Loss-of-function variants in *ATM* confer risk of gastric cancer

Hannes Helgason[1,2,10], Thorunn Rafnar[1,10], Halla S Olafsdottir[3], Jon G Jonasson[4–6], Asgeir Sigurdsson[1], Simon N Stacey[1], Adalbjorg Jonasdottir[1], Laufey Tryggvadottir[5,6], Kristin Alexiusdottir[3], Asgeir Haraldsson[6,7], Louise le Roux[1], Julius Gudmundsson[1], Hrefna Johannsdottir[1], Asmundur Oddsson[1], Arnaldur Gylfason[1], Olafur T Magnusson[1], Gisli Masson[1], Thorvaldur Jonsson[6,8], Halla Skuladottir[9], Daniel F Gudbjartsson[1,2], Unnur Thorsteinsdottir[1,6], Patrick Sulem[1] & Kari Stefansson[1,6]

Gastric cancer is a serious health problem worldwide, with particularly high prevalence in eastern Asia. Genome-wide association studies (GWAS) in Asian populations have identified several loci that associate with gastric cancer risk. Here we report a GWAS of gastric cancer in a European population, using information on 2,500 population-based gastric cancer cases and 205,652 controls. We found a new gastric cancer association with loss-of-function mutations in *ATM* (gene test, $P = 8.0 \times 10^{-12}$; odds ratio (OR) = 4.74). The combination of the loss-of-function variants p.Gln852*, p.Ser644* and p.Tyr103* (combined minor allele frequency (MAF) = 0.3%) also associates with pancreatic and prostate cancers (OR = 3.81 and 2.18, respectively) and gives an indication of risk of breast and colorectal cancers (OR = 1.82 and 1.97, respectively). Cancers in those carrying loss-of-function *ATM* mutations are diagnosed at a significantly earlier age than in non-carriers. Our results confirm an association between gastric cancer in Europeans and three loci previously reported in Asians, *MUC1*, *PRKAA1* and *PSCA*, refine the association signal at *PRKAA1* and support a pathogenic role for the tandem repeat identified in *MUC1*.

Although the incidence and mortality rates of gastric cancer have gradually declined over the last five decades, the disease still presents a major worldwide health burden[1,2]. Gastric cancer is the fifth most common cancer in the world and the third leading cause of cancer-related death, with age-standardized incidence rates highest in eastern Asia[1].

Gastric cancer risk is influenced both by environmental and genetic factors. The most important exogenous gastric cancer risk factor is infection with *Helicobacter pylori*, but other risk factors include consumption of red and processed meats, gastric reflux and smoking[3]. Family history of gastric cancer increases risk, with reported relative risks ranging from about 1.3 to 3.0 for the first-degree relatives of gastric cancer cases[4]. Recently, GWAS conducted in Asian populations have identified associations between the disease and common variants located at 8q24 (*PSCA*)[5], 1q22 (*MUC1*)[6], 10q23 (*PLCE1*)[7], 3q13.31 (*ZBTB20*)[8], 5p13.1 (*PTGER4-PRKAA1*)[8], 6p21 (*LRFN2*)[9] and 7p15.3 (*DNAH11*)[9]. The associations at 1q22 (*MUC1*) and 8q24 (*PSCA*) have been replicated in European populations[6,10–12], but the other loci have yet to be confirmed in non-Asian populations. In particular, a study in two European-ancestry populations (from the United States and Poland) failed to detect an association between gastric cancer and rs2274223 in the *PLCE1* locus[12].

This study is based on 104,220 Icelanders who have been genotyped on microarray platforms. In addition, we have sequenced the whole genomes of 2,636 Icelanders (including 18 gastric cancer cases) to a median depth of 20× and found 20.6 million SNPs and 4.4 million small indels[13]. Using long-range phasing, we imputed these sequence variants into the genomes of those genotyped on microarray platforms and also into the ungenotyped close relatives of genotyped individuals, using genealogy information on all Icelanders[14]. We thus can infer genotypes for individuals with gastric cancer who were diagnosed and died a long time ago, allowing us to stretch the study back to the time when the disease had much higher incidence than at present.

To search for sequence variants that associate with gastric cancer, we performed a GWAS using information on all gastric cancer diagnoses in the Icelandic Cancer Registry[15] (ICR; spanning the period from 1955 to 2012) and genotype information on 25.0 million sequence variants (**Supplementary Figs. 1a** and **2a**). The gastric cancer associations were based on 400 chip-typed gastric cancer cases and 80,629 chip-typed controls in addition to 2,100 cases and 125,023 controls who had at least one of their first- or second-degree relatives chip typed (Online Methods). Of the 3,363 patients with gastric cancer in the ICR, 2,935 (87%) were deceased before the start of recruitment in 2001, reflecting the fact that gastric cancer incidence was tenfold higher in 1955 than it is today (Online Methods). In the first years of

[1]deCODE Genetics/Amgen, Reykjavik, Iceland. [2]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. [3]Department of Medicine, Landspitali-University Hospital, Reykjavik, Iceland. [4]Department of Pathology, Landspitali-University Hospital, Reykjavik, Iceland. [5]Icelandic Cancer Registry, Reykjavik, Iceland. [6]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [7]Children's Hospital Iceland, Landspitali-University Hospital, Reykjavik, Iceland. [8]Department of Surgery, Landspitali-University Hospital, Reykjavik, Iceland. [9]Department of Oncology, Regional Hospital West Jutland, Herning, Denmark. [10]These authors contributed equally to this work. Correspondence should be addressed to H.H. (hannes.helgason@decode.is) or K.S. (kstefans@decode.is).

Received 10 December 2014; accepted 29 May 2015; published online 22 June 2015; doi:10.1038/ng.3342

cancer registration, a large fraction of cases received an unspecified histology code corresponding to Systematized Nomenclature of Medicine–Clinical Terminology (SNOMED-CT) code 8000 or 8010 (neoplasm malignant and carcinoma NOS (not otherwise specified), respectively). We therefore performed a second GWAS scan using only gastric cancer cases with histologically verified adenocarcinomas (2,043 adenocarcinomas and 202,533 controls) (**Supplementary Figs. 1b** and **2b**). We used the weighted Holm-Bonferroni method[16] to allocate a family-wise error rate of 0.05 equally among three annotation-based classes of sequence variation. This allocation yielded significance thresholds of 3.1×10^{-6} for loss-of-function variants (including stop-gain, frameshift, and splice acceptor or donor variants; n = 5,432), 1.5×10^{-7} for missense, splice region and in-frame indel variants (n = 109,370) and 6.7×10^{-10} for other (noncoding) variants (n = 24,873,426); variant annotation was based on protein-coding transcripts from RefSeq release 56 (ref. 17) (Online Methods).

