**Supplemental Materials for**

**“Exploring genetic influences on adverse outcome pathways using heuristic simulation and graph data science”**

**Authors:**

**Joseph D. Romano1,2,3, Liang Mei4, Jonathan Senn4, Jason H. Moore5, Holly M. Mortensen6,\*.**

**Author Affiliations:**

**1** Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA.

**2** Center of Excellence in Environmental Toxicology, University of Pennsylvania, Philadelphia, PA.

**3** Department of Biostatistics, Epidemiology, & Informatics, University of Pennsylvania, Philadelphia, PA.

**4** Oak Ridge Associated Universities, Oak Ridge, TN.

**5** Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA.

**6** United States Environmental Protection Agency, Office of Research and Development, Center for Public Health and Environmental Assessment, Research Triangle Park, NC.

**\*Corresponding Author**: mortensen.holly@epa.gov

**Propensity Score Matching and Patient Demographics**

The figures below – produced using the ‘pymatch’ package for the Python programming language – show the balance of demographic and clinical covariates before and after applying propensity score matching (PSM) to the case (LC) and control (healthy) datasets (Benedetto et al. 2018). PSM is considered successful when covariate distributions that are significantly different (in this case, determined via chi-square test for independence) are no longer significantly different following matching.

For the purposes of this study, we applied PSM to 5 covariates that are highly relevant to LC risk: **a.)** Sex, **b.)** diabetes status, **c.)** patient-reported ethnicity, **d.)** age, and **e.)** Townsend deprivation index (a composite score representing socioeconomic deprivation, which is well-represented in the UKBB) (Townsend 1987). Sex and diabetes status are modeled as binary (0/1) features, while Townsend deprivation index is a continuous measure with a minimum value of 0 (there is no fixed maximum, but the UKBB reports a maximum value of 11.0013). Age is modeled as a continuous variable while ethnicity is modeled as a categorical variable.

In general, LC is more prevalent in males and in individuals with diabetes mellitus (Li, Wang, and Gao 2017). LC is also more prevalent in socioeconomically deprived populations, and tends to be at a more advanced stage upon initial diagnosis. Additionally, since deprived and disadvantaged populations tend to be less well represented in observational health data sets overall, PSM on the Townsend index improves generalizability of the results to those populations and advances social justice.

**a.) Sex:**





**b.) Diabetes**:





**c.) Ethnicity:**





The UK Biobank uses a proprietary coding system to specify ethnicity concepts. The meaning of these codes (used in the original dataset and in the figure above) are defined as follows:

|  |  |
| --- | --- |
| Code | Ethnicity |
| -3 | Prefer not to answer |
| -1 | Do not know |
| 1 | White |
| 2 | Mixed |
| 3 | Asian or Asian British |
| 4 | Black or Black British |
| 5 | Chinese |
| 6 | Other ethnic group |
| 1001 | British |
| 1002 | Irish |
| 1003 | Any other white background |
| 2001 | White and Black Caribbean |
| 2002 | White and Black African |
| 2003 | White and Asian |
| 2004 | Any other mixed background |
| 3001 | Indian |
| 3002 | Pakistani |
| 3003 | Bangladeshi |
| 3004 | Any other Asian background |
| 4001 | Caribbean |
| 4002 | African |
| 4003 | Any other Black background |

**d.) Age**:



**e.) Deprivation Score:**



**Table S1: Full list of LC AOPs with associated SNPs**

Below, we provide the full list of liver cancer AOPs resulting from our query of “hepatocellular carcinoma” and “liver cancer” in the AOP-DB. Also included are the AOP’s ‘wiki\_status’, indicating the AOP’s current stage of development. Finally, all SNPs of interest are listed for each AOP. It should be noted that some AOPs have no associated SNPs at the current time, although this may be an artifact of the manual development process for AOPs that is based on current (and often incomplete) mechanistic knowledge.

