Libertas Academica
FREEDOM TO RESEARCH

**Open Access:** Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

**Supplementary Issue: Array Platform Modeling and Analysis (A)**

# Stratified Pathway Analysis to Identify Gene Sets Associated with Oral Contraceptive Use and Breast Cancer

Herbert Pang[1,2] and Hongyu Zhao[3]

[1]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC, USA. [2]School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong, China. [3]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA.

**ABSTRACT:** Cancer biomarker discovery can facilitate drug development, improve staging of patients, and predict patient prognosis. Because cancer is the result of many interacting genes, analysis based on a set of genes with related biological functions or pathways may be more informative than single gene-based analysis for cancer biomarker discovery. The relevant pathways thus identified may help characterize different aspects of molecular phenotypes related to the tumor. Although it is well known that cancer patients may respond to the same treatment differently because of clinical variables and variation of molecular phenotypes, this patient heterogeneity has not been explicitly considered in pathway analysis in the literature. We hypothesize that combining pathway and patient clinical information can more effectively identify relevant pathways pertinent to specific patient subgroups, leading to better diagnosis and treatment. In this article, we propose to perform stratified pathway analysis based on clinical information from patients. In contrast to analysis using all the patients, this more focused analysis has the potential to reveal subgroup-specific pathways that may lead to more biological insights into disease etiology and treatment response. As an illustration, the power of our approach is demonstrated through its application to a breast cancer dataset in which the patients are stratified according to their oral contraceptive use.

**KEYWORDS:** cancer, random forests, pathways, progesterone receptor

**SUPPLEMENT:** Array Platform Modeling and Analysis (A)

**CITATION:** Pang and Zhao. Stratified Pathway Analysis to Identify Gene Sets Associated with Oral Contraceptive Use and Breast Cancer. *Cancer Informatics* 2014:13(S4) 73–78 doi: 10.4137/CIN.S13973.

**RECEIVED:** June 15, 2014. **RESUBMITTED:** August 15, 2014. **ACCEPTED FOR PUBLICATION:** August 19, 2014.

**ACADEMIC EDITOR:** JT Efird, Editor in Chief

**TYPE:** Technical Advance

**FUNDING:** Support was provided in part by NIH grants GM 59507, CA 154295, and P30CA016359. The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

**COMPETING INTERESTS:** Authors disclose no potential conflicts of interest.



**CORRESPONDENCE:** pathwayrf@gmail.com

Paper subject to independent expert blind peer review by minimum of two reviewers. All editorial decisions made by independent academic editor. Prior to publication all authors have given signed confirmation of agreement to article publication and compliance with all applicable ethical and legal requirements, including the accuracy of author and contributor information, disclosure of competing interests and funding sources, compliance with ethical requirements relating to human and animal study participants, and compliance with any copyright requirements of third parties.

## Introduction

Combining high-throughput microarray data with pathway information has proved to be a fruitful approach to uncovering biological networks from genomics data. Compared to single-gene based analysis, pathway-based methods can identify more subtle changes in expression.[1,35] Furthermore, pathway-based methods can help generate biological hypotheses, which can be readily tested using complementary approaches, such as proteomics and metabolomics technologies. Each pathway used in such analyses generally serves a particular cellular or physiologic function, and these annotated pathways usually come from various external databases, such as KEGG[2] and BioCarta (http://www.biocarta.com/). Wang et al.[3] recently reviewed pathway-based methods, including enrichment analysis, nonparametric regression, discriminant analysis, partial least squares, and random forests. Despite these recent progresses in pathway-based analysis, not much attention has been paid to sample heterogeneity that may be partly captured by observable clinical variables. The incorporation of such clinical information may help identify relevant pathways unique to a subpopulation, leading to novel insights into disease etiology and more specific treatment schemes.

