nature genetics

Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk

The timing of puberty is a highly polygenic childhood trait that is epidemiologically associated with various adult diseases. Using 1000 Genomes Project–imputed genotype data in up to ~370,000 women, we identify 389 independent signals ($P < 5 \times 10^{-8}$) for age at menarche, a milestone in female pubertal development. In Icelandic data, these signals explain ~7.4% of the population variance in age at menarche, corresponding to ~25% of the estimated heritability. We implicate ~250 genes via coding variation or associated expression, demonstrating significant enrichment in neural tissues. Rare variants near the imprinted genes *MKRN3* and *DLK1* were identified, exhibiting large effects when paternally inherited. Mendelian randomization analyses suggest causal inverse associations, independent of body mass index (BMI), between puberty timing and risks for breast and endometrial cancers in women and prostate cancer in men. In aggregate, our findings highlight the complexity of the genetic regulation of puberty timing and support causal links with cancer susceptibility.

Puberty is the developmental stage of transition from childhood to physical and sexual maturity, and its timing varies markedly between individuals[1]. This variation reflects the influence of genetic, nutritional and other environmental factors and is associated with the subsequent risks for several diseases in adult life[2]. Our previous large-scale genomic studies identified 113 independent regions associated with age at menarche (AAM), a notable milestone of puberty in females[3,4]. The vast majority of those signals have concordant effects on the age at voice breaking (genome-wide genetic correlation between traits $r_g = 0.74$), a corresponding milestone in males[5]. Those genetic findings implicated a diverse range of mechanisms in the regulation of puberty timing, identified significant enrichment of AAM-associated variants in/near genes disrupted in rare disorders of puberty, and highlighted shared etiological factors between puberty timing and metabolic disease outcomes[2,3].

However, those previous studies were based on genome-wide association data that were imputed to the relatively sparse HapMap 2 reference panel or used gene-centric arrays. Consequently, the reported genetic signals explained only a small fraction of the population variance, suggesting that several hundred or thousand signals are involved[3,4]. Here we report an enlarged genomic analysis for AAM in a nearly twofold-larger sample of women than previously studied[3] and using more densely imputed genomic data. Our findings increase by more than threefold the number of independently associated signals and indicate likely causal effects of puberty timing on risks of various sex-steroid-sensitive cancers in men and women.

RESULTS

Genome-wide array data, imputed to the 1000 Genomes Project reference panel, were available in up to 329,345 women of European ancestry. These comprised 40 studies from the ReproGen consortium (N = 179,117), in addition to the 23andMe (N = 76,831) and UK Biobank (N = 73,397) studies (**Supplementary Table 1**). The distribution of genome-wide test statistics demonstrated significant inflation ($\lambda_{GC} = 1.75$); however, linkage disequilibrium (LD) score regression analyses confirmed that this inflation was solely due to polygenicity rather than population structure (LD score intercept = 1.00, standard error (SE) = 0.02). In total, 37,925 variants were associated with AAM at $P < 5 \times 10^{-8}$, which were resolved to 389 statistically independent signals (**Supplementary Fig. 1** and **Supplementary Table 2**). Per-allele effect sizes ranged from ~1 week to 5 months, 16 index variants were classed as low frequency (minor allele frequency (MAF) <5%; minimum observed frequency 0.5%), and 26 were insertion/deletion polymorphisms. Signals were distributed evenly across all 23 chromosomes with respect to chromosome size (**Supplementary Fig. 2**). Of the previously reported 106 autosomal, 5 exome array and 2 X-chromosome signals for AAM, all remained associated at genome-wide significance, except for two common loci (reported as *SCRIB–PARP10* ($P = 5 \times 10^{-4}$) and *FUT8* ($P = 5.4 \times 10^{-7}$)) and one rare variant not captured by the 1000 Genomes Project reference panel (p.Trp275*, *TACR3*).

Independent replication in the deCODE study (N = 39,543 women) showed that 367 (94.3%) of the 389 signals had directionally concordant effects (187 at $P < 0.05$), and 368 retained genome-wide significance in a combined meta-analysis (**Supplementary Table 3**). In aggregate, the top 389 index SNPs explained 7.4% of the trait variance in deCODE and 7.2% in UK Biobank (the latter estimate used weights derived from a meta-analysis excluding UK Biobank). These estimates are double the variance explained by the 106 previously reported signals[3] (3.7% in deCODE) and are equivalent to one-quarter of the total chip-captured heritability (h^2_{SNP} = 32%, SE = 1%) for AAM, estimated in UK Biobank.

A full list of authors and affiliations appears at the end of the paper.

Received 20 September 2016; accepted 17 March 2017; published online 24 April 2017; doi:10.1038/ng.3841

Consistent with our previous reports, we found strong sharing between the genetic architectures of AAM in women and age at voice breaking in men (considered as a continuous trait in 55,871 men in 23andMe) (genetic correlation (r_g) = 0.75, $P = 1.2 \times 10^{-79}$). Of the 389 AAM signals, 327 demonstrated directionally consistent trends or associations with age at voice breaking in men (binomial $P = 1.4 \times 10^{-44}$), and 18 signals reached a conservative multiple-test-corrected significance threshold ($P < 1 \times 10^{-4}$; 0.05/389) (**Supplementary Table 4**). Similarly, in UK Biobank where age at voice breaking was recorded using only three categories, 277 and 297 of the 377 autosomal loci demonstrated directionally consistent trends or associations with 'relatively early voice breaking' ($N = 2,678$ cases, $N = 55,763$ controls, binomial $P = 2.4 \times 10^{-20}$) and 'relatively late voice breaking' ($N = 3,566$ cases, $P = 1.9 \times 10^{-30}$), respectively (**Supplementary Table 5**).

Implicated genes and tissues

We used a number of analytical techniques to implicate genes in the regulation of AAM. These included mapping of nonsynonymous SNPs, expression quantitative trait locus (eQTL) analysis and integration of Hi-C chromatin-interaction data. Eight of the 389 lead variants were nonsynonymous, and a further 24 genes were implicated by highly correlated nonsynonymous variants ($r^2 > 0.8$) (**Supplementary Table 6**). These include genes disrupted in rare disorders of puberty: aromatase (*CYP19A1*, signal 307), gonadotropin-releasing hormone (*GNRH1*; signal 178), kisspeptin (*KISS1*; signal 31); and fucosyltransferase 2 (*FUT2*; signal 357), in which a stop-gain variant confers blood group secretor status.

Two approaches were used to interrogate publicly available gene expression data sets, both of which use one or more SNPs (not restricted to lead SNPs) to infer patterns of gene expression based on imputation reference panels (Online Methods). First, to maximize power, we analyzed data from the largest available eQTL data set for any tissue (whole blood, $N = 5,311$)[6], under the assumption that some causal genes and regulatory mechanisms might be ubiquitously expressed or functionally involved in blood tissues. Systematic eQTL integration using the Summary Mendelian Randomization approach[7] prioritized 113 transcripts, 60 of which had evidence for causal or pleiotropic effects, rather than coincidental overlap of signal (as indicated by HEIDI heterogeneity test $P > 0.009$) (**Supplementary Table 7**). Second, we used LD score regression applied to specifically expressed genes (LDSC-SEG)[8] to identify tissues and cell types that are enriched for AAM heritability. Five of the 46 GTEx tissues were positively enriched for AAM-associated variants (**Fig. 1**). Notably, all of these were central nervous system tissues, including the pituitary, and the hypothalamus was just below the significance threshold for enrichment ($P = 9.8 \times 10^{-3}$), consistent with the key role of this central axis[2]. Targeted assessment of these five enriched brain tissues using MetaXcan identified 205 genes whose expression was regulated by AAM-associated variants (**Supplementary Table 8**). Of note, later AAM was associated with higher transcript levels of *LIN28B* (signal 147) in the pituitary, *NCOA6* (nuclear receptor coactivator 6; signal 365) in the cerebellum, and *HSD17B12* (hydroxysteroid-(17-β)-dehydrogenase 12; signal 250) in various tissues.

To identify possible distal causal genes, we interrogated reported Hi-C data to assess whether any of the AAM loci are located in regions of chromatin looping[9]. 335 of the 389 loci were located within a topologically associating domain (TAD; a defined boundary region containing chromatin contact points), each of which contained on average ~5 genes (**Supplementary Table 9**). These included 22 of the 31 gene desert regions (nearest protein-coding gene >300 kb away), where TADs contained notable distal candidate genes such as *INHBA* (signal 158), *BDNF* (signal 248), *JARID2* (signal 128) and several γ-aminobutyric acid receptors (signal 91). We also observed several regions where multiple independent AAM signals all resided within one TAD containing the same single gene—*RORB* (signal 200, intronic; signal 199, ~200 kb downstream; signal 198, ~1.2 Mb downstream), *THRB* (signal 67, intronic; signal 68, ~180 kb upstream) and *TACR3* (signal 96, 5′UTR; signal 97, ~25 kb upstream; signal 98, ~133 kb upstream; signal 95, ~263 kb downstream).

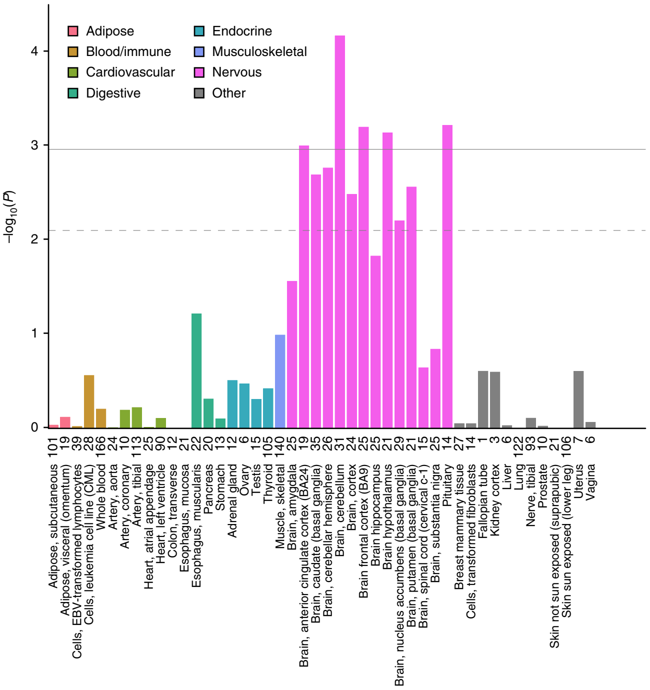

Figure 1 GTEx tissue enrichment using LD score regression. Numbers on the *x* axis correspond to sample number for each tissue. The dashed line represents significance at FDR < 5%, and the solid horizontal line represents Bonferroni-corrected significance for the number of tissues tested.

Sixty-six AAM signals were located in a specific contact point (from 5–25 kb in size) within the 335 TADs, indicating a direct physical connection between these signals and a distal genomic region, on average ~320 kb away. This included the previously reported example of the BMI-associated (and AAM-associated) *FTO* SNP and a distal *IRX3* promoter ~1 Mb away (signal 326)[10]. The longest chromatin interaction observed was ~38.6 Mb, where two distinct AAM signals, located ~300 kb apart (signals 206 and 207), were both in contact with the same distal genomic region ~38.6 Mb away that contains only one gene, *PTGES2*, encoding prostaglandin E synthase 2.

Transcription factor binding enrichment

To identify functional gene networks implicated in the regulation of AAM, we tested for enriched co-occurrence of AAM associations and predicted regulators within 226 putative enhancer modules, combining DNase–1 hypersensitive sites and chromatin states in 111 cell types and tissues. In total, we tested 2,382 transcription factor–enhancer module combinations. Sixteen transcription factor binding motifs were enriched for co-occurrence with AAM-associated variants within enhancer regions at study-level significance (false discovery rate (FDR) < 0.05) (**Supplementary Table 10**). Furthermore, 5 of the 16 genes encoding transcription factors associated with these motifs also mapped within 1 Mb of an index AAM-associated SNP. These included notable

Table 1 Parent-of-origin-specific associations between sequence variants at *MKRN3*, *DLK1* and *MEG9* and AAM in Iceland (N = 39,543)

		Allele		Freq.		Additive		Maternal		Paternal		
Marker	Position (hg38)	A1	A2	A1 (%)	Locus	P	β[a]	P	β[a]	P	β[a]	$P_{\text{mat vs. pat}}$[b]
rs530324840[c]	15:23,565,461	A	C	0.80	*MKRN3*	4.4×10^{-4}	−0.206	2.0×10^{-1}	0.098	6.4×10^{-11}	−0.523	1.3×10^{-7}
rs184950120[c]	15:23,565,696	T	C	0.26	*MKRN3*	1.0×10^{-2}	−0.265	9.8×10^{-1}	0.003	1.5×10^{-4}	−0.502	4.9×10^{-2}
rs12148769[c]	15:23,906,947	A	G	10.1	*MKRN3*	5.8×10^{-6}	−0.078	3.4×10^{-1}	−0.022	9.2×10^{-8}	−0.120	2.3×10^{-3}
rs138827001[d]	14:100,771,634	T	C	0.36	*DLK1*	6.8×10^{-6}	−0.387	8.8×10^{-1}	−0.018	4.7×10^{-10}	−0.704	1.4×10^{-4}
rs10144321[d]	14:100,416,068	G	A	23.0	*DLK1*	5.6×10^{-6}	−0.056	4.0×10^{-1}	−0.014	1.9×10^{-7}	−0.084	9.7×10^{-3}
rs7141210[d]	14:100,716,133	T	C	38.2	*DLK1*	4.5×10^{-2}	0.021	1.5×10^{-1}	−0.021	2.3×10^{-5}	0.059	4.0×10^{-4}
rs61992671[e]	14:101,065,517	A	G	48.5	*MEG9*	4.7×10^{-3}	−0.029	6.0×10^{-8}	−0.077	2.7×10^{-1}	0.015	1.9×10^{-5}

[a]The effect of allele A1 in years per allele. [b]P value for heterogeneity between paternal and maternal allele associations. [c]rs530324840 is a new variant identified by the parent-of-origin-specific analysis. rs184950120 is the rare variant identified by the meta-analysis. rs12148769 is the previously reported intergenic common signal (ref. 3). [d]rs138827001 is a new variant identified by the parent-of-origin-specific analysis. rs10144321 and rs7141210 are previously reported common variants (ref. 3). [e]rs61992671 is a suggestive new parent-of-origin-specific association signal.

candidates; first, *PITX1* (pituitary homeobox 1) is located within 50 kb of genome-wide significant SNPs (~500 kb from lead index signal 114). Second, *SMAD3*, a gene recently implicated in susceptibility to dizygous twinning[11], is located within 600 kb of an index SNP and its expression in several GTEx brain tissues is genetically correlated with AAM. Third, *RXRB* is located within ~500 kb of a new index SNP (signal 133), and it represents the fifth (out of nine) retinoid-related receptor gene implicated by genome-wide significant AAM variants. This set now includes all three retinoid X receptor–encoding genes (*RXRA*, *RXRB* and *RXRG*), and retinoid-related receptor genes are the nearest gene to the index SNP at three AAM loci (*RXRA*, *RORA* and *RORB*).

Pathway analyses

To identify other mechanisms that regulate pubertal timing, we tested all SNPs across the genome for enrichment of AAM associations with genes in predefined biological pathways. Ten pathways reached study-wise significance (FDR < 0.05). Five pathways were related to transcription factor binding, and the other pathways were peptide hormone binding, PI3-kinase binding, angiotensin-stimulated signaling, neuron development and γ-aminobutyric acid (GABA)-type B receptor signaling (**Supplementary Table 11**).

All of our previously reported custom pathways (**Supplementary Table 12**)[3] remained significant in this expanded data set: nuclear hormone receptors ($P = 2.4 \times 10^{-3}$); Mendelian pubertal disorder genes ($P = 1.9 \times 10^{-3}$); and JmjC-domain-containing lysine-specific demethylases ($P = 1 \times 10^{-4}$). Notably, new genome-wide significant signals mapped to lysine-specific demethylase genes: *JMJD1C* (signal 223), *PHF2* (208), *KDM4B* (347), *KDM6B* (332) and *JARID2* (128); or to Mendelian pubertal disorder genes: *CYP19A1* (307), *FGF8* (230), *GNRH1* (178), *KAL1* (378), *KISS1* (31), *NR5A1* (215) and *NR0B1* (379). The strongest AAM signal remained at *LIN28B*[3,12,13], which encodes a key repressor of let-7 microRNA (miRNA) biogenesis and cell pluripotency[14]. Transgenic *Lin28a/b* mice demonstrate both altered pubertal growth and glycaemic control[15], suggesting that the *LIN28*–let-7 axis could link puberty timing to type 2 diabetes susceptibility in humans. let-7 miRNA targets are reportedly enriched for variants associated with type 2 diabetes[16]. We tested the same set of computationally predicted and experimentally derived mRNA/protein let-7 miRNA targets[16] and observed significant enrichment of AAM-associated variants at miRNA targets that are downregulated by let-7b overexpression in primary human fibroblasts ($P_{\min} = 1 \times 10^{-3}$; **Supplementary Table 12**).

Imprinted genes and parent-of-origin effects

We previously reported an excess of parent-of-origin-specific associations for those AAM variants that map near imprinted genes, as defined primarily from animal studies[3]. Recent data from the GTEx consortium now allow a more systematic assessment of imprinted gene enrichment using genes defined from human transcriptome-wide analyses[17]. Consistent with our previous observations, imprinted genes were enriched for AAM-associated variants (MAGENTA $P = 4 \times 10^{-3}$), with a concordant excess of parent-of-origin-specific associations for the 389 index AAM variants (**Supplementary Fig. 3** and **Supplementary Table 3**).

Systematic assessment of the 389 AAM gene regions in the Icelandic deCODE study identified rare variants in two imprinted gene regions with robust parent-of-origin-specific associations with AAM. First, we identified a rare 5′ UTR variant, rs530324840 (MAF = 0.80% in Iceland), in *MKRN3* that was associated with AAM under the paternal model ($P = 6.4 \times 10^{-11}$, $\beta = -0.52$ years) but not the maternal model ($P = 0.20$, $\beta = 0.098$; $P_{\text{het}} = 1.3 \times 10^{-7}$) (**Table 1** and **Supplementary Table 13**). rs530324840 was by far the most significant variant at the *MKRN3* locus and is uncorrelated with our previously reported common variant rs12148769 at the same locus ($r^2 < 0.001$ in deCODE)[3] (**Supplementary Fig. 4**). We note that the rare 5′ UTR variant rs184950120 detected in the current genome-wide association study (GWAS) meta-analysis also showed paternal-specific association in deCODE and, despite being in close proximity (235 bp from rs530324840), is uncorrelated with rs530324840 ($r^2 < 0.0001$ in deCODE).

