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

nature medicine

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]

ARISING FROM X. Wei & R. Nielsen *Nature Medicine* https://doi.org/10.1038/s41591-019-0637-6 (2019)

A recent study reported that a 32-base-pair deletion in the *CCR5* gene (*CCR5*-Δ32) is deleterious in the homozygous state in humans. Evidence for this came from a survival analysis in the UK Biobank cohort, and from deviations from Hardy–Weinberg equilibrium at a polymorphism tagging the deletion (rs62625034). Here, we carry out a joint analysis of whole-genome genotyping data and whole-exome sequencing data from the UK Biobank, which reveals that technical artifacts are a more plausible cause for deviations from Hardy–Weinberg equilibrium at this polymorphism. Specifically, we find that individuals homozygous for the deletion in the sequencing data are under-represented in the genotyping data due to an elevated rate of missing data at rs62625034, possibly because the probe for this single-nucleotide polymorphism overlaps with the Δ32 deletion. Another variant, which has a higher concordance with the deletion in the sequencing data, shows no associations with mortality. A phenome-wide scan for effects of variants tagging this deletion shows an overall inflation of association *P* values, but identifies only one trait at $P < 5 \times 10^{-8}$, and no mediators for an effect on mortality. These analyses show that the original reports of a recessive deleterious effect of *CCR5*-Δ32 are affected by a technical artifact, and that a closer investigation of the same data provides no positive evidence for an effect on lifespan.

CCR5-Δ32 is a deletion in the coding region of the *CCR5* gene, and homozygous deletion of *CCR5*-Δ32 (Δ32/Δ32) has been reported to confer resistance against human immunodeficiency virus infections in humans[1–3]. A recent study (now retracted[4]) suggested that Δ32/Δ32 individuals have a 21% increased mortality rate, and that the increased mortality rate leads to deviations from Hardy–Weinberg equilibrium (HWE) at this site[4,5]. Here, we re-analyze the data on which these results were based, and find that the variant that most closely tags Δ32/Δ32 shows no evidence for an effect on mortality or a deviation from HWE. Our findings show that the previously reported effect on mortality was probably spurious and that the observed deviation from HWE was caused by a technical artifact.

Our work consists of four parts. First, we investigate which variants are most accurately tagging Δ32. Second, we re-examine the evidence for deviation from HWE at these variants. Third, we re-examine the evidence for effects on mortality at these variants. Fourth, we extend previous association tests to identify phenotypes that could potentially mediate an effect of Δ32/Δ32 on mortality.

The original study by Wei and Nielsen (now retracted[4]) investigated potential deleterious effects of Δ32/Δ32 using genetic data and mortality data from the UK Biobank resource. The genotyped single-nucleotide polymorphism (SNP) rs62625034 was used as a proxy for Δ32. However, in an article posted on his online blog[6], S. Harrison showed that the results do not replicate at the nearby correlated SNP rs113010081. Building on this, we compare two genotyped and two imputed variants with the *CCR5*-Δ32 deletion as called in the recently released UK Biobank exome sequencing data (rs333_sequenced), which we treat as the ground truth (Supplementary Tables 1 and 2). The genotyped SNP rs113010081 is a better proxy for Δ32 than rs62625034, as indicated by a higher concordance across all genotype classes (+/+, Δ32/+ and Δ32/Δ32), as well as higher sensitivity and specificity to distinguish Δ32/Δ32 from +/+ and Δ32/+ (Fig. 1, Extended Data Figs. 1 and 2 and Supplementary Tables 3 and 4). In addition, the three genotype classes show better separation in the probe intensity scatter plots (Fig. 1). rs113010081 was not used as a proxy for Δ32 in the original study due to its high missingness (10.3%). However, the overall high missingness rate is caused by the absence of this variant from the UK BiLEVE Axiom array, which was used to genotype the first ~10% of genotyped samples in the UK Biobank. On the UK Biobank Axiom array, which was used for the remaining ~90% of samples, this variant has a missingness rate of 0.08%, while rs62625034 has a missingness rate of 3.6%. Thus, the genotypes of rs113010081 provide a better proxy for Δ32 than those for rs62625034. As the imputed variants tested here are less correlated with Δ32 than the two genotyped variants, we refer to the genotyped variants unless otherwise specified.

