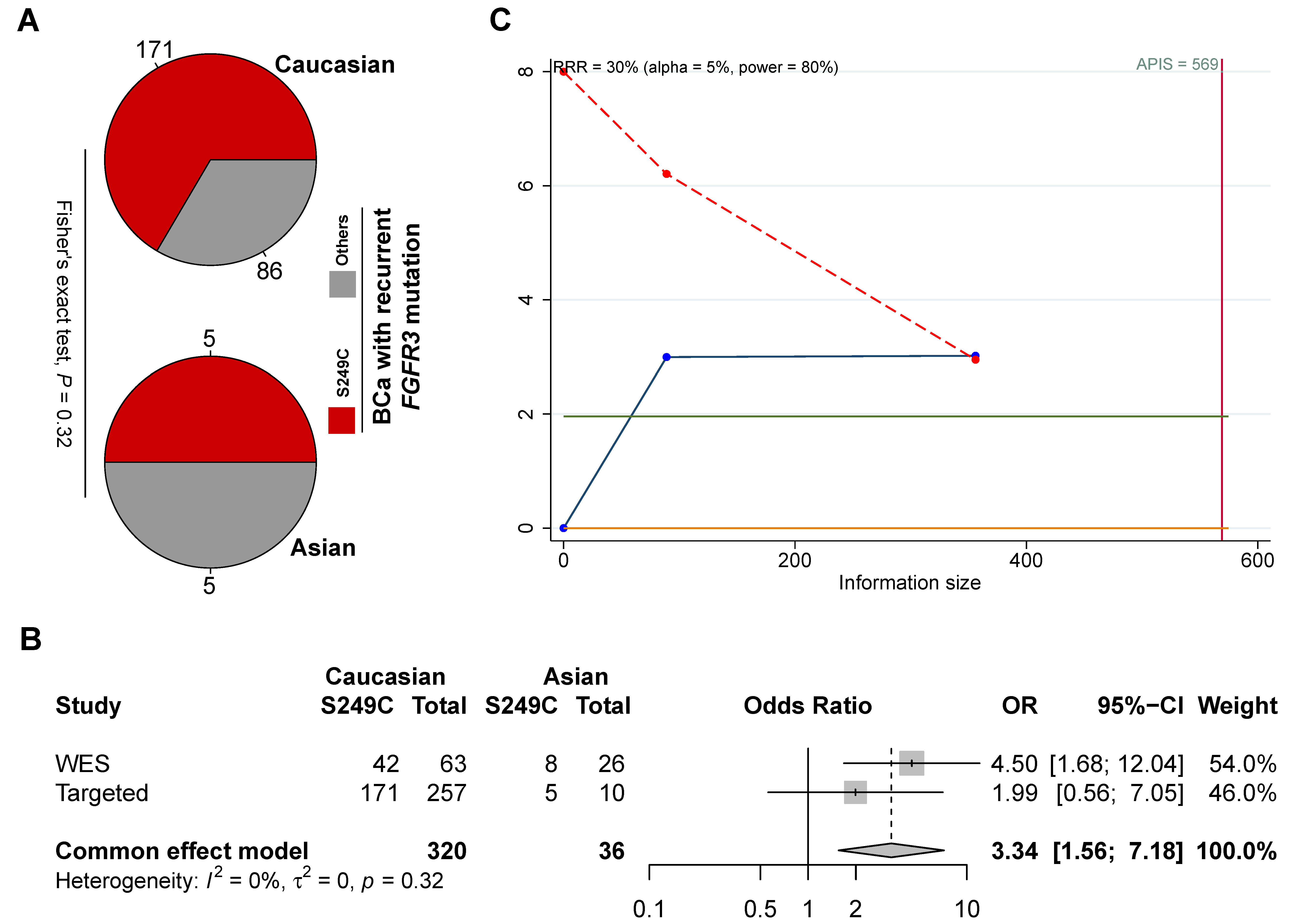