The second new robust parent-of-origin-specific signal is indicated by a rare intergenic variant at the *DLK1* locus (rs138827001; MAF = 0.36% in Iceland) that associated with AAM under the paternal model ($P = 4.7 \times 10^{-10}$, $\beta = -0.70$ years) but not the maternal model ($P = 0.88$, $\beta = -0.018$ years; $P_{\text{het}} = 1.4 \times 10^{-4}$) (**Table 1** and **Supplementary Fig. 5**). rs138827001 is uncorrelated with the two previously reported common variants rs10144321 and rs7141210 at the *DLK1* locus ($r^2 < 0.01$ in Iceland) that both also showed paternal-allele-specific associations[3]. At this locus, we observed a further common variant, rs61992671 (MAF = 48.5% in Iceland), 4.4 kb upstream of the *MEG9* (maternally expressed 9) gene (~300 kb from *DLK1*) that was associated with AAM under the maternal model ($P = 6.0 \times 10^{-8}$, $\beta = -0.077$ years) but not the paternal model ($P = 0.27$, $\beta = 0.015$ years; $P_{\text{het}} = 1.9 \times 10^{-5}$). rs61992671 is uncorrelated ($r^2 < 0.05$) with the two common signals identified in the meta-analysis (rs10144321 and rs7141210) and replicated with a consistent magnitude of effect in our GWAS meta-analysis (additive model, $P = 5.1 \times 10^{-6}$).

Disproportionate genetic effects on early or late puberty timing

Family-based studies in twins have suggested age-related differences in the impacts of genetic and environmental factors on AAM[18]. To test for asymmetry in the genetic effects on puberty timing, we defined two groups of women in the UK Biobank study on the basis of approximated quintiles for AAM—'early' (8–11 years inclusive;

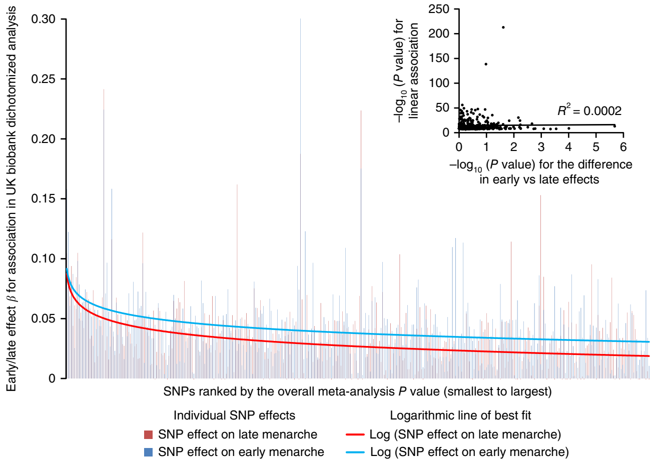

Figure 2 Stronger effects of AAM-associated signals on early menarche than late menarche in women. The 377 index AAM-associated SNPs are ordered from smallest to largest *P* value for their continuous association with AAM. The *y* axis shows the log-transformed odds ratio for each SNP on early menarche (blue; ages 8–11 years, inclusive) or late menarche (red; ages 15–19 years, inclusive). The reference group comprises women with menarche at 13 years. Insert, $-\log_{10}$ *P* values for heterogeneity (based on Cochran's *Q*) between the associations for early and late menarche plotted against the $-\log_{10}$ *P* value for the continuous AAM association.

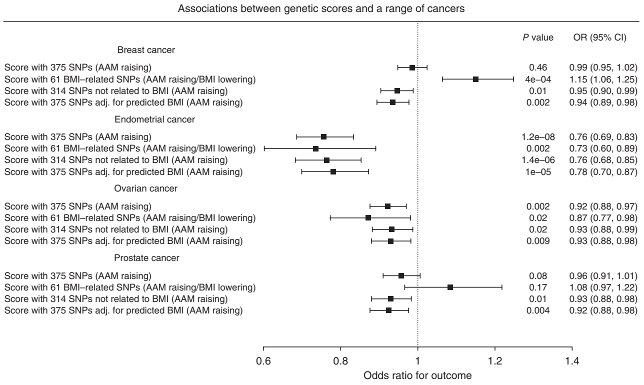

Figure 3 Effects and 95% confidence intervals of genetically predicted AAM on risks for various sex-steroid-sensitive cancers, adjusted for the effects of the same AAM variants on BMI. AAM was predicted by all 375 autosomal AAM-associated SNPs, and models were adjusted for the genetic effects of the same AAM variants on BMI. Three further genetic score associations are shown as sensitivity analyses for each outcome: first, AAM predicted by the 314 AAM-associated SNPs that were not also associated with BMI in the BCAC iCOGS sample (at a nominal level of $P < 0.05$); second, AAM predicted by the 61 AAM-associated SNPs that were also associated with BMI in this sample; and, third, AAM predicted by all 375 autosomal AAM-associated SNPs (unadjusted for BMI).

N = 14,922) and 'late' (15–19 years; N = 12,290). Each group was compared to the same median-quintile AAM reference group (age 13; N = 17,717). Estimated genome-wide heritability was higher for early AAM (h^2_{SNP} = 28.8%, SE = 2.3%) than for late AAM (h^2_{SNP} = 21.5%, SE = 2.5%; P_{dif} = 0.03). Accordingly, 217 of 377 (57.7%) autosomal index SNPs had larger effect estimates on early than on late AAM (binomial P = 0.004 versus 50% expected), and the aggregated effect of the 377 SNPs also differed between strata ($P = 2.3 \times 10^{-4}$) (**Fig. 2** and **Supplementary Table 14**). These differences remained when matching the early- and late-AAM strata for sample size and phenotype ranges (**Supplementary Table 15**).

In contrast, we observed the opposite pattern of disproportion in the genetic effects on male voice breaking in UK Biobank ('relatively early' N = 2,678, 'relatively late' N = 3,566). Genome-wide heritability estimates tended to be higher for relatively late voice breaking (7.8%, SE = 1.2%) than for relatively early voice breaking (6.9%, SE = 1.3%), and 227 of 377 (60.2%) index SNPs had larger effect estimates for relatively late than on relatively early voice breaking (binomial $P = 4.3 \times 10^{-5}$).

BMI-independent effects of puberty timing on cancer risks

Traditional (non-genetic) epidemiological studies have reported complex associations between puberty timing, BMI and adult cancer risks. For example, large studies using historical growth records identified lower adolescent BMI and earlier puberty timing (estimated by the age at peak adolescent growth) as predictors of higher breast cancer risk in women[19,20]. Conversely, BMI is positively associated with breast cancer risk in postmenopausal women[21]. Furthermore, the strong inter-relationship between puberty timing and BMI limits the ability to consider their distinct influences on disease risks in traditional observational studies. Consistent with our previous report[5], we observed a strong inverse genetic correlation between AAM and BMI (r_g = −0.35, $P = 1.6 \times 10^{-72}$). Thirty-nine AAM loci overlapped with reported loci for adult BMI[22], yet even those AAM signals with weak individual associations with adult BMI still contributed to BMI when considered in aggregate: the 237 AAM variants without a nominal individual association with adult BMI (all $P > 0.05$) were collectively associated with adult BMI ($P = 4.2 \times 10^{-9}$) (**Supplementary Fig. 6**). This finding precludes an absolute distinction between BMI-related and BMI-unrelated AAM variants.

In Mendelian randomization analyses, we therefore included adjustment for genetically predicted BMI (as predicted by the 375 autosomal AAM variants) to assess the likely direct (BMI-independent) effects of AAM on the risks for various sex-steroid-sensitive cancers (Online Methods). In these BMI-adjusted models, increasing AAM was associated with lower risk for breast cancer (odds ratio (OR) = 0.935 per year, 95% confidence interval (CI) = 0.894–0.977; $P = 2.6 \times 10^{-3}$) and in particular with estrogen receptor (ER)-positive but not ER-negative breast cancer (P_{het} = 0.02) (**Fig. 3** and **Supplementary Table 16**). Similarly, increasing AAM adjusted for genetically predicted BMI was associated with lower risks for ovarian cancer (OR = 0.930, 95% CI = 0.880–0.982; $P = 9.3 \times 10^{-3}$), in particular serous ovarian cancer (OR = 0.917, 95% CI = 0.859–0.978; $P = 8.9 \times 10^{-3}$) and endometrial cancer (OR = 0.781, 95% CI = 0.699–0.872; $P = 9.97 \times 10^{-6}$). Assuming an equivalent per-year effect of the current AAM variants on age at voice breaking, as we reported for the 106 previously identified AAM variants[5], we could also infer a protective effect of later puberty timing, independent of BMI, on risk for prostate cancer in men (OR = 0.925, 95% CI = 0.876–0.976; $P = 4.4 \times 10^{-3}$).

These findings were supported by sensitivity tests using subgroups of AAM signals stratified by their individual associations with adult BMI. The BMI-unrelated variant score (comprising 314 variants) supported a direct effect of AAM timing on breast cancer risk in women (OR = 0.946, 95% CI = 0.904–0.988; $P = 1.3 \times 10^{-2}$). In contrast, a score using only the 61 BMI-related AAM variants gave a significant result in the opposite direction (OR = 1.15, 95% CI = 1.06–1.25; $P = 4.3 \times 10^{-4}$) (**Supplementary Table 16**), consistent with the recently reported inverse association between genetically predicted BMI and breast cancer risk[23,24]. Further sensitivity tests (heterogeneity and MR-Egger tests) using the BMI-unrelated variant score suggested that additional subpathways might link AAM to risk of ovarian cancer (MR-Egger intercept P = 0.036), but, reassuringly, these tests indicated no further pleiotropy (beyond the effects of BMI) in our analyses of breast, endometrial and prostate cancers (for all, $I^2 < 23\%$ and MR-Egger intercept $P > 0.1$) (**Supplementary Fig. 7** and **Supplementary Table 16**).

DISCUSSION

In a substantially enlarged genomic analysis using densely imputed genomic data, we have identified 389 independent, genome-wide significant signals for AAM. In aggregate, these signals explain ~7.4% of the population variance in AAM, corresponding to ~25% of the estimated heritability. While assigning possible causal genes to associated loci is an ongoing challenge for GWAS findings, we adopted a number of recently described methods to implicate the underlying genes and tissues. Thirty-three genes were implicated by nonsynonymous variants and >200 genes were implicated by transcriptome-wide association in the five neural tissues enriched for AAM-associated gene activation. Transcriptome-wide association analyses also enabled the estimation of direction of gene expression in relation to AAM, notably indicating the likely delaying effect of *LIN28B* gene expression on AAM, which is consistent with inhibitory effects of this gene on developmental timing in animal and cell models[14,15].

Our findings add to the growing evidence for a significant role of imprinted genes in the regulation of puberty timing[3]. In a recent family study, rare coding mutations (two frameshift, one stop gain and one missense) in *MKRN3* were shown to cause central precocious puberty when paternally inherited[25]. Taken together, three distinct types of variants at *MKRN3* appear to influence puberty timing when paternally inherited: (i) multiple rare loss-of-function mutations with large effects[25]; (ii) a common intergenic variant (rs530324840) with small effect; and (iii) two 5′ UTR variants (rs184950120 and rs12148769) with intermediate allele frequencies (1 in 95 Icelandic women) and effects (~0.5 years per allele). Similarly, we found allelic heterogeneity at the imprinted *DLK1* locus where, as at *MKRN3*, a low-frequency paternally inherited allele conferred a substantial decrease in the age of puberty timing. At the same locus, maternal-allele-specific association with an unrelated variant near the maternally expressed gene *MEG9* is consistent with multiple imprinting control centers at this imprinted gene cluster[26].

The strong collective influence of the identified loci on AAM allowed informative stratification of AAM-associated variants in causal analyses to distinguish between BMI-related and BMI-unrelated pathways linking puberty timing to risk of sex-steroid-sensitive cancers. These findings were supported in BMI-adjusted models and, except for ovarian cancer, by additional tests for pleiotropy and indicate causal influences of both lower adolescent BMI and earlier AAM on later cancer risks. The association between BMI and breast cancer risk is complex; directionally opposing associations have been reported with adolescent and adult BMI, and with differing associations with pre- and postmenopausal breast cancer[19–21]. Recent Mendelian randomization studies report a consistent protective effect of higher BMI on pre- and postmenopausal breast cancer[23,24]. Some studies have reported on the association between later puberty timing and lower risk of prostate cancer in men, but such data on puberty timing in men are scarcely recorded[27]. The influences of earlier puberty timing, independent of BMI, on higher risks of breast, ovarian and endometrial cancers in women and prostate cancer in men could be mediated by a longer duration of exposure to sex steroids. Alternatively, mechanisms that confer earlier puberty timing might also promote higher levels of hypothalamic–pituitary–gonadal axis activity, as exemplified by a variant in *FSHB* that confers earlier AAM, higher circulating follicle-stimulating hormone concentrations in women and higher susceptibility to dizygous twinning[11].

We identified disproportionate effects of AAM variants on early or late puberty timing in a sex-discordant pattern. In females, variant effect estimates and heritability were higher for early versus late puberty timing, but the opposite was seen in males. These findings are concordant with clinical observations of sex-dependent penetrance of abnormal early and late puberty timing, even when accounting for presentation bias. Girls are more susceptible than boys to start puberty at abnormally young ages[28], whereas boys are more susceptible than girls to have delayed onset of puberty[29]. These findings suggest some yet-to-be-identified sex-specific gene–environment interactions. Future studies should systematically explore the potential influence of AAM-associated variants on rare disorders of puberty. In summary, our findings suggest unprecedented genetic complexity in the regulation of puberty timing and support new causal links with susceptibility to sex-steroid-sensitive cancers in women and men.

METHODS

Methods, including statements of data availability and any associated accession codes and 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

This research has been conducted using the UK Biobank Resource under applications 5122 and 9797. Full study-specific acknowledgments can be found in the **Supplementary Note**.

AUTHOR CONTRIBUTIONS

All authors reviewed the original and revised manuscripts. Statistical analysis: F.R.D., D.J.T., H.H., D.I.C., H.F., P.S., K.S.R., S.W., A.K.S., E. Albrecht, E. Altmaier, M.A., C.M.B., T. Boutin, A. Campbell, E.D., A.G., C. He, J.J.H., R.K., I.K., P.-R.L., K.L.L., M.M., B.M., G.M., S.E.M., I.M.N., R.N., T.N., L.P., N. Perjakova, E.P., L.M.R., K.E.S., A.V. Segrè, A.V. Smith, L.S., A.T., J.R.B.P. Sample collection, genotyping and phenotyping: I.L.A., S. Bandinelli, M.W.B., J.B., S. Bergmann, M.B., E.B., S.E.B., M.K.B., J.S.B., H. Brauch, H. Brenner, L.B., T. Brüning, J.E.B., H.C., E.C., S.C., G.C.-T., T.C., F.J.C., D.L.C., A. Cox, L.C., K.C., G.D.S., E.J.C.N.d., R.d., I.D.V., J.D., P.D., I.d.-S.-S., A.M.D., J.G.E., P.A.F., L.F.-R., L. Ferrucci, D.F.-J., L. Franke, M.G., I.G., G.G.G., H.G., D.F.G., P.G., P.H., E.H., U.H., T.B.H., C.A.H., G.H., M.J.H., J.L.H., F.H., D.J.H., H.K.I., M.-R.J., P.K.J., D.K., Z.K., G.L., D.L., C.L., L.J.L., J.S.E.L., S. Lenarduzzi, J. Li, P.A.L., S. Lindstrom, Y.L., J. Luan, R.M., A. Mannermaa, H.M., M.I.M., C. Meisinger, T.M., C. Menni, A. Metspalu, K.M., L.M., R.L.M., G.W.M., A.M.M., M.A.N., P.N., H.N., D.R.N., A.J.O., T.A.O., S.P., A. Palotie, N. Pedersen, A. Peters, J.P., P.D.P.P., A. Pouta, P.R., I. Rahman, S.M.R., A.R., F.R.R., I. Rudan, R.R., D.R., C.F.S., M.K.S., R.A.S., M. Shah, R.S., M.C.S., U.S., M. Stampfer, M. Steri, K. Strauch, T. Tanaka, E.T., N.J.T., M.T., T. Truong, J.P.T., A.G.U., D.R.V.E., V.V., U.V., P.V., Q.W., E.W., K.W.v.D., G.W., R.W., B.H.R.W., J.H.Z., M. Zoledziewska, M. Zygmunt. Individual study principal investigators: B.Z.A., D.I.B., M.C., F.C., T.E., N.F., C.G., V.G., C. Hayward, P.K., D.A.L., P.K.E.M., N.G.M., D.O.M.-K., E.A.N., O.P., D.P., A.L.P., P.M.R., H.S., T.D.S., D.S., D.T., S.U., J.A.V., H.V., N.J.W., J.F.W., A.B.S., U.T., K.S.P., D.F.E., J.Y.T., J.C.-C., D.H., A. Murray, J.M.M., K. Stefansson, K.K.O., J.R.B.P. Working group: F.R.D., D.J.T., H.H., D.I.C., H.F., P.S., K.S.R., S.W., A.K.S., A.B.S., U.T., K.S.P., D.F.E., J.Y.T., J.C., D.H., A. Murray, J.M.M., K. Stefansson, K.K.O., J.R.B.P.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Parent, A.S. *et al.* The timing of normal puberty and the age limits of sexual precocity: variations around the world, secular trends, and changes after migration. *Endocr. Rev.* **24**, 668–693 (2003).
2. Perry, J.R., Murray, A., Day, F.R. & Ong, K.K. Molecular insights into the aetiology of female reproductive ageing. *Nat. Rev. Endocrinol.* **11**, 725–734 (2015).
3. Perry, J.R. *et al.* Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* **514**, 92–97 (2014).

