two-SNP additive model above (data not shown).

For the GENECLUSTER analysis, we used a number of reference panels, including the 120 HapMap CEU haplotypes and various subsets of phased case and control haplotypes with

Supplementary Table 10: **Two-SNP additive disease model for T2D, *FTO*.**

SNP	MAF	Relative risk
rs17817449	0.42	1.23 (1.14–1.33)
rs8063946	0.055	1.37 (1.14–1.64)

size ranging from 20 cases and 40 controls to 650 cases and 1300 controls. Supplementary Figure 16 shows the results with the HapMap haplotypes as the reference panel. The top left plot shows the $\log_{10}$ BF across the region for the 1-mutation model (red points) and 2-mutation model (green points), and $\log_{10}$ BF of the additive test of association at genotyped SNPs (black points). The fine-scale recombination rates and cumulative genetic distance are displayed below in red and purple respectively. The maximum $\log_{10}$ BF under the 2-mutation model is 6.76 and occurs at position 52371000, which we refer to as the *focal* position; at the same position the $\log_{10}(\mathrm{BF}_{\text{1-mutation}}) = 6.22$, which leads to $\log_{10}(\mathrm{BF}_{\text{secondary}}) = 0.54$ and a posterior probability for a secondary effect of 0.78 (assuming a prior probability of 0.5). The bottom right plot shows the most likely marginal genealogical tree of the HapMap haplotypes at the focal position, which is used to construct the genealogical tree of the full case-control sample. The root of the HapMap tree is the ancestral root. The large blue dot and the red and green dots indicate the most likely positions, a posteriori, for a disease mutation under the 1-mutation model and a pair of mutations under the 2-mutation model, respectively. The bottom left plot shows the HapMap haplotypes, where each row is a haplotype and the order of the haplotypes corresponds to the leaves of the tree to the right. Each column is a SNP, and at each SNP one of the allele types is coloured in white and the other is coloured according to the alleles carried at the two SNPs of interest (rs17817449 and rs8063946): haplotypes carrying the ancestral alleles at both SNPs (G at rs17817449 and C at rs8063946) are coloured yellow, those carrying both the derived alleles (T at rs17817449 and T at rs8063946) are coloured purple and those carrying the derived allele at rs17817449 (T) and the ancestral allele at rs8063946 (C) are coloured green. The vertical blue dashed line indicates the focal position and the red lines indicate the positions of the SNPs rs17817449 and rs8063946. Comparing the SNP data in the HapMap with the three risk haplotypes induced by the best-fitting two mutations (red and green) under the 2-mutation model, we find that the risk haplotypes are perfectly correlated with the haplotypes induced by the SNPs rs17817449 and rs8063946, as can be appreciated visually. Specifically, the haplotypes carrying the red mutation are perfectly correlated with the haplotypes that carry the T allele at rs8063946 and the haplotypes carrying the green mutation are perfectly correlated with the haplotypes that carry the G allele at rs17817449. The contingency table in the top right of Supplementary Figure 16 lists in each column the expected number of mutant alleles carried by the case and control individuals. The text colour of each column corresponds to the colour of the mutation that it refers to, for example the first column of the top table lists the expected number of case and control haplotypes carrying the blue mutation under the 1-mutation model, and the last column (in black text) lists the expected number of haplotypes carrying no mutations. The relative risks of each mutation, relative to the lack of a mutation, can be calculated based on the allele counts and is listed in the last row of each table. The red mutation (corresponding to the T allele at rs8063946) is estimated to be protective and the green mutation (corresponding to the G allele at rs17817449) is estimated to be deleterious.

We also used subsets of the phased case-control haplotypes as the reference panel with GENECLUSTER, which give similar results as above. The maximum $\log_{10}(\mathrm{BF}_{\text{2-mutation}})$ varies from 6.52 to 7.12 and $\log_{10}(\mathrm{BF}_{\text{secondary}})$ varies from 0.5 to 1.26, which translates to a posterior probability of 0.76 to 0.95 (assuming a prior probability of 0.5). In all analyses, the risk haplotype groups induced by the pair of most likely mutations under the 2-mutation model are perfectly correlated with the haplotypes induced by the SNPs rs17817449 and rs8063946, with estimated relative risks similar to those reported above.

We carried out further analyses to try to account for the risk haplotypes identified by GENECLUSTER using a single mutation model. A superior model fit from a single mutation is likely to result from a mutation that falls above all of the low risk TT (or the high risk GC) haplotypes and a subset of the intermediate risk TC haplotypes. Supplementary Figure 17(a) shows the original HapMap tree used in our analyses. The blue and the pair of red and green dots show the locations of the most likely mutations (a posteriori, out of all possible locations for a single mutation and pair of mutations) under the 1-mutation and 2-mutation models respectively. There is another branch in this tree that falls above all of the low risk TT haplotypes and a subset of the intermediate risk TC haplotypes, which is indicated by the blue dot in Supplementary Figure 17(b). To account for potential inaccuracies in the way that the HapMap tree was estimated, we also re-ordered the coalescent events near the root of the original tree to create a new tree with a branch, which has the required property and is illustrated in Supplementary Figure 17(c). If we assume an association model where the red and green mutations each confer independent haplotype penetrances with beta prior distributions, then $\log_{10}\mathrm{BF} = 8.25$ against the null model. Similarly, if we assume that the blue mutation confers a haplotype penetrance with same beta prior, then the single mutations in Supplementary Figures 17(a), 17(b) and 17(c) result in $\log_{10}\mathrm{BF}$ of 6.80, 5.49 and 2.92 against the null model respectively. Therefore, the best-fitting two mutations model remains superior over all single mutation models that we have encountered. Similar analyses with trees of other reference panels, consisting of subsets of phased case-control haplotypes, give the same result.

The secondary SNP, rs8063946, correlates with a mutation that is near the root in our trees, suggesting that it is older than one might expect given its MAF of $\sim 5\%$. The minor allele, T, is also the protective allele at this locus contributing to the lowest risk haplotype. To examine this further, we looked at the haplotype frequencies among the most recent release of HapMap (merged phaseII and phaseIII data). It is clear that the T allele, along with the TT haplotype containing it, is at much higher frequency in both JPT+CHB and YRI lending support to the idea that it is not a recent mutation (Table 11). It remains unclear why this allele is at such low frequency in CEU if it is both old and protective.

The secondary SNP, rs8063946, has a large variation in allele frequency across different populations, as shown in Supplementary Table 11. To illustrate this we used the HapMap YRI haplotypes as the reference panel and the results are summarised in Supplementary Figure 18. Although, the results are similar to those in Supplementary Figure 16, the TT haplotypes (coloured purple) are substantially more common in the YRI panel.

Given that we observed a non-additive effect in this region earlier, we explored the possible relationship to the secondary SNP effect observed here. Comparing the additive and non-additive models at the main SNP across a range of possible joint models with the secondary SNP (null, additive, general) consistently gave similar p-values, in the range 0.08 to 0.1. In

Supplementary Table 11: **HapMap II + III merged and phased haplotype frequencies for *FTO*.**

RR	rs17817449	rs8063946	CEU (2N=234)	JPT+CHB (2N=340)	YRI (2N=230)
1.68 (1.40–2.02)	G (ancestral)	C (ancestral)	0.453	0.162	0.396
1.37 (1.15–1.65)	T (derived)	C (ancestral)	0.496	0.371	0.191
1.00 (ref.)	T (derived)	T (derived)	0.051	0.468	0.413

Supplementary Table 12: **Haplotype disease model for T2D, *CDKN2A*.** The haplotypes are defined by rs10811661 and rs10217762 and are ordered by increasing risk.

Haplotype	Frequency	Relative risk
00	0.16	1.00 (ref.)
11	0.59	1.19 (1.06–1.34)
10	0.25	(1.31–1.69)

other words, there is suggestive evidence of a deviation irrespective of the secondary SNP. Doing the same with the roles of the SNPs reversed showed there was no evidence of a deviation at the secondary SNP (p-values in the range 0.7 to 0.8).

7.3.2 T2D, *CDKN2A*

The best SNP in this region is rs12555274 and the best SNP on the conditional tests is rs10965250 with a p-value of 4.1×10^{-5}.

It turns out that a haplotypic effect has previously been observed in this region, with three haplotype backgrounds conferring different disease risks.[6,7] The two SNPs rs10811661 and rs10217762 together distinguish these haplotypes well. The SNPs highlighted by the conditional analysis capture essentially the same effects, although not quite as well as the pair just given. This best SNP manages to distinguishes the high-risk haplotype fairly well, but blurs the distinction between the other two somewhat (in a way that allows it to capture a part of the risk difference), so is not an optimal choice when part of a pair of SNPs.

Taken together, rs10811661 and rs10217762 result in only 6 different genotype-pair combinations in the data. This therefore corresponds to 3 underlying haplotypes, for which phase is unambiguous, thus a haplotype-based analysis using these two SNPs is equivalent to one based on genotypes. Details of this model are shown in Supplementary Table 12.

Comparing the two-SNP additive model to the two-SNP general model (here, 3 parameters vs 6 parameters) gives a p-value of 0.5, thus the additive model provides an adequate fit. Details of this model are shown in and Supplementary Figure 14.

