Literature mining

Marina Diachenko

April 2019

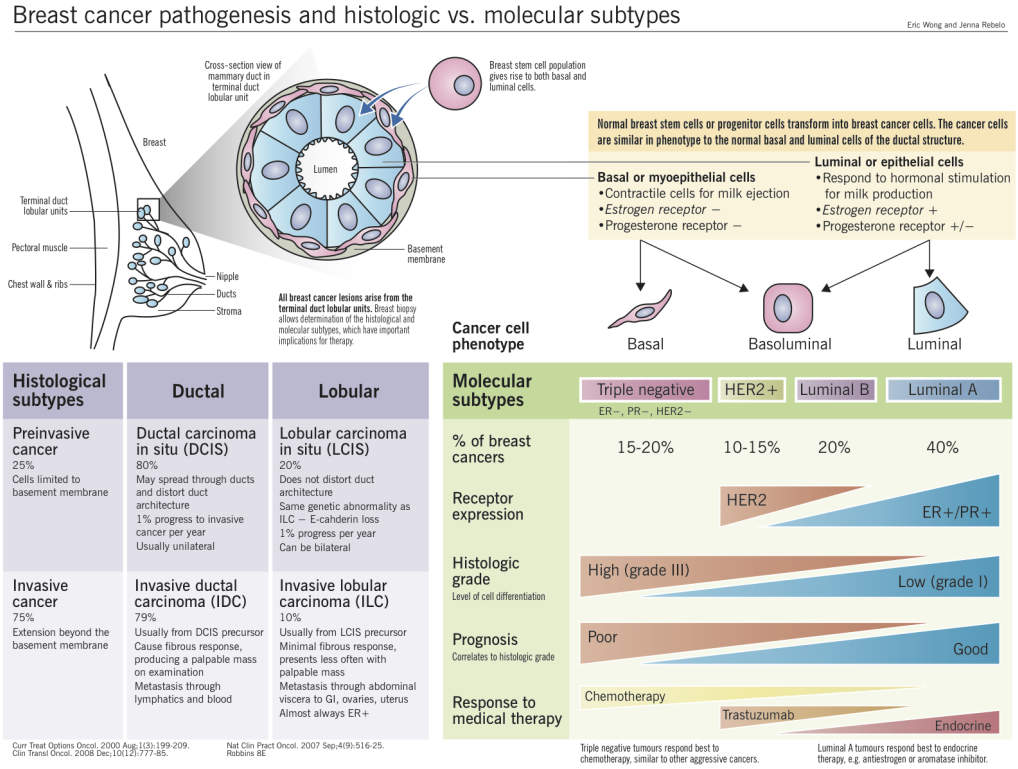

Figure 1: Breast cancer subtype overview.

1. **Identification of Novel Breast Cancer Subtype-Specific Biomarkers by Integrating Genomics Analysis of DNA Copy Number Aberrations and miRNA-mRNA Dual Expression Profiling (2015)**

   -They used an integrative approach. CNA + expression data

   - Tumors within gene expression subtypes share similar DNA copy number aberrations (CNA) which can be used to further subdivide subtypes.
   - For the copy number data in the first dataset, we used a sparse Bayesian learning (SBL) model and backward elimination (BE) procedure to find candidate CNA regions. A SBL model was used to find the most likely candidate break-points for the copy number state, and the BE procedure was used to remove sequentially the least significant breakpoints estimated by the SBL model, allowing a flexible adjustment of the false discovery rate (FDR). We used

package of R software (http://www.r-project.org/) to implement this analysis.

- *Determining Subtype-Specific CNAs*. To determine subtype-specific CNAs, the segment output file which had arisen from the SBL model and BE procedure was converted into an indicator matrix, where, for each sample, each genes copy state was represented as 1 = loss, 0 = no change, and 1 = gain. The counts of state for luminal-A subtype and basal-like subtype were compared to identify subtype-specific CNAs. We performed a Chi-square test on the subtype for each gene. Genes with P below 0.05 were selected as the subtype-specific genes.

