nature genetics

Bayesian refinement of association signals for 14 loci in 3 common diseases

The Wellcome Trust Case Control Consortium[1,2]

To further investigate susceptibility loci identified by genome-wide association studies, we genotyped 5,500 SNPs across 14 associated regions in 8,000 samples from a control group and 3 diseases: type 2 diabetes (T2D), coronary artery disease (CAD) and Graves' disease. We defined, using Bayes theorem, credible sets of SNPs that were 95% likely, based on posterior probability, to contain the causal disease-associated SNPs. In 3 of the 14 regions, *TCF7L2* (T2D), *CTLA4* (Graves' disease) and *CDKN2A-CDKN2B* (T2D), much of the posterior probability rested on a single SNP, and, in 4 other regions (*CDKN2A-CDKN2B* (CAD) and *CDKAL1*, *FTO* and *HHEX* (T2D)), the 95% sets were small, thereby excluding most SNPs as potentially causal. Very few SNPs in our credible sets had annotated functions, illustrating the limitations in understanding the mechanisms underlying susceptibility to common diseases. Our results also show the value of more detailed mapping to target sequences for functional studies.

Genome-wide association studies (GWAS) have had unprecedented success in identifying genomic regions and candidates for causal genes where genetic variation, namely SNP markers, is most strongly associated with phenotypic variation and disease susceptibility. To date, over 1,500 such regions have been identified across more than 200 diseases and phenotypes (Catalog of Published Genome-Wide Association Studies; see URLs). By comparing results and candidate genes across multiple GWAS regions, these studies can indicate new candidates for causal pathways, such as autophagy in Crohn's disease[1], triggering a new wave of functional studies and providing genetic validation of therapeutic strategies[1]. Even though GWAS genotyping chips, including the Affymetrix Human 500K chip we used in an initial GWAS of seven common diseases[2], provide very good mapping coverage of common variation of the genome in Europeans, it is possible that with a much denser set of SNP markers we can refine association signals to particular candidate genes or even detect missing disease association signals that were poorly tagged on the GWAS chip.

It is now appreciated that most common variants that are associated with altered disease risk do not have obvious functional consequences, leading to the plausible conclusion that they affect the regulation of gene expression in some way. Nevertheless, most *cis*-acting regulatory variation lies within or very proximal to the structural genes[3] affected by those functional polymorphisms, and, hence, identifying all or most of the SNPs with the strongest disease associations within a region will help to identify candidate genes with greater confidence. This follow-up genotyping, often referred to as fine mapping, can identify new candidate genes and causal variants in GWAS-identified regions, for example, the nonsynonymous SNP rs3184504 in *SH2B3* in type 1 diabetes[4]. In addition, it can lead to the identification of regions and SNPs that are highly unlikely to be causal. In the present study, as a follow-up to the Wellcome Trust Case Control Consortium (WTCCC) GWAS[2], we studied 14 associated regions in 3 diseases, T2D, CAD and Graves' disease, by attempting to genotype all known SNPs in the associated regions. In five of these regions, we had undertaken SNP discovery through resequencing of controls. As a result, we not only have refined some of the association signals to certain genes and SNPs, but we have also developed and applied an informative way of analyzing and interpreting the results of dense SNP mapping using a Bayesian approach.

RESULTS

Experimental design

We genotyped all known SNPs for which we could design assays across 13 GWAS-associated regions using an Illumina iSELECT assay. These were typed on 7,894 samples in total: 1,930 common controls and ~2,000 cases each with T2D, CAD and the common autoimmune thyroid disease, Graves' disease. The three diseases were chosen to reflect a range of putative disease etiologies, and the regions were selected to include both strong and weaker signals in our original GWAS. A breakdown of the numbers of SNPs and their sources is given in **Table 1**. Our fine-mapping experiment was preceded by resequencing of 32 unrelated Utah residents of Northern and Western European ancestry (CEU) HapMap individuals across 5 of the fine-mapping regions (and 11 other regions showing association in WTCCC-studied diseases), to identify putative causal mutations in these regions; this information added to existing SNP variation resources. Further details of these resequencing experiments are given in **Supplementary Figures 1–7**, **Supplementary Tables 1** and **2** and the **Supplementary Note**.

Overall findings

In the subsequent analyses for a particular disease, we leveraged our design by comparing cases for that disease against an expanded

[1]A full list of authors and affiliations appears at the end of the paper. [2]A list of members and affiliations appears in the **Supplementary Note**.

Received 15 June 2011; accepted 11 September 2012; published online 28 October 2012; doi:10.1038/ng.2435

Table 1 Region definitions with SNP counts and sources

					Affymetrix 500K			HapMap			Resequencing			DbSNP		
Phenotype	Chr.	Start (Mb)	End (Mb)	Gene	Total	Genotyped	Passed quality control	Polymorphic	Total	Genotyped	Passed quality control	Polymorphic	Total	Genotyped	Passed quality control	Polymorphic
CAD	1	109.54	109.65	*SORT1*	49	19	17	13	76	75	50	50	141	77	42	22	539	239	173	127
	9	21.92	22.13	*CDKN2A-CDKN2B*	107	40	37	30	176	174	128	128	691	443	184	64	984	464	314	206
	10	44.01	44.15	*CXCL12*	115	42	40	33	161	159	111	111	74	35	23	16	957	416	303	238
	1	220.78	221.04	1q41	68	23	23	22	11	11	6	6	109	56	28	25	970	399	305	266
	2	226.73	226.91	2q36	67	25	24	18	100	96	72	72	378	216	121	41	691	330	223	138
T2D	16	52.35	52.41	*FTO*	42	15	14	13	75	74	53	52	356	199	98	59	313	128	97	88
	10	94.19	94.49	*HHEX*	48	18	16	14	152	145	115	114	974	564	248	162	1,019	457	309	253
	6	20.63	20.84	*CDKAL1*	84	30	28	26	241	229	167	167	573	364	135	74	1,115	503	347	265
	10	114.71	114.82	*TCF7L2*	42	14	14	14	53	53	34	34	75	32	24	19	369	159	118	92
	7	28.01	28.23	*JAZF1*	101	37	34	30	157	152	104	104	63	34	15	14	873	369	290	214
Graves' disease	2	204.38	204.53	*CTLA-4*	91	35	33	23	135	133	104	104	59	33	13	13	754	322	246	186
	10	6.06	6.17	*CD25-IL2RA*	97	38	32	27	111	111	71	71	17	13	2	2	1,222	566	391	265
	1	155.81	156.09	*FCRL3*	125	46	44	35	206	200	131	130	43	28	10	5	1,637	743	552	342
Total					1,036	382	356	298	1,654	1,612	1,146	1,143	3,553	2,094	943	516	11,443	5,095	3,668	2,680

