# Biomarkers Discovery in Breast Cancer by Interactome-Transcriptome Integration

Maxime Garcia*, Raphaële Millat-Carus¤, Pascal Finetti¤, Daniel Birnbaum¤, François Bertucci¤§, Ghislain Bidaut*

* Integrative Bioinformatics Platform ¤ Molecular Oncology § Medical Oncology

Centre de Recherche en Cancérologie de Marseille, Institut Paoli-Calmettes, Aix-Marseille Université, U1068 Inserm, UMR 7258 CNRS

**KEYWORDS:**

- Breast Cancer
- Interactome
- Transcriptome
- Massive Data Integration
- Biomarkers
- Predictive Signature
- Subnetworks
- Analysis Pipeline

## INTRODUCTION

**Breast cancer** is the most common and the **most deadly cancer type** in women[1]. Patients undergo **avoidable adjuvant chemotherapy** in 70-80% of cases in node-negative early breast cancer[2].

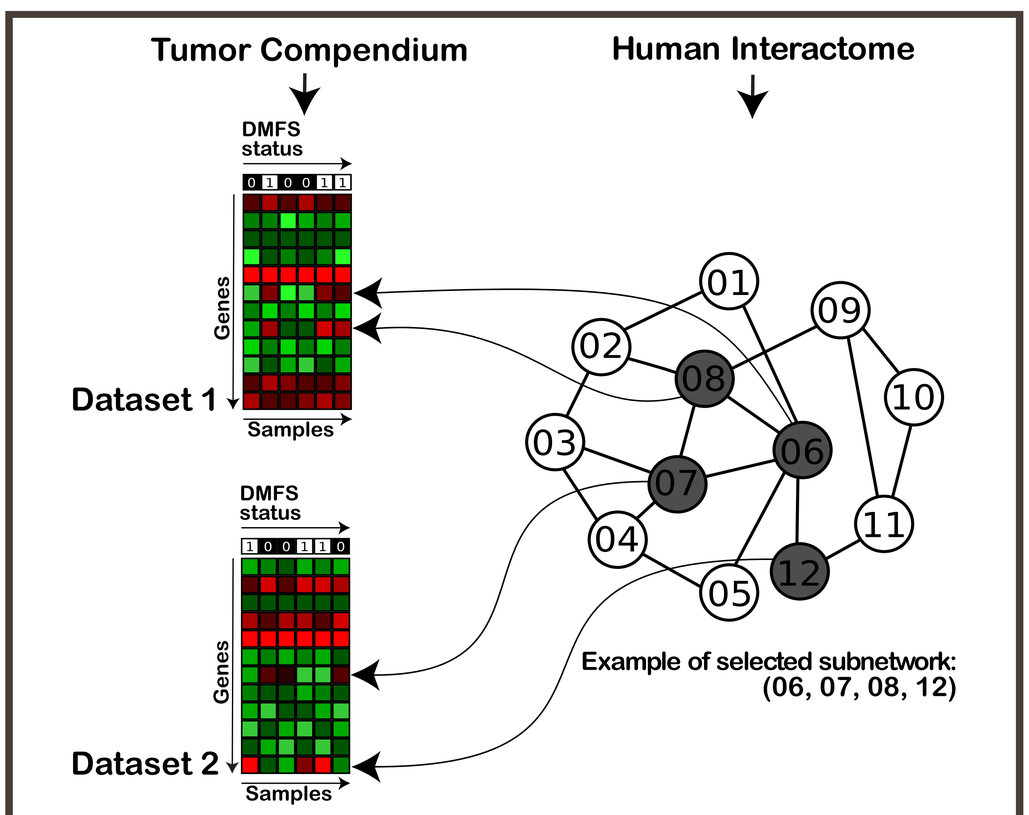
High-throughput gene-expression profiling technologies yield genomic signatures to **predict clinical conditions** or patient outcome and help refine a therapeutic decision.

Two independent studies found a **70-genes signature** (Van't veer et al.[3] 2002) and a **76-genes signature** (Wang et al.[4] 2005) predicting breast cancer relapse. They only have **3 genes in common** (less than **5%** of all genes in the two signatures).

It has been proved that **more than one 70-genes signature** with the same predictive power exists[5]. Such signatures show dependency on training set, **lack of generalization and instability**.

## METHODS

We propose an **interactome-based algorithm**[6,7]: ITI (Interactome-Transcriptome Integration) to find a **generalizable signature** for predicting 5 years relapse free survival in breast cancer.




**Integration Transcriptome-Interactome Algorithm**
Principle of subnetwork selection with the integration of transcriptomic datasets and interactome. Subnetworks countain biomarkers and are characterized by proteins whose expression is correlated with the clinical status in at least one dataset

ITI re-implements the Chuang et al.[8] algorithm, by extracting discriminative regions in the interactome (**subnetworks**), with the additional capability to integrate several gene-expression data sets simultaneously.

