Published in final edited form as:
Clin Cancer Res. 2013 June 15; 19(12): 3114–3120. doi:10.1158/1078-0432.CCR-12-2093.

Molecular Pathways: Extracting Medical Knowledge from High Throughput Genomic Data

Theodore Goldstein[1,2], **Evan O. Paull**[1,2], **Matthew J. Ellis**[3,4,*], and **Joshua M. Stuart**[1,2,*]

[1]Department of Biomolecular Engineering, University of California, Santa Cruz, CA, USA

[2]Center for Biomolecular Science and Engineering, Santa Cruz, CA, USA

[3]Department of Internal Medicine, Division of Oncology, Section of Breast Oncology, Washington University, St. Louis, MO, USA

[4]Siteman Cancer Center Breast Cancer Program, Washington University, St. Louis, MO, USA

Abstract

High-throughput genomic data that measures RNA expression, DNA copy number, mutation status and protein levels provide us with insights into the molecular pathway structure of cancer. Genomic lesions (amplifications, deletions, mutations) and epigenetic modifications disrupt biochemical cellular pathways. While the number of possible lesions is vast, different genomic alterations may result in concordant expression and pathway activities, producing common tumor subtypes that share similar phenotypic outcomes.

How can these data be translated into medical knowledge that provides prognostic and predictive information? First generation mRNA expression signatures such as Genomic Health's Oncotype DX already provide prognostic information, but do not provide therapeutic guidance beyond the current standard of care – which is often inadequate in high-risk patients. Rather than building molecular signatures based on gene expression levels, evidence is growing that signatures based on higher-level quantities such as from genetic pathways may provide important prognostic and diagnostic cues. We provide examples of how activities for molecular entities can be predicted from pathway analysis and how the composite of all such activities, referred to here as the "activitome," help connect genomic events to clinical factors in order to predict the drivers of poor outcome.

Background

Tumor subtypes define clinically relevant and molecularly recognizable classifications of cancer

Cancers manifest in different subtypes defined by a set of characteristic attributes such as mutations, cell lineage markers, and histology. Classifying tumors into clinically relevant subtypes is a major step in identifying therapeutic strategies. The distinctions between subtypes may reflect differences in the originating cells transformed by oncogenesis. For example, luminal breast cancers are often more differentiated than basal breast tumors and have a higher proportion of estrogen receptor expression. Subtype distinctions may also reflect different etiologies at work in similar cells due to the nature of the genomic damage.

[*]Corresponding authors: *Matthew J. Ellis MB. BChir, Ph.D. FRCP., Washington, University School of Medicine, Department of Medicine, Division of Oncology, Section of Breast Oncology, Siteman Cancer Center, Breast Cancer Program, Campus Box 8056, 660 South Euclid Ave, St Louis MO 63110, Tel 314 747 7502, Fax 314 747 9320: *Josh Stuart Ph.D., Biomolecular Engineering, 1156 High Street, Mail Stop: SOE2UC Santa Cruz Santa Cruz, CA 95064, Tel 831 459 1344, Fax 831 459 4829.

Dr M.J. Ellis reports patents, licensing and income from the PAM50 through Bioclassifier LLC and University Genomics

For example, colorectal tumors can exhibit a global DNA methylation phenotype thought to silence DNA repair genes such as MLH1, which then leads to an associated higher background mutation rate compared to other colorectal cancer subtypes. Tumors respond variably to small molecule inhibition and the differences in drug sensitivity between subtypes persist even when the tumors are transformed into cell line models [1]. New high-throughput technologies will aid in the characterization and recognition of established and novel subtypes to better tailor therapeutics.

Genome-wide expression levels, organized as mathematical vectors of statistically differential gene levels, can be used as *signatures* of tumor subtypes. Signatures allow the detection of correlations between tumor characteristics such as the possibility that two different mutations may affect the same cellular wiring or that a particular mutation is associated with a clinical outcome. Signatures based only on gene expression may overlook signals from other *cis-* and *trans*-regulatory logic. Thus, we seek to find a comprehensive cellular pathway activity description we call the *activitome*. Just as the *genome* is the comprehensive description of a cell's genetic information, the activitome is a comprehensive description of a cell's functional and dysfunctional activity based on expression, methylation, copy number, and other high-throughput assay technologies. Here we give a set of examples of data-driven approaches for predicting patient therapy using signatures based on the activitome inferred from global pathway analysis.

