

Supplemental information

Germline *HLA* heterozygosity is associated with decreased lung cancer risk

Taotao Tan, Vikram R. Shaw, Jinyoung Byun, Hyun-Sung Lee, Younghun Han, Yafang Li, Rayjean J. Hung, David C. Christiani, Xin-An Wang, Mattias Johansson, Xiangjun Xiao, David Zaridze, Stig Egil Bojesen, Sanjay Shete, Demetrios Albanes, Melinda C. Aldrich, Adonina Tardon, Guillermo Fernandez-Tardon, Loïc Le Marchand, Gad Rennert, Heike Bickeböller, H.-Erich Wichmann, Angela Risch, John K. Field, Michael Davies, Penella Woll, Lambertus A. Kiemeny, Aage Haugen, Shanbeh Zienolddiny, Stephen Lam, Mikael Johansson, Kjell Grankvist, Matthew B. Schabath, Angeline Andrew, Philip Lazarus, Susanne M. Arnold, Dakai Zhu, Maria Teresa Landi, James McKay, Christopher Amos, and Chao Cheng

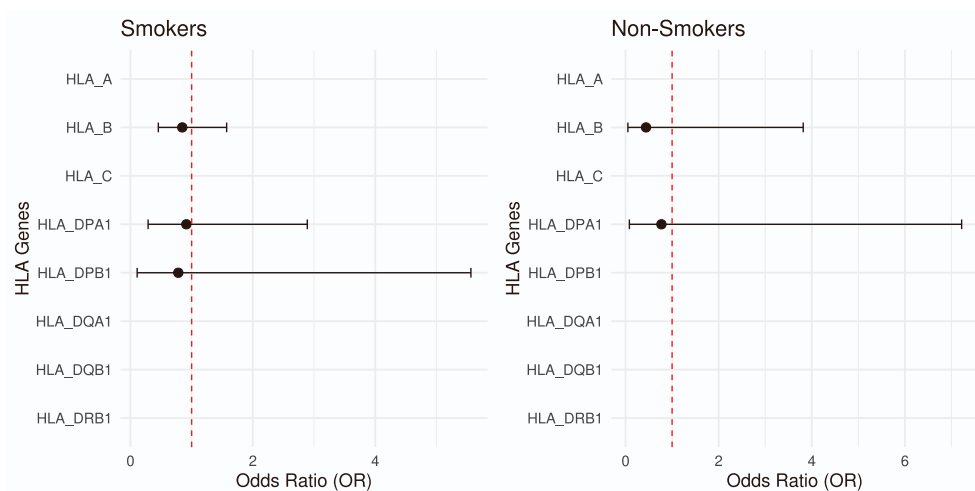


Figure S1A. Effects of Gain of HLA allele (GoA). We ran logistic regression for each HLA gene using: $\text{logit} [\text{Disease}] = \text{HLA-gene} + \text{PC} + \text{age} + \text{sex}$. The “Disease” is a binary variable to indicate if an individual has any form of lung cancer. The “HLA-gene” is a factor variable includes “Homozygosity”, “Heterozygosity”, “GoA”, “LoA”. The OR for GoA in this figure is established by comparing GoA to homozygosity. Missing values lacked a large enough sample size for a reliable estimate. 95% confidence intervals are shown in the panels.