**Two data types are fed to the algorithm:**

- **Large scale interaction data**. We integrated five existing human protein-protein interaction (PPI) maps by **uniqueness of NCBI EntrezGene identifiers**, leading to a final set of 70,530 interactions among 13,202 proteins (HPRD[9], Ramani[10], MINT[11], IntAct[12] and DIP[13]).

- **Gene expression profiles (GEP)**. We built a compendium of breast cancer tumors profiles by examining datasets available with **clinical information** on the NCBI GEO database. Each dataset was downloaded from GEO as raw data and normalized within Bioconductor using affy and gcrma packages. Tumors without relapse information were removed, leading to a final compendium of 5 datasets containing 787 tumors[4,14-17].

**Workflow:**

One dataset was left out for **cross-validation** purpose with **independent testing**. Pearson correlation is computed between GEPs and clinical information (Distant Metastasis Free Survival [DMFS] status) for each dataset.

Interactome regions whose gene expression is highly correlated with DMFS status are then detected. Random distributions of score are drawn to assign p-values to the subnetworks and perform a **statistical validation**.

Finally, the discriminative power of statistically significant subnetworks is tested against an independent dataset.

These are stored and available for exploration in a **bioinformatics resource** : the ITI web site ( bioinformatique.marseille.inserm.fr/iti )

## RESULTS & DISCUSSION

**Intrinsic biology** of extracted subnetworks was examined using annotation information from the **NCBI EntrezGene** database and the **Gene Ontology** Consortium.

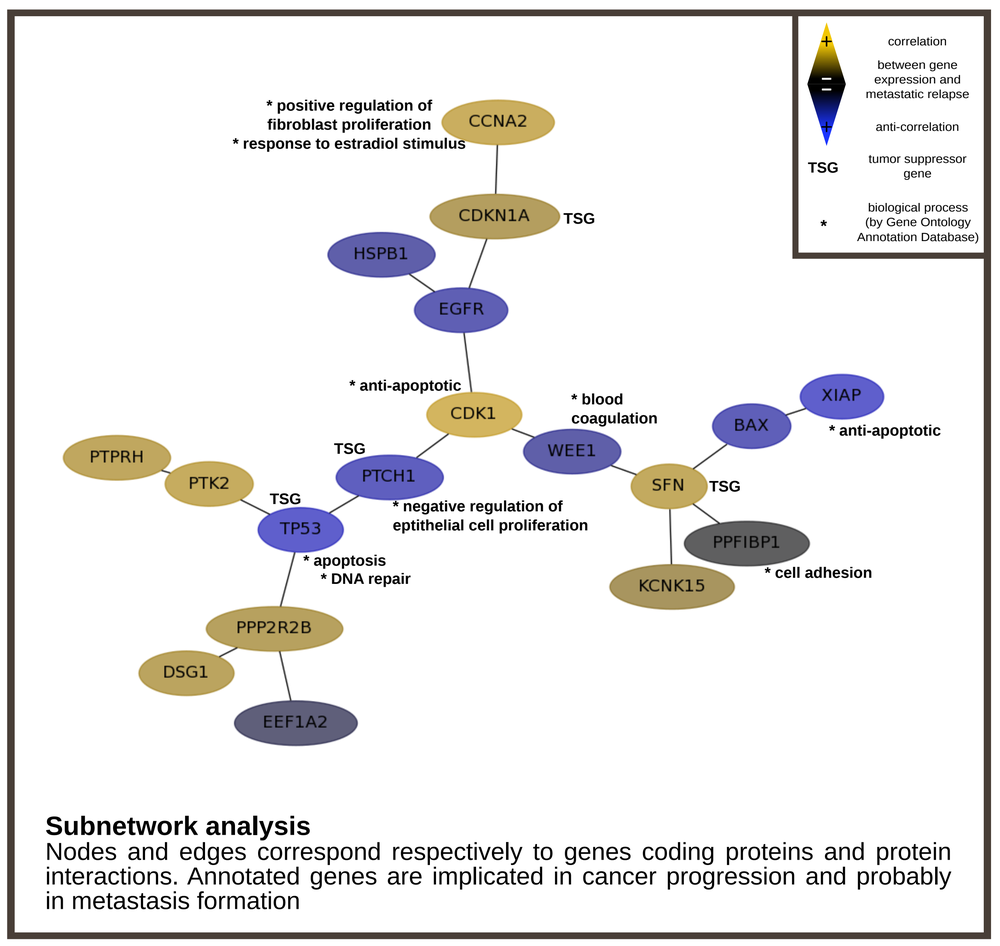
We found that **subnetworks** formed complexes functionally supporting the studied disease for **metabolism**, **cell cycle control**, **proliferation**, **cell-cell adhesion** and **immunological response**, which are known **mechanisms** of **cancer** and **metastatic** process. Several **driver genes** were detected, including **CDK1**, **TP53**, **PDGFB** and some **not previously linked to breast cancer relapse**[7].

**Four studies ER status specific** were made. We had **ER-** or **ER+** tumors and left out Desmedt[14] or van de Vijver[3] dataset. **SVM classification** was made base on a **set of subnetworks** against an **independent dataset** (Desmedt[14] or van de Vijver[3]).

| Dataset | Desmedt | | | van de Vijver | | |
|---|---|---|---|---|---|---|
| signature | 76G | 70G | **ITI** | 76G | 70G | **ITI** |
| ER- | 37.7 % | 44.2 % | **54.1 %** | 55.6 % | 52.8 % | **52.8 %** |
| ER+ | 60.4 % | 41.1 % | **73.6 %** | 63.2 % | 62.3 % | **51.8 %** |

**Classification results** comparison for accuracy between ITI and other signatures[3,4] on the two test datasets of Desmedt and van de Vijver, for ER+ and ER- tumors

We **compared** performances between signatures found with ITI algorithm and **previously established signatures** : the Mammaprint 70 genes signature[3] (70G) and the ER Status 76 genes signature[4] (76G). This test shows that not only ITI has **better predictive power**, but that ITI is **more generalizable**[7]. Our signatures have **11.5 to 32.8%** genes in common, so they are also **less unstable**[7].




**Subnetwork analysis**
Nodes and edges correspond respectively to genes coding proteins and protein interactions. Annotated genes are implicated in cancer progression and probably in metastasis formation

**Subnetworks are stored** in the ITI web site (bioinformatique.inserm.fr/iti). This resource is the **first of its kind** to allow **linking a human interactome to diseases** or clinical situations[6,7]. It can be mined for **isolating** potential **drug targets**, **tumor suppressor genes** or **oncogenes**, as well as **prognostic signatures** for metastasis of breast cancer as well as other diseases. It has the **potential** of becoming **the starting point to establish finer disease models** by **systems biology techniques**.

## BIBLIOGRAPHY

1 World Cancer Report. *International Agency for Research on Cancer*. 2008.
2 F Bertucci et al. *Prognosis of breast cancer and gene expression profiling using DNA arrays*. Annals of the New York Academy of Sciences 2002.
3 MJ van de Vijver et al. *A gene-expression signature as a predictor of survival in breast cancer*. The New England Journal of Medicine 2002.
4 Y Wang et al., *Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer*. The Lancet 2005.
5 L Ein-Dor et al. *Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer*. Proceedings of the National Academy of Sciences of the United States of America 2006.
6 M Garcia et al. *Linking interactome to disease: a network-based analysis of metastatic relapse in breast cancer*. Handbook of Research on Computational and Systems Biology: Interdisciplinary Applications. 2011.
7 M Garcia et al. *Interactome-Transcriptome integration for predicting distant metastasis in breast cancer*. Bioinformatics 2012.
8 HY Chuang et al. *Network-based classification of breast cancer metastasis*. Molecular Systems Biology 2007.
9 TSK Prasad et al. *Human protein reference database-2009 update*. Nucleic Acids Research 2009.
10 AK Ramani et al. *Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome*. Genome Biology 2005.
11 A. Ceolet al. *Mint, the molecular interaction database: 2009 update*. Nucleic Acids Research 2010.
12 B. Aranda et al. *The intact molecular interaction database in 2010*. Nucleic Acids Research 2010.
13 L. Salwinski et al. *The database of interacting proteins: 2004 update*. Nucleic Acids Research 2004.
14 C. Desmedt et al. *Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes*. Clinical Cancer Research 2008.
15 S Loi et al. *Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen*. BMC Genomics 2008.
16 R Sabatier et al. *A gene expression signature identifies two prognostic subgroups of basal breast cancer*. Breast Cancer Research and Treatment 2011
17 M Schmidt et al. *The humoral immune system has a key prognostic impact in node-negative breast cancer*. Cancer Research2008.

## ACKNOWLEDGEMENTS

Research is funded by the **Institut National du Cancer** and the Institut National de la Santé et de la Recherche Médicale (**Inserm**). Our Beowulf cluster was funded by a **Fondation pour la Recherche Médicale** grant.

Maxime Garcia is funded by a fellowship from **Inserm** and the **Provence-Alpes-Côte d'Azur** Region.

**CONTACTS:**

Maxime Garcia:
maxime.garcia@inserm.fr

Pascal Finetti:
finettip@ipc.unicancer.fr

Daniel Birnbaum:
daniel.birnbaum@inserm.fr

François Bertucci:
bertuccif@ipc.unicancer.fr

Ghislain Bidaut:
ghislain.bidaut@inserm.fr

**AVAILABILITY:**

bioinformatique.inserm.fr/iti

**PERSPECTIVES:**

- CGHarray Integration
- Methylome Integration
- Test other Cancers, Diseases
- Data Repository Extension
- Canonical Pathways Inclusion