First considering noncoding variants, no variant exceeded the significance threshold of 6.7×10^{-10}; however, variants at 1q22, a well-established gastric cancer locus covering the *MUC1* gene, were close to the threshold (**Table 1** and **Supplementary Fig. 1**). The top gastric cancer association signal at the *MUC1* locus is represented by two strongly correlated variants ($r^2 = 0.97$): rs140081212[A] downstream of *MTX1* ($P = 7.9 \times 10^{-10}$, OR = 0.79, allelic frequency (AF) = 35.4%) and the missense variant rs760077[A] in *MTX1* ($P = 1.1 \times 10^{-9}$, OR = 0.79, AF = 35.1%, p.Thr63Ser). Incidentally, the missense variant rs760077 is the only variant in the class of missense, splice region and in-frame deletion variants that passed the significance threshold of 1.5×10^{-7} for its class (**Table 1**). rs140081212 is not present in 1000 Genomes Project Pilot 1 data and will not be discussed further.

In the CHB and JPT (Han Chinese and Japanese) panel in 1000 Genomes Project Pilot 1 data (see URLs), the variant rs760077 is perfectly correlated ($r^2 = 1.00$) with the variant rs4072037 that was

Table 1 Summary of single-marker associations for gastric cancer in Iceland

									GC all		GC verified adenocarcinomas	
rs ID	Chr.	Position (hg18; bp)	A1[a]	A2[a]	Freq., (%)[a]	Info[d]	Gene	Coding change[e]	*P* value	OR (95% CI)[a]	*P* value	OR (95% CI)[a]
Associations with $P < 1 \times 10^{-7}$												
rs140081212	1	153,451,599	A	G	35.4	1.00	–	–	7.9×10^{-10}	0.79 (0.73–0.85)	1.8×10^{-8}	0.79 (0.73–0.86)
rs760077	1	153,445,406	A	T	35.1	0.99	*MTX1*[c]	p.Thr63Ser	1.1×10^{-9}	0.79 (0.73–0.85)	2.6×10^{-8}	0.79 (0.73–0.86)
_	1	153,461,695[b]	ATTAT	!ATTAT	35.9	0.98	–	–	6.6×10^{-9}	0.80 (0.74–0.86)	1.1×10^{-7}	0.80 (0.74–0.87)
rs6676150	1	153,390,461	C	G	33.3	1.00	–	–	1.5×10^{-8}	0.80 (0.74–0.86)	4.8×10^{-8}	0.79 (0.73–0.86)
_	11	102,118,158[b]	A	!A	0.15	0.96	–	–	3.4×10^{-8}	4.9 (2.79–8.62)	3.6×10^{-7}	4.88 (2.65–8.98)
_	2	85,874,332	T	G	0.04	0.81	–	–	1.4×10^{-7}	11.1 (4.53–27.19)	7.6×10^{-8}	12.7 (5.03–32.08)
Missense or splice region variants with $P < 1 \times 10^{-5}$												
rs760077	1	153,445,406	A	T	35.1	0.99	*MTX1*[c]	p.Thr63Ser	1.1×10^{-9}	0.79 (0.73–0.85)	2.6×10^{-8}	0.79 (0.73–0.86)
rs35001694	1	153,524,442	T	C	2.9	0.98	*HCN3*[c]	p.Pro630Leu	2.6×10^{-6}	0.55 (0.43–0.71)	2.4×10^{-5}	0.56 (0.43–0.73)
rs55864139	20	39,482,993	A	T	0.29	0.99	*CHD6*	p.Asn1899Ile	2.4×10^{-5}	2.78 (1.73–4.46)	6.9×10^{-6}	3.12 (1.90–5.11)
Loss-of-function variants with $P < 1 \times 10^{-3}$												
–	11	107,643,195	T	C	0.16	1.00	*ATM*	p.Gln852*	5.5×10^{-7}	4.27 (2.42–7.54)	2.0×10^{-7}	4.84 (2.67–8.77)
–	11	107,629,783	A	C	0.07	0.88	*ATM*	p.Ser644*	2.7×10^{-6}	6.87 (3.07–15.37)	1.4×10^{-6}	7.78 (3.38–17.91)
rs138377917	8	143,760,533	A	G	2.68	0.98	*PSCA*	p.Trp109*	6.1×10^{-4}	0.65 (0.51–0.83)	4.0×10^{-4}	0.61 (0.46–0.80)
–	11	11,330,411	A	G	0.49	0.99	*CSNK2A3*	p.Arg278*	7.6×10^{-4}	2.04 (1.35–3.09)	5.7×10^{-3}	1.91 (1.21–3.03)
–	9	132,794,879	A	G	0.04	0.96	*FIBCD1*	p.Arg150*	9.4×10^{-4}	5.67 (2.03–15.85)	1.4×10^{-3}	6.01 (2.00–18.02)
–	22	27,421,856	A	AG	0.13	0.98	*CHEK2*	p.Thr410Metfs*15	1.4×10^{-3}	3.04 (1.54–5.99)	6.6×10^{-4}	3.42 (1.68–6.95)
rs146753414	12	50,998,014	A	C	0.64	0.98	*KRT83*	p.Glu201*	9.7×10^{-3}	0.47 (0.26–0.83)	7.3×10^{-4}	0.26 (0.12–0.57)

The associations are based on 2 overlapping sets: 2,500 gastric cancer cases and 205,652 controls (GC all) and 2,043 verified gastric cancer adenocarcinomas and 202,533 controls (GC verified adenocarcinomas) (Online Methods). Shown are associations that reached the prescribed significance threshold for either set; ordering of the results is with respect to the *P* values for the GC all set. The variants considered in the genome-wide scans had imputation information (info) greater than 0.8. Chr., chromosome.

[a] A1 and A2 are the two alleles tested for the marker. Odds ratios (ORs) and population allele frequencies (Freq.) are given for allele A1. [b] The marker is multi-allelic, and the association is for the allele shown in column A1 against all other alleles for the marker (!A1 in column A2). [c] The genes *MTX1* and *HCN3* are at the 1q22 (*MUC1*) locus. [d] Imputation information for markers estimated by the ratio of the variance in imputed expected allele counts and the variance in the actual allele counts (Online Methods). [e] RefSeq accession identifiers: *MTX1*, NP_002446.2; *HCN3*, NP_065948.1; *ATM*, NP_000042.3; *PSCA*, NP_005663.2; *CSNK2A3*, NP_001243615.1; *FIBCD1*, NP_001138578.1; *CHD6*, NP_115597.3; *CHEK2*, NP_001005735.1; *KRT83*, NP_002273.3.

previously reported to associate with gastric cancer in Asians[7,18] and Europeans[6,12]. The variants rs760077 and rs4072037 are correlated to a lesser extent in Europeans than in Asians (Iceland, $r^2 = 0.75$; CEU panel (Utah residents of Northern and Western European ancestry) in 1000 Genomes Project Pilot 1 data, $r^2 = 0.64$). It has been suggested that the variant rs4072037[T] reported in Asians may be pathogenic and acting through its effect on alternative splicing of *MUC1* (refs. 6,18,19). However, the gastric cancer risk associated with the *MUC1* locus has also been linked to the number of tandem repeats present in exon 2 of the gene[20]. Conditional analysis showed that both variants, rs760077 and the reported rs4072037, became more significant after conditioning on each other; however, whereas the protective effect for rs760077[A] became stronger ($P = 3.4 \times 10^{-11}$, OR = 0.59) when adjusting for rs4072037, the effect of rs4072037[C] went from being protective to conferring risk ($P = 2.7 \times 10^{-5}$, OR = 1.37) when adjusting for rs760077 (**Supplementary Table 1**). Haplotype analysis showed that the gastric cancer signal at the *MUC1* locus cannot be explained by rs4072037 alone and that in Europeans the variant rs760077, along with rs4072037, better represents the signal (Online Methods and **Supplementary Table 2**). When analyzing these two variants in relation to sequencing coverage of *MUC1*, we conclude that our data are consistent with a role for the tandem repeat in *MUC1* in gastric cancer pathogenesis (Online Methods).