In this paper, we propose to perform stratified pathway analysis for subpopulations, with attention to the analysis

results that overlap across strata as well as those that are unique to one of the stratum. If the pathways are found to be significant only to one of the subpopulations, it means that potentially those pathways are switched on for that subgroup. If the pathways are found to be significant for both subpopulations, it suggests that the genes are consistently switched on for both subpopulations, ie, regardless of the stratifier. We primarily study a breast cancer dataset to illustrate the usefulness of this approach. Estrogen receptor (ER) and progesterone receptor (PR) statuses are commonly used to estimate the risk of breast cancer, design therapy, and predict survival rate.[4–10] For example, breast cancer patients are usually treated either with hormone therapy or chemotherapy depending on the hormone status of ER and PR. However, not all breast carcinomas are responsive to the treatment, and pathway analysis may help identify novel therapeutic targets and develop new agents. Pathways or genes that can predict the PR status could potentially tell us more about the biological mechanism of the disease. PRs have been found to provide prognostic information as well.[11,12] Oral contraceptives are known to increase the risk of pre-menopausal breast cancer.[13] Additional evidence supporting hormonal use as a confounding factor toward the risk of breast cancer was found based on the data from the Women's Health Initiative.[14] Another study found that women who have taken contraceptive pills are less likely to die of cancer or heart disease.[15] This is strong evidence that oral contraceptive use can be a confounder in classifying breast cancer samples using hormonal status. Therefore, we perform stratified pathway analysis based on oral contraceptive use to explore potentially different pathways involved in breast cancer. More specifically, we aim to identify pathways involved in distinguishing PR status for users and non-users of oral contraceptives based on gene expression data. Top pathways may contain genes with expression that are good at distinguishing receptor status. For example, progesterone expression may predict progesterone status as it is related to genetic loss of heterozygosity.[16] Other genes may serve as surrogates for that process.

For pathway analysis, we use the Random Forests approach, which has been found to perform well among a number of machine learning methods in pathway-based analysis.[17,18]

The rest of the paper is organized as follows. The detailed methodology is discussed in the Materials and methods section. In the Results section, we demonstrate the usefulness of this approach through the application of our method to a breast cancer microarray dataset. We conclude the paper in the Discussion section.

## Materials and Methods

Our approach is built on the previous proposal of adopting the Random Forests approach for pathway analysis.[17] We describe below how we perform stratified pathway analysis, build pathway connections, compare overlapped pathway similarities, and discover how genes are shared among them.

**Stratified pathway analysis.** The stratified pathway analysis considers important covariates in data analysis. In the case of the breast cancer dataset to be analyzed, we use information about the oral contraceptive use of each patient. In the simplest case, we partition all the samples into subgroups based on oral contraceptive use status, and analyze each subpopulation separately. For each pathway, we build a Random Forest to predict an individual PR+/PR– status based on the gene expression levels within this pathway. To understand Random Forests, we first need to understand how to build a classification tree.

A classification tree is built as follows:

Step (I): For each pathway, draw a bootstrap sample from the original data.

Step (II): A classification tree is grown for each bootstrap sample.

Step (IIa): At each node of the tree, select predictors ($\sqrt{p}$ for classification) at random for splitting.

Step (IIb): Using the gini impurity criterion described later, a node is split using the single predictor from step (IIa). Gini impurity criterion for a binary classification problem is $1 - p_1^2 - p_2^2$, with $p_1$ and $p_2$ being the proportions of individuals in class 1 and 2, respectively.

Step (IIc): Repeat steps (IIa) and (IIb) until each terminal node contains samples in the same class or only one sample.

Random Forests construct many classification trees and thus the name 'forest'. Every tree is built using a deterministic algorithm that differs from ordinary tree algorithms in two regards. First, at each node, a best split is chosen from a random subset of predictors rather than all of them. Second, every tree is built using a bootstrap sample of the original observations. For more details, see Breiman.[19] The default parameters in R's Random Forest implementation are used, except for running 20,000 trees.

For calculating the classification error, we employ five-fold cross-validation to make it more stable across the stratified subgroup. For each cross-validation, four-fifths of the samples are used to build Random Forests, and the other samples are used to estimate the classification error. In doing so, each subject from the left out set of the cross-validation iteration is put down every tree in the forest for classification using the input vector of gene expression for genes within a particular pathway. Each tree gives a classification for this subject, and the forest chooses the class that gives the majority votes for this subject. Small classification error based on genes in a given pathway would indicate the pathway as potentially interesting.[17] Multiple-testing adjusted permutation *P*-values are calculated. The permutation *P*-value is the proportion of observed cross-validation errors smaller than the cross-validation errors obtained from 500 Random Forest runs of randomly permuted labels of patients. The top pathways with a permutation

*P*-value of less than 0.05 from both users and non-users of oral contraceptives will be presented for overlapped and non-overlapping pathways. Important genes, ie, those having strong discrimination power and selected based on the importance measure in Random Forests, are also investigated.

