More than just hormones: H295R cells as predictors of reproductive toxicity

Jodi M. Maglich a, Max Kuhn b, Robert E. Chapin c, Mathew T. Pletcher d,∗

a Compound Safety Prediction, Pfizer Global Research and Development, Pfizer Inc., Cambridge, MA 02420, United States
b Non-Clinical Statistics, Pfizer Global Research Development, Pfizer Inc., Groton, CT 06340, United States
c Developmental and Reproductive Toxicology Center of Expertise, Pfizer Global Research and Development, Pfizer Inc., Groton, CT 06340, United States
d Rare Diseases Research Unit, Pfizer Global Research Development, Pfizer Inc., Cambridge, MA 02420, United States

ARTICLE INFO

Article history:
Received 23 August 2013
Received in revised form 20 November 2013
Accepted 18 December 2013
Available online 13 January 2014

Keywords:
Steroidogenesis
Reproductive toxicity
H295R
Predictive screening

ABSTRACT

Many of the commonly observed reproductive toxicities associated with therapeutic compounds can be traced to a disruption of the steroidogenic pathway. We sought to develop an in vitro assay that would predict reproductive toxicity and be high throughput in nature. H295R cells, previously validated as having an intact and functional steroidogenic pathway, were treated with 83 known-positive and 79 known-negative proprietary and public-domain compounds. The assay measured the expression of the key enzymes STAR, 3βHSD2, CYP17A1, CYP1B2, CYP19A1, CYP21A2, and CYP11A1 and the hormones DHEA, progesterone, testosterone, and cortisol. We found that a Random Forest model yielded a receiver operating characteristic area under the curve (ROC AUC) of 0.845, with sensitivity of 0.724 and specificity of 0.758 for predicting in vivo reproductive toxicity with this in vitro assay system.

© 2014 The Authors. Published by Elsevier Inc. Open access under CC BY-NC-SA license.

1. Introduction

Toxicology is in the middle of a profound change, from a descriptive science to a predictive science, mediated by the visionary Toxicity Testing in the 21st Century publication [1]. This approach has certainly found fertile ground in pharmaceuticals [2]. Some form of safety evaluation often occurs very early in the development of a compound series. Selection of key compounds for further development uses data from multiple assays, each for an individual pathway of toxicity or key biological process [3,4]. The body depends on steroid hormones to regulate or influence the immune system, response to stress, gluconeogenesis, components of behavior, and the many complex parts of male and female reproduction. Based on the importance of this pathway in many physiological processes, it would be beneficial to be able to choose between two candidate drug molecules, one of which showed an unwanted impact on steroidogenesis, and the other of which did not. To accommodate the needs of a screening program early in the candidate selection process, such an assay should require minimal amount of test compound (de novo synthesized drug candidate molecules), be at least relatively high-throughput, and not require extended exposures.

In steroidogenesis, cholesterol is first shuttled to the inner mitochondrial membrane in a rate-limiting step by steroidalogenic acute regulatory protein (STAR) [5] (Fig. 1). Cholesterol is then converted to pregnenolone by side chain cleavage enzyme (CYP11A1). Progesterone is produced by 3 beta-hydroxysteroid dehydrogenase (3βHSD2) action on pregnenolone. CYP17A1 catalyzes the 17α-hydroxylatation of pregnenolone and progesterone to 17α-hydroxy intermediates and the 17,20-lyase reactions leading to DHEA and along with 17β-hydroxysteroid dehydrogenase (17βHSD) activity, to testosterone. Cortisol is synthesized from the 17α-hydroxy intermediates by the enzymes 3βHSD, 21-hydroxylase (CYP21A2), and 11β-hydroxylase (CYP11B1). Estradiol is converted from testosterone by the enzyme aromatase (CYP19A1). Estradiol can alternatively be converted by 17βHSD from estrone, a hormone produced by aromatase activity on androstenedione.

H295R cells are a transformed human adrenal cell line which secretes all the steroid intermediates of the steroidogenesis pathway, and has been found useful for studying steroidogenesis [6–9]. These cells are zonally undifferentiated, i.e., they produce the steroids of each of the three zones normally segregated in the adult adrenal cortex [10–12]. Because H295R cells uniquely express all of the enzymes in the steroidogenesis pathway, they allow the...
simultaneous testing of all components of steroidogenesis to the test exposure in vitro. These cells have been widely used in other testing efforts: the US EPA endocrine disruptor screening program (EDSP), ToxCast, European REACH, as well as the global Organization for Economic Cooperation and Development (OECD) [13–15] have all employed hormone measurement in the H295R cell model for identification of endocrine disrupting chemicals.