When testing for deviations from HWE, we confirm that rs62625034 shows a highly significant deviation from HWE, caused by a deficiency of individuals with two copies of the rare (deletion-tagging) allele (Supplementary Table 5). However, neither rs113010081 nor rs333_sequenced shows a significant deviation from HWE under a chi-squared HWE test. rs62625034 does show a significant HWE deviation, even in the subset of samples with sequencing data, which shows that a difference in power does not cause this discrepancy.

The missingness rate of rs62625034 differs by Δ32 genotype class, as called in the sequencing data (17.3, 4.6 and 2.9% for Δ32/Δ32, Δ32/+ and +/+, respectively; Fig. 1). The HWE deviation at this SNP is fully explained by this bias in missingness (Supplementary

[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

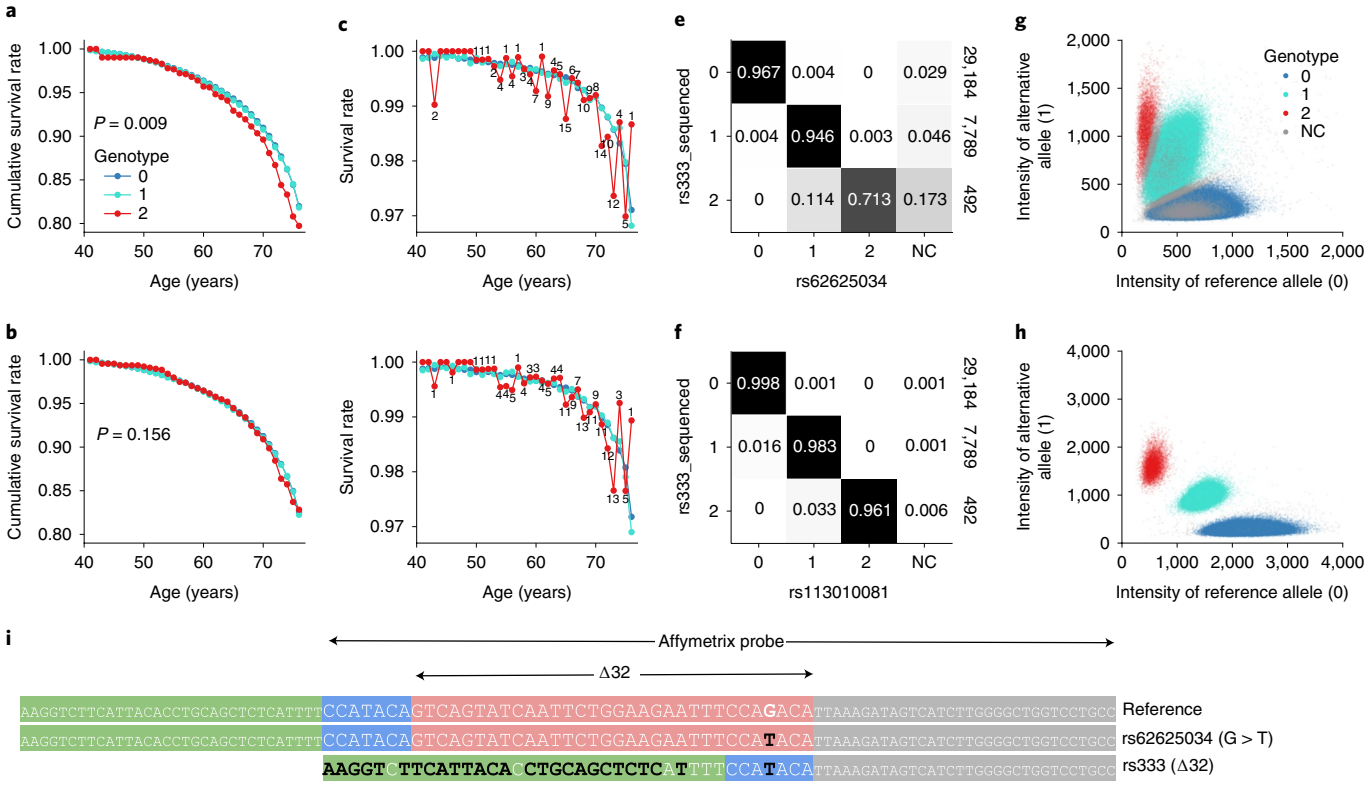


Fig. 1 | Survival rates for individuals with zero, one and two copies of the rare allele for two variants tagging the *CCR5*-Δ32 deletion. a,b, Cumulative survival rates show that the evidence for increased mortality of individuals homozygous for the variant allele in rs62625034 (**a**) does not replicate in rs113010081 (**b**). One-sided *P* values are from a Cox proportional hazard model comparing survival rates of individuals with zero or one allele(s) with those with two alleles. **c,d**, Non-cumulative survival rates for rs62625034 (**c**) and rs113010081 (**d**), which show the large year-to-year variability in the data caused by small sample counts. Numbers indicate how many Δ32/Δ32 individuals died in each year or age. **e,f**, Distribution of genotypes at rs62625034 (**e**) and rs113010081 (**f**) (including missing genotypes) conditioned on rs333_sequenced genotypes. The total count for each row is shown to the right. Missing data are strongly correlated with genotype class for rs62625034, which fully explains the deviation from HWE at this site. No such bias is present at rs113010081. Numbers are based only on samples genotyped on the UK Biobank Axiom array, as rs113010081 data are only available for this array. **g,h**, Allele intensity clusters for UK Biobank genotyping data, showing the poorer separation of genotype classes for rs62625034 (**g**) compared with rs113010081 (**h**). **i**, Different haplotypes at the *CCR5*-Δ32 locus. Black nucleotides differ from the reference. The site of the very rare SNP rs62625034 (G > T) is located within the Δ32 deletion. Due to the sequence similarity at the 3′ end, the probe tags the deletion instead. However, the rs62625034 probes match the reference genotype better than the deletion, leading to higher missingness in the presence of the deletion. NC, no call.

Table 6). Individuals with missing data at rs113010081 are not similarly biased with respect to rs333_sequenced (Fig. 1). The nonrandom missingness of rs62625034 with respect to Δ32 may be caused by the fact that the probe for this SNP overlaps with the deletion region but matches it only imperfectly (Fig. 1).