|  |  |  |  |
| --- | --- | --- | --- |
| **aop\_id** | **aop\_name** | **SNPs of interest** | **Key Events (KEs)****\*: KE is a Molecular Initiating Event (MIE)****†: KE is an Adverse Outcome (AO)** |
| 1 | Uncharacterized liver damage leading to hepatocellular carcinoma | N/A | \*Unknown;Hyperplasia;Cell proliferation in the absence of cytotoxicity;†Promotion; Hepatocellular carcinoma |
| 27 | Cholestatic Liver Injury induced by Inhibition of the Bile Salt Export Pump (ABCB11) | Rs10172795, rs78391298, rs4619568Rs72884586, rs72887636, rs200569798Rs6433102, rs4143279, rs57495486Rs13386120, rs11895864, rs78545355Rs2239596, rs185809151, rs16856247Rs16856332, rs552976, rs2287623Rs3755157, rs2241339, rs72623176Rs2287622, rs563694, rs569805Rs11568377, rs16823014, rs10177080Rs2389606, rs853789, rs10176901Rs7576531, rs2287621, rs3770585Rs112259853, rs7306032, rs1500658Rs397964285, rs12312626, rs137855310Rs1156438, rs1828891, rs11110067Rs2195249, rs11109935, rs61945564Rs11109969, rs35724, rs12305294Rs113120051, rs12306800, rs2253039Rs58050306, rs12230262, rs139792030Rs35738, rs4357750, rs551783606rs75061399 | \*Inhibition, Bile Salt Export Pump (ABCB11);Activation of specific nuclear receptors, Transcriptional change;Bile accumulation, Pathological condition;Release, Cytokine;Increase, Inflammation;Production, Reactive Oxygen species;Peptide Oxidation;†Cholestasis, Pathology |
| 31 | Oxidation of iron in hemoglobin leading to hematotoxicity | N/A | \*Parent compound is converted to the reactive metabolite and forms free radicals leading to oxidation of heme iron(II) in hemoglobin to iron(III);Altered regulation, Alpha hemoglobin;Propagation, Oxidative stress;Damaging, Red blood cells - Hemolysis;Formation of hemoglobin adducts;Downregulation, Glucose-6-phosphate dehydrogenase;Increase, RBC congestion in liver;Increase, Liver and splenic hemosiderosis;Methemoglobinemia, decreased hemoglobin, hematocrit, red blood cell number;†Cyanosis occurs |
| 37 | PPARalpha activation leading to hepatocellular adenomas and carcinomas in rodents | Rs4253772, rs9615264, rs9626751Rs4253735, rs4253755, rs28669215 | \*Activation, PPARα;Increase, Phenotypic enzyme activity;Increase, cell proliferation (hepatocytes);Increase, Clonal Expansion of Altered Hepatic Foci;†Increase, hepatocellular adenomas and carcinomas |
| 38 | Protein Alkylation leading to Liver Fibrosis | N/A | \*Alkylation, Protein;Cell injury/death;Tissue resident cell activation;Increased Pro-inflammatory mediators;Activation, Stellate cells;Accumulation, Collagen;†Liver fibrosis |
| 41 | Sustained AhR Activation leading to Rodent Liver Tumours | Rs60202136, rs79596953, rs61182683Rs2731557, rs1031275, rs75014006Rs847424, rs500907, rs28393181Rs17344091, rs55880145, rs11773342Rs115256444, rs11772601, rs139191970Rs77945258, rs4719525, rs7459272Rs58221005, rs7803752, rs7810190Rs10499477, rs73679834, rs847375Rs71540771, rs4476901, rs62446922Rs80207581, rs6978003, rs62446372Rs62441123, rs76937714, Rs57496705, rs143872172, rs1149515, rs200062544Rs17137033, rs149459500, rs76281236Rs706042, rs4721601, rs117263259Rs1178373, rs191334862, rs138995679Rs12670403, rs4410790, Rs6968554, rs10275488, rs2892838, Rs11400459, rs3944100, Rs28549925, rs1525735,Rs17137566, rs1636744, Rs117132860, rs13236243, Rs6968865, rs6961860, Rs145145722, rs1029576, Rs74500669, rs79060196, Rs10252701, rs10277582.Rs10281571, rs2106727, Rs62444550, rs866423, Rs1721043, rs73683618 | \*Activation, Long term AHR receptor driven direct and indirect gene expression changes;Hepatotoxicity, Hepatopathy, including a constellation of observable effects;Changes/Inhibition, Cellular Homeostasis and Apoptosis;Alterations, Cellular proliferation / hyperplasia;†Formation, Hepatocellular and Bile duct tumors |
| 46 | AFB1: Mutagenic Mode-of-Action leading to Hepatocellular Carcinoma (HCC) | Rs78378222, rs35850753, rs1625895Rs1641549, rs35119871 | \*Formation, Pro-mutagenic DNA Adducts;Increased, Induced mutations in critical genes;Metabolism of AFB1, Production of Reactive Electrophiles;Clonal Expansion/Cell Proliferation, to form Altered Hepatic Foci (AHF);Increased, Insufficient repair or mis-repair of pro-mutagenic DNA adducts;†Tumorigenesis, Hepatocellular carcinoma |
