Published in final edited form as:

Breast Cancer Manag. 2012 May 1; 1(1): 57–64. doi:10.2217/BMT.12.14.

Using large-scale molecular data sets to improve breast cancer treatment

Chad J Creighton*

*Department of Medicine & Dan L Duncan Cancer Center Division of Biostatistics, Baylor College of Medicine, 1 Baylor Plaza, BCM 600, Houston, TX 77030, USA

SUMMARY

The treatment of breast cancer patients could potentially be advanced by having a more complete understanding of breast cancer biology, including a catalog of recurrently altered genes. Over the last decade, thousands of human breast tumors have been profiled for gene expression and DNA copy number alterations, and ongoing efforts in DNA sequencing are establishing the set of somatically mutated genes. Much of the molecular data being generated resides in the public domain, available as a resource for further research. The challenge then becomes how to best utilize all these data, to mine them for candidate biomarkers and targets, which is the subject of this review. Some examples of integrative analysis and combining results from diverse data sets are also presented.

To date, breast cancer research has yielded two major success stories in the use of targeted therapy – namely, blocking the estrogen receptor (ER) pathway in ER-positive cancers and blocking the HER2 pathway in HER2-positive breast cancers. These therapies have contributed to a steady decline in breast cancer death rates, being credited with improved survivorship for hundreds of thousands of women [1,2]. For all the progress that has been made, however, more is left to be done. In 2011, approximately 39,520 breast cancer deaths were expected to occur among women in the USA alone [3]. Currently, breast cancer patients with ER-positive or progesterone receptor (PR)-positive tumors (who can benefit from anti-estrogen therapy) have a better prognosis as compared with ER⁺/PR⁺ patients, with a 5-year overall survival of approximately 83 versus 69%, respectively [4]; however, for both patient groups, there remains a substantial number who are not cured. In particular, for the 15% of patients with triple-negative (i.e., ER[−]/PR[−]/HER2[−]) tumors, there is currently no targeted therapy available, which contributes to the worse prognosis for this group [5].

To improve treatment of breast cancer patients, more research is needed, both to optimize the use of existing therapies and to develop new therapies. A deeper understanding of the molecular biology of breast cancer ought to contribute substantially towards these ends. Since approximately the year 2000, molecular profiling of breast cancers has been carried out in earnest by the research community, beginning with mRNA-expression profiling using DNA microarrays [6]. This has led to the present-day comprehensive efforts of large consortiums, such as The Cancer Genome Atlas (TCGA) and International Cancer Genomics Consortium (ICGC), which are profiling cancers at multiple levels including

© 2012 Future Medicine Ltd

creightc@bcm.edu.

Financial & competing interests disclosure The author has no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

mRNA, miRNA, DNA copy, DNA methylation and DNA mutation. All of this work has resulted in an incredible amount of molecular data being continually deposited into the public domain, which may be repeatedly mined for additional insights.

This review will examine the potential for using large-scale molecular data sets as a resource towards improving breast cancer treatment. We will focus here on analysis and mining of molecular profile data on human breast tumors and cell lines. Recent developments in personalized oncology through high-throughput sequencing of the individual patient's tumor [7,8] are a worthy topic for examination elsewhere but are outside the present scope. This review will also feature a number of examples of 'integrative analysis' using public molecular data sets, whereby results from two or more different profile data sets are combined in order to yield new information.

Clinical biomarker discovery

In the clinical setting, biomarkers are needed that accurately predict clinical outcomes and response to therapy. In cancer treatment, there are two basic types of biomarkers: prognostic biomarkers, which inform on the patient's overall cancer outcome, regardless of therapy; and predictive biomarkers, which inform on the potential effect of a therapeutic intervention [9]. For breast cancer, ER and HER2 are used to predict responses to antiestrogen and anti-HER2 therapy, respectively. ER expression (as well as PR expression) is considered an independent prognostic factor in addition to being a therapeutic predictor. On the other hand, HER2 was initially considered a prognostic factor; however, subsequent successful targeting of the HER2 pathway has largely neutralized the biomarker's prognostic ability [9].