We carried out a simulation study showing that for two variants in high linkage disequilibrium, strong deviations from HWE at one variant, but not the other, cannot be induced by ascertaining samples on one variant alone (Extended Data Fig. 3). However, correlated ascertainment on both variants (which can occur through technical artifacts) can create this pattern.

When analyzing survival rates, we recapitulate the findings of Wei and Nielsen[4,5], and find that for rs62625034, carriers of two copies of the rare allele tend to have a lower survival rate (Fig. 1, Extended Data Fig. 1 and Supplementary Table 7). However, none of the other tested variants shows any association with survival rate. The fact that the highly correlated rs113010081 SNP shows no association with survival, and the small number of deaths per year on which the signal is based (Fig. 1), make this finding uncompelling. The power to detect a 20% increased mortality rate at this SNP at a 0.05 significance level is only 75% (Extended Data Fig. 4 and Supplementary Information), which means that we cannot rule out that the deletion does affect survival based on the available data. We note that samples with missing genotypes at rs113010081 have greatly increased mortality rates ($P = 2.7 \times 10^{-32}$) due to a batch effect that is described in the Supplementary Information.

To identify phenotypes that could potentially mediate an effect of Δ32/Δ32 on mortality, we tested 3,911 phenotypes for associations with Δ32/Δ32, tagged by rs113010081. We identify 'lymphocyte count' as the only trait that is significant at a *P* value smaller than the classic threshold for declaring genome-wide statistical significance: 5×10^{-8} (Supplementary Table 8 and Extended Data Figs. 5 and 6). At less stringent *P* value thresholds, we find associated phenotypes that are similar to the previously reported associations from additive tests (Supplementary Tables 8 and 9). These are consistent with the role of C–C chemokine receptor type 5 in the immune system, and suggest that Δ32/Δ32 has effects besides conferring resistance to human immunodeficiency virus. However, we do not observe, on any diseases, effects that are large enough to explain a substantially increased mortality rate (Supplementary Information).

In summary, our analyses show no evidence that Δ32/Δ32 individuals have increased mortality rates. Similar findings have also been reported in other recent manuscripts[7–9]. This provides a case example of the subtle pitfalls that can produce false positive results, even in an extraordinarily high-quality and relatively uniformly generated dataset such as the UK Biobank.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All the data used in this study are available with the permission of the UK Biobank.