We will provide some biological interpretations of our results with the help of PubMatrix,[20] software for comparing a list of terms against any other list of terms in PubMed.

**Datasets.** *Pathways.* A total of 285 pathways from BioCarta were used for the analysis. Most of these are related to signal transduction in humans with a smaller group of metabolic pathways.

*Microarray data.* A breast cancer microarray dataset accompanied by oral contraceptive use data was analyzed. This dataset used Affymetrix GeneChip® hgu-133 plus 2.0 with 54,613 probe sets. The INTEGEN (http://www.integen.org/) dataset (accession number GSE2109 from GEO) consists of 199 breast tissue samples with clinical status of PR. A total of 123 of the 199 samples were taken from patients with oral contraceptive use and the remaining 76 from patients who did not take oral contraceptives.

We chose the breast cancer dataset and PR positive/negative status (PR+/PR–) to study because breast cancer has been extensively studied in the literature and tumor samples are normally classified on the basis of hormone receptor status. A recent publication described PR as a stronger predictor of treatment response of adjuvant tamoxifen than the ER.[21] More recently, additional studies suggested that measurement of PR status in conjunction with ER status may help identify patients that benefit from therapy.[22,23] Daniel et al.[24] provide strong rationale for targeting PR and ER in combination. The PR status has also been used to guide breast cancer therapy, breast cancer survival rate, and estimate breast cancer risk.[4,5,9]

*Software.* R was used to perform stratified pathway analysis. The R code is based on pathway analysis using Random Forests and it is available here: http://people.duke.edu/~hp44/r_code.htm. For pathway visualizations, Cytoscape[25] was used.

## Results

**Top pathways.** In this section, we show the results of the analyses. Top pathways for both users and non-users of oral contraceptives that are good at distinguishing between PR+ and PR– with permutation *P*-values of less than 0.05 are presented. Table 1 lists the top four pathways for non-users of oral contraceptives while the top five pathways for users of oral contraceptives are shown in Table 2. The top pathways in Table 1 are: (i) Eph Kinases and ephrins support platelet aggregation and (ii) IL 10 Anti-inflammatory Signaling Pathway. It has been found that Eph–ephrin signaling and IL 10 signaling are related to cancer[26,27] and to breast cancer in particular.[28] The following genes are important for the classification between PR+ and PR– in the pathways in Table 1: *ADCY1*, *MAP3K7*, and *PAK2*. And for Table 2, these are the important genes that are shared in some of the pathways listed: JUN, MAPK3, PIK3C, PIK3R1, and SOS1. Moreover, oral ethinylestradiol, an active estrogen compound found in oral contraceptives, decreased expression of chemokine receptors such as CCR3.[29] This potentially explains why CCR3 signaling in eosinophils pathway was found as one of the top pathways among oral contraceptive users. About 27% of the important genes among the top pathways for non-users of oral contraceptives have literature citations compared with 39% of the important genes among the top pathways for users of oral contraceptives. While some pathway names and relationships to cancer are less apparent, the top genes in them may be

**Table 1.** Top pathways for non-users of oral contraceptives.

| PATHWAY NAME | *P*-VALUE* |
|---|---|
| Eph Kinases and ephrins support platelet aggregation | 0.030 |
| IL 10 Anti inflammatory signaling pathway | 0.010 |
| Regulation of spermatogenesis by CREM | 0.012 |
| TNFR1 signaling pathway | 0.018 |

**Note:** *From permutation.

**Table 2.** Top pathways for users of oral contraceptives.

| PATHWAY NAME | *P*-VALUE* |
|---|---|
| CCR3 signaling in eosinophils pathway | 0.012 |
| Neuropeptides VIP and PACAP inhibit the apoptosis of activated T cells | 0.004 |
| PDGF signaling pathway | 0.008 |
| The IGF 1 receptor and longevity | 0.024 |
| Transcription factor CREB and its extracellular signals | 0.010 |

**Note:** *From permutation.

**Table 3.** Overlapping top pathways for both users and non-users of oral contraceptives.

| PATHWAY NAME | *P*-VALUE 1 | *P*-VALUE 2 |
|---|---|---|
| IL 2 receptor beta chain in T cell activation | 0.006 | 0.002 |
| IL 6 signaling pathway | 0.022 | 0.032 |
| Keratinocyte differentiation | 0.002 | 0.002 |
| Pelp1 modulation of estrogen receptor activity | 0.004 | 0.002 |
| Rho cell motility signaling pathway | 0.014 | 0.006 |
| Estrogen-responsive Efp controls cell cycle breast tumors growth | 0.008 | 0.026 |
| Role of ERBB2 in signal transduction and oncology | 0.002 | 0.002 |