For breast cancer, molecular profiling efforts have yielded a number of additional prognostic biomarker candidates, which may consist of individual genes, as well as gene 'signatures' (i.e., entire sets of genes) [10–13]. Many of these prognostic gene signatures have been examined together and found to perform similarly overall [11], and so it is unclear whether any far superior prognostic classifier remains to be discovered. As a result of profiling studies, several gene-expression-based reference laboratory tests, the most well known of these being Oncotype DX® (Genomic Health, CA, USA) and MammaPrint™ (Agendia, The Netherlands), are now available for prognostication of patients diagnosed with breast cancer. Both assays summarize the expression values of a panel of genes (Oncotype DX: 16 cancer-related genes; MammaPrint: 70 genes) into a single value to assign a risk index to the patient sample (the index being prognostic for the likelihood of disease recurrence). In addition to its role as a prognostic test, Oncotype DX is currently being evaluated as a predictive test for chemotherapy response in women with node-positive disease [14], whereby patients with a low enough predicted risk of recurrence might avoid chemotherapy altogether.

While prognostic gene markers and signatures have been extensively mined from molecular data sets, more work remains to be done in identifying useful predictive markers, which would have the potential to guide patient treatment. Towards this end, additional molecular-profiling data sets may be needed over those currently available, ideally including both untreated and treated patients, along with robust measures of response (including long follow-up times); much work is required to generate such data sets. One expression-profile data set from Loi *et al.* includes clinical follow-up for 439 ER⁺ tumors from patients receiving tamoxifen and 172 ER⁺ tumors from untreated patients [15]. The Loi *et al.* data set might be a useful starting point for defining candidate predictive markers of tamoxifen response (e.g., defining genes that are prognostic for the tamoxifen-treated cohort but not for the untreated cohort). Such an approach, however, would have its limitations, including difficulty in distinguishing therapeutic predictive patterns from prognostic patterns, as well as having to meaningfully interpret a negative result of no significant pattern in the untreated

group (which may simply be a function of, for example, statistical power). In order to be useful in the clinical setting, any predictive biomarker candidates obtained from profiling data would first need to be extensively validated in independent patient cohorts.

Besides data from human tumors, molecular data on breast cancer cell lines can be mined to identify potential predictive gene markers. For a host of drugs, it may be difficult to readily obtain treatment data from human patients, whereas cell lines are much more accessible for a variety of testing. In a recent study by Heiser *et al.*, gene-expression profiles on 45 breast cancer cell lines were examined, for each of which the 50% growth inhibition concentrations (GI₅₀) were assessed for 77 different compounds [16,101]. These data are public and would allow for identifying gene correlates of the GI₅₀ for a particular drug (e.g., genes that are high when GI₅₀ is low might represent drug-sensitivity markers). As an example of the types of integrative analysis that may use these data, for this review, each of the cell line profiles from the Heiser *et al.* data were scored for similarities to gene-transcription signatures of either HER2 or EGFR activation *in vitro*, the data for which were generated and examined elsewhere [17,18]. In addition to individual genes, a pathway-based signature could represent a potential biomarker, whereby cells that show more pathway activity may be more sensitive to pathway inhibition. Consistent with this hypothesis, inferred pathway activities for HER2 and EGFR correlated with response to lapatinib (a HER2 inhibitor) and erlotinib (an EGFR inhibitor), respectively (Figure 1). While the Heiser *et al.* data sets are a valuable resource, one limitation is the relatively low number of cases (e.g., tamoxifen GI₅₀ tends to correlate with a predefined estrogen-pathway signature, but with insufficient numbers for reaching statistical significance [data not shown]).

Identifying new therapeutic targets

Further progress in treating cancer could be made by identifying additional molecular targets, against which therapies not currently in use could be applied, particularly for triple-negative breast cancers, as well as the subset of ER⁺ cancers that are unresponsive to hormone therapy. There are conceivably a number of ways one might define a therapeutic target candidate using existing molecular data. A desirable candidate gene or associated pathway would be potentially druggable (e.g., involving a cell receptor), and would be something for which cancer cells and not normal cells have a unique sensitivity towards inhibition. Molecular data on human cancers can therefore be mined for genes that show evidence of dysregulation in cancer. Such genes may be aberrantly expressed at the mRNA level, or may show genetic or genomic targeting by somatic mutation or copy number alteration (at least in a subset of breast cancers). The premise that is entertained here is that there are druggable targets, in addition to ER and HER2, that remain to be found. A number of growth factor receptors, including IGF-1R and EGFR, are already under investigation [19–21], although it remains to be seen just how many more potential targets remain that are either unknown or understudied.

The search for new therapeutic targets may be aided by comprehensive molecular profiling of breast cancer for all types of molecular aberrations. Initial efforts to profile breast cancer at the gene-expression level have been useful in defining clinically relevant molecular subtypes of the disease [6,22], although most differentially expressed genes do not represent critical disease drivers. Profiling cancers for somatic mutation and copy alterations, in addition to expression, can uncover genetic or genomic events that natural selection has retained as a benefit to the cancer. TCGA consortium is one initiative currently engaged in profiling human breast cancers at multiple levels, including mRNA (by gene-expression arrays and by RNA sequencing), protein (by reverse-phase protein array assay), miRNA (by small RNA sequencing), gene-promoter methylation (by methylation arrays), DNA copy (by single-nucleotide polymorphism [SNP] arrays), and somatic mutation (by whole-exome

sequencing), all for the same set of at least 1000 cases [23,24]. As of January 2012, data were available on at least one platform for over 800 cases. The data are made publicly available upon generation, most of it residing at the TCGA data portal [102].

Efforts by the TCGA and other initiatives will soon yield a near-complete catalog of molecular aberrations in breast cancer, which will provide opportunities for various integrative analyses that put together results from the various data sets in meaningful ways, providing a more complete molecular picture of the disease. For one thing, we can readily identify which genes are influenced by copy alteration or promoter methylation. As an exercise, for this review, TCGA data on 185 human breast cancers were analyzed, for which mRNA expression, copy and promoter methylation were assayed for over 11,000 genes, all on the same samples. By computing the pairwise correlation between gene mRNA and gene copy, copy alterations were found to have a widespread effect on gene expression, as expected [25], with 46% of genes showing significant positive correlations ($p < 0.001$ on two-sided Spearman's correlation); positive correlations included cases where a gene had copy gain, leading to increased expression, and cases where a gene had copy loss, leading to decreased expression (Figure 2A). On the other hand, there was a subset of genes (19% at $p < 0.001$ on two-sided Spearman's correlation) for which increased gene-promoter methylation correlated with a decrease in expression (Figure 2B), which would be expected in cases of gene silencing.

While integrative analysis can help filter for alterations that have an observable impact on gene expression, it may be difficult in many cases to distinguish driver alteration events from passenger events [26,27], particularly with copy profiling. Work is currently underway by TCGA analysis groups and others to correlate molecular aberrations with the histological parameters of breast tumors, including histological type, histological grade, lymph node status and immunohistochemical study of ER, PR and HER2. While correlation patterns do not themselves establish cause-and-effect, having a catalog of such correlations can be valuable for research purposes.

Establishing the relevance of experimental models to human disease

As much as we currently know about breast cancer and its biology, much more remains to be learned, including potential new ways to target breast cancer cells. Over recent decades, basic research has greatly contributed to our understanding of how molecular pathways are altered in cancer, including the ER, HER2 and PI3K/Akt/mTOR pathways. While much of the present day emphasis on translational research (i.e., research leading to a foreseeable impact in the clinic) is certainly warranted, an initial investment in basic research can potentially reap huge rewards later on, although patience may be necessary.