The objective of our work reported here was to develop a multi-parameter assay to detect chemical disruption of the steroidogenesis pathway. The assay would measure quantitative differences in the expression of enzymes in the steroidogenic pathway and levels of hormones secreted into the media following treatment with known in vivo toxicants and non-toxic compounds. This approach then allowed us to explore whether there was any relationship between changes in these H295R endpoints and male or female reproductive system pathology in rats in preclinical safety studies. Thus, we posed the question “how well does one or a combination of steroidogenic endpoints in vitro predict any male or female reproductive toxicity in vivo?”

2. Materials and methods

2.1. Compound selection

Compounds that had previously been evaluated in vivo for their adverse effects on reproductive tissues were selected from Pfizer’s internal library and from published in vivo studies (Table 1). We took an unbiased approach and chose compounds that displayed reproductive toxicity in either male or females, and incorporated all reproductive findings, not just those considered to be hormonally driven. Not requiring a known mechanism of action for the in vivo reproductive toxicity was necessary to ensure sufficient power for the planned model building. It also allowed for us to determine if an assay focusing on steroidogenic regulation could play a role in predicting more general reproductive toxicity especially since most toxicity testing ends at that level of pathological resolution. The vast majority of these findings were of structural abnormalities (i.e., pathology or lesions) noted in reproductive organs after dosing for periods ranging from 2 weeks to 6 months. We compiled 83 compounds with adverse reproductive findings and 79 compounds with no in vivo adverse finding. Forskolin (Sigma) and prochloraz (Sigma) served as positive controls for the assay. All compounds were dissolvered in DMSO and DMSO (1%) was used as vehicle control.

2.2. Cell culture and toxicity assay

H295R (ATCC) cells were cultured at 37°C in 5% CO2 in DMEM:F12 (Invitrogen) medium supplemented with 2.5% Nu-Serum (BD Biosciences), 1% ITS+ Premix (BD Biosciences), L-glutamine (Invitrogen), and penicillin-streptomycin (Invitrogen). Exposures were conducted in 2 phases: dose-range finding, and the response phase. For dose-range finding, cells were plated at 10 × 10^4 per well in 96 well plates for 48 h before treatment with compounds dissolved in DMSO for 30h in 8 point curves (2-fold dilutions, starting at 300 μM). Compound effect on cell viability was determined by using the Cell Titer Glo (Promega) assay kit on compound-treated cells and results were graphed using the IDBS XLFit version 4.2.1 plug-in for Microsoft Excel to determine a TC10 (toxic concentration at which 10% cell death is observed) for the concentration used in the response experiment. If no cytotoxicity was observed, the cells were treated at 300 μM. The response phase exposure was run with cells plated at 10 × 10^4 cells per well in 96-well plates for 48 h before compound exposure at the TC10 for an additional 30h.

We wanted to determine if the readout from our assay was correlated with cytotoxicity. We employed an assay which is commonly used in our lab as a general readout of cytotoxicity: the transformed human liver epithelium cell line [16,17]. THLE-2 cells (ATCC; CRL2706) were cultured at 37°C in 5% CO2 in BEBM (Lonza) supplemented with 10% HI-FBS, 5 ng/ml hEGF, 70 ng/ml phospho-ethanolamine, and the supplied BEBM bullet kit (Lonza). Cells were plated at 2.5 × 10^3 cells/well in 384-well plates for 24 h before compound treatment. Compounds were added dissolved in DMSO in 10-point curves with 300 μM as the highest concentration. After a 72-h incubation, Cell Titer Glo reagent (Promega) was added. Luminescence values were read on a luminometer and graphed using the IDBS XLFit version 4.2.1 plug-in for Microsoft Excel to determine an IC50 values for cytotoxicity.