Table 2 Associations in *ATM* loss-of-function burden tests for cancer phenotypes

					Age at diagnosis (years)		
Phenotype	*n* affected[a]	*n* controls[a]	*P* value	OR (95% CI)[b]	Average (s.d.)[c]	*P* value	β (95% CI)[b]
Gastric cancer[d]	2,500	205,652	8.0×10^{-12}	4.74 (3.03–7.40)	69.6 (12.4)	0.0078	−6.11 (−10.61, −1.61)
Gastric cancer, verified adenocarcinoma[d]	2,043	202,533	1.0×10^{-12}	5.42 (3.41–8.63)	69.1 (12.3)	0.013	−6.05 (−10.84, −1.26)
Pancreatic cancer	1,074	174,982	6.4×10^{-5}	3.81 (1.98–7.34)	70.3 (11.7)	0.028	−7.25 (−13.72, −0.78)
Prostate cancer	4,858	83,103	5.5×10^{-4}	2.18 (1.40–3.39)	72.2 (9.2)	0.095	−2.63 (−5.72, 0.46)
Breast cancer	5,318	280,808	0.0041	1.82 (1.21–2.74)	60.6 (14.3)	0.48	−1.67 (−6.30, 2.96)
Colorectal cancer	3,546	236,404	0.0071	1.97 (1.20–3.23)	69.8 (12.7)	0.080	−4.76 (−10.09, 0.57)

In total, 25 cancers were tested, and the cancers not listed in this table had $P > 0.10$. See **Supplementary Table 9** for the full list of association results.

[a] Number of individuals used in the association tests. [b] Odds ratios (ORs) and allelic effects (β) correspond to the *ATM* loss-of-function genotype (MAF for the collapsed variant = 0.26%). [c] The average for age at diagnosis was calculated for the whole group of cancer cases, irrespective of genotype. [d] The gastric cancer associations are based on two overlapping sets, one including all individuals who have been diagnosed with gastric cancer and the other limiting analysis to individuals with verified gastric cancer adenocarcinomas (Online Methods).

Two loss-of-function variants passed the significance threshold 3.1×10^{-6}. The most strongly associated variant was a rare stop-gain variant in *ATM* at chr11:107643195[T] (hg18) (p.Gln852*, AF = 0.16%) with a high OR ($P = 5.5 \times 10^{-7}$, OR = 4.27). The second strongest loss-of-function variant was another rare stop-gain variant in *ATM* at chr11:107629783[A] (hg18) (p.Ser644*, AF = 0.07%), also with a high OR ($P = 2.7 \times 10^{-6}$, OR = 6.87) (**Table 1**). These two loss-of-function variants are uncorrelated ($r^2 < 1 \times 10^{-6}$) and never occurred on the same chromosome in our data (**Supplementary Figs. 3** and **4**, and **Supplementary Tables 3** and **4**). The two significant loss-of-function variants in *ATM* provide two independent associations of variants in the gene with gastric cancer; the two variants individually reached genome-wide significance.

Gene-level tests, such as burden tests, provide an approach to jointly test a set of variants in a given gene[21–23]. We performed a burden test association scan[21] where we collapsed all rare (MAF < 1%) loss-of-function variants in a gene into one allele (Online Methods). The most significant gene in the scan was *ATM* ($P = 8.0 \times 10^{-12}$, OR = 4.74), where three uncorrelated stop-gain variants, p.Gln852*, p.Ser644* and p.Tyr103* ($r^2 < 1 \times 10^{-6}$ for all pairs), were collapsed (cumulative MAF = 0.26%) (**Supplementary Table 5**); all three variants were on distinct haplotypes. *ATM* was the only gene in the scan that passed a significance threshold accounting for the total number of genes in the genome ($P = 0.05/20{,}000 = 2.5 \times 10^{-6}$; **Supplementary Table 5**). All three loss-of-function variants resulted in an increased risk for gastric cancer, and the effects of the three variants were not significantly different (heterogeneity *P* value (P_{het}) = 0.52, I^2 (measure of degree of inconsistency) = 0.0; **Supplementary Table 6**). To assess the quality of the imputation, we genotyped all imputed carriers of the rare *ATM* variants in our chip-genotyped population of 104,220 Icelanders using single-SNP assays. In total, we genotyped 744 individuals (59 imputed carriers of p.Tyr103*, 124 imputed carriers of p.Ser644*, 334 imputed carriers of p.Gln852* and 227 randomly selected individuals imputed not to carry any of the 3 variants). The correlation between imputed and directly measured genotypes was 0.91, 0.96 and 1.00 for p.Tyr103*, p.Ser644* and p.Gln852*, respectively (**Supplementary Table 7**). The genotyping of these carriers also confirmed that the three variants never occurred on the same chromosome in our data, as none of the genotyped individuals carried more than one of the variants.

Information on the type of gastric cancer was available for a subset of the cases. We tested the association of the collapsed loss-of-function *ATM* variant with four gastric cancer subphenotypes—intestinal, diffuse, cardia and non-cardia gastric cancers. In all instances, we observed increased risk for carriers of the loss-of-function *ATM* variant, showing that the gastric cancer association is not limited to a particular subtype of this cancer (**Supplementary Table 8**).

ATM has a key role in the DNA damage response, and homozygous and compound-heterozygous loss-of-function (including some missense) mutations in *ATM* cause ataxia telangiectasia (A-T) syndrome. Previous studies have demonstrated an increased risk of breast cancer in heterozygotes for *ATM* mutations and given suggestive evidence of excess risk of colorectal, gastric and pancreatic cancers in relatives of A-T cases and kindreds with familial cancer[24–26]. We tested association between the collapsed loss-of-function variant in *ATM* and 24 additional cancer types using population-based samples (**Table 2** and **Supplementary Table 9**). Two cancer types passed a Bonferroni significance threshold for 24 tests ($P < 0.05/24 = 0.0021$): pancreatic cancer ($P = 6.4 \times 10^{-5}$, OR = 3.81) and prostate cancer ($P = 5.5 \times 10^{-4}$, OR = 2.18). A linear model assuming an additive effect of the collapsed loss-of-function *ATM* variant on age at diagnosis indicated that loss-of-function *ATM* mutation carriers are diagnosed with these cancers at a considerably earlier age than non-carriers; the allelic effects (β) ranged from −1.7 to −7.3 years, with the association for gastric cancer being most significant ($P = 0.0078$, $\beta = -6.1$ years) (**Table 2**, Online Methods, **Supplementary Figs. 5–10** and **Supplementary Table 10**). When considering the general Icelandic population, carriers of the collapsed loss-of-function *ATM* variant had a shorter lifespan than non-carriers among deceased individuals who lived to be 50 years old and were born after 1890 ($P = 0.0017$, $\beta = -3.0$ years, $n = 62{,}558$).

