nature medicine

SUPPLEMENTARY INFORMATION

https://doi.org/10.1038/s41591-019-0710-1

In the format provided by the authors and unedited.

No statistical evidence for an effect of *CCR5*-Δ32 on lifespan in the UK Biobank cohort

Robert Maier[1,2,7*], Ali Akbari[1,2,7*], Xinzhu Wei[3], Nick Patterson[2,4], Rasmus Nielsen[3,5] and David Reich[1,2,4,6]

[1]Department of Genetics, Harvard Medical School, Boston, MA, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [3]Department of Integrative Biology and Statistics, University of California, Berkeley, Berkeley, CA, USA. [4]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA. [5]GeoGenetics Centre, University of Copenhagen, Copenhagen, Denmark. [6]Howard Hughes Medical Institute, Harvard Medical School, Boston, MA, USA. [7]These authors contributed equally: Robert Maier, Ali Akbari. *e-mail: rmaier@broadinstitute.org; Ali_Akbari@hms.harvard.edu

Supplementary Information

Markers tagging CCR5-∆32

Genotype data in the UK Biobank is available in three different forms: (1) Allele counts as inferred from the genotyping array intensity values; (2) Imputed genotype dosages which are commonly rounded to best guess allele count integer values. These are based on the array data genotype calls but for many genotyped variants are not equal to the array data genotype calls; and (3) Whole exome sequencing data, currently for a pilot sample of around 10% of the total sample size. Two different pipelines were used to call variants from the read data. Here we only use the variant calls from the GATK pipeline.

We analyze five variants in total, two from the array data, two imputed and one sequenced (Supplementary Table 1):

rs62625034: This is the genotyped variant which has been used as a proxy for the CCR5-∆32 deletion in Wei and Nielsen[1].

rs113010081: A genotyped variant in high LD to the CCR5-∆32 deletion.

rs113010081_imputed: The imputed data for the same variant.

3:46414943_TACAGTCAGTATCAATTCTGGAAGAATTTCCAG_T: The CCR5-∆32 deletion as called in the imputed data. For brevity, we refer to it as **rs333_imputed**, even though this rs ID is not used in the raw data.

3:46373452:D:32: The CCR5-∆32 deletion as called in the exome sequencing data. rs62625034 is not present among the set of imputed SNPs. We refer to it as **rs333_sequenced**, even though this rs ID is not used in the raw data.

Concordance rates across variants

As the genotyped and imputed variants do not directly target the Δ32 deletion, we treated the exome sequencing data CCR5-Δ32 variant itself (rs333_sequenced) as the ground truth. We then assess the accuracy of the genotype array variants by comparing them to the exome sequencing variant. Figure 1, Supplementary Figures 1 and 2, and Supplementary Tables 2 and 3 show concordance rates, sensitivity, and specificity of the genotyped variants and the exome sequencing variant.

While rs113010081 has a higher Pearson correlation coefficients (r^2) with rs333_sequenced than rs62625034 (0.977 compared to 0.968, Supplementary Table 3), these r^2 values are mostly influenced by the concordance of the more common genotypes Δ32/+ and +/+. As we are specifically interested in Δ32/Δ32 individuals and for the purpose of the present analysis are less concerned by misclassification of the two other much more common genotypes, we also computed sensitivity and specificity based on a comparison of Δ32/Δ32 individuals to the union of Δ32/+ and +/+ individuals. The sensitivity and specificity to correctly identify individuals with Δ32/Δ32 in the WES data is 0.934 and 0.998 for rs62625034, and 0.998 and 1 for rs113010081. In addition, out of all individuals identified as Δ32/Δ32 by the WES data, 11.4% are classified as Δ32/+ at rs62625034, compared to 3.3% at rs113010081 (Figure 1). This suggests that rs113010081 more accurately tags CCR5-Δ32 deletion than rs62625034 (Supplementary Table 3).

Supplementary Table 2 shows conditional genotype counts for all individuals, as well as for only those individuals genotyped on the UK Biobank Axiom array. We observed differences in missingness between the two array types, but no relative differences in genotype counts. Other Supplementary Tables only show results from both arrays, as these numbers change very little when restricting to samples genotyped on the UK Biobank Axiom array.