**Note:** *From permutation *P*-value 1 and *P*-value 2 for non-users and users, respectively.

biologically meaningful. The most important genes in Table 3 will be explored in more detail in the next section. Given that the sample size for oral contraceptives users is larger than that for the non-users, we investigated whether the sample size had an impact on the number of top pathways by sampling only 76 from the 123 oral contraceptive patients. All the pathways listed in Table 2 remain significant and have permutation *P*-values less than or equal to 0.05. If we did not stratify by the users and non-users of oral contraceptives, then four of the five top pathways are different from what have been found, including MAPKinase Signaling Pathway; Telomeres, Telomerase, Cellular Aging, and Immortality; CARM1 and Regulation of the Estrogen Receptor pathway; and Cell to Cell Adhesion Signaling pathway. Keratinocyte Differentiation pathway is the only one found in Table 3.

**Important genes and biological implications.** Figure 1 contains two plots, each showing the set of pathways listed in Table 3 as well as the corresponding important genes. The top and bottom halves of Figure 1 correspond to non-users and users of oral contraceptives, respectively. The genes are hexagon shaped and are shaded according to their discriminative power in distinguishing PR+/PR– samples, with darker

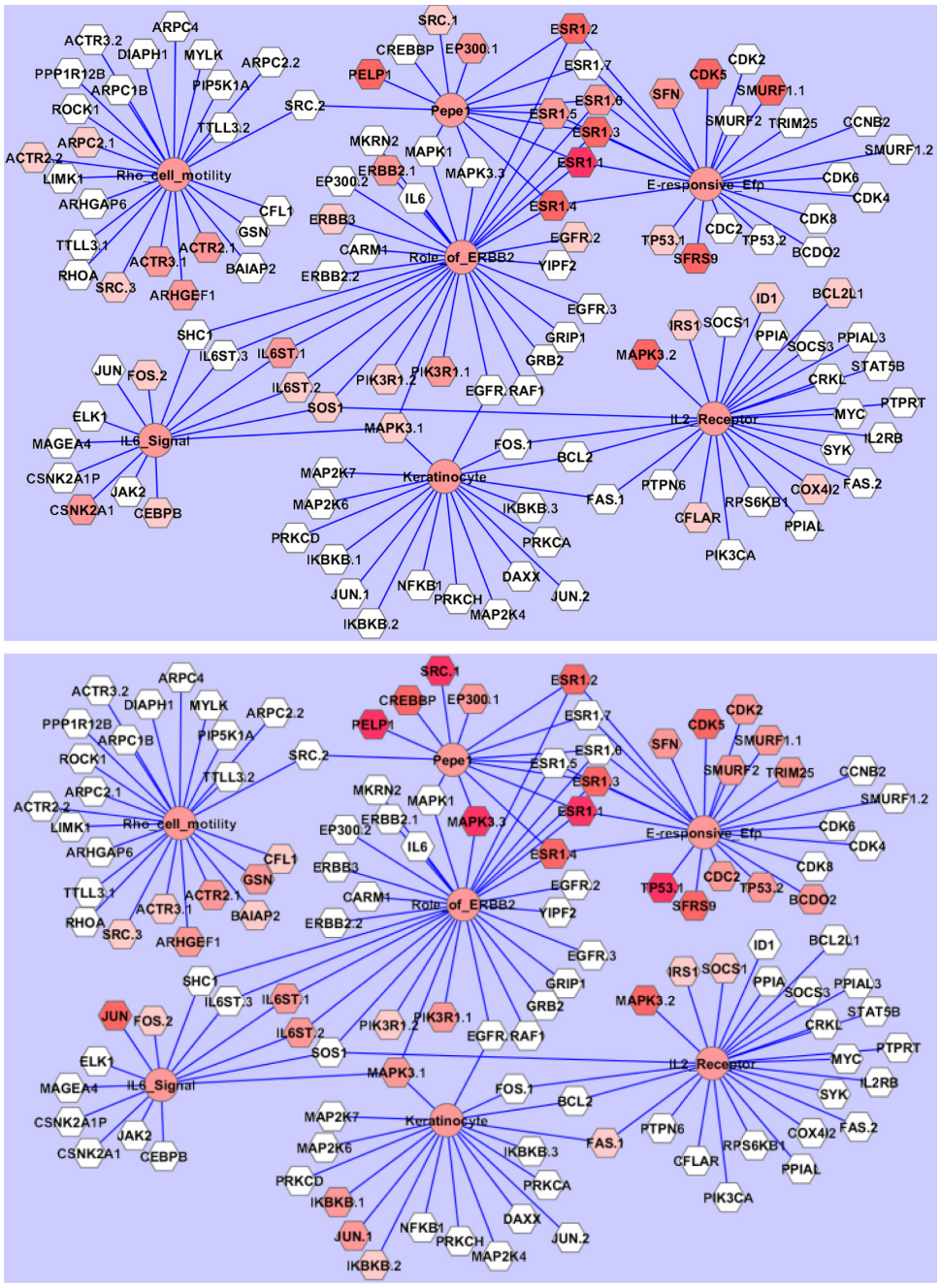


**Figure 1.** Top overlapped pathways for non-users (top) and users (bottom) of oral contraceptives.
**Notes:** Hexagon shaped are genes. Dark red as most important, white as least important.

red indicating more discriminative power. Clearly as an overview of the plot, we can see that users of oral contraceptives have darker red genes than non-users of oral contraceptives. Moreover, it is important to note that there are genes that are important for distinguishing PR status in non-users but not in users of oral contraceptives and vice versa.

The genes that are colored red and common among them include: *ESR1*, *SFN*, *TP53*, *MAPK3*, *IRS1*, *IL6ST*, *ACTR2*, *ACTR3*, *SRC1*, and *PELP1*. Of these genes, *ESR1*, *IL6ST*, and *MAPK3* are shared among pathways, and may therefore help facilitate pathway crosstalk. Now, let us look at the unique top important genes separately. We make use of PubMatrix, web-based software, to compare a list of