Supplementary Information for:

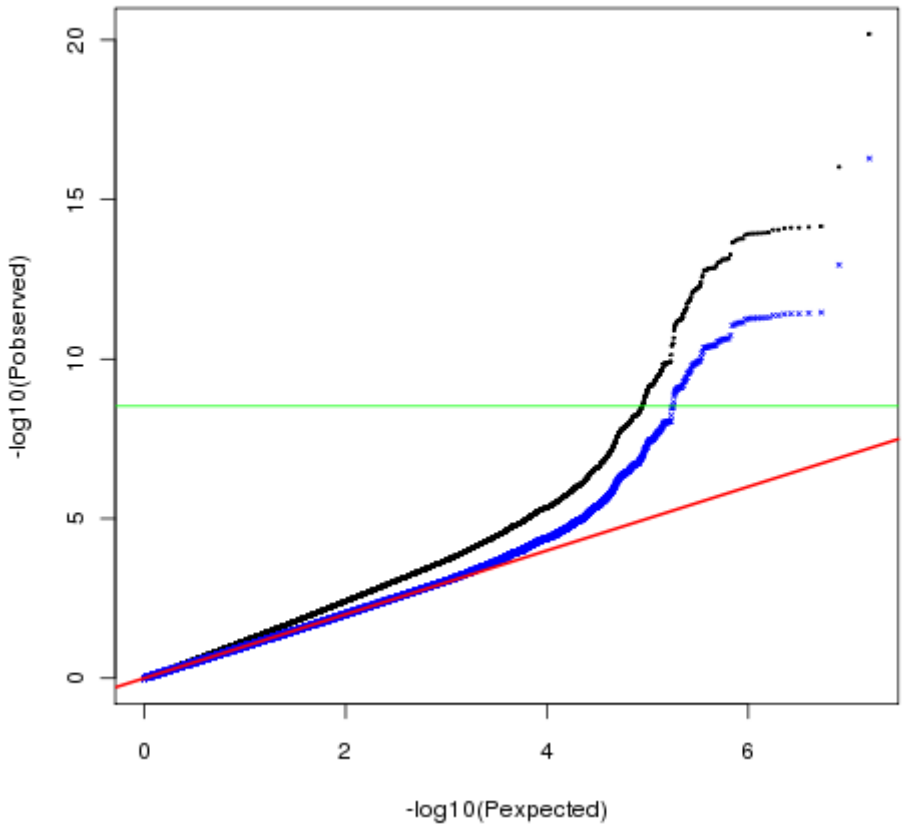
A Germline Variant in the *TP53* Polyadenylation Signal Confers Cancer Susceptibility.

S.N. Stacey et al.

Contents:

Nature Genetics: doi:10.1038/ng.926

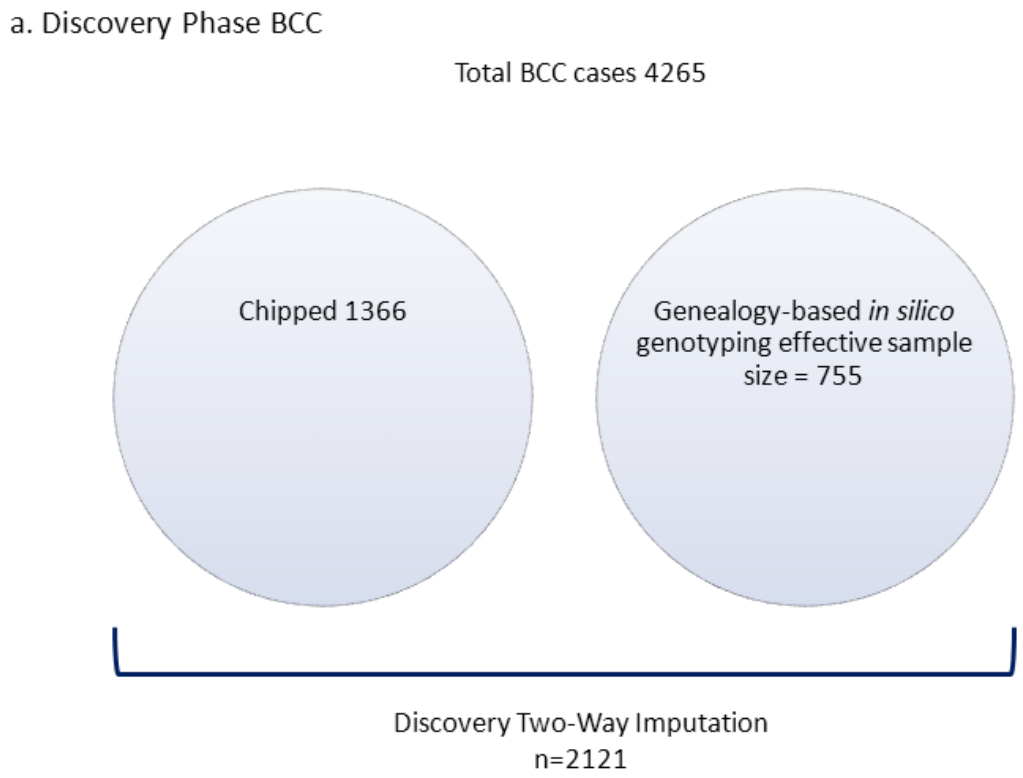
Supplementary Figure 1: Q-Q Plot



Legend: Q-Q plot for association of 16 million SNPs with BCC, based on the Icelandic Discovery Two-Way Imputation data. Unadjusted P values are shown in black and P values adjusted using genomic control are shown in blue (λ=1.25). The expected (diagonal) values are indicated by a red line. The green line indicates the threshold for genome-wide significance based on Bonferroni correction for the 16 million SNPs tested ($P = 3.0 \times 10^{-9}$). All SNPs exceeding this threshold mapped to the chromosome 17 locus discussed in this report, or to previously published loci: 1p36 (*PADI6*), strongest signal $P = 3.5 \times 10^{-12}$; 1q42 (*RHOU*), strongest signal $P = 9.0 \times 10^{-11}$; and 5p15 (*TERT-CLPTM1L*), strongest signal $P = 9.0 \times 10^{-12}$ [Stacey et al., Nat Genet. 2008 Nov;40(11):1313-8; Rafnar et al., Nat Genet. 2009 Feb;41(2):221-7]. In order to roughly estimate the numbers of other loci detected at sub-genome wide significant signal levels, we divided the genome into 1Mb bins and determined how many such bins contained a SNP that exceeded a particular significance threshold. We observed that 475 bins contained SNPs with $P < 10^{-4}$ and 74 bins contained SNPs with $P < 10^{-5}$.

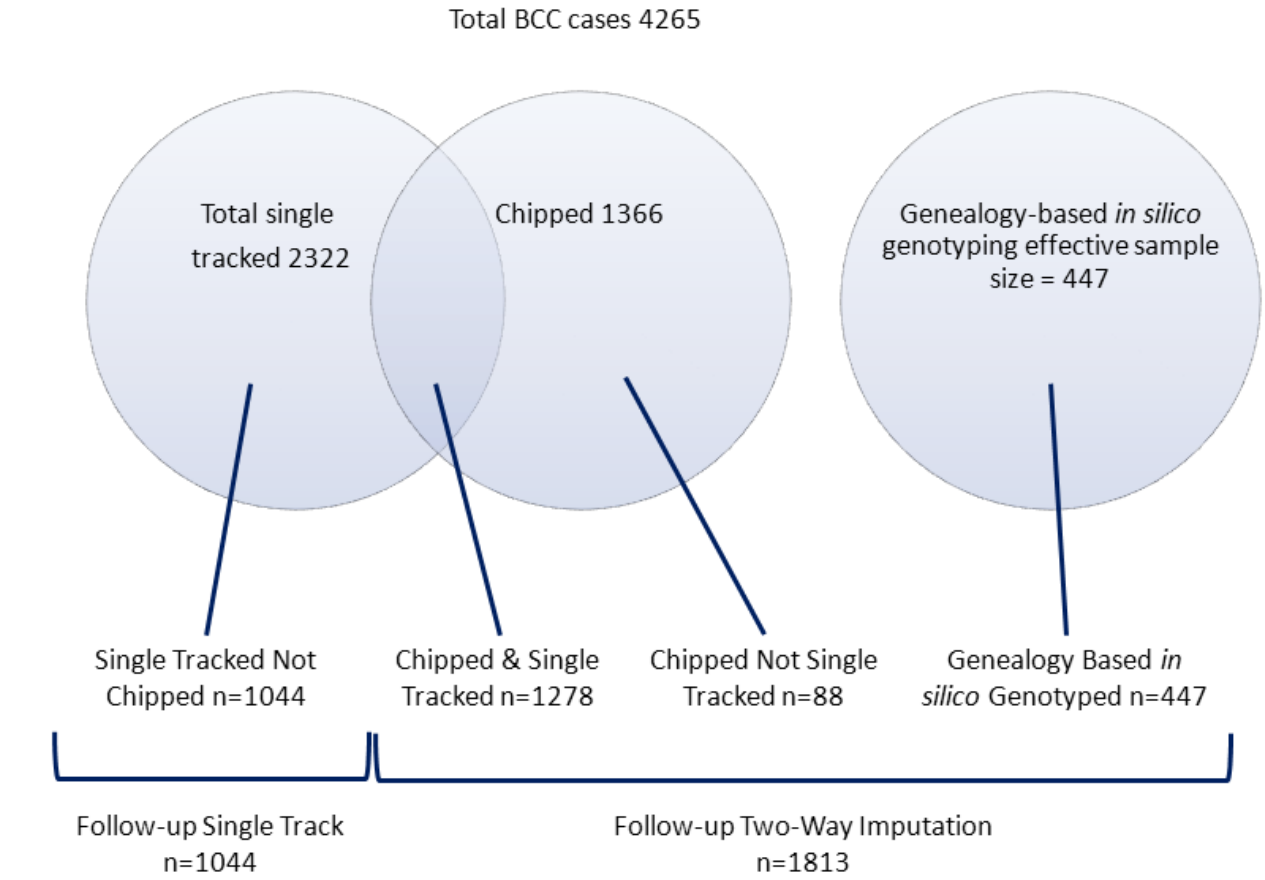
Nature Genetics: doi:10.1038/ng.926

Supplementary Figure 2: Venn diagram of Discovery Phase and Follow-up Phase Icelandic samples for BCC