4. Lunetta, K.L. *et al.* Rare coding variants and X-linked loci associated with age at menarche. *Nat. Commun.* **6**, 7756 (2015).
5. Day, F.R. *et al.* Shared genetic aetiology of puberty timing between sexes and with health-related outcomes. *Nat. Commun.* **6**, 8842 (2015).
6. Westra, H.J. *et al.* Systematic identification of *trans* eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
7. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
8. Finucane, H.K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Preprint at *bioRxiv* https://dx.doi.org/10.1101/103069 (2017).
9. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
10. Smemo, S. *et al.* Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*. *Nature* **507**, 371–375 (2014).
11. Mbarek, H. *et al.* Identification of common genetic variants influencing spontaneous dizygotic twinning and female fertility. *Am. J. Hum. Genet.* **98**, 898–908 (2016).
12. Ong, K.K. *et al.* Genetic variation in *LIN28B* is associated with the timing of puberty. *Nat. Genet.* **41**, 729–733 (2009).
13. Perry, J.R. *et al.* Meta-analysis of genome-wide association data identifies two loci influencing age at menarche. *Nat. Genet.* **41**, 648–650 (2009).
14. Zhang, J. *et al.* LIN28 regulates stem cell metabolism and conversion to primed pluripotency. *Cell Stem Cell* **19**, 66–80 (2016).
15. Zhu, H. *et al.* *Lin28a* transgenic mice manifest size and puberty phenotypes identified in human genetic association studies. *Nat. Genet.* **42**, 626–630 (2010).
16. Zhu, H. *et al.* The *Lin28/let-7* axis regulates glucose metabolism. *Cell* **147**, 81–94 (2011).
17. Baran, Y. *et al.* The landscape of genomic imprinting across diverse adult human tissues. *Genome Res.* **25**, 927–936 (2015).
18. van den Berg, S.M. & Boomsma, D.I. The familial clustering of age at menarche in extended twin families. *Behav. Genet.* **37**, 661–667 (2007).
19. Ahlgren, M., Melbye, M., Wohlfahrt, J. & Sørensen, T.I. Growth patterns and the risk of breast cancer in women. *N. Engl. J. Med.* **351**, 1619–1626 (2004).
20. Collaborative Group on Hormonal Factors in Breast Cancer. Menarche, menopause, and breast cancer risk: individual participant meta-analysis, including 118 964 women with breast cancer from 117 epidemiological studies. *Lancet Oncol.* **13**, 1141–1151 (2012).
21. Bhaskaran, K. *et al.* Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults. *Lancet* **384**, 755–765 (2014).
22. Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
23. Gao, C. *et al.* Mendelian randomization study of adiposity-related traits and risk of breast, ovarian, prostate, lung and colorectal cancer. *Int. J. Epidemiol.* **45**, 896–908 (2016).
24. Guo, Y. *et al.* Genetically predicted body mass index and breast cancer risk: Mendelian randomization analyses of data from 145,000 women of European descent. *PLoS Med.* **13**, e1002105 (2016).
25. Abreu, A.P. *et al.* Central precocious puberty caused by mutations in the imprinted gene *MKRN3*. *N. Engl. J. Med.* **368**, 2467–2475 (2013).
26. da Rocha, S.T., Edwards, C.A., Ito, M., Ogata, T. & Ferguson-Smith, A.C. Genomic imprinting at the mammalian *Dlk1–Dio3* domain. *Trends Genet.* **24**, 306–316 (2008).
27. Giles, G.G. *et al.* Early growth, adult body size and prostate cancer risk. *Int. J. Cancer* **103**, 241–245 (2003).
28. de Vries, L., Kauschansky, A., Shohat, M. & Phillip, M. Familial central precocious puberty suggests autosomal dominant inheritance. *J. Clin. Endocrinol. Metab.* **89**, 1794–1800 (2004).
29. Wehkalampi, K., Widén, E., Laine, T., Palotie, A. & Dunkel, L. Patterns of inheritance of constitutional delay of growth and puberty in families of adolescent girls and boys referred to specialist pediatric care. *J. Clin. Endocrinol. Metab.* **93**, 723–728 (2008).

Felix R Day[1,163], Deborah J Thompson[2,163], Hannes Helgason[3,4,163], Daniel I Chasman[5,6], Hilary Finucane[7,8], Patrick Sulem[3], Katherine S Ruth[9], Sean Whalen[10], Abhishek K Sarkar[11,12], Eva Albrecht[13], Elisabeth Altmaier[14,15], Marzyeh Amini[16], Caterina M Barbieri[17], Thibaud Boutin[18], Archie Campbell[19], Ellen Demerath[20], Ayush Giri[21,22], Chunyan He[23,24], Jouke J Hottenga[25], Robert Karlsson[26], Ivana Kolcic[27], Po-Ru Loh[7,28], Kathryn L Lunetta[29,30], Massimo Mangino[31,32], Brumat Marco[33], George McMahon[34], Sarah E Medland[35], Ilja M Nolte[16], Raymond Noordam[36], Teresa Nutile[37], Lavinia Paternoster[34,38], Natalia Perjakova[39], Eleonora Porcu[40], Lynda M Rose[5], Katharina E Schraut[41,42], Ayellet V Segrè[43], Albert V Smith[44,45], Lisette Stolk[46], Alexander Teumer[47], Irene L Andrulis[48,49], Stefania Bandinelli[50], Matthias W Beckmann[51], Javier Benitez[52,53], Sven Bergmann[54,55], Murielle Bochud[56], Eric Boerwinkle[57], Stig E Bojesen[58–60], Manjeet K Bolla[2], Judith S Brand[26], Hiltrud Brauch[61–63], Hermann Brenner[63–65], Linda Broer[46], Thomas Brüning[66], Julie E Buring[5,6], Harry Campbell[42], Eulalia Catamo[67], Stephen Chanock[68], Georgia Chenevix-Trench[69], Tanguy Corre[54–56], Fergus J Couch[70], Diana L Cousminer[71,72], Angela Cox[73], Laura Crisponi[40], Kamila Czene[26], George Davey Smith[34,38], Eco J C N de Geus[25], Renée de Mutsert[74], Immaculata De Vivo[7,75], Joe Dennis[2], Peter Devilee[76,77], Isabel dos-Santos-Silva[78], Alison M Dunning[79], Johan G Eriksson[80], Peter A Fasching[51,81], Lindsay Fernández-Rhodes[82], Luigi Ferrucci[83], Dieter Flesch-Janys[84,85], Lude Franke[86], Marike Gabrielson[26], Ilaria Gandin[33], Graham G Giles[87,88], Harald Grallert[14,15,89], Daniel F Gudbjartsson[3,4], Pascal Guénel[90], Per Hall[26], Emily Hallberg[91], Ute Hamann[92], Tamara B Harris[93], Catharina A Hartman[94], Gerardo Heiss[82], Maartje J Hooning[95], John L Hopper[88], Frank Hu[75,96], David J Hunter[7,75,96], M Arfan Ikram[97], Hae Kyung Im[98], Marjo-Riitta Järvelin[99–103], Peter K Joshi[42], David Karasik[6,104], Manolis Kellis[11,12], Zoltan Kutalik[54,56], Genevieve LaChance[31], Diether Lambrechts[105,106], Claudia Langenberg[1], Lenore J Launer[93], Joop S E Laven[107], Stefania Lenarduzzi[67], Jingmei Li[26], Penelope A Lind[35], Sara Lindstrom[108], YongMei Liu[109], Jian'an Luan[1], Reedik Mägi[39], Arto Mannermaa[110–112], Hamdi Mbarek[25], Mark I McCarthy[113–115], Christa Meisinger[14,116], Thomas Meitinger[117], Cristina Menni[31], Andres Metspalu[39], Kyriaki Michailidou[2,118], Lili Milani[39], Roger L Milne[87,88], Grant W Montgomery[119], Anna M Mulligan[120,121], Mike A Nalls[122], Pau Navarro[18], Heli Nevanlinna[123], Dale R Nyholt[124], Albertine J Oldehinkel[125], Tracy A O'Mara[69], Sandosh Padmanabhan[126], Aarno Palotie[28,127–131], Nancy Pedersen[26], Annette Peters[14,89], Julian Peto[78], Paul D P Pharoah[2,79], Anneli Pouta[132], Paolo Radice[133], Iffat Rahman[134], Susan M Ring[34,38], Antonietta Robino[67], Frits R Rosendaal[74], Igor Rudan[42], Rico Rueedi[54,55], Daniela Ruggiero[37], Cinzia F Sala[17], Marjanka K Schmidt[135,136], Robert A Scott[1], Mitul Shah[79], Rossella Sorice[37], Melissa C Southey[137], Ulla Sovio[99,138], Meir Stampfer[7,75], Maristella Steri[40], Konstantin Strauch[13,139], Toshiko Tanaka[83], Emmi Tikkanen[131,140],

