T1000: A reduced gene set prioritized for toxicogenomic studies

Othman Soufan [1], **Jessica Ewald** [2], **Charles Viau** [1], **Doug Crump** [3], **Markus Hecker** [4], **Niladri Basu** [Corresp., 2], **Jianguo Xia** [Corresp. 5]

[1] Institute of Parasitology, McGill University, Montreal, Canada

[2] Faculty of Agricultural and Environmental Sciences, McGill University, Montreal, Canada

[3] Ecotoxicology and Wildlife Health Division, Environment and Climate Change Canada, National Wildlife Research Centre, Carleton University, Ottawa, Canada

[4] School of the Environment & Sustainability and Toxicology Centre, University of Saskatchewan, Saskatoon, Canada

[5] Institute of Parasitology and Department of Animal Science, McGill University, Montreal, Canada

Corresponding Authors: Niladri Basu, Jianguo Xia
Email address: niladri.basu@mcgill.ca, jeff.xia@mcgill.ca

There is growing interest within regulatory agencies and toxicological research communities to develop, test, and apply new approaches, such as toxicogenomics, to more efficiently evaluate chemical hazards. Given the complexity of analyzing thousands of genes simultaneously, there is a need to identify reduced gene sets. Though several gene sets have been defined for toxicological applications, few of these were purposefully derived using toxicogenomics data. Here, we developed and applied a systematic approach to identify 1000 genes (called Toxicogenomics-1000 or T1000) highly responsive to chemical exposures. First, a co-expression network of 11,210 genes was built by leveraging microarray data from the Open TG-GATEs program. This network was then re-weighted based on prior knowledge of their biological (KEGG, MSigDB) and toxicological (CTD) relevance. Finally, weighted correlation network analysis was applied to identify 258 gene clusters. T1000 was defined by selecting genes from each cluster that were most associated with outcome measures. For model evaluation, we compared the performance of T1000 to that of other gene sets (L1000, S1500, Genes selected by Limma, and random set) using two external datasets based on the rat model. Additionally, a smaller (T384) and a larger version (T1500) of T1000 were used for dose-response modeling to test the effect of gene set size. Our findings demonstrated that the T1000 gene set is predictive of apical outcomes across a range of conditions (e.g., *in vitro* and *in vivo*, dose-response, multiple species, tissues, and chemicals), and generally performs as well, or better than other gene sets available.

1   **T1000: A reduced gene set prioritized for toxicogenomic studies**

2   [1]Othman Soufan, [2]Jessica Ewald, [1]Charles Viau, [3]Doug Crump, [4]Markus Hecker, [2,*]Niladri Basu

3   and [1,5,*]Jianguo Xia

4   [1]Institute of Parasitology, McGill University, Montreal, Quebec, Canada; [2]Faculty of

5   Agricultural and Environmental Sciences, McGill University, Montreal, Quebec, Canada;

6   [3]Ecotoxicology and Wildlife Health Division, Environment and Climate Change Canada,

7   National Wildlife Research Centre, Carleton University, Ottawa, Canada; [4]School of the

8   Environment & Sustainability and Toxicology Centre, University of Saskatchewan, Saskatoon,

9   Canada; [5]Department of Animal Science, McGill University, Montreal, Quebec, Canada.

11   *Correspondance: Niladri Basu

12   Email: niladri.basu@mcgill.ca

13   Tel: (514) 398-8642

14   Address: CINE Building,

15   21111 Lakeshore Road

16   Ste. Anne de Bellevue

17   QC, Canada

18   H9X 3V9.

20   *Correspondence: Jianguo Xia

21   Email: jeff.xia@mcgill.ca

22   Tel: (514) 398-8668

23   Address: Office P107, Parasitology Building,

24   21111 Lakeshore Road

25   Ste. Anne de Bellevue

26   QC, Canada

27   H9X 3V9.

Abstract

34   There is growing interest within regulatory agencies and toxicological research communities to

35   develop, test, and apply new approaches, such as toxicogenomics, to more efficiently evaluate

36   chemical hazards. Given the complexity of analyzing thousands of genes simultaneously, there is

37   a need to identify reduced gene sets. Though several gene sets have been defined for

38   toxicological applications, few of these were purposefully derived using toxicogenomics data.

39   Here, we developed and applied a systematic approach to identify 1000 genes (called

40   Toxicogenomics-1000 or T1000) highly responsive to chemical exposures. First, a co-

41   expression network of 11,210 genes was built by leveraging microarray data from the Open TG-

42   GATEs program. This network was then re-weighted based on prior knowledge of their

43   biological (KEGG, MSigDB) and toxicological (CTD) relevance. Finally, weighted correlation

44   network analysis was applied to identify 258 gene clusters. T1000 was defined by selecting

45   genes from each cluster that were most associated with outcome measures. For model evaluation,

46   we compared the performance of T1000 to that of other gene sets (L1000, S1500, Genes selected

47   by Limma, and random set) using two external datasets based on the rat model. Additionally, a

48   smaller (T384) and a larger version (T1500) of T1000 were used for dose-response modeling to

49   test the effect of gene set size. Our findings demonstrated that the T1000 gene set is predictive of

50   apical outcomes across a range of conditions (e.g., *in vitro* and *in vivo*, dose-response, multiple

51   species, tissues, and chemicals), and generally performs as well, or better than other gene sets

52   available.

Introduction

53
54

55   Over the past decade there have been profound steps taken across the toxicological sciences and

56   regulatory communities to help transform conventional toxicity testing largely based on animal

57   models and apical outcome measurements to an approach that is founded on systems biology and

58   predictive science (Kavlock et al. 2018; Knudsen et al. 2015; Villeneuve & Garcia-Reyero

59   2011). On the scientific side, efforts are being exemplified by emergent notions such as the

60   Adverse Outcome Pathway framework (AOP; Ankley et al., 2010) and New Approach Methods

61   (ECHA 2016). On the regulatory side, these are exemplified by changes to, for example,

62   chemical management plans in Canada, the United States and REACH (ECHA 2007) across the

63   European Union.

64

65   A core tenet underlying the aforementioned transformations, as catalyzed by the 2007 U.S.

66   National Research Council report "Toxicity Testing in the 21st Century" (Andersen & Krewski

67   2009), is that perturbations at the molecular-level can be predictive of those at the whole

68   organism-level. Though whole transcriptome profiling is increasingly popular, it still remains

69   costly for routine research and regulatory applications. Additionally, building predictive models

70   with thousands of features introduces problems due to the high dimensionality of the data and so

71   considering a smaller number of genes has the potential to increase classification performance

72   (Alshahrani et al. 2017; Soufan et al. 2015b). Identifying smaller panels of key genes that can be

73   measured, analyzed and interpreted conveniently remain an appealing option for toxicological

74   studies and decision making

75

76   In recent years, several initiatives across the life sciences have started to identify reduced gene

77   sets from whole transcriptomic studies. For example, the Library of Integrated Network-Based

78   Cellular Signatures (LINCS) project derived L1000, which is a gene set of 976 'Landmark'

79   genes chosen to infer the expression of 12,031 other highly connected genes in the human

80   transcriptome (Subramanian et al. 2017). In the toxicological sciences, the U.S. Tox21 Program

81   recently published S1500+, which is a set of 2,753 genes designed to be both representative of

82   the whole-transcriptome, while maintaining a minimum coverage of all biological pathways in

83   Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al. 2007) and Molecular

84   Signatures Database (MSigDB) (Liberzon et al. 2015a). The first 1,500 genes were selected by

85   analyzing microarray data from 3,339 different studies, and the rest were nominated by members

86   of the scientific community (Mav et al. 2018). L1000 and S1500 gene sets were originally

87   proposed to serve a different purpose. The 978 landmark genes of L1000 are chosen to infer

88   expression of other genes more accurately, while genes of S1500 are selected to achieve more

89   biological pathway coverage. Compared to L1000, the S1500 gene set attains more toxicological

90   relevance through the gene nomination phase, though its data-driven approach relies upon

91   microarray data primarily derived from non-toxicological studies. It worth nothing that about

92   33.7% of genes are shared between both signatures. Even though some differences can be

93   realized between L1000 and S1500, they are both strong candidates of gene expression modeling

94   and prediction (Haider et al. 2018).

95

96   The objectives of the current study were to develop and apply a systematic approach to identify

97   highly-responsive genes from toxicogenomic studies, and from these to nominate a set of 1000

98   genes to form the basis for the T1000 (Toxicogenomics-1000) reference gene set. Co-expression

99   network analysis is an established approach using pairwise correlation between genes and

100   clustering methods to group genes with similar expression patterns (van Dam et al. 2018). First, a

101   co-expression network was derived using *in vitro* and *in vivo* data from human and rat studies

102   from the Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (Open TG-

103   GATEs) database. Next, the connections within the co-expression network were adjusted to

104   increase the focus on genes in KEGG pathways, the MSigDB, or the Comparative

105   Toxicogenomics Database (CTD) (Davis et al. 2017). This incorporation of prior biological and

106   toxicological knowledge was motivated by loose Bayesian inference to refine the

107   computationally-prioritized transcriptomic space. Clusters of highly connected genes were

108   identified from the resulting co-expression network, and machine learning models were applied

109   to prioritize clusters based on their association with apical endpoints. Clustering genes based on

110   expression data has been shown to be instrumental in functional annotation and sample

111   classification (Necsulea et al. 2014), with the rationale that genes with similar expression

112   patterns are likely to participate in the same biological pathways (Budinska et al. 2013). From

113   each cluster key genes were identified for inclusion in T1000. Testing and validation of T1000

114   was realized through two separate datasets (one from Open TG-GATEs and one from the U.S.

115   National Toxicology Program) that were not used for gene selection. The current study is part of

116   the larger EcoToxChip project (Basu et al. 2019). For the processed data, user can download all

117   samples processed from https://zenodo.org/record/3359047#.XUcTwpMzZ24. We also deposited

118   source codes and scripts used for the study at https://github.com/ecotoxxplorer/t1000.

119   **Materials & Methods**

120   **Databases and datasets preparation**

121   The derivation of T1000 was based on five public microarray datasets of toxicological relevance

122   (**Table 1**): four datasets from Open TG-GATEs (Igarashi et al. 2014b), and one dataset generated

123   by Thomas *et al* (referred to as the dose-response dataset in this manuscript; GSE45892) (Thomas

124   et al. 2013). **Table 1** provides a summary of all microarray datasets used in this study. For building

125   the initial T1000 gene set, we used three of the four Open TG-GATEs datasets (see datasets 1-3 in

126   **Table 1**).

127
128   **Open TG-GATEs**
129
130   Open TG-GATEs is one of the largest publicly accessible toxicogenomics resources (Igarashi et

131   al. 2014b). This database comprises data from 170 compounds (mostly drugs) with the aim of

132   improving and enhancing drug safety assessment. It contains gene expression profiles and

133   traditional toxicological data derived from *in vivo* (rat) and *in vitro* (primary rat hepatocytes and

134   primary human hepatocytes) studies. To process the raw gene expression data files of Open TG-

135   GATEs, the Affy package (Gautier et al. 2004) was used to produce Robust Multi-array Average

136   (RMA) probe set intensities (Irizarry et al. 2003b). Gene annotation for human and rat was

137   performed using Affymetrix Human Genome U133 Plus 2.0 Array annotation data and

138   Affymetrix Rat Genome 230 2.0 Array annotation data, respectively. Genes without annotation

139   were excluded. When the same gene was mapped multiple times, the average value was used.

140   Finally, all profiles for each type of experiment were joined into a single matrix for downstream

141   analysis.

142   From the training datasets, specific samples were labelled binary as "dysregulated" or "non-

143   dysregulated". Dysregulated refers to exposure cases with potential toxic outcomes and non-

144   dysregulated included controls and exposures with non-toxic outcomes. For the *in vitro* datasets,

145   gene expression changes were associated with lactate dehydrogenase (LDH) activity (%). The

146   activity of LDH, which serves as a proxy for cellular injury or dysregulation, was binarized such

147   that values above 105% and below 95% were considered "dysregulated". While conservative,

148   we note that these cut-off values were situated around the 5% and 95% marks of the LDH

149   distribution curve (see **Supplemental Figure S1** and **Supplemental Information S1** for more

150   details).

151
152   For the *in vivo* datasets (kidney and liver datasets from Open TG-GATEs), gene expression

153   changes were associated with histopathological measures. The magnitude of pathologies was

154   previously annotated into an ordinal scale: present, minimal, slight, moderate and severe

155   (Igarashi et al. 2014a). This scale was further reduced into a binary classification with the first

156   three levels considered "non-dysregulated" while the latter two were considered "dysregulated".

157

158   **Dose-response dataset and benchmark dose (BMD) calculation**

159   The dose-response dataset (Accession No. GSE45892), was used to externally evaluate the

160   ability of T1000 genes to predict apical endpoints (Thomas et al. 2013). Briefly, this dataset

161   contains Affymetrix HT Rat230 PM microarray data following *in vivo* exposure of rats to six

162   chemicals (TRBZ: 1,2,4-tribromobenzene, BRBZ: bromobenzene, TTCP: 2,3,4,6-

163   tetrachlorophenol, MDMB: 4,4'-methylenebis(*N,N'*-dimethyl)aniline, NDPA: N-

164   nitrosodiphenylamine, and HZBZ: hydrazobenzene). In exposed animals, both gene expression

165   and apical outcomes (liver: absolute liver weight, vacuolation, hypertrophy, microvesiculation,

166   necrosis; thyroid: absolute thyroid weight, follicular cell hypertrophy, follicular cell hyperplasia;

167   bladder: absolute bladder weight, increased mitosis, diffuse transitional epithelial hyperplasia,

168   increased necrosis epithelial cell) were measured, permitting the comparison of transcriptionally-

169   derived benchmark doses (BMD_t) with traditional benchmark doses derived from apical

170   outcomes (Yang et al. 2007). The apical outcome-derived benchmark dose (BMD_a) for each

171   treatment group was defined as the benchmark dose from the most sensitive apical outcome for

172   the given chemical-duration group.

173   Raw gene expression data (CEL files) for the dose-response dataset were downloaded from GEO

174   (Accession No. GSE45892), organized into chemical-exposure-duration treatment groups, and

175   normalized using the RMA method (Irizarry et al. 2003a). Only expression measurements

176   corresponding to genes in the T1000 gene (or T384 and T1500) set were retained, resulting in

177   reduced gene expression matrices for each treatment group ($t = 24$). The reduced gene

178   expression matrices were analyzed using BMDExpress 2.0 to calculate a toxicogenomic

179   benchmark dose (BMD_t) for each treatment group (Yang et al. 2007). Here, the BMD_t was

180   calculated as the dose that corresponded to a 10% increase in gene expression compared to the

181   control (Farmahin et al. 2017). Within BMDExpress 2.0, genes were filtered using one-way

182   ANOVA (FDR adjusted p-value cut-off = 0.05). A BMD_t was calculated for each differentially

183   expressed gene by curve fitting with exponential (degree 2-5), polynomial (degree 2-3), linear,

184   power, and Hill models. For each gene, the model with the lowest Akaike information criterion

185   (AIC) was used to derive the BMD_t.

186

187   The BMD_ts from individual genes were used to determine a treatment group-level BMD_t using

188   functional enrichment analysis with Reactome pathways (Farmahin et al. 2017). Note, we chose

189   here to functionally enrich with Reactome since we utilized KEGG to derive the T1000 list.

190   After functional enrichment analysis, significantly enriched pathways (p-value < 0.05) were

191   filtered such that only pathways with > 3 genes and > 5% of genes in the pathway were retained.

192   The treatment group-level BMD_t was calculated by considering the mean gene-level BMD_t for

193   each significantly enriched pathway and selecting the lowest value. If there were no significantly

194   enriched pathways that passed all filters, no BMD_t could be determined for that treatment group.

195   The similarity of the BMD_t to the benchmark dose derived from apical outcomes (BMD_a) was

196   assessed by calculating the BMD_t/BMD_a ratio and the correlation between BMD_t and BMD_a for

197   all treatment groups (Farmahin et al. 2017). Following the same procedures, BMD_t/BMD_a ratio

198   and correlation statistics were determined from genes belonging to L1000, S1500, and Linear

199   Models for Microarray Data (Limma) (Smyth 2005) to provide a reference for the performance

200   of T1000 genes.

201

202   **Databases for Computing Prior Knowledge**

203   The CTD, KEGG, and Hallmark databases were mined to integrate existing toxicogenomics and

204   broader biological knowledge into one network that represents the prior knowledge space. CTD is

205   manually curated from the literature to serve as a public source for toxicogenomics information,

206   currently including over 30.5 million chemical-gene, chemical-disease, and gene-disease

207   interactions (Davis et al. 2017). Following the recommendations of Hu et al. (2015), only

208   "mechanistic/marker" associations were extracted from the CTD database, thus excluding

209   "therapeutic" associations that are presumably less relevant to toxicology. The extracted subgraph

210   contained 2,889 chemicals, 950 diseases annotated as toxic endpoints (e.g. neurotoxicity,

211   cardiotoxicity, hepatotoxicity and nephrotoxicity), and 22,336 genes. KEGG pathways are a

212 popular bioinformatics resource that help to link, organize, and interpret genomic information

213 through the use of manually drawn networks describing the relationships between genes in specific

214 biological processes (Kanehisa et al. 2007). The MSigDB Hallmark gene sets have been developed

215 using a combination of automated approaches and expert curation to represent known biological

216 pathways and processes while limiting redundancy (Liberzon et al. 2015b).

217 Each feature vector consisted of 239 dimensions, representing information encoded from

218 Hallmark, KEGG and CTD. For the Hallmark and KEGG features, we used "1" or "0" to

219 indicate if a gene was present or absent for each of the 50 Hallmark gene sets (Liberzon et al.

220 2015b) and 186 KEGG pathways (Kanehisa & Goto 2000). These features were transformed into

221 z-scores. For the CTD features, we computed the degree, betweenness centrality, and closeness

222 centrality of each gene, based on the topology of the extracted CTD subgraph. The topology

223 measures were log-scaled for each gene in the network. The resulting prior knowledge space

224 consisted of a 239-dimension vector for each of the 22,336 genes, with each vector containing 50

225 z-score normalized Hallmark features, 186 z-score normalized KEGG features, and three log-

226 scaled CTD network features.

227

228 Reactome database

229 To understand the biological space covered by T1000, we analyzed T1000's top enriched

230 Reactome pathways (as KEGG was used to develop T1000). Reactome is a manually curated

231 knowledgebase of human reactions and pathways with annotations of 7,088 protein-coding genes

232 (Croft et al. 2014).

Performance evaluation

234 For the performance evaluation and testing phase, we leveraged the fourth dataset from Open

235 TG-GATEs (see dataset 4 in **Table 1**), which was not used for gene ranking or selection so that it

236 could serve as an external validation dataset. The dose-response dataset was used for an

237 additional external validation (see dataset 5 in **Table 1**).

238
239 In this step, we applied five supervised machine learning methods to the TG-GATES rat kidney

240 *in vivo* dataset, with the objective to predict which exposures caused significant "dysregulation",

241 according to the criteria defined in step 4. This dataset was purposefully not used earlier when

242 deriving T1000 so that it could serve later as a validation and testing dataset. The five machine

243 learning models used were K-nearest neighbors (KNN; K = 3) (Cover & Hart 1967), Decision

244 Trees (DT), Naïve Bayes Classifier (NBC), Quadratic Discriminant Analysis (QDA) and

245 Random Forests (RF).

246

247 The performance of each method was evaluated with five-fold cross-validation and measured

248 using six different metrics (Equations 2 – 7). TP represents the number of true positives, FP the

249 number of false positives, TN the number of true negatives and FN the number of false

250 negatives. The F_1 score (also called the balanced F-score) is a performance evaluation measure

251 that computes the weighted average of sensitivity and precision (He & Garcia 2009), and is well-

252 suited for binary classification models. The $F_{0.5}$ score (Davis & Goadrich 2006; Maitin-Shepard

253 et al. 2010; Santoni et al. 2010) is another summary metric that gives twice as much weight to

254 precision than sensitivity. The evaluation was performed on a Linux based workstation with 16

255 cores and 64 GB RAM for processing the data and running the experiments.

256

257    $sensitivity = TP/(TP + FN)$ (1)

258    $specificity = TN/(TN + FP)$ (2)

259    $precision = TP/(TP + FP)$ (3)

260    $GMean = \sqrt{sensitivity \times specificity}$ (4)

261    $F_1Score = 2 \times \frac{precision \times sensitivity}{precision + sensitivity}$ (5)

262    $F_{0.5}Score = 1.25 \times \frac{precision \times sensitivity}{0.25 \times precision + sensitivity}$ (6)

263

264   Proposed T1000 Framework

265   The work of T1000 was conducted in four discrete phases as follows (see **Figure 1**): I) data

266   preparation and gene co-expression network generation; II) network clustering to group relevant

267   genes; III) gene selection and prioritization; and IV) external testing and performance evaluation.

268   The goal of phase I was to construct two network representations of the interactions between

269   toxicologically-relevant genes, with one based on TG-GATES microarray data (step 1&2) and

270   the other based on the KEGG, MSigDB, and CTD databases (step 3). In a co-expression

271   network, nodes represent genes and edges represent the Pearson's correlation of expression

272   values of pairs of genes. In the current study, we constructed three separate co-expression

273   networks using gene expression profiles from Open TG-GATEs datasets (human *in vitro*, rat *in*

274   *vitro*, and rat *in vivo*) (**Table 1**). If an interaction with a correlation coefficient of 60% or higher

275   was present in all three networks, that gene-gene interaction was then accepted and mapped into

276   one integrated co-expression network by averaging the absolute values of the pairwise

277   correlation coefficients between individual genes. Matching between rat and human genes was

278   based on gene symbols (e.g., Ddr1 in rat is matched with DDR1 in human using BiomaRt R

279   package (Durinck et al. 2009)) and ignored when no match exists. This is a more conservative

280   approach to maintain perfect matching orthologues in the networks although other computational

281   approaches to match orthologues can be used (Wang et al. 2015). The final integrated co-

282   expression network had 11,210 genes from a total of 20,502 genes.

283   To build the prior knowledge space (step 3), we encoded information from the Hallmark, KEGG

284   and CTD databases into feature vectors composed of 239 features describing each gene (see

285   Materials section). Then, we projected the data onto a two-dimensional space using principle

286   component analysis (PCA) and clustered using K-means (K=3) to detect those genes that

287   contributed most to the prior knowledge space. Regarding K-means, we initially experimented

288   with K=1, K=3 and K=5 and after visual inspection of summarized information as

289   **Supplemental Information S2 Figure 1**, we chose K=3.

290   Genes that were furthest from the centroids (i.e., highest contributing ones) of the K-means

291   clusters were more enriched with pathways and gene-chemical-disease interactions (see

292   **Supplemental Information S2**). Based on step 3, a ranked list of all genes was generated such

293   that the first ranked gene would have a prior score of 100% and the last, a prior score close to

294   0%. In phase II, we re-weighted the interactions in the co-expression network based on the prior

295   knowledge space and then detected clusters of highly connected genes in the updated network

296   (step 4). In a Bayesian fashion, the pairwise connections between genes in the co-expression

297   network were re-weighted by multiplying the correlation with the mean prior score. For example,

298   given $P(A)$ and $P(B)$ as prior scores of genes A and B, the correlation score $S(A, B)$ is re-

299   weighted as follows (Eq. 7):

300

301    $S(A, B)_{new} = S(A, B) * ((P(A) + P(B))/2)$ (7)

302

303   It should be noted that in Eq. 7, the product of joint distribution could have been considered for

304   the update such that $S(A, B)_{new} = S(A, B) * ((P(A) * P(B))$.

305   After re-weighting the connections, we detected clusters of highly connected genes using the

306   Markov Cluster Algorithm (MCL) (Van Dongen & Abreu-Goodger 2012). The MCL approach

307   groups together nodes with strong edge weights and then simulates a random flow through a

308   network to find more related groups of genes based on the flow's intensity of movement. It does

309   not require the number of clusters to be pre-specified. An inflation parameter controls the

310   granularity of the output clustering and several values within a recommended range (1.2-5.0)

311   were tried (Van Dongen & Abreu-Goodger 2012). To optimize for the granularity of the

312   clustering, a systematic analysis for the MCL inflation parameter was performed with values in

313   range (1.2-5.0) (see **Supplemental Information S3**). After examining closely efficiency and

314   mass fraction, a value of 3.3 was chosen. This generated 258 clusters that consisted of 11,210

315   genes. The average number of genes in each cluster was 43.4 with the min-max ranging from 1

316   to 8,423.

317   The goal of phase III of gene selection and prioritization was to select the top genes from each

318   cluster to form T1000 (step 5), and then produce a final ranking of the 1000 selected genes (step

319   6). For each of the 258 gene clusters, random forest (RF) classifiers were used to rank genes

320   based on their ability to separate changes in gene expression labelled as "dysregulated" from

321   those labelled "non-dysregulated", using the Gini impurity index of classification (Nguyen et al.

322   2013; Qi 2012; Tolosi & Lengauer 2011). RF is one of the most widely used solutions for feature

323   ranking, and as an ensemble model, it is known for its stability (Chan & Paelinckx 2008). In

324   order to cover more biological space and ensure selected genes represent the whole

325   transcriptome, a different RF classifier is built for each cluster and used to select representative

326   genes (Sahu & Mishra 2012).

327   We selected the top genes from each cluster based on the performance of the RF classifier. For

328   example, when selecting the 1,000 top genes from two clusters (A and B), if the cross-validation

329   prediction accuracy estimated for models A and B were 60% and 55%, respectively, then 522

330   ((60%/(60%+55%))*1000) and 478 ((55%/(60%+55%))*1000) genes would be selected from

331   clusters A and B. However, if cluster A contained only 520 genes, the remaining two genes

332   would be taken from group B, if possible. So, the cluster size is only used if it contains

333   insufficient genes. We repeated this process until 1000 genes were selected. After choosing top k

334   genes from each cluster, we aggregated them into a single list of 1000 genes and built a final RF

335   model to get a global ranking of the genes. We refer to this final ranked list as T1000 (see

336   **Supplemental Table S1** for a full list of selected genes and summary annotation; see

337   **Supplemental Information S4** for the cluster assignment of the genes). The goal of phase IV

338   was to test the performance of the T1000 gene set using external datasets, and thus transition

339   from gene selection activities to ones that focus on the evaluation of T1000. Phase IV is

340   discussed in the following Results section. To discuss factors that characterizes and distinguishes

341   T1000 from L1000 and S1500, Table 2 is provided. As summarized in Table 2, T1000 is more

342   toxicogenomic tailored by selecting genes that optimizes for endpoint predictions and using

343   toxicogenomic datasets. Incorporating the prior knowledge space is critical for T1000 in ranking

344   genes with more contribution to toxic effects. L1000 aims at finding a set of genes that can be

345   used to extrapolate for the full expression space of all other genes. S1500 has considered an

346   optimization for the number of covered pathways. T1000, L1000 and S1500 have considered

347   using PCA and clustering during the selection process. In T1000, however, this step is part of

348   computing the prior only.

349

350   **Results**

351   **Overview of T1000 and biological relevance**

352   The genes comprising T1000 cover a wide biological space of toxicological relevance. For

353   illustration, co-expression networks, before and after applying Steps 2 and 3 (i.e., networks built

354   on the Open TG-GATEs data that are subsequently updated with prior information from KEGG,

355   MSigDb, and CTD), are shown in **Figure 2**. In part (a) of **Figure 2**, a sample co-expression

356   network composed of 150 genes (i.e., 150 for visualization purposes only; of the 11,210 genes

357   identified) has, in general, similar color and size of all the nodes of the network. While this

358   covers a broad toxicological space, it does not necessarily identify or prioritize the most

359   important genes. After subjecting the data to steps 2 and 3, two clusters of genes with different

360   node sizes and colors were identified (**Figure 2b**). Through this refined network, we then

361   applied a prediction model to each cluster to identify the most representative genes resulting in

362   the final co-expression network of the T1000 genes (**Figure 2c**).

363
364   The complete list of T1000 genes with their gene symbols and descriptions, as well as their

365   regulation state (up- or down-regulated) is provided in **Supplemental Table S1**.

366

367   Visual examination of the Reactome enrichment map (**Supplemental Figure S2**) reveals that

368   'biological oxidations' (largest circle in **Supplemental Figure S2**) contained the most enriched

369   pathways followed by 'fatty acid metabolism'. This is logical given that xenobiotic and fatty acid

370 metabolism, mediated by cytochrome P450 (CYP450) enzymes, feature prominently across the

371 toxicological literature (Guengerich 2007) (Hardwick 2008).

372

373 We further examine two genes that are ranked among the top up- and down-regulated gene sets,

374 respectively. We observed that CXCL10 (ranked 2nd in up-regulated genes) and IGFALS (ranked

375 3rd in down-regulated genes) had reported links in the literature in response to exposure to toxic

376 compounds. Upregulation of CXCL10, the ligand of the chemokine receptor CXCR3 found on

377 macrophages, has been observed in the bronchiolar epithelium of patients with Chronic

378 Obstructive Pulmonary Disease (COPD) compared to non-smokers or smokers with normal lung

379 function (Saetta et al. 2002). Smokers develop COPD after exposure to the many chemicals found

380 in cigarette smoke, which include oxidants that cause inflammation (Foronjy & D'Armiento 2006).

381 Although TG-GATEs does not comprise any cigarette toxicants within its database, the general

382 pathways by which toxicants disrupt tissue function are represented by T1000.

383 A gene that was found to be significantly downregulated by T1000 was the gene encoding for

384 Insulin Like Growth Factor Binding Protein Acid Labile Subunit or IGFALS, which is an Insulin

385 growth factor-1 (IGF-1) binding protein (Amuzie & Pestka 2010). Interestingly, the mRNA

386 expression of IGFALS was reported to be significantly downregulated when experimental animals

387 were fed deoxynivalenol, a mycotoxin usually found in grain (Amuzie & Pestka 2010). By

388 reducing IGFALS, the half-life of circulating IGF-1 is reduced, causing growth retardation

389 (Amuzie & Pestka 2010). Many compounds in the TG-GATEs database are of organismal origin,

390 and thus, as the data suggest, they have a similar mode of action as deoxynivalenol in reducing

391 expression of important effectors such as IGFALS.

392

393   Regarding potential clinical applications, we discuss the use of T1000 signature for screening

394   drugs that may show toxic adverse effects in **Supplemental Information S5**. The experiment is

395   motivated by the connectivity map project for connecting small molecules, genes, and disease

396   using gene-expression signatures (Lamb et al. 2006).

397

398   **Benchmark dose-response results**

399
400   Overall, the aim of the evaluation was to assess the ability of T1000 gene sets to predict apical

401   outcomes according to previously published methods (Farmahin et al. 2017). Additionally, we

402   repeated step 4 of the T1000 approach to select the top 384 (T384; i.e., a number conducive to

403   study in a QCPR microplate format as per the EcoToxChip project; (Basu et al. 2019)) and 1,500

404   (T1500 see **Supplemental Information S6**; i.e., a number pursued in other endeavours like

405   S1500) genes to investigate the effect of gene set size on apical outcome prediction. To benchmark

406   the performance of T1000 against other notable gene sets, we considered S1500 (Merrick et al.

407   2015) and L1000 (Subramanian et al. 2017).

408
409
410   BMD_t analysis (see Materials section) of the dose-response dataset was performed with the

411   T1000 gene list and the BMDExpress software program (Yang et al. 2007). The maximum

412   number of BMDs calculated was 21 because for three of the experimental groups a BMD_a

413   (benchmark dose, apical outcome) did not exist due to a lack of observed toxicity (**Table 3**). The

414   T384 gene set performed similarly with Limma; however, increasing the size of this gene set to

415   T1000 resulted in performance evaluation metrics that rivaled that of all other gene sets of the

416   same size or larger (L1000, Limma, and S1500). Further increasing the size of T1000 to T1500

417   did not increase the performance as the correlation slightly decreased while the average ratio of

418   BMD$_t$/BMD$_a$ got slightly closer to one. **Figure 3** provides a visual summary of the comparison

419   based on the BMD$_t$/BMD$_a$ ratios.

420

421   **Prediction results**

422   In a second validation study, we applied T1000 to study the Rat Genome 230 2.0 Array for

423   Kidney dataset from the Open TG-GATEs program. This dataset was not included in any model

424   training or parameter tuning steps. This helped to establish another external validation of T1000

425   in terms of its generalized ability to predict apical outcomes for datasets derived from different

426   tissues. When compared to baseline gene sets mapped using Limma and L1000, T1000 achieved

427   a relative improvement of the F$_1$Score by 6.9% and 27.56%, respectively, thus outperforming

428   comparison gene sets (**Table 4)**. When considering the absolute difference of F$_1$Score between

429   T1000 and the second best (i.e., Limma), T1000 achieved an improvement of 1.59%. The

430   improvement was 1.54% for F$_{0.5}$Score confirming that T1000 led to fewer false positive

431   predictions.

432   Another baseline we compare with is Random-500, where a set of 1000 features are selected

433   randomly and the performance is reported for the five classifiers considered (i.e., LDA, NBC,

434   KNN, QDA and RF). This experiment is repeated for 500 times and the average and standard

435   deviation scores are reported in Table 4. GMean, F1Score and F0.5Score of T1000 are

436   significantly higher (alpha = 0.1) than the random scores. However, an interesting observation is

437   that the Random-500 results outperformed several gene sets. This can be due to the fact that

438   some machine learning models are less sensitive to the type of selected features (e.g., RF). On

439   average, we found that a randomly generated set would outperform other models with a chance

440   of about 30% only.

441 In the context of high throughput screening, such small improvements in F_1Score or $F_{0.5}$Score

442 may represent large cost savings (Soufan et al. 2015a) as false positives may lead to added

443 experiments that would otherwise be unnecessary. Detailed performance scores of each

444 individual machine learning model are provided in **Supplemental Table S2**. Please refer to

445 **Supplemental Information S7** for more comparisons including expression space visualization

446 using PCA and gene set coverage evaluation.

447

448 **Discussion**

449 There is great interest across the toxicological and regulatory communities in harnessing

450 transcriptomics data to guide and inform decision-making (Basu et al. 2019; Council 2007;

451 ECHA 2016; Mav et al. 2018; Thomas et al. 2019). In particular, transcriptomic signatures hold

452 great promise to identify chemical-specific response patterns, prioritize chemicals of concern,

453 and predict quantitatively adverse outcomes of regulatory concern, in a cost-effective manner.

454 However, the inclusion of full transcriptomic studies into standard research studies faces

455 logistical barriers and bioinformatics challenges, and thus, there is interest in the derivation and

456 use of reduced but equally meaningful gene sets.

457

458 Our approach to select T1000 followed the same rationale of how the LINCS program derived

459 the L1000 dataset (Liu et al. 2015), though here we purposefully included additional steps to

460 bolster the toxicological relevance of the resulting gene set. Generating a list of ranked genes

461 based on toxicologically relevant input data and prior knowledge is another key feature of

462 T1000.

463   There are some limitations associated with our current study. For instance, the co-expression

464   network was based on data from the Open TG-GATEs program. While this is arguably the

465   largest toxicogenomics resource available freely, the program is founded on one *in vivo* model

466   (rat), two *in vitro* models (primary rat and human hepatocytes), 170 chemicals that are largely

467   drugs, and microarray platforms. Thus, there remain questions about within- and cross- species

468   and cell type differences, the environmental relevance of the tested chemicals, and the biological

469   space captured by the microarray. The multi-pronged and -tiered bioinformatics approach was

470   designed to yield a toxicologically robust gene set, and the approach can be ported to other

471   efforts that are starting to realize large toxicogenomics databases such as our own EcoToxChip

472   project (Basu et al. 2019). In addition, our approach in selecting T1000 genes was purely data-

473   driven without considering input from scientific experts as was done by the NTP to derive the

474   S1500 gene set (Mav et al. 2018). It is unclear how such gene sets (e.g., T1000, S1500) will be

475   used by the community and under which domains of applicability, and thus there is a need to

476   perform case studies in which new approach methods are compared to traditional methods

477   (Kavlock et al. 2018). It worth to mention that T1000 had 259 and 90 genes in common with

478   S1500 and L1000, respectively and 741 unique genes.

479

Conclusions

481   Here we outlined a systematic, data-driven approach to identify highly-responsive genes from

482   toxicogenomics studies. From this, we prioritized a list of 1,000 genes termed the T1000 gene

483   set. We demonstrated the applicability of T1000 to 7,172 expression profiles, showing great

484   promise in future applications of this gene set to toxicological evaluations. We externally

485   validated T1000 against two *in vivo* datasets of toxicological prominence (a kidney dataset of

486   308 experiments on 41 chemicals from Open TG-GATEs and a dose-response study of 30

487   experiments on six chemicals (Thomas et al. 2013). We compared the performance of T1000

488   against existing gene sets (Limma, L1000 and S1500) as well as panels of randomly selected

489   genes. In doing so, we demonstrate T1000's versatility as it is predictive of apical outcomes

490   across a range of conditions (e.g., *in vitro* and *in vivo*), and generally performs as well, or better

491   than other gene sets available. Our approach represents a promising start to yield a

492   toxicologically-relevant gene set. We hope that future efforts will start to use and apply T1000

493   in a diverse range of settings, and from these we can then start to make updates to the

494   composition of the T1000 gene set based on improved understanding of its performance

495   characteristics and user experiences.

497   **Supplemental data**

498   Supplemental data are available at PeerJ online.

499
500
501   **Acknowledgements**

502   We acknowledge the support of all members of the EcoToxChip project as well as the project's

503   financial sponsors (Genome Canada, Génome Québec, Genome Prairie, the Government of

504   Canada, Environment and Climate Change Canada, Ministère de l'Économie, de la Science et de

505   l'Innovation du Québec, the University of Saskatchewan, and McGill University). We are

506   grateful to the guidance offered by our project's program officer (Micheline Ayoub,

507   Génome Québec) and members of our Research Oversight Committee (Chair: Nancy Denslow;

508   Members: Kevin Crofton, Dan Schlenk, Roy Suddaby, and Carole Yauk).

509
510   **Funding**

511   This study was funded by Genome Canada, Génome Québec, Genome Prairie, the Government

512   of Canada, Environment and Climate Change Canada, Ministère de l'Économie, de la Science et

513   de l'Innovation du Québec, the University of Saskatchewan, and McGill University.

514

515   **References**

516   Alshahrani M, Soufan O, Magana-Mora A, and Bajic VB. 2017. DANNP: an efficient artificial
517   neural network pruning tool. *PeerJ Computer Science* 3:e137.
518   Amuzie CJ, and Pestka JJ. 2010. Suppression of insulin-like growth factor acid-labile subunit
519   expression--a novel mechanism for deoxynivalenol-induced growth retardation. *Toxicol*
520   *Sci* 113:412-421. 10.1093/toxsci/kfp225
521   Andersen ME, and Krewski D. 2009. Toxicity testing in the 21st century: bringing the vision to
522   life. *Toxicol Sci* 107:324-330. 10.1093/toxsci/kfn255
523   Basu N, Crump D, Head J, Hickey G, Hogan N, Maguire S, Xia J, and Hecker M. 2019.
524   EcoToxChip: A next-generation toxicogenomics tool for chemical prioritization and
525   environmental management. *Environ Toxicol Chem* 38:279-288. 10.1002/etc.4309
526   Budinska E, Popovici V, Tejpar S, D'Ario G, Lapique N, Sikora KO, Di Narzo AF, Yan P,
527   Hodgson JG, Weinrich S, Bosman F, Roth A, and Delorenzi M. 2013. Gene expression

528   patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J Pathol*
529   231:63-76. 10.1002/path.4212
530   Chan JC-W, and Paelinckx D. 2008. Evaluation of Random Forest and Adaboost tree-based
531   ensemble classification and spectral band selection for ecotope mapping using airborne
532   hyperspectral imagery. *Remote Sensing of Environment* 112:2999-3011.
533   Council NR. 2007. *Toxicity testing in the 21st century: a vision and a strategy*: National
534   Academies Press.
535   Cover TM, and Hart PE. 1967. Nearest neighbor pattern classification. *Information Theory,*
536   *IEEE Transactions on* 13:21-27.
537   Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, Caudy M, Garapati P, Gillespie M,
538   Kamdar MR, Jassal B, Jupe S, Matthews L, May B, Palatnik S, Rothfels K, Shamovsky
539   V, Song H, Williams M, Birney E, Hermjakob H, Stein L, and D'Eustachio P. 2014. The
540   Reactome pathway knowledgebase. *Nucleic Acids Res* 42:D472-477.
541   10.1093/nar/gkt1102
542   Davis AP, Grondin CJ, Johnson RJ, Sciaky D, King BL, McMorran R, Wiegers J, Wiegers TC,
543   and Mattingly CJ. 2017. The comparative toxicogenomics database: update 2017. *Nucleic*
544   *acids research* 45:D972-D978.
545   Davis J, and Goadrich M. 2006. The relationship between Precision-Recall and ROC curves.
546   Proceedings of the 23rd international conference on Machine learning: ACM. p 233-240.
547   Durinck S, Spellman PT, Birney E, and Huber W. 2009. Mapping identifiers for the integration
548   of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* 4:1184-1191.
549   10.1038/nprot.2009.97
550   ECHA. 2007. Understanding REACH. *Available at*
551   *https://echa.europa.eu/regulations/reach/understanding-reach* (accessed April 11, 2019.
552   ECHA. 2016. New Approach Methodologies in Regulatory Science. In: Agency EC, editor.
553   European Chemicals Agency (ECHA). Helsinki.
554   Farmahin R, Williams A, Kuo B, Chepelev NL, Thomas RS, Barton-Maclaren TS, Curran IH,
555   Nong A, Wade MG, and Yauk CL. 2017. Recommended approaches in the application of
556   toxicogenomics to derive points of departure for chemical risk assessment. *Archives of*
557   *toxicology* 91:2045-2065.
558   Foronjy R, and D'Armiento J. 2006. The Effect of Cigarette Smoke-derived Oxidants on the
559   Inflammatory Response of the Lung. *Clin Appl Immunol Rev* 6:53-72.
560   10.1016/j.cair.2006.04.002
561   Gautier L, Cope L, Bolstad BM, and Irizarry RA. 2004. affy--analysis of Affymetrix GeneChip
562   data at the probe level. *Bioinformatics* 20:307-315. 10.1093/bioinformatics/btg405
563   Guengerich FP. 2007. Mechanisms of cytochrome P450 substrate oxidation: MiniReview. *J*
564   *Biochem Mol Toxicol* 21:163-168.
565   Haider S, Black MB, Parks BB, Foley B, Wetmore BA, Andersen ME, Clewell RA, Mansouri K,
566   and McMullen PD. 2018. A Qualitative Modeling Approach for Whole Genome
567   Prediction Using High-Throughput Toxicogenomics Data and Pathway-Based Validation.
568   *Front Pharmacol* 9:1072. 10.3389/fphar.2018.01072
569   Hardwick JP. 2008. Cytochrome P450 omega hydroxylase (CYP4) function in fatty acid
570   metabolism and metabolic diseases. *Biochem Pharmacol* 75:2263-2275.
571   10.1016/j.bcp.2008.03.004
572   He H, and Garcia EA. 2009. Learning from imbalanced data. *Knowledge and Data Engineering,*
573   *IEEE Transactions on* 21:1263-1284.

574 Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, Urushidani T, and Yamada H. 2014a.
575   Open TG-GATEs - Pathological items. *Available at*
576   *https://dbarchive.biosciencedbc.jp/en/open-tggates/data-12.html*2017).
577 Igarashi Y, Nakatsu N, Yamashita T, Ono A, Ohno Y, Urushidani T, and Yamada H. 2014b.
578   Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic acids research*
579   43:D921-D927.
580 Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, and Speed TP. 2003a. Summaries of
581   Affymetrix GeneChip probe level data. *Nucleic acids research* 31:e15-e15.
582 Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, and Speed TP.
583   2003b. Exploration, normalization, and summaries of high density oligonucleotide array
584   probe level data. *Biostatistics* 4:249-264. 10.1093/biostatistics/4.2.249
585 Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S,
586   Okuda S, and Tokimatsu T. 2007. KEGG for linking genomes to life and the
587   environment. *Nucleic acids research* 36:D480-D484.
588 Kanehisa M, and Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids*
589   *research* 28:27-30.
590 Kavlock RJ, Bahadori T, Barton-Maclaren TS, Gwinn MR, Rasenberg M, and Thomas RS. 2018.
591   Accelerating the Pace of Chemical Risk Assessment. *Chem Res Toxicol* 31:287-290.
592   10.1021/acs.chemrestox.7b00339
593 Knudsen TB, Keller DA, Sander M, Carney EW, Doerrer NG, Eaton DL, Fitzpatrick SC,
594   Hastings KL, Mendrick DL, Tice RR, Watkins PB, and Whelan M. 2015. FutureTox II:
595   in vitro data and in silico models for predictive toxicology. *Toxicol Sci* 143:256-267.
596   10.1093/toxsci/kfu234
597 Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP,
598   Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ,
599   Clemons PA, Wei R, Carr SA, Lander ES, and Golub TR. 2006. The Connectivity Map:
600   using gene-expression signatures to connect small molecules, genes, and disease. *Science*
601   313:1929-1935. 10.1126/science.1132939
602 Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, and Tamayo P. 2015a. The
603   Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* 1:417-
604   425. 10.1016/j.cels.2015.12.004
605 Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, and Tamayo P. 2015b. The
606   molecular signatures database hallmark gene set collection. *Cell systems* 1:417-425.
607 Liu C, Su J, Yang F, Wei K, Ma J, and Zhou X. 2015. Compound signature detection on LINCS
608   L1000 big data. *Mol Biosyst* 11:714-722. 10.1039/c4mb00677a
609 Maitin-Shepard J, Cusumano-Towner M, Lei J, and Abbeel P. 2010. Cloth grasp point detection
610   based on multiple-view geometric cues with application to robotic towel folding.
611   Robotics and Automation (ICRA), 2010 IEEE International Conference on: IEEE. p
612   2308-2315.
613 Mav D, Shah RR, Howard BE, Auerbach SS, Bushel PR, Collins JB, Gerhold DL, Judson RS,
614   Karmaus AL, and Maull EA. 2018. A hybrid gene selection approach to create the
615   S1500+ targeted gene sets for use in high-throughput transcriptomics. *PLoS One*
616   13:e0191105.
617 Merrick BA, Paules RS, and Tice RR. 2015. Intersection of toxicogenomics and high throughput
618   screening in the Tox21 program: an NIEHS perspective. *International journal of*
619   *biotechnology* 14:7-27.

620 Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F,
621   and Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in
622   tetrapods. *Nature* 505:635-640. 10.1038/nature12943
623 Nguyen C, Wang Y, and Nguyen HN. 2013. Random forest classifier combined with feature
624   selection for breast cancer diagnosis and prognostic. *Journal of Biomedical Science and*
625   *Engineering* 6:551.
626 Qi Y. 2012. Random forest for bioinformatics. *Ensemble machine learning*: Springer, 307-323.
627 Saetta M, Mariani M, Panina-Bordignon P, Turato G, Buonsanti C, Baraldo S, Bellettato CM,
628   Papi A, Corbetta L, Zuin R, Sinigaglia F, and Fabbri LM. 2002. Increased expression of
629   the chemokine receptor CXCR3 and its ligand CXCL10 in peripheral airways of smokers
630   with chronic obstructive pulmonary disease. *Am J Respir Crit Care Med* 165:1404-1409.
631 Sahu B, and Mishra D. 2012. A novel feature selection algorithm using particle swarm
632   optimization for cancer microarray data. *Procedia Engineering* 38:27-31.
633 Santoni FA, Hartley O, and Luban J. 2010. Deciphering the code for retroviral integration target
634   site selection. *PLoS computational biology* 6:e1001008.
635 Smyth GK. 2005. Limma: linear models for microarray data. *Bioinformatics and computational*
636   *biology solutions using R and Bioconductor*: Springer, 397-420.
637 Soufan O, Ba-alawi W, Afeef M, Essack M, Rodionov V, Kalnis P, and Bajic VB. 2015a.
638   Mining Chemical Activity Status from High-Throughput Screening Assays. *PLoS One*
639   10:e0144426. 10.1371/journal.pone.0144426
640 Soufan O, Kleftogiannis D, Kalnis P, and Bajic VB. 2015b. DWFS: a wrapper feature selection
641   tool based on a parallel genetic algorithm. *PLoS One* 10:e0117988.
642   10.1371/journal.pone.0117988
643 Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli
644   AA, Asiedu JK, Lahr DL, Hirschman JE, Liu Z, Donahue M, Julian B, Khan M, Wadden
645   D, Smith IC, Lam D, Liberzon A, Toder C, Bagul M, Orzechowski M, Enache OM,
646   Piccioni F, Johnson SA, Lyons NJ, Berger AH, Shamji AF, Brooks AN, Vrcic A, Flynn
647   C, Rosains J, Takeda DY, Hu R, Davison D, Lamb J, Ardlie K, Hogstrom L, Greenside
648   P, Gray NS, Clemons PA, Silver S, Wu X, Zhao WN, Read-Button W, Wu X, Haggarty
649   SJ, Ronco LV, Boehm JS, Schreiber SL, Doench JG, Bittker JA, Root DE, Wong B, and
650   Golub TR. 2017. A Next Generation Connectivity Map: L1000 Platform and the First
651   1,000,000 Profiles. *Cell* 171:1437-1452 e1417. 10.1016/j.cell.2017.10.049
652 Thomas RS, Bahadori T, Buckley TJ, Cowden J, Deisenroth C, Dionisio KL, Frithsen JB, Grulke
653   CM, Gwinn MR, Singh A, Richard AM, Williams AJ, Deisenroth C, Grulke CM,
654   Patlewicz G, Shah I, Cowden J, Wambaugh JF, Harrill JA, Paul-Friedman K, Houck KA,
655   Gwinn MR, Linnenbrink M, Setzer RW, Sams R, Judson RS, Simmons SO, Knudsen TB,
656   Thomas RS, Lambert JC, Bahadori T, Swank A, Wetmore BA, Ulrich EM, Sobus JR,
657   Phillips KA, Dionisio KL, Isaacs KK, Strynar M, Tornero-Valez R, Newton SR, Buckley
658   TJ, Frithsen JB, Villeneuve DL, Hunter ES, III, Simmons JE, Higuchi M, Hughes MF,
659   Padilla S, Shafer TJ, and Martin TM. 2019. The next generation blueprint of
660   computational toxicology at the U.S. Environmental Protection Agency.
661   10.1093/toxsci/kfz058
662 Thomas RS, Wesselkamper SC, Wang NCY, Zhao QJ, Petersen DD, Lambert JC, Cote I, Yang
663   L, Healy E, and Black MB. 2013. Temporal concordance between apical and
664   transcriptional points of departure for chemical risk assessment. *toxicological sciences*
665   134:180-194.

666   Tolosi L, and Lengauer T. 2011. Classification with correlated features: unreliability of feature
667   ranking and solutions. *Bioinformatics* 27:1986-1994. 10.1093/bioinformatics/btr300
668   van Dam S, Vosa U, van der Graaf A, Franke L, and de Magalhaes JP. 2018. Gene co-expression
669   analysis for functional classification and gene-disease predictions. *Brief Bioinform*
670   19:575-592. 10.1093/bib/bbw139
671   Van Dongen S, and Abreu-Goodger C. 2012. Using MCL to extract clusters from networks.
672   *Bacterial Molecular Networks*: Springer, 281-295.
673   Villeneuve DL, and Garcia-Reyero N. 2011. Vision & strategy: Predictive ecotoxicology in the
674   21st century. *Environ Toxicol Chem* 30:1-8. 10.1002/etc.396
675   Wang Y, Coleman-Derr D, Chen G, and Gu YQ. 2015. OrthoVenn: a web server for genome
676   wide comparison and annotation of orthologous clusters across multiple species. *Nucleic*
677   *Acids Res* 43:W78-84. 10.1093/nar/gkv487
678   Yang L, Allen BC, and Thomas RS. 2007. BMDExpress: a software tool for the benchmark dose
679   analyses of genomic data. *BMC genomics* 8:387.
680

Figure 1

Framework of the T1000 approach for gene selection and prioritization.

Phase I is composed of Steps [1-3]. After data is prepared in Step 1, the co-expression network is generated through Step 2. The prior knowledge scores are computed using (KEGG, MSigDB) and toxicological (CTD) relevance graphs in Step 3. Phase II involves Step 4 for re-weighting of the co-expression scores based on prior knowledge of biological and toxicological relevance graphs. In addition, the graph is clustered during Step 4. In Phase III, in Step 5, a prediction model is trained for each cluster. Then, after selecting top genes from each cluster in Step 5, a one final prediction model called global is trained to rank all selected genes (Step 6). Phase IV is a focused on external evaluation of the prioritized gene list.

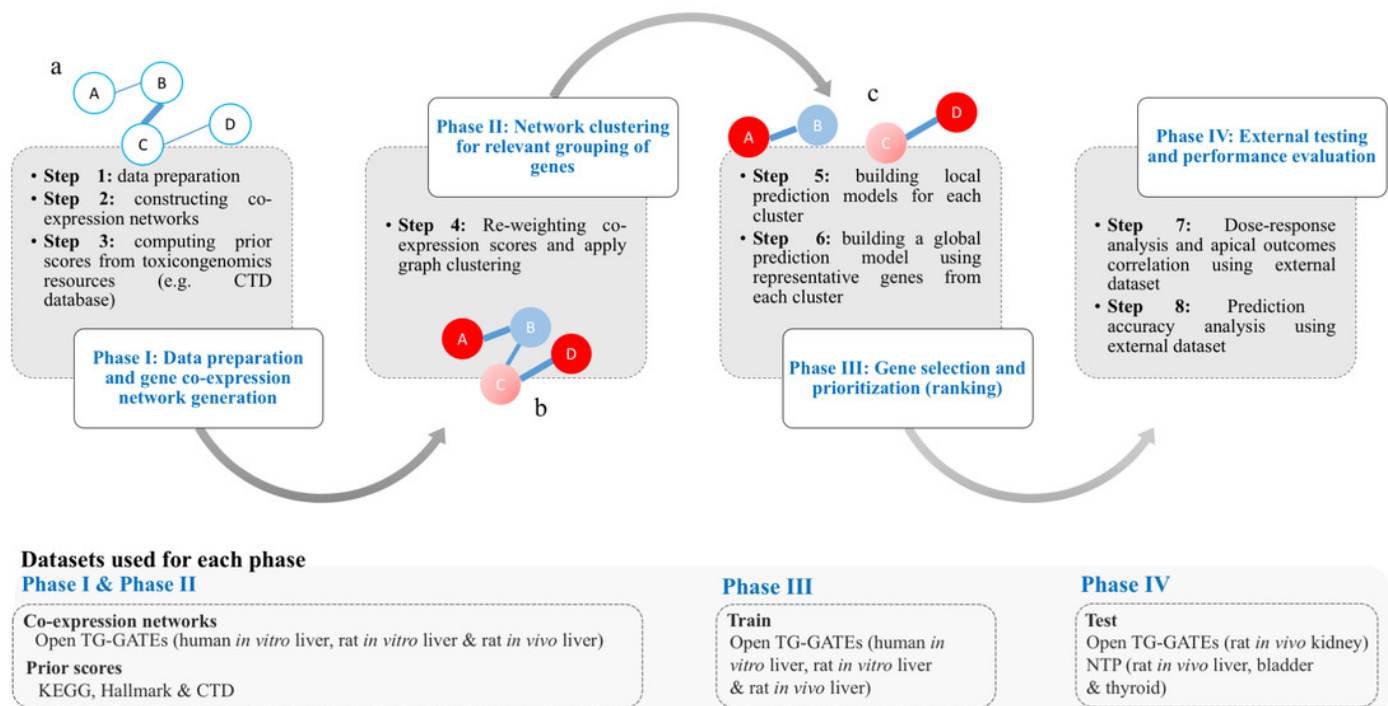
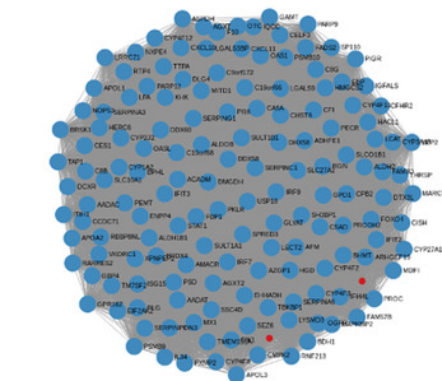
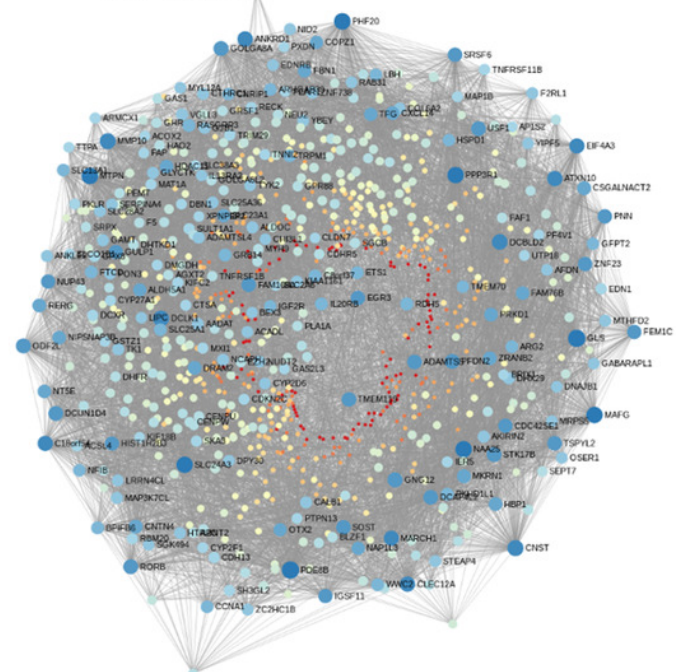



Figure 2

Visual representation of co-expression networks before and after performing Steps 2 and 3 of the T1000 selection process.

Visual representation of co-expression networks before and after performing Steps 2 and 3 of the T1000 selection process. A sample co-expression network of a group of 150 genes such that each pair of genes would have a connection is provided in Part (a). After re-weighting the correlation scores using the prior knowledge of biological and toxicological relevance graphs and performing clustering through Steps [1-4] of T1000 framework (see Figure 1), the graph in Part (a) is evolved to the one in Part (b). In Part (b), a pair of genes would have a link only if they hold enough confidence after applying prior scores. From part (b), nodes representing genes gain different levels of colors summarizing different levels of structural representations in the graph. Therefore, it is more relevant to cluster the graph at this stage after applying prior weights instead of the stage of Part (a). We can visually detect two separate clusters of genes in Part (b). After executing T1000 framework, we visualize the generated co-expression graph of all selected 1000 genes in Part (c). Compared to Part (a), we see variant levels of colors indicating different structural relevance. The colors in Parts (a), (b), and (c) reflect structural statistics using betweenness centrality and node degree. Part (a) holds a very similar statistics while Parts (b), (c) exploits and shows variant levels. A more contributing gene would have a larger node and a darker blue color while a less important one would have a very small node with a red color intensity. Please note that Parts (b) and (c) are realized only after executing steps from T1000 framework while Part (a) shows the generic representation of the co-expression graph.

a) Co-expression network of a group of genes before clustering

c) T1000 generated co-expression network after clustering

b) Co-expression network after applying prior weights and clustering

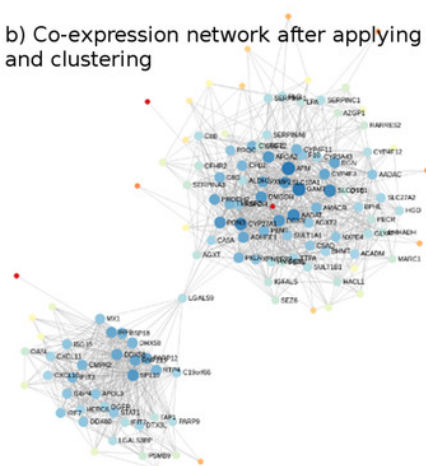
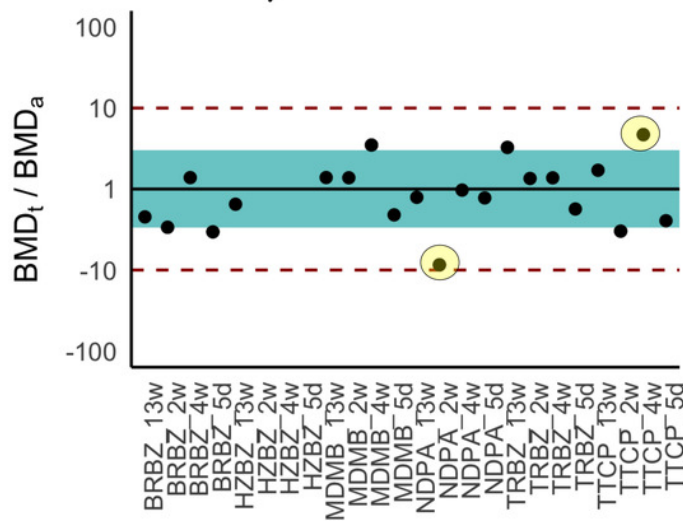
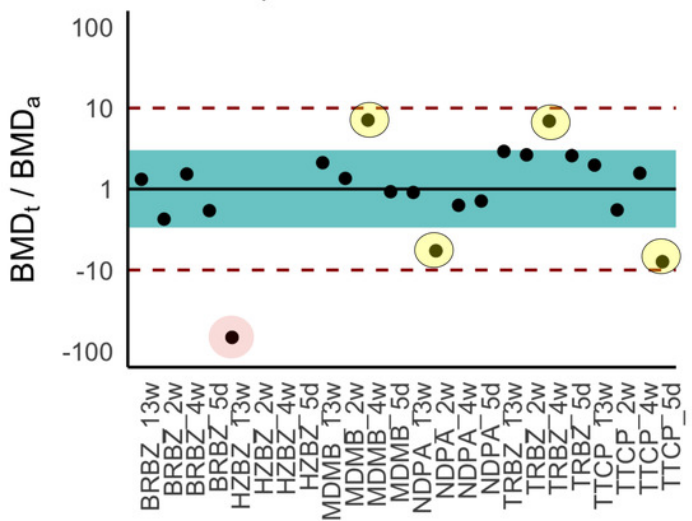
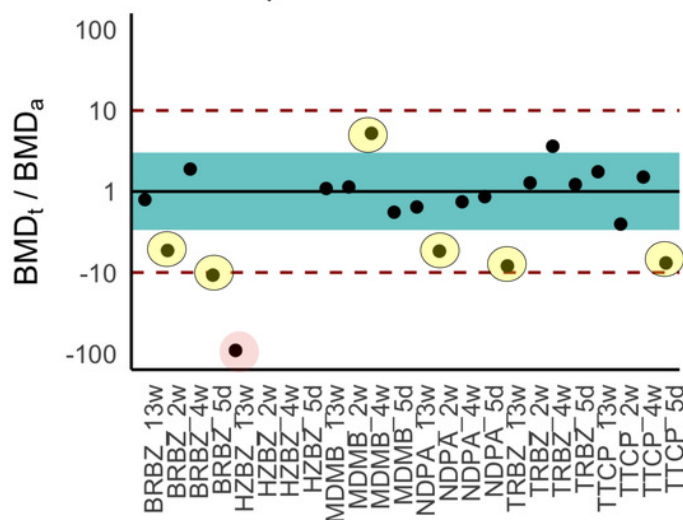
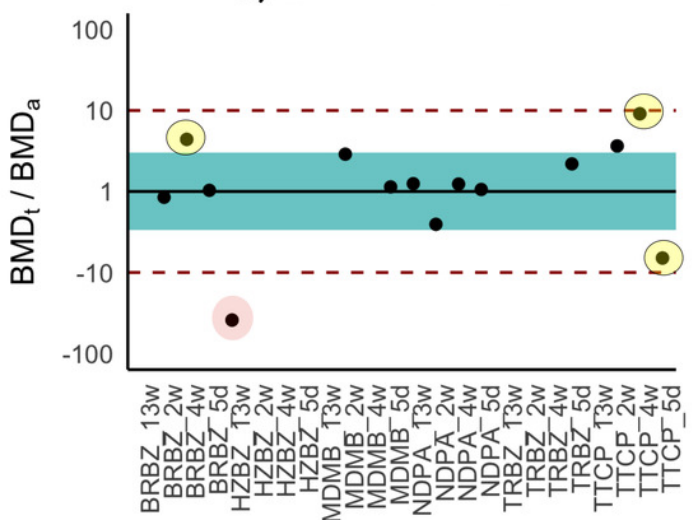



Figure 3

Ratios of BMDt/BMDa for each experimental group determined with various gene sets as indicated atop the plots.

Ratios of BMDt/BMDa represents ratio of transcriptionally-derived benchmark doses BMDt using gene signatures to apical outcome-derived benchmark dose BMDa serving as the ground truth. The limits of the blue rectangular band and dotted lines represent 3-fold and 10-fold of unity, respectively. Ratios could not be calculated for three experimental groups (hydrazobenzene (HZBZ): 5 day, 2 week, 4 week) due to a lack of apical outcomes. Red circles represent mean ratios greater than 10-fold, while the yellow ones represent ratios greater than 3-fold. The fewer circles, the more the gene set is indicative of potential relevance to the examined apical endpoints (see Supplementary Figure 3 and 4 for T384 and T1500 plots, respectively). In Part a), the T1000 results are highlighted such that in only two experiments, the ratio of difference from the ground truth was greater than 3 folds and less than 10. In Parts b,c&d, the results of L1000, S1500 and Limma are illustrated, respectively, with each having a single experiment (i.e., red circle) with 10-fold difference from the ground truth. All of them had more yellow circles as compared to Part a of T1000.

a) T1000 Ratios

b) L1000 Ratios

c) S1500 Ratios

d) Limma Ratios

Summary of datasets used in the current study.

Datasets 1-3 were used to develop T1000 (see Phase I, II & III in Methods Section) and datasets 4 and 5 (see Phase IV in Methods Section) were used to evaluate the performance of the gene sets.

Dataset #	Dataset	Organism	Organ	Exposure Type	Number of chemicals	Matrix size (% missing values)	Purpose in Current Study
1	Open TG-GATEs	Human	Liver	*in vitro*	158 chemicals	2,606 experiments x 20,502 genes (8.9%)	Training
2	Open TG-GATEs	Rat	Liver	*in vitro*	145 chemicals	3,371 experiments x 14,468 genes (11.6%)	Training
3	Open TG-GATEs	Rat	Liver	*in vivo* (single dose)	158 chemicals	857 experiments x 14,400 genes (11.5%)	Training
4	Open TG-GATEs	Rat	Kidney	*in vivo* (single dose)	41 chemicals	308 experiments x 14,400 genes (12.2%)	Testing
5	Dose-response (GSE45892)	Rat	Liver, Bladder, Thyroid	*in vivo* (repeated dose)	6 chemicals	30 experiments x 14,400 genes (0%)	Testing (external validation)
Total						7,172 experiments	

Descriptive comparison of T1000 against existing gene sets.

For the 'selection criteria' column, expression space coverage refers to the goal of finding a subset of genes that would achieve high correlation with the original full set of genes.Pathway coverage refers to finding a subset of genes that cover more pathways in a reference library.

Gene set	Selection criteria	Ranked gene list	Species	Data	Approach	Number of genes
L1000	Expression space coverage	No	Human	L1000 data	PCA and clustering (Data mining)	978
S1500 (NTP 2018)	Pathway coverage that combines data-driven and knowledge-driven activities	No	Human	Public GEO expression datasets (mainly GEO 3339 gene expression series)	PCA, clustering, and other data-driven steps (Data mining)	2861 (includes L1000 genes)
T1000	Toxicological relevance using endpoint prediction	Yes	Human and Rat	Open TG-GATEs that is founded on co-expression networks from CTD, KEGG and Hallmark	Co-expression network and prior knowledge (Graph mining). PCA and clustering are used only for the prior knowledge.	1000

Summary of correlation of apical endpoints to 24 experimental groups (6 chemicals x 4 exposure durations).

	T384 (n = 384)	T1000 (n = 1000)	T1500 (n = 1500)	L1000 (n = 976)	S1500 (n = 2861)	Limma (n = 1000)
# of BMD$_t$s	18	21	21	21	21	14
Mean ratio (BMD$_t$/BMD$_a$)	2.2	1.2	1.1	1.8	1.1	2.1
Correlation (BMD$_t$, BMD$_a$)	0.83 ($p <$ 0.001)	**0.89** ($p <$ 0.001)	0.83 ($p <$ 0.001)	0.76 ($p <$ 0.001)	0.78 ($p <$ 0.001)	0.73 ($p < 0.01$)

Summary comparison of average classification performance using the testing RatKidney dataset. Scores are based on average results from five classifiers (LDA, NBC, QDA, DT and RF) and the standard deviation is reported to highlight variance of estimate.

*Statistically significant at an alpha level of 0.1 using T-test and considering comparison with Random results.

	Sensitivity	Specificity	Precision	Gmean	F1Measure	F0.5Measure
T1000	**29.25% (±11.64)***	71.33% (±4.74)	**21.51% (±4.45)**	**44.7% (±7.8)***	**24.58% (±7.11)***	**22.6% (±5.36)**
Limma	27.76% (±16.3)	70.75% (±6.33)	20% (±9.96)	41.84% (±14.81)	22.99% (±12.04)	21.06% (±10.64)
CD	21.79% (±15.39)	68.08% (±10.97)	13.94% (±6.64)	34.79% (±13.3)	16.65% (±9.96)	14.83% (±7.82)
L1000	22.99% (±12.82)	70.42% (±5.78)	16.84% (±7.29)	38.33% (±11.46)	19.27% (±9.27)	17.71% (±7.97)
S1500	21.79% (±7.65)	**72.67% (±3.98)***	17.87% (±3.99)	39.19% (±6.2)	19.53% (±5.42)	18.48% (±4.48)
Random-500	27.83% (±11.69)	70.89% (±5.09)	20.31% (±4.89)	42.81% (±8.38)	18.41% (±12.03)	21.29% (±5.79)
P-value (T1000 vs. Random)	0.0555	0.3454	0.1283	0.0504	0.0192	0.1112
Best Model (Limma_NBC)	44.78%	68.75%	28.57%	55.48%	34.88%	30.80%
Worst Model (Limma_QDA)	4.48%	72.08%	4.29%	17.97%	4.38%	4.32%