Received: 3 October 2019; Accepted: 19 November 2019;
Published online: 23 December 2019

References

1. Samson, M. et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the *CCR-5* chemokine receptor gene. *Nature* **382**, 722–725 (1996).
2. Hütter, G. et al. Long-term control of HIV by *CCR5* delta32/delta32 stem-cell transplantation. *N. Engl. J. Med.* **360**, 692–698 (2009).
3. Gupta, R. K. et al. HIV-1 remission following CCR5Δ32/Δ32 haematopoietic stem-cell transplantation. *Nature* **568**, 244–248 (2019).
4. Wei, X. & Nielsen, R. Retraction note: *CCR5*-Δ32 is deleterious in the homozygous state in humans. *Nat. Med.* **25**, 1796 (2019).
5. Wei, X. & Nielsen, R. Deviations from Hardy Weinberg equilibrium at *CCR5*-Δ32 in large sequencing data sets. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/768390v2 (2019).
6. Harrison, S. "CCR5-Δ32 is deleterious in the homozygous state in humans"—is it? *Sean Harrison: Blog* https://seanharrisonblog.com/2019/06/20/ccr5-%e2%88%8632-is-deleterious-in-the-homozygous-state-in-humans-is-it/ (2019).
7. Gudbjartsson, D. et al. CCR5-del32 is not deleterious in the homozygous state in humans. Preprint at *bioRxiv* https://doi.org/10.1101/788117 (2019).
8. Karczewski, K. J., Gauthier, L. D. & Daly, M. J. Technical artifact drives apparent deviation from Hardy–Weinberg equilibrium at CCR5-Δ32 and other variants in gnomAD. Preprint at *bioRxiv* https://doi.org/10.1101/784157 (2019).
9. Tanigawa, Y. & Rivas, M. A. Reported CCR5-Δ32 deviation from Hardy-Weinberg equilibrium is explained by poor genotyping of rs62625034. Preprint at *bioRxiv* https://doi.org/10.1101/791517 (2019).

Acknowledgements

This research has been conducted using the UK Biobank Resource under application no. 31063. We acknowledge the participants in the UK Biobank. We are grateful to B. Neale and A. Price for critical comments, and to S. 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 K. Karczewski, K. Stefansson and M. 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.

Author contributions

R.M. and A.A. performed all of the analyses except for the one presented in Supplementary Table 4, which was carried out by X.W. N.P., R.N. and D.R. supervised the study. R.M., A.A. and D.R. wrote the manuscript, with critical review from all co-authors.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-019-0710-1.

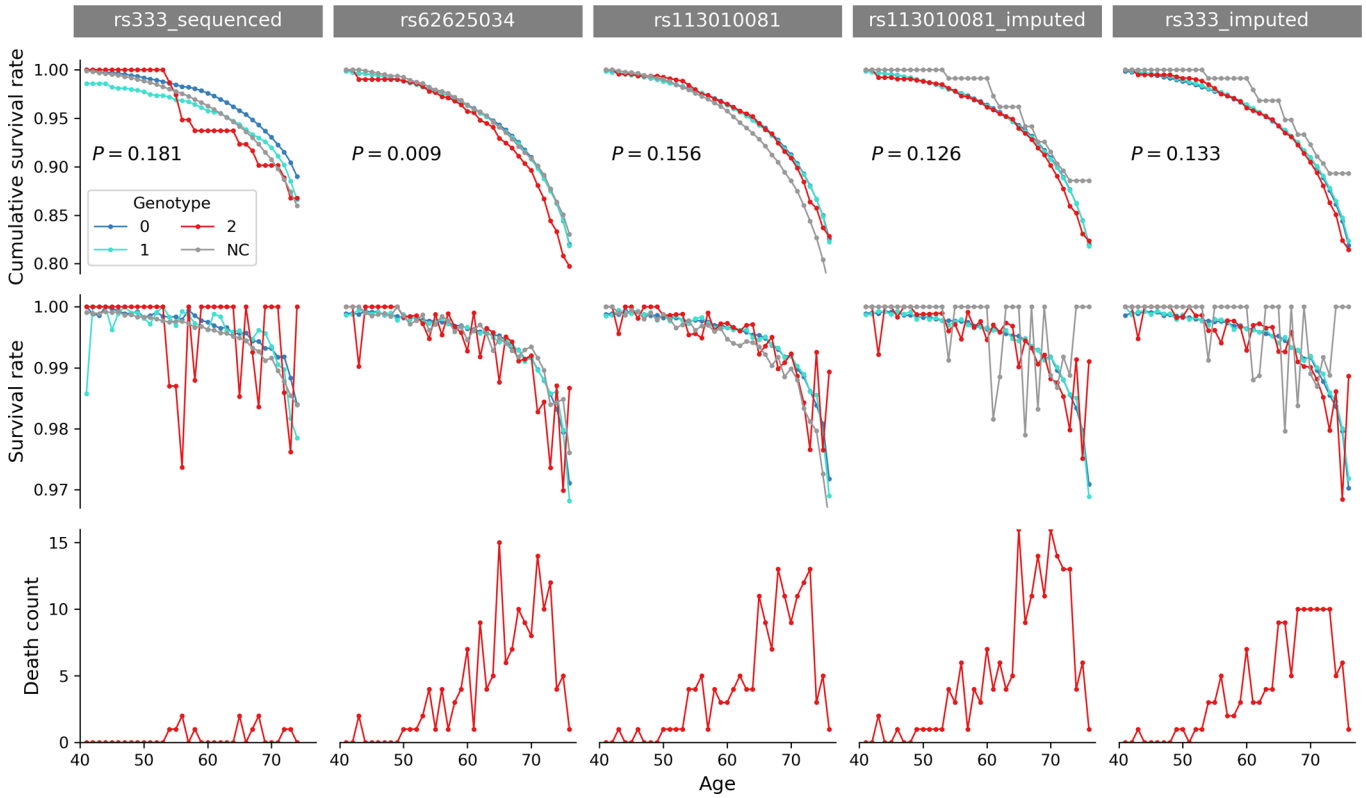
Supplementary information is available for this paper at https://doi.org/10.1038/s41591-019-0710-1.