Nature Genetics: doi:10.1038/ng.926

b. Follow-up Phase BCC

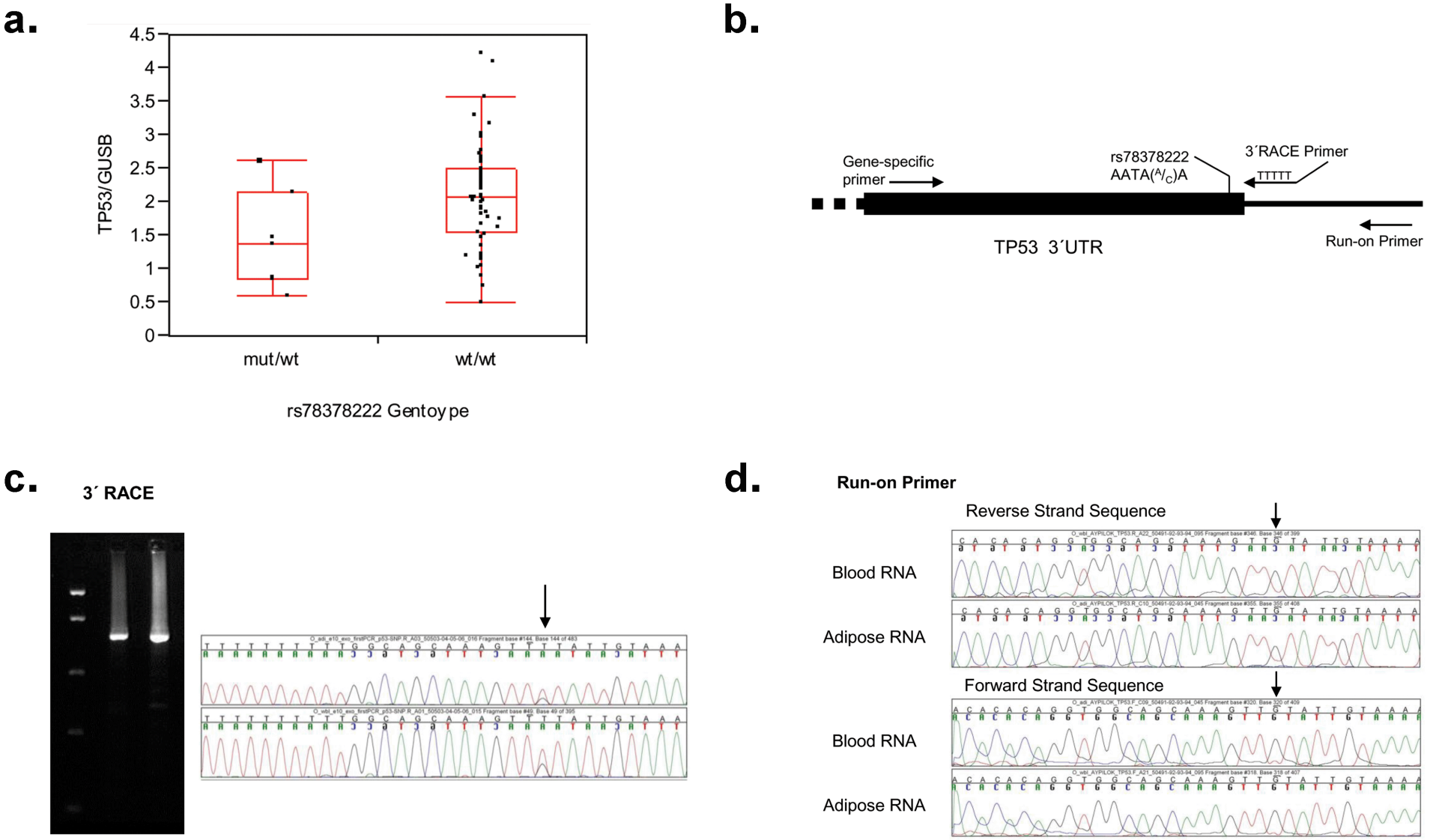


Nature Genetics: doi:10.1038/ng.926

Supplementary Figure 2 Legend: A total of 4265 patients with BCC were ascertained from the Icelandic Cancer Registry. Of these 1366 were genotyped on Illumina SNP array platforms ("chipped"). In the Discovery Phase (panel a) SNP genotypes were determined by imputation for the 1366 cases who had been chip typed or by genealogy-based *in silico* genotyping for a further 755 cases (effective sample size, ESS, estimates are given for *in silico* genotyped sets). The combined sample that was genotyped by two-way imputation and used in the association analysis was 2121 cases (ESS). In the Follow-up Phase (panel b), all available cases (n = 2322) were genotyped for rs78378222 by Centaurus single-track assay. Of these, 1044 had not been typed on the Illumina SNP arrays and 1278 had. Genealogy-based *in silico* genotyping was used to obtain genotypes for a further 447 individuals (ESS). Follow-up Phase two-way imputation and association analysis was carried out using all of the individuals who had been chip typed or who had been genotyped by genealogy-based *in silico* genotyping, yielding a total sample of 1813 cases (ESS). The Follow-up Phase single-track genotyping and association analysis was based on individuals who were directly genotyped by Centaurus assay and had not been chip typed (n=1044). Two different aspects of validation are examined in the Follow-up Phase: Biological validation (*i.e.* a further investigation of whether the allele frequencies differ between cases and controls) is examined by the 1044 single-track genotyped cases compared with single-track genotyped controls, as shown in *Table 1*. (Note that this is not a fully independent replication of the Discovery Phase result since some of the individuals who were assigned *in silico* genotypes in the Discovery Phase were subsequently single-track genotyped in the Follow-up Phase). Technical validation refers to how well the imputed genotypes match the genotypes determined by direct single-track genotyping. This was assessed by examining rs78378222 genotypes for the 1278 cases (along with a further 8413 controls) who had been typed on both Illumina SNP array and Centaurus single-track platforms. The r^2 between the results of the two methods was 0.92. The process of defining Discovery Phase and Follow-up Phases for all other disease groups studied was carried out in an analogous manner to that described here for BCC. The numbers of cases in each group are detailed in *Supplementary Table 4*.