Inferring the *activitome* using global pathway analysis

Increases in computational power and the availability of comprehensive genetic networks make possible a systematic pathway analysis of tumor cells. Rather than focusing on one or a few known pathways, developments in probabilistic graphical models allow *all known pathways of a cell* to be computationally represented and used for multiplatform data analysis. We developed an integrated pathway approach called PARADIGM [2]. In this framework, each type of omics measurement is mapped to a graphical model based on the central dogma of molecular biology (DNA is transcribed to RNA which is translated into amino-acids and hence proteins, and that protein may exist in passive and active forms). We enrich the model with the knowledge that proteins and RNA may regulate DNA. PARADIGM uses a merged set of constituent pathways from various databases called the SuperPathway. PARADIGM then infers the maximum likelihood integrated pathway level (IPL) of pathway elements including genes, proteins and protein complexes. The algorithm currently incorporates four types of high-throughput gene-level data: mRNA expression levels (including microarray and RNA-Seq), genomic copy number measures, epigenetic methylation data, and protein level data (such as from the new reverse phase protein arrays, or from mass-spectroscopy approaches).

Figure 1 illustrates how gene activities can be inferred for a “small toy” pathway, i.e. a pared-down model, simpler than reality. The PARADIGM graphical model centered on a particular gene is shown in detail in Figure 1A. Multiple different data measurements of a tumor sample are connected into a graphical model as observed variables (shaded ellipses). Unobserved states of gene expression and activity are connected into the graph as hidden variables (open ellipses). A computationally intensive method called *Bayesian belief propagation* is then used on the underlying factor graph to set the internal probabilities of the graphical model to a configuration that has a high likelihood according to the observed data[3].

The toy pathway in Figure 1B shows a small pathway involving a single kinase, PAK2 that post-translationally inhibits the MYC/MAX complex. The transcription factor complex MYC/MAX in turn activates CCNB1 and ENO1 and represses WNT5A.Figure 1C illustrates how belief propagation could set the internal “active state” of PAK2 based on the downstream evidence. In this example, it finds that the PAK2 kinase is inactivated using the evidence downstream that suggests that MYC/MAX, a complex that is inhibited by PAK2, is active because one of its known activated targets (CCNB1) is highly expressed while one of its repressed targets (WNT5A) has lower expression. Note that even though ENO1 is an activated target of the complex and it is not highly expressed, the model can *explain away* this apparent discrepancy using the information that ENO1's promoter seems to be epigenetically silenced and therefore the lower expression of ENO1 does not reflect on a lower activity of the MYC/MAX transcription factor complex. The set of all inferred quantities of gene-encoded features in a complex wiring diagram (right hand side of Figure 1C), which together form a quantitative state description of tumor cells that we refer to here as the *activitome*. The integrated measure of the activitome not only represents expression states of genes but multimeric protein complexes, gene family roles, and higher-level cellular processes, which encapsulates both the molecular function and the information transmission aspect of genes and proteins.

Activitome signatures reveal the signaling layer that interlink genomic perturabations and transcriptional changes that characterize tumor subtypes

If activitome signatures provide an accurate view of tumor cell circuitry then one can ask if they explain why observed genomic lesions are associated with concomitant transcriptional changes. For example, basal breast cancers are often associated with TP53 mutations and are also characterized by major transcriptional “hubs” involving transcription factors such as FOXM1, MYC/MAX, and HIF1A. What is the regulatory logic that leads to alterations in these major programs as a result of loss of TP53? If we can determine the network that resolves this question then such a network may provide an accurate representation of the cellular wiring of the tumor that can be used as a surrogate to identify potential targets.

One approach is to identify “linking genes” that connect observed genomic aberrations, such as copy number gains/losses or mutations, to observed expression-based activities such as the up-regulation of transcription factor hubs would be to adapt previously published heat-diffusion approaches such as HotNet [4]. As an extension of the heat-diffusion approach, a linking set of genes also can be identified by applying multiple input gene sets to identify nodes that interconnect a set of “sources” (genomic perturbations) to a distinct set of “targets” (transcription factors). One can then find essential paths that resolve the effect of genomic alterations with phenotypic changes in the tumor state. Sub-networks can then be identified that interconnect protein level activitome data to gene expression level data using protein-protein interactions, predicted transcription factor to target connections, and curated interactions from literature. Permutation-based analysis can then be used to gauge the significance of the solutions.