Supplementary Figure 19 summarises the results of the GENECLUSTER analysis with the

HapMap CEU haplotype reference panel. The focal position is at 22124000, where $\log_{10}(\text{BF}_{\text{2-mutation}}) = 6.10$ and $\log_{10}(\text{BF}_{\text{1-mutation}}) = 4.50$, which leads to $\log_{10}(\text{BF}_{\text{secondary}}) = 1.6$ and a posterior probability for a secondary effect of 0.98 (assuming a prior probability of 0.5). The haplotypes in the reference panel on the bottom left are coloured according to the alleles at rs10811661 and rs10217762: haplotypes carrying the TC alleles (T at rs10811661 and C rs10217762), CC and TT are coloured green, yellow and purple respectively. We find that the haplotypes induced by the SNPs rs10811661 and rs10217762 are highly correlated with the risk haplotypes induced by the most likely mutations (red and green) under the 2-mutation model, as can be appreciated visually. Specifically, the haplotypes carrying the green mutation are highly correlated ($r^2 = 0.97$) with the haplotypes that has the C allele at rs10217762, and the haplotypes carrying both the red and green mutations are perfectly correlated with the haplotypes that has the C allele at rs10811661. The relative risk estimates lie within the confidence intervals listed in Supplementary Table 12 and indicate that the TC haplotypes are deleterious and the CC haplotypes are protective relative to the TT haplotypes.

Using subsets of the phased case and control haplotypes as the reference panel gives similar results as above. The maximum $\log_{10}(\text{BF}_{\text{2-mutation}})$ varies from 5.62 to 6.80 and $\log_{10}(\text{BF}_{\text{secondary}})$ varies from 0.61 to 1.25, which translates to a posterior probability for secondary effects of 0.80 to 0.95 (assuming a prior probability of 0.5). In all analyses, the most likely mutations under the 2-mutation model induce haplotypes that are highly correlated with the haplotypes induced by SNPs rs10811661 and rs10217762 ($r^2 > 0.95$).

As with FTO, we looked at possible single mutation models, either by placing a mutation in the estimated HapMap tree, or a slightly altered tree, at the focal position that can account for the haplotype risk groups that GENECLUSTER identified. However, we did not find a single mutation that provides a better model fit than the most likely mutations under the 2-mutation model.

7.3.3 T2D, *CDKAL1*

The best SNP in this region is rs7756992. The best SNP on the conditional tests (and the only one passing our specified threshold) is rs6456360 with a p-value of 8.7×10^{-5}.

We explored a range of models with these two SNPs. All 9 possible genotype pairs are observed in the data, thus a saturated model has 9 parameters. The simpler, additive model has 3 parameters. Comparing the two gives a p-value of 0.8, thus the additive model adequately explains the data. Details of this model are shown in Supplementary Table 13 and Supplementary Figure 15.

Using phased haplotypes, we explored the possibility that a haplotype model might better explain the observed effects. There are 10 possible two-SNP haplotype combinations and all of these are observed in the data (the extra parameter is due to the phase ambiguity at the double heterozygote). An additive model on haplotypes has 4 parameters (one for each possible haplotypes). Comparing the saturated and additive models gives a p-value of 0.8. Furthermore, comparing the additive model on haplotypes to that of SNPs (4 parameters vs 3 parameters) gives a p-value of 0.9. Thus, we conclude that the SNPs on their own adequately describe the effect.

Supplementary Table 13: **Two-SNP additive disease model for T2D, *CDKAL1*.**

SNP	MAF	Relative risk
rs7756992	0.28	1.28 (1.18–1.40)
rs6456360	0.49	1.17 (1.08–1.26)

The GENECLUSTER analysis of *CDKAL1*, using 120 HapMap CEU haplotypes as the reference panel, resulted in a maximum $\log_{10}(\mathrm{BF}_{\text{2-mutation}})$ of 6.17 and $\log_{10}(\mathrm{BF}_{\text{2-mutation}})$ of 0.16, which translates to a modest posterior probability of 0.59 (assuming a prior probability of 0.5). Using subsets of the phased cases-control haplotypes as the reference panel, which are genotyped at a denser set of SNPs than the HapMap, gave a maximum $\log_{10}(\mathrm{BF}_{\text{2-mutation}})$ between 6.42 and 6.88, and $\log_{10}(\mathrm{BF}_{\text{secondary}})$ between 0.25 and 0.58, which translates to a posterior probability of between 0.64 and 0.79 for a secondary effect (assuming a prior probability of 0.5). In all analyses, the most likely mutation under the 1-mutation model and one of the best fitting mutations under the 2-mutation model induces a risk haplotype in the reference panel that is highly correlated with the SNP rs7756992 ($r^2 > 0.99$), but a second SNP that corresponds to the other most likely mutation could not be found. The lack of support for the conditional tests from GENECLUSTER (compared to the other two regions) can, at least in part, be explained by the fact that the two identified SNPs, rs7756992 and rs6456360, are separated by recombination event(s) in the case-control sample, which means that the mutations that create those SNPs do not occur on the same marginal tree.

8 Region Specific Results

We turn now to findings in each specific fine mapping region. We focus here on *TCF7L2, CDKN2A/B, FTO, CDKAL1, HHEX*, for T2D, *CDKN2A/B*, and *SORT1* for CAD, and *CTLA4* for GD, namely the regions where the fine mapping experiment added information. Further information on the other regions follows (*JAZF1* for T2D, *1q41, 2q36*, and *CXCL12* for CAD, *FCRL3* and *IL2RA* for GD).

8.1 Type 2 diabetes

8.1.1 *TCF7L2*

Sequencing and fine-mapping efforts to date have shown that the T2D-association signal,[8] which remains the strongest common variant effect for T2D-predisposition,[9–13] maps to a 64kb interval including exon 4 and flanking introns, and have implicated rs7903146 as the strongest causal candidate.[14]

In our analysis, rs7903146 remains the best SNP after fine mapping (relative risk $RR = 1.40$; $95\%CI = 1.29 - 1.52$), accounting for 75% of the posterior weight. There are four other SNPs in the 95% credible set (6 in the 99% set), all mapping within the largest intron of *TCF7L2*. rs7903146 maps to an enhancer region active in pancreatic islets identified using FAIRE

(Formaldehyde-Assisted Isolation of Regulatory Elements), and has allele-specific differences in both islet enhancer activity and chromatin accessibility.[15] This region showed no evidence for a secondary signal, nor for departures from a multiplicative disease model. There is evidence for extended haplotype homozygosity in East Asian samples in the Human Genome Diversity Panel (HGDP) in the association interval which is suggestive of a selective sweep (see SoM for details).

8.1.2 *CDKN2A/B*

T2D association with variants in the *9p21* region close to *CDKN2A/B* was identified through the GWA analysis of WTCCC, DGI and FUSION samples.[9,10,12,13] Based on functional data from mouse,[16] the T2D-association signal is likely to involve gain of function of *CDKN2A*, potentially mediated via the non-coding RNA *ANRIL* (or *CDKN2BAS*).[17] The variants influencing susceptibility to CAD[18–20] and aneurysm formation[21] at this locus are distinct from those involved in T2D.

In our analysis, the best T2D-associated SNP after the fine mapping is rs12555274 ($RR = 1.26$; $95\% CI = 1.15 - 1.37$), accounting for 68% of the posterior weight. There are 4 other SNPs in the 95% credible set.

However, more detailed analysis indicates a more complex picture. Previous suggestions that this region contains multiple signals[13,22] are confirmed following this denser fine-mapping. There are two key sets of SNPs, (rs10811661, rs2383208, rs10965250, rs10811660) and (rs10217762, rs10757283, rs7019778), the SNPs within each set displaying strong correlations. The model including rs12555274 alone fits the data significantly less well than that based around pairs of SNPs.

As expected in a region of very low recombination, only three of the four possible haplotypes involving these two sets of SNPs occur with any substantial frequency in the data, and haplotype phase at the pair of SNPs is uniquely determined by the SNP genotypes. As a result, statistical models in which disease risk at this locus depends on the two-SNP haplotypes cannot be distinguished from those in which it depends on the pair of genotypes at the two SNPs. For convenience we describe results in terms of the haplotypes, described explicitly by the pair of SNPs rs10811661 and rs10217762. The three haplotypes present are CC, TT, and TC with frequencies 16%, 59%, and 25%, respectively, listed in increasing order of disease risk, with the unobserved CT haplotype being ancestral (that is, the haplotype which was present before the mutation events giving rise to the SNPs at these positions). There is no evidence at this locus of departures from the simple model in which each additional copy of a haplotype increases disease risk in a multiplicative way. Estimated relative risks for the three haplotypes are 1, 1.19 ($95\% CI = 1.06 - 1.34$), and 1.49 ($95\% CI = 1.31 - 1.69$). We note that the relative risk for the high-risk haplotype, namely 1.49, is at the high end for recent GWAS findings for common complex diseases.