Nature Genetics: doi:10.1038/ng.926

Supplementary Figure 3: The rs78378222 mutation impairs correct 3´ end processing of TP53



Nature Genetics: doi:10.1038/ng.926

Supplementary Figure 3 Legend: (a) Tukey box plot of quantitative RT-PCR of *TP53* RNA abundance in peripheral blood samples. RNA levels were quantitated by RT-PCR and normalized against expression of *GUSB*. Normalized expression levels (y-axis) are plotted separately for each genotype (x-axis). The central horizontal line indicates the median of each distribution, upper and lower boundaries of the boxes indicate the 25^{th} and 75^{th} percentiles and the whiskers indicate the 5^{th} and 95^{th} percentiles. N = 7 [C/A] heterozygotes and 51 [A/A] homozygotes. **(b)** Schematic diagram (not to scale) showing the locations of primers used for investigation of termination and polyadenylation of *TP53* rs78378222 mutant and wild-type alleles. For the 3´ RACE, a gene specific primer was paired with the 3´ RACE primer in order to amplify specifically RNA species that were correctly polyandenylated. For the Run-on experiments, a gene specific primer was paired with the Run-on primer in order to amplify specifically RNA species that failed to terminate and polyadenylate correctly. **(c)** 3´ RACE of RNA samples from peripheral blood and adipose tissue from rs78378222 heterozygotes produced 1300bp bands as expected for correctly terminated and polyadenylated mRNA. Sequencing of the 3´ RACE products showed a predominance of mRNAs bearing the wild type [A] allele and a reduced abundance of the mutant [C] allele for rs78378222 (arrowed). Data from 6 replicate samples yielded an average ratio of 73% wild-type to 27% mutant transcripts ($P = 1.6 \times 10^{-6}$). Sequence traces of the reverse strand (and in genomic orientation, with the TP53 transcript running right to left) are shown for blood (upper) and adipose (lower) tissue-derived mRNA. **(d)** Sequence analysis of run-on transcription from an rs78378222 heterozygote. RT-PCR was conducted on blood and adipose-derived RNA using gene-specific and Run-on primers. Sequence analysis of the products on both forward and reverse strands shows that the mutant [C] allele of rs78378222 (arrowed) is predominant on run-on transcription products. Sequence traces are shown in genomic orientation with the TP53 transcript running right to left.

Nature Genetics: doi:10.1038/ng.926

Supplementary Table 1: Association data for rs78378222[C] derived from all individuals who were genotyped by Centaurus single-track method[a]

Tumour Type	P-value	OR	95% Confidence Interval	Number of Cases	Frequency in Cases	Number of Controls	Frequency in Controls
Basal Cell Carcinoma	4.74×10^{-11}	2.11	(1.69, 2.64)	2322	0.0383	7200	0.0185
Colorectal Adenoma	0.0031	1.42	(1.13, 1.79)	2396	0.0261	7200	0.0185
Prostate Cancer	0.0096	1.36	(1.08, 1.72)	2445	0.0249	7200	0.0185
Glioma	0.1460	1.81	(0.81, 4.03)	121	0.0331	7200	0.0185

[a]This set differs from the Iceland Follow-up Phase Single-Track Genotyped set in that it includes all cases who were single-track genotyped, without regard to whether or not they were also typed by Illumina chip. See *Supplementary Figure 2.*

Nature Genetics: doi:10.1038/ng.926

Supplementary Table 2: Overview of the replication sample sets used in this study

Sample Set	Disease/Phenotype	Cases[a]	Controls[a]	Type	Location	Reference
Denmark	Basal Cell Carcinoma	308	3,441	Nested Case Control Study from EPIC with supplementary population-based controls from Inter99 cohort	Copenhagen, Denmark	a,b
Eastern Europe	Basal Cell Carcinoma	528	533	Multi-center Hospital-based Case: Population-based control	Hungary, Romania, Slovakia	c
Spain	Basal Cell Carcinoma	628	3,928	Multi-center Hospital-based Case: Population-based control	Zaragoza & Valencia, Spain	d
Netherlands	Prostate Cancer	1,085	1,796	Registry Ascertained Case: Population-based Control	Eastern Netherlands	e
Romania	Prostate Cancer	639	815	Hospital-Based Case:Population Based Control	Bucharest, Romania	e
Spain	Prostate Cancer	785	1,787	Hospital-Based Case:Population Based Control	Zaragoza, Spain	e
U.K.	Prostate Cancer	521	1,407	PSA based testing and treatment trial	Nine locations in the U.K.	e
U.S.A.	Prostate Cancer	1,454	1,293	Hospital-Based Case:Control	Chicago, U.S.A.	e
U.S.A. UCSF	Glioma	658	573	Multi-center Hospital-based an Population-based Case: Population-based control	San Francisco Bay Area, U.S.A.	f,g
U.S.A. Mayo Clinic	Glioma	530	283	Hospital-Based Case:Control	Minnesota, U.S.A.	f,g
Netherlands	Colorectal Cancer	464	1,796	Hospital-Based Case:Population Based Control	Eastern Netherlands	h
Spain	Colorectal Cancer	184	1,940	Hospital-Based Case:Population Based Control	Zaragoza, Spain	i
Sweden	Colorectal Cancer	1,781	1,737	Hospital-Based Case:Population Based Control	Stockholm, Sweden	j
US	Colon Cancer	475	807	Hospital-Based Case:Population Based Control	North Carolina, U.S.A.	k
US	Rectal Cancer	942	922	Hospital-Based Case:Population Based Control	North Carolina, U.S.A.	k
Netherlands	Breast Cancer	725	1,796	Registry Ascertained Case:Control	Eastern Netherlands	l
Spain	Breast Cancer	1,007	1,940	Hospital-Based Case:Population Based Control	Zaragoza, Spain	l
Netherlands	Melanoma	683	1,796	Registry Ascertained Case:Control	Eastern Netherlands	d
Spain Valencia	Melanoma	823	1,988	Hospital-Based Case:Population Based Control	Valencia, Spain	d