2.3. RNA isolation and quantitative RT-PCR analysis

At time of harvest, the cell supernatants were removed and frozen, and 100 μl lysis buffer was added to each well in the 96-well plate. The RNA was then extracted using the SV96 RNA kit (Promega) according to manufacturer’s instructions. cDNA was generated using HiCapacity RT kit (ABI #4368813) according to
<table>
<thead>
<tr>
<th>Compound</th>
<th>On market name</th>
<th>In vivo observed toxicity</th>
<th>Compound</th>
<th>On market name</th>
<th>In vivo observed toxicity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compound002</td>
<td></td>
<td>Mucification of vagina, subnuclear vacuoles in uterus</td>
<td>Compound085</td>
<td></td>
<td>Inflammation in epididymides</td>
</tr>
<tr>
<td>Compound004</td>
<td>Clomiphene</td>
<td>Seminal vesicle and prostate gland weights decreased, reduced testosterone [38]</td>
<td>Compound092</td>
<td></td>
<td>Inflammation in epididymides</td>
</tr>
<tr>
<td>Compound007</td>
<td>Cefoperazone</td>
<td>Testicular degeneration [39]</td>
<td>Compound093</td>
<td></td>
<td>Degeneration in testis, abnormal content in epididymides</td>
</tr>
<tr>
<td>Compound009</td>
<td>Vincristine</td>
<td>Malformation of late spermatids and arrest of cell division of spermatocytes and spermatagonia</td>
<td>Compound094</td>
<td></td>
<td>Dilatation tubules in testis</td>
</tr>
<tr>
<td>Compound010</td>
<td></td>
<td>Hypertrophy of interstitial cells in the testis</td>
<td>Compound096</td>
<td></td>
<td>Estrous cycle disruption, mammary hyperplasia, decreased seminal vesicles</td>
</tr>
<tr>
<td>Compound012</td>
<td></td>
<td>Testicular degeneration</td>
<td>Compound098</td>
<td></td>
<td>Inflammation in Epididymides</td>
</tr>
<tr>
<td>Compound013</td>
<td></td>
<td>Testicular degeneration, decreased ovary weight</td>
<td>Compound099</td>
<td></td>
<td>Low testosterone</td>
</tr>
<tr>
<td>Compound017</td>
<td>Colchicine</td>
<td>Degraded microtubules in the testis, caused abnormalities of the head and acrosome of testicular spermatids [40]</td>
<td>Compound100</td>
<td></td>
<td>Inflammation in Epididymides</td>
</tr>
<tr>
<td>Compound018</td>
<td></td>
<td>Inhibits sexual differentiation in gonads</td>
<td>Compound101</td>
<td></td>
<td>Testicular degeneration (dog), cervical polyps</td>
</tr>
<tr>
<td>Compound019</td>
<td></td>
<td>Testicular atrophy, disturbance in rat estrus cyte</td>
<td>Compound102</td>
<td></td>
<td>Degenerative spermatid cell debris,</td>
</tr>
<tr>
<td>Compound022</td>
<td></td>
<td>Degeneration of seminiferous tubules, germlinal cell degeneration/depletion, hyposperma, interstitial cell hyperplasia in the testis, Decrease ovary weights</td>
<td>Compound104</td>
<td></td>
<td>Testicular lesions: germ cell degeneration and necrosis</td>
</tr>
<tr>
<td>Compound024</td>
<td></td>
<td>Testicular degeneration/atrophy</td>
<td>Compound105</td>
<td></td>
<td>Testicular lesions: germ cell degeneration and necrosis</td>
</tr>
<tr>
<td>Compound026</td>
<td></td>
<td>Vaginal mucosal atrophy, Vaginal parakeratosis due to senescence</td>
<td>Compound110</td>
<td></td>
<td>Decreased epididymides and testes weights</td>
</tr>
<tr>
<td>Compound027</td>
<td></td>
<td>Abnormal luminal content, increased giant cells in seminiferous, necrosis of corpora lutea in ovaries</td>
<td>Compound111</td>
<td></td>
<td>Testic/epididymis, seminiferous tubules, spermatocytic degeneration, necrosis, mammary increases</td>
</tr>
<tr>
<td>Compound028</td>
<td>Acyclovir</td>
<td>Testicular atrophy, spermatogenesis [mouse] [41]</td>
<td>Compound112</td>
<td></td>
<td>Decreased prostate/semenal vesicle wt, delayed sperm release, delayed estrus cycle</td>
</tr>
<tr>
<td>Compound029</td>
<td>Nifedipine</td>