Correspondence and requests for materials should be addressed to R.M. or A.A.

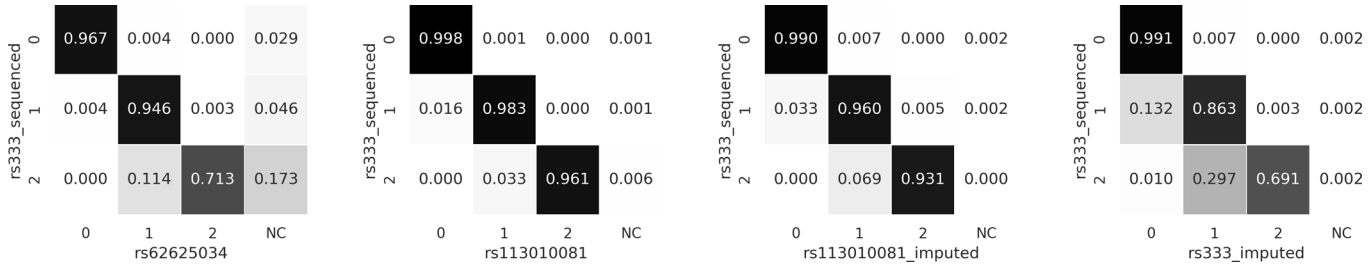
Peer review information Kate Gao was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.

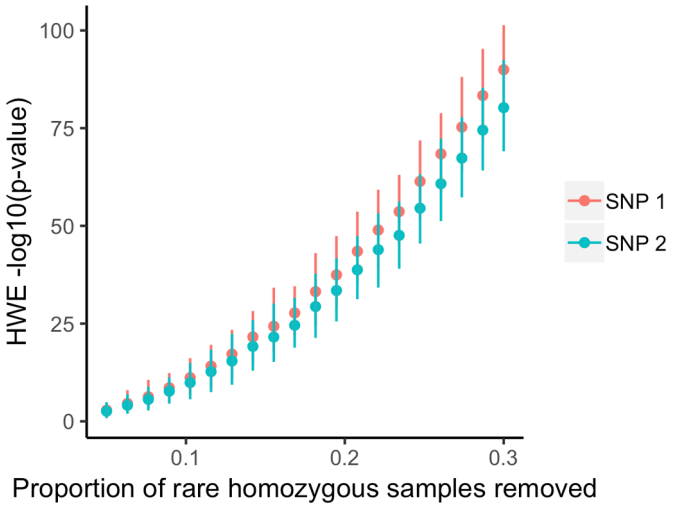
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