In basic research studies, one challenge is to demonstrate how experimental results, obtained using some model system, are pathologically relevant to breast cancer as it presents itself in human patients. The limitations of cancer cell line cultures as a model system have been debated at length by the research community, and the case can be made that cancer cells growing on plastic differ in many ways from the cancer cells in a patient, one notable difference being the influence of the tissue microenvironment on tumor growth and progression [28,29]. As required, researchers may overcome the limitations of 2D cell cultures by using more complicated model systems, including mouse models and 3D cultures. Nevertheless, any model, no matter how complicated, is still a model, and therefore how it relates to human cancer may be questioned. However, for all of their perceived drawbacks, experimental models allow us to elucidate the roles of specific genes and pathways, in ways that would not be feasible with human patient data. In general, breast cancer cell lines have been found to carry the same types of genomic aberrations found in

human tumors [30], and cell line cultures are arguably suitable for many studies that do not concern the influence of the microenvironment.

Molecular data on human tumors, obtained from large-scale profiling data sets, can complement data from experimental models. Experimental data reveal cause-and-effect relationships, although their relevance to human disease is not always clear, while data from human tumors reveal correlative (although not causal) relationships, which would seem relevant to human disease. For example, if perturbing a gene or pathway *in vitro* shows a measurable effect on cancer cell behavior, the human data sets can also be examined for evidence of the gene or pathway being a target of somatic mutation or copy number alteration, or whether expression or activation of the gene or pathway is associated with patient outcome. For another example, if one gene is found to functionally influence another gene, the human data sets can be examined so as to determine whether the genes appear correlated with each other across tumors.

Studies using gene-expression profiling have been able to demonstrate concordance between results from experimental models and results from human tumors [17,18,31,32]. For example, genes upregulated in cell cultures by overexpression of cyclin-D1 tend to be positively correlated with *CCND1* in human tumors [32]. This type of concordance has also been demonstrated for multiple oncogenes in human prostate tumors [17]. To illustrate how these types of analysis can be applied to breast cancer, for this review, the author examined multiple sets of genes – each of which were found to be upregulated by a specific oncogene – against a large compendium of human breast tumor-expression profiles (gene signatures obtained from [17] and tumor data previously compiled for [33], and available at [103,104]). For each oncogene considered, all 12,722 genes profiled in the human tumors were ranked by correlation with the oncogene (from high to low correlation). If we consider the positions of the oncogene-induced genes (experimental) along the oncogene-correlated (human tumor) ranked list, we find that many of these genes tend to be positively rather than negatively correlated (Figure 3). These patterns are particularly statistically strong (as well as visually apparent) for *MYC*, *ERBB2*, *EGFR* and *AKT1* genes; however, no such patterns are evident for *CTNNB1* or *CCND1*. This latter finding could perhaps be due to insufficient data or an experimental model system not relevant to breast cancer.

Analyses such as those presented in Figure 3 can help the investigator focus on the subset genes that appear most relevant in the human tumor setting, out of all the genes affected in the experimental setting. Furthermore, concerns regarding the limitations of profiling a single cell line model, for example, may be lessened if the transcriptional program observed experimentally appears manifested in a sizable number of human tumor specimens [18]. When translating experimental results to human tumors, one should keep in mind that human cancer represents a highly complicated system, one in which many different genes may be altered in any given sample, unlike an experimental model in which we are able to perturb a single factor in a controlled setting. For many experimental findings, the ideal human tumor molecular data set may not be available (e.g., results of testing a drug on cell lines, a drug that has never been tested in patients).

Conclusion & future perspective

Numerous large-scale molecular data sets on human breast cancers and cell lines are currently available in the public domain, and these provide a rich resource for the discovery of new clinical biomarkers and therapeutic targets, and for helping establish the relevance of experimental models to human disease. With these data, and with a clear hypothesis and access to bioinformatics expertise, one can readily generate candidate genes of interest for subsequent validation and study, although these subsequent steps can take substantial time

and effort. To maximize the potential of molecular profiling, molecular biologists, pathologists and oncologists need greater access to them. User-friendly software tools can make it easier for researchers (who may not have extensive computational skills) to ask a limited set of analysis questions; however, the trade-off in this case is invariably between usability and flexibility. By virtue of their computer and analysis skills, bioinformatics experts have greater flexibility in how they can access and query the data, although biologists, pathologists and oncologists often know which specific questions to ask. Through greater collaboration between researchers of all disciplines, great advances might be made in improving breast cancer treatment.

While molecular profiling has yielded new classifications of breast cancer relevant to our understanding of the biology of the disease [6,34], it remains to be seen how large-scale data sets can contribute to the development of new therapies having a substantial impact on breast cancer patients. To date, there is no compelling evidence that an approach to mining molecular data sets for novel targets will be fruitful, and the entire endeavor might therefore be considered somewhat speculative. Thus far, we have ER and HER2 as success stories in targeting; however, neither of these was discovered through high-throughput approaches. Even with these vast data resources at our disposal, finding new cures for breast cancer will be most challenging and will involve much preclinical and clinical work.

Acknowledgments

This work was supported in part by NIH/NCI grants P30 CA125123, P50 CA58183 and U24CA143843.

References

Papers of special note have been highlighted as:

■ of interest

■ ■ of considerable interest

1. Jordan V. Tamoxifen: catalyst for the change to targeted therapy. *Eur. J. Cancer.* 2008; 44(1):30–38. [PubMed: 18068350]
2. Baselga J. Treatment of HER2-overexpressing breast cancer. *Ann. Oncol.* 2010; 21(Suppl. 7):vii36–vii40. [PubMed: 20943641]
3. Desantis C, Siegel R, Bandi P, Jemal A. Breast cancer statistics, 2011. *CA Cancer J. Clin.* 2011; 61(6):409–418. [PubMed: 21969133]
4. Grann V, Troxel A, Zojwalla N, Jacobson J, Hershman D, Neugut A. Hormone receptor status and survival in a population-based cohort of patients with breast carcinoma. *Cancer.* 2005; 103(11): 2241–2251. [PubMed: 15844176]
5. Foulkes W, Smith I, Reis-Filho J. Triplenegative breast cancer. *N. Engl. J. Med.* 2010; 363(20): 1938–1948. [PubMed: 21067385]
6. Perou CM, Sorlie T, Eisen MB, et al. Molecular portraits of human breast tumours. *Nature.* 2000; 406(6797):747–752. [PubMed: 10963602] ■ ■ First gene-expression profiling study to classify human breast cancer using cDNA microarrays.
7. Roychowdhury S, Iyer M, Robinson D, et al. Personalized oncology through integrative high-throughput sequencing: a pilot study. *Sci. Transl. Med.* 2011; 3(111):111ra121.
8. Jones S, Laskin J, Li Y, et al. Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. *Genome Biol.* 2010; 11(8):R82. [PubMed: 20696054]
9. Oldenhuis C, Oosting S, Gietema J, de Vries E. Prognostic versus predictive value of biomarkers in oncology. *Eur. J. Cancer.* 2008; 44(7):946–953. [PubMed: 18396036]
10. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* 2002; 347(25):1999–2009. [PubMed: 12490681] ■

Demonstration of a prognostic gene signature for human breast cancer, which is the basis for the now commercially available MammaPrint™ test.

