Computational Approaches to Assessing Clinical Relevance of Pre-clinical Cancer Models

Cancer is one of the leading causes of death worldwide, and therefore a priority in biomedical research. However, multiple disease causing mechanisms and histopathological heterogeneity mean that developing treatments is a highly complex problem [1]. One of the main goals of oncology-related research is the development of anti-cancer drugs. Before a potential drug can be used as a therapy it must go through several stages of research to ensure its effectiveness, safe dosage and side effects are elucidated. With 7% combined likelihood of approval, phase 2 and 3 drug failure rates in oncology are higher than those in other medical disciplines [2]. Not only is this problematic for the search of an effective treatments for patients, but also carries high financial and logistical price. Typically, phase 2 and phase 3 trials involve hundreds of participants and have combined average cost of around 60 million USD [3]. This makes failed drug trials costing 56 million USD. One proposed cause of the high failure rate is inefficiency of current pre-clinical models in modelling cancer heterogeneity [4].

Pre-clinical cancer models, such as tumour-derived cell-lines and animal models, are essential in cancer research. Consistently used as a platform to investigate mechanism of action, they can identify potential biomarkers prior to clinical trials where similar exploration is more complicated, unethical and expensive. However, whilst cell-lines are the most used pre-clinical model, their applicability in certain cancer settings is questioned because of the difficulty of aligning the appropriate cell-line with a clinically relevant disease segment. Likely caused by some of the known biological differences between cell-lines and tumours (up-regulation of genes related to *in vitro* survival, low heterogeneity, lacking of tumour micro-environment and vascular system).

We aim to develop computational methods which would determine, for a given pre-clinical model, its suitability for evaluating novel therapeutics, and the most appropriate disease segment represented by that model. Ultimately these would take form of am easily accessible web-based tool.
This would enable researchers to score their cell-lines prior to experiments, helping them select the most appropriate ones, thus increasing the chance of finding effective therapeutics and reducing costs and time associated with failure.

Genomics profiling data from patient tumours and cell-lines were used to train and test the method. Machine learning techniques (including random forests, principal component analysis, Gaussian processes) were applied to create predictive models based on patient training data. Their accuracy was evaluated on the patient test set and then applied to cell-line data.

Endometrial and breast cancer types were used and their available corresponding subtypes. For the endometrial cancer, 516 cancer samples and 28 cancer cell lines were used, while for the breast cancer it was 937 cancer samples and 59 cell lines. Genomics profiling data (copy number, expression) and clinical data from patient tumours and cell lines were used to train and test the method. These have been obtained from The Cancer Genome Atlas (TCGA) [5] and Cancer Cell Line Encyclopaedia (CCLE) [6] through http://firebrowse.org/ and https://portals.broadinstitute.org/ccle/. Data types used were GISTIC2.0 copy number, RSEM and Kallisto quantified expression. The datasets contained gene measurements (depending on the data type) of a list of genes for a cancer sample or a cell line. Genes that were not present in both the CCLE and TCGA data were eliminated from the analysis. CCLE copy number data was comprised of 23316 different genes while the expression data had 18988 genes. Starting TCGA data contained 24776 different genes in case of copy number of both cancer types, 20533 for endometrial expression and 20532 for breast cancer expression.

Several machine learning techniques were employed to build a model out of the tumour data which was afterwards used to evaluate the cell-lines.
Random forests (R package 'randomForest') and Gaussian process (python package 'GPy') classifiers use classes associated with samples in the training set to learn how to differentiate between classes based on the features of the samples. Afterwards their effectiveness is evaluated on the test set to see how well they can assign classes to new, unseen data.
Classes for the classifier were obtained from TCGA clinical data files. Endometrioid and serous subtypes for endometrial cancer and triple negative or non-triple negative for breast cancer. Classifiers performed extremely well (0.90+ AUC) on the tumour test set for the data type that was more relevant for the cancer type (copy number for endometrial, expression for breast cancer). When applied to cell-lines, most of them scored in line with their assignment found in the literature.