Nicholas J Timpson[34,38], Michela Traglia[17], Thérèse Truong[90], Jonathan P Tyrer[79], André G Uitterlinden[46,97], Digna R Velez Edwards[22,141,142], Veronique Vitart[18], Uwe Völker[143], Peter Vollenweider[144], Qin Wang[2], Elisabeth Widen[131], Ko Willems van Dijk[77,145,146], Gonneke Willemsen[25], Robert Winqvist[147,148], Bruce H R Wolffenbuttel[149], Jing Hua Zhao[1], Magdalena Zoledziewska[40], Marek Zygmunt[150], Behrooz Z Alizadeh[16], Dorret I Boomsma[25], Marina Ciullo[37], Francesco Cucca[40,151], Tõnu Esko[28,39], Nora Franceschini[82], Christian Gieger[14,15,89], Vilmundur Gudnason[44,45], Caroline Hayward[18], Peter Kraft[7,152], Debbie A Lawlor[34,38], Patrik K E Magnusson[26], Nicholas G Martin[35], Dennis O Mook-Kanamori[74,153], Ellen A Nohr[154], Ozren Polasek[27], David Porteous[19], Alkes L Price[7,8,28], Paul M Ridker[5,6], Harold Snieder[16], Tim D Spector[31], Doris Stöckl[14,155], Daniela Toniolo[17], Sheila Ulivi[67], Jenny A Visser[46], Henry Völzke[47], Nicholas J Wareham[1], James F Wilson[18,42], The LifeLines Cohort Study[156], The InterAct Consortium[156], kConFab/AOCS Investigators[156], Endometrial Cancer Association Consortium[156], Ovarian Cancer Association Consortium[156], PRACTICAL consortium[156], Amanda B Spurdle[69], Unnur Thorsteindottir[3,44], Katherine S Pollard[10,157], Douglas F Easton[2,79], Joyce Y Tung[158], Jenny Chang-Claude[159,160], David Hinds[158], Anna Murray[9], Joanne M Murabito[30,161], Kari Stefansson[3,44,164], Ken K Ong[1,162,164] & John R B Perry[1,164]

[1]MRC Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK. [2]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [3]deCODE Genetics/Amgen, Inc., Reykjavik, Iceland. [4]School of Engineering and Natural Sciences, University of Iceland, Reykjavik, Iceland. [5]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, Massachusetts, USA. [6]Harvard Medical School, Boston, Massachusetts, USA. [7]Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA. [8]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [9]Genetics of Complex Traits, University of Exeter Medical School, University of Exeter, Exeter, UK. [10]Gladstone Institutes, San Francisco, California, USA. [11]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [12]Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [13]Institute of Genetic Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [14]Institute of Epidemiology II, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [15]Research Unit of Molecular Epidemiology, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [16]Department of Epidemiology, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [17]Genetics of Common Disorders Unit, IRCCS San Raffaele Scientific Institute and Vita-Salute San Raffaele University, Milan, Italy. [18]Medical Research Council Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [19]Medical Genetics Section, Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [20]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, Minnesota, USA. [21]Division of Epidemiology, Institute for Medicine and Public Health, Vanderbilt University, Nashville, Tennessee, USA. [22]Vanderbilt Genetics Institute, Vanderbilt University, Nashville, Tennessee, USA. [23]Department of Epidemiology, Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, Indiana, USA. [24]Indiana University Melvin and Bren Simon Cancer Center, Indianapolis, Indiana, USA. [25]Department of Biological Psychology, VU University Amsterdam, Amsterdam, the Netherlands. [26]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [27]Faculty of Medicine, University of Split, Split, Croatia. [28]Program in Medical and Population Genetics, Broad Institute, Cambridge, Massachusetts, USA. [29]Department of Biostatistics, Boston University School of Public Health, Boston, Massachusetts, USA. [30]NHLBI's and Boston University's Framingham Heart Study, Framingham, Massachusetts, USA. [31]Department of Twin Research and Genetic Epidemiology, King's College London, London, UK. [32]National Institute for Health Research (NIHR) Biomedical Research Centre at Guy's and St Thomas' Foundation Trust, London, UK. [33]Department of Clinical Medical Sciences, Surgical and Health, University of Trieste, Trieste, Italy. [34]School of Social and Community Medicine, University of Bristol, Bristol, UK. [35]QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [36]Department of Internal Medicine, Section of Gerontology and Geriatrics, Leiden University Medical Center, Leiden, the Netherlands. [37]Institute of Genetics and Biophysics, CNR, Naples, Italy. [38]MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK. [39]Estonian Genome Center, University of Tartu, Tartu, Estonia. [40]Institute of Genetics and Biomedical Research, National Research Council, Cagliari, Italy. [41]Centre for Cardiovascular Sciences, Queen's Medical Research Institute, University of Edinburgh, Royal Infirmary of Edinburgh, Edinburgh, UK. [42]Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK. [43]Cancer Program, Broad Institute, Cambridge, Massachusetts, USA. [44]Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [45]Icelandic Heart Association, Kopavogur, Iceland. [46]Department of Internal Medicine, Erasmus MC, Rotterdam, the Netherlands. [47]Institute for Community Medicine, University Medicine Greifswald, Greifswald, Germany. [48]Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Ontario, Canada. [49]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [50]Geriatric Unit, Azienda Sanitaria di Firenze, Florence, Italy. [51]Department of Gynaecology and Obstetrics, University Hospital Erlangen, Friedrich Alexander University Erlangen-Nuremberg, Erlangen, Germany. [52]Human Genetics Group, Human Cancer Genetics Program, Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [53]Centro de Investigación en Red de Enfermedades Raras (CIBERER), Valencia, Spain. [54]Swiss Institute of Bioinformatics, Lausanne, Switzerland. [55]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [56]Institute of Social and Preventive Medicine, University Hospital of Lausanne, Lausanne, Switzerland. [57]Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, Texas, USA. [58]Copenhagen General Population Study, Herlev Hospital, Copenhagen University Hospital, University of Copenhagen, Copenhagen, Denmark. [59]Department of Clinical Biochemistry, Herlev Hospital, Copenhagen University Hospital, University of Copenhagen, Copenhagen, Denmark. [60]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [61]Dr. Margarete Fischer-Bosch Institute of Clinical Pharmacology, Stuttgart, Germany. [62]University of Tübingen, Tübingen, Germany. [63]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. [64]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany. [65]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany. [66]Institute for Prevention and Occupational Medicine of German Social Accident Insurance, Institute of Ruhr University Bochum (IPA), Bochum, Germany. [67]Institute for Maternal and Child Health, IRCCS "Burlo Garofolo", Trieste, Italy. [68]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, Maryland, USA. [69]Department of Genetics, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [70]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA. [71]Division of Genetics, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [72]Department of Genetics, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [73]Academic Unit of Molecular Oncology, Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK. [74]Department of Clinical Epidemiology, Leiden University Medical Center, Leiden, the Netherlands. [75]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [76]Department of Pathology, Leiden University Medical Center, Leiden, the Netherlands. [77]Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. [78]Non-Communicable Disease Epidemiology Department, London School of Hygiene and Tropical Medicine, London, UK. [79]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK. [80]Department of General Practice and Primary Health Care, University of Helsinki, Helsinki, Finland. [81]David Geffen School of Medicine, Department of Medicine, Division of Hematology and Oncology, University of California at Los Angeles, Los Angles, California, USA. [82]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, Chapel Hill, North Carolina, USA.