terms against any other list of terms in PubMed. The top important genes *SOCS1*, *CREBBP*, *IKBKB*, *CDC2*, *CDK2*, *FAS*, *JUN*, and *FOS* of the bottom pathway cluster in Figure 1 have 102 literature citations relating them to oral contraceptives from PubMatrix. Whereas for non-users of oral contraceptives, the top important genes, *PRKCD*, *CFLAR*, *EGFR*, *BCL2L1*, and *ERBB2*, only have 30 literature citations relating them to oral contraceptives from PubMatrix. Interestingly, these genes are unique among pathways and are not shared. A Mann–Whitney *U* test gives a two-sided *P*-value of 0.075 when comparing the number of literature in user and non-user groups for the top five genes. Furthermore, using GeneGO, we identified therapeutic drug targets for eight of the top genes identified, see Table 4.

## Discussion

Stratified pathway analysis refers to performing pathway analysis and measuring prediction accuracy within subgroups or strata of the experimental population. Clinical covariates that may confound the pathway analysis or gene-set enrichment analysis should be taken into account. If necessary, stratified pathway analysis should be performed. In this paper, we highlight the importance of incorporating covariates in pathway analysis. We have described a Random Forests-based approach to perform stratified pathway analysis in distinguishing PR positive and negative breast cancer patients. The novel method presented in the article was able to identify unique pathways specific to oral contraceptives users and non-users as well as shared pathways among those groups. In addition, we were able to tease out the important genes that relate to outcome of interests that are biologically meaningful. Although we used Random Forests for classification, other methods, such as support vector machines, can also be employed here for stratified analysis.

We have demonstrated the biological relevance of our approach using PubMatrix. The number of important genes identified with literature agreed well with those shortlisted by our analyses. Furthermore, with the aid of network visualization tools, we can allow biologists to investigate how the important genes are related to each other within a set of pathways. One limitation of our approach is that the computational intensity of our approach increases linearly with the number of levels of the confounding variable.