The best single SNP in our data (rs12555274) has a substantially elevated BF compared to those previously typed including those that define these haplotypes. However, whilst rs12555274 tags the high-risk haplotype reasonably well, it fails to distinguish between the two lower-risk haplotypes. Performing a conditional analysis with this top SNP identifies a secondary signal at rs10965250 which is highly significant. This pair of SNPs capture

essentially the same effects as those described in terms of haplotypes above, although not quite as well.

Although the best single SNP in our data does not explain the data as well as the pair of SNPs rs1081161 and rs10217762, we cannot in principle exclude the possibility that there is another, untyped, single SNP which explains the data better than this pair of SNPs. However, careful examination of the estimated or likely genealogical tree at this locus reveals that it is unlikely that the effect we see could be explained by any single SNP (see SoM for details). None of the SNPs described has compelling annotations.

8.1.3 *CDKAL1*

The association between variants mapping close to *CDKAL1* and T2D was identified in several GWA studies.[9,10,12,13,23] The mechanism of action is unclear though *CDKAL1* has homology with *CDK5RAP1*, a known inhibitor of *CDK5* activation; in turn *CDK5* has been implicated in the regulation of pancreatic beta cell function.[24,25] *CDKAL1* variants implicated in psoriasis and Crohns disease[26,27] are distinct from those associated with T2D.

The best SNP after the fine mapping is rs7756992 ($RR = 1.29; 95\% CI = 1.19 - 1.40$), accounting for 35% of the posterior weight. The risk-variant previously highlighted at this locus, rs10946398 ($r^2 = 0.73$ with rs7756992) accounts for only 1.7% of the posterior in our analysis. There are 32 other SNPs in the 95% credible set. This region showed no evidence for departures from a multiplicative disease model.

Conditional analysis reveals an additional signal at SNP rs6456360 ($RR = 1.16; 95\% CI = 1.08 - 1.25$), with MAF in controls of 0.47 ($p = 8.7x10^{-5}$ for comparing the model with both the primary signal at SNP rs7756992 and rs6456360 vs. the model with just the primary signal). The data are consistent with a model in which risks at the primary SNP(s) and this second SNP combine multiplicatively. Unlike the *CDKN2A/B* locus, because all genotype combinations occur in the data, we could check here whether a haplotype model, potentially with interactions between the pair of haplotypes in an individual, fitted the data better than the multiplicative 2-SNP model, but there was no evidence that it did (see SoM). Analyses of the joint risks at rs7756992 and rs6456360 estimate the relative risk for each additional copy of both risk alleles (that is comparing the risk of a haplotype carrying risk alleles at both SNPs to that of a haplotype carrying protective alleles at both SNPs) as 1.50 ($95\% CI = 1.34 - 1.68$), large by GWAS standards, and considerably larger than the effect considering the top SNP alone.

All SNPs in this region are located in the third and fourth introns of *CDKAL1*. 17 SNPs in the 95% credible set are within a block with evidence of extended haplotype homozygosity in the HGDP samples from the Mideast, Europe, and South Asia. Four SNPs (rs9460544, rs9460545, rs4712526, rs35456723) in the 95% set show conservation and regulatory potential, with one (rs35456723) among the most conserved regions in a 28-way alignment of mammalian genomes. The entire 99% credible set is within a broad domain containing histone modifications associated with active transcription.

8.1.4 *FTO*

The association between *FTO* variants and T2D[12] is mediated through a primary effect on body mass index and risk of obesity.[28–30] The associated variants map to a 50kb interval in intron 1 of the *FTO* gene, and recent evidence from both mouse and human studies, implicates *FTO* as the gene through which the effect is mediated.[31,32] In the analyses here, we investigate SNPs in the region in terms of their effect on the endpoint of T2D, rather than directly on obesity (as BMI information was not available on many of our samples). We would expect use of T2D rather than BMI as an endpoint to attenuate the primary signal somewhat, but would not expect it to affect broad conclusions.

There is a set of 33 SNPs, for which all pairwise r^2 values are at least 0.95: together, these account for 95% of the posterior weight after fine mapping. The *RR* associated with the top SNPs is 1.26 ($95\%CI = 1.16-1.36$). A single common haplotype spans the entire region, with multiple SNPs segregating with the haplotype (see Supplementary Figure 10). In settings such as this, fine mapping focuses attention on the haplotype background, but will not be able to distinguish between the correlated SNPs unless very large numbers of samples are typed or transethnic analyses provide access to distinct haplotypic structures. Functional annotations for the associated SNPs offer few clues to which of these might be causal.

Conditional analysis on rs17817449 reveals an additional signal at SNP rs8063946 with MAF in controls of 0.06 and *RR* 1.37 ($95\%CI = 1.14 - 1.64; p = 5.9 \times 10^{-4}$ comparing the model with both the primary signal at SNP rs17817449 and rs8063946 vs. the model with just the primary signal). Since, as at *CDKN2A*, only three of the four possible haplotypes occur in the data, phase can be determined unambiguously, and it is impossible to distinguish haplotype-based models from SNP-based models. The three haplotypes at rs17817449 and rs8063946 are GC (ancestral), TC, and TT, with respective relative risks of 1.68 ($95\%CI = 1.40-2.02$), 1.37 ($95\%CI = 1.15 - 1.65$), and 1. Note that the relative risk of the high risk haplotype, at 1.66, is high by GWAS standards, although the low frequency (5% in CEU Hapmap) of the low-risk TT haplotype makes precise estimation of this *RR* difficult. The frequencies of these haplotypes differ substantially between populations. Most obviously, the derived, protective TT haplotype emerges as the most common haplotype in JPT/CHB (47%) and YRI (41%) (Supplementary Table 11). Four SNPs in the 95% set (rs62033408, rs10468280, rs7202296, rs7202116) have high Fst in the Human Genome Diversity Panel (HGDP), adding further weight to the evidence for population differentiation at this locus. Investigation of disease models at the top SNP (rs17817449), shows some evidence of departure from the simple multiplicative model ($p = 0.099$) with estimates suggesting any possible departure lies in the direction of a dominant model (Supplementary Figure 11).

We also undertook extensive additional haplotype-based analyses at *FTO* (see SoM for details). If there were untyped SNPs affecting disease risk, these should manifest themselves through distinct haplotypes having different risks. It has recently been suggested that many GWAS associations with common SNPs are in fact synthetic, with the signal being driven by rare or low frequency SNPs that have been escaped detection and genotyping to date.[33] The clear haplotype structure of the *FTO* region facilitates an examination of this issue. Our haplotype analyses in effect recapitulated the effects of the primary and secondary SNPs and did not reveal evidence for additional SNPs, rare or common.

We note that because we have undertaken analyses with T2D as the endpoint, we cannot assess whether the effect of the second SNP on T2D is mediated entirely via an effect on BMI (as for the primary signal), or partially, or is independent of BMI.

8.1.5 *HHEX*

The association between variants mapping within an interval containing the genes *HHEX, KIF11* and *IDE* was first reported in a GWA scan of French subjects[11] and subsequently confirmed in other studies.[9,10,13] The recombination interval contains two genes with strong biological candidacy for a role in T2D pathogenesis: *HHEX* encodes a homeobox protein implicated in pancreatic development,[34] and *IDE* (insulin degrading enzyme) has a role in insulin metabolism.[35,36] Fine-mapping will be important in helping to define which of these is responsible for the association effect.

The best SNP after the fine mapping is rs10882098 ($RR = 1.21; 95\% CI = 1.12 - 1.31$), accounting for 20% of the posterior weight. This SNP is well correlated ($r^2 = 0.73$) with the previously reported lead SNP for this region. There are 13 other SNPs in the 95% credible set, of which two (rs10882099, rs10882106) appear to lie in polymerase II binding sites: rs10882106 is also in a region of high conservation and suggestive DNase hypersensitivity. As regards signals of selection and population differentiation, one SNP (rs7923866) in the 95% credible set (11 in 99% set) shows high Fst in HGDP, and the entire associated region, including *HHEX, IDE*, and *KIF11*, shows extended haplotype homozygosity among East Asian samples, suggesting the effect of a selective sweep. This region showed no evidence for a secondary signal or departure from a multiplicative disease model.

8.2 Coronary artery disease

8.2.1 *CDKN2A/B*

The association of the chromosome *9p21* locus with CAD has been confirmed in multiple studies in several ethnic groups.[37] The locus has also been shown to be associated with risk of abdominal aortic and intra-cranial aneurysms.[21] The association signal overlaps the location of a gene for a non-coding RNA, *ANRIL*, with the strongest signals in the GWAS studies located at the 3 end of the *ANRIL* gene. The function of *ANRIL* is unknown but the correlation of its transcript level with those of the adjacently located cyclin-dependent kinase inhibitors (*CDKN2A/B*)[17] suggests that it may be involved in the regulation of these important cell-cycle regulators. As cell proliferation is a key feature of atherosclerosis, this could be the mechanism for the association of the locus with CAD, and a recent study reported that deletion of the homologous region in mice is associated with reduced *CDKN2A/B* expression and higher proliferation rates of vascular smooth muscle cells from such mice.[38]

The best SNP after the fine mapping is rs1537370 ($RR = 1.40; 95\% CI = 1.29 - 1.51$), accounting for 26% of the posterior weight. Two additional SNPs highly correlated with the top SNP (rs10116277 and rs6475606) each account for similar posterior weight, and there are 12 other SNPs in the 95% credible set. This region showed no evidence for a secondary signal. Estimation of effect sizes suggests there may be a departure from the multiplicative disease

model (Supplementary Figure 11), but this is not statistically significant in our data. We do note that under the more general 2df statistical model for effects at this locus, the peak of the signal moves somewhat (See SoM for details).