In addition to the 3 *ATM* loss-of-function variants used in the imputation, we had evidence from whole-genome sequencing of 3 other loss-of-function variants in *ATM* that were not imputable in our data set (imputation information < 0.4); one of these variants was observed in 2 of the 2,636 sequenced individuals, and the other 2 were observed in only one individual each (**Supplementary Table 11**).

To assess the pathogenicity of the predicted loss-of-function variants in *ATM* described here, we checked for their presence in A-T cases in Iceland. One A-T case was found to be a compound-heterozygous carrier of p.Tyr103* and p.Arg457* (confirmed using Sanger sequencing), and a second A-T case was homozygous for p.Gln852* (confirmed using direct genotyping). These data, along with the previous report of p.Ser644* in an A-T case[27], give evidence for the loss-of-function effects of the sequence variants p.Gln852*, p.Ser644* and p.Tyr103* that were used in our association tests.

In addition to the GWAS and gene tests, we scanned six previously reported gastric cancer susceptibility loci, *PSCA*[5], *PLCE1* (ref. 7), *ZBTB20* (ref. 8), *PRKAA1-PTGER4* (ref. 8), *LRFN2* (ref. 9) and *DNAH11* (ref. 9), to search for evidence of association in the Icelandic population. We found significant associations at the *PSCA* and *PRKAA1-PTGER4* loci. However, the reported SNPs at the remaining loci did not replicate, and no biallelic variants were significant across each of these loci in the Icelandic data in a 500-kb window centered on the reported SNP (Online Methods).

The 5′ UTR variant rs2294008[T] in *PSCA*, which has previously been reported to associate with gastric cancer, replicated in Icelanders ($P = 2.4 \times 10^{-7}$, OR = 1.21) and showed an effect both for diffuse- and intestinal-type gastric cancers (**Supplementary Table 8**). Our best gastric cancer association within a 500-kb window centered on *PSCA* was with the variant rs2920295[G] ($P = 1.0 \times 10^{-7}$, OR = 1.21), which in Iceland is almost perfectly correlated with the reported variant rs2294008 ($r^2 = 0.99$) (**Supplementary Table 12**). No other variant remained significant in the locus after adjusting for the top variant rs2920295 (**Supplementary Fig. 11**). A stop-gain variant, rs138377917[A], in *PSCA* ($P = 6.1 \times 10^{-4}$, OR = 0.65) passed a Bonferroni significance threshold for an association scan limited to 31 coding variants in the locus; this variant was not fully explained by the reported variant rs2294008 ($P = 0.011$, OR = 0.72 when adjusting for rs2294008) (Online Methods and **Supplementary Table 13**).

The variant rs13361707[C] in *PRKAA1*, which was previously reported in the Han Chinese population[8], associated with gastric cancer in the Icelandic data ($P = 2.7 \times 10^{-4}$, OR = 1.16). The variant that associated most significantly with gastric cancer in Iceland at this locus was an intronic variant in *PTGER4* (rs10036575[C], $P = 4.8 \times 10^{-6}$, OR = 0.81) that turned out to be a refinement in Europeans of the signal previously reported with rs13361707 (Online Methods, **Supplementary Fig. 12** and **Supplementary Table 14**).

In conclusion, a GWAS using whole-genome sequence data confirmed *ATM* as a high-risk gene for gastric cancer in a European population. The association between mutations in *ATM* and gastric cancer is intriguing, as it has been shown that *H. pylori* can induce DNA double-strand breaks in target cells and trigger a DNA damage response that involves upregulation of *ATM* expression[28]. ATM has a key role in maintaining the integrity of the genome, and loss-of-function mutations in the gene are very rare. Notably, the Exome Aggregation Consortium (ExAC) database reports 111 loss-of-function *ATM* variants that are diverse geographically and carried by 171 individuals who are all heterozygotes (**Supplementary Table 15**). The combined MAF of all these variants in ExAC is 0.14%, and the most frequent variant in the list has a frequency of 0.0066% (observed on 8 of 121,104 chromosomes). This scarcity of loss-of-function mutations in *ATM* has complicated studies on the association of the gene with cancer. It is possible to observe associations in Iceland because of a founder effect that allows these rare variants to reach a sufficiently high allele frequency for detection.

URLs. Ensembl, release 76 (accessed 12 August 2014), http://www.ensembl.org/; SNAP v.2.2 (for CHB and JPT proxies; r^2 threshold of 0.8 and distance limit of 500 kb; accessed 14 August 2014), https://www.broadinstitute.org/mpg/snap/ldsearch.php; GTEx (accessed 28 October 2014), http://www.gtexportal.org/; Icelandic Cancer Registry (ICR; accessed 4 November 2014), http://www.krabbameinsskra.is/; Exome Aggregation Consortium (ExAC) database (accessed 5 March 2015), http://exac.broadinstitute.org/; Picard, http://picard.sourceforge.net/.

METHODS

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

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the individuals who participated in the study and whose contribution made this work possible. We also thank the personnel at the recruitment center. We acknowledge the Icelandic Cancer Registry (ICR) for assistance in the ascertainment of patients with cancer.

AUTHOR CONTRIBUTIONS

H.H., T.R., P.S. and K.S. designed the study and interpreted the results. H.S.O., J.G.J., L.T., K.A., A.H., T.J. and H.S. carried out the subject ascertainment, recruitment and collection of clinical data. H.H., T.R., A.S., S.N.S., A.J., L.I.R., J.G., H.J., A.O., O.T.M., G.M. and U.T. performed the sequencing, genotyping and expression analyses. H.H., A.G., D.F.G. and P.S. performed the statistical and bioinformatics analyses. H.H., T.R., P.S. and K.S. drafted the manuscript. All authors contributed to the final version of the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2014).
2. Parkin, D.M., Stjernswärd, J. & Muir, C.S. Estimates of the worldwide frequency of twelve major cancers. *Bull. World Health Organ.* **62**, 163–182 (1984).
3. de Martel, C., Forman, D. & Plummer, M. Gastric cancer: epidemiology and risk factors. *Gastroenterol. Clin. North Am.* **42**, 219–240 (2013).
4. Hemminki, K., Sundquist, J. & Ji, J. Familial risk for gastric carcinoma: an updated study from Sweden. *Br. J. Cancer* **96**, 1272–1277 (2007).
5. Study Group of Millennium Genome Project for Cancer. Genetic variation in *PSCA* is associated with susceptibility to diffuse-type gastric cancer. *Nat. Genet.* **40**, 730–740 (2008).
6. Jia, Y. *et al.* A comprehensive analysis of common genetic variation in *MUC1*, *MUC5AC*, *MUC6* genes and risk of stomach cancer. *Cancer Causes Control* **21**, 313–321 (2010).
7. Abnet, C.C. *et al.* A shared susceptibility locus in *PLCE1* at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nat. Genet.* **42**, 764–767 (2010).
8. Shi, Y. *et al.* A genome-wide association study identifies new susceptibility loci for non-cardia gastric cancer at 3q13.31 and 5p13.1. *Nat. Genet.* **43**, 1215–1218 (2011).
9. Jin, G. *et al.* Genetic variants at 6p21.1 and 7p15.3 are associated with risk of multiple cancers in Han Chinese. *Am. J. Hum. Genet.* **91**, 928–934 (2012).
10. Lochhead, P. *et al.* Genetic variation in the prostate stem cell antigen gene and upper gastrointestinal cancer in white individuals. *Gastroenterology* **140**, 435–441 (2011).
11. Sala, N. *et al.* Prostate stem-cell antigen gene is associated with diffuse and intestinal gastric cancer in Caucasians: results from the EPIC-EURGAST study. *Int. J. Cancer* **130**, 2417–2427 (2012).

