Supplementary information

Supplementary materials and methods

Effect of Δ32 in non-British ancestry individuals

Similarly to the results in the British ancestry individuals, the survival probabilities of Δ32+ and +/+ in the non-British ancestry individuals in UK Biobank are not significantly different from each other (bootstrap two-tail $P = 0.45$). Unfortunately, only 560 non-British ancestry volunteers are homozygous at Δ32, among them 15 have death records, so the sample size is too small to reliably estimate survival probabilities. Whether Δ32 also is deleterious in the homozygous state in other populations than the British, cannot be determined by our study.

Deviation from Hardy-Weinberg Equilibrium (HWE) with age

We calculate the deviation from HWE for Δ32/Δ32 at the age at recruitment using the following steps: First, we estimate the frequency of individuals recruited at age i, p_i, for each genotype. Next, we estimate the allele frequency of Δ32 at age i as: $p_{i,\Delta 32} = p_{i,\Delta 32/+}/2 + p_{i,\Delta 32/\Delta 32}$. We then use $p_{i,\Delta 32/\Delta 32}/p^2_{i,\Delta 32}$ to measure the observed deviation from HWE for each age. Because the age used is the age at recruitment, this age range is smaller than the age range for survival probability. We generate 1000 bootstrap samples of the genotypes for each age i to obtain a 95% confidence interval for the deviation from Hardy-Weinberg equilibrium.

We also calculate the predicted deviation from HWE using the observed survival probabilities. First, we estimate the total observed fraction of each genotype among all volunteers $p_{0,+/+}, p_{0,\Delta 32/+}, p_{0,\Delta 32/\Delta 32}$ and use them as the baseline at age 0. This baseline may slightly affect the predicted deviations. The predicted population survival probability to each age is:

$$w_i = p_{0,+/+} S_{C,i,+/+} + p_{0,\Delta 32/+} S_{C,i,\Delta 32/+} + p_{0,\Delta 32/\Delta 32} S_{C,i,\Delta 32/\Delta 32}$$

We then calculate the predicted frequency q for each genotype at each age:

$$q_{i,+/+} = p_{0,+/+} S_{C,i,+/+} / w_i$$

$$q_{i,\Delta 32/+} = p_{0,\Delta 32/+} S_{C,i,\Delta 32/+} / w_i$$

$$q_{i,\Delta 2/\Delta 32} = p_{0,\Delta 32/\Delta 32} S_{C,i,\Delta 32/\Delta 32} / w_i$$

$$q_{i,\Delta 32} = q_{i,\Delta 32/+} / 2 + q_{i,\Delta 32/\Delta 32}$$

The predicted deviation from HWE is therefore: $q_{i,\Delta 32/\Delta 32} / q_{i,\Delta 32}^2$.

CCR5-Δ32 and HIV-infection

The observed deviation from HWE by age of recruitment generally agrees with the predicted deviation based on the corrected survival probability (Spearman's $\rho = 0.67, P = 1.4 \times 10^{-4}$; Fig. S1), even though the two curves differ from each other in a number of respects (Fig. S1a and Fig. S1b). The largest discrepancy between the two curves occurs at around age 50, where the observed curve first increases and then decreases much more rapidly than the predicted curve. One possible explanation might be that the prediction is based on death reports from the recent ten years, which could be different from the death rate in earlier years. The protective effect of Δ32 against HIV infection[1] might have a smaller effect on survivability in recent years due to the improvement in treatment options for HIV. The difference in mortality rates between HIV-infected individuals and the general population narrowed in every calendar period from 1996 onward[2], with a majority of deaths from HIV occurring in the early 90s before the introduction of HAART treatment. In the UK Biobank, the volunteers who were at around age 50 at recruitment were at age around 30 in 1990. Therefore, the protective effect of Δ32

might be the largest for the cohort recruited at age approximately 50, creating a bigger discrepancy between the observed and predicted curves (Fig. S1a and Fig. S1b). We also note that the deviation of Δ32/Δ32 from HWE is already very large even before age of 40, suggesting that Δ32 in the homozygous state might also be deleterious in early life or before birth. This prediction could be reevaluated when an appropriate cohort with younger volunteers is available. However, a likely explanation of the current curve is that the deleterious effects of the Δ32/Δ32 genotype disproportionately affects young people (perhaps increasing infant mortality) and old people, while there is an overall protective effect in the middle years likely due to protection against HIV.

Controlling for population structure

To exclude the possibility that population structure or particular ancestry affects the estimated mortality of Δ32/Δ32 individuals, we apply a Cox-model for left truncated and right censored data[3] on all the British ancestry individuals that are genotyped for Δ32. We use the function "coxph" in R package "survival"[4], and provide the start time, end time, and event. The start time is the estimated age at recruitment(see Materials and Methods). For volunteers who do not have a recorded death in the data, the event is coded 0, and the end time is the estimated age on 2016-02-16. For those who have a recorded death, the event is coded 1, and the age at death is the end time. This function assumes that each predictor has an homogeneous effect on the event across all time, but different ages could have different baseline death rates. We code the genotype of Δ32/Δ32 as 1, and the other two genotypes as 0. We first test this model using the genotype as the only predictor, and find Δ32/Δ32 has an 21.4% elevated death rate, with 95% confidence interval 3.4% and 42.6%. We then incorporate all the 40 principal components from UK Biobank[5] as covariates into the model to control for the effect of population structure, and find that Δ32/Δ32 still have an effect of 21.2% (CI: 3.2% to 42.3)%. Although five of the 40 PCs show significant effects (at two-tail $P = 0.05$ level) on death rate, Δ32/Δ32 still has an elevated death rate of 21.0% (CI: of 3.1% to 42.1%) when these five significant PCs (PC5, PC16, PC19, PC29, PC40) are included as covariates. When controlling for