| 107 | Constitutive androstane receptor activation leading to hepatocellular adenomas and carcinomas in the mouse and the rat | Rs4477295, rs140578039, rs12094940Rs3813620, rs116791819, rs72714971Rs11586556, rs12028921, rs80168109Rs11265567, rs10157070, rs75131185Rs2025516, rs191551653, rs60315407,Rs2165088, rs12133979. Rs8179404. Rs201221767, Rs181431636, rs71517297Rs115827564, rs200745199, Rs77032824, rs78736195, rs80147258, rs74555614Rs12069336, rs3829793. rs111308195Rs7530560, rs112608874, rs143637329Rs7512905, rs12029217, rs16832317Rs3813611, rs352688, Rs138860776, rs9427063, rs77434773, Rs147608570, rs115624142, rs397797989. rs4073054rs2502815 | \*Activation, Constitutive androstane receptor;Altered gene expression specific to CAR activation, Hepatocytes;Increase, cell proliferation (hepatocytes);Increase, Preneoplastic foci (hepatocytes);†Increase, hepatocellular adenomas and carcinomas |
| 108 | Inhibition of pyruvate dehydrogenase kinase leading to hepatocellular adenomas and carcinomas (in mouse and rat) | N/A | \*Inhibition, Pyruvate dehydrogenase kinase (PDK) enzyme;Increase, Cytotoxicity;Peptide Oxidation;Increased, Induction of pyruvate dehydrogenase (PDH);Increase, Oxidative metabolism;†Increase, hepatocellular adenomas and carcinomas |
| 117 | Androgen receptor activation leading to hepatocellular adenomas and carcinomas (in mouse and rat) | rs5031002 | \*Activation, Androgen receptor;Increase, cell proliferation (hepatocytes);Increase, Preneoplastic foci (hepatocytes);†Increase, hepatocellular adenomas and carcinomas |
| 130 | Phospholipase A inhibitors lead to hepatotoxicity | N/A | \*Inhibition, Phospholipase A;Damage, Lipid bilayer;Disturbance, Lysosomal function;Injury, Mitochondria;Occurrence, Cytoplasmic vacuolization (hepatocyte);Occurrence, Ballooning degeneration (hepatocyte);Occurrence, Cytoplasmic vacuolization (kupffer cell);Induction, Microvesicular fat;Formation, Mallory body;Occurrence, Cytoplasmic vacuolization (Bile duct cell);†Formation, Liver fibrosis |
| 144 | Endocytic lysosomal uptake leading to liver fibrosis | N/A | \*Endocytotic lysosomal uptake;Disruption, Lysosome;Mitochondrial dysfunction 1;Cell injury/death;Increased Pro-inflammatory mediators;Leukocyte recruitment/activation;Activation, Stellate cells;Accumulation, Collagen;†Liver fibrosis |
| 220 | Cyp2E1 Activation Leading to Liver Cancer | rs8005745 | \*Activation of Cyp2E1;Oxidative stress;Hepatocytotoxicity;Induction, persistent proliferation/sustained proliferation;†Liver Cancer |
| 273 | Mitochondrial complex inhibition leading to liver injury | N/A | \*Increase, Mitochondrial complex III antagonism;\*Mitochondrial Complex IV inhibition;\*Mitochondrial Complex V inhibition;\*Binding of inhibitor, NADH-ubiquinone oxidoreductase (complex I);\*Inhibition, NADH-ubiquinone oxidoreductase (Complex I);\*Binding of inhibitor to mitochondrial complex III;\*Binding of inhibitor to mitochondrial complex IV;\*Binding of inhibitor to mitochondrial complex V;Decrease in mitochondrial oxidative phosphorylation;Increased reactive oxygen species (in the mitochondria);Mitochondrial injury;Impaired, Proteostasis;Cell injury/death;Mitochondrial dysfunction 1;†Necrotic Tissue;†Liver Injury |
| 278 | IKK complex inhibition leading to liver injury | N/A | \*Inhibition, IKK complex;Inhibition, Nuclear factor kappa B;Activation, Caspase 8 pathway;Cell injury/death;Activation, Tissue resident cells (Kuppfer cells);Increase, proinflammatory mediators (TNFalpha);†Necrotic tissue;†Liver Injury |
| 285 | Inhibition of N-linked glycosylation leads to liver injury | N/A | \*Inhibition of N-linked glycosylation;Accumulation of misfolded proteins;Unfolded protein response;Apoptosis;Activation of hepatic stellate cells;Increase, Liver steatosis;†Liver injury |