<td>Reduced weight of the testis and epididymis, reduced sperm count [42]</td>
<td>Compound116</td>
<td></td>
<td>Grossly small testes and epididymides</td>
</tr>
<tr>
<td>Compound031</td>
<td>Spironolactone</td>
<td>Low testosterone levels [43]</td>
<td>Compound118</td>
<td></td>
<td>Spermatogenic arrest in seminiferous tubules, spermatogenic epithelium was attenuated</td>
</tr>
<tr>
<td>Compound032</td>
<td></td>
<td>Low testosterone, leydig and sertoli cells were decreased</td>
<td>Compound119</td>
<td></td>
<td>Testic/epididymis, seminiferous tubules, spermatocytic degeneration, necrosis</td>
</tr>
<tr>
<td>Compound037</td>
<td>Chlorambucil</td>
<td>Damage spermatogenesis and cause testicular damage [44]</td>
<td>Compound122</td>
<td></td>
<td>Testicular seminiferous tubular degeneration, epididymal sperm granulomas, oligosperma and increased intraluminal cell debris in epididymides</td>
</tr>
<tr>
<td>Compound039</td>
<td>Flutamide</td>
<td>Increases in plasma testosterone level and Leydig cell hyperplasia, seminiferous tubular atrophy and degeneration [45]</td>
<td>Compound126</td>
<td></td>
<td>Seminiferous tubule and epididymal lesions</td>
</tr>
<tr>
<td>Compound042</td>
<td>Chlorpromazine</td>
<td>Decrease in testicular weight [46]</td>
<td>Compound128</td>
<td></td>
<td>Arteriopathy in Epididymides (dog)</td>
</tr>
<tr>
<td>Compound048</td>
<td>Divalproex</td>
<td>Reduced spermatogenesis and testicular atrophy [47]</td>
<td>Compound129</td>
<td></td>
<td>Testicular giant cells, epididymal spermatocytic granuloma</td>
</tr>
<tr>
<td>Compound053</td>
<td>Busulfan</td>
<td>Testes, tubules and germinal epithelia were decreased significantly, disrupt spermatogenesis through affecting both germ and somatic cells [48]</td>
<td>Compound130</td>
<td></td>
<td>Germ cell degeneration and necrosis</td>
</tr>
<tr>
<td>Compound057</td>
<td>Theobromine</td>
<td>Testicular atrophy with extensive spermatogenetic cell degeneration and necrosis [49]</td>
<td>Compound135</td>
<td></td>
<td>Spermatogenic tubular degeneration, decreased numbers of spermatids, sloughing of degenerate and necrotic cells in the tubular lumen, presence of giant cells and vacuolation of Sertoli cells</td>
</tr>
<tr>
<td>Compound059</td>
<td></td>
<td>Vacuolation in testis</td>
<td>Compound137</td>
<td></td>
<td>Seminiferous tubule degeneration</td>
</tr>
<tr>
<td>Compound060</td>
<td></td>
<td>Spermatogenic Retention in testis, Abnormal content in Epididymides</td>
<td>Compound138</td>
<td></td>
<td>Testicular degeneration</td>
</tr>
<tr>
<td>Compound061</td>
<td></td>
<td>Testicular degeneration</td>
<td>Compound139</td>
<td></td>
<td>Testicular degeneration and muscle necrosis</td>
</tr>
<tr>
<td>Compound063</td>
<td></td>
<td>Multifocal dilation of the seminiferous tubules, stops rat cycles</td>
<td>Compound140</td>
<td></td>
<td>Testicular degeneration, Seminiferous tubule degeneration</td>
</tr>
<tr>
<td>Compound064</td>
<td></td>
<td>Decreased uterine weights, subnuclear vacuolation, Increased estrus cycle length, increased number of animals in metestrus decreased conception, corpora lutea</td>
<td>Compound142</td>
<td></td>
<td>Delayed sperm release</td>
</tr>
</tbody>
</table>
A variety of statistical models were explored to find the best option for predicting the probability of toxicity. A detailed description of each of these models is given in Hastie et al. [21] and Kuhn and Johnson [22]. Many of these models have “tuning parameters” which cannot be estimated directly from the data. For these, the cross-validation procedure was executed for each value of the tuning parameter(s) and the best model was selected based on the area under the ROC curve. The models, and their tuning parameters, used here were:

- Random Forests (RF) is a tree-based ensemble model [23]. Five values of the tuning parameter, the number of random selected analytes at each split, were used to optimize the model.
- Support Vector Machines (SVM) [24] is a kernel-based machine learning method (the radial basis kernel was used in these models). For this model, five values of the cost parameter were tuned. Additionally, the method of Caputo et al. [25] was used to analytically estimate the radial basis function parameter.
- Nearest Shrunken Centroids (NSC) is a linear classification model with built-in feature selection. Forty values of the tuning parameter, the shrinkage threshold, were evaluated.
- Naïve Bayes (NB) uses the distribution of each analyte independently to estimate the probability of toxicity [21]. Two different methods for estimating the distributions (parametrically and non-parametrically) were tested.
- Logistic Regression is a well-known statistical technique that produces linear decision boundaries [26]. It will be demonstrated that there are many significant between-analyte correlations, which can cause instability in the model and may negatively affect performance. Because of this, a second order (L2) penalty will be used during model fitting to compensate for the collinearity of the analytes similar to classical ridge regression [27].

These models used all of the analytes data produced in the model (degree of change in gene expression for the steroid synthesis genes, as well as medium levels of steroids). Recursive feature...
elimination (RFE) was also used in conjunction of several of these models (except NCS). This backwards selection routine ranks the analytes by their importance to the model [24]. After the model with all analytes was created, the least important analytes were removed one at a time, the model was re-fit and performance was tracked. Using this profile of performance, the optimal number of predictors was determined and the final model was fit using only the top analytes.

Three measures of performance were used to characterize the effectiveness of the models:

- **Sensitivity:** given that a compound induces reproductive toxicity, what is the probability that the compound is predicted as toxic (a.k.a. true positive rate).
- **Specificity:** given that a compound does not induce reproductive toxicity, what is the probability that the compound is predicted as non-toxic (true negative rate).
- **Area under the ROC curve:** using the estimated probability of reproductive toxicity, the ROC curve tracks the sensitivity and specificity over a continuum of probability cut-offs [28]. The area under the curve can be used as a measure that combines sensitivity and specificity.

To obtain useful estimates of model performance, cross-validation was used on the training set samples [29] given the size of the data set. This is a form of “resampling” where random subsets of the samples are methodically held back from the data set. The model was fit to the majority of the data and the held-out samples were predicted. This was repeated multiple times and performance was estimated for each of the held-out samples and these values were aggregated into a single estimate of performance. In these analyses, five repeats of ten-fold cross-validation were used because of its attractive bias and variance properties. In the end, 50 different hold-out samples of roughly 10% were used to estimate performance. For recursive feature elimination, the same cross-validation scheme was utilized to estimate the uncertainty of the feature selection [30] while an internal cross-validation was used to tune the models.

3. Results

3.1. Assay performance using prototypical inducers

Development of the steroidogenesis assay in H295R cells included systematically evaluating different cell plating densities and times of sampling (Supplementary material related to this article can be found, in the online version, at http://dx.doi.org/10.1016/j.reprotox.2013.12.009). Because the goal of this work was to develop a cell-based screen that could be utilized for the evaluation of large compound sets, it would not be possible to use multiple timepoints in the final version of the assay. Conditions were optimized so that the two model compounds, forskolin and prochloraz, produced the greatest effect in the assay endpoints. These two compounds were then used as the assay’s positive controls and included in each subsequent run. Optimal conditions were determined to be 100,000 cells plated per well plated 48 h before compound addition and exposure for an additional 30 h.

Forskolin is an adenylyl cyclase activator that raises intracellular cAMP levels. It has been demonstrated to effect the enzymes of the steroidogenesis pathway as well as increase secreted progesterone and cortisol levels [31]. Treatment of cells with 10 μM forskolin for 30 h resulted in significant changes in gene expression. Robust induction of CYP11B2 (270-fold, p-value <0.0001) was observed. Forskolin exposure also produced strong inductions of CYP19A1 (26-fold, p-value <0.0001), CYP21A2 (12.3-fold, p-value <0.0005), and HSD3B2 (24.4-fold, p-value <0.0005), as well as more mild, but still significant increases in CYP11A1 (3.3-fold, p-value <0.00001), CYP17A1 (6.1-fold, p-value <0.0001), and STAR (5.1-fold, p-value <0.00001) expression [Fig. 2A]. These findings are similar in rank order to those reported by others [6]. Based on such gene changes, one would expect to see increased product being secreted, and this was the case: the expected increases in the secreted levels of cortisol (290%, p-value <0.0001) and progesterone (392%, p-value <0.0005) into the culture medium were observed [Fig. 2B]. Minimal effects on testosterone and DHEA were observed at this time point. Prochloraz is an antifungal compound that produces instances of male reproductive toxicity when administered to rats during sexual differentiation [32,33]. It has been reported to increase progesterone levels whilst decreasing testosterone levels in vivo [34]. Exposure of the cells to 4 μM prochloraz resulted in a robust induction of CYP11A2 (97-fold, p-value <0.0001) and modest inductions of CYP19A1 (3.4-fold, p-value <0.0001), CYP21A1 (3.4-fold, p-value <0.0005), and HSD3B2 (5.2-fold, p-value <0.001) (Fig. 2C). Treatment of H295R cells with prochloraz resulted in considerable increases in progesterone secretion (1838%, p-value <0.001) while substantially decreasing the secretion of cortisol (3%, p-value <10^-18), DHEA (7%, p-value <10^-18), and testosterone (9%, p-value <10^-17) after 30 h.