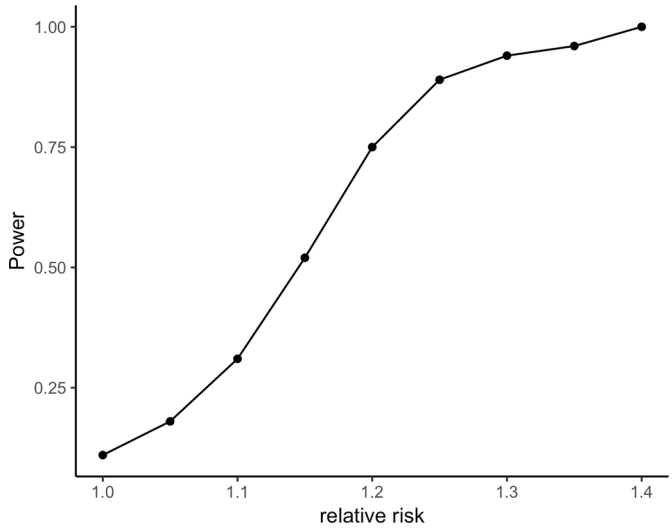
Extended Data Fig. 1 | Survival analysis. Survival rates for individuals with 0, 1, or 2 copies of the rare allele or No Call (NC) for variants tagging the CCR5-Δ32 deletion. First row: Cumulative survival rates. Numbers are one-sided p-values of a Cox proportional hazard model which compares survival rates of individuals with 0 or 1 alleles to those with 2 alleles. Second row: non-cumulative survival rates. Third row: Number of individuals who have died in any given year with 2 copies of rare allele (see also Supplementary Table 7).



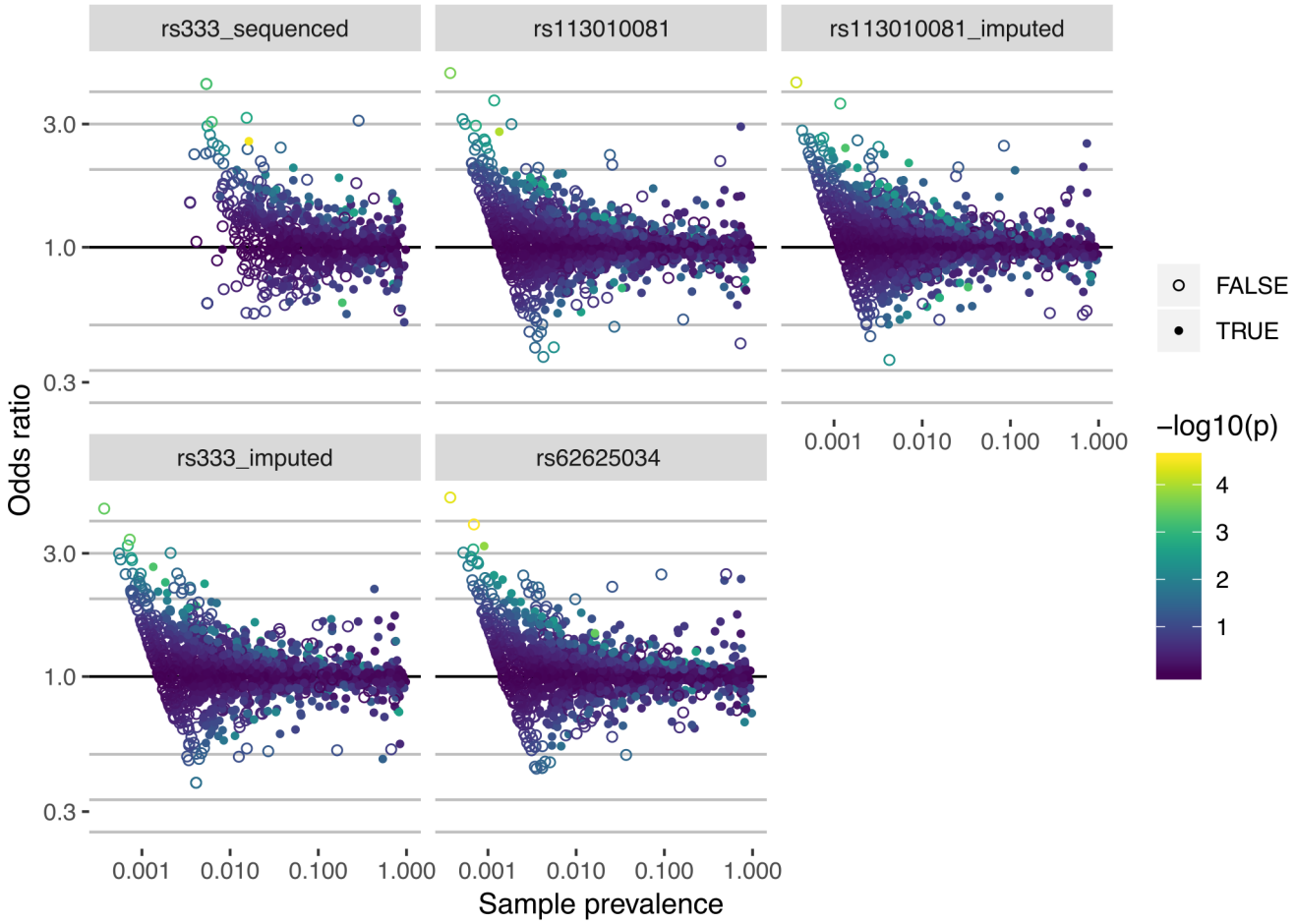
Extended Data Fig. 2 | Concordance analysis. Confusion matrix for different markers with missing data. The last column of the first panel shows that individuals with missing genotype at rs62625034 are enriched for Δ32/Δ32 according to rs333_sequenced. This can lead to a violation of HWE at rs62625034. All white British samples of UK Biobank WES data shared with UK Biobank Axiom array data are used in this figure.