Nature Genetics: doi:10.1038/ng.926

Spain Zaragoza	Melanoma	290	1,787	Hospital-Based Case:Population Based Control	Zaragoza, Spain	d
Iceland Pigmentation	Pigmentation Traits	9,805	NA	Population-Based Self Reported Questionnaire	Nationwide, Iceland	m
Netherlands	Pigmentation Traits	1,326	NA	Population-Based Self Reported Questionnaire	Eastern Netherlands	m

[a]Numbers successfully genotyped for rs78378222 are given

References:
(a) Vogel U. et al. Mutat Res 617, 138-46 (2007)
(b) Glumer C. et al. Diabetes Care 26, 2335-40 (2003)
(c) Scherer, D et al. Int. J. Cancer 122, 1787-1793 (2008)
(d) Stacey, S.N. et al., Nat Genet 41, 909-14 (2009)
(e) Gudmundsson, J. et al. Sci Transl Med 2, 62ra92 (2010)
(f) Wrensch M. et al., Nat Genet 41, 905-8 (2009)
(g) Jenkins R.B., et al., Cancer Genetics 204, 13-18 (2011)
(h) van der Logt, E.M.J. et al. Carcinogenesis 25, 2407-15 (2004)
(i) Rafnar, T et al. Nat Genet 41, 221-7 (2009)
(j) Ghazi S. et al. Am J Pathol 177, 2688-93 (2010)
(k) Satia J.A. et al. Cancer Epidemiol Biomarkers Prev 14, 429-36 (2005)
(l) Stacey, S.N. et al. Nat Genet 39, 865-9 (2007)
(m) Sulem, P. et al. Nat Genet 39, 1443-52 (2007)

Nature Genetics: doi:10.1038/ng.926

Supplementary Table 3: Discovery Phase Two-Way Imputation association results for rs78378222[C] with BCC, colorectal adenoma, prostate cancer and brain cancers[a]

Tumour Type	P-value	OR	95% Confidence Interval	Number of Cases	Frequency in Cases	Number of Controls	Frequency in Controls
Basal Cell Carcinoma	5.2×10^{-17}	2.36	(1.93, 2.89)	2121[b]	0.0442	>39,614[b]	0.0192[c]
Colorectal Adenoma	3.3×10^{-4}	1.62	(1.24, 2.11)	3776[b]	0.0306	>37,417[b]	0.0192[c]
Prostate Cancer	0.0016	1.39	(1.13, 1.71)	2708[b]	0.0265	>39,060[b]	0.0192[c]
All Brain Cancers[d]	0.0018	2.18	(1.34, 3.56)	327[b]	0.0409	>20,824[b]	0.0192[c]

[a]Results are shown for tumour types that yielded significant associations after Bonferroni adjustment for the 20 types of tumour tested.[b]Effective sample size estimate taking into account efficiency of Icelandic genealogy-based *in silico* genotyping.[c]The control frequency given is derived from 40,309 directly chip-typed non-BCC control individuals.[d]ICD-10 codes C70-C72

Nature Genetics: doi:10.1038/ng.926

Supplementary Table 4: Numbers of samples analysed in Icelandic Discovery and Follow-up Phases

Phase:	General			Discovery Phase			
	B	C	D	E	F	G	H
Phenotype	Total Number of Cases Ascertained[a]	Cases Chipped[b]	Cases Not Chipped (B-C)	Cases Not Chipped but with a FSDR[c] on Chip	Cases Genealogy-Based *in silico* Genotyped ESS[d]	Cases Total ESS (C+F)	Matched Control ESS[e]
BCC	4265	1366	2899	1986	755	2121	>39,614
Prostate Cancer	5046	1860	3186	2231	848	2708	>39,060
Colorectal Adenoma	6703	2201	4502	3245	1233	3776	>37,417
All Brain Cancers	682	137	545	413	157	327	>20,824
Glioma	425	51	374	274	104	182	>10,353
Colorectal Cancer	3888	1128	2760	1836	698	1950	>40,547
Melanoma	1030	602	428	292	111	724	>41,073
Breast Cancer	5456	2414	3042	1952	742	3253	>39,261
ER Negative Breast Cancer	435	368	67	45	17	385	>41,216
High Risk Breast Cancer[f]	1739	875	864	563	220	1095	>40,250

Phase:	Follow-up Phase							
	I	J	K	L	M	N	O	P
Phenotype	Total Number of Cases Single Track Genotyped	Cases Single Tracked but not Chipped	Cases Single Tracked and Chipped	Cases neither Chipped nor Single Tracked	Cases neither Chipped nor Single Tracked but with a FSDR on Chip	Cases Genealogy-Based *in silico* Genotyped ESS	2-Way Imputation Cases Total ESS (C+N)	2-Way Imputation Matched Control ESS
BCC	2322	1044	1278	1855	1115	447	1813	>36,709
Prostate Cancer	2445	635	1810	2550	1749	811	2671	>36,331
Colorectal Adenoma	2396	1038	1358	2106	2237	856	3057	>36,022
Glioma	121	72	49	302	207	84	135	>37,881

[a]Case ascertainment was through the Icelandic Cancer Registry or National Pathology Department registers. [b]"Chipped" means that the samples were genotyped with Illumina Human Hap300, HapCNV370, Hap610, 1M or Omni-1 Quad bead chips. [c]"FSDR" means first or second degree relative. [d]"ESS" means effective sample size estimate. [e]The matched control set (see *Online Methods*) was drawn from a total number of 437,218 population based controls, of whom 40,309 were chip typed and did not have a recorded BCC diagnosis. [f]Probands with breast cancer diagnosed under 50 years of age or a record of multiple independent primary breast cancers.