3.2. Compound test set administration

H295R cells were plated at 100,000 cells per well in 96 well plates for 48 h before compound administration in 8-point curves with a 300 μM starting maximal concentration. Cell Titer Glo (Promega) assay kit, which measures ATP, was used as a measure of cytotoxicity. TC10 (10% unspecified cytotoxicity) values were calculated from the curves and used as the treatment dose in the subsequent effect assay. A single concentration point was employed in order to enable this assay to be high throughput. The TC10 concentration was chosen as many known endocrine disruptors were shown to produce significant effects in the assay system at the TC10 concentration (Ulleras et al.) If no cytotoxicity was observed, the cells were treated with the compounds at 300 μM. Cells were treated in duplicate wells. The experiment was run under the previously described optimized conditions. Compound effects on gene expression of the enzymes in the steroidogenesis pathway as well as the level of secreted progesterone, testosterone, DHEA, and cortisol were measured and the results (see supplemental data Table 1) were used to build the statistical model.

3.3. Predictive modeling

ROC curve analysis was conducted for each analyte separately to obtain an initial indication of its utility as a predictor. For each predictor, a series of thresholds are examined to see if the toxic and non-toxic compounds can be classified using a simple cut-off point. The sensitivities and specificities resulting from the various cut-offs are then used to form the predictor’s ROC curve. The area under that curve is then used to quantify the predictive ability of each assay. If the assay can differentiate between toxic and non-toxic compounds on its own, the area under the ROC curve should be close to one. The closer to 1, the greater the value that endpoint has in predicting reproductive toxicity. The areas under the ROC curves for each individual analyte were: CYP11B1 (0.77), CYP11A1 (0.77), CYP21A1 (0.74), CYP19A1 (0.7), STAR (0.67), CYP11B2 (0.65), HSD3B2 (0.62), CYP17A1 (0.61), progesterone (0.55), testosterone (0.54), cortisol (0.54), DHEA (0.46). Here, the RT-PCR assays of gene expression showed higher individual predictivity of toxicity than the hormonal assays. However, many of the predictive models that were used here are able to model complex relationships between the outcomes and multiple predictors and, because of this, the
individual ROC curve results are poorly indicative of the relative importance of an individual assay to a final comprehensive model.

The cross-validation results are shown in Table 2. The mean and corresponding standard error were computed for the 50 estimates of performance produced during the cross-validation process. For example, the Naive Bayes model had area under the ROC curve values that were on average 0.756 but, given that the standard error of the mean was 0.0178, a 95% confidence interval for the mean of the cross-validation results was (0.720, 0.792). Many of the predictive models showed performance characteristics that were roughly equivalent, given the samples size. Unlike the other models, the Nearest Shrunken Centroid model achieved a reasonable area under the ROC curve (0.77), but did so by sacrificing model specificity (0.52) for sensitivity (0.79).

The performance profiles for the models when used in conjunction with recursive feature elimination are shown in Fig. 3. Although Random Forest shows a small decrease in performance as predictors are removed, the model with the largest area under the ROC curve (random forest) uses all the endpoints as useful predictors. The Naive Bayes model showed an increase in performance up to 3 analytes (CYP11B1, CYP11A1, and CYP21A1). Logistic Regression had small gains when eliminating predictors.

Of these models, we focused on the basic Random Forest model, which appeared to perform the best and Naive Bayes with feature selection classifier. Naive Bayes (RFE) was chosen as the comparator because it performs the best with fewer endpoints. This model would allow the assay to maintain predictive power while eliminating the need for hormone measurements and reducing the transcriptional endpoints, further enabling the high throughput ability of the assay. Random Forest can compute a built-in variable importance metric that quantifies the loss in performance if each predictor was coerced to be non-informative by randomly scrambling the data. For example, suppose the testosterone values were randomly scrambled and new model predictions were calculated. If performance did not drop by altering this predictor, it would not be considered important to the model. This same process would then be applied independently to each predictor. Using this process, the list of predictors from most important to least important is CYP11B1, CYP11A1, CYP21A1, CYP19A1, STAR, CYP11B2, HSD3B2, CYP17A1, progesterone, testosterone, cortisol, and DHEA. For the two other models, the number of times that each predictor was selected for each of the 1 cross-validation iterations can be used to measure the importance to the models. For Naive Bayes, the same assays were selected in each cross-validation iteration: CYP11B1, CYP11A1 and CYP21A1.