Chr., chromosome. Genomic coordinates are from NCBI Build 36, and references to dbSNP correspond to Build 128. The numbers of SNPs for each region come from various sources: the Affymetrix 500K genotyping chip; HapMap 2; targeted resequencing experiments (either within WTCCC, provided by collaborators or through mining of the Watson and Venter genomes); and dbSNP. Each SNP was attributed to the left-most source in which it was found. For each SNP source, counts are shown for the total number of SNPs, the number for which successful genotyping assays were designed and performed, the number that passed quality control filters and the number that were polymorphic in our samples (MAF > 0.01).

reference group consisting of the common controls and the cases with unrelated diseases: Graves' disease cases were compared against common controls and T2D and CAD cases, and T2D and CAD cases were each compared against common controls and Graves' disease cases. Comparing cases only against the common controls resulted in weaker signals, as expected, but no change in the pattern of signals (data not shown). One region on chromosome 9p21, containing *CDKN2A-CDKN2B*, is associated with both CAD and T2D, such that there were 14 association regions in total analyzed in the fine-mapping experiment.

We developed and applied a new Bayesian method for the statistical analysis of the fine-mapping data, which seems to have several advantages over other approaches. In the Bayesian framework, the evidence for association at a SNP is measured by the Bayes factor. Under certain assumptions, these Bayes factors can be used to calculate the posterior probability for each SNP, that is, the probability when taking into consideration the fine-mapping data that the SNP is driving the association signal (Online Methods). These posterior probabilities can be directly compared between SNPs within and across regions, in a way that is not straightforward, for example, with association *P* values.

For each region, the posterior probability associated with each SNP can be calculated (**Fig. 1** and **Supplementary Fig. 8**). Uncertainty after fine mapping as to the identity of the causal SNPs is conveniently handled by describing credible sets. In each region, one can choose the smallest set of SNPs accounting for 95% or 99% of the posterior probability. These credible sets are somewhat analogous to confidence intervals: in a particular region, if the true causal SNP has been genotyped, we can be reasonably confident that it will be in the relevant credible set. If the credible set for a particular region contains a small number of SNPs, the fine-mapping experiment has been informative in narrowing down the set of SNPs that might be causal. In the other direction, SNPs not contained in the credible set can be excluded from being responsible for the primary association signal, such that, if the credible set is large, the fine-mapping experiment has added little or no resolution in the search for the causal variant(s).

Our results identify several features. First, there are only three regions in which much of the posterior probability after the fine mapping rests on a single SNP: *TCF7L2* in T2D, *CTLA4* in Graves' disease and *CDKN2A-CDKN2B* in T2D (**Fig. 1**). The finding at the third region is misleading, as there is a substantially stronger signal generated by pairs of SNPs at *CDKN2A-CDKN2B* in T2D (**Supplementary Note**). Second, in 7 of the 14 regions (*CDKN2A-CDKN2B* in CAD, *CTLA4* in Graves' disease and *CDKN2A-CDKN2B*, *CDKAL1*, *FTO*, *HHEX* and *TCF7L2* in T2D), including the 3 just mentioned, the credible sets are relatively small, including fewer than 34 SNPs, and the fine-mapping experiment has provided useful information, at least in excluding large numbers of SNPs from being causal (**Supplementary Table 3**). In the other 7 regions, 1q41, 2q36, *IL2RA*, *CXCL12*, *FCRL3*, *JAZF1* and *SORT1*, the project has not excluded many of the SNPs as potentially being causal (or being good markers)—for each region, the 95% credible set contains at least 75 and typically over 100 SNPs. Nonetheless, two of these seven regions (*SORT1* and *IL2RA*) have the property that, even though large numbers of SNPs cannot be excluded, substantial posterior weight rests on one or a few SNPs, and, therefore, typing additional samples in these regions should lead to a much smaller credible set.

The seven regions for which our additional genotyping did not eliminate many of the SNPs are distinguished by the fact that they were the regions where the initial association signal in our data was weakest.