Nature Genetics: doi:10.1038/ng.926

Supplementary Table 5 : Association between rs78378222[C] and colorectal cancer, breast cancer and melanoma

Sample Set	Tumour Type	P-value	OR	95% Confidence Interval	Number of Cases	Frequency in Cases	Number of Controls	Frequency in Controls	Phet
Iceland Discovery Phase Two-Way Imputation	Colorectal Cancer	0.31	1.14	(0.89, 1.47)	1950[a]	0.0218	>40,547[a]	0.0192[b]	
Netherlands	Colorectal Cancer	0.87	0.95	(0.53,1.72)	464	0.0140	1796	0.0148	
Spain	Colorectal Cancer	0.052	3.54	(1.00,12.53)	184	0.0109	1940	0.0031	
Sweden	Colorectal Cancer	0.56	0.90	(0.63,1.27)	1781	0.0171	1737	0.0190	
US	Colon Cancer	0.20	1.95	(0.70,5.40)	475	0.0084	807	0.0043	
US	Rectal Cancer	0.50	0.81	(0.44,1.49)	942	0.0106	922	0.0130	
Combined Icelandic Discovery and Replication	**Colorectal Cancer**	**0.51**	**1.06**	**(0.89, 1.27)**	**5796[a]**	**NA**	**>47,749[a]**	**NA**	**0.23**
Iceland Discovery Phase Two-Way Imputation	Breast Cancer	0.95	0.99	(0.80, 1.23)	3253[a]	0.0191	>39,261[a]	0.0192[b]	
Netherlands	Breast Cancer	0.33	1.27	(0.79, 2.03)	725	0.0186	1794	0.0148	
Spain	Breast Cancer	0.27	1.61	(0.69, 3.77)	1007	0.0050	1940	0.0031	
Combined Icelandic Discovery and Replication	**Breast Cancer**	**0.57**	**1.06**	**(0.88, 1.27)**	**4985[a]**	**NA**	**>42,995[a]**	**NA**	**0.40**
Iceland Discovery Phase Two-Way Imputation	Melanoma	0.61	0.90	(0.60, 1.35)	724[a]	0.0173	>41,073[a]	0.0192[b]	
Netherlands	Melanoma	0.042	1.60	(1.02, 2.52)	683	0.0234	1796	0.0148	
Spain Valencia	Melanoma	0.25	0.60	(0.25, 1.43)	823	0.0036	1988	0.0060	
Spain Zaragoza	Melanoma	0.62	0.62	(0.09, 4.10)	290	0.0017	1787	0.0028	
Combined Icelandic Discovery and Replication	**Melanoma**	**0.64**	**1.07**	**(0.81, 1.42)**	**2520[a]**	**NA**	**>46,644[a]**	**NA**	**0.12**

[a]Effective sample size taking into account efficiency of Icelandic genealogy-based *in silico* genotyping. [b]The control frequency given is derived from 40,309 directly chip-typed non-BCC control individuals.

Nature Genetics: doi:10.1038/ng.926

Supplementary Table 6 : Specification of Nucleic Acid Sequences

Description	Sequence
Sequence context of novel SNP chr17:7640788	cttcctgcccacgcccaccaagatgcattacctcttcaaccttcgagacatctccaaggtgactcgcggcctgaccttgccccttctgcttggcccagcctccg cggaggctttctcttctcaaactaagccttaacactcactagcatg[C/T]gcaccaaaagtcacccccatgctgaagtgccacactccctggccttaccttta aaacttctgggccaagtgcggtggctcacacctgtaattccagcactttgggaggccaacgcaggcagatcacctgaggttaggagttcaagaccag
TP53 3´ RACE Gene-specific primer	GAATGAGGCCTTGGAACTCAAGGAT
TP53 Sequencing primer	TTCCCCTCCTTCTCCCTTTTT
Run-on primer	TCCCGTAATCCTTGGTGAGA
TP53 Internal primer for Run-on experiments	TGCAAGCACATCTGCATTTT

Nature Genetics: doi:10.1038/ng.926

Supplementary Note: Methods for Genotype Imputation

Long range phasing: Long range phasing of all chip-genotyped individuals was performed with methods described previously[1,2]. In brief, phasing is achieved using an iterative algorithm which 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 genome-wide for all chip-typed Icelanders. For long range phased haplotype association analysis, we then partitioned the genome into non-overlapping fixed 0.3cM bins. Within each bin, we observed the haplotype diversity described by the combination of all chip-typed markers in the bin. Haplotypes with frequencies over 0.001 were tested in a case : control analysis.

Genotype imputation: We imputed the SNPs identified and genotyped through sequencing into all Icelanders who had been phased with long range phasing using the same model as used by IMPUTE[1]. The genotype data from sequencing can be ambiguous due to low sequencing coverage. In order to phase the sequencing genotypes, an iterative algorithm was applied for each SNP with alleles 0 and 1. We let H be the long range phased haplotypes of the sequenced individuals and applied the following algorithm:

1. For each haplotype h in H, use the Hidden Markov Model of IMPUTE to calculate for every other k in H, the likelihood, denoted $\gamma_{h,k}$, of h having the same ancestral source as k at the SNP.

2. For every h in H, initialize the parameter θ_h, which specifies how likely the one allele of the SNP is to occur on the background of h from the genotype likelihoods obtained from sequencing. The genotype likelihood L_g is the probability of the observed sequencing data at the SNP for a given individual assuming g is the true genotype at the SNP. If L_0, L_1 and L_2 are the likelihoods of the genotypes 0, 1 and 2 in the individual that carries h, then set $\theta_h = \frac{L_2+\frac{1}{2}L_1}{L_2+L_1+L_0}$.