- Generally, the copy number alteration at gene promoters typically does not alter the coding sequences of genes but contributes to cancer by influencing gene expression [21]. The genes residing in CNA regions are expected to cause the corresponding expression changes. Therefore, genes amplified or deleted as well as overexpressed or underexpressed in a subtype-specific manner are good candidate genes.

- **Results**:

  In regard to the luminal-A subtype, the total counts of gains and losses were 3,650 and 3,544, respectively. For the basal-like subtype, the total counts of gains and losses were 4,024 and 3,980, respectively. In particular, the highest counts of CNA regions for a luminal-A subtype sample and a basal-like subtype sample were 663 and 547, respectively (see Figure S2). Moreover, the copy number changes detected high frequency regions (more than 40 % across the breast cancer subtype samples) included 8 regions (2 gain regions and 6 loss regions) for the luminal-A subtype and 18 regions (6 gain regions and 12 loss regions) for the basal-like subtype, respectively (see Figure 2).

	Gains			Losses	
Subtype	Chromosome	Frequency (%)	Subtype	Chromosome	Frequency (%)
Luminal-A	5p15.3-q11.1	46.20	Luminal-A	1p36.3-31.2	55.80
Luminal-A	16p13.3-13.1	61.50	Luminal-A	11p15.5-15.4	42.30
Basal-like	2p25.3-24.2	42.50	Luminal-A	17p13.3-11.2	59.60
Basal-like	6p25.3-21.2	52.50	Luminal-A	18p11.3-q12.1	40.40
Basal-like	9p24.3-22.1	47.50	Luminal-A	19p13.3-13.2	42.30
Basal-like	10p15.3-11.1	62.50	Luminal-A	22q11.1-13.1	50.00
Basal-like	12p13.3-13.1	47.50	Basal-like	1p36.3-35.3	55.00
Basal-like	21q11.1-21.3	60.00	Basal-like	3p26.3-25.2	42.50
			Basal-like	4p16.3-15.3	50.00
			Basal-like	7p22.3-21.3	45.00
			Basal-like	8p23.3-23.1	67.50
			Basal-like	11p15.5-15.2	57.50
			Basal-like	13q11-14.3	57.50
			Basal-like	14q11.2-q13.1	42.50
			Basal-like	16p13-11.2	45.00
			Basal-like	17p13.3-11.2	67.50
			Basal-like	19p13.3-13.1	65.00
			Basal-like	22q11.1-12.1	50.00

Figure 2: Results from CNA (aCGH).

