SNPs to predict Penicillin Allergy

Xavier Suriol
xavier.suriol@amsafis.com

The Harvard Personal Genome Project (PGP) website (https://pgp.med.harvard.edu) publishes individual genomic data as well as their personal health records, freely available. The current research downloaded 262 individual '23andMe' datasets of this website and merged all of them into a big dataset for statistical purposes.

The results show that the combination of SNPs rs1323826=AG AND rs12483887=CC (both simultaneously) is significant at p_value Fisher Exact Test 5.82442e-12 to predict the allergy to penicillin. An estimate of the Odds ratio is 166.019, 95% confidence interval is between 21.40169 and 7265.44838.

The big dataset (R file) is in the Open Science Framework website (https://osf.io/). Also, all PGP IDs are below and anybody can download individual files from the PGP website.

Participants with penicillin allergy in the dataset:

ID at PGP	Dataset size	rs1323826	rs12483887
hu1A4F2E	Small	n.a.	n.a.
hu1A7894	Small	n.a.	n.a.
hu002B3C	Large	AG	CC
hu4DE348	Large	AG	CC
hu8F82E3	Large	AG	CC
hu41F03B	Large	AG	CC
hu90B053	Large	AG	CC
hu499ED5	Small	n.a.	n.a.
hu840B0B	Large	AG	CC
hu16360E	Small	n.a.	n.a.
huC1C7D0	Large	AG	CC
huDB1635	Large	AG	CC
hu8D99F6	Large	AG	CC
hu11603C	Large	AG	CC
huE58004	Large	--	CC
huFE71F3	Large	AG	CC
hu443DE9	Small	n.a.	n.a.

hu4CFB79	Large	AG	CC
hu9FEC32	Large	AG	CT
huDD1522	Small	n.a.	n.a.

Small datasets have around 500,000 positions, and large ones almost 1,000,000. Both SNPs are only in large datasets and so 6 out of 20 having the disease (in yellow) were discontinued. 67 without the disease were discontinued for the same reason.

For hu9FEC32, the mentioned combination of SNPs fails (in red). For huE58004, one of the SNP is missing (in grey).

Such combination of rs1323826 = AG and rs12483887 = CC (both simultaneously) is low frequent for those not suffering from penicillin allergy. Next two tables show all the values for both SNPs.

Cross tabulation of the above table (only large datasets) is...

Yes Penicillin allergy

		rs12483887		
		CC	CT	TT
rs1323826	--	1	0	0
	AA	0	0	0
	AG	12	1	0
	GG	0	0	0

... and the cross tabulation for the ones without the disease (only large datasets) is:

No Penicillin allergy

		rs12483887		
		CC	CT	TT
rs1323826	--	0	1	0
	AA	58	52	17
	AG	11	23	4
	GG	4	4	1

The combined cross tabulation in a 2 x 2 table: (not included huE58004 - in grey- due to its missing rs1323826 since it cannot be confirmed the combination of SNPs is true or false; instead, the missing rs1323826 '--' of the No Penicillin allergy is included because rs12483887=CT doesn't belong to that combination).

	rs12483887=CC AND rs1323826=AG	NOT (rs12483887=CC AND rs1323826=AG)
Yes Penicillin allergy	12	1
No Penicillin allergy	11	164

The Odds Ratio and the Fisher's Exact Test (of the 2x2 contingency table) were estimated:

- At MedCalc (statistical software) https://www.medcalc.org/calc/odds_ratio.php:

Odds ratio: 178.9091

95% Confidence Interval: 21.2736 to 1504.6107

Z statistics: 4.774

Significance level: P < 0.0001

- R version 2.15.2.:

Fisher Exact Test p_value: 5.824424e-12.

Estimate of the Odds ratio: 166.019 (conditional Maximum Likelihood Estimate is used)

95% Confidence Interval: 21.40169 to 7265.44838

99% Confidence Interval: 14.23404 to 4.503600e+15

- R package statmod version 1.4.30 under R version 3.4.1.:

Power of the Fisher's Exact Test: 100% (up to 3 decimals) chance of detecting difference at 1% significance level and using 10,000 simulated datasets.

In the Phenotype-Genotype Integrator database of The National Center for Biotechnology Information (NCBI) (https://www.ncbi.nlm.nih.gov/gap/phegeni), for both SNPs it is reported "No association data found" in the "Association Results" section.

I added participants in 4 phases, the last one in October-November 2014.

The rest of participants are the following ones:

11 participants with the SNPs combination and no disease (in blue in tables above): hu1BD549 hu83E9B9 hu627574 huD50D1C hu14262A hu2D53F2 huE8E4FC hu05F442 hu3C86DB hu3EEE3A hu85E6EC

164 participants without the SNPs combination and no disease: hu0C0779 hu0DEE68 hu1AF744 hu1BDBA5 hu2BC187 hu2DEBA7 hu2E413D hu3CAB43 hu3D355A hu3F864B hu4AEB32 hu4B07B3 hu4B11A3 hu4BE6F2 hu4C3094 hu4D2239 hu05FD49 hu5A2074 hu5CD2C6 hu5FCE15 hu5FF6B0 hu6FECE9 hu7B594C hu7DE7FD hu8A7E4B hu8B3865 hu9D7C95 hu16A1B3 hu019BBA hu25E94B hu28F39C hu33F35D hu48C4EB hu56B3B6 hu63A000 hu63DA55 hu066C78 hu67B84E hu75BE2C hu76CAA5 hu84B706 hu91BD69 hu00147A hu340C7A hu363FD6 hu448C4B hu459AD0 hu0515BA hu524B5B hu654B61 hu781EE2 hu1097B2 hu1187FF hu3458D8 hu7504E8 hu9689C4 hu30888B hu59141C hu96713F hu297562 hu345185 hu352868 hu775356 hu868880 hu993257 huA5FD8B huA720D3 huA62230 huAE4A11 huB2A9E7 huB7EC37 huB59C05 huB63C0C huB828CB huBC03A7 huBD9C9B huBFEDCE huC4A276 huC82AA9 huCCA261 huD0D79A huD3E181 huD4F7DB huD57BBF huD0449C huD7960A huD52556 huE4CA90 huE24396 huE31062 huEBD467 huF2DA6F huF9E138 huE9E777 huEAA57B hu5AE862 hu02AB06 hu53075C hu60AB7C hu69073E hu6A0E65 hu939B7C hu9A0F06 huAA16BD huAF3C63 huBE28C7 huC92BC9 huDF04CC huEDF7DA hu34D5B9 hu2FEC01 hu35071E hu5A0DFE hu619F51 hu66330E hu8D6607 huD37D14 huD58ABC huDF9008 huE4DAE4 huDDEC1D huF06AD0 hu77CC58 hu394092 huAC827A hu0A6570 hu0C1563 hu0D2DBE hu1213DA hu155D20 hu1712BC hu2A4D22 hu30F119 hu3800D8 hu3DDADF hu3F7CB4 hu4B0812 hu4C20BC hu4CA5B9 hu4F8813 hu57C8BE hu5BB600 hu5CABA7 hu5DE8CD hu72110E hu786B4C hu79E3C7 hu8A5FBF hu8D1A62 hu925B56 hu92F252 huA27736 huA33758 huAA53E0 huC170B1 huCFD853 huD00199 huEC6EEC huF0B9DB huF9E172 huFAA0CF hu41D90F hu998A3D huDD6E7A

LIMITATIONS

An independent big enough dataset is required to validate results.

DECLARATIONS OF INTERESTS

The author has no competing interests to declare.

FUNDING

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.