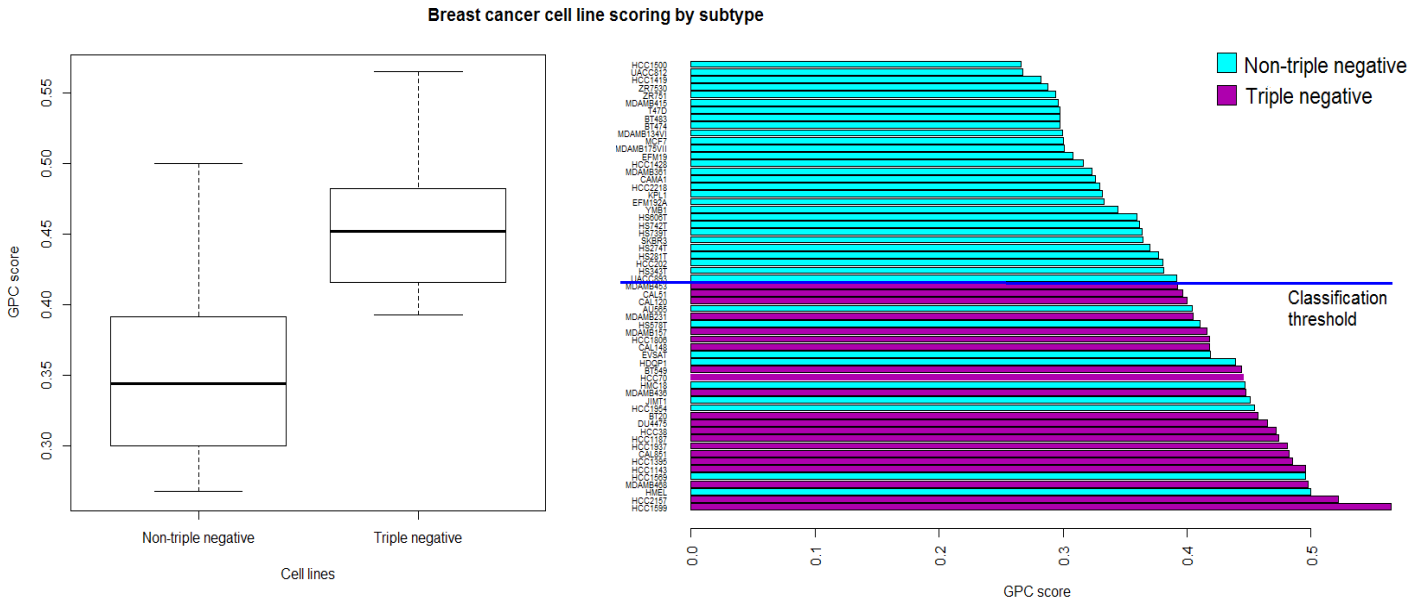

Figure 1: Breast cancer cell lines scored with GPy classifier based on expression. Colour indicates what the literature search says about their classes

Principal component analysis (PCA) transforms the features of the samples into new ones (principal components) which have the highest possible variance making it easier to see meaningful patterns in the data.

Since it doesn't produce a simple score like the mentioned classifiers, but a large number of principal components, it allows for a better exploration of relevant (new) features. Also, without classes, it doesn't have the problem of determining whether two classes are appropriate or not for the model and avoids forcing a class on a sample which might be unfitting for either classes. However, just by itself, it doesn't assign classes to new samples, so it doesn't have a predictive functionality.

Cancer subtypes have significantly different principal component distributions. While cell-line position does go well with the scores of the used classifiers, PCA does unveil an interesting pattern - according to distance from centroid of a subtype cluster, cell-lines inappropriate for one subtype are likely to be inappropriate for the other one too.