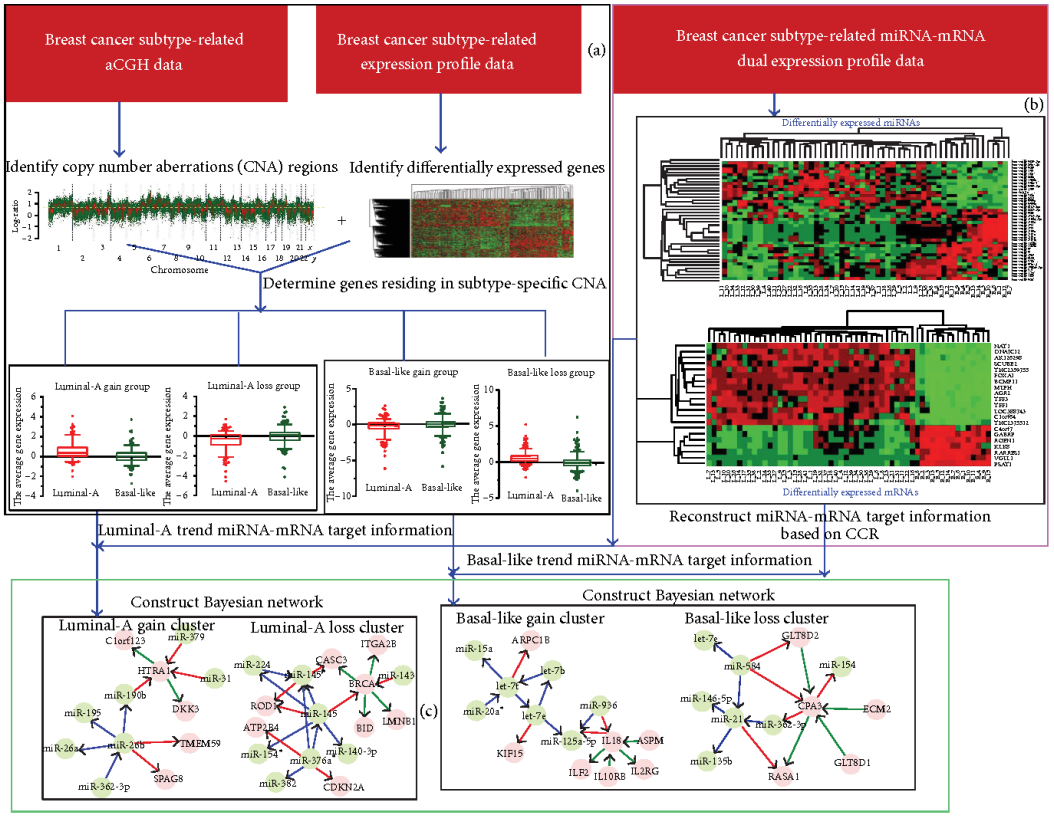

Figure 3: The workflow presented in the paper.

2. **Novel Data Mining Techniques in aCGH based Breast Cancer Subtypes Profiling: the Biological Perspective (2007).**

   Here only 2 classes were included

   - This is a comparative study among well established data mining algorithm (namely J48 and Nave Bayes Tree) and novel machine learning paradigms like **Ant Miner and Gene Expression Programming**. The aim of this study was to discover significant rules discriminating ER+ and ER- cases of breast cancer.
   - More on aCGH technique:

     -Results of aCGH screening are in the form of microarray images (Fig. 1); spot intensities are evaluated as ratios of fluorescent tag concentration and corresponding values are associated to specific probe copy number. Bacterial Artificial Chromosome (BACs) have been commonly used as probes in order to observe copy number changes of regions of the genome that share the same relative copy number on average.

     -It is important to note that although the biological entity (copy number) is intrinsically discrete, the signal under investigation is considered as being continuous; this inconsistency is due to the fact that quantification of copy number levels is based on fluorescence measurement that is of an analogue source.
   - Here we present a new experimental pipeline that takes advantage of some of the most promising among emergent methods for rule induction (namely Ant Colony Optimization Ant Miner -and Gene Expression Programming -GEP) establishing a comparative study with well known data mining algorithms (namely J48 and Nave Bayes Tree). The choice of rule induction algorithm for bioinformatics and biomedical data mining will be explained in one of the next paragraphs.
   - We built a consensus criterion for predictive power estimation of each of the BACs sod the reduced input set using three main methods: Students T-test, Receiver Operating Curve (ROC) and entropy (Kullback-Lieber divergence).
   - In this study we considered a cohort of 124 patients with breast cancer at different stages.

   CLASSIFIERS:

- **J48 Classifier** J48 classifier forms rules from pruned partial decision trees built using C4.5s heuristics. C4.5 is Quinlans most recent non-commercial tree-building algorithm. The main goal of this scheme is to minimize the number of tree levels and tree nodes, thereby maximizing data generalization. It uses a measure taken from information theory to help with the attribute selection process. For any choice point in the tree, it selects the attribute that splits the data so as to show the largest mount of gain in information. The J48 classifier described above builds a C4.5 decision tree. Each run of J48 it creates an instance of this class by allocating memory for building and storing a decision tree classifier. The algorithm, the classifier it builds, and a procedure for outputting the classifier, are all part of that instantiation of the J48 class. J48 class does not actually contain any code for building a decision tree. It includes references to instances of other classes that do most of the work. It also combines the divide-and-conquer strategy for decision tree and separate divide-and-conquer one for rule learning. Such approach adds flexibility and speed.
- **Nave Bayes Tree** is a hybrid between decision trees and Nave Bayes. This algorithm creates trees whose leaves are Nave Bayes classifiers for instances that reach the leaf. When constructing the tree, cross-validation is used to decide whether the node should be split further or a Nave Bayes model should be used instead. The algorithm is similar to the classical recursive partitioning schemes, except that the leaf nodes created are Nave Bayes categorizers instead of nodes predicting a single class. A threshold for continuous attributes is chosen using the standard entropy minimization technique, as is done for decision-trees. The utility of a node is computed by discretising the data and computing the 5-fold cross- validation accuracy estimate of using Nave-Bayes at the node. The utility of a split is the weighted sum of the utility of the nodes, where the weight given to a node is proportional to the number of instances that go down to that node. Intuitively the algorithm tries to approximate whether the generalization accuracy for Nave-Bayes classifier at each leaf is higher than the single Nave-Bayes classifier at the current node. To avoid splits with little value, we define a split to be significant if the relative (not absolute) reduction in error is greater than 5 % and there are at least 30 instances in the node.
- **Ant Miner** Ant based algorithms or ant colony optimization

(ACO) have been applied successfully to combinatorial optimization problems. More recently Parpinelli and colleagues have applied ACO to data mining classification problems, where they introduced a classification algorithm called Ant Miner. The goal of Ant miner is to extract classification rules from data [REF: Parpinelli 2002, 12] this is accomplished by leaving agents (ant) exploring the space of attributes looking for best combination of antecedents that predict a given class. An overview of the Ant Miner algorithm is given in Figure 4. Ant Colony Algorithms have been recently used in classification problems in bioinformatics by Chan and Freitas in [A New Ant Colony Algorithm for Multi-Label Classification with Applications in Bioinformatics (2006).].

TS = all training cases;
WHILE (No. of cases in TS > max_uncovered_cases)
i=0;
REPEAT
i=i+1; Ant_i incrementally constructs a rule;
Prune the just constructed rule;
Update the pheromone of the trail followed by Ant_I;
UNTIL $(i \geq No\_of\_Ants)$ or (Ant_i constructed the same rule as the previous No_Rules_Converg-1 Ants)
Select the best rule among all constructed rules;
Remove the cases correctly covered by the selected rule from the training set;
END

Fig. 4 Pseudocode of the Ant Miner algorithm in Parpinelli's implementation.

Figure 4: Ant Miner algorithm.

-**Feature set**:

- Feature set can be greatly optimized, eliminating redundancy of

co-regulated genes, for example, or considering subsets of genes that minimize inter correlation. Many different approaches are documented in literature; the most recent contribute to this field of optimal feature set finding comes from [14]. Many relevant suggestions can be found in this work in particular about the covariance structure of data and its impact on the optimality of feature set. Other feasible approaches include sensitivity analysis by removing attributes, proportion correct use in rules, ratio of features Between- category to Within-category sums of squares, Signal-to-Noise scores in Onve-versus-Rest or One-versus-All fashion, Kruskall-Wallis non parametric test (ANOVA) and number of appearances in models [A. Stanikov et al. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis (2004).]

- However the scientific community seems to agree that the optimal feature set simply doesnt exist but, instead, it should be measured on the single classification approach and, in general, on the single experiment [15]. For this reason we developed a consensus scheme for attribute selection that takes advantage of three well established statistical methods, they are: Students T-Test, Receiver Operating Characteristic and Entropy (Kullback-Lieber divergence). T-Test checks for mean of a distribution and allow to establish a comparison of diversity between two populations through mean comparison; this test returns a value that can be easily translated in the probability that the sets of data are drawn from the same distribution or from different distribution with the same means. In ROC the Area Under The Curve is estimated as an indicator of class separation; the more separated the classes, the higher the AUC. Kullback-Lieber divergence, instead, is a principle drawn for information theory that accounts for inner information contained in each attribute being a good metrics for more expressive attributes selection. All of these techniques can be used to compile a ranking of the features that accounts for the power of a single attribute to discriminate between the output classes.
- Results:

  All the systems showed good performances, however J48, Ant Miner and Gene Expression Programming algorithm were characterized by comparable and higher accuracy levels when compared with Nave Bayes Tree approach. Multiple runs of the non deterministic algorithms were carried out in order to overcome the

variable nature of the results returned by these approaches.