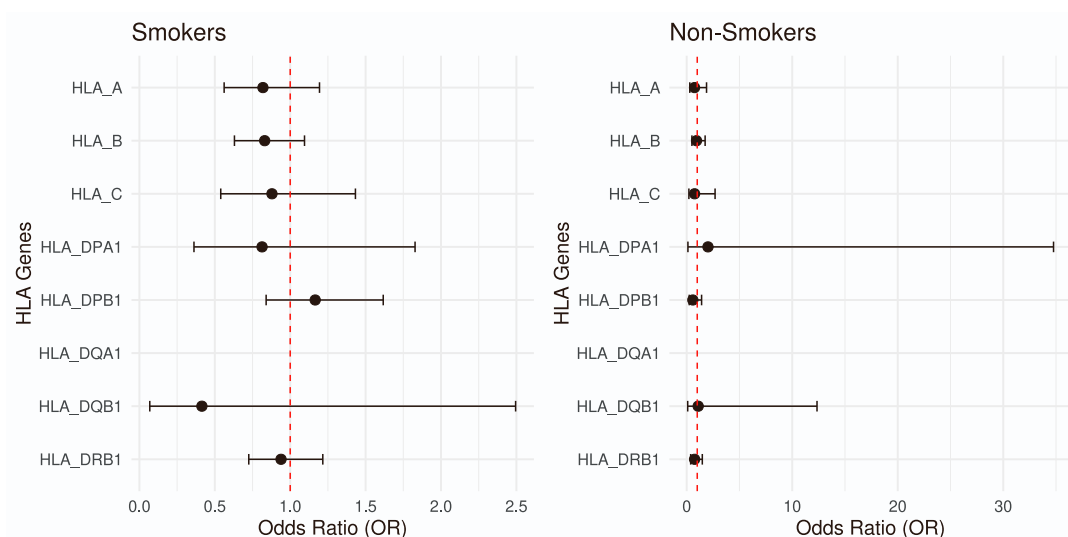


Figure S1B. Effects of Loss of HLA allele (LoA). We ran logistic regression for each HLA gene using: $\text{logit} [\text{Disease}] = \text{HLA-gene} + \text{PC} + \text{age} + \text{sex}$. The “Disease” is a binary variable to indicate if an individual has any form of lung cancer. The “HLA-gene” is a factor variable includes “Homozygosity”, “Heterozygosity”, “GoA”, “LoA”. The OR for LoA in this figure is established by comparing LoA to homozygosity. Missing values lacked a large enough sample size for a reliable estimate. 95% confidence intervals are shown in the panels.

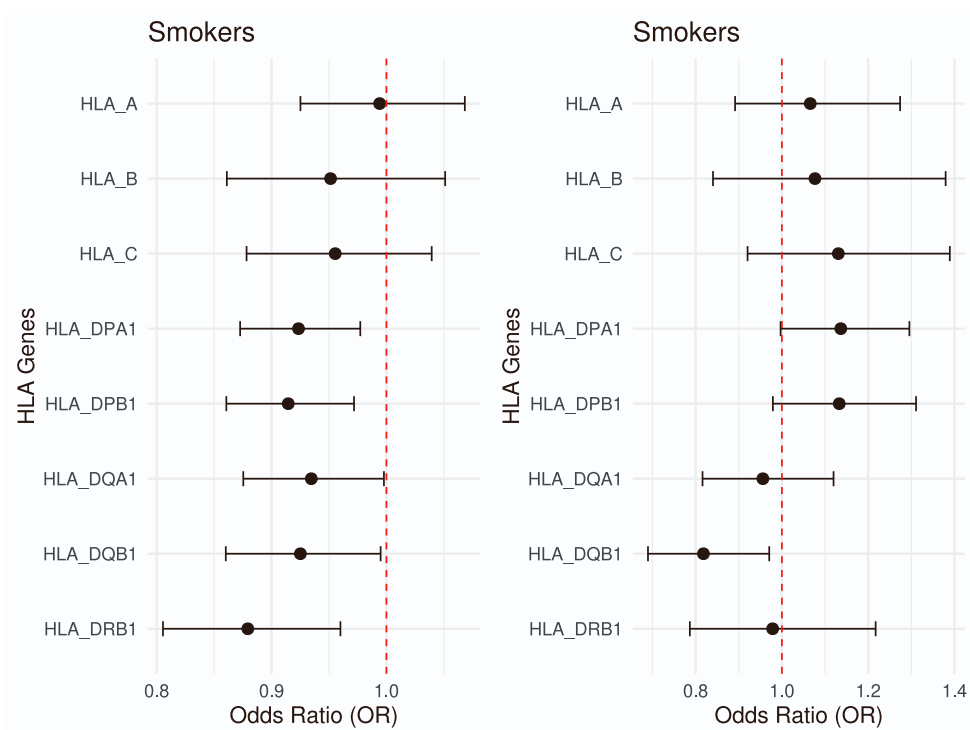


Figure S1C. Effects of Heterozygosity for HLA, removing GoA and LoA samples. After removing GoA and LoA individuals, we repeated the logistic regression found in Figure 2 for each HLA gene using: $\text{logit} [\text{Disease}] = \text{HLA-gene} + \text{PC} + \text{age} + \text{sex}$. The “Disease” is a binary variable to indicate if an individual has any form of lung cancer. The “HLA-gene” is a factor variable including only “Homozygosity” and “Heterozygosity”. 95% confidence intervals are shown in the panels.