In this work, we do not focus on the imputed variants, as they do not tag the Δ32 deletion as well as the genotyped variants (Supplementary Figure 2 and Supplementary Table 3). In addition, Supplementary Table 4 shows that imputation quality differs by genotype at rs11301008_imputed.

Hardy-Weinberg disequilibrium

As population heterogeneity can induce deviations from HWE, we limit all of our analyses to individuals classified as “white British” in the UK Biobank. We do not exclude related individuals for the results shown here, though our results remain qualitatively the same when excluding related individuals. We compute approximate HWE p-values using a Chi-squared test. To be consistent with Wei and Nielsen, we also compute HWE deviation p-values in two alternative ways: First, we compute P1, which measures where the B-statistic (observed/expected ratio of the rare homozygous genotype) falls relative to the distribution of frequency matched control SNPs. This corrects for the fact that different data sets have different average deviations from HWE due to the Wahlund-effect and due to differences in genotype calling data and algorithms. This test tends to be more conservative than the Chi-squared test. Second, we compute P2, which tests whether the B-statistic of a given variant falls below the median B-statistic of the frequency matched control SNPs (Supplementary Table 5). This test was argued to be the preferred test as it provides some protection against outliers. The null-hypothesis of this test is that a given SNP has a B-statistic equal to or greater than the median across all control SNPs, and so this test is less conservative than a Chi-squared test.

For rs333_sequenced, the P2 p-value is 0.0276, similar to the previously reported value of 0.0272. For rs62625034 and rs113010081, P2 is < 0.0001 and 0.0023, respectively. P1 is 0.0032 for rs62625034, but not significant for the other SNPs. When subsetting to the samples for which we have exome sequencing data, P1 and P2 remain qualitatively similar, however P1 for rs113010081 is 0.0242.

Computing p-values corrected for missingness

In order to compute HWE p-values which are corrected for the differential missingness at rs333_sequenced, we computed the number of counts expected in each genotype class if missingness was independent of $\Delta 32$ genotype in the sequencing data. We then computed HWE Chi-squared p-values on those expected counts (Supplementary Table 6).

Probe design may cause differential missingness at ∆32

Figure 1 provides a plausible explanation for why rs62625034 exhibits higher missingness rates in individuals with the ∆32 deletion. The Affymetrix probe for rs62625034 is targeting a very rare G>T SNP which is located at the 3' end of the site of the ∆32 deletion. Since this variant is rare, almost all of the called non-reference alleles indicate the presence of the ∆32 deletion, which at its 3' end closely resembles the targeted G>T SNP. Since the probe overlaps with the ∆32 deletion but matches it only imperfectly, ∆32 individuals have a higher missingess rate. In contrast, the probe for rs113010081 is 42 kb downstream of ∆32 and suffers from no such problems.

Simulating the effect of sample ascertainment on HWE at two SNPs in high LD

We carried out a simulation study to test whether increased mortality or other negative ascertainment on ∆32/∆32 individuals can plausibly create a highly significant HWE deviation at this deletion, but no HWE deviation at a SNP with an r^2 of 0.95 relative to the deletion. We find that ascertainment on one variant induces similarly high deviations from HWE at other variants in high LD (Supplementary Figure 3). Thus, if one variant shows a high degree of deviation from the null expectation of HWE, and another variant in high linkage disequilibrium with it shows no significant deviation from HWE, it is highly likely that a technical artifact is affecting the genotyping of at least one of the variants.

Survival rate analysis

To study the effect on survival rates, we extend the phenotypic association analysis by exploring the effects of all variants tagging the CCR5-∆32 deletion on all phenotypes available to us.

For each of the variants, we assess the impact on mortality as previously described in Wei and Nielsen[1,2]. We use five different UK Biobank variables - age at recruitment (ID 21022), Date of attending assessment centre (ID 53), year of birth (ID 34), month of birth (ID 52), and the age at death (ID 40007) - to compute the number of individuals who are ascertained from age i to age i + 1 (N_i), and the occurrence of death observed from these N_i individuals during the interval of age i to age i + 1 (O_i). The death rate per year is calculated separately for each ∆32 genotype class as $h_i = \frac{O_i}{N_i}$ and the probability of surviving to age i + 1, $S_i = \prod_{n=1}^{n=i} h_\mathrm{n}$. h_{77} is grouped together with h_{76}.

To compute p-values for the survival rate analysis, we run Cox proportional hazard models using the 'coxph' function in the R-package 'survival'. We do not use binning into age groups, as described in the previous paragraph, for this analysis. Instead we use only age at recruitment and reported age at death or, if no age at death is reported, the inferred age at time t, where t is the date of the last reported age at death in the entire cohort (16 February 2016).

Survival rate power calculation

We estimate the power to detect effects on mortality rate in the following way. First, we extract for each sample age at death, or, if age at death has not been reported, the inferred age at time t, where t is the date of the last reported death in the entire cohort (16 February 2016). Next, we randomly draw a genotype (0 or 1) for each person from a Bernoulli distribution with a probability that depends on whether or not this person has died, in proportion to a given relative risk (RR). For individuals who have died, this probability is P(G=1|D) = P(D|G=1) * P(G=1) / P(D), where P(G=1) is the frequency of Δ32/Δ32 (0.012), P(D) is the fraction of samples with a reported age at death (0.029), and P(D|G=1) = RR * P(D|G=0) = RR * P(D|G=0) * P(D) / (P(G=1)*P(D|G=1) + (1-P(G=1)) * P(D|G=0)) = RR * P(D) / (P(G=1)*RR + (1-P(G=1))). Similarly, for individuals who are still alive, this probability is P(G=1|A) = P(A|G=1) * P(G=1) / P(A), where P(A) = 1 - P(D) and P(A|G=1) = 1 - P(D|G=1). We then obtain a p-value from a Cox proportional hazard model for each random draw, repeat this 100 times for 9 different RR values, and compute the fraction of random draws with p-value smaller than 0.05 at each value of RR.

Genotyping array - missingness batch effect the UK Biobank

We find that samples with missing genotypes at rs113010081 show greatly increased mortality rates (p-value $2.7\text{x}10^{-32}$). This is a genotyping batch effect: rs113010081 is absent from the UK BiLEVE Axiom array, and the individuals who were genotyped on this array were ascertained to be smokers[3]. This association disappears when restricting to individuals genotyped on the UK Biobank Axiom array. The same sample restriction does not explain the increased mortality rate seen for two carriers of the rare allele in rs62625034 (though the p-value increases to 0.016), but this example cautions against reporting associations between variants from the array data and mortality without controlling for possible genotyping array batch effects. We have only observed these batch effects in the array data, but not in the imputed data. Further, we only observed differences in missingness rates between the two array types, but no differences in the relative proportion of called genotypes (Supplementary Table 2).

Associations with other phenotypes in the UK Biobank

If a genetic variant has a substantial effect on early mortality then that effect is likely to act through specific phenotypes. We therefore tested whether Δ32/Δ32 individuals were at higher risk for 3,331 diseases or disorders than Δ32/+ and +/+ individuals. We tested each of the five variants for associations with 3,911 phenotypes in the UK Biobank. We used the following logistic regression model: $y \sim x_{01,2} + c$. Here, y is a vector of phenotypes; $x_{01,2}$ is the vector of genotypes, recoded so that each sample with zero or one copy of the deletion is 0 and each sample with two copies of the deletion is 1; and c is a set of covariates, including age, sex, genotyping array, and PC 1 to PC 20, calculated on a set of European individuals[4]. We similarly tested an additional 580 continuous phenotypes using a linear regression model.

Associations with other phenotypes - results

"Lymphocyte count" is the only trait which is significant at a p-value smaller than the classic threshold for declaring genome-wide statistical significance, $5x10^{-8}$. However, it can be argued that the genome-wide significance threshold is too stringent, since we only test the effect at one locus. When we instead apply Bonferroni multiple testing correction for 3,911 tested phenotypes, we find one additional phenotype, "Mean sphered cell volume", which is associated at a p-value smaller than $1.27x10^{-5}$ (which corresponds to 0.05 after Bonferroni correction for 3,911 phenotypes; Supplementary Table 8, Supplementary Figures 5 and 6). A large number of traits are nominally significant at $p < 0.05$, which is consistent with the null-hypothesis of no effect. The phenotype "Overall health rating" is associated with rs113010081 at a nominal p-value of $5.22x10^{-3}$. On average, Δ32/Δ32 individuals are 7% more likely to rate their health as "poor" or "fair" compared to other individuals (Supplementary Table 9). We also obtain p-values of $4.47x10^{-3}$ and

5.74×10^{-3} for two collections of diagnosis codes described as "Certain infectious and parasitic diseases"[5]. Given that ∆32/∆32 has previously been reported to be a risk factor for symptomatic West Nile virus infection[6], this is noteworthy.

We single these phenotypes out because they relate to previously reported effects of ∆32/∆32, but we highlight that we tested almost 4,000 phenotypes. Many other phenotypes with more significant nominal p-values seem unrelated to any relevant health outcomes, and the false positive rate in this set of traits is likely very high. Despite the large overall sample size, many disease phenotypes are rare, which further limits the power to detect effects of a genotype present in only 1% of the population at a reasonable significance level.

Cause of death

We conducted Poisson tests to check whether any ICD10 diagnosis codes were overrepresented as the reported cause of death in ∆32/∆32 compared to all other individuals. We find no ICD10 codes which are overrepresented in ∆32/∆32 individuals compared to all other individuals, but similar power considerations as in the survival rate analysis apply here.

Acknowledgements

This research has been conducted using the UK Biobank Resource under Application Number 31063. We acknowledge the participants in the UK Biobank. We are grateful to Benjamin Neale and Alkes Price for critical comments, and to Sean Harrison for a blog posting that showed how the association results and HWE p-values at rs113010081 were qualitatively discordant with those at rs62625034, which prompted us to re-examine these issues. We thank Konrad Karczewski, Kári Stefansson and Mark Daly for sharing with us early versions of two other manuscripts re-examining the evidence of association to mortality at *CCR5*, and working with us to post all manuscripts together. This work was funded in part by NIH grants GM100233 and HG006399, the Paul Allen Family Foundation, the John Templeton Foundation grant 61220, and the Howard Hughes Medical Institute.

References

1. Wei, X. & Nielsen, R. CCR5-∆32 is deleterious in the homozygous state in humans. Nature Medicine 25, 909–910 (2019).

2. Wei, X. & Nielsen, R. Deviations from Hardy Weinberg Equilibrium at CCR5-Δ32 in Large Sequencing Data Sets. bioRxiv 768390 (2019). doi:10.1101/768390

3. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209 (2018).

4. UK Biobank — Neale lab. Neale lab Available at: http://www.nealelab.is/uk-biobank. (Accessed: 24th September 2019)

5. Clinical endpoints | FinnGen. Available at: https://www.finngen.fi/en/researchers/clinical-endpoints. (Accessed: 28th September 2019)

6. Glass, W. G. et al. CCR5 deficiency increases risk of symptomatic West Nile virus infection. J. Exp. Med. 203, 35–40 (2006).

Supplementary Tables

Variant ID in this study	Variant ID in UK Biobank	type	GRCh37 position	alleles	non-missing	MAF
rs62625034	rs62625034	genotyped	46414975	T/G	395,656	0.116
rs113010081	rs113010081	genotyped	46457412	C/T	364,602	0.118
rs113010081_imputed	rs113010081	imputed	46457412	C/T	408,911	0.119
rs333_imputed	3:46414943_TACAGTCAGTATCAATTCTGGAAGAATTTCCAG_T	imputed	46414943	T/TACAGTCAGTATCAATTCTGGAAGAATTTCCAG	408,897	0.106
rs333_sequenced	3:46373452:D:32	sequenced	46414943	T/TACAGTCAGTATCAATTCTGGAAGAATTTCCAG	41,059	0.117

Supplementary Table 1: Variants tagging the CCR5-Δ32 deletion.

Variant	Allele count non-sequenced	Allele count sequenced (rs333_sequenced) All samples			Allele count sequenced (rs333_sequenced) Only UK Biobank Axiom array		
		0	1	2	0	1	2
rs62625034	0	30984	35	0	28212	33	0
rs62625034	1	129	8094	66	120	7370	56
rs62625034	2	0	27	380	0	25	351
rs62625034	NA	872	381	91	852	361	85
rs113010081	0	29127	126	0	29127	126	0
rs113010081	1	34	7658	16	34	7658	16
rs113010081	2	0	1	473	0	1	473
rs113010081	NA	2824	752	48	23	4	3
rs113010081_imputed	0	31671	279	0	28901	257	0
rs113010081_imputed	1	246	8200	37	217	7480	34
rs113010081_imputed	2	NA	42	500	0	37	458
rs113010081_imputed	NA	68	16	0	66	15	0
rs333_imputed	0	31696	1150	5	28918	1032	5
rs333_imputed	1	221	7347	156	200	6720	146
rs333_imputed	2	0	24	375	0	22	340
rs333_imputed	NA	68	16	1	66	15	1

Supplementary Table 2: Cross-tabulation of allele counts for genotyped variants tagging the CCR5-Δ32 deletion against rs333_sequenced.

Variant	P	N	TP	TN	Sensitivity (TP/P)	Specificity (TN/N)	r^2
rs62625034	407	39308	380	39242	0.934	0.998	0.968
rs113010081	474	36961	473	36945	0.998	1.000	0.977
rs113010081_imputed	542	40433	500	40396	0.923	0.999	0.930
rs333_imputed	399	40575	375	40414	0.940	0.996	0.818

Supplementary Table 3: Sensitivity and specificity of genotyped variants to distinguish Δ32/Δ32 from the other two genotypes (+/+ and Δ32/+) in the exome sequencing data. The last column is Pearson correlation coefficients (r^2) between variants and the CCR5-Δ32 deletion across all genotype classes (+/+, Δ32/+, Δ32/Δ32).

	rs113010081 genotype	rs333_sequenced		
		Δ32/Δ32	Δ32/+	+/+
a: decimal dosage	C/C	373	36	0
	C/T	30	4024	161
	T/T	0	127	901
b: integer dosage	C/C	126	5	0
	C/T	7	4178	85
	T/T	0	152	30770

Supplementary Table 4: Genotype calls at rs113010081_imputed and rs333_sequenced in the UK Biobank White British. **a**, Individuals with imputed dosage (0,0.5] as C/C, (0.5,1.5) as C/T, and [1.5,2) as T/T. **b**, Individuals with imputed dosage 0 as C/C, 1 as C/T, and 2 as T/T. Notice the relative increase in Δ32/Δ32, Δ32/+ genotypes with decimal dosage (low confidence imputation) relative to integer dosage (high confidence imputation), and the relative large discrepancy between the exome sequencing data and imputation based genotyping data for decimal dosage genotypes. For example, within the class of genotypes with decimal dosage, 30/403 homozygous minor genotypes in the exome sequencing data are called as heterozygous in the UK Biobank decimal dosage imputation data, and 36/409 homozygous minor genotypes in the UK Biobank decimal dosage imputation data are called as heterozygous in the exome sequencing data.

	Chi-squared HWE p-values		P1 and P2 p-values (genomic control corrected)	
Variant	All samples	Samples with WES data	All samples	Samples with WES data
rs333_sequenced	0.22 (537, 562)	0.22 (537, 562)	N/A	0.0764, 0.0276
rs62625034	4.8e-51 (4348, 5317)	6.1e-09 (421, 540)	0.0032, < 0.0001	0.0022, < 0.0001
rs113010081	0.36 (4979, 5036)	0.23 (496, 520)	0.0941, 0.0023	0.0242, 0.0326
rs113010081_imputed	0.78 (5759, 5778)	0.48 (565, 580)	N/A	N/A
rs333_imputed	1.4e-05 (4301, 4563)	0.02 (416, 461)	N/A	N/A

Supplementary Table 5: HWE p-values for variants tagging the CCR5-Δ32 deletion. In brackets: observed and expected number of samples with two copies of the rare allele. We report Chi-squared p-values, and P1 and P2 p-values, which were used in the original study. The latter two tests attempt to correct for the Wahlund-effect and other genome wide effects, such as systematic genotyping errors, and are described in the Supplementary Information.

Variant	Observation from Genotyping data					Corrected Values			
	GT = 0	GT = 1	GT = 2	GT = NC	HWE *P*	GT = 0	GT = 1	GT = 2	HWE *P*
rs62625034	308,274	83,034	4,348	13,989	4.8E-51	318,295	85,683	5,668	0.25
rs113010081	283,877	75,746	4,979	45,043	0.36	318,088	85,835	5,722	0.43
rs113010081_imputed	317,457	85,695	5,759	734	0.78	317,764	86,194	5,687	0.07
rs333_imputed	326,808	77,788	4,301	748	1.4E-05	318,142	85,831	5,672	0.17

Supplementary Table 6: Correcting for bias can explain the extreme p-value for the violation of HWE for rs62625034. Unbiased genotype counts is the expected number of true genotypes conditioned on observations in the genotyping array data (including missing genotypes). Conditional distribution is estimated by the joint distribution of genotyping array and UK Biobank WES data. UK Biobank WES data is considered as the ground truth. This table includes all white British samples in the UK Biobank.

Variant	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76
rs333_sequenced	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	2	0	1	0	0	0	0	0	0	2	0	1	2	0	0	0	1	1	0	1	0
rs62625034	0	0	2	0	0	0	0	0	0	1	1	1	2	4	1	4	1	3	4	7	1	9	4	5	15	6	7	10	9	8	14	10	12	4	5	1
rs113010081	0	0	1	0	0	1	0	0	0	1	1	1	1	4	4	5	1	4	3	3	4	5	4	4	11	9	7	13	11	9	11	12	13	3	5	1
rs113010081_imputed	0	0	2	0	0	1	0	0	1	1	1	1	1	4	3	6	1	4	3	7	4	6	4	5	16	9	11	14	11	16	14	13	13	4	6	1
rs333_imputed	0	0	1	0	0	0	0	0	1	1	0	1	1	3	3	5	2	2	3	7	3	3	4	4	9	9	5	10	10	10	10	10	10	5	6	1

Supplementary Table 7: Number of samples who have died, for each variant and age group. Values correspond to the red dots in the third row of Supplementary Figure 1.

Phenotype ID	beta	SE	p-value	type	count	type	description
30120	0.087	0.015	1.27E-08	continuous	4251	continuous_irnt	Lymphocyte count
30270	-0.074	0.015	1.65E-06	continuous	4188	continuous_irnt	Mean sphered cell volume
30180	0.066	0.015	1.75E-05	continuous	4251	continuous_irnt	Lymphocyte percentage
30260	-0.066	0.015	1.88E-05	continuous	4188	continuous_irnt	Mean reticulocyte volume
670_3	1.030	0.264	9.44E-05	binary	15	binary	Type of accommodation lived in: Mobile or temporary structure (i.e. caravan)
5119	0.129	0.034	1.29E-04	continuous	881	continuous_irnt	3mm cylindrical power (left)
30050	-0.057	0.015	1.65E-04	continuous	4257	continuous_irnt	Mean corpuscular haemoglobin
L12_HIDRADENITISSUP	1.554	0.422	2.32E-04	binary	6	categorical	Hidradenitis suppurativa
30040	-0.056	0.015	2.35E-04	continuous	4257	continuous_irnt	Mean corpuscular volume
30190	-0.050	0.015	7.72E-04	continuous	4251	continuous_irnt	Monocyte percentage
20003_1140868408	0.605	0.180	7.90E-04	binary	32	*NA*	*NA*
V_PREGNANCY_BIRTH	-0.364	0.111	1.01E-03	binary	111	*NA*	*NA*
30300	0.050	0.015	1.05E-03	continuous	4188	continuous_irnt	High light scatter reticulocyte count
102280	0.078	0.024	1.29E-03	continuous	618	ordinal	Milk chocolate intake
F5_SOMATOFORM	1.084	0.341	1.46E-03	binary	9	*NA*	*NA*
2744	-0.088	0.028	1.66E-03	continuous	1874	ordinal	Birth weight of first child
6149_1	0.149	0.047	1.67E-03	binary	512	binary	Mouth/teeth dental problems: Mouth ulcers
4294_9	1.309	0.420	1.85E-03	binary	6	binary	Final attempt correct: abandon
30010	0.041	0.013	1.92E-03	continuous	4257	continuous_irnt	Red blood cell (erythrocyte) count
L12_SCARCONDITIONS	0.562	0.182	2.06E-03	binary	31	categorical	Scar conditions and fibrosis of skin
20003_1140922174	-0.578	0.191	2.50E-03	binary	28	binary	Treatment/medication code: alendronate sodium
20003_1140852948	0.541	0.180	2.63E-03	binary	32	binary	Treatment/medication code: calcium+vitamin d 500units tablet
KRA_PSY_ANXIETY	0.623	0.207	2.67E-03	binary	24	categorical	Anxiety disorders
30250	0.046	0.015	2.71E-03	continuous	4188	continuous_irnt	Reticulocyte count
20003_1141169520	0.962	0.323	2.88E-03	binary	10	binary	Treatment/medication code: cosopt 2%/0.5% eye drops

30000	0.046	0.015	2.92E-03	continuous	4257	continuous_irnt	White blood cell (leukocyte) count
30200	-0.046	0.015	3.10E-03	continuous	4251	continuous_irnt	Neutrophill percentage
103990	-0.291	0.099	3.25E-03	binary	483	binary	Vegetable consumers
30220	-0.045	0.015	3.32E-03	continuous	4251	continuous_irnt	Basophill percentage
CHRONLARGE	0.946	0.322	3.32E-03	binary	10	categorical	Crohn's disease of large intestine
30290	0.045	0.016	3.62E-03	continuous	4188	continuous_irnt	High light scatter reticulocyte percentage
20003_1140865564	1.099	0.386	4.41E-03	binary	7	binary	Treatment/medication code: imodium 2mg capsule
AB1_INFECTIONS	0.269	0.095	4.47E-03	binary	117	categorical	Certain infectious and parasitic diseases
AB1_OTHER_VIRAL	0.577	0.203	4.49E-03	binary	25	categorical	Other viral diseases
2316	0.107	0.038	4.51E-03	binary	923	binary	Wheeze or whistling in the chest in last year
2030	-0.100	0.035	4.63E-03	binary	1131	binary	Guilty feelings
20002_1077	-0.892	0.317	4.97E-03	binary	10	binary	Non-cancer illness code, self-reported: heart arrhythmia
2178	0.031	0.011	5.23E-03	continuous	4365	ordinal	Overall health rating
I_INFECT_PARASIT	0.239	0.087	5.74E-03	binary	140	categorical	Certain infectious and parasitic diseases
X_EXTERNAL_MORB_MORT	0.885	0.322	5.97E-03	binary	10	NA	NA
L12_ATROPHICSKIN	0.478	0.174	6.10E-03	binary	34	categorical	Atrophic disorders of skin
20003_1140862438	1.142	0.417	6.22E-03	binary	6	binary	Treatment/medication code: uniphyllin continus 200mg m/r tablet
1628	0.031	0.012	6.53E-03	continuous	4050	ordinal	Alcohol intake versus 10 years previously
30670-0.0	0.056	0.021	7.33E-03	continuous	4175	NA	NA
41231_2	-0.313	0.119	8.53E-03	binary	93	NA	NA
22601_51112476	1.098	0.419	8.73E-03	binary	6	binary	Job coding: farmer, farming contractor, herd manager, smallholder, bailiff
20003_1140883504	0.324	0.124	9.12E-03	binary	67	binary	Treatment/medication code: cetirizine
CHRONNAS	0.640	0.246	9.31E-03	binary	17	categorical	Crohn's disease NAS
M13_SYNOTEND	0.275	0.106	9.70E-03	binary	92	categorical	Disorders of synovium and tendon

20002_1157	0.831	0.322	9.75E-03	binary	10	binary	Non-cancer illness code, self-reported: non-infective hepatitis
20003_1140916282	-0.978	0.379	9.82E-03	binary	7	binary	Treatment/medication code: venlafaxine
20002_1113	0.306	0.119	9.85E-03	binary	74	binary	Non-cancer illness code, self-reported: emphysema/chronic bronchitis

Supplementary Table 8: Association results for rs113010081 showing phenotypes with p-value < 0.01. Phenotypes with p-value < $1.27x10^{-5}$ are significant after Bonferroni correction for 3,911 phenotypes. The count column lists the number of Δ32/Δ32 individuals who are cases (for binary phenotypes) or who have non-missing phenotype information (for all other phenotypes).

Overall health rating	Δ32/+ and +/+	Δ32/Δ32 observed	Δ32/Δ32 expected
Excellent	54436	676	751.59
Good	185795	2592	2569.13
Fair	62757	913	868.30
Poor	12720	184	175.98

Supplementary Table 9: Contingency table of self-reported health rating and Δ32 status inferred from rs113010081. The odds ratio of “Fair” or “Poor” health vs “Excellent” or “Good” health is 1.068. Adjusted and unadjusted p-values are 0.0052 and 1.