Fig. 4 shows the distributions of the class probabilities for the Random Forest and Naive Bayes models. The histograms show the probability of reproductive toxicity for each compound that was held-out during the cross-validation process. The bottom pair of histograms shows the probability of reproductive toxicity for compounds that were truly non-toxic in vivo. The Random Forest model shows the largest proportion of points between 20% and 40% probability. As the predicted probability of reproductive toxicity increases, fewer non-toxic compounds are misclassified as toxic. For the Bayesian model, a large number of non-toxic compounds have low probabilities (<20% probability) of being called toxic but, compared to Random Forest; more compounds are confidently mis-predicted as toxic in the Bayesian model. The upper
panels show similar results for the truly reproductive toxicants. Random Forest shows a broader probability distribution for reproductive toxicants while Naive Bayes tends to confidently predict reproductive toxicity (correctly or incorrectly).

Since the Random Forest model has a wider distribution of class probability values, one potential method for improving its performance is to institute an equivocal or indeterminate zone for random forest predictions. Here, we define a range of probability values that are too uncertain to confidently predict the compounds. For example, if the predicted probability of reproductive toxicity was between 40% and 60%, our confidence in the correctness of that prediction is low and the compound would not be classified as reproductive toxicant or non-toxic. For the Random Forest model, an equivocal zone between 40% and 60% would exclude, on average, 28.9% of the compounds but would increase the area under the ROC curve from 0.838 to 0.888. This approach would not be as effective for the Naive Bayes model since fewer compounds are predicted inside of the potential equivocal zone (as indicated by the zone between the dotted lines in Fig. 3).

The value of this approach increases with the criticality of the prediction. If one or more independent techniques will be used to verify the model prediction for a given compound, the equivocal zone may increase the cost and time of a model prediction. However, if very little evidence exists related to the toxicity of the compound (as is the case with new pharma candidates), the equivocal zone can play a pivotal role in increasing our confidence in the information used to judge a compound.

3.4. Reproductive toxicity predictions are independent of cytotoxicity

General cytotoxicity assays have been utilized by the pharmaceutical industry as a tool to predict in vivo toxicities. One commonly used assay employs THLE transformed liver cells and measures the amount of viable cells present by quantification of ATP after treatment with compounds [17]. We wished to determine if the predictions from this assay and algorithm was being driven more by drug effects on cellular health and cytotoxicity than drug-induced effects on steroidogenesis. THLE cells were treated with the compound test set in 10-point curves starting at 300 μM. Values were normalized to DMSO and IC50 were determined. The H295R random forest predictions were compared to the cytotoxicity IC50s in THLE cells (Fig. 5). No correlation was observed between the cytotoxicity measurement and the prediction of reproductive toxicity by random forest modeling. This comparison provides confidence that the H295R assay is predicting specific toxicity due to

### Table 2
Mean performance values estimated using cross-validation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Predictors</th>
<th>ROC Mean</th>
<th>ROC Std. err.</th>
<th>Sensitivity Mean</th>
<th>Sensitivity Std. err.</th>
<th>Specificity Mean</th>
<th>Specificity Std. err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Reg</td>
<td>12</td>
<td>0.778</td>
<td>0.014</td>
<td>0.708</td>
<td>0.023</td>
<td>0.704</td>
<td>0.025</td>
</tr>
<tr>
<td>Logistic Reg (RFE)</td>
<td>7</td>
<td>0.801</td>
<td>0.014</td>
<td>0.745</td>
<td>0.021</td>
<td>0.731</td>
<td>0.025</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>12</td>
<td>0.756</td>
<td>0.018</td>
<td>0.657</td>
<td>0.023</td>
<td>0.754</td>
<td>0.023</td>
</tr>
<tr>
<td>Naive Bayes (RFE)</td>
<td>3</td>
<td>0.812</td>
<td>0.015</td>
<td>0.655</td>
<td>0.019</td>
<td>0.780</td>
<td>0.022</td>
</tr>
<tr>
<td>NSC</td>
<td>4</td>
<td>0.773</td>
<td>0.017</td>
<td>0.789</td>
<td>0.025</td>
<td>0.524</td>
<td>0.026</td>
</tr>
<tr>
<td>Random Forest</td>
<td>12</td>
<td>0.847</td>
<td>0.013</td>
<td>0.724</td>
<td>0.019</td>
<td>0.758</td>
<td>0.023</td>
</tr>
<tr>
<td>Random Forest (RFE)</td>
<td>11</td>
<td>0.845</td>
<td>0.013</td>
<td>0.751</td>
<td>0.018</td>
<td>0.745</td>
<td>0.025</td>
</tr>
<tr>
<td>SVM</td>
<td>12</td>
<td>0.852</td>
<td>0.012</td>
<td>0.748</td>
<td>0.024</td>
<td>0.791</td>
<td>0.019</td>
</tr>
<tr>
<td>SVM (RFE)</td>
<td>10</td>
<td>0.837</td>
<td>0.011</td>
<td>0.753</td>
<td>0.020</td>
<td>0.771</td>
<td>0.020</td>
</tr>
</tbody>
</table>

**Fig. 3.** Performance profiles for different models when conducting recursive feature elimination.
alteration in the steroidogenesis pathway and not merely general cytotoxicity mechanisms.

