The Systems Toxicology Challenge

The computational challenge: Markers of Exposure Response Identification

TECHNICAL DOCUMENT

Table of Contents

1 Dataset description

1.1 Clinical studies

1.1.1 QASMC study (dset1 provided as training set)

The Queen Ann Street Medical Center (QASMC) clinical case–control study (1) was conducted between July 2011 and December 2012 at The Heart and Lung Centre (London, UK), after approval from the National Health Service (NHS) Black County Ethics Committee and in strict compliance with the International Conference on Harmonisation—Good Clinical Practice (ICH-GCP) guidelines. The study was registered at ClinicalTrials.gov with the identifier NCT01780298. The study aimed to identify a biomarker or a panel of biomarkers that would enable differentiation between smokers with chronic obstructive pulmonary disease (COPD) (i.e., cigarette smoke (CS) with a ≥10 pack/year smoking history and COPD disease classified as GOLD Stage 1 or 2) and three comparative groups of matched subjects: smokers (S), former smokers (FS), and never smokers (NS). All smoking subjects (S and FS) had a smoking history of at least 10 pack-years. FS have quitted smoking for at least 1 year (~78% of FS have quitted for more than 5 years). Sixty subjects in each group were enrolled (240 subjects in total). The 240 patients included males (58%) and females (42%) aged between 40 and 70 years. All subjects were matched by ethnicity, gender, and age (within 5 years) with the recruited COPD subjects.

1.1.2 BLD-SMK-01 (dset2 provided as test set)

The blood gene expression dataset, BLD-SMK-01, was produced from PAXgene blood samples obtained from a banked repository (BioServe Biotechnologies Ltd., Beltsville, MD, USA) based on well-defined inclusion criteria. At the time of sampling, the subjects were between 23 and 65 years of age. Subjects with a disease history and those taking prescription medications were excluded. Smokers (S) had smoked at least 10 cigarettes daily for at least three years. Former smokers (FS) had ceased smoking at least two years prior to sampling and before quitting had smoked at least 10 cigarettes daily for at least three years. Smokers and never smokers (NS) were matched by age and gender, while former smokers could not be properly matched due to lower number of samples available for this group.

1.1.3 ZRHR-Reduced exposure (REX)C-03-EU and -04-JP studies (dset3a and b provided as verification sets)

Both REXC studies were randomized, controlled, open-label, 3-arm parallel group, and single-center studies. These studies were performed to demonstrate reductions in exposure to selected smoke constituents in smoking, healthy subjects switching to the tobacco heating system THS2.2 (Switch), a candidate modified-risk tobacco product, or smoking abstinence (Cess), compared with continuing to use conventional cigarettes (CC, considered

as smokers (S)), for 5 days in confinement. The studies were conducted in Europe and Japan and were registered at ClinicalTrials.gov with the identifier NCT01959932 and NCT01970982, respectively.

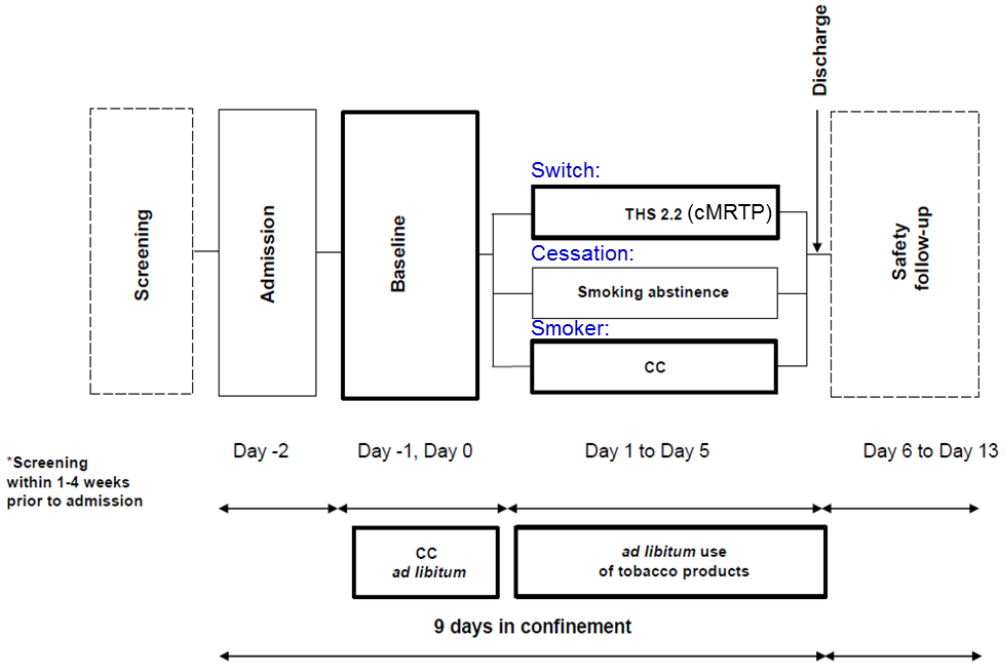

Figure 1: Schema of the ZRHR-REXC-03-EU and ZRHR-REXC-04-JP clinical study design

1.2 Mouse inhalation studies

1.2.1 Mouse C57Bl6-pMRTP-SW inhalation study (dset4 provided as training and verification sets)

A 7-month cigarette smoke inhalation study was conducted with female C57BL/6 mice (2). The study design included 5 different groups. Among those, 3 groups are provided as training set: reference cigarette 3R4F-exposed mice (3R4F), mice exposed to air after 2months of 3R4F exposure (Cessation or Cess), and mice continuously exposed to air (Sham); 2 groups are provided as verification set: a group in which mice were continuously exposed to a prototype modified-risk tobacco product (pMRTP) as mentioned in (2)), and a group of mice which switched to a pMRTP after 2 months of exposure to 3R4F (Switch). For each group, blood samples were collected from 7 to 10 animals at 2 months, 3 months, 5 months and 7 months. For the 3R4F exposure, pMRTP, and Sham arms, samples were also collected after 4 months.

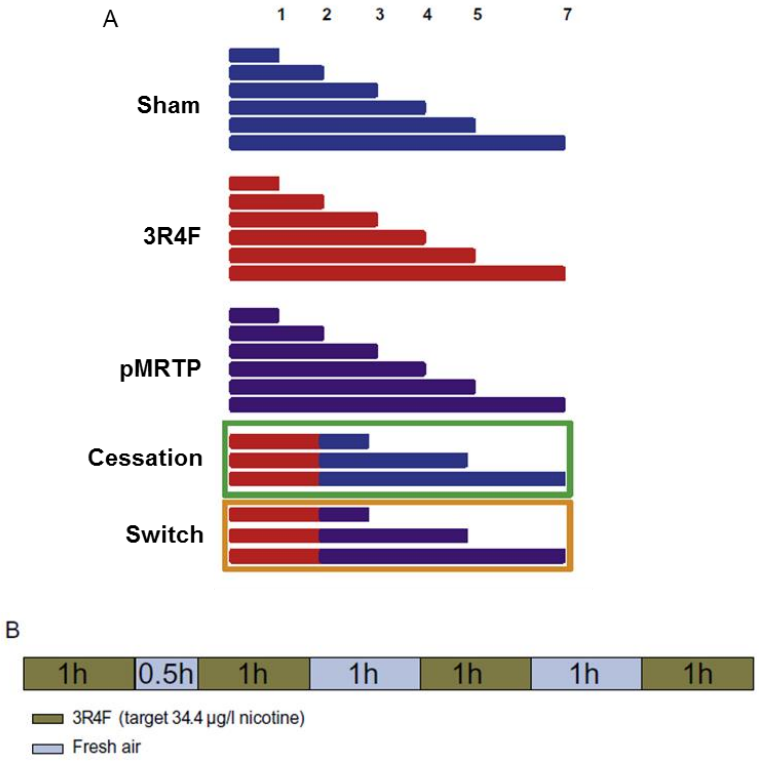

Figure 2: Experimental design and allocation of mice to the exposure groups. Mice were exposed to filtered, fresh air (sham), to CS from 3R4F, or aerosol from a prototype MRTP for up to 7 months; additional groups were exposed to 3R4F for 2 months and then to fresh air (Cessation or Cess) or to a prototype MRTP aerosol (Switch). (B) Daily exposure schedule. Four 1-h blocks of CS/aerosol exposure were separated by exposure breaks with fresh air to avoid a build-up of excessive COHb concentrations.

1.2.2 Mouse ApoE-THS2.2-SW inhalation study (dset5 provided as test and verification set)

For this study, female Apoe−/− mice were randomized into five groups (Figure 3): sham (exposed to air), 3R4F (exposed to CS from the reference cigarette 3R4F), THS2.2 (exposed to mainstream aerosol from THS2.2 at

nicotine levels matched to those of 3R4F), smoking cessation (Cess), and switching to THS2.2 (Switch). Mice from the sham, 3R4F, and THS2.2 groups were exposed to fresh air, CS from 3R4F, or THS2.2 aerosol, respectively, for up to eight months. To model the effects of smoking cessation and switching to THS2.2, mice from the cessation and switch groups were first exposed to 3R4F for 2 months and then switched to air or THS2.2 aerosol, respectively, for up to six additional months (Figure 3). Female mice were chosen because of their possible increased susceptibility to develop emphysema (3,4).

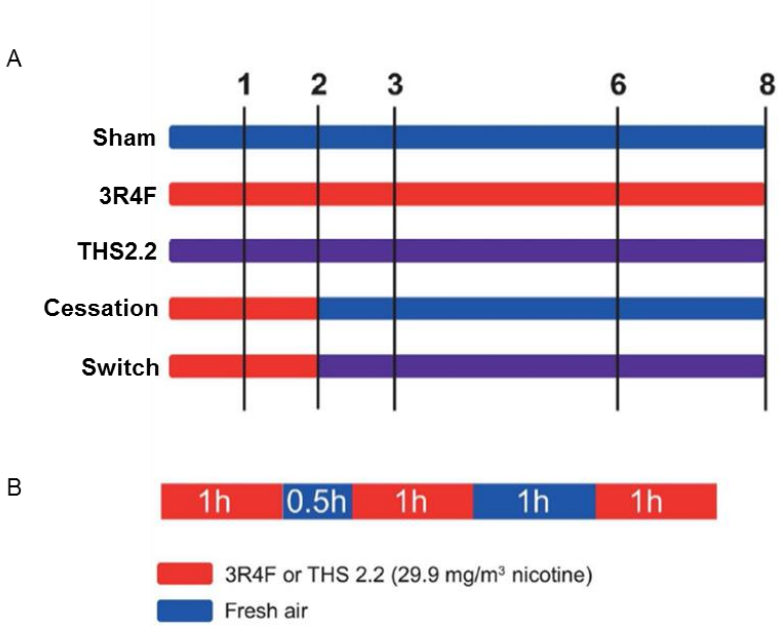

Figure 3: Experimental design and allocation of mice to the exposure groups. Mice were exposed to filtered, fresh air (sham), to CS from 3R4F, or aerosol from THS2.2 for up to 8 months; additional groups were exposed to 3R4F for 2 months and then to fresh air (Cessation or Cess) or to THS2.2 aerosol (Switch). (B) Daily exposure schedule.

Of note, the size of each group is provided for training datasets, but not for test and verification datasets before the challenge on purpose to limit the release of information that may bias/influence class prediction.

1.2.3 Additional public and/or private datasets as training sets (optional)

The participants have the freedom to use additional relevant public and/or private gene expression datasets to train their model(s). As suggestions, blood-cell related gene expression datasets such as GSE42057 and GSE15289 can be found in GEO (http://www.ncbi.nlm.nih.gov/gds/?term=GEO).

2 Transcriptomics data generation and processing

2.1 RNA isolation from human blood samples

For each clinical and in vivo study, a randomization plan has been prepared and followed for the RNA extraction and gene expression workflows.

For clinical (human) studies, total RNAs (including microRNAs) were isolated using the PAXgene Blood miRNA Kit (catalog number 763134; Qiagen) according to the manufacturer's instructions.

For in vivo (mouse) studies, total RNA (including microRNAs) were isolated using the Qiagen RNeasy Protect Animal Blood Kit (Cat. 73224) according to the manufacturer's instructions.

The concentration and purity of the RNA samples were determined using a UV spectrophotometer (NanoDrop® 1000 or Nanodrop 8000; Thermo Fisher Scientific, Waltham, MA, USA) by measuring the absorbance at 230, 260, and 280 nm. RNA integrity was further checked using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Only RNAs with an RNA integrity number >6 were processed for further analysis.

2.2 RNA preparation and hybridization on Affymetrix chip

Targets were prepared from 80 ng of RNA using the Ovation® Whole Blood Reagent and Ovation RNA Amplification System V2 (NuGEN, AC Leek, The Netherlands). The quantity of cDNA was measured with a SpectraMax® 384Plus microplate reader (Molecular Devices, Sunnyvale, CA, USA). The cDNA quality was determined by assessing the size of unfragmented cDNA using the Fragment analyzer (Advanced analytical, Ankeny, IA, USA). The size distribution of the final fragmented and biotinylated product was also monitored using electropherograms on the Agilent 2100 Bioanalyzer (Santa Clara, CA, USA). After fragmentation and labeling the cDNA fragments were hybridized on a GeneChip® Human Genome U133 Plus 2.0 Array or GeneChip® Mouse Genome 430 2.0 Array (Affymetrix) according to the manufacturer's guidelines.

For the QASMC study, the target preparation from blood samples and its hybridization on GeneChip® Human Genome U133 Plus 2.0 Array (Affymetrix) were performed by AROS Applied Biotechnology AS (Aarhus, Denmark).

2.3 Raw data preprocessing and QC

Raw data (CEL files) were processed and normalized per dataset in the R environment (v3.1.2, (5)) using frozen Robust Microarray Analysis, fRMA v1.18 (6). Frozen parameter vectors for mouse (mouse4302frmavecs v1.3.0, (7)) and human (hgu133plus2frmavecs v1.3.0, (8)) were used by the frma and GNUSE functions. The custom brainarray cdf files for mouse (mouse4302mmentrezgcdf v16.0.0) and human (hgu133plus2hsentrezgcdf v16.0.0 (9)) were used for affymetrix probe-to-entrez gene ID mapping and resulting in one probe set for one gene relationship.

Data were also quality checked. The quality check step removed all CEL files which did not pass one of the following cutoffs for the criteria described below.

- For a given probeset *j*, the Normalized Unscaled Standard Error (NUSE) provides a measure of the precision of its expression estimate on a given array, *i*, relative to other arrays in the batch (dataset). Problematic arrays result in higher Standard Error (SE) than the median SE. Arrays are suspected to be of poor quality if either the NUSE median exceeds 1 or arrays have a large interquartile range (IQR) (10). Arrays with NUSE values higher that 1.05 were removed (11,12).
- The Relative Log Expression (RLE) compares for each array the level of intensity of a given probe relative to the median level of intensity for that probe across all *j* arrays. The array-specific distribution of RLE is used to determine if a particular array has predominately low- or high- expressed features. A median RLE not near zero indicates that the number of up-regulated genes does not approximately equal the number of down-regulated genes, and a large RLE IQR indicates that most of the genes are differentially expressed (10). An array with median RLE > 0.1 (in absolute value) was considered an outlier and removed (13).
- Arrays with Median Absolute RLE (MARLE) greater than the median absolute deviation of all array dataset MARLEs divided by the square root of 0.01 (or median(MARLE)/(1.4826*mad(MARLEs)) > 1/sqrt(0.01)) were considered bad quality chip and removed.

2.4 Human-Mouse homology mapping procedure

Mouse genes were homologized to human genes using the NCBI/HCOP mapping file (14). In case of mouse genes mapping to multiple human genes, only human genes matching capitalized mouse genes were retained. To facilitate the dataset handling, human and mouse gene expression datasets are provided with mouse gene symbols for both. However, the final table that describes orthologous genes between human and mouse retained for the datasets is given as additional information on the website.

3 Metadata

For human and mouse studies, the gender is provided for each sample in the metadata, however the use of this information by participants is optional.

4 Test and verification datasets release for prediction

The test and verification datasets will be released altogether, however split into two subsets (A and B) provided sequentially as described in the section Get started. As described on the website, only samples from the test datasets will be used to score participants' class predictions.

5 References

1. Titz, B., Sewer, A., Schneider, T., Elamin, A., Martin, F., Dijon, S., Luettich, K., Guedj, E., Vuillaume, G., Ivanov, N. V., Peck, M. J., Chaudhary, N. I., Hoeng, J., and Peitsch, M. C. (2015) Alterations in the sputum proteome and transcriptome in smokers and early-stage COPD subjects. *Journal of proteomics* **128**, 306-320

2. Phillips, B., Veljkovic, E., Peck, M. J., Buettner, A., Elamin, A., Guedj, E., Vuillaume, G., Ivanov, N. V., Martin, F., Boue, S., Schlage, W. K., Schneider, T., Titz, B., Talikka, M., Vanscheeuwijck, P., Hoeng, J., and Peitsch, M. C. (2015) A 7-month cigarette smoke inhalation study in C57BL/6 mice demonstrates reduced lung inflammation and emphysema following smoking cessation or aerosol exposure from a prototypic modified risk tobacco product. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association* **80**, 328-345

3. Hantos, Z., Adamicza, A., Janosi, T. Z., Szabari, M. V., Tolnai, J., and Suki, B. (2008) Lung volumes and respiratory mechanics in elastase-induced emphysema in mice. *Journal of applied physiology* **105**, 1864-1872

4. March, T. H., Wilder, J. A., Esparza, D. C., Cossey, P. Y., Blair, L. F., Herrera, L. K., McDonald, J. D., Campen, M. J., Mauderly, J. L., and Seagrave, J. (2006) Modulators of cigarette smoke-induced pulmonary emphysema in A/J mice. *Toxicological sciences : an official journal of the Society of Toxicology* **92**, 545-559

5. Team, R. D. C. (2008) R: A language and environment for statistical computing. *R Foundation for Statistical Computing*

6. McCall, M. N., Bolstad, B. M., and Irizarry, R. A. (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics* **11**, 242-253

7. Matthew N. McCall, R. A. I. mouse4302frmavecs: Vectors used by frma for microarrays of type mouse4302. *R package version 1.3.0.*

8. Matthew N. McCall, R. A. I. hgu133plus2frmavecs: Vectors used by frma for microarrays of type hgu133plus2. *R package version 1.3.0.*

9. http://brainarray.mbni.med.umich.edu/Brainarray/default.asp.

10. McCall, M. N., Murakami, P. N., Lukk, M., Huber, W., and Irizarry, R. A. (2011) Assessing affymetrix GeneChip microarray quality. *BMC bioinformatics* **12**, 137

11. Heber, S., and Sick, B. (2006) Quality assessment of Affymetrix GeneChip data. *Omics : a journal of integrative biology* **10**, 358-368

12. Julia Brettschneider, F. C., Benjamin M. Bolstad, Terence P. Speed. (2008) Quality assessment for short oligonucleotide microarray data. *Technometrics* **50**, 241-264

13. Kauffmann, A., Gentleman, R., and Huber, W. (2009) arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. *Bioinformatics* **25**, 415-416

14. Eyre, T. A., Wright, M. W., Lush, M. J., and Bruford, E. A. (2007) HCOP: a searchable database of human orthology predictions. *Briefings in bioinformatics* **8**, 2-5