As an example of this linking diffusion approach, the method was applied to the Cancer Genome Atlas (TCGA) breast cancer dataset, which included patient tumor/matched-normal samples for 533 patients, each with genomic sequencing data and microarray expression. To find the significant pathway differences between Luminal and Basal cancer subtypes, we performed a differential analysis between 99 Basal and 235 Luminal A samples. As a noise-filtering step, we used MEMo [5] calls to get 117 genes with significant numbers of amplification, deletion, methylation and mutation events. We then used a chi-square proportions test to find the genomic perturbations that occur with significantly different frequency between tumor subtypes. Significance Analysis of Microarrays [6] was used to compare the differential expression between Basal and Luminal-A tumor subtypes. A test network that included curated transcriptional, protein-level and complex interactions for nearly 5,000 genes and abstract concepts, with roughly 100,000 interactions was used for the analysis. The interlinking diffusion approach was run using the 12 differentially occurring genomic perturbations as the first “source” set, and the 370 differentially transcribed transcription factors as the second or “target” set. 1000 random permutations of the upstream sources were conducted while keeping the downstream set fixed to assess the significance of the networks. A sub-network connecting 11 of the 12 genomic perturbations to 336 of the 370 differentially expressed transcription factors, with 57 intermediate “linker” nodes and 5238 network edges was found to be the most significant (Figure 2).

Clinical-Translational Advances

Without translational applicability, inferring gene activities with pathway knowledge would be no more than an academic exercise. We describe how the inferences encoded in the activitome can be used to predict patient outcomes.

Activitome-derived features predict patient outcomes more reliably than gene expression

Specific pathway components in an activitome signature can reveal important aspects of tumor biology. For example, PARADIGM uncovered the FOXM1 transcription factor network in ovarian serous cancers, published in *Nature* with the TCGA marker paper [7]. This indicated previously unknown crosstalk between proliferation and DNA damage repair regulated by distinct isoforms of the FOXM1 transcription factor. This crosstalk may explain in part why these tumor cells proliferate in response to DNA repair signals.

Evidence suggests activitomes may improve our ability to predict patient outocomes. When PARADIGM was applied to the TCGA glioblastoma multiforme (GBM) dataset higher accuracy predictors of overall survival in the patients could be obtained compared to those using gene expression signatures [2]. This strongly suggests that the pathway-level information provides biologically relevant clues about the intrinsic state of tumor cells. Thus, using pathway-inferred levels to build activitome signatures shows promise for predicting biomedical outcomes.

Activitome signatures provide clues about cellular targets

Gene expression signatures have been used to identify putative drug targets. Some examples include the Connectivity-Map project pursued by the Golub lab at the Broad Institute, the Ailun project by the Butte lab at Stanford [8], and the Disease Diagnostic Gene Expression database by the Zhou lab at USC [9]. In the same way, activitome-based signatures can be used to build predictive signatures. Activitomes provide potentially much richer information because the pathway interactions can reveal cryptic signals such as active transcription factors and signaling molecules that could go unnoticed by looking at gene expression alone.

The merit of using pathway-based signatures for prediction was tested in a proof-of-concept demonstration in cell lines[1]. In this case, gene expression and copy number data on 50 breast cancer cell lines, half of which were of the basal (more aggressive) subtype and the other half were of the luminal (less aggressive) subtype. PARADIGM was run on all of the *in vitro* data and produced inferred pathway activity levels. A two-class (dichotomized) Significance Analysis of Microarrays test[6] was used to produce an association score for every feature in the SuperPathway. In this application, positive association scores reflect higher activity in basal tumors while negative associations reflect higher activity in luminal tumors. An *activitome signature* contrasting basal from luminal tumors was then constructed as the vector of all association scores across the entire SuperPathway.