All SNPs in the 99% credible set for this region are within the 3 half of *ANRIL* and within a region showing weak evidence of selective sweep in the Bantu samples of HGDP. Our top SNP (rs1537370) appears to be in a nuclease accessible domain and has evidence for polymerase II binding. It also has evidence of histone modifications associated with active transcription including H3K4me3 and H3K9ac in human mammary epithelial cells (HMEC) and H3K27ac in HMEC, normal human epithelial keratinocytes (NHEK) and normal human lung fibroblasts (NHLF). Four SNPs (rs1333045, rs10757278, rs10757279, rs10757277) in the 95% credible set appear to be in regions that bind *TCF7L2* and one shows some evidence of conservation (rs10757278).

A recent study of this region[39] involved sequencing and subsequent assessment of potential functional roles for variants correlated with the top CAD SNPs from GWAS. It then focussed on two variants, rs10811656 and rs10757278, located in one of the many enhancers in the region. These SNPs disrupt a binding site for *STAT1*, a member of the Signal Transducers and Activators of Transcription family of transcription factors. *STAT1* is involved in upregulating genes due to a signal by either type I, type II or type III interferons.[40] In lymphoblastoid cell lines homozygous for the CAD risk haplotype, which has a disrupted predicted *STAT* binding site, no binding of *STAT1* occurs. In lymphoblastoid cell lines homozygous for the CAD non-risk haplotype, binding of *STAT1* inhibits *CDKN2BAS* expression, which is reversed by siRNA knockdown of *STAT1*.[39] Taken together, the data suggests a link between the CAD susceptibility and the response to inflammatory signalling.

We attempted to genotype both of these SNPs as part of our fine mapping experiment, but only obtained high quality genotype data for rs10757278. The second SNP, rs10811656, is almost perfectly correlated with several SNPs we typed, including rs10757278, so it is easy to impute well, and the signal and posterior weights of both SNPs will be very similar. Our fine mapping experiment revealed a set of three SNPs to have much higher posterior weights (25% each) than rs10757278 and rs10811656 (<5% each, correlation of rs10757278 and rs10811656 with the top FM SNPs is $r^2 = 0.78$).

8.2.2 *SORT1*

Although studied here in the context of CAD, the locus exerts its primary effect on levels of plasma low density lipoprotein cholesterol (LDL-C), a major risk factor for CAD. The locus provides a useful positive control for our analysis, because the functional variant has been identified[41] (independently of our genotyping experiment).

The functional variant, rs12740374, has been shown to create a *C/EBP* (CCAAT/enhancer binding protein) transcription factor binding site which affects the expression of *SORT1* which in turn affects intracellular apolipoprotein B processing.[41] We designed for the functional SNP in our assay, but unfortunately the SNP failed genotyping. Figure 2 shows the result of imputation from the 1000 Genome data set (June 2011 release) into our fine mapping data in this region. With CAD as the phenotype of interest, the imputed version of the functional SNP is the sixth strongest association signal (posterior probability 0.035), and the SNP would

have been included in the 95% credible set based on the imputed data (see Supplementary Figure 9). Our data do not allow an assessment of the relative contributions of imputation error, and our use of an indirect phenotype, to the failure to see a large posterior probability at the causal SNP.

8.3 Graves disease

8.3.1 *CTLA4*

Early studies of *CTLA4* in Graves disease (GD) reported association of a limited number of variants, including, a dinucleotide (AT)n repeat within the 3 untranslated region of exon 3,[42] an A/G SNP within exon 1 (Ala/Thr position 17)[43] a C/T SNP within the *CTLA4* promoter region -318bp from the ATG start codon[44] and a C/T SNP within the non-coding intron 1 of *CTLA4*.[45] A more detailed fine mapping study genotyped 108 SNPs across 330 kb containing *CD28, CTLA4* and *ICOS*.[46] Association (maximum $OR = 1.51$) was refined to within a 6.1 kb region, 3 of *CTLA4*, with four SNPs, rs3087243 (CT60), rs11571302 (JO31), rs7565213 (JO30) and rs11571297 (JO27_1) highly associated with GD.[46] In 672 cases and 844 controls the rs3087243 A>G SNP was nominally more associated with GD and the T1D susceptibility allele (A) of rs3087243 was associated with reduced mRNA levels of a soluble *CTLA4* isoform.[46] A recent meta-analysis of 10 studies (4,906 GD subjects and controls) and two of the SNPs studied commonly in the literature supported findings that rs3087243 is the most associated SNP and calculated a combined $OR = 1.49$ for rs3087243.[47] It is noted, however, that accurate genotyping of the functional candidate, the polymorphic (AT)n microsatellite variants near the main polyA site of the CTLA4 3UTR, in large numbers of samples has not been reported, and the entire region of linkage disequilibrium (LD) has not been thoroughly re-sequenced to reveal all the variants that might be strongly associated with GD in the region.

The best SNP after our fine mapping is rs11571297 ($RR = 1.39; 95\% CI = 1.29 - 1.50$), accounting for 77% of the posterior weight. This SNP was not typed in many of the studies included the earlier meta-analysis. There are five other SNPs in the 95% credible set, with rs3087243, the top SNP from the earlier meta-analysis ranking in a group of three similarly ranked SNPs behind the top SNP. Its posterior weight was 5.2%. rs3087243, located near the 3 end of *CTLA4* also has some evidence of regulatory potential and conservation. This region showed no evidence for a secondary signal, nor for departures from a multiplicative disease model.

For this region, we were able to follow up rs11571297 and rs3087243 in additional samples (2,415 cases and 10,749 controls in the total analysis). This led to broadly similar conclusions, including for analyses in which sex and geographical location were used as covariates: both SNPs remained highly associated, with very similar estimates of RR, and with a comparison of Bayes Factors suggesting considerably stronger evidence for rs11571297 over rs3087243. We note the major functional candidacy of the 3UTR (AT)n microsatellite repeat which is in the main *CTLA4* transcript very close the polyA site. Unfortunately, owing to a PCR amplification bias for the smallest repeat allele, we have been unable to develop a genotyping assay for this variant that is in HWE in large control populations.

9 Region Specific Results for Additional Regions

9.1 Type 2 diabetes

9.1.1 *JAZF1*

Variants around the *JAZF1* gene were implicated in T2D-susceptibility through a meta-analysis of European GWA studies.[48] This association has since been confirmed in studies from non-European populations,[49] and the evidence supports mediation via reduced insulin secretion.[50] *JAZF1* is a transcriptional repressor of *NR2C2* and variants at this locus have also been described with associations to both height[51,52] and prostate cancer.[53,54]

Our fine mapping data failed to refine the signal in this region. The top SNP (rs12531540) accounted for 9% of the posterior weight, and 252 SNPs were included in the 95% credible set.

9.2 Coronary artery disease

9.2.1 *CXCL12*

The association of a locus at 10q11.21 with CAD observed in the combined analysis of the WTCCC and German MI I GWA studies[19] has been supported by further studies.[55,56] The signal lies upstream of the *CXCL12* gene which codes for stromal cell-derived factor-1 (*SDF-1*), a chemokine which plays a key role in stem-cell homing and tissue regeneration in ischemic cardiomyopathy[57] and in promoting angiogenesis through recruitment of endothelial progenitor cells.[58] There is some evidence that the effect of the locus on CAD risk may be stronger in women than in men[56] although this requires further confirmation.

Our fine mapping data failed to refine the signal in this region. The top SNP (rs34161818) accounted for 7% of the posterior weight, and 266 SNPs were included in the 95% credible set.

9.2.2 1q41

The association signal for CAD 1q41 has also been confirmed in additional studies.[55,56] It lies within the melanoma inhibitory activity family, member 3 (*MIA3*) gene. This gene has not been very well characterised functionally but may play a role in cell growth or inhibition.[59]

Our fine mapping data failed to refine the signal in this region. The top SNP (rs2936023) accounted for 4% of the posterior weight, and 240 SNPs were included in the 95% credible set.

9.2.3 2q36

The chromosome 2q36 association signal for CAD seen in the combined analysis of the WTCCC and German MI I Study[4] lies in a non-genic region. Further studies[55,56] have

shown nominally significant associations of the locus with CAD but not provided definite confirmation.

Our fine mapping data failed to refine the signal in this region. The top SNP (rs2673145) accounted for 6% of the posterior weight, and 86 SNPs were included in the 95% credible set.