4. Discussion

This study evaluated a mechanistic approach to identify reproductive toxicants by examining their effects on steroidogenic gene expression and selected steroid output in the H295R cell culture system. The H295R in vitro cell system is being employed throughout the world for use in screening environmental agents for possible endocrine disruption activity that may result in developmental and reproductive problems in both humans and wildlife. The H295R cell line expresses each of the enzymes in the steroidogenic pathway and secretes the hormones of interest at detectable levels into the culture medium making this an ideal in vitro model for studying the steroidogenic pathway. The steroidogenic pathway is made up of multiple enzymes and there are numerous control points in the production of steroids. The ability to simultaneously assess the effect of compounds upon multiple steps within the pathway in vitro and employ a statistic model to correlate these effects with a toxic/non-toxic in vivo prediction is of great use to those in the drug development process [35].

We used forskolin and prochloraz, two compounds well-studied for their effects on the steroidogenesis pathway, to determine the effectiveness and robustness of this cell system. Both compounds were able to differentially induce the transcriptional levels of the studied enzymes. Forskolin induced a robust increase in progesterone and cortisol levels secreted by the H295R cells. Prochloraz induced progesterone, but resulted in dramatic decreases in testosterone, DHEA and cortisol. This demonstrated the cell system’s ability to identify compounds that both increase and decrease hormone levels. It should be noted that modulators of enzyme expression are not the same as modulators of enzyme activity. This assay system does not detect compounds which alter the activity of an enzyme, except by inference from changes in steroid levels themselves.

Indeed the robust increases in progesterone observed when treating the H295R cells with prochloraz can not be explained by the transcriptional changes of the enzymes. CYP17A1, responsible for its conversion, is increased in expression. However, others have reported that the hydrolase activity of CYP17A1 is severely inhibited by similar levels of prochloraz treatment in H295R cells [36]. This decreased activity is therefore likely the cause for the increases progesterone levels observed in our system.

We were able to create a model that predicts reproductive toxicants with 76% sensitivity and 72% specificity for an ROC of 0.85.
using the Random Forest model with all 12 predictors. Using the Naïve Bayes model with the recursive feature elimination including only 3 predictors, the ROC drops to only 0.82, yet makes the assay easier and more amenable to a highthroughput system by eliminating the hormone measurement and limiting the amount of Q-PCR required, while maintaining predictive power. Combining multiple endpoints from this assay has yielded a model which predicts better than any one endpoint alone, demonstrating the multiparameter approach necessary and beneficial. The majority of publications describe the use of screening H295R cells for hormone measurements only [9, 14, 37]. Those enlisting these cells to predict toxic outcomes without incorporating the gene expression arm are likely missing important information and could likely benefit from inclusion of a genomics assay.

Finally, we should point out that this assay does not predict only compounds with hormonal effects in vivo. Indeed, many of the compounds used here have no evidence of hormonal changes (Table 1). We recognize that it’s possible that many of those toxics with no detectable hormonal changes may, indeed, produce some subclinical alterations in the reproductive hormone economy, changes which would not be picked up by the relatively insensitive methods available to investigators (pathology, single-timepoint hormone measurements, organ weights). In any case, it’s worth noting that we view this assay as being useful for identifying all reproductive toxins, regardless of their degree of hormonal impact. We recognize the interpretational challenges this poses, but we feel the data are compelling.

The field of predictive toxicology has been rapidly developing. It includes epigenetics, toxicokinetics, genomic biomarkers, systems biology, in vitro assays, predictive modeling and computational science. These techniques are being incorporated more often in industry as we strive to reduce the costs of drug development by limiting expensive animal studies to only the most promising compounds. This has added a benefit of a reduction in animal usage. Our goal was to develop and implement an in vitro method to employ as early as the drug discovery process as possible to predict reproductive toxicity. We characterized the H295R system as an impressive model for reproductive toxicity that can be employed to rank order compounds and to select the safest possible compounds to be used in animal studies.

**Conflict of interest**

All authors were employees of Pfizer Global Research and Development, which funded this study, at the time the experimental work was conducted.

Transparency document associated to this article can be found in the online version, at http://dx.doi.org/10.1016/j.pertox.2013.12.009.

**References**