11. Fan C, Oh D, Wessels L, et al. Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med.* 2006; 355(6):560–569. [PubMed: 16899776]
12. Wang Y, Klijn JG, Zhang Y, et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet.* 2005; 365:671–679. [PubMed: 15721472]
13. Paik S, Shak S, Tang G, et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* 2004; 351(27):2817–2826. [PubMed: 15591335] ■
Demonstration of the prognostic ability of the Oncotype DX® multigene assay, developed in part using gene-expression array data.
14. Kim C, Paik S. Gene-expression-based prognostic assays for breast cancer. *Nat. Rev. Clin. Oncol.* 2010; 7(6):340–347. [PubMed: 20440284]
15. Loi S, Haibe-Kains B, Desmedt C, et al. Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *J. Clin. Oncol.* 2007; 25(10):1239–1246. [PubMed: 17401012]
16. Heiser L, Sadanandam A, Kuo W, et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. *Proc. Natl Acad. Sci. USA.* 2011; 109(8):2724–2729. [PubMed: 22003129] ■ Good data resource of gene-expression profiles on breast cancer cell lines, with corresponding growth-inhibition data for 77 drug compounds.
17. Creighton C. Multiple oncogenic pathway signatures show coordinate expression patterns in human prostate tumors. *PLoS ONE.* 2008; 3(3):E1816. [PubMed: 18350153]
18. Creighton CJ, Hilger AM, Murthy S, Rae JM, Chinnaiyan AM, El-Ashry D. Activation of mitogen-activated protein kinase in estrogen receptor alpha-positive breast cancer cells in vitro induces an *in vivo* molecular phenotype of estrogen receptor alpha-negative human breast tumors. *Cancer Res.* 2006; 66(7):3903–3911. [PubMed: 16585219]
19. 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):7721–7735. [PubMed: 16234531]
20. Litzenburger B, Creighton C, Tsimelzon A, et al. High IGF-IR activity in triple-negative breast cancer cell lines and tumorgrafts correlates with sensitivity to anti-IGF-IR therapy. *Clin. Cancer Res.* 2011; 17(8):2314–2327. [PubMed: 21177763]
21. Osborne C, Neven P, Dirix L, et al. Gefitinib or placebo in combination with tamoxifen in patients with hormone receptor-positive metastatic breast cancer: a randomized Phase II study. *Clin. Cancer Res.* 2011; 17(5):1147–1159. [PubMed: 21220480]
22. Sorlie T, Tibshirani R, Parker J, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl Acad. Sci. USA.* 2003; 100(14):8418–8423. [PubMed: 12829800]
23. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature.* 2008; 455(7216):1061–1068. [PubMed: 18772890]
24. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature.* 2011; 474(7353):609–615. [PubMed: 21720365] ■■ References [23] and [24] are comprehensive genomic studies of glioblastoma and serous ovarian carcinoma from The Cancer Genome Atlas. The Cancer Genome Atlas is currently performing molecular profiling of over 20 different cancer types, including breast cancer. As a public resource, these data will have a tremendous influence within the research community.
25. Pollack J, Sørli T, Perou C, et al. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl Acad. Sci. USA.* 2002; 99(20):12963–12968. [PubMed: 12297621]
26. Greenman C, Stephens P, Smith R, et al. Patterns of somatic mutation in human cancer genomes. *Nature.* 2007; 446(7132):153–158. [PubMed: 17344846]
27. Liu E. Functional genomics of cancer. *Curr. Opin. Genet. Dev.* 2008; 18(3):251–256. [PubMed: 18691651]
28. Hanahan D, Weinberg R. The hallmarks of cancer. *Cell.* 2000; 100(1):57–70. [PubMed: 10647931]

29. Tuxhorn J, Ayala G, Rowley D. Reactive stroma in prostate cancer progression. *J. Urol.* 2001; 166(6):2472–2483. [PubMed: 11696814]
30. Neve RM, Chin K, Fridlyand J, et al. A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell.* 2006; 10(6):515–527. [PubMed: 17157791]
31. Sweet-Cordero A, Mukherjee S, Subramanian A, et al. An oncogenic *KRAS2* expression signature identified by cross-species gene-expression analysis. *Nat. Genet.* 2005; 37:48–55. [PubMed: 15608639]
32. Lamb J, Ramaswamy S, Ford HL, et al. A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. *Cell.* 2003; 114(3):323–334. [PubMed: 12914697] ■ One of the first demonstrations of an overall concordance of gene-expression profiles between cell lines and human tumors, in this case regarding putative transcriptional targets of cyclin D1.
33. Kessler J, Kahle K, Sun T, et al. A SUMOylation-dependent transcriptional subprogram is required for *Myc*-driven tumorigenesis. *Science.* 2012; 335(6066):348–353. [PubMed: 22157079]
34. Prat A, Perou C. Deconstructing the molecular portraits of breast cancer. *Mol. Oncol.* 2011; 5(1): 5–23. [PubMed: 21147047]