The activitome signature and SuperPathway together were used to identify significantly large sub-networks that connect high-scoring pathway components to “druggable” biomarkers. Sub-networks were created by retaining any interaction that connected two features both of which had absolute association scores higher than the average absolute association. Among the largest of the hubs in the resulting network were a central DNA damage hub with the second highest connectivity (55 regulatory interactions; 1% of the network) and TP53 with the 14th highest connectivity (26 connections; 0.5% of the network) The sub-network identified several pathways of interest including the FOXM1-related network. Several genes upstream of FOXM1 are known targets of available drugs, including PLK3, suggesting polo-kinase inhibitors may disrupt basal tumors. Indeed polo-kinase inhibitors on the cell lines were found to sensitize basal cells to a higher degree compared to luminal cells, consistent with the prediction encoded in the network.

Comparing activitome signatures reveals novel connections between mutations and drug response in luminal breast cancers

Activitome signatures can be used to connect mutations, clinical outcoms, and other “events” present in tumor samples. It is often of interest to know whether a particular mutation is associated with elevated risk or the possibility of developing resistance to a particular treatment option. Molecular signatures derived from such events can be used as proxies to predict such tendencies. As an example, pathway-based activitome signatures were used to analyze a set of patient tumors of the Luminal breast cancer subtype (both Luminal A and Luminal B subtypes) [10]. In this study, clinical and genomic data on samples were assessed from a neoadjuvant aromatase inhibitor (AI) clinical trial designed to assess the responsiveness of samples to these estrogen-lowering agents [11]. PARADIGM was used to build a predictive model for AI therapy and to develop links between gene mutations and clinical outcomes. PARADIGM analysis revealed that multiple pathways are affected by a phalanx of mutations including caspase/apoptosis, ErbB signalling, Akt/PI3K/mTORsignalling, TP53/RB signalling and MAPK/JNK pathways. Several “hubs” such as ESR1 and FOXA1 were activated cohort-wide while other hubs exhibited high but differential changes in aromatase-inhibitor-resistant tumors including MYC, FOXM1 and MYB.

A method called Differential Pathway Signature Correlation (DiPSC) was developed for this study to compare signatures while accounting for the confounding that stems from sample overlap. Mutations in different genes may cause disruptions in the same pathway, which may lead to similar disruptions in the activitome. By comparing the vectors of activitome signatures of different mutations and clinical outcomes intrinsic connections between these events may be uncovered. DiPSC randomly splits the patient cohort in half. In each half, two different activitome signatures are calculated from two distinct contrasts. A contrast corresponds to the dichotomy defined by the presence versus the absence of a particular “event,” such as a mutation or a clinical outcome. The event is used as a dichotomous variable in a to two-sample SAM analysis to derive an activitome signature. The activitome signatures computed from each disjoint half are compared to one another. This guarantees that the comparison of the signatures is not polluted by any overlapping samples. The procedure of randomly splitting the cohort, re-deriving the activitome signatures with SAM, and comparing the signatures is repeated 1000 times. The final correlation is then computed as a mean and standard deviation across the random trials.

DiPSC was applied to the 77 luminal samples using the PARADIGM-derived activitome signatures to uncover phenomena that underlie the resistance of some cancers to aromatase inhibitors. All pairs of associations were scored across all of the cohorts. An example of the association of mutations to subtype is illustrated in the DiPSC (*dipstick)* shown in Figure 3, which plots the correlation of all activitome signatures against the Luminal B vs Luminal A activitome signature. From this visualization one can immediately see what patient groups lead to common signatures. The analysis revealed for example that mutated MALAT1 (a small non-coding RNA) had activitome signatures similar to TP53 mutations and are also associated with both high Ki-67 and high preoperative endocrine prognostic index (PEPI) scores that are indicators of resistance to drug treatment. Ki-67 is a prognostic indicator of proliferation in breast cancers. Because MALAT1 is mutated in only a handful of samples, this precluded several analyses used to detect such relations because of the low samples size whereas DiPSC was able to leverage the robustness of the pathway signatures to find a significantly indicative pattern. On the other hand, PIK3CA, MLL3 and CDH1 do not enrich for either Luminal subtype. ATR, and MAP2K4 are slightly enriched for Luminal A and MAP3K1 mutations are overwhelmingly enriched for Luminal A. Thus, relating cancer outcomes to those based on pathway-inferred signatures teased out novel connections not available to standard approaches.