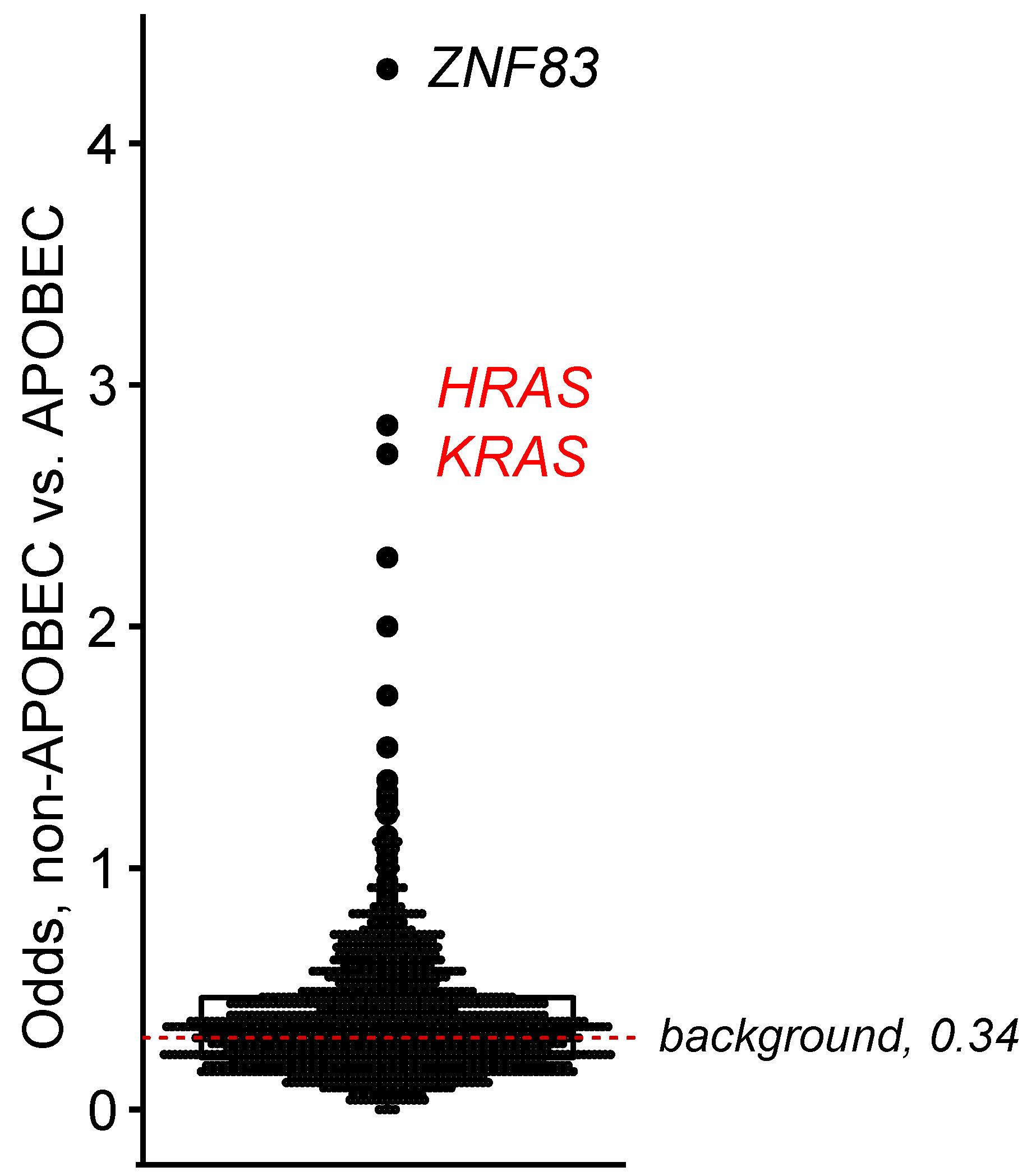