3. For every pair of haplotypes h and k in H that are carried by the same individual, use the other haplotypes in H to predict the genotype of the SNP on the backgrounds of h and k: $\tau_h = \sum_{l\in H\setminus\{h\}} \gamma_{h,l}\theta_l$ and $\tau_k = \sum_{l\in H\setminus\{k\}} \gamma_{k,l}\theta_l$. Combining these predictions with the genotype likelihoods from sequencing gives un-normalized updated phased genotype probabilities: $P_{00} = (1-\tau_h)(1-\tau_k)L_0$, $P_{10} = \tau_h(1-\tau_k)\frac{1}{2}L_1$, $P_{01} = (1-\tau_h)\tau_k\frac{1}{2}L_1$ and $P_{11} = \tau_h\tau_k L_2$. Now use these values to update θ_h and θ_k to $\theta_h = \frac{P_{10}+P_{11}}{P_{00}+P_{01}+P_{10}+P_{11}}$ and $\theta_k = \frac{P_{01}+P_{11}}{P_{00}+P_{01}+P_{10}+P_{11}}$.

4. Repeat step 3 when the maximum difference between iterations is greater than a convergence threshold ε. We used $\varepsilon=10^{-7}$.

Given the long range phased haplotypes and θ, the allele of the SNP on a new haplotype h not in H, is imputed as $\sum_{l\in H} \gamma_{h,l}\theta_l$.

Nature Genetics: doi:10.1038/ng.926

The above algorithm can easily be extended to handle simple family structures such as parent-offspring pairs and triads by letting the *P* distribution run over all founder haplotypes in the family structure. The algorithm also extends trivially to the X-chromosome. If source genotype data are only ambiguous in phase, such as chip genotype data, then the algorithm is still applied, but all but one of the *L*s will be 0. In some instances, the reference set was intentionally enriched for carriers of the minor allele of a rare SNP in order to improve imputation accuracy. In this case, expected allele counts will be biased toward the minor allele of the SNP. Call the enrichment of the minor allele *E* and let θ' be the expected minor allele count calculated from the naïve imputation method, and let θ be the unbiased expected allele count, then $\theta' = \frac{E\theta}{1-\theta+E\theta}$ and hence $\theta = \frac{\theta'}{E+(1-E)\theta'}$.

This adjustment was applied to all imputations based on enriched imputations sets. We note that if θ' is 0 or 1, then θ will also be 0 or 1, respectively.

Using a sample of 9691 individuals who had been typed both on chip and by direct genotyping for rs78378222, we compared the imputed genotype expectation values with direct single track genotypes. The r^2 between the results of the two methods was 0.92.

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

$$\frac{Var(E\ \theta\ chip\ data\ )}{Var(\theta)},$$

where $\theta \in \{0, 1\}$ is the allele count. $Var(E\ \theta\ chip\ data\ )$ was estimated by the observed variance of the imputed expected counts and $Var(\theta)$ was estimated by $p(1-p)$, where p is the allele frequency. 78.2 % of SNPs were imputed with information values $\geq$ 0.8 and a further 16.6% were imputed with information values $\geq$ 0.6 and < 0.8. Thus 97.4% of SNPs were imputed with information values $\geq$ 0.6. The information value for rs78378222 was 0.97.

Genealogy-based in silico genotyping: In addition to imputing sequence variants from the whole genome sequencing effort into chip genotyped individuals, we also performed a second imputation step where genotypes were imputed into relatives of chip genotyped individuals, creating *in silico* genotypes. The inputs into the second imputation step are the fully phased (in particular every allele has been assigned its parent of origin[3]) imputed and chip type genotypes of the available chip typed individual. The algorithm used to perform the second imputation step consists of:

1. For each ungenotyped individual (the proband), find all chip genotyped individuals within two meioses of the individual. The six possible types of two meiotic distance relatives of the proband are (ignoring more complicated relationships due to pedigree loops): Parents, full and half siblings, grandparents, children and grandchildren. If all pedigree paths from the proband to a genotyped relative go through other genotyped relatives, then that relative is excluded. E.g. if a parent of the proband is genotyped, then the proband's grandparents through that parent are excluded. If the number of meiosis in the pedigree around the proband exceeds a threshold (we used 12), then

Nature Genetics: doi:10.1038/ng.926

relatives are removed from the pedigree until the number of meiosis falls below 12, in order to reduce computational complexity.

2. At every point in the genome, calculate the probability of each genotyped relative sharing with the proband based on the autosomal SNPs used for phasing. A multipoint algorithm based on the hidden Markov model Lander-Green multipoint linkage algorithm using fast Fourier transforms is used to calculate these sharing probabilities[4,5]. First single point sharing probabilities are calculated by dividing the genome into 0.5cM bins and using the haplotypes over these bins as alleles. Haplotypes that are the same, except at most at a single SNP, are treated as identical. When the haplotypes in the pedigree are incompatible over a bin, then a uniform probability distribution was used for that bin. The most common causes for such incompatibilities are recombinations within the pedigree, phasing errors and genotyping errors. Note that since the input genotypes are fully phased, the single point information is substantially more informative than for unphased genotyped, in particular one haplotype of the parent of a genotyped child is always known. The single point distributions are then convolved using the multipoint algorithm to obtain multipoint sharing probabilities at the center of each bin. Genetic distances were obtained from the most recent version of the deCODE genetic map[6].