Conclusion: toward patient-specific models

Clearly activitome informatics tool development for clinical care is in an early stage of development. The data-driven discovery approach requires pooling data from many patients and data sources to build functional inferences. The challenge ahead of us is to develop single sample predictors that can guide therapeutic decisions in individual cases. One can imagine building a database of activitome signatures to represent all known cancer subtypes. Each will span the range of possible genomic, epigenomic, and transcriptomic activities that characterize all samples of a particular subtype. To identify a patient-specific model then would require two conceptual steps: 1) From the database of subtype signatures identify the most representative for a particular patient sample; and 2) Refine the model to best fit the particular set of genomic, epigenomic, and proteomic changes observed in the patient's data. The approach leverages on the statistical power of multiple samples to define the starting subtype models but also encompasses the flexibility to adapt to a particular form of the disease. Just as in gene expression-based models, activitome-based models will require the careful acquisition of samples from well conducted clinical trials that are sufficiently powered for the full suite of genomic and proteomic analysis pipeline executed at the clinical grade testing level. While we are years from clinical utility, the pathway-based approaches we describe provide the basis for a discussion on progress towards this goal and underscore the value of the deeply collaborative environment provided by our rapidly growing bioinformatics and computational biology discipline and many teams of clinicians and genome and proteome centers that provide us with data to analyze.

Acknowledgments

Dr. MJ Ellis is supported by NIH grants U24 CA160035, R01 CA095614, P30 CA91842, A Susan G Komen for the Cure Promise Grant and a BCRP-Idea Award-BC112014. He is also supported by the Breast Cancer Research Fund. Dr. JM Stuart is supported by an NSF CAREER Award and co-leads a genome data analysis center supported by the Cancer Genome Atlas (TCGA) project.

References

1. Heiser, LM.; Sadanandam, A.; Kuo, WL.; Benz, SC.; Goldstein, TC.; Ng, S., et al. Subtype and pathway specific responses to anticancer compounds in breast cancer. Proceedings of the National Academy of Sciences of the United States of America; 2011;
2. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010; 26:i237–245. [PubMed: 20529912]
3. Kschischang, Frey, Loeliger. Factor graphs and the sum-product algorithm. IEEE Transactions on Information Theory. 2001:47.
4. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. Journal of computational biology: a journal of computational molecular cell biology. 2011; 18:507–522. [PubMed: 21385051]
5. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome research. 2012; 22:398–406. [PubMed: 21908773]

6. Tusher, VG.; Tibshirani, R.; Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America; 2001; p. 5116-5121.

7. Network CGA. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474:609–615. [PubMed: 21720365]

8. Chen R, Li L, Butte AJ. AILUN: reannotating gene expression data automatically. Nature methods. 2007; 4:879. [PubMed: 17971777]

9. Huang, H.; Liu, CC.; Zhou, XJ. Bayesian approach to transforming public gene expression repositories into disease diagnosis databases. Proceedings of the National Academy of Sciences of the United States of America; 2010; p. 6823-6828.

10. Ellis MJ, Ding L, Shen D, Luo J, Suman VJ, Wallis JW, et al. Whole-genome analysis informs breast cancer response to aromatase inhibition. Nature. 2012; 486:353–360. [PubMed: 22722193]

11. Ellis MJ, Suman VJ, Hoog J, Lin L, Snider J, Prat A, et al. Randomized phase II neoadjuvant comparison between letrozole, anastrozole, and exemestane for postmenopausal women with estrogen receptor-rich stage 2 to 3 breast cancer: clinical and biomarker outcomes and predictive value of the baseline PAM50-based intrinsic subtype--ACOSOG Z1031. Journal of clinical oncology: official journal of the American Society of Clinical Oncology. 2011; 29:2342–2349. [PubMed: 21555689]