Websites

101. Stand Up to Cancer (SU2C)/University of California, Santa Cruz Cancer Genome Browser. <https://genome-cancer.soe.ucsc.edu>
102. Data Portal for The Cancer Genome Atlas (TCGA). <http://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp>
103. Gene Expression Omnibus. www.ncbi.nlm.nih.gov/geo
104. ArrayExpress, TABM-158 (breast cancer data set). www.ebi.ac.uk/arrayexpress

Practice Points

- Efforts are underway to comprehensively profile human breast tumors and cell lines at various molecular levels, including gene expression, somatic mutation, copy number alterations and methylation.
- Large-scale molecular data sets are available in the public domain and may be reanalyzed with new questions in mind.
- Large-scale molecular data sets may be mined to uncover candidate gene markers with either prognostic or therapeutic predictive ability.
- Large-scale molecular data sets may be mined to uncover additional candidates for therapeutic targeting strategies, although it remains to be seen whether this type of approach will bear fruit.
- Using public molecular data sets, genes arising from experimental models may be examined in human tumors, in order to help establish the pathological relevance of these genes in human disease.
- Bioinformatics experts, with the ability to access and query large-scale molecular data sets, should collaborate with oncologists, pathologists and molecular biologists in generating candidate gene lists of interest for further validation and study.

Figure 1. In breast cancer cell lines, gene-transcription signatures for HER2 and EGFR pathways correlate with response to lapatinib (HER2 inhibitor) and erlotinib (EGFR inhibitor), respectively

(A) The correlation between inferred HER2 pathway activity score (on the basis of gene-expression profiling) and response to lapatinib ($n = 37$ breast cancer cell lines; $p = 0.01$ by two-sided Pearson's correlation) and (B) between EGFR score and response to erlotinib ($n = 42$; $p = 0.0005$). Breast cancer cell line data were obtained from [16,101]. The gene signature scoring method was described in [24]. *HER2* and *EGFR* gene-transcription signatures were obtained from [17]. A high- $\log_{10}(\text{GI}_{50})$ indicates that a lower drug concentration is needed to inhibit cell growth. GI_{50} : 50% growth inhibition concentration.

Figure 2. In human breast tumors, gene-expression patterns are widely influenced by both copy number alteration and methylation of promoter regions

Data are from The Cancer Genome Atlas (TCGA) [102], comprising 185 breast tumors profiled for gene expression, gene copy and gene-promoter methylation (data accessed February 2011; 11,750 genes shared between all three profiling platforms). Plots show Spearman's correlations between gene expression and copy (A) and between gene expression and gene-promoter methylation (B). The order of genes is the same for both (A) and (B). (C) Positions of genes residing in the 17q genomic region (which includes the *HER2* gene). Correlation t-statistic of ± 3 approximately corresponds to $p = 0.001$.

Figure 3. Genes upregulated by a given oncogene in the experimental setting (using cell lines or mouse models) are often coexpressed with the same oncogene in human breast tumors Nine different oncogenes were considered, each of which was previously manipulated in an experimental model, in order to determine the set of genes being upregulated by the oncogene (gene signatures obtained from [17]). For each oncogene, a set of 1340 human breast tumor-expression profiles ([103,104], previously compiled for [33]) was examined and all 12,722 genes profiled were ranked according to Pearson's correlation with the oncogene (left: high correlation with the oncogene; right: low correlation/high anticorrelation). Dashed lines indicate the positions within each (tumor-based) ranked list of the genes in the corresponding experimental signature. For a given oncogene, clustering of the lines towards the left end of the plot is indicative of significant overlap between the experimental gene set and the top gene correlates within the tumors. R-values and p-values were obtained by one-sided Pearson's correlation. NS: Not significant.