Bioinformaticians and biologists can make use of this method to analyze specific subgroup of patients, focus on a few sets of genes, identify pathway targets, and find out how important genes are shared among the top pathways. This allows researchers to obtain results that are more closely tied to the biological mechanism of diseases. This analysis can also be applied to ER status. In this case, a weighted random forests algorithm should be used to deal with the unbalanced proportion of ER positive and negative groups.[30] Other machine learning gene selection strategies may be incorporated.[31,32] Moreover, it may be applied to genotyping data as well.[33,34]

**Table 4.** Therapeutic drug target for top genes.

| KNOWN DRUGS (EXCLUDING EXPERIMENTAL) | GENE SYMBOL |
|---|---|
| Afatinib, Neratinib, Lapatinib, Canertinib, Gefitinib | ERBB2 & EGFR |
| Roniciclib, Alvocidib | CDK1 & CDK2 |
| Erlotinib, Falnidamol, Vandetanib, Genistein, Cediranib, Varlitinib, Pelitinib, Suramin | EGFR |
| Ingenol | PRKCD |
| Raloxifene, Afimoxifene, Megestrol, Diethylstilbestrol, Clomifene | ESR1 |
| Riviciclib | CDK1 |
| Bardoxolone, Xanthohumol | IKBKB |
| Seliciclib | CDK2 |
| Navitoclax | BCL2L1 |
| Masoprocol | ERBB2 |

## Acknowledgments

We thank the International Genomics Consortium (IGC) and Expression Project For Oncology (expO) for making the breast cancer dataset available to us. We also thank Michael Klein for proofreading our article.

## Author Contributions

Conceived and designed the experiments: HP, HZ. Analyzed the data: HP. Wrote the first draft of the manuscript: HP, HZ. Contributed to the writing of the manuscript: HP, HZ. Agree with manuscript results and conclusions: HP, HZ. Jointly developed the structure and arguments for the paper: HP, HZ. Made critical revisions and approved final version: HP, HZ. Both authors reviewed and approved of the final manuscript.

## REFERENCES

1. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;34(3):267–73.
2. Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, Tanabe M. Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*. 2014;42(1):D199–205.
3. Wang X, Dalkic E, Wu M, Chan C. Gene module level analysis: identification to networks and dynamics. *Curr Opin Biotechnol*. 2008;19:482–91.