**Table S2: Synthetic dataset produced by HIBACHI**

Although our results primarily focus on the generative models produced by HIBACHI when fitness is evaluated on our balanced LC data, the original purpose of HIBACHI was to produce synthetic datasets with interactions between features that imitate those generative models. The following is the first 50 lines of a synthetic dataset produced by HIBACHI using the most fit model as reported in **Table 2**. Columns 1-25 represent the 25 LC AOP SNPs included in our analysis (identified here by RSIDs), and column 26 (labeled ‘target’) represents LC status. Note that only synthetic patients with LC (‘1’) are shown in this portion of the synthetic dataset; non-LC (‘0’) patients are present in subsequent lines not included here. The full dataset is available on Figshare, at <https://doi.org/10.6084/m9.figshare.20698504.v1>. Although we do not use this dataset in our analyses, a likely use would be for sharing datasets to train predictive models of LC without exposing patient privacy.



**Table S3: Linkage disequilibrium analysis**

Analysis of linkage disequilibrium using the LDLink web service (Machiela and Chanock 2015) in the 4 SNPs hibachi represented as most important in liver cancer. Also included are the chromosome positions of each SNP, and allele frequencies as represented by the 1000 Genomes Project. R2 is a measure of allele correlation for two genetic variants. D’ is an indicator of allelic segregation of two genetic variants.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **SNP1** | **Chromosome position** | **SNP2** | **Chromosome position** | **Allele frequency** | **r^2** | **D’** |
| Rs563694 | Chr2: 168917561 | Rs552976 | Chr2: 168934928 | Snp1: C=0.158 (792/5008, 1000G)A = 0.842SNP2: A=0.252(1264/5008, 1000G)C/G = 0.748 | 0.419 | 0.868 |
| Rs563694 | Chr2: 168917561 | Rs16856247 | Chr2: 168927903 | SNP1: C=0.158(792/5008, 1000G)A/G = 0.842SNP2: T = 0.137 (688/5008, 1000G) | 0.017 | 0.752 |
| Rs552976 | Chr2: 168934928 | Rs16856247 | chr2:168927903 | snp1:A=0.252 (1264/5008, 1000G)C/G = 0.748 Snp2: T=0.137 (688/5008, 1000G) | 0.054 | 1 |
| rs4410790 | Chr7:17244953 | rs563694 | Chr2: 168917561 | SNP1: C = 0.467(2339/5008)T = 0.533snp2:A=0.252 (1264/5008, 1000G)C/G = 0.748 | 0.0012 | 0.076 |
| rs4410790 | Chr7:17244953 | rs16856247 | chr2:168927903 | SNP1: C = 0.467(2339/5008)T = 0.533SNP2: T = 0.137 (688/5008, 1000G)C = 0.863 | 0.0001 | 0.0228 |
| rs4410790 | Chr7:17244953 | Rs552976 | Chr2: 168934928 | SNP1: C = 0.467(2339/5008)T = 0.533snp2:A=0.252 (1264/5008, 1000G)C/G = 0.748  | 0.0009 | 0.047 |

**Figure S1: Liver Cancer AOP network**

A network detailing all the AOPs, genes, and SNPs involved in liver cancer using data from the AOP-DB. Red indicates AOPs, blue indicates SNPs, and green indicates genes.



**Bibliography**

Benedetto, Umberto, Stuart J. Head, Gianni D. Angelini, and Eugene H. Blackstone. 2018. “Statistical Primer: Propensity Score Matching and Its Alternatives.” *European Journal of Cardio-Thoracic Surgery: Official Journal of the European Association for Cardio-Thoracic Surgery* 53 (6): 1112–17. https://doi.org/10.1093/ejcts/ezy167.

Li, Xu, Xiaocong Wang, and Pujun Gao. 2017. “Diabetes Mellitus and Risk of Hepatocellular Carcinoma.” *BioMed Research International* 2017: 5202684. https://doi.org/10.1155/2017/5202684.

Machiela, Mitchell J., and Stephen J. Chanock. 2015. “LDlink: A Web-Based Application for Exploring Population-Specific Haplotype Structure and Linking Correlated Alleles of Possible Functional Variants.” *Bioinformatics (Oxford, England)* 31 (21): 3555–57. https://doi.org/10.1093/bioinformatics/btv402.

Townsend, Peter. 1987. “Deprivation.” *Journal of Social Policy* 16 (2): 125–46. https://doi.org/10.1017/S0047279400020341.