12. Palmer, A.J. *et al.* Genetic variation in *C20orf54*, *PLCE1* and *MUC1* and the risk of upper gastrointestinal cancers in Caucasian populations. *Eur. J. Cancer Prev.* **21**, 541–544 (2012).
13. Gudbjartsson, D.F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat. Genet.* **47**, 435–444 (2015).
14. Kong, A. *et al.* Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* **40**, 1068–1075 (2008).
15. Sigurdardottir, L.G. *et al.* Data quality at the Icelandic Cancer Registry: comparability, validity, timeliness and completeness. *Acta Oncol.* **51**, 880–889 (2012).
16. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
17. Pruitt, K.D., Tatusova, T., Brown, G.R. & Maglott, D.R. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.* **40**, D130–D135 (2012).
18. Saeki, N. *et al.* A functional single nucleotide polymorphism in mucin 1, at chromosome 1q22, determines susceptibility to diffuse-type gastric cancer. *Gastroenterology* **140**, 892–902 (2011).
19. Ng, W., Loh, A.X.W., Teixeira, A.S., Pereira, S.P. & Swallow, D.M. Genetic regulation of *MUC1* alternative splicing in human tissues. *Br. J. Cancer* **99**, 978–985 (2008).
20. Carvalho, F. *et al. MUC1* gene polymorphism and gastric cancer—an epidemiological study. *Glycoconj. J.* **14**, 107–111 (1997).
21. Lange, L.A. *et al.* Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am. J. Hum. Genet.* **94**, 233–245 (2014).
22. Majithia, A.R. *et al.* Rare variants in *PPARG* with decreased activity in adipocyte differentiation are associated with increased risk of type 2 diabetes. *Proc. Natl. Acad. Sci. USA* **111**, 13127–13132 (2014).
23. Liu, D.J. *et al.* Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* **46**, 200–204 (2014).
24. Renwick, A. *et al. ATM* mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat. Genet.* **38**, 873–875 (2006).
25. Thompson, D. *et al.* Cancer risks and mortality in heterozygous *ATM* mutation carriers. *J. Natl. Cancer Inst.* **97**, 813–822 (2005).
26. Roberts, N.J. *et al. ATM* mutations in patients with hereditary pancreatic cancer. *Cancer Discov.* **2**, 41–46 (2012).
27. Li, A. & Swift, M. Mutations at the ataxia-telangiectasia locus and clinical phenotypes of A-T patients. *Am. J. Med. Genet.* **92**, 170–177 (2000).
28. Toller, I.M. *et al.* Carcinogenic bacterial pathogen *Helicobacter pylori* triggers DNA double-strand breaks and a DNA damage response in its host cells. *Proc. Natl. Acad. Sci. USA* **108**, 14944–14949 (2011).

ONLINE METHODS

The Icelandic study population. This study is based on whole-genome sequence data from the whole blood of 2,636 Icelanders participating in various disease projects at deCODE genetics (**Supplementary Tables 16** and **17**). In addition, a total of 104,220 Icelanders have been genotyped using Illumina SNP chips (**Supplementary Table 18**).

Affected individuals were identified through the ICR. The ICR contains all cancer diagnoses in Iceland since 1 January 1955; over 90% of diagnoses are histologically confirmed. The ICR contains records of 3,363 Icelandic patients with gastric cancer (65% male) diagnosed through 31 December 2012. Of the 3,363 Icelandic patients with gastric cancer, 2,935 (87%) were deceased before the start of recruitment in 2001, reflecting the high incidence of gastric cancer at the time cancer registration was initiated. Recruitment of gastric cancer cases was initiated in 2001 and included all prevalent cases as well as newly diagnosed cases from that time onward. Of the 386 gastric cancer cases diagnosed from 1 January 2001 to 31 December 2012, 271 (70%) participated in our study. Patients are recruited by trained nurses on behalf of the physicians treating the patients, through special recruitment clinics. Participants in the study sign an informed consent form, donate a blood sample and answer a lifestyle questionnaire.

The great majority of gastric cancer cases (about 90%) are adenocarcinomas, and most of the genetic studies published thus far focus on this histological form of the disease. Our study population differs from other gastric cancer populations in that a large fraction of our cases were diagnosed decades ago, when the incidence of gastric cancer was much higher than it is today (in 1955, age-standardized incidence of gastric cancer per 100,000 was 78 for males and 30 for females, whereas it is currently 7.5 and 4.3 for males and females, respectively). In the first years of cancer registration, a large fraction of cases received an unspecified histology code corresponding to SNOMED code 8000 or 8010 (neoplasm malignant and carcinoma NOS, respectively). For example, for the first 5 years of registration, 234 of 452 (52%) gastric cancer diagnoses received such a code. To use the full power of our population, 2 GWAS scans were performed: the first scan included all cases regardless of histology (2,500 gastric cancer cases and 205,652 controls) and the second scan used only gastric cancer cases with histologically verified adenocarcinomas (2,043 adenocarcinomas and 202,533 controls). Results from both types of analysis are presented in **Tables 1** and **2**, and **Supplementary Tables 3**–**6**, **8** and **9**.

A total of 400 cases (59% male; diagnosed from 1958 to 2012) were included in a genome-wide SNP genotyping effort, using the Infinium II assay method (Illumina) and one of the BeadChips listed in **Supplementary Table 18**. The median age at diagnosis for all consenting cases was 71 years (range of 20–100 years), the same as that for all patients with gastric cancer in the ICR. In addition to the chip-genotyped cases, we used information on 2,100 gastric cancer cases (median age of 70 years; 68% male) without chip information whose genotype probabilities were imputed using methods of familial imputation[14]. The 205,652 controls (80,629 chip typed and 125,023 with familial imputation) used in this study consisted of individuals from other ongoing GWAS at deCODE genetics. No individual disease group represented more than 10% of the total control group. Samples from other cancer cases used in the cross-risk analysis came from other ongoing projects at deCODE genetics. All subjects were of European ancestry.

All sample identifiers were encrypted in accordance with the regulations of the Icelandic Data Protection Authority. Approval for the study was granted by the Icelandic National Bioethics Committee (reference 99/096) and the Icelandic Data Protection Authority (reference 2001120906).