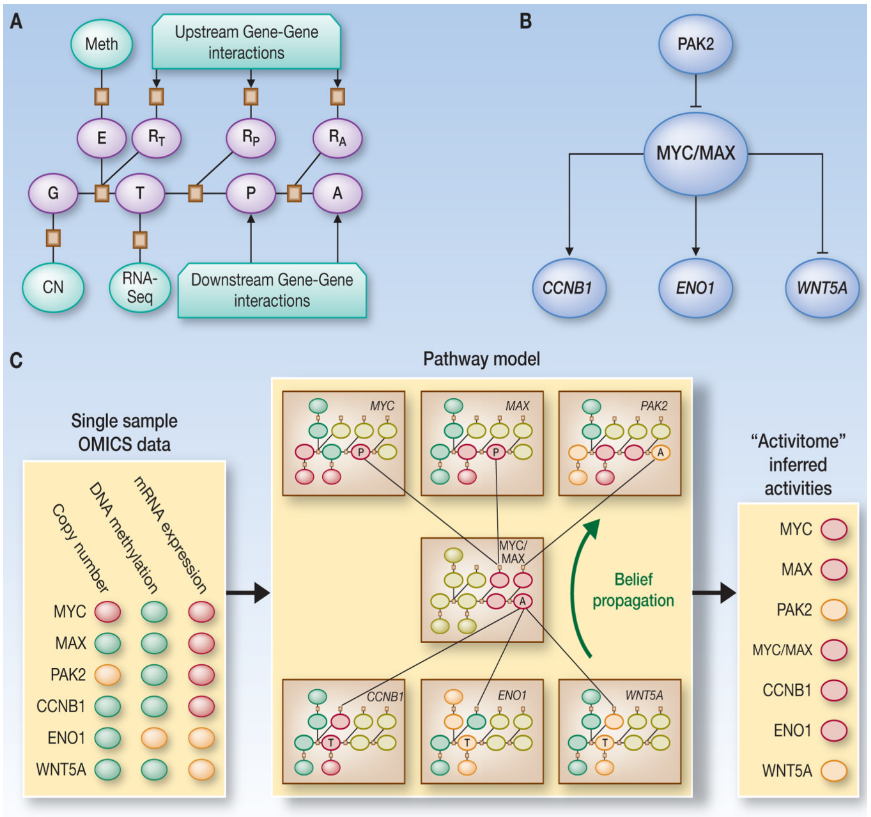

Figure 1. PARADIGM model for integrative data analysis

A. Factor graph model oriented around a single gene Hidden states in a tumor sample (open ellipses) for genomic copies (G), epigenetic promoter state (E), mRNA transcripts (T), peptide (P), and active protein (A). Regulation gene expression (open ellipses) include transcriptional (R_T), translational (R_P), and post-translational (R_A) control. Sample data (filled circles, constrain gene states through *factors* (boxes). **B. Toy example of a MYC/ MAX-associated pathway**. Two transcription factors (MYC and MAX) form a complex (MYC/MAX) that is inhibited by PAK2, a protein kinase. MYC/MAX activates two target genes (CCNB1, ENO1) and inactivates a third (WNT5A). **C. Single patient data converted to inferred activities for toy pathway**. Measurements and inferred levels are either higher (red), lower (blue), or comparable (purple) to levels in matched normal. Belief propagation infers the kinase is inactive based on inferred higher activity of MYC/MAX.

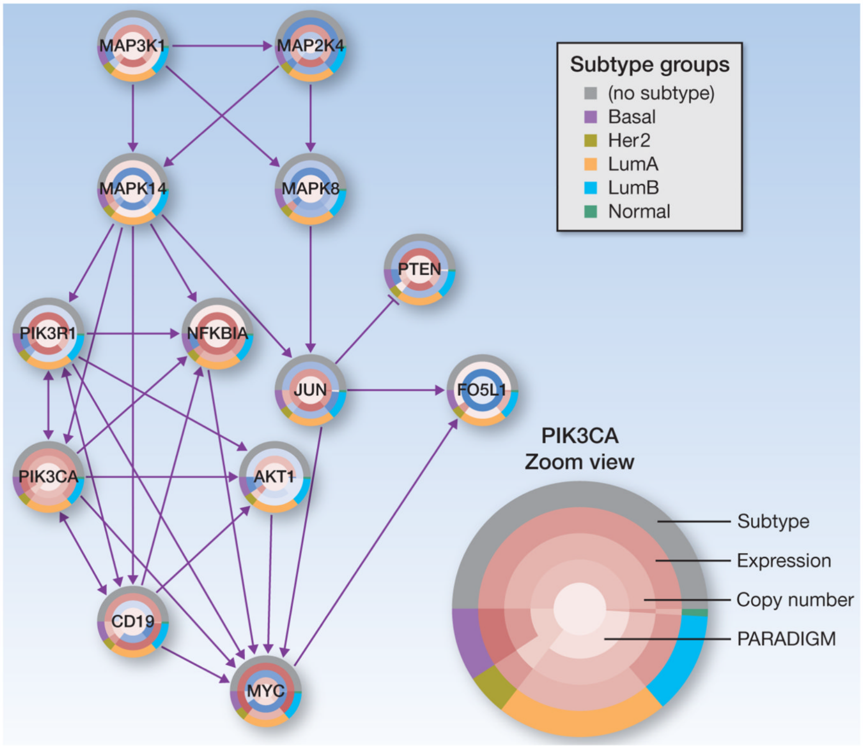

Figure 2. HotLink Result for TCGA Breast Cancer

Rings of data depict different measurements about genes or proteins as higher activity (red) or lower activity (blue) compared to normal controls. Pathway illustrates part of the HotLink solution inferred from the TCGA basal and luminal breast tumors with various data available through TCGA including (inner to outer): pathway levels inferred by PARADIGM, Copy number alterations, RNA-Seq RSEM levels, and RPPA data. The outermost ring depicts the patient subtypes. Segments display aggregated levels for samples within each grouping defined by the breast cancer subtype (indicated in the outermost ring).

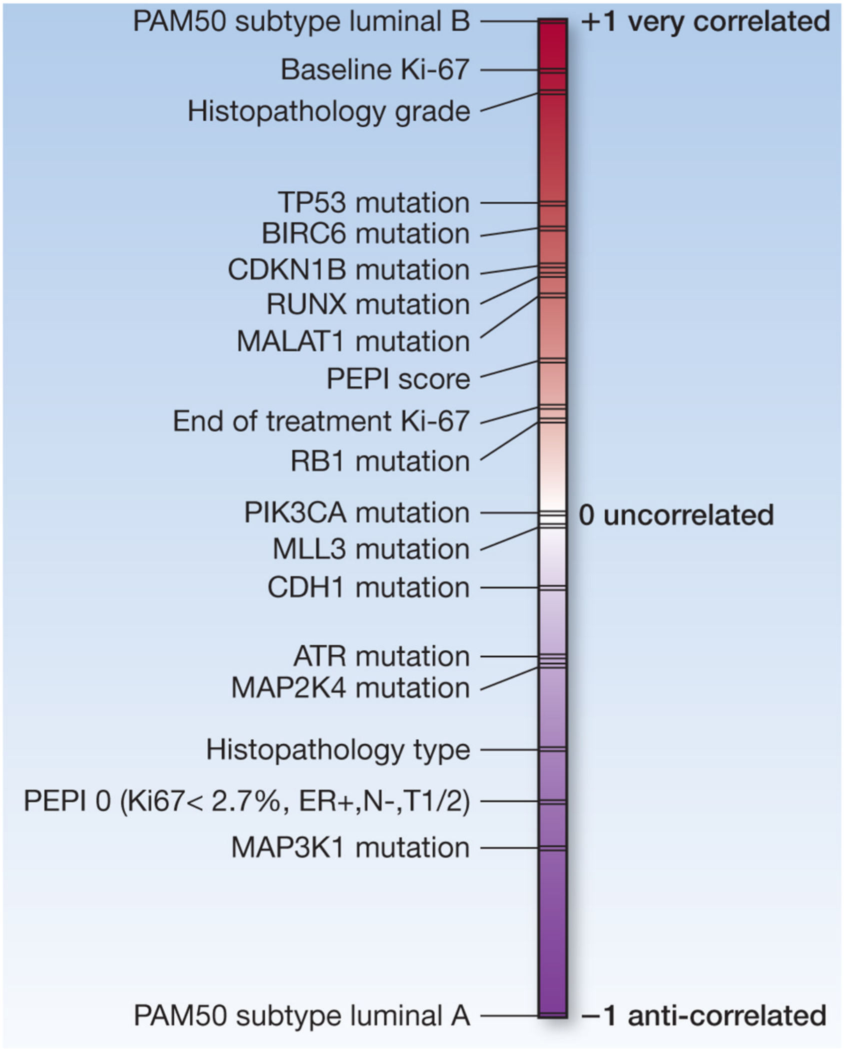

Figure 3.

DiPSC(Dipstick) depicts correlations comparing mutations, and biomarkers along the Luminal-A/Luminal B dichotomy.