3. Based on the sharing probabilities at the center of each bin, all the SNPs from the whole genome sequencing are imputed into the proband. To impute the genotype of the paternal allele of a SNP located at x, flanked by bins with centers at x_{left} and x_{right}. Starting with the left bin, going through all possible sharing patterns v, let I_v be the set of haplotypes of genotyped individuals that share identically by descent within the pedigree with the proband's paternal haplotype given the sharing pattern v and $P(v)$ be the probability of v at the left bin – this is the output from step 2 above – and let e_i be the expected allele count of the SNP for haplotype i. Then $e_v = \frac{\sum_{i \in I_v} e_i}{\sum_{i \in I_v} 1}$ is the expected allele count of the paternal haplotype of the proband given v and an overall estimate of the allele count given the sharing distribution at the left bin is obtained from $e_{left} = \sum_v P(v) e_v$. If I_v is empty then no relative shares with the proband's paternal haplotype given v and thus there is no information about the allele count. We therefore store the probability that some genotyped relative shared the proband's paternal haplotype, $O_{left} = \sum_{v, I_v = \emptyset} P(V)$ and an expected allele count, conditional on the proband's paternal haplotype being shared by at least one genotyped relative: $c_{left} = \frac{\sum_{v, I_{v \neq \emptyset}} P(v) e_v}{\sum_{v, I_{v \neq \emptyset}} P(v)}$. In the same way calculate O_{right} and c_{right}. Linear interpolation is then used to get an estimates at the SNP from the two flanking bins:

$$O = O_{left} + \frac{x - x_{left}}{x_{right} - x_{left}} \left(O_{right} - O_{left} \right),$$

$$c = c_{left} + \frac{x - x_{left}}{x_{right} - x_{left}} \left(c_{right} - c_{left} \right).$$

Nature Genetics: doi:10.1038/ng.926

If θ is an estimate of the population frequency of the SNP then $Oc + (1 - O)\theta$ is an estimate of the allele count for the proband's paternal haplotype. Similarly, an expected allele count can be obtained for the proband's maternal haplotype.

Case : control association testing: Logistic regression was used to test for association between SNPs and disease, treating disease status as the response and expected genotype counts from imputation or allele counts from direct genotyping as covariates. Testing was performed using the likelihood ratio statistic. The conditional analysis of rs78378222 and *chr17:7640788* was performed by adding rs78378222 as a covariate while testing *chr17:7640788* for association with BCC. When testing for association using the *in silico* genotypes, controls were matched to cases based on the informativeness of the imputed genotypes, such that for each case C controls of matching informativeness where chosen. Failing to match cases and controls will lead to a highly inflated genomic control factor, and in some cases may lead to spurious false positive findings. The informativeness of each of the imputation of each one of an individual's haplotypes was estimated by taking the average of

$$a\ e,\theta\ = \begin{cases} \dfrac{e-\theta}{1-\theta}, & e \geq \theta \\ \dfrac{\theta-e}{\theta}, & e < \theta \end{cases}$$

over all SNPs imputed for the individual, where e is the expected allele count for the haplotype at the SNP and θ is the population frequency of the SNP. Note that $a\ \theta,\theta\ = 0$ and $a\ 0,\theta\ = a(1,\theta) = 1$. The mean informativeness values cluster into groups corresponding to the most common pedigree configurations used in the imputation, such as imputing from parent into child or from child into parent. Based on this clustering of imputation informativeness we divided the haplotypes of individuals into seven groups of varying informativeness, which created 27 groups of individuals of similar imputation informativeness; 7 groups of individuals with both haplotypes having similar informativeness, 21 groups of individuals with the two haplotypes having different informativeness, minus the one group of individuals with neither haplotype being imputed well. Within each group we calculate the ratio of the number of controls and the number of cases, and choose the largest integer C that was less than this ratio in all the groups. For example, if in one group there are 10.3 times as many controls as cases and if in all other groups this ratio was greater, then we would set $C = 10$ and within each group randomly select ten times as many controls as there are cases. For the different tumour types the value of $C$was always higher than 15.

Inflation Factor Adjustment: In order to account for the relatedness and stratification within the case and control sample sets we applied the method of genomic control based on chip typed markers[7]. The adjustment factors ranged from 1.06 (for PBC) to 1.27 (for Prostate Cancer). Quoted *P* values have been adjusted accordingly.

Effective sample size estimation: In order to estimate the effective sample size of the case control association analyses, we compared the variances of the logistic and generalized linear regression parameter estimates based on the *in silico* genotypes to their one step imputation counterparts. For the quantitative trait association analysis, assume that a single step imputation (SNPs are imputed, but *in silico* genotypes are not used) association analysis with

Nature Genetics: doi:10.1038/ng.926

n_1 subjects leads on average to an estimate of the regression parameter with variance σ_1^2 and that the corresponding *in silico* genotype association analysis leads to an estimate of the regression parameter with variance σ_2^2, then assuming that variance goes down linearly with sample size we estimate the effective sample size in the *in silico* genotype association analysis as $n_2 = \frac{\sigma_1^2}{\sigma_2^2} n_1$. For the case control association analysis, the number of controls is much greater than the number cases and we use the same formula to estimate the effective number of cases, with the n-s representing the number of cases and the σ^2-s representing the variances of the logistic regression coefficient.

References:

1. Kong, A. et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* 40, 1068-75 (2008).
2. Holm, H. et al. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet* 43, 316-20 (2011).
3. Kong, A. et al. Parental origin of sequence variants associated with complex diseases. *Nature* 462, 868-74 (2009).
4. Lander, E.S. & Green, P. Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S A* 84, 2363-7 (1987).
5. Kruglyak, L. & Lander, E.S. Faster multipoint linkage analysis using Fourier transforms. *J Comput Biol* 5, 1-7 (1998).
6. Kong, A. et al. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467, 1099-103 (2010).
7. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* 55, 997-1004 (1999).

Nature Genetics: doi:10.1038/ng.926