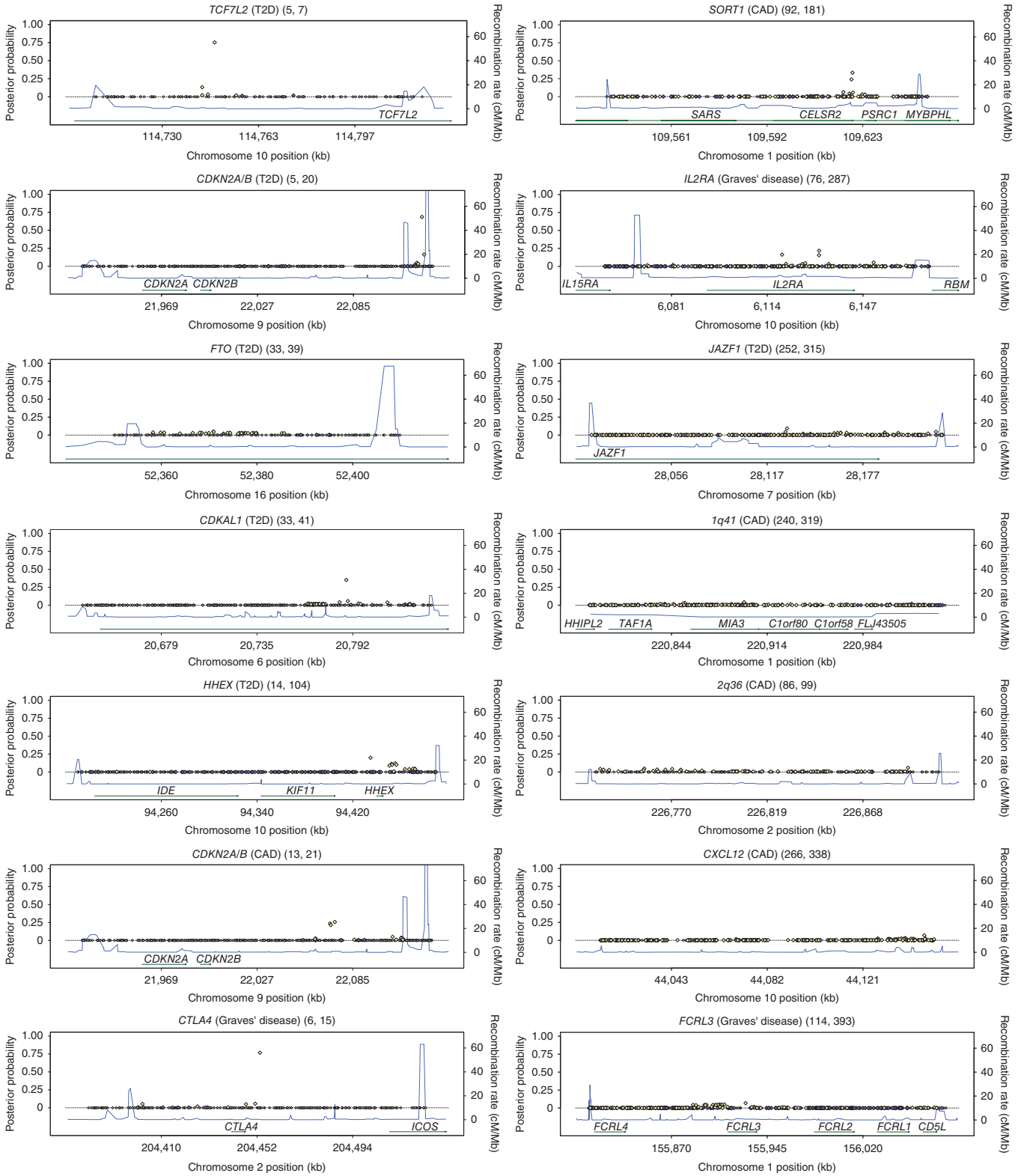

Figure 1 Association plots showing the signal strength in each region as the posterior probability of each SNP passing quality control. The estimated recombination rate is shown in blue (right *y* axis). Genomic position is shown on the *x* axis with Human Genome Build 36 coordinates. SNPs are colored according to membership in credible sets: yellow, 95% credible set; purple, 99% credible set; gray outline, neither set. Genes in the region are shown at the bottom in green. For each plot, the region name, phenotype and sizes of the 95% and 99% credible sets are indicated. All SNPs in the 95% credible set are a subset of the 99% credible set. SNPs with large posterior probabilities represent those most likely to be causal among the SNPs typed.

Table 2 Comparison of association evidence between top SNPs after fine mapping and top SNPs on GWAS chips

Disease	Region	SNP set	SNP ID	Risk allele frequency (controls)	Risk allele	Effect size (relative risk)	log Bayes factor	Bayes factor ratio	λ_s
CAD	*SORT1*	Genotyped in fine mapping	rs3832016	0.79	A	1.22 (1.11–1.35)	2.68	1.00	1.006
		Affymetrix 500K	rs599839	0.78	A	1.17 (1.07–1.29)	1.59	0.08	1.004
		Illumina 550K	rs611917	0.69	A	1.17 (1.07–1.27)	1.96	0.19	1.005
		Illumina 1.2M	rs660240	0.79	G	1.22 (1.10–1.34)	2.54	0.72	1.006
	CDKN2A-CDKN2B	Genotyped in fine mapping	rs1537370	0.47	A	1.40 (1.29–1.51)	14.03	1.00	1.028
		Affymetrix 500K	rs6475606	0.47	A	1.40 (1.29–1.51)	13.95	0.83	1.028
		Illumina 550K	rs10116277	0.47	A	1.40 (1.29–1.51)	13.99	0.92	1.028
		Illumina 1.2M	rs1537370	0.47	A	1.40 (1.29–1.51)	14.03	1.00	1.028
	CXCL12	Genotyped in fine mapping	rs34161818	0.84	A	1.21 (1.08–1.35)	1.56	1.00	1.004
		Affymetrix 500K	rs977754	0.85	A	1.18 (1.06–1.33)	1.16	0.39	1.003
		Illumina 550K	rs977754	0.85	A	1.18 (1.06–1.33)	1.16	0.39	1.003
		Illumina 1.2M	rs977754	0.85	A	1.18 (1.06–1.33)	1.16	0.39	1.003
	1q41	Genotyped in fine mapping	rs2936023	0.85	T	1.21 (1.08–1.36)	1.58	1.00	1.004
		Affymetrix 500K	rs17464857	0.85	A	1.17 (1.05–1.31)	0.93	0.22	1.003
		Illumina 550K	rs17464857	0.85	A	1.17 (1.05–1.31)	0.93	0.22	1.003
		Illumina 1.2M	rs11485123	0.85	G	1.17 (1.05–1.31)	0.93	0.22	1.003
	2q36	Genotyped in fine mapping	rs2673145	0.41	A	1.20 (1.11–1.30)	3.76	1.00	1.008
		Affymetrix 500K	rs2943646	0.64	G	1.19 (1.10–1.29)	3.01	0.18	1.007
		Illumina 550K	rs2972153	0.67	G	1.21 (1.11–1.32)	3.45	0.49	1.008
		Illumina 1.2M	rs2972153	0.67	G	1.21 (1.11–1.32)	3.45	0.49	1.008
T2D	*FTO*[a]	Genotyped in fine mapping	rs17817449	0.40	C	1.26 (1.17–1.36)	6.69	1.00	1.013
		Combined	rs17817449,rs8063946						1.018
		Affymetrix 500K	rs8050136	0.40	A	1.26 (1.17–1.36)	6.49	0.62	1.013
		Illumina 550K	rs8050136	0.40	A	1.26 (1.17–1.36)	6.49	0.62	1.013
		Illumina 1.2M	rs17817449	0.40	C	1.26 (1.17–1.36)	6.69	1.00	1.013
	CDKN2A-CDKN2B[a]	Genotyped in fine mapping	rs12555274	0.23	C	1.26 (1.15–1.37)	4.75	1.00	1.011
		Combined	rs10811661,rs10217762						1.012
		Affymetrix 500K	rs10811661	0.83	A	1.27 (1.14–1.41)	3.46	0.05	1.007
		Illumina 550K	rs2383208	0.82	A	1.26 (1.13–1.40)	3.20	0.03	1.007
		Illumina 1.2M	rs2383208	0.82	A	1.26 (1.13–1.40)	3.20	0.03	1.007
	HHEX	Genotyped in fine mapping	rs10882098	0.59	G	1.21 (1.12–1.31)	4.01	1.00	1.009
		Affymetrix 500K	rs5015480	0.59	G	1.20 (1.11–1.30)	3.80	0.61	1.008
		Illumina 550K	rs5015480	0.59	G	1.20 (1.11–1.30)	3.80	0.61	1.008
		Illumina 1.2M	rs5015480	0.59	G	1.20 (1.11–1.30)	3.80	0.61	1.008
	CDKAL1[a]	Genotyped in fine mapping	rs7756992	0.27	G	1.29 (1.19–1.40)	6.71	1.00	1.014
		Combined	rs7756992,rs6456360						1.021
		Affymetrix 500K	rs9460546	0.31	C	1.25 (1.15–1.35)	5.38	0.05	1.012
		Illumina 550K	rs7756992	0.27	G	1.29 (1.19–1.40)	6.71	1.00	1.014
		Illumina 1.2M	rs7756992	0.27	G	1.29 (1.19–1.40)	6.71	1.00	1.014
	TCF7L2	Genotyped in fine mapping	rs7903146	0.30	A	1.40 (1.29–1.52)	13.61	1.00	1.027
		Affymetrix 500K	rs4506565	0.32	T	1.37 (1.27–1.49)	12.24	0.04	1.024
		Illumina 550K	rs7903146	0.30	A	1.40 (1.29–1.52)	13.61	1.00	1.027
		Illumina 1.2M	rs7903146	0.30	A	1.40 (1.29–1.52)	13.61	1.00	1.027
	JAZF1	Genotyped in fine mapping	rs12531540	0.51	G	1.14 (1.06–1.23)	1.75	1.00	1.004
		Affymetrix 500K	rs1859687	0.06	C	1.25 (1.08–1.45)	1.12	0.23	1.003
		Illumina 550K	rs498475	0.37	G	1.14 (1.05–1.23)	1.44	0.49	1.004
		Illumina 1.2M	rs498475	0.37	G	1.14 (1.05–1.23)	1.44	0.49	1.004

(continued)

Table 2 Comparison of association evidence between top SNPs after fine mapping and top SNPs on GWAS chips (continued)

Disease	Region	SNP set	SNP ID	Risk allele frequency (controls)	Risk allele	Effect size (relative risk)	log Bayes factor	Bayes factor ratio	λ_s
Graves' disease	*CTLA4*	Genotyped in fine mapping	rs11571297	0.51	A	1.39 (1.29–1.50)	16.08	1.00	1.027
		Affymetrix 500K	rs3087243	0.55	G	1.38 (1.28–1.49)	14.89	0.06	1.025
		Illumina 550K	rs231804	0.57	A	1.37 (1.27–1.48)	13.58	0.00	1.023
		Illumina 1.2M	rs11571291	0.58	A	1.38 (1.28–1.48)	13.87	0.01	1.024
	CD25-IL2RA	Genotyped in fine mapping	rs10905669	0.23	A	1.20 (1.10–1.30)	3.10	1.00	1.007
		Affymetrix 500K	rs10905669	0.23	A	1.20 (1.10–1.30)	3.10	1.00	1.007
		Illumina 550K	rs7090530	0.60	A	1.16 (1.08–1.25)	2.49	0.25	1.005
		Illumina 1.2M	rs10905669	0.23	A	1.20 (1.10–1.30)	3.10	1.00	1.007
	FCRL3	Genotyped in fine mapping	rs11264798	0.52	C	1.17 (1.09–1.26)	3.00	1.00	1.006
		Affymetrix 500K	rs2785663	0.58	C	1.17 (1.08–1.26)	2.80	0.63	1.006
		Illumina 550K	rs2785665	0.58	A	1.17 (1.08–1.26)	2.76	0.58	1.006
		Illumina 1.2M	rs11264798	0.52	C	1.17 (1.09–1.26)	3.00	1.00	1.006

For each region, shown is the top SNP (based on Bayes factor) after the fine-mapping experiment among various sets of SNPs, including SNPs passing quality control in the fine-mapping experiment and SNPs from the Affymetrix Human 500K SNP array, the Illumina 550K SNP array and the Illumina 1.2M SNP array. Shown are risk allele frequency, risk allele, relative risk (and 95% confidence interval (CI)), $\log_{10}$ of the Bayes factor for the SNP, the ratio of the Bayes factor for the SNP to the Bayes factor for the top fine-mapping SNP and the contribution of that SNP to the sibling relative risk (λ_s).

[a]For the three regions with evidence for a second associated SNP, the combined contribution of the pair of SNPs to the sibling relative risk is shown.

Our results indicate that, in these regions, these smaller signals are due to the causal variant(s) having only very modest effect sizes, explaining the difficulty in fine mapping and, therefore, requiring much larger sample sizes to begin to resolve the linkage disequilibrium in the regions and refine the associations. There was the possibility of a considerably stronger signal at one of the new SNPs genotyped if one of the new SNPs was a causal variant of large effect that was tagged only poorly in the original GWAS or a good tag marker for such a variant. We note that none of the 14 regions showed evidence of such a variant.

Imputation

Although we aimed to genotype all known SNPs in the regions at the time that the experiment was designed, one limitation of our study was that the catalog of variation was not complete. In addition, we were only able to design assays for 78% of SNPs, and only obtained genotypes that successfully passed quality control for 94% of the SNPs for which assays were designed. To recover information for SNPs not genotyped in our experiment, we also imputed genotypes at all SNPs present in 1000 Genomes Project data (June 2011 release) for which we did not have genotype data (**Supplementary Note**). The resulting association plots are shown in **Supplementary Figure 9**. The signal at the top genotyped SNP is compared with that at the top imputed SNP in **Supplementary Table 4**. There was only one region, 2q36 in CAD, where the top SNP after imputation was not one of the genotyped SNPs, and, even in this region, imputation did not greatly affect the overall results: the top imputed SNP had a posterior weight of 0.051, only slightly greater than the weight (0.049 in the imputed data set) for the top genotyped SNP, and the size of the 95% (99%) credible region increased from 86 (99) SNPs to 107 (122). No imputed SNP had higher posterior probability than the top genotyped SNP in any other region, and the imputed SNPs did not change the overall results in these regions, either by identifying any 1000 Genomes Project SNPs with large signals in the seven regions where our additional genotyping did not exclude many of the SNPs or by greatly changing the composition of the credible sets in the regions where our efforts were informative.

Mapping information gained

We can also ask whether the fine-mapping effort added to our understanding from the original GWAS study. The evidence for association of the top SNP after the fine-mapping experiment was compared with that from the top SNP in the region on each of three commercial genotyping chips (Affymetrix Human 500K SNP array and Illumina 550K and 1.2M SNP arrays) (**Table 2**). The Bayes factor ratio allows a direct comparison of the relative evidence for association of the top SNP from a particular chip to the top SNP from our new results: for example, a Bayes factor ratio of 0.22 means that the posterior probability for the top SNP from the chip is smaller by a factor of 0.22 than that for the new top SNP. The new top SNP improved on the top SNP from the Affymetrix 500K and Illumina 550K and 1.2M arrays, respectively, in 13, 12 and 8 of the 14 regions, with comparisons among SNP chips differing across regions. Our results, not unexpectedly, indicate that the top GWAS SNPs are typically not the best markers for the causal variant. Another comparison of GWAS results and our findings is given in **Supplementary Table 5**. For the Affymetrix 500K array, we show how many of the SNPs in the credible set after the fine-mapping experiment were typed in the original GWAS. In most regions, the vast majority of the SNPs in credible sets after the fine-mapping experiment were not typed in the original association study.

The new top SNPs do not greatly change the proportion of heritability explained by the locus (**Table 2**). There can be strong evidence in favor of one SNP over another, even when there are only slight differences in effect sizes (most notably, in *CTLA4* in Graves' disease, but similar results were also found in several other SNP comparisons) and, hence, only slight differences in heritability explained by the locus.

Bioinformatic annotation

In most of the regions where the fine mapping considerably refined the association signal, there was still a set of correlated SNPs that could not be separated on the basis of our current genetic data in terms of the possibility that a SNP might be functional and causal. We therefore cross-referenced our most credible SNPs against all relevant annotation tracks found in the UCSC Genome Browser.

Table 3 Biological annotations for the 109 and 247 SNPs making up the 95% and 99% credible sets, respectively, across 7 fine-mapped regions

		Proportion of SNPs in credible set		Proportion of posterior probability in credible set	
Annotation type[a]	Annotation	95%	99%	95%	99%
dbSNP130 functions	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
	Unknown (intergenic)	0.31	0.44	0.56	0.56
Other gene prediction methods	Alternative splicing events	0	0	0	0
	Non-coding RNA	0	0	0	0
	miRNA (miRBase)	0	0	0	0
ChIP-seq transcription factor–binding site summary	ENCODE ChIP-seq transcription factor–binding sites	0.23	0.32	0.21	0.20
DNase I hypersensitivity	Duke/UW DNase I hypersensitivity ($P < 0.05$)	0.10	0.11	0.07	0.08
	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 predictions	tfbsCons	0.01	0.01	0	0
	miRNA target site	0	0	0	0
	7× reg score >0.1	0.14	0.1	0.07	0.07
Sequence conservation	17-way most conserved vertebrate	0.05	0.04	0.02	0.02
	28-way most conserved mammal	0.05	0.03	0.02	0.02

UW, University of Washington.
[a]See the **Supplementary Note** for details on specific annotation classes.

We first considered annotations based on primary DNA sequence, before turning to more recent tissue-specific annotations relating principally to properties of chromatin.

The results for annotations based primarily on DNA sequence for the seven regions where the fine mapping excluded most SNPs in the region are shown in **Table 3**. The lack of putative functional annotations is noteworthy. Of the 109 (247) SNPs in the 95% (99%) credible sets across the 7 regions, there were no SNPs in coding exons and no (5) SNPs in 5′ and 3′ RefGene UTRs. Two SNPs were found in the *CDKN2B-AS1* gene (also known as *ANRIL*) for a non-coding antisense RNA, but they each had very low posterior probability. Many SNPs were intronic, but none of these were in canonical splice sites. Cross-referencing our SNPs against other gene annotation tracks, including non-coding and microRNA (miRNA) tracks, did not indicate that any other SNPs were genic.

Next, we examined recently derived annotations associated with histone modification and chromatin accessibility[2]. This analysis was complicated by several factors. The first is that these domains tend to be modified in a tissue-specific manner. Relevant tissues for particular diseases are not always clear and are often not unique, and data for these tissues may in any event not be available. In addition, the precision of these annotations is sometimes either limited or unclear. Overall summaries of the annotations are provided in the **Supplementary Table 6** and the **Supplementary Note**, but, at what seems to be an early stage in understanding the mechanisms involved in chromatin accessibility, this analysis did not provide compelling information to distinguish SNPs highlighted in the fine-mapping experiment.

Secondary signals

In addition to single-SNP analyses, in each region, we performed conditional and other analyses to look for secondary signals. There are several regions with clear evidence for secondary signals (*CDKN2A-CDKN2B*, *FTO* and *CDKAL1*; **Supplementary Note**). We also assessed the best model for relating SNP genotype to disease risk. For most reported GWAS associations, there is no significant evidence for departure from the simple model in which each additional copy of the risk allele increases disease risk by the same multiplicative factor. In this simple model, the log odds of disease increase in an additive manner with each additional copy of the risk allele, such that the model is sometimes referred to as the additive model. But, as has been noted elsewhere[5], the power to detect departures from this model is limited unless the true causal variant or a SNP highly correlated with it is typed in the study. With the much more extensive genotyping in this study, it is natural to revisit this question. We found no compelling evidence for departures from the additive model.

Table 4 Proportion of 1000 Genomes Project SNPs captured in this experiment (by direct genotyping or imputation) stratified by MAF

MAF	Genotyped in fine mapping	Genotyped or well imputed[a]	1000 Genomes Project	Proportion captured
MAF < 0.01	125	1,665	20,306 (12,263)[b]	0.082 (0.14)[c]
0.01 ≤ MAF < 0.02	92	838	1,920	0.44
0.02 ≤ MAF < 0.04	240	1,305	1,601	0.82
0.04 ≤ MAF < 0.06	256	675	794	0.85
0.06 ≤ MAF < 0.10	260	808	901	0.90
0.10 ≤ MAF < 0.20	614	972	1,083	0.90
0.20 ≤ MAF	1,784	2,824	3,196	0.88

Imputation and MAFs are based on a reference panel of 143 European haplotypes in the 1000 Genomes Project June 2011 data release. See the **Supplementary Note** for imputation details.
[a]SNPs were deemed to be well imputed if average maximum posterior (as returned by IMPUTE1) was greater than 0.98, the IMPUTE1 info score was above 0.8 and there was less than 2% missing data. [b]Count of SNPs with at least two observations of the minor allele in the reference panel. [c]Proportion of SNPs with at least two observations of the minor allele in the reference panel.

Detailed results for the 14 regions studied are described in the **Supplementary Note** (see also **Supplementary Figs. 8–21** and **Supplementary Tables 3–13**).

DISCUSSION

We have found that the use of Bayesian statistical approaches for assigning posterior probabilities and credible sets of SNPs is informative in refining the association signals from GWAS-detected loci with denser genotyping. The calculated value of the posterior probability depends on a number of assumptions, including the assumption that there is a single causal SNP that is typed in the study. This will not always be the case, and, indeed, our data show this is not true for several of the regions we studied, most notably *CDKN2A-CDKN2B* in T2D (**Supplementary Note**). Nonetheless, the ratio of these posterior probabilities for a pair of SNPs reduces to the ratio of the Bayes factors for the pair, and, in a Bayesian framework, this ratio is the quantity that summarizes the weight of evidence in the data for one SNP compared to the other (as shown, for example, in **Table 2**). The calculated posterior probabilities can be thought of as the ratio of the evidence for the particular SNP compared to that for all SNPs typed, such that they are helpful summaries of the overall evidence for each individual SNP based on the fine-mapping data, whether or not there is a single causal SNP or a very good marker typed in the study.

In addition to the detailed analyses of particular regions (**Supplementary Note**), there are several general conclusions from our fine-mapping experiments. First, it seems that, in general, weaker GWAS signals are not driven by poor tagging of a variant with a larger effect size but simply by genuinely smaller effects. Our fine-mapping experiment was powered to find larger effect variants that were genotyped or tagged well, but such variants were not identified in any region with a relatively small effect in the original GWAS. Consequently and because the required sample size depends on the true effect size, in these regions, our additional genotyping did not add greatly to the resolution of the signal for experiments of our size (2,000 cases and, in effect, 4,000 or 6,000 controls). For smaller effect sizes, much larger sample sets are required, and, even then, if linkage disequilibrium is very strong and extended, very large sample sizes can still be uninformative.

Evidence for the involvement of multiple SNPs in a single region was obtained in 3 of the 14 regions, even though our sample size for each disease was small. It thus seems likely that allelic heterogeneity is not uncommon at GWAS loci, and this knowledge will not only contribute to the evidence for a particular candidate gene but will also provide an allelic series to aid functional studies and help explain the familial clustering of the disease. Taken together, the pairs of SNPs at these three loci substantially increase the strength of the genetic effect and also the heritability explained by the locus (**Table 2**).

One limitation with our study is that we were restricted to genotyping only those SNPs known at the time of the study design and our specific sequencing efforts, and we therefore could simply have missed the causal variants or the most associated SNP markers, particularly at lower minor allele frequencies (MAFs). However, we used data from the 1000 Genomes Project to assess our coverage of variants in the regions studied (**Table 4**). For example, in the June 2011 1000 Genomes Project data release, there were 1,920 and 1,601 SNPs in samples from individuals of European descent in our fine-mapping regions with MAFs from 1–2% and 2–4%, respectively. Of these, we directly typed 92 (4.8%) and 240 (15%), respectively, and, in addition, obtained good-quality imputation data on a further 746 and 1,065 SNPs. Our project typed a much denser set of SNPs than GWAS chips, leading to improved imputation. Thus, in these MAF ranges, we obtained actual or well-imputed data on 44% and 82% of these variants. For all higher MAF ranges, the proportion captured or well imputed was at least 85% (**Table 4**). Coverage via genotyping or imputation of variants with MAF of <1% was much lower, but we do not believe this undermines our main conclusions (**Supplementary Note**).

Ranking of SNPs by posterior probability is in general different from ranking based on *P* values, complicating direct comparisons of our findings with those from other studies. In comparing our top SNPs with those in a recent large follow-up study of T2D[6], we found that the top SNPs were identical in three regions (*CDKN2A-CDKN2B*, *CDKAL1* and *TCF7L2*) and almost perfect proxies in the other two regions for which our experiment was informative (*FTO*, $r^2 = 0.995$; *HHEX*, $r^2 = 0.967$) (**Supplementary Table 9**; this table also compares our findings with those described in a recent review of the GWAS results for CAD[7], for which comparison is complicated because our top SNPs were not necessarily typed in the studies reviewed).

This study provides a statistical platform for intuitive and informative analysis and interpretation of GWAS and denser mapping data. We have succeeded via resequencing, dense genotyping and statistical analysis in substantially refining the association signals at several risk-associated loci. This fundamental knowledge will aid future functional studies of specific fine-mapped sequences and candidate genes, help explain heritability and lead to the identification of functional, causal pathways in disease that could be safely modulated for the prevention or intervention of common and rare diseases.

URLs. Catalog of Published Genome-Wide Association Studies, http://www.genome.gov/GWAStudies/; 1000 Genomes Project reference panel haplotypes, http://www.1000genomes.org/.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to D. Altshuler and colleagues at the Broad Institute for provision of SNP discovery information in some of our fine-mapping regions. We acknowledge the many physicians, research fellows and research nurses who contributed to the various case collections and the collection teams and senior management of the UK Blood Services responsible for the UK Blood Services Collection. The principal funder of the project was the Wellcome Trust. For the 1958 Birth Cohort, venous blood collection was funded by the UK Medical Research Council, and cell line production and DNA extraction and processing were funded by the Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory. Many of the authors of this paper received funding from the NIHR Biomedical Research Centers.

AUTHOR CONTRIBUTIONS

The study was conceptually designed by M.A.B., P.R.B., M.J.C., A.C., M.F., A.S.H., A.T.H., A.V.S.H., C.G.M., M. Pembrey, J.S., M.R.S., J.W., N.C., M.H., W.O., M. Parkes, N.R., A.D., J.A.T., D.P.K., N.J.S., S.C.L.G., M.I.M., P. Deloukas and P. Donnelly. The study was implemented by P. Deloukas, J.B.M., S.M., A.M., G.M., D.P.K., M.I.M., M.A.B., N.R., J.A.T., N.J.S. and P. Donnelly. Statistical analyses were performed by J.B.M., J.B., D.V., Z.S., K.P., J.M.M.H., A.A., M. Pirinen, G.M. and P. Donnelly. The paper was written by P. Donnelly, G.M., M.I.M., N.J.S., S.C.L.G., J.A.T., J.M.M.H. and J.B.M.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/ng.2435.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Steinman, L. Mixed results with modulation of T_H-17 cells in human autoimmune diseases. *Nat. Immunol.* **11**, 41–44 (2010).
2. 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).
3. Stranger, B.E. *et al.* Patterns of *cis* regulatory variation in diverse human populations. *PLoS Genet.* **8**, e1002639 (2012).
4. Dendrou, C.A. *et al.* Cell-specific protein phenotypes for the autoimmune locus *IL2RA* using a genotype-selectable human bioresource. *Nat. Genet.* **41**, 1011–1015 (2009).
5. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
6. Morris, A.P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44**, 981–990 (2012).
7. Peden, J.F. & Farrall, M. Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour. *Hum. Mol. Genet.* **20**, R198–R205 (2011).

The authors of this paper are:

Julian B Maller[1,2], Gilean McVean[1,2], Jake Byrnes[1], Damjan Vukcevic[1], Kimmo Palin[3], Zhan Su[1], Joanna M M Howson[4,5], Adam Auton[1], Simon Myers[1,2], Andrew Morris[1], Matti Pirinen[1], Matthew A Brown[6,7], Paul R Burton[8,9], Mark J Caulfield[10], Alastair Compston[11], Martin Farrall[12], Alistair S Hall[13], Andrew T Hattersley[14,15], Adrian V S Hill[1], Christopher G Mathew[16], Marcus Pembrey[17], Jack Satsangi[18], Michael R Stratton[3,19], Jane Worthington[20], Nick Craddock[21], Matthew Hurles[3], Willem Ouwehand[3,22,23], Miles Parkes[24], Nazneen Rahman[19], Audrey Duncanson[25], John A Todd[5], Dominic P Kwiatkowski[1,3], Nilesh J Samani[26,27], Stephen C L Gough[28,29], Mark I McCarthy[1,28,29], Panagiotis Deloukas[3] & Peter Donnelly[1,2]