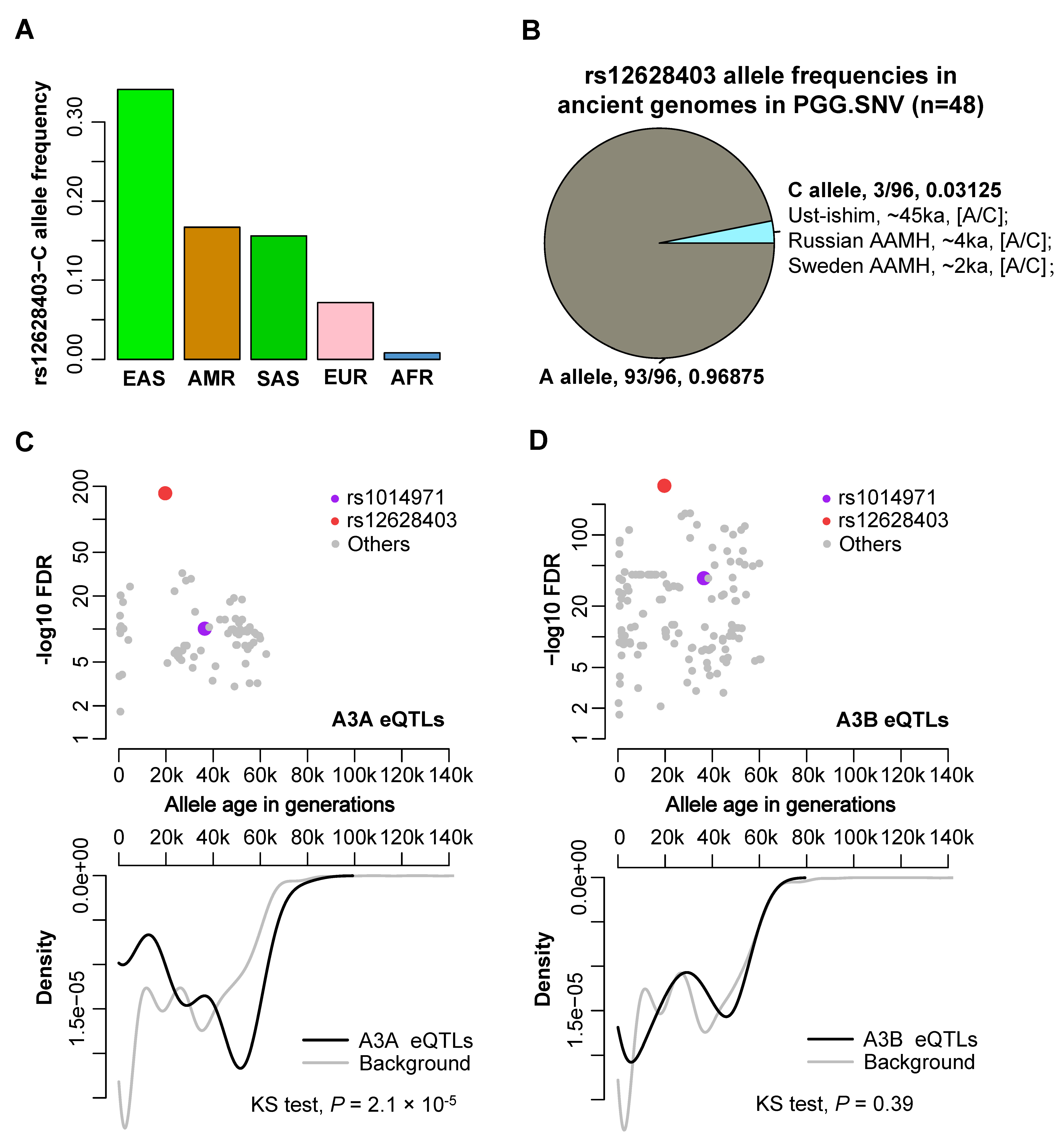
**SUPPLEMENTARY FIGURES**