[83]Longitudinal Studies Section, Translational Gerontology Branch, National Institute on Aging, Baltimore, Maryland, USA. [84]Institute for Medical Biometrics and Epidemiology, University Clinic Hamburg-Eppendorf, Hamburg, Germany. [85]Department of Cancer Epidemiology/Clinical Cancer Registry, University Clinic Hamburg-Eppendorf, Hamburg, Germany. [86]Department of Genetics, University of Groningen, University Medical Centre Groningen, Groningen, the Netherlands. [87]Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Victoria, Australia. [88]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Victoria, Australia. [89]German Center for Diabetes Research, Neuherberg, Germany. [90]Cancer and Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris–Sud, University Paris–Saclay, Villejuif, France. [91]Division of Epidemiology, Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA. [92]Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany. [93]Laboratory of Epidemiology and Population Sciences, National Institute on Aging, Intramural Research Program, US National Institutes of Health, Bethesda, Maryland, USA. [94]Department of Psychiatry, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [95]Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, the Netherlands. [96]Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, USA. [97]Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands. [98]Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, Illinois, USA. [99]Department of Epidemiology and Biostatistics, MRC Health Protection Agency (HPA) Centre for Environment and Health, School of Public Health, Imperial College London, London, UK. [100]Biocenter Oulu, University of Oulu, Oulu, Finland. [101]Department of Children and Young People and Families, National Institute for Health and Welfare, Oulu, Finland. [102]Institute of Health Sciences, University of Oulu, Oulu, Finland. [103]Unit of Primary Care, Oulu University Hospital, Oulu, Finland. [104]Hebrew SeniorLife Institute for Aging Research, Boston, Massachusetts, USA. [105]Laboratory for Translational Genetics, Department of Oncology, University of Leuven, Leuven, Belgium. [106]Vesalius Research Center (VRC), VIB, Leuven, Belgium. [107]Division of Reproductive Medicine, Department of Obstetrics and Gynaecology, Erasmus MC, Rotterdam, the Netherlands. [108]Department of Epidemiology, School of Public Health, University of Washington, Seattle, Washington, USA. [109]Center for Human Genetics, Division of Public Health Sciences, Wake Forest School of Medicine, Winston-Salem, North Carolina, USA. [110]Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland. [111]Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland. [112]Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio, Finland. [113]NIHR Oxford Biomedical Research Centre, Churchill Hospital, Oxford, UK. [114]Oxford Centre for Diabetes, Endocrinology, and Metabolism, University of Oxford, Churchill Hospital, Oxford, UK. [115]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [116]Central Hospital of Augsburg, MONICA/KORA Myocardial Infarction Registry, Augsburg, Germany. [117]Institute of Human Genetics, Helmholtz Zentrum München–German Research Center for Environmental Health, Neuherberg, Germany. [118]Department of Electron Microscopy/Molecular Pathology, Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus. [119]Institute for Molecular Bioscience, University of Queensland, Brisbane, Queensland, Australia. [120]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. [121]Laboratory Medicine Program, University Health Network, Toronto, Ontario, Canada. [122]Laboratory of Neurogenetics, National Institute on Aging, Bethesda, Maryland, USA. [123]Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. [124]Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia. [125]Interdisciplinary Center Psychopathology and Emotion Regulation, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands. [126]British Heart Foundation Glasgow Cardiovascular Research Centre, Institute of Cardiovascular and Medical Sciences, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, UK. [127]Psychiatric and Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts, USA. [128]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. [129]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, UK. [130]Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA. [131]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. [132]National Institute for Health and Welfare, Helsinki, Finland. [133]Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Preventive and Predictive Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan, Italy. [134]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [135]Division of Molecular Pathology, Netherlands Cancer Institute–Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands. [136]Division of Psychosocial Research and Epidemiology, Netherlands Cancer Institute–Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands. [137]Department of Pathology, University of Melbourne, Melbourne, Victoria, Australia. [138]Department of Obstetrics and Gynaecology, University of Cambridge, Cambridge, UK. [139]Institute of Medical Informatics, Biometry and Epidemiology, Chair of Genetic Epidemiology, Ludwig-Maximilians-Universität, Munich, Germany. [140]Department of Public Health, University of Helsinki, Helsinki, Finland. [141]Vanderbilt Epidemiology Center, Institute for Medicine and Public Health, Vanderbilt University, Nashville, Tennessee, USA. [142]Department of Obstetrics and Gynecology, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. [143]Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany. [144]University Hospital of Lausanne, Lausanne, Switzerland. [145]Department of Internal Medicine, Division of Endocrinology, Leiden University Medical Center, Leiden, the Netherlands. [146]Einthoven Laboratory for Experimental Vascular Medicine, Leiden University Medical Center, Leiden, the Netherlands. [147]Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Oulu, Finland. [148]Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre NordLab, Oulu, Finland. [149]Department of Endocrinology, University of Groningen, University Medical Centre Groningen, Groningen, the Netherlands. [150]Department of Obstetrics and Gynecology, University Medicine Greifswald, Greifswald, Germany. [151]Department of Biomedical Sciences, University of Sassari, Sassari, Italy. [152]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [153]Department of Public Health and Primary Care, Leiden University Medical Center, Leiden, the Netherlands. [154]Research Unit for Gynaecology and Obstetrics, Department of Clinical Research, University of Southern Denmark, Odense, Denmark. [155]Department of Obstetrics and Gynaecology, Campus Grosshadern, Ludwig Maximilians University, Munich, Germany. [156]A full list of members and affiliations appears in the **Supplementary Note**. [157]Division of Biostatistics, Institute for Human Genetics, and Institute for Computational Health Sciences, University of California, San Francisco, San Francisco, California, USA. [158]23andMe, Inc., Mountain View, California, USA. [159]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [160]University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [161]Boston University School of Medicine, Department of Medicine, Section of General Internal Medicine, Boston, Massachusetts, USA. [162]Department of Paediatrics, University of Cambridge, Cambridge, UK. [163]These authors contributed equally to this work. [164]These authors jointly directed this work. Correspondence should be addressed to J.R.B.P. (john.perry@mrc-epid.cam.ac.uk) or K.K.O. (ken.ong@mrc-epid.cam.ac.uk).

ONLINE METHODS

GWAS meta-analysis for age at menarche in women. Each individual study tested SNPs using a two-tailed additive linear regression model for association with AAM, including age at study visit and other study-specific covariates. Insertion/deletion polymorphisms were coded as "I" and "D" for data storage efficiency and to allow harmonization across all studies. Genetic variants and individuals were filtered on the basis of study-specific quality control metrics. Association statistics for each SNP were then uploaded by study analysts for central processing. Study-level results files were assessed following a standardized quality control pipeline[30], and results for each SNP were subjected to meta-analysis across studies using an inverse-variance-weighted model with METAL[31] in a two-stage process. First, results from ReproGen consortium studies (**Supplementary Table 1**) were combined and then filtered so that only SNPs that appeared in over half of these studies were taken forward. Second, aggregated ReproGen consortium results were combined with data from the UK Biobank[32,33] and 23andMe[5] studies. Variants were only included in the final results file if they had results from at least two of these three sources, and combined MAF >0.1%. We assessed potential inflation of test statistics due to sample relatedness and population stratification using LD score regression[34]. Here an intercept value not significantly different from 1 indicates no such inflation, with a value over 1 indicating inflation.

A final list of index variants was first defined using a distance-based metric, by which any SNPs passing the two-tailed threshold of significance ($P < 5 \times 10^{-8}$) within 1 Mb of another significant SNP were considered to be located in the same locus. This list of signals was then further augmented using approximate conditional analysis in GCTA, using an LD reference panel from the UK Biobank study. Only secondary signals that were uncorrelated ($r^2 < 0.05$) were included in the final list.

Replication and parent-of-origin testing. Replication of identified hits was performed in an independent sample of 39,486 women of European ancestry from the deCODE study, Iceland. Main effects and parent-of-origin association testing were performed using the same methodology as previously reported[3,4]. The fraction of variance explained by a variant associating under the additive model was calculated using the formula $2f(1-f)\beta_a^2$, where f denotes the MAF of the variant and β_a is the additive effect. For variants associating under the recessive model, the formula $f_h(1-f_h)\beta_r^2$ was used, where f_h denotes the homozygous frequency of the variant and β_r denotes the recessive effect. For variants associating under parent-of-origin models, fraction of variance explained was computed using the formulas $f(1-f)\beta_m^2$ for the maternal model and $f(1-f)\beta_p^2$ for the paternal model, where f denotes the MAF of the variant, β_m denotes the effect under the maternal model and β_p denotes the effect under the paternal model. Variance explained across multiple SNPs was calculated by summing the individual variances for all uncorrelated variants. We also estimated variance explained for top hits in UK Biobank using a combined allele score of all 377 autosomal genetic variants. Each individual variant was weighted using effect estimates derived from a meta-analysis excluding UK Biobank.

Age at voice breaking in men. Data on male voice breaking were available from two sources. First, the 23andMe study recorded recalled age at voice breaking in a sample of 55,871 men, as previously described[5]. This was recorded as a quantitative trait into predefined 2-year age bins by online questionnaire in response to the question "How old were you when your voice began to crack/deepen?" (ref. 5). Individual SNP effect estimates from the 2-year age bins were rescaled to 1-year estimates for both voice breaking and AAM as reported previously.

Age at voice breaking was also recalled in the UK Biobank study, as previously described[33]. This was recorded as a categorical trait—"younger than average," "about average age," "older than average," "do not know" or "prefer not to answer"—in response to the question "When did your voice break?" In separate models, the earlier or later voice breaking groups were compared to the average group (used as the reference group).

Disproportionate effects on early or late puberty timing. Disproportionate effects on early or late puberty timing of AAM-associated SNPs were tested for AAM in UK Biobank. The distribution of AAM was divided into approximate quintiles, as previously reported[33]. Odds ratios for being in the youngest quintile (range 8–11 years) or the oldest quintile (range 15–19 years) were compared to the middle quintile (age 13 years) as the reference, for each AAM-associated SNP and also for a combined weighted AAM-increasing allele score, with weights derived from a meta-analysis of all studies except UK Biobank. Sensitivity tests were performed by dividing UK Biobank individuals into broad strata on the basis of birth year (before or after 1950) and geographical location (attendance at a study assessment center in the north or south of the UK, as indicated by a line joining Mersey and Humber).

Genetic correlation and genome-wide variance analysis. Genome-wide genetic correlations with adult BMI[22] and voice breaking[5] were estimated using LD score regression implemented in LDSC[34]. The total trait variance of all genotyped SNPs was calculated using restricted estimate maximum likelihood (REML) implemented in BOLT[35]. This was estimated using the same UK Biobank study sample in the discovery analysis, excluding any related individuals. The proportion of the heritability explained by index SNPs was estimated by dividing the variance explained by the index SNPs by the total variance explained by all genotyped SNPs across the genome.

Mendelian randomization analyses. Individual genotype data on cancer outcomes were available from the Breast Cancer Association Consortium (BCAC) and the Endometrial Cancer Association Consortium (ECAC). In addition, summary-level results for ovary and prostate cancer were made available from the Ovarian Cancer Association Consortium (OCAC) and the Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome (PRACTICAL) consortium, respectively. Total analyzed numbers were as follows: 47,800 breast cancer cases and 40,302 controls, 4,401 endometrial cancer cases and 28,758 controls, 18,175 ovarian cancer cases and 26,134 controls, and 20,219 prostate cancer cases and 20,440 controls (from the PRACTICAL iCOGS data set).

We performed Mendelian randomization analyses to assess the likely causal effects of puberty timing on the risks for various sex-steroid-sensitive cancers. Hence, AAM was predicted by a weighted genetic risk score of all 375 autosomal AAM-associated SNPs, and genetically predicted AAM was tested for association with each cancer in a logistic regression model. The individual SNP genotype dosages comprising this score were imputed using the 1000 Genomes Project reference panel (minimum imputation $r^2 = 0.43$, median = 0.95). To avoid potential confounding by effects of the AAM genetic risk score on BMI, we performed BMI-adjusted analyses by including the same AAM genetic risk score in models as a covariate, but weighting each SNP for its effect on BMI (rather than on AAM) in the same study sample. Hence, we estimated the effect of genetically predicted AAM, controlling for genetically predicted BMI by the same SNPs. BMI weighting was based on the association between each SNP and adult BMI in this sample (childhood BMI measurements were not available, but there is reportedly high genetic correlation between adult and childhood obesity ($r_g = 0.73$)[36]. We did not adjust for measured BMI because such measurements in prevalent cancer cases are likely to introduce bias. As sensitivity tests, three further genetic score associations were performed for each cancer outcome: first, AAM predicted by the 314 AAM-associated SNPs that were not also individually associated with BMI in the BCAC iCOGS sample (at a nominal level of $P < 0.05$); second, AAM predicted by the 61 AAM-associated SNPs that were also associated with BMI in this sample ($P < 0.05$); and, finally, AAM predicted by all 375 autosomal AAM-associated SNPs (unadjusted for BMI). To further consider pleiotropy, we tested for the presence of heterogeneity between AAM-associated SNPs and analyzed MR-Egger regression models[37].

Pathway analyses. MAGENTA was used to explore pathway-based associations in the full GWAS data set. MAGENTA implements a gene set enrichment analysis (GSEA)-based approach, as previously described[38]. Briefly, each gene in the genome is mapped to a single index SNP with the lowest P value within a 110 -kb upstream, 40-kb downstream window. This P value, representing a gene score, is then corrected for confounding factors such as gene size, SNP density and LD-related properties in a regression model. Genes within the human leukocyte antigen (HLA) region were excluded from analysis because of difficulties in accounting for gene density and LD patterns. Each mapped

doi:10.1038/ng.3841

gene in the genome is then ranked by its adjusted gene score. At a given significance threshold (95th and 75th percentiles of all gene scores), the observed number of gene scores in a given pathway with a ranked score above the specified threshold percentile is calculated. This observed statistic is then compared to 1,000,000 randomly permuted pathways of identical size. This generates an empirical GSEA P value for each pathway. Significance was determined when an individual pathway reached FDR < 0.05 in either analysis. In total, 3,216 pathways from Gene Ontology, PANTHER, KEGG and Ingenuity were tested for enrichment of multiple modest associations with AAM. MAGENTA software was also used for enrichment testing of custom gene sets.

Gene expression data integration. To identify which tissues and cell types were likely to be most relevant to genes involved in pubertal development, we applied LD score regression[39] to specifically expressed genes (LDSC-SEG)[8]. For each tissue, we ranked genes by a t statistic for differential expression, using sex and age as covariates and excluding all samples in related tissues. For example, we compared expression in hippocampus samples to expression in all non-brain samples. We then took the top 10% of genes by this ranking, formed a genome annotation including these genes (exons and introns) plus 100 kb on either side, and used stratified LD score regression to estimate the contribution of this annotation to per-SNP AAM heritability, adjusting for all categories in the baseline model[39]. We computed significance using a block jackknife over SNPs and corrected for 46 hypotheses tested at $P = 0.05$.

To identify specific eQTL-linked genes, we used two complementary approaches to systematically integrate publicly available gene expression data with our genome-wide data set: summary mendelian randomization (SMR) and MetaXcan.

SMR uses summary-level gene expression data to map potentially functional genes to trait-associated SNPs[7]. We ran this approach against the publicly available whole-blood eQTL data set published by Westra *et al.*[6], giving association statistics for 5,950 transcripts. A conservative significance threshold was set at $P < 8.4 \times 10^{-6}$, in addition to a HEIDI test statistic $P > 0.009$ for any variants that surpass the main threshold.

MetaXcan, a meta-analysis extension of the PrediXcan method[40], was used to infer the association between genetically predicted gene expression (GPGE) and AAM. PrediXcan is a gene-based data aggregation and integration method that incorporates information from gene expression and GWAS data to translate evidence of association with a phenotype from the SNP level to the gene. Briefly, PrediXcan first imputes gene expression at an individual level using prediction models trained on measured transcriptome data sets with genome-wide SNP data and then regresses the imputed transcriptome levels with the phenotype of interest. MetaXcan extends its application to allow inference of the direction and magnitude of GPGE–phenotype associations with only summary GWAS statistics, which is advantageous when SNP–phenotype associations result from a meta-analysis setting and also when individual-level data are not available. As input, we used GWAS meta-analysis summary statistics for AAM, LD matrix from the 1000 Genomes Project and, as weights, gene expression regression coefficients for SNPs from models trained with transcriptome data (V6p) from the GTEx Project[41]. GTEx is a large-scale collaborative effort where DNA and RNA from multiple tissues were sequenced from almost 1,000 deceased individuals of European, African and Asian ancestry. MetaXcan analyses were targeted to those tissue types with previous evidence of association with AAM (based on the GTEx enrichment analyses described above). The threshold for statistical significance was estimated using the Bonferroni method for multiple-testing correction across all tested tissues ($P < 2.57 \times 10^{-6}$).

Motif enrichment testing. We identified transcription factors whose binding could be disrupted by AAM-associated variants in enhancer regions by combining predicted enhancer regions across 111 human cell types and tissues with predicted motif instances for 651 transcription factor families as previously described[42].

Briefly, we defined enhancer regions by first applying ChromHMM[43], training a 15-state model for each reference epigenome on 5 histone modifications: H3K4me1, H3K4me3, H3K36me3, H3K9me3 and H3K27me3. We then produced a higher-confidence set of predicted enhancer regions in each reference epigenome by intersecting DNase I hypersensitive sites (taking the union over 53 reference epigenomes for which DNase–seq was performed) with enhancer-like chromatin states predicted in that reference epigenome[42]. We defined 226 disjoint enhancer modules with distinct patterns of activity by hierarchically clustering the high-confidence regions according to their patterns of activity (presence/absence) across the 111 reference epigenomes.

We predicted motif instances by first building a database of position weight matrices (PWMs) combining known motifs from Transfac and Jaspar with *de novo*–discovered motifs in 427 ChIP–seq experiments for 123 transcription factors from ENCODE[44]. We predicted active regulators in each enhancer module by computing the enrichment of true PWM matches in the set of regions assigned to that module against the background of shuffled PWM matches. We only considered PWMs with a conservation score of at least 0.3 and used $\log_2$ (fold enrichment) > 1.5 as the significance cutoff.

We used the full set of AAM association summary statistics, excluding the 23andMe component, to identify a heuristic P-value threshold[42]. Briefly, we pruned a set of 8,094,080 variants to 432,550 independent loci (pairwise $r^2 < 0.1$). We scored each locus as the proportion of variants in the locus overlapping a predicted enhancer region, ranked loci by the best P value in the locus and then plotted enrichment curves comparing the cumulative score every 100 loci against the expected score for that total number of loci under the null where the score increases uniformly to the genome-wide value. We defined the right-most elbow point (inflection point) among all the enrichment curves as the heuristic P-value cutoff.

For each combination of enhancer module and predicted regulator, we constructed a 2 × 2 contingency table counting enhancer regions in that module partitioned by presence of that motif and orthogonally by presence of an AAM association (based on the heuristic P-value cutoff described above). We restricted the set of regions to the domain on which motifs were discovered (excluding coding regions, 3′ UTRs, transposons and repetitive regions) and additionally to the subset of regions that harbor an imputed SNP for the disease. We computed one-sided P values using Fisher's exact test.

Hi-C integration. Significant Hi-C interactions and contact domains were obtained from Rao *et al.* (GEO accession GSE63525) for six ENCODE cell lines: K562, GM12878, HeLa-S3, IMR90, NHEK and HUVEC. Their Juicer pipeline assigns statistical significance to each Hi-C interaction at resolutions ranging from 5–25 kb, depending on coverage, at a 10% FDR. Contact domains are genomic regions enriched for regulatory interactions and are more conserved across cell types than are specific interactions. They are conceptually similar to TADs, but with improved resolution (median length of 185 kb versus 880 kb). We used the intersect command of BEDtools to produce a list of significantly interacting Hi-C fragments containing one or more of our identified SNPs in either fragment from any of the six cell lines. For each SNP-containing fragment, genes present in the corresponding interacting fragment were identified as potential regulatory targets. As a second approach, we also scored genes on the basis of the number of ENCODE cell types in which they were in the same contact domain as a SNP.

Data availability. GWAS meta-analysis summary statistics from the ReproGen consortium are available to download from the ReproGen website (http://www.reprogen.org/).

30. Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat. Protoc.* **9**, 1192–1212 (2014).
31. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
32. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
33. Day, F.R., Elks, C.E., Murray, A., Ong, K.K. & Perry, J.R. Puberty timing associated with diabetes, cardiovascular disease and also diverse health outcomes in men and women: the UK Biobank study. *Sci. Rep.* **5**, 11208 (2015).
34. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
35. Loh, P.R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

36. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
37. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
38. Segrè, A.V., Groop, L., Mootha, V.K., Daly, M.J. & Altshuler, D. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet.* **6**, e1001058 (2010).
39. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
40. Gamazon, E.R. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
41. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
42. Sarkar, A., Ward, L.D. & Kellis, M. Functional enrichments of disease variants across thousands of independent loci in eight diseases. Preprint at *bioRxiv* http://dx.doi.org/10.1101/048066 (2016).
43. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
44. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).