3. **Array comparative genomic hybridization and its applications in cancer (2005).**

   - DNA copy number affects gene expression and function
   - Array comparative genomic hybridization has proven its value for analyzing DNA copy-number variations.
   - The major technical challenge of array CGH is the generation of hybridization signals that are sufficiently intense, specific and quantitative that copy-number changes can be detected
   - signal intensity on an array element is affected by a number of factors, including base composition, proportion of repetitive sequence content and amount of hybridizable DNA in the array element. Intensities may vary by a factor of 30 or more among array elements even if there are no copy-number changes. If the entire hybridization and measurement process is well behaved (i.e., the signals are linearly proportional to sequence abundance), then the comparative hybridization strategy provides ratios that are quantitatively proportional to relative copy number. Most importantly, production variability among different arrays, such as the amount of DNA in array elements or element morphology, is accurately compensated.
   - Identification of important genes in regions of copy-number change is a complex task. If narrow regions of highly elevated copy number or total deletion that contain previously known cancer genes or genes with suggestive function are found, high-probability candidates may be immediately evident.
   - In many cases, however, even minimally defined aberrant regions resulting from combining data from many specimens may contain several attractive candidates (or none), or the copy-number aberrations may be complex, making it difficult to determine how many different loci may be under selection. If a gain is greater than a single copy, it is possible that more than one evolutionary step was involved in its formation. This sometimes results in a profile that resembles a peak with sloped sides, suggesting, but not proving, that the critical genes are located near the center of the peak 24,54 . Therefore, it is sometimes useful to interpret the amplitude of copy-number changes in addition to noting their locations.

- Evaluation of genes in regions of copy-number losses is also complex. In some cases, the decrease in expression caused by deletion of a single copy of a gene contributes to tumor development. But in the classic case of tumor-suppressor genes, function is totally abrogated by deletion of all copies of a gene, deletion of one copy and mutation or epigenetic alteration of the other 73 , or alteration of one copy and replacement of the other by a duplicate of the altered copy. Deletions of all copies of a genomic region are easily detectable in cell lines by array CGH and other techniques 74 , but their reliable detection in tumor specimens is complicated by the likely inclusion of normal cells. The finding of focal homozygous deletions in regions of frequent heterozygous deletion or loss of heterozygosity can provide crucial information to focus searches for important genes. Aberrations that result in loss of heterozygosity but no copy-number change are not detectable by array CGH.

4. **Distinct Patterns of DNA Copy Number Alteration Are Associated with Different Clinicopathological Features and Gene-Expression Subtypes of Breast Cancer (2006).**

   -From previous studies:

   - Genomic DNA copy number alterations (CNAs) also provide potentially useful molecular markers for breast cancer prognostication or prediction of treatment response. Frequently observed CNAs include gain of chromosomal regions 1q, 8q, 17q, and 20q, and loss of 1p, 8p, 13q, and 17p. (Knuutila et al., 2000). Sites of localized high-level DNA amplification harboring known oncogenes include 7p12 (EGFR), 8q24 (MYC), 11q13 (CCND1), 12q14 (MDM2), 17q12 (ERBB2), 20q12 (AIB1), and 20q13 (ZNF217) [(Al-Kuraya et al., 2004), and references therein]. Deletions with known tumor suppressor genes (TSGs) include 13q12 (BRCA2), 17p13 (TP53), and 17q21 (BRCA1). Cytogenetic studies have identified gains on 8q, 17q12, and 20q13 to be associated with poor overall survival (Isola et al., 1995; Tanner et al., 1995; Ross and Fletcher, 1998). DNA amplification of ERBB2 at 17q12 also predicts response to trastuzumab and high-dose anthracyclines. Since genomic DNA is more stable than mRNA, and since CNAs define key genetic events driving tumorigenesis, such genomic alterations are potentially advantageous as prognostic/predictive factors.