4. Bardou V-J, Arpino G, Elledge RM, Kent Osborne C, Clark GM. Progesterone receptor status significantly improves outcome prediction over estrogen receptor status alone for adjuvant endocrine therapy in two large breast cancer databases. *J Clin Oncol*. 2003;21(10):1973–9.
5. Colditz GA, Rosner BA, Chen WY, Holmes MD, Hankinson SE. Risk factors for breast cancer according to estrogen and progesterone receptor status. *J Natl Cancer Inst*. 2004;96(3):218–28.
6. Arpino G, Weiss H, Lee AV, et al. Estrogen receptor–positive, progesterone receptor–negative breast cancer: association with growth factor receptor expression and tamoxifen resistance. *J Natl Cancer Inst*. 2005;97(17):1254–61.
7. Goss PE, Ingle JN, Martino S, et al. Efficacy of letrozole extended adjuvant therapy according to estrogen receptor and progesterone receptor status of the primary tumor: National Cancer Institute of Canada Clinical Trials Group MA.17. *J Clin Oncol*. 2007;25(15):2006–11.
8. Yu K-D, Liu G-Y, Di G-H, et al. Progesterone receptor status provides predictive value for adjuvant endocrine therapy in older estrogen receptor-positive breast cancer patients. *Breast*. 2007;16(3):307–15.
9. Kyndi M, Sørensen FB, Knudsen H, Overgaard M, Nielsen HM, Overgaard J. Estrogen receptor, progesterone receptor, HER-2, and response to postmastectomy radiotherapy in high-risk breast cancer: the Danish Breast Cancer Cooperative Group. *J Clin Oncol*. 2008;26(9):1419–26.
10. Bagaria SP, Ray PS, Sim MS, et al. Personalizing breast cancer staging by the inclusion of ER, PR, and HER2. *JAMA Surg*. 2014;149(2):125–9.
11. Feeley LP, Mulligan AM, Pinnaduwage D, Bull SB, Andrulis IL. Distinguishing luminal breast cancer subtypes by Ki67, progesterone receptor or TP53 status provides prognostic information. *Mod Pathol*. 2014;27(4):554–61.
12. Purdie CA, Quinlan P, Jordan LB, et al. Progesterone receptor expression is an independent prognostic variable in early breast cancer: a population-based study. *Br J Cancer*. 2014;110(3):565–72.
13. Kahlenborn C, Modugno F, Potter DM, Severs WB. Oral contraceptive use as a risk factor for premenopausal breast cancer: a meta-analysis. *Mayo Clin Proc*. 2006;81(10):1290–1302.
14. Heiss G, Wallace R, Anderson GL, et al. Health risks and benefits 3 years after stopping randomized treatment with estrogen and progestin. *JAMA*. 2008; 299(9):1036–45.
15. Hannaford PC, Iversen L, Macfarlane TV, Elliott AM, Angus V, Lee AJ. Mortality among contraceptive pill users: cohort evidence from Royal College of General Practitioners' Oral Contraception Study. *BMJ*. 2010;340:c927.
16. Cui X, Schiff R, Arpino G, Osborne CK, Lee AV. Biology of progesterone receptor loss in breast cancer and its implications for endocrine therapy. *J Clin Oncol*. 2005;23(30):7721–35.
17. Pang H, Lin A, Holford M, et al. Pathway analysis using random forests classification and regression. *Bioinformatics*. 2006;22(16):2028–36.
18. Pang H, Datta D, Zhao H. Pathway analysis using random forests with bivariate node-split for survival outcomes. *Bioinformatics*. 2010;26(2):250–58.
19. Breiman. *Manual on Setting up, Using, and Understanding Random Forests V*4.0; 2003. Available at http://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf.
20. Becker KG, Hosack DA, Dennis G Jr, et al. PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics*. 2003;4:61.
21. Stendahl et al. (2006)
22. Dowsett M, Houghton J, Iden C, et al. Benefit from adjuvant tamoxifen therapy in primary breast cancer patients according oestrogen receptor, progesterone receptor, EGF receptor and HER2 status. *Ann Oncol*. 2006;17(5):818–26.
23. Punglia RS, Kuntz KM, Winer EP, Weeks JC, Burstein HJ. The impact of tumor progesterone receptor status on optimal adjuvant endocrine therapy for postmenopausal patients with early-stage breast cancer. *Cancer*. 2006;106(12):2576–82.
24. Daniel AR, Gaviglio AL, Knutson TP, et al. Progesterone receptor-B enhances estrogen responsiveness of breast cancer cells via scaffolding PELP1- and estrogen receptor-containing transcription complexes. *Oncogene*. January 27, 2014. doi:10.1038/onc.2013.579. [Epub ahead of print].
25. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13:2498–504.
26. Sato Y, Takahashi S, Kinouchi Y, et al. IL-10 deficiency leads to somatic mutations in a model of IBD. *Carcinogenesis*. 2006;27(5):1068–73.
27. Merlos-Suárez A, Batlle E. Eph-ephrin signalling in adult tissues and cancer. *Curr Opin Cell Biol*. 2008;20(2):194–200.
28. Zheng M, Bocangel D, Doneske B, et al. Human interleukin 24 (MDA-7/IL-24) protein kills breast cancer cells via the IL-20 receptor and is antagonized by IL–10. *Cancer Immunol Immunother*. 2007;56(2):205–15.
29. Subramanian S, Matejuk A, Zamora A, Vandenbark AA, Offner H. Oral feeding with ethinyl estradiol suppresses and treats experimental autoimmune encephalomyelitis in SJL mice and inhibits the recruitment of inflammatory cells into the central nervous system. *J Immunol*. 2003;170(3):1548–55.
30. Pang H, Zhao H. Building pathway clusters from Random Forests classification using class votes. *BMC Bioinformatics*. 2008;6(9):87.
31. Mao KZ, Tang W. Recursive Mahalanobis separability measure for gene subset selection. *IEEE/ACM Trans Comput Biol Bioinform*. 2011;8(1):266–72.
32. Pang H, George SL, Hui K, Tong T. Gene selection using iterative feature elimination random forests for survival outcomes. *IEEE/ACM Trans Comput Biol Bioinform*. 2012;9(5):1422–31.
33. Han X, Li Y, Huang J, et al. Identification of predictive pathways for non-hodgkin lymphoma prognosis. *Cancer Inform*. 2010;9:281–92.
34. Pang H, Hauser M, Minvielle S. Pathway-based identification of SNPs predictive of survival. *Eur J Hum Genet*. 2011;19:704–09.
35. Kim I, Pang H, Zhao H. Bayesian semiparametric regression models for evaluating pathway effects on continuous and binary clinical outcomes. *Stat Med*. 2012;31(15):1633–51.