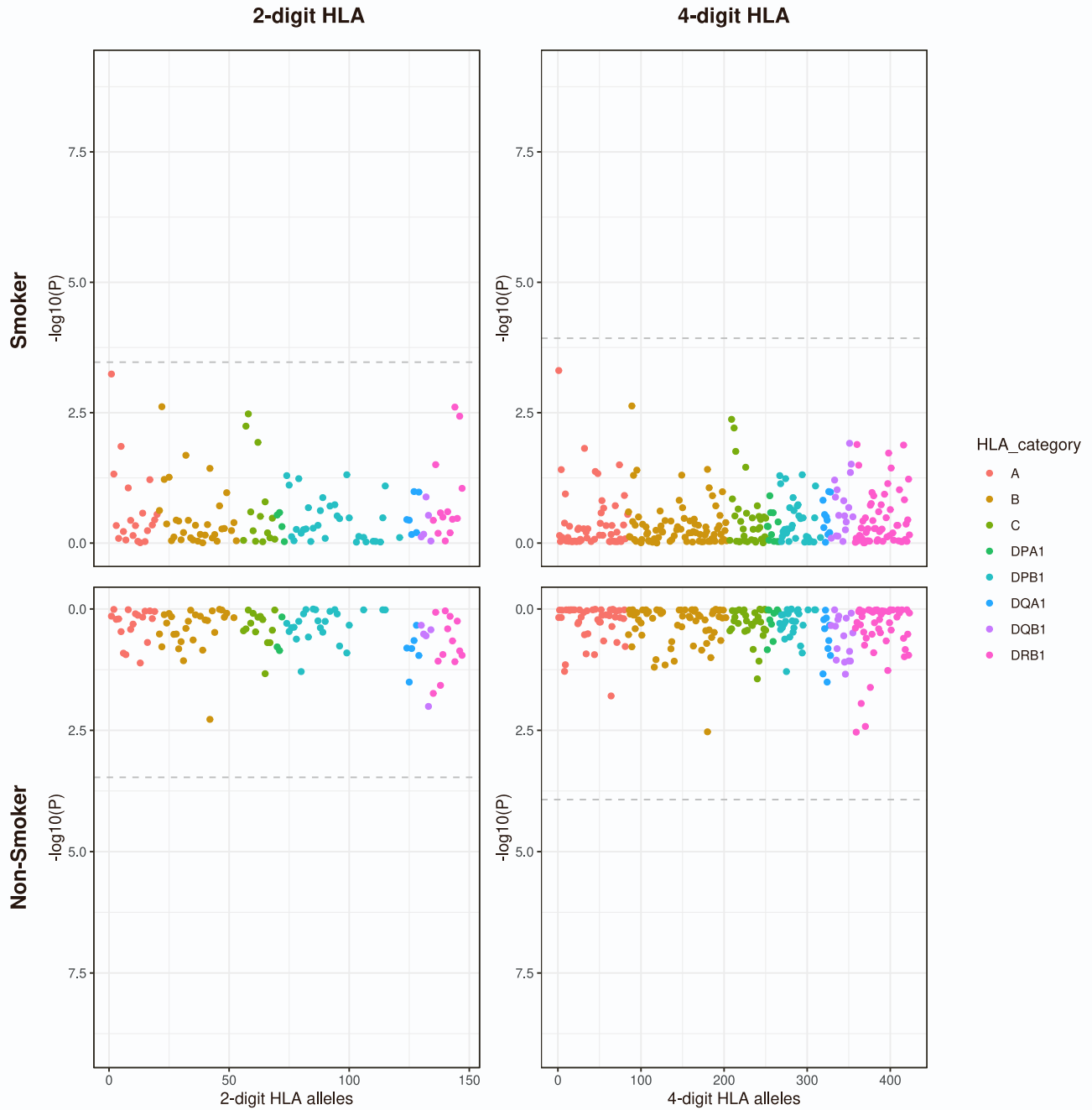


Figure S2. Logistic regression analysis for each 2-digit/4-digit HLA alleles in ADC: $\text{logit}[\text{Disease}] = \text{HLA-allele} + \text{HLA-gene} + \text{PC} + \text{age} + \text{sex}$. “HLA-allele” is the dosage of any 2-digit/4-digit HLA alleles (e.g. *HLA:A*01* dosage, which consists of 0, 1, 2). “HLA-gene” is a factor variable for the corresponding HLA gene (e.g. *HLA:A*, which consists of “Homozygous”, “Heterozygous”, “GoA”, “LoA”). The OR and p-values are reported for all 2-digit/4-digit HLA alleles. The $-\log_{10}(\text{p-value})$ are visualized for each HLA alleles (top row: smokers; bottom row: non-smokers; left column: 2-digit HLA alleles; right column: 4-digit HLA alleles).