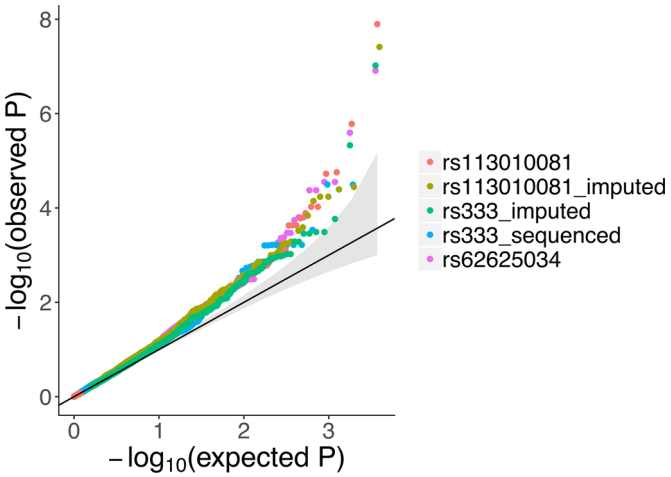
Extended Data Fig. 3 | HWE p-values of linked variants. Simulated HWE Chi-squared p-values at two variants with minor allele frequency of 11% with r^2 of 0.95, in a sample of 400,000 individuals. Both variants are initially in HWE. We then remove a subset of samples which are homozygous for the rare allele at SNP 1. This leads to a deviation from HWE at SNP 1, but it also leads to a similar deviation from HWE at SNP 2. Only simultaneous selection acting in the opposing direction on SNP 2, or technical artifacts which create a dependence of missingness in one SNP on genotype in the other SNP explain a situation where HWE p-values are very different at both SNPs. Error bars denote the 5th and 95th percentile out of 100 replicates in each bin.



Extended Data Fig. 4 | Power analysis. Power to detect effects on mortality of a genotype with the frequency of Δ32/Δ32 in a sample of the same total size and mortality rate as the cohort studied here, as a function of relative risk. The power to detect a 20% increase in mortality rate at a 0.05 significance level is 75%.



Extended Data Fig. 5 | Odds ratios against sample prevalence. Odds ratios (e^{β}) for all case-control phenotypes in five variants as a function of sample prevalence. Colors represent uncorrected p-values. Open circles represent case-control phenotypes with 10 or fewer cases in Δ32/Δ32 individuals. Only phenotypes with more than five cases in Δ32/Δ32 individuals are shown.



Extended Data Fig. 6 | QQ-plot of the associations. QQ-plot of the associations across all phenotypes. Each variant is plotted in a different color. Only phenotypes with more than five cases in Δ32/Δ32 individuals are shown.

nature research

Corresponding author(s): Ali Akbari

Last updated by author(s): 2019/10/04

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No custom code or mathematical algorithm that is deemed central to the conclusions have been used.
Data analysis	No custom code or mathematical algorithm that is deemed central to the conclusions have been used.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All the data used in this study are available with the permission of the UK Biobank.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Approximately 500,000 individuals from UK Biobank
Data exclusions	We only analyzed the white British individuals in this study (84% of the UK Biobank data)
Replication	NA
Randomization	NA
Blinding	NA

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging