Lieven.Verbeke@intec.ugent.be

Identifying relevant pathways for different breast cancer subtypes using network based data integration

Lieven P.C. Verbeke[1], Anna Carolina Fierro[2], Jimmy Van Den Eynden[1,3], Piet Demeester[1], Jan Fostier[1], and Kathleen Marchal[1,3]

[1]IBCN - iMinds, Ghent University, Belgium, [2]Department of Microbial and Molecular Systems, KU Leuven, Belgium, [3]Department of Plant Biotechnology and Bioinformatics, Ghent University, Belgium

Introduction

Breast cancer is the main cause of cancer death in women worldwide. It is a heterogeneous cancer, with different subtypes and associated therapies and prognosis. We investigated the well adopted (expression based) PAM50 tumor classification and tried to assess pathway activity for each subtype. We developed a new network based method that integrates, in a single step, different types of data and prior knowledge in the form of known gene interactions. Using the method, 210 KEGG pathways were given a score that expresses the affinity of the pathway with each of the PAM50 subtypes. Our results recapitulate to a large extent what is known about pathway activity in breast cancer, confirming the potential of the proposed method to select relevant pathways from a number of candidate pathways.

The data

We used data from The Cancer Genome Atlas (TCGA). Data were downloaded for 463 patients with a known PAM50 based subtype. The following data types were available **(Figure 1)**:

- Gene expression (1700 differentially expressed genes selected)
- Somatic mutations (522 mutations selected using an external procedure)
- Copy number variation (tumor vs. normal, 141 genes selected)
- 210 KEGG pathways (disease pathways were not included)

Method

All data were integrated in a single dataset using a network based approach **(Figure 2).** The integrated network contains different types of entities as nodes:

- Patients
- Genes
- Mutations
- Copy number aberrations

These entities are connected using different types of relations. We incorporated prior knowledge in the form of known gene – gene interactions.

Once the integrated network is constructed, network based similarity measures can be used to calculate the average similarity between a group of patients and the genes in a particular pathway. These similarities can then be converted to an affinity score (between 0 – no affinity and 1 – very high affinity) representing the importance of each pathway for a given breast cancer subtype. We expect the best scoring pathways to be involved in signaling, cell cycle and apoptosis.

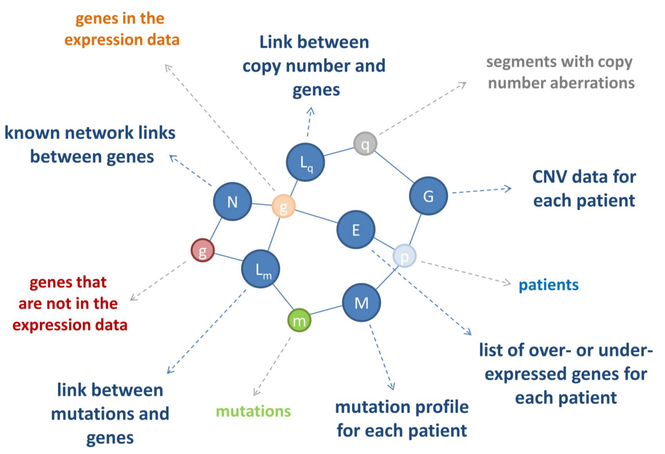

Figure 1. overview and interrelation of the different data sources

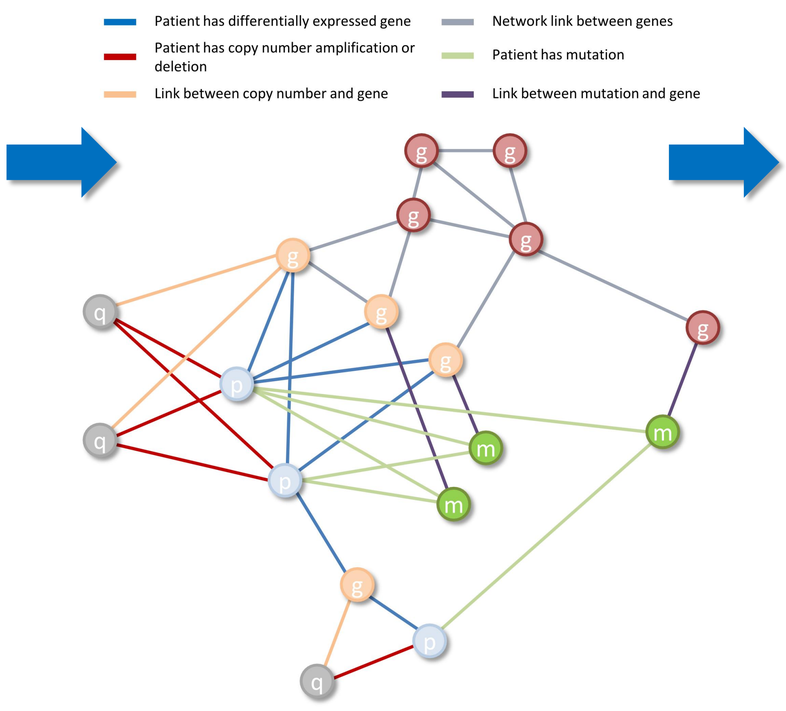

Figure 2. The integrated network containing different types of entities and relations

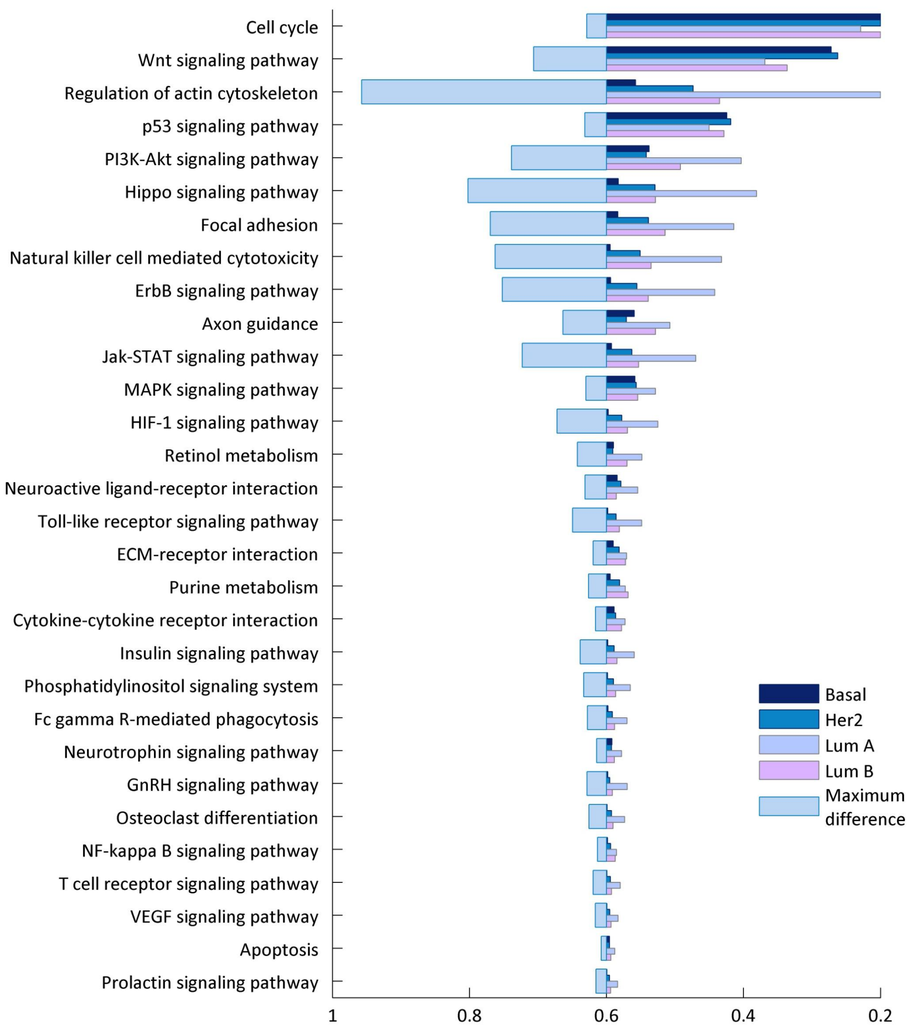

Figure 3. The top 30 best-scoring KEGG pathways

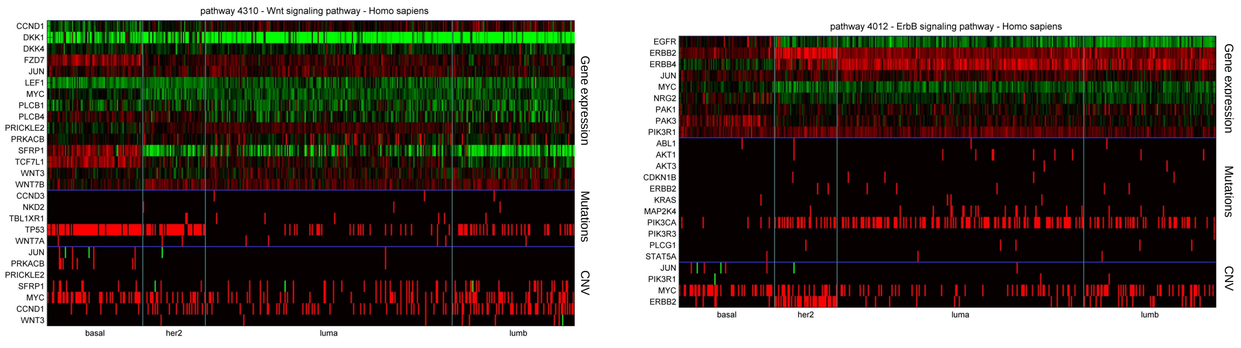

Figure 4. Two interesting high-scoring pathways illustrating differences between breast cancer subtypes (red=high expression, mutated or copy number amplification, green=low expression or copy number deletion)

Results

We calculated a similarity score for each subtype vs. each of 210 KEGG pathways, and ranked the pathways according to their average similarity score over all subtypes (Basal, Her2, Luminal A and Luminal B).

Figure 3 displays the top 30 best scoring pathways, together with an indication how variable the score was between subtypes. Our results are in line with the results of PARADIGM (Vaske et al. ,2010, Bioinformatics 26).

Some interesting pathways were cherry-picked and displayed in relation to all available data in more detail in **Figure 4**.