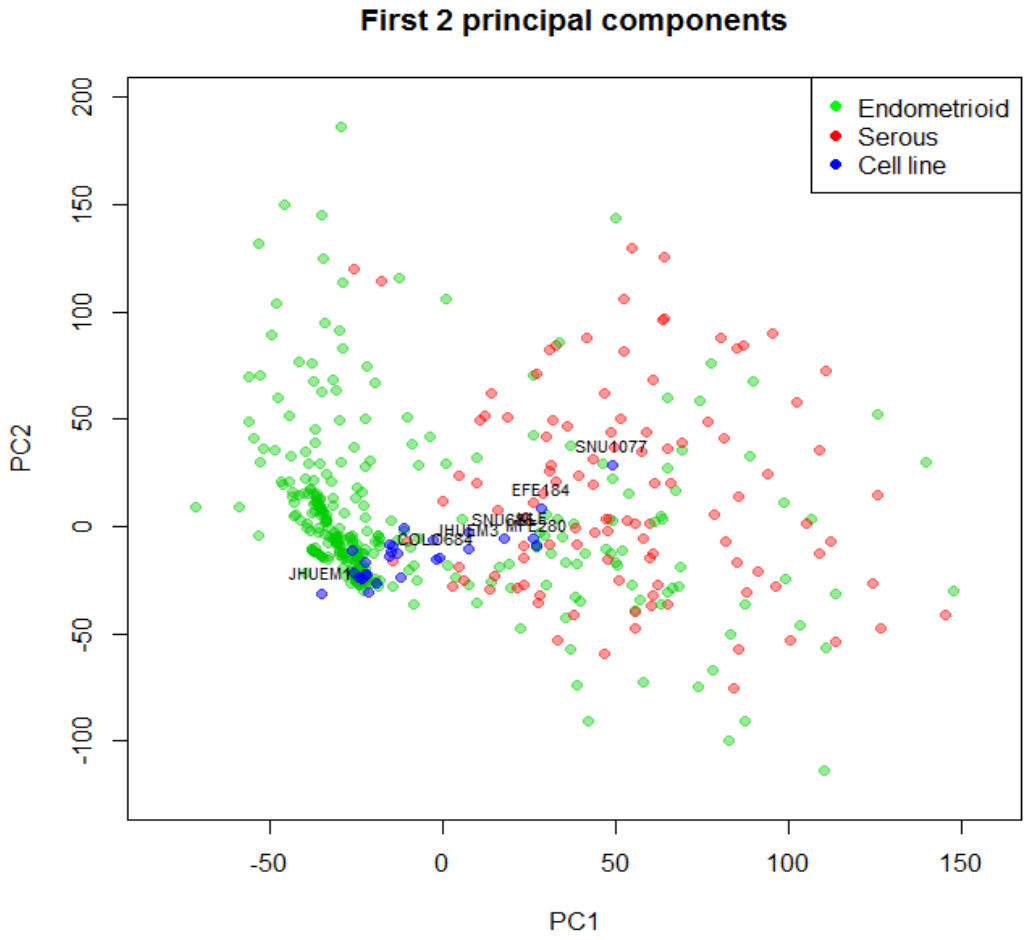

Figure 2: Values of first 2 principal components for endometrial cancer subtypes and endometrial cancer cell lines.

Classifiers scoring high on test set and cell-line scores and outliers being mostly in line with the literature suggest that methodology is appropriate. Misaligned cell-lines might also indicate potential misclassification in the literature, which, considering the age of some cell-lines, could be a realistic scenario. Ambiguity of cell-line scores with respect to differences in the subtype score (PCA distance to centroid) might indicate the general genetic drift of cell-lines away from tumours. Further improvements include investigating biological causes, selection of more subtle classes, extending to other tumours and building a web-tool for systematic evaluation of cell-line suitability.

[1] Gazdar, A. F., L. Girard, W. W. Lockwood, W. L. Lam and J. D. Minna (2010). "Lung Cancer Cell Lines as Tools for Biomedical Discovery and Research." Jnci-Journal of the National Cancer Institute 102(17): 1310-1321.

[2] Hay, M., D. W. Thomas, J. L. Craighead, C. Economides and J. Rosenthal (2014). "Clinical development success rates for investigational drugs." Nature Biotechnology 32(1): 40-51.

[3]US Department of Health and Human Services. https://aspe.hhs.gov/report/examination-clinical-trial-costs and-barriers-drug-development

[4] Yang, W., J. Soares, P. Greninger, E. J. Edelman, H. Lightfoot, S. Forbes, N. Bindal et al. (2013). "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells." Nucleic Acids Research 41(D1): D955-D961.

[5]Weinstein, John N.. et al. (2013). ”The Cancer Genome Atlas Pan-Cancer analysis project.” Nature Genetics 45(10): 1113-1120., available at http://firebrowse.org/

[6]Baretina, J. et al. (2012). ”The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.” Nature 483(7391): 603-607., available at http://www.broadinstitute.org/ccle/