The Icelandic genealogy. The Icelandic genealogical database contains 819,410 individuals going back to 740 CE. Of the 471,284 Icelanders recorded to have been born in the twentieth century, 91.1% had a recorded father and 93.7% had a recorded mother in the database. Similarly, of the 183,896 Icelanders recorded to have been born in the nineteenth century, 97.5% had a recorded father and 97.8% had a recorded mother.

Single-track variant genotyping. Genotyping of single SNPs was carried out by deCODE genetics in Reykjavik, Iceland, applying the Centaurus (Nanogen) platform. The quality of imputation was evaluated by comparing imputed genotypes to genotypes obtained using the assay.

Illumina SNP chip genotyping. The chip-typed Icelandic samples were assayed with several types of Illumina BeadChips at the deCODE genetics facility (**Supplementary Table 18**). SNPs were excluded if they had (i) a yield lower than 95%, (ii) a MAF of less than 0.01% in the population, (iii) excessive deviation from Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$), (iv) an excessive inheritance error rate (>0.001) or (v) if there was a substantial difference in AF between chip types (in which case, the SNP was removed from a single chip type if that resolved the difference; however, if it did not, then the SNP was removed from all chip types). All samples with a call rate of <97% were removed from the analysis. All samples with a call rate below 97% were excluded from the analysis.

Long-range phasing. Long-range phasing of all chip-genotyped individuals was performed with methods described previously[14,29]. In brief, phasing is achieved using an iterative algorithm that phases a single proband at a time given the available phasing information about everyone else that shares a long haplotype identically by state with the proband. Given the large fraction of the Icelandic population that has been chip typed, accurate long-range phasing is available across the genome for all chip-typed Icelanders. For long-range phased haplotype association analysis, we then partitioned the genome into non-overlapping fixed 0.3-cM bins. Within each bin, we observed the haplotype diversity described by the combination of all chip-typed markers in the bin.

Sample preparation for whole-genome sequencing. Paired-end libraries for sequencing were prepared according to the manufacturer's instructions (Illumina, TruSeq). In short, approximately 1 μg of genomic DNA, isolated from frozen blood samples, was fragmented to a mean target size of 300 bp using a Covaris E210 instrument. The resulting fragmented DNA was end repaired using T4 and Klenow polymerases and T4 polynucleotide kinase with 10 mM dNTP, followed by the addition of an adenine base at the ends using Klenow exo fragment ($3'$-to-$5'$ exo^{-}) and dATP (1 mM). Sequencing adaptors containing thymine overhangs were ligated to the DNA products, and fragments were separated by agarose (2%) gel electrophoresis. Fragments of about 400–500 bp were isolated from the gels (Qiagen Gel Extraction kit), and the adaptor-modified DNA fragments were PCR enriched for ten cycles using Phusion DNA polymerase (Finnzymes Oy) and a PCR primer cocktail (Illumina). Enriched libraries were further purified using AMPure XP beads (Beckman Coulter). The quality and concentrations of the libraries were assessed with the Agilent 2100 Bioanalyzer using the DNA 1000 LabChip (Agilent Technologies). Barcoded libraries were stored at −20 °C. All steps in the workflow were monitored using an in-house laboratory information management system with barcode tracking of all samples and reagents.

Whole-genome sequencing. Template DNA fragments were hybridized to the surface of flow cells (Genome Analyzer Paired-End cluster kit (v2) or HiSeq Paired-End cluster kit (v2.5 or v3)) and amplified to form clusters using the Illumina cBot. In brief, DNA (2.5–12 pM) was denatured and then hybridized to the grafted adaptors on the flow cell. Isothermal bridge amplification using Phusion polymerase was followed by linearization of the bridged DNA, denaturation, blocking of $3'$ ends and hybridization of the sequencing primer. Sequencing by synthesis (SBS) was performed on Illumina Genome Analyzer IIx and/or HiSeq 2000 instruments. Paired-end libraries were sequenced for 2 × 101 (HiSeq) or 2 × 120 (Genome Analyzer IIx) cycles of incorporation, with imaging performed using the appropriate TruSeq SBS kits. Each library or sample was initially run on a single Genome Analyzer IIx lane for quality control validation and then underwent further sequencing on either the Genome Analyzer IIx (≥4 lanes) or HiSeq (≥1 lane) with targeted raw cluster densities of 500,000–800,000/mm^2, depending on the version of the data imaging and analysis packages (SCS2.6-2–9/RTA1.6–1.9, HCS1.3.8–1.4.8/RTA1.10.36–1.12.4.2). Real-time analysis involved conversion of image data to base calling in real time.

Whole-genome alignment, SNP and indel calling, and genotype imputation. In this study, we sequenced 2,636 Icelanders to a median depth of 20×. Reads were aligned to NCBI Build 36 (hg18) of the human reference sequence using Burrows-Wheeler Aligner (BWA) 0.5.7–0.5.9 (ref. 30). Alignments were merged into a single BAM file and marked for duplicates using Picard 1.55. Only non-duplicate reads were used for downstream analyses. Resulting BAM files were realigned and recalibrated using Genome Analysis Toolkit (GATK) version 1.2-29-g0acaf2d[31,32]. Multi-sample calling was performed with GATK version 2.3.9 using all 2,636 BAM files together. A detailed description of genotype imputation is given in the **Supplementary Note**.

Genotype imputation information. The informativeness of genotype imputation is estimated by the ratio of the variance of imputed expected allele counts and the variance of the actual allele counts:

$$\frac{\mathrm{Var}\left(E\left(\theta|\text{chip data}\right)\right)}{\mathrm{Var}(\theta)}$$

where θ is the allele count. Var($E(\theta|$chip data)) is estimated by the observed variance in the imputed expected counts.

Gene and variant annotation. Variants were annotated with information from Ensembl release 70 using Variant Effect Predictor (VEP) version 2.8 (refs. 33,34). Only protein-coding transcripts from RefSeq release 56 (ref. 17) were considered.

Association testing, case-control. Logistic regression was used to test for association between sequence variants and disease, treating disease status as the response and genotype counts as covariates. Other available individual characteristics that correlate with disease status were also included in the model as nuisance variables. These characteristics included sex, county of birth, current age or age at death (first- and second-order terms included), blood sample availability for the individual and an indicator function for the overlap of the lifetime of the individual with the timespan of phenotype collection.

Given genotype counts for n individuals, $g_1, g_2, \ldots, g_n \in \{0,1,2\}$, their phenotypes $y_1, y_2, \ldots, y_n \in \{0,1\}$ and a list of vectors of nuisance parameters $x_1, x_2, \ldots, x_n$, the logistic regression model states that:

$$L_i(\alpha,\beta,\gamma) = P\left(y_i = 1|g_i, x_i\right)$$

$$\mathrm{logit}\left(P\left(y_i = 1|g_i, x_i\right)\right) = \alpha + \beta g_i + \gamma^T x_i, \text{ for all } i \in \{1,2,\ldots,n\}$$

where α, β and γ are the regression coefficients and L_i is the contribution of the ith individual to the likelihood function:

$$L(\alpha,\beta,\gamma) = \prod_{i=1}^{n} L_i(\alpha,\beta,\gamma)$$

It is then possible to test for association on the basis of the asymptotic assumption that the likelihood-ratio statistic follows a χ^2 distribution with 1 degree of freedom:

$$2\log\left(\frac{\max\limits_{\alpha,\beta,\gamma} L(\alpha,\beta,\gamma)}{\max\limits_{\alpha,\gamma} L(\alpha,\beta=0,\gamma)}\right) \sim \chi_1^2$$

Maximizing over the nuisance parameters at every marker in the genome would be prohibitively expensive. We therefore choose to maximize the likelihood under the null hypothesis of $\beta = 0$, which is the same for all markers, and use the maximizer of γ, $\tilde{\gamma}$, under the alternative. Because $\max\limits_{\alpha,\beta,\gamma} L(\alpha,\beta,\gamma) \geq \max\limits_{\alpha,\beta} L(\alpha,\beta,\gamma=\tilde{\gamma})$, this approach will lead to a smaller likelihood ratio than if we were to maximize over γ for every marker.

Our analysis is based on imputed genotype values where the values of g_i are not known. Instead, we use $P(g_i = j|I_i)$ for $j \in \{0,1,2\}$, where I_i represents the information about g_i. Given the logistic regression model above, this approach allows us to calculate:

$$P\left(y_i = 1|I_i\right) = \sum_{j=0}^{2} P\left(g_i = j \mid I_i\right) P\left(y_i = 1|g_i\right), \text{ for all } i \in \{1,2,\ldots,n\}$$

We note that this approach differs from the common approximation of substituting g_i with $Eg_i|I_i$, the expectation of g_i given I_i, in the logistic regression equation above. This approach has a straightforward mathematical justification and requires fewer assumptions than the approximate method, and it seems to be more robust to very uninformative imputations. This approach requires a special implementation of logistic regression and is slightly more computationally taxing.

Association testing, quantitative traits. A generalized form of linear regression was used to test for association for age at diagnosis with indels and SNPs. Let y be the vector of quantitative measurements and let g be the vector of expected allele counts for the variant being tested. We assume that the quantitative measurements follow a normal distribution with a mean that depends linearly on the expected allele at the variant and a variance-covariance matrix proportional to the kinship matrix:

$$y \sim \mathrm{N}\left(\alpha + \beta g, 2\sigma^2 \Phi\right)$$

where

$$\Phi_{ij} = \begin{cases} \frac{1}{2}, & i = j \\ 2k_{ij}, & i \neq j \end{cases}$$

is the kinship matrix as estimated from the Icelandic genealogical database. It is not computationally feasible to use this full model, and we therefore split the individuals with trait values into smaller clusters for the calculation. The maximum-likelihood estimates for the parameters α, β and σ^2 involve inverting the kinship matrix. If there are n individuals in the cluster, then this inversion requires $O(n^3)$ calculations; however, because these calculations only need to be performed once, the computational cost of carrying out a genome-wide association scan will only be $O(n^2)$ calculations per variant, which is the cost of calculating the maximum-likelihood estimates if the kinship matrix has already been inverted.

Adjusting for relatedness. To account for relatedness and stratification within the case and control sample sets, we applied the method of genomic control[35]. The inflation factor λ_g of the χ^2 statistics was estimated on the basis of a set of about 300,000 common variants distributed across the genome, and P values were adjusted by dividing the corresponding χ^2 values by this factor.

Inheritance models in association testing. When testing for association with the additive model, we used the sum of the two imputed haplotype probabilities as covariates for both the logistic regression and the generalized linear regression. Thus, if an individual is imputed to have the minor allele of a sequence variant with probability a_f on his paternal chromosome and a_m on his maternal chromosome, then $a_\mathrm{f} + a_\mathrm{m}$ was used as a covariate.

Gene test by variant collapsing. For each gene, we considered the set of observed loss-of-function sequence variants in that gene with MAF less than 1%. We then estimated, for each imputed individual, the probability that he or she carried a minor allele of at least one of the rare loss-of-function variants that were observed in the gene; the probability was estimated on the basis of the individual's imputed and phased loss-of-function genotypes. Because the collapsed alleles are close together in terms of genomic coordinates, we employed the same imputation procedure into untyped close relatives and performed association testing as for single markers. The method for variant collapsing and gene testing involving rare coding variants (loss-of-function, nonsynonymous and splice region variants of MAF less than 1%) was performed in a similar fashion.

Analysis of the gastric cancer association signal at *MUC1*. We scrutinized the association signal at *MUC1* in the context of previous reports from Asian gastric cancer GWAS. In the CHB and JPT panel in 1000 Genomes Project Pilot 1 data (see URLs), the variant rs760077, corresponding to the strongest gastric cancer signal in our data, is perfectly correlated (r^2 = 1.00) with the variant rs4072037 that has previously been reported to associate with gastric cancer in Asians[7,18] and Europeans[6,12] and fully accounted for the signal at the locus (**Supplementary Fig. 13**). The variants rs760077 and rs4072037 seem to be correlated to a lesser extent in Europeans than in Asians (Iceland, $r^2 = 0.75$; CEU panel of 1000 Genomes Project Pilot 1 data, $r^2 = 0.64$).

In Iceland, the variant rs760077 had a stronger association with gastric cancer than the reported variant rs4072037[C] ($P = 7.6 \times 10^{-4}$, OR = 0.88, AF = 41.6%); it has been suggested that rs4072037[T] might be pathogenic and act through its effect on alternative splicing of *MUC1* (refs. 6,18,19). Both variants rs760077 and rs4072037 became more significant after conditioning on each other; however, whereas the protective effect for rs760077[A] became stronger (adjusted OR = 0.59), the effect of the reported variant rs4072037[C] went from being protective to conferring risk (adjusted OR = 1.37) (**Supplementary Table 1**). To more closely examine these unexpected results, we investigated the haplotypes involving the two variants rs4072037 and rs760077. In brief, both haplotypes that carried the protective allele rs760077[A] were protective and both haplotypes that had the risk allele rs760077[T] conferred risk of gastric cancer; note that the allele rs4072037[C] occurred both on a protective and a risk haplotype (**Supplementary Table 2**).

The gastric cancer risk associated with the *MUC1* locus has been linked to the number of tandem repeats present in exon 2 of the gene; specifically, low numbers of repeats associate with gastric cancer risk, whereas higher numbers confer protection[20]. Exploiting the whole-genome sequencing data to assess this association, we considered a proxy of the tandem repeat in *MUC1* on the basis of the normalized average sequencing read coverage over its region (based on Build 36 and $n = 2,636$ sequenced individuals), where lower read coverage is expected to correlate with short alleles of the tandem repeat (see below for the method used to calculate the proxies). In a full linear model, we observed a significant association of the marginal risk alleles rs760077[T] and rs4072037[T] with lower and higher read coverage (shorter and longer tandem repeat lengths), respectively (rs760077[T]: $P = 1.5 \times 10^{-5}$, $\beta = -0.25$ s.d.; rs4072037[T]: $P = 0.0077$, $\beta = 0.15$ s.d.). The result for rs760077 was supported by Southern blot analysis for a small set of randomly chosen individuals ($n = 18$), for whom a strong correlation ($\rho = 0.85$, $P = 8 \times 10^{-6}$) was demonstrated between the risk allele rs760077[T] and shorter alleles of the tandem repeat in *MUC1*.

The above analyses show that the gastric cancer signal in the *MUC1* locus cannot be explained by rs4072037 alone and that in Europeans the variant rs760077, along with rs4072037, better represents the signal. Our data are also consistent with the role of the tandem repeat in *MUC1* in gastric cancer pathogenicity.

Proxy for *MUC1* tandem repeat genotype based on whole-genome sequencing data. We used sequencing coverage to evaluate the correlation between the reported tandem repeat in *MUC1* and the variants rs4072037 and rs760077. For each of the 2,636 whole genome–sequenced individuals, we found the average sequencing depth over the region chr. 1: 153,427,725–153,428,337 (hg18); this region covers the *MUC1* tandem repeat in the Build 36 reference genome. The average sequencing depth over the region was then normalized by dividing it by the total number of reads mapped over the whole genome for the individual. To evaluate correlation between the tandem repeat proxy and the risk alleles of the variants rs4072037 and rs760077, we performed linear regression using the normalized average sequencing depth (in standardized units) as a response variable and imputed genotypes for rs4072037 and rs760077 as explanatory variables.

Analysis of the effect of the collapsed *ATM* loss-of-function variant on age at diagnosis of five cancers. We tested the effect of the collapsed loss-of-function variant on the age at diagnosis for the five most significantly associated cancers: gastric, pancreatic, prostate, breast and colorectal cancers. A linear model assuming an additive effect of the collapsed *ATM* loss-of-function variant on age at diagnosis indicated that individuals with cancer who carried loss-of-function mutations in *ATM* were diagnosed with the cancer at a considerably earlier age than non-carriers; the effects (β) ranged from −1.7 to −7.3 years, with the association for gastric cancer being most significant ($P = 0.0078$, $\beta = -6.1$ years) (**Supplementary Table 2**). We note that, for breast cancer, the association for the linear model was not significant. Upon inspection of the distribution of the ages at diagnosis of chip-typed breast cancer cases, comparing patients carrying loss-of-function mutations in *ATM* to non-carriers, we noticed a nonlinear effect of genotype on age at diagnosis (**Supplementary Figs. 5–10**). In particular, we noticed that the change in the distribution was not a simple shift in the case of breast cancer. To adjust for such nonlinear effects of the collapsed *ATM* loss-of-function variant on age at diagnosis, we split each of the five cancer groups into sets of patients with age at diagnosis lower and higher than the first quartile of the age at diagnosis for the cancer group. The first quartile for age at onset was comparable for all five cancers except for breast cancer, where it was considerably lower (breast cancer, 50 years; other four cancers, 62–66 years). We then tested the collapsed loss-of-function variant separately for each of the ten resulting subgroups against groups of controls. Consistent with the results for the linear model above, the individuals with cancer who carried loss-of-function variants in *ATM* appeared to begin developing the disease at a much earlier age than the average cancer patient (**Supplementary Table 10**).

Analysis of coding variants at the *PSCA* locus. To further examine the *PSCA* locus, we restricted our scan to coding variants that are predicted to affect protein structure (loss-of-function, missense and splice region variants). In total, we observed 31 such variants within the 500-kb window centered on *PSCA* (**Supplementary Table 13**). A stop-gain variant, rs138377917[A], in *PSCA* ($P = 6.1 \times 10^{-4}$, OR = 0.65) was the only variant passing a Bonferroni significance threshold for an association scan limited to these coding variants. In 1000 Genomes Project Phase 1 data, the protective A allele for the stop-gain variant rs138377917 has a frequency of 3% in Europeans but is absent from Asians (according to Ensembl release 76; see URLs). The gastric cancer association for rs138377917 was not fully explained by the reported variant rs2294008 ($P = 0.011$, OR = 0.72 for rs138377917[A] when adjusting for rs2294008). According to the GTEx database[36], the protective allele of the reported variant rs2294008 associates with decreased *PSCA* mRNA expression in stomach (see URLs; **Supplementary Fig. 14**); this is concordant with the protective effect of the stop-gain variant rs138377917[A].

Refinement of a gastric cancer association signal at the *PRKAA1-PTGER4* locus. The variant rs13361707[C] in *PRKAA1*, which was previously reported in the Han Chinese population[8], associated with gastric cancer in the Icelandic data ($P = 2.7 \times 10^{-4}$, OR = 1.16). There are 33 proxy variants for rs13361707 identified on the basis of the CHB and JPT panel in 1000 Genomes Project Pilot 1 data (variants with correlation $r^2 > 0.8$ with rs13361707 and <500 kb away; see URLs). Of these, the variant that associated most significantly with gastric cancer in Iceland was an intronic variant in *PTGER4* (rs10036575[C], $P = 4.8 \times 10^{-6}$, OR = 0.81), which is also correlated with the reported variant rs13361707 in Iceland ($r^2 = 0.79$; in the CHB an JPT panel, $r^2 = 0.93$). rs10036575[C] was among the strongest gastric cancer variants in this region and had similar effect in diffuse- and intestinal-type gastric cancers (**Supplementary Table 8**). The gastric cancer association of the reported variant rs13361707 did not remain significant when adjusting for rs10036575 ($P = 0.12$, OR = 1.16), whereas rs10036575 remained significant after adjustment for rs13361707 ($P = 2.8 \times 10^{-4}$, OR = 0.68). Hence, the variant rs10036575 is a refinement in Europeans of the signal previously reported with rs13361707 (see **Supplementary Table 14** for variants in the locus correlated with rs10036575). No variants remained significant in the locus after adjusting for rs10036575 (**Supplementary Fig. 12**), and none of the 37 coding variants observed within a 500-kb window centered on *PRKAA1* was significant after correcting for multiple testing ($P > 0.05/37 = 0.0014$).

Four reported gastric cancer susceptibility loci that do not replicate in Iceland. None of the reported SNPs at the remaining four gastric cancer susceptibility loci, *ZBTB20* (3q13.31), *PLCE1* (10q23), *LRFN2* (6p21.1) and *DNAH11* (7p15.3), replicated in the Icelandic population. Because the gastric cancer GWAS reporting these loci were based on populations of non-European ancestry, we performed an ancestry shift refinement[37] by also checking the results at linkage disequilibrium proxies of reported variants, as we did in the case of *PRKAA1*; the proxy variants were based on the CHB and JPT panel in 1000 Genomes Project Pilot 1 data ($r^2 > 0.8$ and <500 kb from the reported variant; see URLs). None of the reported variants had proxies that were

doi:10.1038/ng.3342

significant after accounting for the number of proxies tested, and no biallelic variants were significant across each locus in the Icelandic data in a 500-kb window centered on the reported SNP.

29. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).
30. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
31. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
32. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
33. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
34. Flicek, P. *et al.* Ensembl 2012. *Nucleic Acids Res.* **40**, D84–D90 (2012).
35. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
36. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
37. Stacey, S.N. *et al.* Ancestry-shift refinement mapping of the *C6orf97-ESR1* breast cancer susceptibility locus. *PLoS Genet.* **6**, e1001029 (2010).