-this study:

- Map positions for arrayed cDNA clones were assigned using the NCBI genome assembly, accessed through the UCSC genome browser database (NCBI Build 35)
- Overall frequencies of gain/loss varied among breast tumors with different clinicopathological features (Table below). In particular, gains/losses were more frequent (borderline-significant) in ER-negative tumors (P 14 0.06, Students t test), and high-level DNA amplifications were more common (strong trend) in high-grade (P 14 0.08) and TP53-mutant (P 14 0.13) tumors.

Table 1. Average Total CNAs for Clinicopathological Parameters

	Grade		ER		*TP53*		Subtypes			
	Low	High	Pos	Neg	WT	Mut	Lum A	Lum B	ERBB2	Basal-like
Gain	40	48	40	59	41	45	46	45	31 [a]	62 [b]
Loss	54	60	52	80	54	59	59	46	39 [a]	89 [b]
Gain/loss	94	108	92	140	95	103	105	91	70 [a]	152 [b]
Amplification	13	18	14	18	13	17	10	24 [c, d]	8	13

a $P < 0.05$ (vs. Lum-A or Basal-like).

b $P < 0.05$ (vs. ERBB2).

c $P < 0.001$ (vs. Lum-A or ERBB2).

d $P < 0.05$ (vs. Basal-like).

Figure 5

- To identify associations between specific CNAs and pathological parameters, we used the SAM method (Tusher et al., 2001), which corrects for multiple hypothesis (loci) testing in determining statistical significance. Since CNAs are known to often span cytobands, the finding of two or more adjacent cytobands associated with a particular clinicopathological parameter (emphasized in the results below) further increased our confidence in the results (as being biologically sensible).

- We identified several CNAs associated with tumor grade, including loss at 3p14, 4q31-q35, and 5q13-q23 in high-grade tumors. We also found CNAs associated with ER status, where ER-negative tumors exhibited more frequent loss at 5q11-q35 and 12q14-23, and gain at 6p21-p25 and 7p12. Additionally, we identified loci associated with TP53 mutation status, including gain at 1q21-q32 with wild-type TP53 and loss at 5q14-q23 with mutant TP53.
- To determine whether different gene-expression subtypes were associated with distinct CNAs, we used the two-class SAM method (i.e. one subtype versus all others). Significant associations between cytobands and clinicopathological parameters were identified using the Significance Analysis of Microarrays (SAM) method (Tusher et al., 2001), which is based on a modified t-statistic (for two-class comparisons) or Cox score (for survival analysis), and uses random permutations of class labels to estimate a false discovery rate (FDR). KaplanMeier survival analysis was performed using WinSTAT (R. Finch software).

5. **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation (2002).**

    -

TO BE CONTINUED

References

- Dongguo Li, Hong Xia, Zhen-ya Li, Lin Hua, and Lin Li, Identification of Novel Breast Cancer Subtype-Specific Biomarkers by Integrating Genomics Analysis of DNA Copy Number Aberrations and miRNA-mRNA Dual Expression Profiling, BioMed Research International, vol. 2015, Article ID 746970, 17 pages, 2015. https://doi.org/10.1155/2015/746970.

- Menolascina, F., Tommasi, S., Paradiso, A.V., Cortellino, M., Bevilacqua, V., Mastronardi, G. (2007). Novel Data Mining Techniques in aCGH based Breast Cancer Subtypes Profiling: the Biological Perspective. 2007 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology, 9-16.

- Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. Nat Genet. 2005;37:S1117. doi: 10.1038/ng1569.

- Bergamaschi A, Kim YH, Wang P, Srlie T, Hernandez-Boussard T, Lonning PE, Tibshirani R, Brresen-Dale AL, Pollack JR. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. Genes Chromosomes Cancer. 2006;45:103340. doi: 10.1002/gcc.20366.

- Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic acids research, 30(4), e15.