**Figure S1. Further analysis for the association between *FGFR3* S249C prevalence among *FGFR3* recurrent mutations by ethnicity. (A)** A higher FGFR3 S249C prevalence was observed in Caucasian vs. Asian patients, with non-significant difference (OR = 1.99, Fisher’s test, P = 0.32) likely due to limited number of Asian patients (n = 10). **(B)** Fixed-effects Mantel-Haenszel meta-analysis of the WES and the targeted-sequencing cohorts showed significantly higher S249C prevalence in Caucasians compared with Asians (Statistical heterogeneity I2 = 0; pooled OR = 3.34, *P* = 0.002). (C) Trial sequential monitoring boundaries just crossed the cumulative Z-curve with a priori relative risk reduction (RRR) set at 0.3 and 5% alpha at 80% power, but did not surpass the a priori information size (APIS), suggesting additional samples with *FGFR3* recurrent mutations were needed for further validation with a smaller a priori RRR, i.e. a smaller assumed difference in S249C fraction between Caucasian and Asian patients.



**Fig. S2. Odds of mutation events with non-APOBEC and APOBEC mutagenesis as origins in highly mutated genes (≥20 mutations).** The dashed line indicates the background of all mutations.



**Fig. S3. Allele frequency of rs12628403 in contemporary and ancient genomes; and allele age and eQTL effect of rs1014971 and rs12628403. (A)** Ethnic difference in rs12628403-C allele frequency in 1000Genomes populations (n = 5,008). **(B)** rs12628403 allele frequencies in ancient genomes in the PGG.SNV database. AAMH, ancient anatomically modern humans. **(C-D)** Estimated allele age in generations and adjusted statistical significance (-log10 false discovery rate, -log10 FDR) of *APOBEC3A* (upper part of E) and 3B eQTLs (upper part of F). Density curves showing the distributions of estimated ages of local background SNPs vs. *APOBEC3A* (lower part of E) and 3B eQTLs (lower part of F). A3A, *APOBEC3A*; A3B, *APOBEC3B*; eQTL, expression quantitative trait loci; KS test, Kolmogorov-Smirnov test. Allele age estimations were obtained from the Human Genome Dating database and eQTL summary statistics were obtained from the GTEx database v8 (Metasoft results for all tissues in the V8 release).

**SUPPLEMENTARY TABLES**

**Table S1. Studies considered for validation analysis regarding the relationship between *FGFR3* mutation and ethnicity.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Study** | **YoP** | **Center** | **Cancer type** | **n** | **Methods** | **Inclusion** |
| Lee et al. | 2018 | Columbia Univ. | UC | 130 | targeted | Noa |
| Robinson et al. | 2019 | Cornell/Baylor/MDACC | UTUC | 37 | WES | Noa |
| Su et al. | 2021 | IGBMC | UTUC | 30 | WES | Noa |
| Clinton et al. | 2022 | MSKCC | UC | 1,659 | targeted | Yes |
| Damrauer et al. | 2022 | BCAN/HCRN | UC | 192 | targeted | Yes |

a Excluded due to no data available for patients’ ethnicity. YoP, year of publication. UC, urothelial carcinoma; UTUC, upper tract urothelial carcinoma. MDACC, MD Anderson Cancer Center; IGBMC, Institut de génétique, biologie moléculaire et cellulaire; MSKCC, Memorial Sloan Kettering Cancer Center; BCAN, Bladder Cancer Advocacy Network; HCRN, Hoosier Cancer Research Network.

**Table S2. rs1014971 allele frequency distribution in the present and ancient samples. Data correspond to upper panel forest plot of Figure 1J.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Region** | **Present** | | **Ancient** | | **OR** | **95%CI** | **P-value** |
| **n** | **T** | **n** | **T** |
| West Eurasia | 1,006 | 65.51% | 6,418 | 64.65% | 1.04 | [0.90,1.19] | 0.60 |
| Americas | 694 | 50.10% | 624 | 26.60% | 2.77 | [2.20,3.50] | 5.9×10-18 |
| Africa | 1,322 | 40.24% | 134 | 41.04% | 0.97 | [0.67,1.39] | 0.86 |
| South Asia | 978 | 34.60% | 200 | 56.00% | 0.42 | [0.30,0.57] | 2.4×10-8 |
| East Asia | 1,008 | 28.47% | 768 | 41.41% | 0.56 | [0.46,0.69] | 1.4×10-11 |

n, sample size; T, frequency of rs1014971-T allele.

**Table S3. Number of genomes analyzed and rs1014971-T allele frequency in ancient samples of West Eurasia region by chronologic time points (1k year unit before present). Data correspond to lower panel line plot of Figure 1J.**

|  |  |  |
| --- | --- | --- |
| **Time** | **n** | **T** |
| ≥10k | 48 | 29.17% |
| 9k-10k | 14 | 64.29% |
| 8k-9k | 55 | 63.64% |
| 7k-8k | 157 | 74.20% |
| 6k-7k | 144 | 71.53% |
| 5k-6k | 292 | 71.23% |
| 4k-5k | 449 | 62.36% |
| 3k-4k | 579 | 64.59% |
| 2k-3k | 532 | 64.66% |
| 1k-2k | 703 | 62.02% |
| 0k-1k | 236 | 65.68% |

**Table S4. Fisher’s exact test P-value for rs1014971-T allele frequency in ancient samples of West Eurasia region by chronologic epochs (aggregated 1k-year periods based on similarity of rs1014971 allele frequency).**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Epoch 1** | **Epoch 2** | **Epoch 3** |
| **Epoch 2** | 1.9×10-7 | . | . |
| **Epoch 3** | 1.4×10-16 | 0.047 | . |
| **Epoch 4** | 1.4×10-11 | 1 | 2.5×10-8 |

The epochs are aggregated neighboring 1k-year units by similarity of rs1014971-T allele frequency. Epoch 1: ≥10k, overall rs1014971-T allele frequency 29.17%; Epoch 2: 8k-10k, overall rs1014971-T allele frequency 63.77%; Epoch 3: 5k-8k, overall rs1014971-T allele frequency 72.09%; Epoch 4: 0k-5k, overall rs1014971-T allele frequency 63.58%.