multiple testing, only PC5 remains significant. Positive PC loads on PC5 is correlated with elevated death rate (1 unit in PC load corresponds to 1.1% elevated death rate with 95% confidence interval 0.84% to 1.4% and two-sided $P = 2.5 \times 10^{-16}$), and the Irish ancestry volunteers in the UK Biobank generally have higher loads on PC5[5]. These effect size estimates of Δ32 using the Cox model, with or without controlling for population structure, are similar to those obtained when allowing genotypes to have variable death rate at different ages (see Materials and Methods and Table S1).

Δ32 and the UK Biobank phenotypes

We download the summary statistics from the Global Biobank Engine through links on https://github.com/rivas-lab/public-resources/tree/master/uk_biobank. This data provide the association statistics for UK Biobank phenotypes[6] for all SNPs that pass their filtering criteria for SNP quality control. Because many phenotypes are only measured/recorded in a small proportion of volunteers, and because the SNP quality control is done separately for each phenotype, not all SNPs have statistics available for all phenotypes. The candidate SNP representing Δ32 has statistical record available for 821 phenotypes, eight are significant at significance level $0.05/821 = 6.1 \times 10^{-5}$. The eight phenotypes are: lymphocyte count, high light scatter reticulocyte count, immature reticulocyte fraction, high light scatter reticulocyte percentage, reticulocyte count, reticulocyte percentage, corneal hysteresis, and corneal resistance factor (right). To test whether Δ32 has more significant phenotypic associations than other SNPs, we compare it to the SNPs with matching MAF. A total of 823 phenotypes have statistics available for at least one matching MAF SNP, but only 181 phenotypes have summary statistics available for all 5932 SNPs with matching MAF, as well as Δ32. We use these 181 pheno- types and apply a PheWas significance level of $P = 0.05/181 = 2.76 \times 10^{-4}$ to determine the number of phenotypes associated with each SNP. We find that Δ32 is significantly associated with eight phenotypes including two additional phenotypes: osteoporosis and heart attack/myocardial infarction. Only 76 out of the other 5932 SNPs are associated with eight or more phenotypes, suggesting that Δ32 is associated with more UK Biobank

phenotypes than random SNPs with the same allele frequency ($P = 0.0128$). However, we note that many of the phenotypes are correlated and that this conclusion is highly dependent on the definition of phenotypes.

Supplementary 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 (1996).
2. Bhaskaran, K. *et al.* Changes in the risk of death after HIV seroconversion compared with mortality in the general population. *Jama* **300**, 51–59 (2008).
3. Cox, D. R. *Analysis of survival data* (Routledge, 2018).
4. Therneau, T. M. & Lumley, T. Package 'survival'. *R Top Doc* **128** (2015).
5. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203 (2018).
6. McInnes, G. *et al.* Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. *BioRxiv* 304188 (2018). DOI: https://doi.org/10.1101/304188

Supplementary figure legend

Extended Data Figure 1. The deviation from HWE with age. a, The observed deviation using age at recruitment estimated. Each dot represents one age group. The grey error bars show the 95% confidence intervals estimated from bootstrap the genotypes of individuals recruited at each age 1000 times. The sample size used for each error bar ranges from 15191 to 100117 with a mean of 65479. **b**, The predicted deviation from HWE using the corrected survival probability. A total of 395704 samples are used.

The observed and predicted values are significantly correlated (Spearman's correlation coefficient $\rho = 0.67$, $P = 1.4 \times 10^{-4}$).

Supplementary table

Supplementary Table 1. Δ32/Δ32 **individuals have higher death rate.**

Age	log-rank [1]	Δ32/Δ32 vs. +/+ [2]	Δ32/Δ32 vs. Δ32/+	Δ32/+ vs. +/+
72*	0.021	22.4% [-1.7%,48.3%]	20.6% [-3.8%,46.9%]	1.49% [-4.5%,7.6%]
73*	0.019	21.4% [-0.8%,45.0%]	20.0% [-2.0%,43.3%]	1.10% [-4.5%,6.7%]
74**	0.0074	26.4% [3.0%,49.5%]	25.7% [2.8%,50.1%]	0.58% [-4.9%,5.9%]
75**	0.0089	22.5% [1.4%,45.3%]	24.5% [2.3%,48.2%]	-1.56% [-7.0%,3.6%]
76*	0.0089	21.1% [-1.2%,44.3%]	22.5% [-0.5%,47.0%]	-1.15% [-6.9%,4.6%]

[1] Log-rank one-tail *P* -value reveals an elevated death rate of Δ32/Δ32 vs. the rest. The sample size used varies from 395698 at age 72 to 395704 at age 74 or larger.

[2] The increase in probability of death and the 95% bootstrap confidence interval. Sample sizes: 4349 for Δ32/Δ32, 83038 for Δ32/+, and 308314-308317 for +/+.

* Δ32/Δ32 has a higher probability of death at bootstrap one-tail $P = 0.05$.

**Δ32/Δ32 has a higher probability of death at bootstrap one-tail $P = 0.025$.