Genome-wide meta-analysis associates HLA-DQA1/DRB1 and LPA and lifestyle factors with human longevity

Citation for published version:

Digital Object Identifier (DOI):
10.1038/s41467-017-00934-5

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF; also known as Version of record

Published In:
Nature Communications

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of a file breaches copyright, please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 20. Jul. 2019
Genomic analysis of longevity offers the potential to illuminate the biology of human aging. Here, using genome-wide association meta-analysis of 606,059 parents’ survival, we discover two regions associated with longevity (HLA-DQA1/DRB1 and LPA). We also validate previous suggestions that APOE, CHRNA3/S, CDKN2A/B, SH2B3 and FOXO3A influence longevity. Next we show that giving up smoking, educational attainment, openness to new experience and high-density lipoprotein (HDL) cholesterol levels are most positively genetically correlated with lifespan while susceptibility to coronary artery disease (CAD), cigarettes smoked per day, lung cancer, insulin resistance and body fat are most negatively correlated. We suggest that the effect of education on lifespan is principally mediated through smoking while the effect of obesity appears to act via CAD. Using instrumental variables, we suggest that an increase of one body mass index unit reduces lifespan by 7 months while 1 year of education adds 11 months to expected lifespan.
longevity is of interest to us all, and philosophers have long speculated on the extent to which it is pre-determined by fate. Here we focus on a narrower question—the extent and nature of its genetic basis and how this inter-relates with that of health and disease traits. In what follows, we shall use longevity as an umbrella term. We shall also more specifically refer to lifespan (the duration of life) and long-livedness (living to extreme old age, usually defined by a threshold, such as 90 years). Up to 25% of the variability in human lifespan has been estimated to be genetic, but genetic variation at only three loci (near APOE, FOXO3A and CHRNA3/5) have so far been demonstrated to be robustly associated with lifespan.

Prospective genomic studies of lifespan have been hampered by the fact that subject participation is often only recent, allowing insufficient follow-up time for a well-powered analysis of participant survival. On the other hand, case-control studies of long-livedness have had success and some technical appeal (focussing on the truly remarkable), but such studies can be limited and costly in their recruitment. We recently showed that the extension of the kin-cohort method to parental lifespans, beyond age 40, of genotyped subjects could be used to detect genetic associations with lifespan with some power in genomically British participants in UK Biobank (UKB). Here we extend that approach in a genome-wide association meta-analysis (GWAMA) to discovery across UKB European- and African-ancestry populations and 24 further population studies (LifeGen), mainly from Europe, Australia and North America, to search for further genetic variants influencing longevity. We then use those GWAMA results to measure genetic correlations and carry out Mendelian randomisation (MR) between other traits and lifespan seeking to elucidate the underlying effects of disease and socio-economic traits on longevity, in a framework less hampered by confounding and reverse causality than observational epidemiology.

Results
Genome-wide association study. In total, 606,059 parental lifespans were available for analysis, of which 334,974 were already complete (Table 1).

In our GWAS of 586,626 European parental lifespans, we find four regions HLA-DQA1/DRB1, LPA, CHRNA3/5 and APOE, in which the lead SNPs rs34831921, rs55730499, rs8042849 and rs429358, respectively, associate with survival at genome-wide significance (p < 5 × 10^{-8}) (Table 2, Fig. 1a, b, Fig. 2a–d). The two previously unreported loci, rs34831921 (HLA-DQA1/DRB1) and rs55730499 (LPA), both showed statistically significant, directionally consistent, evidence of association at the proxy SNPs in strongest LD in the largest (5406 cases, 15,112 controls) publicly available set of GWAS summary statistics for extreme long-livedness (CHARGE-EU 90+), with p < 0.0035 for both SNPs. As our GWAS results were of the observed effect of offspring genotype on parent phenotype and the actual effect of carrying an allele for the individual concerned (rather than their parent) is twice that observed in a parent-offspring kin-cohort study, all reported effect sizes (and their standard errors) throughout this manuscript have been doubled to give the estimated effect size in the allele carriers themselves. The hazard ratios for one copy of the minor alleles were 0.942 and 1.074 for rs34831921 (HLA-DQA1/DRB1) and rs55730499 (LPA), respectively, corresponding to an increase/decrease in lifespan of ~0.6/0.7 years for a carrier of one additional copy of the minor allele.

We meta-analysed our results with the CHARGE-EU 90+ longevity GWAMA summary statistics using Z-scores and equal weights for each study, reflecting their similar statistical power. We found strengthened signals, substantially at APOE (rs4420638, p = 5.4 × 10^{-11}) and slightly in the LPA region (rs1045587, p = 2.05 × 10^{-11}). No improvement of statistical significance was observed in the HLA-DQA1/DRB1 region, where there were no SNPs in strong LD with the lead LifeGen SNP, nor was there an increase in significance near CHRNA3/5. However, in this meta-analysis one further region near AKAP7/EPB41L2 on chromosome 6 just reached genome-wide significance (rs1919453, A allele frequency = 0.36, p = 4.34 × 10^{-8}; Fig. 1c, Supplementary Fig. 1), and the observed hazard ratio (SE) for the minor allele was 0.976 (0.0056) in LifeGen alone. In our study of 9359 father and 10,074 mother lifespans in participants with African ancestry, no SNPs were genome-wide significant in the analysis of both parents combined. However, we found one GW significant result (rs10918124, G allele frequency 0.39 in African subjects), in an intergenic region of chromosome 2 associating with lifespan for fathers (HR (SE) for G allele = 1.22 (0.0354), p = 1.66 × 10^{-8}), with a consistent direction of association in all 9 cohorts studied. No association was observed at this SNP in African mothers, or fathers and mothers of European ancestry (HR (SE) = 0.97 (0.038), 1.01 (0.007) and 1.00 (0.008), p = 0.51, 0.21 and 0.77, respectively (Fig. 1d, Supplementary Fig. 2A–D).

Cross-validation of candidate genes. We next attempted to validate 13 candidate genes identified in previous longevity studies. In our study, only three of these genes showed statistically significant, directionally consistent evidence (p < 0.0003, two-sided test) of association; CDKN2A/B, SH2B3 and FOXO3A (Fig. 3, Supplementary Fig. 3 and Supplementary Data 3). For SH2B3 and FOXO3A our estimated effect sizes are concordant with those reported from the most robust (i.e., narrowest 95% confidence interval (CI)) previous study. However, for CDKN2A/B, the 95% CI for our estimate is entirely below that from the more robust of the two studies considered.

Table 1 Summary of the LifeGen parental lifespans

<table>
<thead>
<tr>
<th>Ancestry</th>
<th>Parent</th>
<th>Count</th>
<th>Mean age</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Alive</td>
<td>Dead</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alive</td>
<td>Dead</td>
</tr>
<tr>
<td>African</td>
<td>Father</td>
<td>2435</td>
<td>6924</td>
</tr>
<tr>
<td>African</td>
<td>Mother</td>
<td>4185</td>
<td>5889</td>
</tr>
<tr>
<td>European</td>
<td>Father</td>
<td>113,611</td>
<td>178,017</td>
</tr>
<tr>
<td>European</td>
<td>Mother</td>
<td>150,854</td>
<td>144,144</td>
</tr>
<tr>
<td>ALL</td>
<td></td>
<td>271,085</td>
<td>334,974</td>
</tr>
</tbody>
</table>

Summary statistics for the 606,059 parental lifespans that passed phenotypic QC (in particular, parent age > 40) and were analysed here. In practice, fewer lives than these were analysed for some SNPs, as a SNP may not have passed QC in all cohorts (in particular within cohort MAF > 1%). The mean age of alive parents across European cohorts was reduced by the large iPSYCH cohort, of relatively younger subjects and thus parents, who were predominantly alive (mean father/mother age among the alive parents in iPSYCH was 52.4/50.4).
PhenoScanner8, an online tool which searches 88 complex trait (Table 2) for association with other ageing traits, using the FOXO3 locus had greater effect at older ages (75+) while the effects, while the 1.06 (.01) for men and 1.04 (.01) for women (Fig. 4a). Conversely, for APOE, hazard ratios stratified by age were 1.06 (.01) for ages 40−75 and 1.14 (.01) for ages 75+, whereas for CHRNA3/5 they were 1.08 (.01) for 40−75 and 1.03 (.01) for age 75+ (Fig. 4b), with similar patterns when stratifying by age and sex at the same time, (Fig. 4c), although the distinctions between men and women for CHRNA3/5 disappeared beyond age 75. For LPA, CDKN2B and SH2B3, there was no statistically significant evidence of age-specific or sex-specific effects, while the HLA and FOXP3 variants showed age but not sex-specific effects (Fig. 4a, b), with the HLA locus having a greater effect at younger ages (40−75) while, conversely, the FOXP3 locus had greater effect at older ages (75+).

We tested the four SNPs identified in the discovery phase (Table 2) for association with other ageing traits, using PhenoScanner8, an online tool which searches 88 complex trait GWAMAs and three GWAS catalogues. For the SNP in the LPA region, associations were found with blood lipids and coronary traits. For the SNP in the LPA region, we found associations with rheumatoid arthritis and Crohn’s disease. For the CHRNA3/5 region, we found associations with traits which associate with smoking behaviour: nicotine dependence, lung cancer, chronic obstructive pulmonary disease and schizophrenia. Finally, for the APOE region, we saw associations with Alzheimer’s disease, age-related macular degeneration, blood lipids, adiposity, cardiac and cognitive ageing traits (Supplementary Data 4).

Genetic correlation of complex traits with lifespan. We estimated the genetic correlation between 113 complex quantitative and disease susceptibility traits and lifespan using LD Score regression: 46 showed meaningful genetic correlations (rg) with lifespan (statistically significant, lrg > 0.15). The most strongly correlated with mortality were coronary artery disease (CAD) and cigarettes smoked per day, rg (SE) = 0.66 (0.05) and 0.58 (0.11), respectively. Those most negatively correlated were years of schooling and former vs. current smoker, rg (SE) = −0.47 (0.05) and −0.64 (0.09), respectively (Supplementary Fig. 4, Supplementary Data 5). Lung cancer, type 2 diabetes and insulin resistance also correlated relatively strongly with earlier mortality, while increased age at first birth, openness to experience (a personality trait reflecting curiosity vs. caution, determined by the questionnaire) and high-density lipoproteins (HDL) cholesterol were correlated with later death.

Given the similarity in definition of many traits (e.g., obesity classes) and the strong correlations between others, we clustered the 46 traits which showed a significant and meaningful rg into nine clusters. Positive genetic correlations with mortality for the clusters ranged from 0.68 (smoking) to 0.17 (rheumatoid arthritis and breast cancer), whilst negative correlations varied from −0.50 (education) to −0.15 (age at menarche); (Fig. 5, Supplementary Data 5). We found that the beneficial trait clusters for education and happiness group together, as do a core group of factors (obesity, dyslipidemia/waist-hip ratio (DL/WHR), type 2 diabetes, CAD and smoking) which show stronger correlation not only to mortality but also among each other, while albuminuria and blood pressure seem to form their own risk cluster. We next considered whether and to what extent the observed correlations between mortality and the trait clusters are mediated through other clusters, using partial correlations. In most cases, there was relatively little difference between correlations and partial correlations with mortality (Supplementary Table 1) and the direction of effects remained the same. On the whole, the correlation of each risk cluster is therefore not mainly mediated via other clusters. However, the entire correlation of the DL/WHR cluster with lifespan was 0.41, whereas its partial correlation was −0.18, implying that one or more of the other clusters influenced the genetic correlation, likely CAD with which it is strongly correlated and whose partial correlation did not fall in the same manner. Similarly, the entire correlation of the education cluster with lifespan fell from −0.50 to −0.18 as a partial correlation, in this case apparently due to mediation through smoking behaviour. Blood pressure and age at menarche also showed reductions in partial rg, to near zero for age at menarche, consistent with mediation by other traits.

Causal relationships with lifespan. Finally, we used MRbase and further summary statistics for breast cancer (BCAC11) and C-reactive protein (CHARGE-CRP12) made available to us to

---

Table 2 Four regions associated with lifespan at genome-wide significance and replication via proxy SNPs in CHARGE

<table>
<thead>
<tr>
<th>rsid</th>
<th>Gene</th>
<th>a1</th>
<th>Freq a1</th>
<th>N(000) parent</th>
<th>HR a1</th>
<th>SE</th>
<th>P-value</th>
<th>Years Proxy</th>
<th>r²</th>
<th>CHARGE P</th>
<th>Dir.</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs34831921</td>
<td>HLA-DQA1/DRB1</td>
<td>A</td>
<td>0.09</td>
<td>481</td>
<td>0.942</td>
<td>0.011</td>
<td>4.18 E-08</td>
<td>0.6</td>
<td>rs3129720</td>
<td>0.39</td>
<td>0.003</td>
</tr>
<tr>
<td>rs55730499</td>
<td>LPA</td>
<td>T</td>
<td>0.033</td>
<td>563</td>
<td>1.074</td>
<td>0.011</td>
<td>8.67 E-11</td>
<td>−0.7</td>
<td>rs10455872</td>
<td>0.97</td>
<td>0.002</td>
</tr>
<tr>
<td>rs8042849</td>
<td>CHRNA3/S</td>
<td>C</td>
<td>0.356</td>
<td>567</td>
<td>1.046</td>
<td>0.006</td>
<td>3.75 E-14</td>
<td>−0.4</td>
<td>rs9788721</td>
<td>0.98</td>
<td>0.951</td>
</tr>
<tr>
<td>rs429358</td>
<td>FOXP3</td>
<td>C</td>
<td>0.142</td>
<td>556</td>
<td>1.091</td>
<td>0.008</td>
<td>1.44 E-27</td>
<td>−0.9</td>
<td>rs6857</td>
<td>0.69</td>
<td>2E-20</td>
</tr>
</tbody>
</table>

+a1 the effect allele, CHARGE, CHARGE European GWAS for survivorship beyond age 90 vs. younger controls, CHARGE P, the p-value for the two-sided test of association between proxy and longevity, HR direction of effect of a1 in CHARGE "+" means long-livedness increasing, "−" means long-livedness decreasing, Freq, frequency, N(000) count (thousands of parents with lifespan and subject genotype information), HR, Hazard Ratio, P-value for the Wald test of association between imputed dosage for a1 and lifespan, Proxy, the closest proxy SNP in CHARGE, r² the linkage disequilibrium between the discovery SNP and its CHARGE proxy, in the 1000 genomes EU panel, SE, Standard Error, Years the number of additional years of lifespan expected for a carrier of one additional copy of a1. There are four overlapping cohorts between the two studies; EGCUT, NTR, PROSPER and RS1, but only RS1 contributed cases to the CHARGE: out of all 5406 cases analysed in CHARGE, 892 cases (from RS1) overlapped the 300,000 genotyped subjects studied in discovery and the genotyped individuals were in any case not the same.

---

No statistically significant (p > 0.22, two-sided test) evidence of association was found for the other 10 genes. In all cases (with the possible exceptions of ABO and 5q33) our estimates of the odds ratio were close to 1 and our 95% CI did not include previous estimates, suggesting, at least for the remaining 8 SNPs (at or near CAMK4, C3orf21, GRIK2, IL6, RGS7, CADM2, MINPP1 and ANKRD20A9P), that our non-replication did not arise solely from lack of power.

Consistent with our previous reports4, we found age-specific and sex-specific effects of the lead SNPs in the APOE and CHRNA3/5 loci. For APOE, the hazard ratio (SE) of the lead SNP was 1.07 (.01) for men and 1.13 (.01) for women, whereas for CHRNA3/5 it was 1.07 (.01) for men and 1.04 (.01) for women (Fig. 4a). Conversely, for APOE, hazard ratios stratified by age were 1.06 (.01) for ages 40−75 and 1.14 (.01) for ages 75+, whereas for CHRNA3/5 they were 1.08 (.01) for 40−75 and 1.03 (.01) for age 75+ (Fig. 4b), with similar patterns when stratifying by age and sex at the same time, (Fig. 4c), although the distinctions between men and women for CHRNA3/5 disappeared beyond age 75. For LPA, CDKN2B and SH2B3, there was no statistically significant evidence of age-specific or sex-specific effects, while the HLA and FOXP3 variants showed age but not sex-specific effects (Fig. 4a, b), with the HLA locus having a greater effect at younger ages (40−75) while, conversely, the FOXP3 locus had greater effect at older ages (75+).

We tested the four SNPs identified in the discovery phase (Table 2) for association with other ageing traits, using PhenoScanner8, an online tool which searches 88 complex trait GWAMAs and three GWAS catalogues. For the SNP in the LPA region, associations were found with blood lipids and coronary traits. For the SNP in the LPA region, we found associations with rheumatoid arthritis and Crohn’s disease. For the CHRNA3/5 region, we found associations with traits which associate with smoking behaviour: nicotine dependence, lung cancer, chronic obstructive pulmonary disease and schizophrenia. Finally, for the APOE region, we saw associations with Alzheimer’s disease, age-related macular degeneration, blood lipids, adiposity, cardiac and cognitive ageing traits (Supplementary Data 4).

---

Causal relationships with lifespan. Finally, we used MRbase and further summary statistics for breast cancer (BCAC11) and C-reactive protein (CHARGE-CRP12) made available to us to...

---

NATURE COMMUNICATIONS | DOI: 10.1038/s41467-017-00934-5 ARTICLE
perform two-sample Mendelian randomisation to investigate causal influences on lifespan. Of more than 90 tested phenotypes, seven risk factors (cigarettes smoked per day, HDL cholesterol, LDL cholesterol, fasting insulin, systolic blood pressure and CRP) and six disease susceptibilities (Alzheimer’s disease, breast cancer, CAD, ischaemic stroke, squamous cell lung cancer and type 2 diabetes) significantly associated with mortality (Table 3). Smoking causally reduced lifespan by 6.8 years for lifelong smoking of one pack of 20 cigarettes a day, BMI reduced life by 7 months per unit, while education causally increased lifespan by 11 months for each further year spent studying. In contrast to the genetic correlations (rg CRP: mortality = 0.35), genetically raised CRP seems to have a life-lengthening effect: 5.5 months of increased lifespan per log mg/L.

We compared the relative strengths of these different phenotypic effects on lifespan using a measure independent of scale: extrapolating the genetic effects across the interquartile phenotypic range. Variation in smoking and systolic blood pressure had the strongest causal life-shortening effects (5.3 and 5.2 years, respectively), followed by fasting insulin, body mass index and CAD, while years of education showed by far the most beneficial effect (4.7 years), when comparing the estimated effect of moving from the first to the third quartile of the phenotype distribution. Similarly, we estimate moving from the bottom to the top of the interquartile phenotypic range of CRP increases lifespan by 0.7 years.

**Discussion**

We replicated previous findings of genome-wide significant associations between longevity and variants at *CHRNA3/5* and *APOE* and discovered two further associations, at *LPA* and *HLA-DQA1/DRB1*, with replication of the further associations in a long-livedness study. We found no evidence of our lead SNPs at the *CHRNA3/5, LPA* and *HLA-DQA1/DRB1* loci associating with traits other than smoking behaviour, cardio-metabolism and rheumatoid arthritis, respectively, while finding more pleiotropy at *APOE*. We also robustly replicated previous work suggesting associations with longevity at *CDKN2A/B, SH2B3/ATXN2* and *FOXO3A*. We found no evidence of association between lifespan and the other 10 loci previously found to suggestively associate with lifespan, despite apparent power to do so. We showed strong negative genetic correlation between CAD, smoking and type 2 diabetes and lifespan, while education and openness to experience were positively genetically correlated. Using MR, we found that moving from the 25th to 75th percentile of cigarettes per day, systolic blood pressure, fasting insulin and BMI causally reduced lifespan by 5.3, 5.2, 4.1 and 3.8 years, respectively, and similarly moving from the 25th to 75th percentile of educational attainment causally extended lifespan by 4.7 years. Strikingly, we also found that increased CRP increases lifespan, as a causal effect, the reverse of its correlation.

Lipoprotein(a) is a spherical lipoprotein carrying cholesterol and triglycerides in the bloodstream. Variation in *LPA* has
been extensively studied\textsuperscript{14}, and found to influence cardiovascular disease\textsuperscript{15} and type 2 diabetes\textsuperscript{16}. A close proxy to our lead SNP (rs10455872, \(r^2 = 0.97\)) has been strongly associated with decreased Lp(a) size and increased Lp(a) plasma concentration and is one of the strongest predictors of coronary heart disease risk with an odds ratio of 1.7 per allele, consistent across populations\textsuperscript{17}, all suggesting that rs55730499 affects mortality by increasing Lp(a) levels and susceptibility to cardiovascular events.

The large major histocompatibility complex (MHC) encompasses \textit{HLA-DQA1/DRB1}. MHC class II genes encode components of the antigen-presenting apparatus and are the most polymorphic region of the human genome. Genes within the MHC have previously been associated with many autoimmune conditions and other traits, including psoriasis\textsuperscript{18}, rheumatoid arthritis\textsuperscript{19}, multiple sclerosis\textsuperscript{20} and T1D\textsuperscript{21}. In a recent informed

The \textit{FOXO3A} locus has been repeatedly reported by other studies\textsuperscript{3, 23} as associating with extreme longevity. Variant rs3800231, which exhibits the strongest association in our data, seems to exert its beneficial effect on people aged above 75 but may have a neutral, or deleterious effect at younger ages, supporting the consensus that \textit{FOXO3A} plays a putative role in extreme longevity and general health into old age. This contrasts our findings for the \textit{CHRHA3/S}, \textit{LPA}, \textit{HLA-DQA1/DRB1} loci, where effects appear to be specific to disease susceptibility, rather than general ageing. The \textit{CDKN2A/B} locus at 9p21 has previously been associated with CAD\textsuperscript{24}, while the missense allele rs3184504-T we identified within the \textit{SH2B3/ATXN2} locus has been previously associated with increased risk for type 1 diabetes\textsuperscript{25}, diastolic blood pressure\textsuperscript{26} and several autoimmune conditions\textsuperscript{27–29}.

The failure to replicate previous findings for lifespan increases at \textit{ABO} and 5q33.3/\textit{EBF1} may be due to a combination of limited power in our study, despite its size, and a degree of winner’s curse in previous findings. However, for \textit{CAMK4}, \textit{C3orf71}, \textit{GRK2}, \textit{IL6}, \textit{RGS7}, \textit{CADM2}, \textit{MINPP1} and \textit{ANKRD20A9P}, our findings appear inconsistent with the previous work, suggesting those findings were either false positive associations, or differences in effects are due to the differences between the types of lives studied by us and other studies.

The use of different cohorts from a diverse range of countries with common shared ancestry is common in GWAMA and potentially gives rise to heterogeneity in effect sizes, whatever the trait under consideration. However, a study of lifespan is perhaps particularly susceptible to such effects, as mean lifespans vary by cohort (Supplementary Data 2) and genetic effects might vary by environment. Nonetheless, such heterogeneity is not relevant under the null hypothesis (effect size \(= 0\) in all cohorts) and so will not have induced false positives. On the other hand,
In general the results of the MR analyses appear consistent with those of the LD score regression estimates. This might be expected since the main difference is that MR compares two phenotypes using just a small number of SNPs which the underlying GWAS were powered to find, and LD score regression uses the whole genome. Nevertheless, as a result the latter may indicate a shared heritable confounding factor, rather than a causal effect, which appears to be the case for our CRP results, as the measured effect of CRP on lifespan is in the opposite direction to the genetic correlation. CRP’s effects per se are not well understood, but our results lead us to speculate it may have a protective function, rising in the presence of disease, rather than causing it, despite observational associations with disease and consequent attempts to develop a drug to reduce it35. If true, this pattern is somewhat analogous to findings for the N-terminal fragment of pro-BNP, which is a protective molecule, but observationally positively associates with cardiac failure and adverse cardiovascular outcomes36. Our finding that a reduction in one BMI unit leads to a 7-month extension of life expectancy, appears broadly consistent with the recently published by the Global BMI Mortality Collaboration, where great effort was made to exclude confounding and reverse causality.37 We also found each year longer spent in education translates into approximately a year longer lifespan. When compared using the interquartile distance, risk factors generally exhibited stronger effects on mortality than disease susceptibility. Although both CAD and cigarette smoking show a very similar genetic correlation with lifespan, the measured effect of smoking is twice as large as that of CAD, perhaps because smoking influences mortality through multiple pathways.

Our results show that longevity is partly determined by the predisposition to common diseases and, to an even greater extent, by modifiable risk factors. The genetic architecture of lifespan appears complex and diverse and there appears to be no single genetic elixir of long life.

Methods

Genome-wide association. As is conventional in GWAMA, analysis was carried out locally at each cohort and then meta-analysed centrally. Initial phenotype and genotype quality control were carried out in accordance with local standards, with variants imputed to 1000 Genomes (typically phase 1, version 3). Cohort characteristics, including genotyping and imputation methods and summary statistics of the parental lives analysed are described in Supplementary Data 1 and 2. Study protocols were approved by the relevant committees for each of the local cohorts. Written informed consent was obtained from each participant in each study.

We conducted an association test between parental survival (age and alive/dead status) and offspring genotype. To do so, survival traits were transformed into residuals, permitting analysis as quantitative traits. To facilitate standardisation across the GWAS consortium, residuals for GWAS were calculated in accordance with the analysis plan set out below using a common R protocol distributed to all groups. These residual traits were then tested for association in a GWAS over the imputed SNP panel.

Parents who died before the age of 40 were excluded. Analysis was thus of survivorship beyond the age of 40. Association testing was conducted under the following Cox Proportional Hazards Model38:

$$h(x) = h_0(x)e^{\beta Z_1 \gamma_1 + \ldots + \gamma_k Z_k}$$

$h_0$ is the baseline, $\beta$ the hazard log, ratio associated with $X$ (the effect allele count) and $Z_1,\ldots,Z_k$ the other variables fitted i.e., subject sex, and the first 10 PCs of genetic structure along with each studies’ usual further covariates, such as batch or assessment centre.

Rather than fit the full model in one step, we calculated Martingale residuals of the Cox model (excluding $X$). Martingale residuals39 are:

$$\tilde{M}_i = \tilde{d}_i - \hat{N}_i(t)\tilde{e}_i Z_1 \gamma_1 + \ldots + \gamma_k Z_k$$

where $\tilde{d}_i$ and $\tilde{e}_i$ are the parent status (1—dead, 0—alive at assessment date) and age of the i'th individual, $\gamma_1 \ldots \gamma_k$ are effect estimates of $Z_1,\ldots,Z_k$. Where the allele counts, $X$, has an effect, $\tilde{M}_i$ has a linear association with it34.

However, although these residuals are associated proportionately with the hazard ratio and thus permit statistical hypothesis testing, their relationship with

---

**Fig. 3** Validation of associations reported elsewhere by lookup in LifeGen. A search of recent literature suggested the gene regions shown here were most likely to harbour associations with lifespan, beyond the four loci identified in Table 2, which are further explored in the Discussion. The most powerful LifeGen analysis (i.e., European ancestry, father and mother combined) was used for validation. The odds ratio (OR) for extreme long-livedness is presented for the reported life-shortening allele (i.e., the OR for long-livedness < 1) in the original study, but not necessarily in LifeGen. The LifeGen OR of being long-lived was estimated empirically on the assumption that the relationship between the LifeGen observed hazard ratio (HR) and the OR is stable across allelic effects, with APOE results from LifeGen and CHARGE-EU 90+ -6 being used to estimate the ratio of in HR to ln OR (~4.7). These estimates will only fully align with the published ORs if the shape of the effect on lifespan is similar to APOE, as is true under the proportional hazards assumption, nonetheless the pattern is suggestive. Further details are shown in Supplementary Data 3.
the hazard ratio depends of the (parent) population structure, in particular the proportion dead. The Martingale residuals were therefore scaled up by \( \lambda \) (proportion dead) separately for each parent gender, to give a residual trait with a 1:1 correspondence with the hazard ratio. This transformed trait was then tested for association with each SNP separately under the following (additive) model,

\[
P = \beta X + e
\]

where \( \beta \) is the size effect of the SNP (and an estimate of the HR) and \( X \) the non-reference allele count of the marker, with \( e \) being normally distributed and independent. Despite this efficient approach, runtimes in UK Biobank were still potentially onerous, so RegScan 0.2.25 was used there as it is ideally suited for multiple, residualised traits in large data sets.

For cohorts with significant relatedness, all but one subject amongst relatives with coefficient of kinship > 5% were excluded, to create a (smaller) unrelated population, in preference to conventional (potentially more powerful) mixed modelling among family based studies. This was done because the genomic relationship matrix among offspring does not precisely reflect the genetic covariance among parental traits. As an example, consider the offspring of two brothers: the correlation between genetic values of the father trait is 0.5, but of the mother trait is 0, while the Genetic Relationship Matrix (GRM) entry would be 0.25 in both cases. The GRM thus does not fully express covariance among (parent) trait values across subjects, so the smaller unrelated population was used. Exceptions to this were for CILENTO, ERF, GeneSTAR, MICROS and OGP, where it was impractical to exclude relatives, and mixed modelling was used.

After preparing GWAS results locally, cohorts submitted these to the central team for meta-analysis. The central meta-analysis was carried out in METAL, with QC following that of Easy-QC24, but sometimes more conservative, as follows. UK Biobank data were read into METAL25 first, standardising all subsequent input alleles to that imputation. SNPs with mismatching alleles in other GWAS were rejected. SNPs were removed from a cohort’s GWAS if the minor allele frequency for that cohort was < 0.01. As all studies had in excess of 500 lives, this meant that minor allele count exceeded 10 Alleles with an info score (observed variance in dosage/expected under HWE) < 0.3 were excluded. Each GWAS was checked for systematic errors in allele coding/frequencies and test statistics for SNPs passing QC. After QC, SNP counts were 13,699,988 for European fathers, 13,643,373 for European mothers, 20,305,364, for African fathers and 20,296,065 for African mothers.

African and European ancestries were meta-analysed separately, as were the results for each parental sex, using inverse variance meta-analysis in METAL. Double genomic control was applied. The median \( \lambda \) for 78 GWAS was 0.998 and the maximum was 1.048, suggesting good control for stratification. The highest \( \lambda \) was for UK Biobank—genomically British, the most powered study. After the first level of genomic control, results were meta-analysed by inverse variance, while keeping continental ancestry separate and parental sex separate. The \( \lambda \) applied was 1.034, 1.023, 1.027, 1.028, for European fathers, mothers, African fathers, mothers, respectively. Finally, within continent across parent inverse variance meta-analysis was applied. As expected, due to environmental correlation among spouses, there was some inflation: \( \lambda \) of 1.107 and 1.094, for Europeans and Africans, respectively, giving two final combined meta-analyses (African and European) for both parents combined, subject to double genomic control.

These GWAS results were of the observed effect of offspring genotype on parent phenotype. The actual effect of carrying an allele for the individual concerned (rather than their parent) is twice that observed in a parent-offspring kin-cohort study. All reported effect sizes throughout this manuscript were therefore doubled to give the estimated effect size in the allele carriers themselves. The effect of hazard ratios on lifespan was calculated from survival curves of the Cox model by each cohort. The weighted average effect of hazard ratio on lifespan across all cohorts and both sexes was that a 1% reduction in hazard extended expected lifespan by 0.108 years. To avoid an undue sense of precision, and in accordance with an actuarial rule of thumb, where applicable, hazard ratios were converted to estimated effects on lifespan using a 10% HR: –1 year of lifespan ratio.

Genome-wide signficant European lead SNPs at each QTL were then looked up in the largest independent GWAS of lifespan with published summary statistics, for survivorship beyond age 90 vs. younger controls (CHARGE-EU 90+). None of the lead SNPs were present in that dataset, so proxy SNPs in strongest LD were chosen using LDlink43, with European populations selected. The SNP showing the highest \( r^2 \) with each LifeGen lead SNP was extracted from the CHARGE GWAS. The Rotterdam study was part of both GWAMAs, but the trait measured was in different people. In our study, we considered the lifespan of parents, whereas the longitudinal analysis was in the offspring. The LifeGen and CHARGE-EU 90+ GWAMAs were then meta-analysed using \( p \)-values and direction of effect (after reversing the sign of effect for CHARGE to convert longevity to mortality) with equal weights placed on each GWAMA, using METAL. The choice of equal weights was made, rather than weights reflecting sample size, because (i) the CHARGE extreme case-control approach is more powerful per sample than parent lifespan Cox modelling, and comparison of \( r \) is not straightforward, (ii) the \( Z \)-test statistics for \( r^2 \) (the most significant SNP overlapping in both studies) were similar: 9.4 and 9.5 for CHARGE and LifeGen, respectively, for the same \( r \), indicating similar overall power.

We used PhenoScanner8 to search for other trait associations with our lead SNPs. PhenoScanner settings were: rid, Catalogue = GWAS, \( p \)-value cutoff = 0.001, proxies = 1000 G, \( r = 0.6 \).

A review of recent literature was conducted for SNPs that have been associated with longevity and lifespan by other researchers, to see if we could validate their results. Nine papers published since 2008 were selected: Broer et al.4, Deelen et al.4, Emanuele et al.44, Flachsbart et al.45, Fortney et al.22, Malovini et al.46, Newman et al.47, Wilcox et al.23 and Zeng et al.48. Variants with MAF above 1% in 1000 Genomes, from these papers were taken forward, if they exhibited genome-wide significance (\( p < 5 \times 10^{-8} \)) or they had suggestive associations that were replicated.
Fig. 5 Genetic correlations between trait clusters that associate with mortality. The upper panel shows whole genetic correlations, the lower panel, partial correlations. T2D, type 2 diabetes; BP, blood pressure; BC, breast cancer; CAD, coronary artery disease; Edu, educational attainment; RA, rheumatoid arthritis; AM, age at menarche; DL/WHR Dyslipidaemia/Waist-Hip ratio; BP, blood pressure.
In aggregate, 18 variants in 13 gene regions were identified, of which four were genome-wide significant in the original study, while nine were suggestive (Fig. 3). These lead SNPs were then looked up in our results and compared with the previously reported associations (Supplementary Data 3 and Fig. 3).

### Table 3 Mendelian randomisation associations for the 19 traits with lifespan

<table>
<thead>
<tr>
<th>Exposure</th>
<th>SNPs in the IV</th>
<th>Beta</th>
<th>SE</th>
<th>P-value</th>
<th>Egger pleiotropy P</th>
<th>SD</th>
<th>Years per exposure unit</th>
<th>Interquartile effect in years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Body mass index SD (kg/m²)</td>
<td>65</td>
<td>0.279</td>
<td>0.04</td>
<td>2.26 × 10⁻¹²</td>
<td>0.4</td>
<td>4.77</td>
<td>0.584</td>
<td>3.8</td>
</tr>
<tr>
<td>Years of schooling SD (years)</td>
<td>64</td>
<td>−0.348</td>
<td>0.054</td>
<td>9.42 × 10⁻¹¹</td>
<td>0.039</td>
<td>3.71</td>
<td>−0.937</td>
<td>−4.7</td>
</tr>
<tr>
<td>Cigarettes smoked per day</td>
<td>rs12914385</td>
<td>0.034</td>
<td>0.005</td>
<td>6.47 × 10⁻¹⁰</td>
<td>−</td>
<td>11.7</td>
<td>0.338</td>
<td>5.3</td>
</tr>
<tr>
<td>HDL cholesterol SD (mg/dL)</td>
<td>39</td>
<td>−0.106</td>
<td>0.044</td>
<td>0.017</td>
<td>0.793</td>
<td>15.5</td>
<td>0.068</td>
<td>−1.4</td>
</tr>
<tr>
<td>LDL cholesterol SD (mg/dL)</td>
<td>17</td>
<td>0.101</td>
<td>0.042</td>
<td>0.017</td>
<td>0.82</td>
<td>38.7</td>
<td>0.026</td>
<td>1.4</td>
</tr>
<tr>
<td>Fasting insulin log pmol/L</td>
<td>6</td>
<td>0.389</td>
<td>0.176</td>
<td>0.027</td>
<td>0.823</td>
<td>0.79</td>
<td>3.89</td>
<td>4.1</td>
</tr>
<tr>
<td>SBP mmHg</td>
<td>rs381815</td>
<td>0.02</td>
<td>0.009</td>
<td>0.031</td>
<td>−</td>
<td>18.9</td>
<td>0.204</td>
<td>5.2</td>
</tr>
<tr>
<td>CRP log mg/L</td>
<td>39</td>
<td>−0.046</td>
<td>0.021</td>
<td>0.033</td>
<td>0.073</td>
<td>1.08</td>
<td>−0.458</td>
<td>−0.66</td>
</tr>
<tr>
<td>DBP mmHg</td>
<td>3</td>
<td>0.029</td>
<td>0.015</td>
<td>0.056</td>
<td>0.248</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Omega-3 fatty acids (SD)</td>
<td>rs145717049</td>
<td>−0.229</td>
<td>0.182</td>
<td>0.208</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total cholesterol SD (mg/dL)</td>
<td>11</td>
<td>0.036</td>
<td>0.068</td>
<td>0.597</td>
<td>0.348</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Triglycerides SD (mg/dL)</td>
<td>18</td>
<td>0.034</td>
<td>0.093</td>
<td>0.72</td>
<td>0.185</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apolipoprotein B (SD)</td>
<td>3</td>
<td>0.013</td>
<td>0.067</td>
<td>0.846</td>
<td>0.918</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disease susceptibility</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alzheimer’s disease</td>
<td>18</td>
<td>0.035</td>
<td>0.013</td>
<td>0.009</td>
<td>0.783</td>
<td>−</td>
<td>−</td>
<td>0.77</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>109</td>
<td>0.034</td>
<td>0.007</td>
<td>7.11 × 10⁻⁶</td>
<td>0.318</td>
<td>−</td>
<td>−</td>
<td>0.74</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>26</td>
<td>0.13</td>
<td>0.02</td>
<td>3.22 × 10⁻¹¹</td>
<td>0.125</td>
<td>−</td>
<td>−</td>
<td>2.9</td>
</tr>
<tr>
<td>Ischaemic stroke</td>
<td>rs4984814</td>
<td>0.012</td>
<td>0.003</td>
<td>1.39 × 10⁻⁵</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>0.26</td>
</tr>
<tr>
<td>Squamous cell lung cancer</td>
<td>2</td>
<td>0.073</td>
<td>0.03</td>
<td>0.014</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>1.6</td>
</tr>
<tr>
<td>Type 2 diabetes</td>
<td>22</td>
<td>0.036</td>
<td>0.015</td>
<td>0.02</td>
<td>0.247</td>
<td>−</td>
<td>−</td>
<td>0.79</td>
</tr>
</tbody>
</table>

The 19 traits which were significant in the first step analysis are shown. Exposure, list of exposures tested (for traits in which the betas in the original GWAS were expressed in standard deviations, SD has been added after the name of the exposure). Abbreviations/definitions: SNPs in the IV, the number of variants in the instrumental variable, or the identity of the SNP if <2. Beta, effects of exposure on lifespan expressed as the log hazard ratio of the Cox model, i.e., parent/offspring effect sizes have been doubled. For traits analysed in SD units, the betas refer to a variation of one standard deviation. CRP, C-reactive protein; DBP, diastolic blood pressure; HDL, high-density lipoprotein; LDL, low-density lipoprotein; SE, the standard error of beta. Egger pleiotropy P refers to the p-value and effect estimate. We inferred standard errors, -values were directly apparent, we also wished to compare other traits using our both parents European ancestry GWAMA summary statistics provided to us. Each of these was run through the LDHub server in order to estimate the genetic correlations with the other traits while the genetic correlation with lifespan was estimated by using a local run of LD score regression. The Benjamini and Hochberg multiple correction test procedure was applied to determine the statistical significance of the resulting genetic correlations.

We then defined three categories of traits: (a) Meaningfully genetically correlated to mortality if estimated rg > 0.15 and FDR < 0.05; (b) Not meaningfully genetically correlated to mortality if 95% CI for rg C [−0.15,0.15]; and (c) Otherwise, insufficient evidence.

After subsetting to only those meaningfully genetically correlated to mortality, we estimated all genetic correlations among those traits; some pairs of traits showed very high correlations. For example, many were genetically correlated to BMI and obesity, we thus used the ICLUST clustering algorithm to cluster the most similar ones. The number of clusters was chosen empirically, by visual inspection. The ICLUST algorithm from the psych R package clusters items hierarchically based on the loading of the items on the factors from factor analysis. Two clusters are then merged together only if by their joining their internal consistency increases. As rotation matrix for the factor analysis we used “promax” which is a high efficiency algorithm which allows correlation between the different factors. Other than to define the initial list, mortality was not included in the clustering analysis. At the same time, some highly correlated traits, which the clustering algorithm sought to combine, appeared to capture distinct clinical aspects and these were therefore kept separate. In particular, we split an education/smoking/
rheumatoid arthritis group into three; we separated CAD out from the cluster for dyslipidaemia and waist-to-ratio (DLW/WHR) and we separated breast cancer from age at menarche. The 12 clusters (and their constituent traits) were as follows: obesity (body mass index, body fat, childhood obesity, extreme BMI, obesity class 1, obesity class 2, obesity class 3, overweight, hip circumference, lep- tin_m_ad, adiposity, waist circumference & CRP), smoking (cigarettes smoked per day, former vs current smoker and lung cancer), DLW/WHR (fasting insulin, extreme waist-to-ratio, HDL cholesterol, insulin resistance (from homeostasis modelling assessment), triglycerides and waist-to-ratio), kidney (urinary albumin-to-creatinine ratio and urinary albumin-to-creatinine ratio-non-diabetics), type 2 diabetes (child birth weight, type 2 diabetes, fasting glucose and haemoglobin A1c), blood pressure (systolic and diastolic), happiness (depressive symptoms and subjective well-being), breast cancer, CAD, educational attainment (years of schooling, openness to experience from NEO personality inventory), rheumatoid arthritis and age at menarche.

The resulting correlation matrix amongst clusters was used to estimate partial genetic correlations between the clusters using the matrix inversion method as implemented in the corporc R package (ISBN: 978-0-470-74366-9).

Mendelian randomisation. As a further step to identify which traits affect, rather than merely correlate with, mortality and to determine how much they shorten or lengthen lifespan, we performed a multiple step two-sample Mendelian randomisation (MR) study using summary statistics. We first identified a list of 96 candidate phenotypes selected amongst diseases and disease-associated risk factors (Supplementary Table 2) for which genome-wide association data were publicly available as part of the MRbase package. Data were available for more than one GWAS for a given trait, we selected those which had the largest sample size, were performed on both sexes and were performed in both European or Mixed descent samples. To this list we added other GWAS which were not present in MRbase: diastolic and systolic blood pressure, C-reactive protein and breast cancer. For each of the selected traits, instrumental variables were constructed starting from all SNPs with \( p < 5 \times 10^{-8} \). We then performed LD clumping (10 \( r^2 = 0.1, \) window = 10 Mb) in order to prune all non-independent SNPs. Some traits had no SNPs below the significance threshold and were thus excluded.

MR was performed using the inverse variance method utilising each of the selected traits as exposures and mortality as outcome. Where the instrument for the trait was composed of a single SNP, we used the Wald ratio instead. We then verified the absence of directional pleiotropy using MR Egger regression, but none of the candidate traits showed statistically significant evidence of pleiotropy once corrected for multiple testing (Supplementary Table 2).

Having already corrected for FDR at the previous step, no further adjustment was made for multiple testing in Table 3.

Several traits associated with BMI and obesity were extremely redundant: we thus removed obesity class 1, obesity class 2, obesity class 3, overweight, extreme body mass index, hip circumference, and childhood obesity. Finally, myocardial infarction was removed since 20 of the 22 SNPs composing its instrument were also in the CAD instrument. Supplementary Table 3 summarises the number of SNPs comprising each instrument before and after pruning.

For each significant trait, we estimated the difference in expected years of life between the 25th and 75th percentile phenotypic percentiles. For normally distributed traits this difference corresponds to 1.345 phenotypic standard deviations. For binary traits, the variance of the logistic distribution is constant, this difference instead corresponds to 1.345 phenotypic standard deviations. For binary traits, the variance of the logistic distribution is constant, this difference instead corresponds to 1.345 phenotypic standard deviations.