[1]The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Department of Statistics, University of Oxford, Oxford, UK. [3]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK. [4]Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [5]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Cambridge, UK. [6]Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, Nuffield Orthopaedic Centre, University of Oxford, Oxford, UK. [7]Diamantina Institute of Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Brisbane, Queensland, Australia. [8]Department of Genetics, University of Leicester, Leicester, UK. [9]Department of Health Sciences, University of Leicester, Leicester, UK. [10]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, London, UK. [11]Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. [12]Cardiovascular Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics, Oxford, UK. [13]Multidisciplinary Cardiovascular Research Centre (MCRC), Leeds Institute of Genetics, Health and Therapeutics (LIGHT), University of Leeds, Leeds, UK. [14]Genetics of Complex Traits, Peninsula College of Medicine and Dentistry, University of Exeter, Exeter, UK. [15]Genetics of Diabetes, Peninsula College of Medicine and Dentistry, University of Exeter, Exeter, UK. [16]Department of Medical and Molecular Genetics, King's College London School of Medicine, Guy's Hospital, London, UK. [17]Clinical and Molecular Genetics Unit, Institute of Child Health, University College London, London, UK. [18]Gastrointestinal Unit, School of Molecular and Clinical Medicine, University of Edinburgh, Western General Hospital, Edinburgh, UK. [19]Section of Cancer Genetics, Institute of Cancer Research, Sutton, UK. [20]Arthritis Research UK Epidemiology Unit, University of Manchester, Manchester, UK. [21]Medical Research Council (MRC) Centre for Neuropsychiatric Genetics and Genomics, School of Medicine, Cardiff University, Cardiff, UK. [22]Department of Haematology, University of Cambridge, Cambridge, UK. [23]National Health Service Blood and Transplant, Cambridge Centre, Cambridge, UK. [24]Inflammatory Bowel Disease Genetics Research Group, Addenbrooke's Hospital, Cambridge, UK. [25]The Wellcome Trust, Gibbs Building, London, UK. [26]Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Leicester, UK. [27]Leicester National Institute for Health Research (NIHR) Biomedical Research Unit in Cardiovascular Disease, Glenfield Hospital, Leicester, UK. [28]Oxford Centre for Diabetes, Endocrinology and Medicine, University of Oxford, Churchill Hospital, Oxford, UK. [29]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford, UK. Correspondence should be addressed to P. Donnelly (donnelly@well.ox.ac.uk).

ONLINE METHODS

Bayesian approach. We adopted a Bayesian statistical perspective for analysis, with the strength of evidence for association measured via the Bayes factor for each SNP: the probability of the genotype configuration at that SNP in cases and controls under the alternative hypothesis that the SNP is associated with disease status divided by the probability of the genotype configuration at that SNP in cases and controls under the null hypothesis that disease status is independent of genotype at that SNP (**Supplementary Note**). Large values of the Bayes factor correspond to strong evidence for association, in the way that small *P* values would correspond to strong evidence in a frequentist perspective.

One advantage of the Bayesian perspective is that Bayes factors for different SNPs can be compared quantitatively, which does not seem straightforward with *P* values. For example, for a particular region, under specific assumptions about how many causal SNPs are in the set of genotyped SNPs, it is straightforward to calculate the posterior probability that any particular SNP is causal taking into consideration data from the fine-mapping experiment. For definiteness, we calculate these posterior probabilities under the simple assumption that exactly one of the genotyped SNPs in each region is causal and that it is equally likely, a priori, to be any of the genotyped SNPs in the region. We cannot know that these previous assumptions are true for any particular region, and they may well not be, but the resulting probabilities can be thought of as the relative strength of evidence in favor of each of the SNPs studied. Defining BF_k as the Bayes factor for SNP *k*, we show in the **Supplementary Note** that the posterior probably for SNP *k* is equal to

$$\frac{BF_k}{\sum_j BF_j}$$

where *j* indexes SNPs in the region. Regardless of whether the causal SNP or SNPs have been typed in the fine-mapping experiment, SNPs with a low value for this posterior probability are unlikely to be causal.

SNP selection, genotyping and quality control. The case samples used in this study were as previously described in earlier WTCCC studies[2,8]. Control samples were a subset of those used in the original WTCCC study[2].

The boundaries of the genomic regions for study were determined as follows. We took the top associated SNP in the region from published results at the time of study design (referred to here as the 'focal SNP') and extended the region upstream and downstream by a genetic distance of 0.1 cM, using HapMap fine-scale estimates of recombination rates. Next, we looked in HapMap for any SNPs that lay outside these recombination-determined boundaries with $r^2 > 0.2$ to the focal SNPs in the CEU population and, if necessary, extended the boundaries of the regions to include such SNPs. Finally, for CAD and T2D, we looked in the WTCCC GWAS data for any SNPs with association *P* value within two orders of magnitude of that of the focal SNP and, if necessary, extended the boundaries to include any such SNPs. In most cases, the boundaries defined by genetic distance did not need to be extended to meet the other criteria.

We attempted to design assays for and genotype all SNPs from the Affymetrix Human 500K SNP array used in our original GWAS study, all polymorphic HapMap SNPs and all SNPs from our resequencing pilot and the other (smaller) resequencing data sets we had access to for the regions we studied, including mining of the Watson and Venter genomes. We did the same for any SNP in dbSNP (version 128) that had genotype or frequency data showing variation and any SNPs that had been reported by more than one group (**Table 1**).

Genotype calling was performed in two stages. First, genotypes were called using Illuminus[9]. Only SNPs that Illuminus called with high confidence were taken forward directly; the remainder underwent manual cluster inspection and were called again where appropriate. The first quality control filter applied was to remove individual genotypes with an Illuminus confidence score of less than 0.2, which is the recommended threshold. Samples with call rates lower than 90% were excluded. We excluded SNPs with call rates less than 0.95, Hardy-Weinberg equilibrium *P* values of less than 0.001 or MAFs less than 0.001.

Imputation. For imputation, we used as a reference panel the 286 haplotypes for individuals of European descent (defined as samples from the CEU, British in England and Scotland (GBR), Iberian population in Spain (IBS) and Toscani in Italia (TSI) populations) from the June 2011 release of the 1000 Genomes Project (see URLs). We used the software package IMPUTE v1.1.5 to perform the imputation. Quality control after imputation consisted of excluding imputed SNPs with either (i) average maximum posterior (as returned by IMPUTE1) less than 0.98, (ii) IMPUTE1 info score less than 0.8 or (iii) greater than 2% missing data. Imputed data were analyzed for association in the program SNPTEST.

8. Burton, P.R. *et al.* Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.* **39**, 1329–1337 (2007).
9. Teo, Y.Y. *et al.* A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* **23**, 2741–2746 (2007).

doi:10.1038/ng.2435