9.3 Graves' disease

9.3.1 *FCRL3*

Association of the FC receptor like (*FCRL*) region with autoimmune disease was first detected within a Japanese rheumatoid arthritis cohort and with maximum effect seen with SNP rs7528684 within *FCRL3* (OR=2.15). In the same study association of rs7528684 was reported in a GD cohort (OR=1.79).[60] In a UK Caucasian cohort of 983 GD patients and 733 controls association with rs7528684 was also detected (OR=1.17) but at a much lower level than that seen in the Japanese cohort.[61] Results from the WTCCC 14,500 nsSNP screen detected association of rs7522061 SNP (which tagged rs7528684) $p = 2.1x10^{-4}$. Seven tag SNPs capturing 11 common variants (minor allele frequency > 0.05) within *FCRL3* (including one tag that captured rs7528684) were typed in the National AITD cohort consisting of 5000 Graves cases and controls (WTCCC and TASC, 2007, Nature Genetics, 39, 1329-1337). Out of these 7 tags, four SNPs (rs3761959, rs11264794, rs11264798 and rs11264793) were found to be associated with GD (P = 0.01 – 1.6 x 10^{-5}) with rs11264798 SNP now producing the strongest signal ($p = 1.6 \times 10^{-5}$, OR=1.22).

Our fine mapping data failed to refine the signal in this region. The top SNP, rs11264798, accounted for 7% of the posterior weight, and 114 SNPs were included in the 95% credible set.

9.3.2 *CD25/IL-2Rα*

An LD mapping approach, using tag SNPs, identified genetic association between type 1 diabetes (T1D) and the interleukin-2 receptor alpha (*IL-2Rα*)/*CD25* gene region.[62] Further fine mapping in T1D suggested localization of association to two independent groups of SNPs, spanning overlapping regions of 14 and 40 kb, including *IL2Rα* intron 1 and the 5' regions of *IL2Rα* and *RBM17*, producing a combined odds ratio of 2.[63] Employing the same 20 tag SNPs used in the original T1D study and applying a multilocus test upon a case-control study of 1896 GD cases and 1892 matched controls, evidence for association between GD and the *IL-2Rα* region was found ($p = 4.5x10^{-4}$), with the pattern of association similar to that seen in T1D[64] with SNPs rs7093069 ($OR = 1.15;95\%CI = 1.04 - 1.31$; $p = 2x10^{-3}$) and rs12722592 ($OR = 1.24;95\%CI = 1.09 - 1.45$; $p = 6x10^{-3}$). Subsequently, studies in T1D and in multiple sclerosis (MS) have reported evidence of three association signals in the *IL-2Rα* region, with SNPs, rs41295061, rs11594656 and rs2104286 being required to explain the risk of T1D region, and two of these, rs11594656 and rs2104286, being associated with MS risk.[65,66]

The top SNP after the fine mapping, rs10905669 ($RR = 1.20;95\%CI = 1.10-1.30$), accounted for 21% of the posterior weight, however there were 76 SNPs in the 95% credible set, so that

the experiment failed to exclude many possible SNPs.

10 Annotation

10.1 Methods

We used the Galaxy web platform to parse relevant SNP annotations found in the UCSC genome browser for the human genome release hg18.[67] All 247 SNPs contained in the 99% credible sets for the seven regions in which fine mapping was informative, along with SNPs contributing significant secondary signals to these regions, were cross-referenced for annotations in the following tracks:

- Genes and Gene Prediction Tracks
  - RefSeq Genes
    - refGene (extracted exons and introns)
  - Ensembl Genes
    - ensGene (extracted exons and introns)
  - Alt Events
  - RNA Genes
  - ACEScan
  - EvoFold
  - sno/miRNA
  - Pos Sel Genes
- Variation and Repeats Tracks
  - SNPs (130)
  - HGDP Smoothd Fst
  - HGDP iHS
  - HGDP XP-EHH
- Regulation Tracks
  - Broad Histone (peaks tables only)
  - EIO/JCVI NAS
  - Eponine TSS
  - First EF
  - GIS ChIP-PET (peaks tables only)
  - HAIB Methyl-seq

  - HAIB Methyl27
  - HAIB TFBS (peaks tables only)
  - NHGRI Bi-Pro
  - NHGRI NRE
  - Open Chromatin
  - ORegAnno
    * oreganno
  - SUNY RBP
  - SwitchGear TSS
  - TFBS Conserved
    * tfbsConsSites
  - TS miRNA sites
  - UW DNaseI HS
  - Vista Enhancers
  - Yale TFBS
  - 7X Reg Potential
  - FOX2 CLIP-seq
    * fox2ClipSeq
  - LI TAF1 Sites
  - NKI LADs (Tig3)
  - Nucl Occ: A375
  - Nucl Occ: Dennis
  - Nucl Occ: MEC
  - UU ChIP Sites
- Comparative Genomics Tracks
  - phastConst28wayPlacMammal
  - phastConst17way

Unless otherwise specified, all available tables for each annotation track were used.

Among the **Genes and Gene Prediction** track group, **RefSeq Genes**[68,69] and **Ensembl Genes**[70] tracks were used to annotate SNPs to coding and non-coding (UTR) exons (appearing as **refGene all exons** and **refGene coding exons** in annotation tables). Since both the first and the largest intron in each transcript often contains regulatory information, we wanted to quantify which of our intronic SNPs were in either the first and/or largest introns. A SNP

was annotated as **refGene largest intron** or **refGene first intron** if it was contained in largest or first intron of any annotated transcript from the refGene gene set respectively. The same criteria was used for the **ensGene largest intron** and **ensGene first intron** annotations using the Ensembl gene set. The **Alt Events** track (appearing as **alternative splicing events** in annotation tables), based on analysis from the `txgAnalyse` program by Jim Kent, contains annotations of various alternative splicing events. The **RNA Genes** track (appearing as **noncoding RNA** in annotation tables) contains all non-protein coding RNA genes including tRNAs, rRNAs, miRNAs, etc.[71,72] The **ACEScan** track contains predicted alternative conserved exons.[73] The **EvoFold** track (appearing as **EvoFold secondary structure** in the supplementary annotation tables) contains the predictions of the EvoFold program, a method which uses evolutionary conservation to help locate regions with RNA secondary structure.[74] The **sno/miRNA** track (appearing as **miRNA** in annotation tables) includes annotations for snoRNAs from snoRNABase and miRNAs from miRBase (formerly the miRNA Registry).[75,76] The **Pos Sel Genes** track (appearing as **positive selection on human branch p < 0.05** in the supplementary annotation table) shows the results of a genome-wide scan for positively selected genes from a six species sequence alignment. We focused on genes showing significant evidence of positive selection on the human branch of the phylogenetic tree.[77,78]

Among the **Variation and Repeat** track group, we used **SNPs (130)** to further annotate SNPs to genic functional positions (nonsynonymous, synonymous, intron, splicing, 5' utr, 3' utr, near-gene-5' and near-gene-3', unknown).[79] Human Genome Diversity Panel (HGDP) tracks were used to look for population differentiation and population-specific selection.[80–84]

Among the **Regulation** track group, many tracks were used to search for higher-order functional effects of SNPs. The **Broad Histone** track (appearing, for example, as **Broad H3K4me1 any -logp > 5** in the supplementary annotation table) contains regions with significant enrichment for histones with specific modification tags that affect chromatin conformation and thus local gene expresion.[85–87] The data are provided as part of the ENCODE project and each annotation is available for numerous cell types. For this and all other ENCODE annotation data, we used only data available for unrestricted usage as of February 2010. For our analysis, annotations were combined across cell types. A SNP was scored with an annotation if, for any cell type, it was within a peak with significant ($-\log_{10} p > 5$) ChIP-Seq enrichment. The **EIO/JCVI NAS** track (appearing as **EIO/JCV Nucleosome Accesibility CD34** in the supplementary annotation table) provides maps of nuclease hypersensitive regions, known to correlate with functionally active cis-regulatory elements.[88] The **Eponine TSS** track contains annotations of predicted transcription start sites.[89] The **firstEF** annotation track contains the predictions of the first Exon Finder program for the locations of first exons, promoters, and CpG windows.[90] The **GIS ChIP-PET** track provides transcription factor binding site maps for *p53*, *STAT1*, *c-Myc*, along with histone modifications H3K4me3 and H3K27me3.[91–95] The three **HAIB** tracks provide Hudson Alpha's ENCODE data of transcription factor binding sites mapped using ChIP-Seq.[96,97] **NHGRI BiPro**[98,99] and **NHGRI NRE**[100] tracks contain the National Human Genome Research Institute's maps of bidirectional promoters and negative regulatory elements respectively. The **Open Chromatin** track contains the data from the collaborative Duke/UNC/UT-Austin/EBI ENCODE group mapping accessible chromatin by DNaseI hypersensitivity (appearing as **Duke/UW DNaseHS p < 0.05**), Formaldehyde-Assisted Isolation of Regulatory Elements, and ChIP-chip of transcription factors (appearing combined with Broad ENCODE annotations, for

example, **Broad/Duke CTCF p < 0.05**)[88,101–104].[105–109] The **ORegAnno** track provides literature-curated regulatory regions.[110] **SwitchGear TSS**, created by Nathan Trinklein and Shelley Force Aldred, maps transcription start sites. The **TFBS Conserved** (appearing as **TFBScons**) track contains computationally predicted conserved transcription factor binding sites from a human, mouse, and rat genome alignment. It was created by Matt Weirauch and Brian Raney of UCSC using the Transfac Matrix and Factor databases.[111] **TS miRNA sites** (appearing as **miRNA target site**) contains conserved miRNA target sites in refSeq genes predicted by TargetScanS.[112–114] **UW DNaseI HS** (appearing as **Duke/UW DNaseHS p < 0.05**) provides the University of Washington ENCODE data on DNase hypersensitivity sites.[115,116] The track **Vista Enhancers** contains conserved non-coding sequences that have shown reproducible enhancer activity in a transgenic mouse reporter assay.[117] The **Yale TFBS** track contains the data from the collaborative Yale/UC-Davis/Harvard ENCODE group mapping binding sites for multiple transcription factors across multiple cell types.[118–122] Again, a SNP was scored with an annotation if, for any cell type, it was within a peak with significant ($q < 0.05$) ChIP-Seq enrichment. The **7X Reg Potential** track shows regulatory potential (RP) scores derived from an alignment of human, chimp, macaque, mouse, rat, dog, and cow genomes.[123,124] The score compares the frequency of short regulatory motifs in a region to that of neutral DNA. We include only regions with a score > 0.1, which is the suggested cutoff for regions with high regulatory potential. The **FOX2 CLIP-seq** (appearing as **FOX2 binding sites** in the supplementary table) track contains FOX2 CLIP-seq (cross-linking immunoprecipitation combined with high-throughput sequencing) reads from human embryonic stem cells uniquely mapped to the reference genome.[125] **LI TAF1 Sites** shows TAF1 binding sites.[126] **NKI LADs** shows lamina associated domains. The three **Nucl Occ** (A375, Dennis, and MEC) tracks (appearing as **UW Nucleosome Occupancy**) contain nucleosome occupancy scores determined by MNase digestion assays.[127–129] A375 and Dennis tracks are tuned to locate regions of high occupancy and MEC is tuned to locate regions of low occupancy. The threshold of 1.0 (-1.0 for MEC) corresponds to a confident prediction. Finally, the **phastCons17way** and **phastConst28wayPlacMammal** tracks (appearing as **17-way most conserved Vertebrate** and **28-way most conserved Mammals**) contain predicted conserved elements from a 17-way vertebrate genome alignment and a 28-way placental mammal genome alignment respectively.[130] **UU ChIP Sites** (appearing as **Uppsala USF1**, **Uppsala USF2**, and **Uppsala H3ac**) contains a map of *USF1*, *USF2*, and H3ac sites.[131,132]

10.2 Annotations of interest by region

10.2.1 T2D:*TCF7L2*

All 5 SNPs in the 95% posterior set (all 7 in 99% set) are within the largest intron of *TCF7L2*. The best SNP after the fine mapping, rs7903146 (accounting for 75% of the posterior weight) maps into a putative *FOX2* binding site identified by cross-linking immunoprecipitation coupled with high-throughput sequencing (CLIP-seq).[125] *FOX2* is known to be expressed in muscle and neuronal tissue and plays a role in tissue specific alternative splicing. One of the four other SNPs in the 95% credible set, rs4132670 (accounting for 1% of the posterior), is in a conserved domain with regulatory potential and apparent nucleosome occupancy. This domain also appears to be DNase/nuclease hypersensitive in the CD34- neuroblastoma cell line

SK-N-MC, and in CD34- maturating myeloid cells. There is also evidence for extended haplotype homozygosity in East Asia samples from the Human Genome Diversity Panel (HGDP) in this intron which is suggestive of a selective sweep.

10.2.2 T2D:*CDKN2A*

Looking in to SNP annotations for this region, we find that two SNPs in the 99% credible set (rs7866783 and rs564398 with posterior probabilities of 0.05% and 0.02% respectively) are in exons of the non-coding antisense RNA (*CDKN2BAS*, also known as *ANRIL*).

10.2.3 T2D:*CDKAL1*

All SNPs in this region are located in the third and fourth introns of *CDKAL1*. 17 SNPs in the 95% credible set are within a block with evidence of extended haplotype homozygosity in the HGDP samples from the Mideast, Europe, and South Asia. Four SNPs (rs9460544, rs9460545, rs4712526, rs35456723 accounting for 2%, 2%, 2%, 1% of the posterior respectively) in the 95% set show conservation and regulatory potential, with one (rs35456723) among the most conserved regions in a 28-way alignment of mammalian genomes.

We find that 1 SNP in the 99% credible set (rs35612982 with a posterior probability of 0.4%) shows high Fst in the HGDP.

10.2.4 T2D:*HHEX*

Two SNPs among the 95% set (rs10882099 and rs10882106 accounting for 9% and 4% of the posterior respectively) appear to be in polymerase II binding sites. rs10882106 is also in a region of high conservation and suggestive DNase hypersensitivity. Looking for signals of selection and population differentiation, we find that 1 SNP (rs7923866 accounting for 4% of the posterior) in the 95% credible set (11 in 99% set) shows high Fst in HGDP. Also, the entire associated region, including *HHEX*, *IDE*, and *KIF11* shows extended haplotype homozygosity among East Asian samples, suggesting the effect of a selective sweep.

5 SNPs in the 99% credible set are also in untranslated gene regions (1 in the 5 UTR of *IDE* and 4 in the 3 UTR of *KIF11*).

10.2.5 CAD:*CDKN2A*

All SNPs in the 99% credible set for this region are within the 3 half of *ANRIL* and within a region showing weak evidence of selective sweep in the Bantu samples of HGDP. Our top SNP (rs1537370 accounting for 26% of the posterior) appears to be in a nuclease accessible domain and has evidence for polymerase II binding. Perhaps of interest is the fact that four SNPs (rs1333045, rs10757278, rs10757279, rs10757277 accounting for 5%, 4%, 4%, and 3% of the posterior respectively) in the 95% credible set appear to be in sites that bind *TCF7L2* and one shows some evidence of conservation (rs10757278).

11 *CTLA4* further analysis

For this region we were able to follow-up our top SNP JO27_1 (rs11571297), along with our fourth best SNP CT60 (rs3087243), in additional samples.

In our original analysis including the *extended reference panel*, we found $\log_{10}$BFs of 16.08 and 14.89 for rs11571297 and rs3087243 respectively. A previous fine mapping analysis suggested that these two SNPs were among a set of four highly correlated, nearly indistinguishable SNPs.[46] In our analysis, rs3087243 is ranked fourth, with a $\log_{10}$BF similar to the second (rs11571302, 14.95) and third (rs1968351, 14.93) ranked SNPs.

In our follow-up analysis, we included geographic location and sex as covariates and found that although these covariates account for some of the association signal observed, both SNPs remain highly associated, with rs11571297 still retaining a higher $\log_{10}$BF (17.484 versus 16.496 respectively on 2,415 cases and 10,749 controls).

12 Imputation

For imputation we used as reference panel the 286 haplotypes for Caucasian samples (defined as samples from CEU,GBR,IBS and TSI populations) from the June 2011 release of the 1,000 Genomes project. The haplotypes for the reference panel are available from www.1000genomes.org. We used the software package IMPUTE v1.1.5[2] to perform the imputation. Post-imputation quality control consisted of excluding imputed SNPs with either (i) with average maximum posterior (as returned by IMPUTE1) less than 0.98, or (ii) IMPUTE1 info score less than 0.8, or (iii) greater than 2% missing data. Imputed data was analysed for association in the program SNPTEST.

In addition to the discussion in the main text, we assessed the coverage of variants with lower MAF. Table 4 shows that our coverage, via genotyping or imputation, of variants with MAF $< 1\%$ is much lower, but unless the effect sizes for such variants are large we would have limited power to detect them with our study size even if we had genotyped or imputed them. Whether the lack of coverage for rare ($MAF < 1\%$) variants undermines our main conclusions then depends on how likely it is that many or all of the GWAS signals in our fine mapping regions were driven by rare variants. While it has been suggested that this might be the case for GWAS regions in general,[33] we did not see any evidence for this where we could check indirectly (see *FTO* region), and several published lines of evidence also argue against this possibility.[133,134]

References

[1] Teo, Y. Y. *et al.* A genotype calling algorithm for the illumina beadarray platform. *Bioinformatics* **23**, 2741–2746 (2007). URL `http://bioinformatics.oxfordjournals.org/content/23/20/2741.abstract`. `http://bioinformatics.oxfordjournals.org/content/23/20/2741.full.pdf+html`.

[2] Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* **39**, 906–913 (2007).

[3] Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* **3**, e114 (2007).

[4] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

[5] Su, Z., Cardin, N., the Wellcome Trust Case Control Consortium, Donnelly, P. & Marchini, J. A Bayesian method for detecting and characterizing allelic heterogeneity and boosting signals in genome-wide association studies. *Statistical Science* **24(4)**, 430–450 (2009).

[6] Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).

[7] Su, Z. *Statistical Methods for the Analysis of Genetic Association Studies*. D.Phil. thesis, Balliol College, University of Oxford (2008).

[8] Grant, S. *et al.* Variant of transcription factor 7-like 2 (tcf7l2) gene confers risk of type 2 diabetes. *Nature genetics* **38**, 320–323 (2006).

[9] Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–6 (2007).

[10] Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science* **316**, 1341–5 (2007).

[11] Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–5 (2007).

[12] WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).

[13] Zeggini, E. *et al.* Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–41 (2007).

[14] Helgason, A. *et al.* Refining the impact of tcf7l2 gene variants on type 2 diabetes and adaptive evolution. *Nat Genet* **39**, 218–25 (2007).

[15] Gaulton, K. J. *et al.* A map of open chromatin in human pancreatic islets. *Nat Genet* **42**, 255–9 (2010).

[16] Krishnamurthy, J. *et al.* p16ink4a induces an age-dependent decline in islet regenerative potential. *Nature* **443**, 453–457 (2006).

[17] Pasmant, E. *et al.* Characterization of a germ-line deletion, including the entire ink4/arf locus, in a melanoma-neural system tumor family: identification of anril, an antisense noncoding rna whose expression coclusters with arf. *Cancer Research* **67**, 3963 (2007).

[18] McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488–91 (2007).

[19] Samani, N. *et al.* Genomewide association analysis of coronary artery disease. *New England Journal of Medicine* (2007).

[20] Helgadottir, A. *et al.* A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316**, 1491–3 (2007).

[21] Helgadottir, A. *et al.* The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nature genetics* **40**, 217–224 (2008).

[22] Shea, J. *et al.* Comparing strategies to fine-map the association of common snps at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat Genet* **43**, 801–5 (2011).

[23] Steinthorsdottir, V. *et al.* A variant in cdkal1 influences insulin response and risk of type 2 diabetes. *Nat Genet* **39**, 770–5 (2007).

[24] Ubeda, M., Rukstalis, J. & Habener, J. Inhibition of cyclin-dependent kinase 5 activity protects pancreatic beta cells from glucotoxicity. *Journal of Biological Chemistry* **281**, 28858 (2006).

[25] Wei, F. *et al.* Cdk5-dependent regulation of glucose-stimulated insulin secretion. *Nature medicine* **11**, 1104–1108 (2005).

[26] Quaranta, M. *et al.* Differential contribution of cdkal1 variants to psoriasis, crohn's disease and type ii diabetes. *Genes and Immunity* (2009).

[27] Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nat Genet* **40**, 955–62 (2008).

[28] Frayling, T. M. *et al.* A common variant in the fto gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–94 (2007).

[29] Scuteri, A. *et al.* Genome-wide association scan shows genetic variants in the fto gene are associated with obesity-related traits. *PLoS Genet* **3**, e115 (2007).

[30] Dina, C. *et al.* Variation in fto contributes to childhood obesity and severe adult obesity. *Nat Genet* **39**, 724–6 (2007).

[31] Fischer, J. *et al.* Inactivation of the fto gene protects from obesity. *Nature* **458**, 894–8 (2009).

[32] Boissel, S. *et al.* Loss-of-function mutation in the dioxygenase-encoding fto gene causes severe growth retardation and multiple malformations. *Am J Hum Genet* **85**, 106–11 (2009).

[33] Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS Biol* **8**, e1000294 (2010).

[34] Bort, R., Signore, M., Tremblay, K., Barbera, J. & Zaret, K. Hex homeobox gene controls the transition of the endoderm to a pseudostratified, cell emergent epithelium for liver bud development. *Developmental biology* **290**, 44–56 (2006).

[35] Fakhrai-Rad, H. *et al.* Insulin-degrading enzyme identified as a candidate diabetes susceptibility gene in gk rats. *Human Molecular Genetics* **9**, 2149 (2000).

[36] Farris, W. *et al.* Insulin-degrading enzyme regulates the levels of insulin, amyloid -protein, and the -amyloid precursor protein intracellular domain in vivo. *Proceedings of the National Academy of Sciences* **100**, 4162–4167 (2003).

[37] Samani, N. & Schunkert, H. Chromosome 9p21 and cardiovascular disease: The story unfolds. *Circulation: Cardiovascular Genetics* **1**, 81 (2008).

[38] Visel, A. *et al.* Targeted deletion of the 9p21 non-coding coronary artery disease risk interval in mice. *Nature* **464**, 409–12 (2010).

[39] Harismendy, O. *et al.* 9p21 dna variants associated with coronary artery disease impair interferon- signalling response. *Nature* **470**, 264–8 (2011).

[40] Katze, M. G., He, Y. & Gale, M., Jr. Viruses and interferon: a fight for supremacy. *Nat Rev Immunol* **2**, 675–87 (2002).

[41] Musunuru, K. *et al.* From noncoding variant to phenotype via sort1 at the 1p13 cholesterol locus. *Nature* **466**, 714–9 (2010).

[42] Yanagawa, T., Hidaka, Y., Guimaraes, V., Soliman, M. & DeGroot, L. J. Ctla-4 gene polymorphism associated with graves' disease in a caucasian population. *J Clin Endocrinol Metab* **80**, 41–5 (1995).

[43] Heward, J. M. *et al.* The development of graves' disease and the ctla-4 gene on chromosome 2q33. *J Clin Endocrinol Metab* **84**, 2398–401 (1999).

[44] Braun, J. *et al.* Ctla-4 promoter variants in patients with graves' disease and hashimoto's thyroiditis. *Tissue Antigens* **51**, 563–6 (1998).

[45] Vaidya, B. *et al.* Ctla4 gene and graves' disease: association of graves' disease with the ctla4 exon 1 and intron 1 polymorphisms, but not with the promoter polymorphism. *Clin Endocrinol (Oxf)* **58**, 732–5 (2003).

[46] Ueda, H. *et al.* Association of the t-cell regulatory gene ctla4 with susceptibility to autoimmune disease. *Nature* **423**, 506–11 (2003).

[47] Kavvoura, F. K. *et al.* Cytotoxic t-lymphocyte associated antigen 4 gene polymorphisms and autoimmune thyroid disease: a meta-analysis. *J Clin Endocrinol Metab* **92**, 3162–70 (2007).

[48] 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* **40**, 638–45 (2008).

[49] Omori, S. *et al.* Replication study for the association of new meta-analysis-derived risk loci with susceptibility to type 2 diabetes in 6,244 japanese individuals. *Diabetologia* **52**, 1554–60 (2009).

[50] Grarup, N. *et al.* Association testing of novel type 2 diabetes risk alleles in the jazf1, cdc123/camk1d, tspan8, thada, adamts9, and notch2 loci with insulin release, insulin sensitivity, and obesity in a population-based sample of 4,516 glucose-tolerant middle-aged danes. *Diabetes* **57**, 2534 (2008).

[51] Johansson, A. *et al.* Common variants in the jazf1 gene associated with height identified by linkage and genome-wide association analysis. *Human Molecular Genetics* **18**, 373 (2009).

[52] Soranzo, N. *et al.* Meta-analysis of genome-wide scans for human adult stature identifies novel loci and associations with measures of skeletal frame size. *PLoS Genetics* **5** (2009).

[53] Frayling, T., Colhoun, H. & Florez, J. A genetic link between type 2 diabetes and prostate cancer. *Diabetologia* **51**, 1757–1760 (2008).

[54] Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nature genetics* **40**, 310–315 (2008).

[55] Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet* **41**, 334–341 (2009).

[56] Coronary Artery Disease Consortium. Large scale association analysis of novel genetic loci for coronary artery disease. *Arterioscler Thromb Vasc Biol* **29**, 774–780 (2009).

[57] Askari, A. *et al.* Effect of stromal-cell-derived factor 1 on stem-cell homing and tissue regeneration in ischaemic cardiomyopathy. *The Lancet* **362**, 697–703 (2003).

[58] Zheng, H., Fu, G., Dai, T. & Huang, H. Migration of endothelial progenitor cells mediated by stromal cell-derived factor-1 [alpha]/cxcr4 via pi3k/akt/enos signal transduction pathway. *Journal of Cardiovascular Pharmacology* **50**, 274 (2007).

[59] Bosserhoff, A. & Buettner, R. Expression, function and clinical relevance of mia (melanoma inhibitory activity). *Histology and histopathology* **17**, 289–300 (2002).

[60] Kochi, Y. *et al.* A functional variant in fcrl3, encoding fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. *Nature genetics* **37**, 478–485 (2005).

[61] Simmonds, M. *et al.* Contribution of single nucleotide polymorphisms within fcrl3 and map3k7ip2 to the pathogenesis of graves' disease. *Journal of Clinical Endocrinology and Metabolism* **91**, 1056 (2006).

[62] Vella, A. *et al.* Localization of a type 1 diabetes locus in the il2ra/cd25 region by use of tag single-nucleotide polymorphisms. *The American Journal of Human Genetics* **76**, 773–779 (2005).

[63] Lowe, C. *et al.* Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the il2ra region in type 1 diabetes. *Nature genetics* **39**, 1074–1082 (2007).

[64] Brand, O. *et al.* Association of the interleukin-2 receptor alpha (il-2r [alpha])/cd25 gene region with graves' disease using a multilocus test and tag snps. *Clinical endocrinology* **66**, 508 (2007).

[65] Maier, L. *et al.* Il2ra genetic heterogeneity in multiple sclerosis and type 1 diabetes susceptibility and soluble interleukin-2 receptor production. *PLoS Genetics* **5** (2009).

[66] Dendrou, C. *et al.* Cell-specific protein phenotypes for the autoimmune locus il2ra using a genotype-selectable human bioresource. *Nature genetics* **41**, 1011–1015 (2009).

[67] Giardine, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. *Genome research* **15**, 1451 (2005).

[68] Kent, W. Blat the blast-like alignment tool. *Genome research* **12**, 656–664 (2002).

[69] Pruitt, K., Tatusova, T. & Maglott, D. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* (2006).

[70] Hubbard, T. *et al.* The ensembl genome database project. *Nucleic acids research* **30**, 38–41 (2002).

[71] Pasquinelli, A. *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory rna. *Nature* **408**, 86–89 (2000).

[72] Mourelatos, Z. *et al.* mirnps: a novel class of ribonucleoproteins containing numerous micrornas. *Genes and Development* **16**, 720–728 (2002).

[73] Yeo, G., Van Nostrand, E., Holste, D., Poggio, T. & Burge, C. Identification and analysis of alternative splicing events conserved in human and mouse. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 2850 (2005).

[74] Pedersen, J. *et al.* Identification and classification of conserved rna secondary structures in the human genome. *PLoS Comput Biol* **2**, e33 (2006).

[75] Griffiths-Jones, S. The microrna registry. *Nucleic acids research* **32**, D109 (2004).

[76] Weber, M. New human and mouse microrna genes found by homology search. *FEBS* **272**, 59–73 (2005).

[77] Kosiol, C. *et al.* Patterns of Positive Selection in Six Mammalian Genomes. *PLoS Genetics* **4** (2008).

[78] Yang, Z. & Nielsen, R. Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. *Mol. Biol. Evol* **19**, 908–917 (2002).

[79] Sherry, S. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29**, 308 (2001).

[80] Pickrell, J. *et al.* Signals of recent positive selection in a worldwide sample of human populations. *Genome Research* **19**, 826 (2009).

[81] Li, J. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100 (2008).

[82] Cann, H. *et al.* A Human Genome Diversity Cell Line Panel. *Science* **296**, 261 (2002).

[83] Sabeti, P. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913–918 (2007).

[84] Voight, B., Kudaravalli, S., Wen, X. & Pritchard, J. A map of recent positive selection in the human genome. *PLoS biology* **4**, 446 (2006).

[85] Bernstein, B. *et al.* Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell* **120**, 169–181 (2005).

[86] Bernstein, B. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315–326 (2006).

[87] Mikkelsen, T. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553–560 (2007).

[88] Boyle, A. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311–322 (2008).

[89] Down, T. & Hubbard, T. Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA. *Genome Research* **12**, 458 (2002).

[90] Davuluri, R., Grosse, I. & Zhang, M. Computational identification of promoters and first exons in the human genome. *Nature genetics* **29**, 412–417 (2001).

[91] Ng, P. *et al.* Gene identification signature (GIS) analysis for transcriptome characterization and genome annotation. *Nature methods* **2**, 105–111 (2005).

[92] Chiu, K. *et al.* PET-Tool: a software suite for comprehensive processing and managing of Paired-End diTag (PET) sequence data. *BMC Bioinformatics* **7**, 390 (2006).

[93] Wei, C. *et al.* A global map of p53 transcription-factor binding sites in the human genome. *Cell* **124**, 207–219 (2006).

[94] Choo, A., Padmanabhan, J., Chin, A., Fong, W. & Oh, S. Immortalized feeders for the scale-up of human embryonic stem cells in feeder and feeder-free conditions. *Journal of biotechnology* **122**, 130–141 (2006).

[95] Zeller, K. *et al.* Global mapping of c-Myc binding sites and target gene networks in human B cells. *Proceedings of the National Academy of Sciences* **103**, 17834 (2006).

[96] Johnson, D., Mortazavi, A., Myers, R. & Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **316**, 1497 (2007).

[97] Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nature methods* **5**, 829–834 (2008).

[98] Yang, M., Taylor, J. & Elnitski, L. Comparative analyses of bidirectional promoters in vertebrates. *BMC bioinformatics* **9**, S9 (2008).

[99] Piontkivska, H. *et al.* Cross-species mapping of bidirectional promoters enables prediction of unannotated 5' UTRs and identification of species-specific transcripts. *BMC genomics* **10**, 189 (2009).

[100] Petrykowska, H., Vockley, C. & Elnitski, L. Detection and characterization of silencers and enhancer-blockers in the greater CFTR locus. *Genome research* **18**, 1238 (2008).

[101] Bhinge, A., Kim, J., Euskirchen, G., Snyder, M. & Iyer, V. Mapping the chromosomal targets of STAT1 by sequence tag analysis of genomic enrichment (STAGE). *Genome research* **17**, 910 (2007).

[102] Boyle, A., Guinney, J., Crawford, G. & Furey, T. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537 (2008).

[103] Buck, M., Nobel, A. & Lieb, J. ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biology* **6**, R97 (2005).

[104] Crawford, G. E. *et al.* Genome-wide mapping of dnase hypersensitive sites using massively parallel signature sequencing (mpss). *Genome Res* **16**, 123–31 (2006).

[105] Crawford, G. *et al.* Dnase-chip: a high-resolution method to identify dnase i hypersensitive sites using tiled microarrays. *Nature methods* **3**, 503–509 (2006).

[106] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature* **447**, 799–816 (2007).

[107] Giresi, P., Kim, J., McDaniell, R., Iyer, V. & Lieb, J. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research* **17**, 877 (2007).

[108] Giresi, P. & Lieb, J. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (formaldehyde assisted isolation of regulatory elements). *Methods* **48**, 233–239 (2009).

[109] Li, H., Ruan, J. & Durbin, R. Mapping short dna sequencing reads and calling variants using mapping quality scores. *Genome Res* **18**, 1851–8 (2008).

[110] Montgomery, S. *et al.* ORegAnno: an open access database and curation system for literature-derived promoters, transcription factor binding sites and regulatory variation. *Bioinformatics* **22**, 637 (2006).

[111] Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in bioinformatics* (2008).

[112] Lewis, B. & Shih, I. Prediction of mammalian microrna targets. *Cell* **115**, 787–798 (2003).

[113] Lewis, B., Burge, C. & Bartel, D. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell* **120**, 15–20 (2005).

[114] Grimson, A. *et al.* Microrna targeting specificity in mammals: determinants beyond seed pairing. *Molecular cell* **27**, 91–105 (2007).

[115] Sabo, P. *et al.* Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays. *Nature methods* **3**, 511–518 (2006).

[116] Sabo, P. *et al.* Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 16837 (2004).

[117] Pennacchio, L. *et al.* In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502 (2006).

[118] Euskirchen, G. *et al.* CREB binds to multiple loci on human chromosome 22. *Molecular and cellular biology* **24**, 3804 (2004).

[119] Euskirchen, G. *et al.* Mapping of transcription factor binding regions in mammalian cells by ChIP: Comparison of array-and sequencing-based technologies. *Genome Research* **17**, 898 (2007).

[120] Martone, R. *et al.* Distribution of NF-κB-binding sites across human chromosome 22. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 12247 (2003).

[121] Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nature methods* **4**, 651–657 (2007).

[122] Rozowsky, J. *et al.* PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nature biotechnology* **27**, 66–75 (2009).

[123] King, D. *et al.* Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. *Genome research* **15**, 1051 (2005).

[124] Kolbe, D. *et al.* Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome research* **14**, 700 (2004).

[125] Yeo, G. *et al.* An rna code for the fox2 splicing regulator revealed by mapping rna-protein interactions in stem cells. *Nature Structural and Molecular Biology* **16**, 130–137 (2009).

[126] Kim, T. *et al.* A high-resolution map of active promoters in the human genome. *Nature* **436**, 876–880 (2005).

[127] Dennis, J. *et al.* Independent and complementary methods for large-scale structural analysis of mammalian chromatin. *Genome research* **17**, 928 (2007).

[128] Ozsolak, F., Song, J., Liu, X. & Fisher, D. High-throughput mapping of the chromatin structure of human promoters. *Nature Biotechnology* **25**, 244–248 (2007).

[129] Gupta, S. *et al.* Predicting human nucleosome occupancy from primary sequence. *PLoS Computational Biology* **4** (2008).

[130] Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034–1050 (2005).

[131] Rada-Iglesias, A. *et al.* Binding sites for metabolic disease related transcription factors inferred at base pair resolution by chromatin immunoprecipitation and genomic microarrays. *Human molecular genetics* **14**, 3435 (2005).

[132] Rada-Iglesias, A. *et al.* Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. *Genome Research* **18**, 380 (2008).

[133] Wray, N. R., Purcell, S. M. & Visscher, P. M. Synthetic associations created by rare variants do not explain most gwas results. *PLoS Biol* **9**, e1000579 (2011).

[134] Anderson, C. A., Soranzo, N., Zeggini, E. & Barrett, J. C. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biol* **9**, e1000580 (2011).