Applying expression profile similarity for discovery of patient-specific functional mutations

Guofeng Meng
Partner Institute of Computational Biology, Yueyang Road No. 333, Shanghai, China
email: menggf@gmail.com

Abstract

The progress of cancer genome sequencing projects yields unprecedented information of mutations for numerous patients. However, the complexity of mutation profiles of patients hinders the further understanding of mechanisms of oncogenesis. One basic question is how to uncover mutations with functional impacts. In this work, we introduce a computational method to predict functional somatic mutations for each of patient by integrating mutation recurrence with similarity of expression profiles of patients. With this method, the functional mutations are determined by checking the mutation enrichment among a group of patients with similar expression profiles. We applied this method to three cancer types and identified the functional mutations. Comparison of the predictions for three cancer types suggested that most of the functional mutations were cancer-type-specific with one exception to p53. By checking prediction results, we found that our method effectively filtered non-functional mutations resulting from large protein sizes. In addition, this methods can also perform functional annotation to each patient to describe their association with signalling pathways or biological processes. In breast cancer, we predicted ”cell adhesion” and other mutated gene associated terms to be significantly enriched among patients.

1 Introduction

Cancer arises from accumulations of somatic mutations and other genetic alterations, which leads to abnormal cell proliferation [1]. With the progress of cancer genome sequencing projects, such as The Cancer Genome Atlas (TCGA) project, mutation information for more and more patients are becoming publicly available [2, 3, 4], which provides the foundation to uncover mechanisms of oncogenesis. However, cancer patients usually carry an average of 33 to 66 mutations in the protein coding regions and those mutations are supposed to take unequal roles in oncogenesis [5, 6]. It remains a great challenge to distinguish the functional mutations that give cells with growth advantages [7], from the ones with non-crucial roles to oncogenesis.

Many computational tools have been developed to predict the functional mutations. One popular strategy is to find recurrently mutated genes. Genes with higher mutation frequency are supposed to take a more essential role in oncogenesis [6, 8, 9]. To improve the accuracy, most computational tools also consider background mutation rates and protein sizes in discovery of recurrent mutations [10, 11, 12]. Another strategy is mutual exclusivity, which is based on an assumption that if one gene is mutated in a patient, other members acting in the same signalling pathway are less likely to be mutated [13]. Using mutual exclusivity, members of signalling pathways are investigated for their coverage across a large number of patients while not co-mutated in the same patients [14, 15, 16, 17]. Besides of functional annotations, expression information is also used in driver mutation discovery. One popular application is to find functional copy number variations (CNVs) by identifying the differentially expressed genes located in the CNV regions [18]. For somatic mutations, correlation of gene expression profiles is also used to identify the functional mutations based on the assumption that somatic mutations will result in altered expression of their downstream targets [19].

Somatic mutations in coding regions are usually supposed to involve oncogenesis by affecting the activities of proteins associated with cell proliferation [20]. However, somatic mutations at different sites have different effects on protein activities. One solution is to study the stability of protein structures after mutations. Amino acid changes in the protein sequences can either stabilize, destabilize or have no effect on protein structures. Methods based on this strategy calculate the protein folding free energy, which will be used to evaluate the protein structural stability [21, 22, 23]. However, those methods still face the problems of establishing the direct connections between protein structures and activities.

In this work, we introduce a novel method to recover functional somatic mutations for each of the studied patients by integrating mutation recurrence and expression profile similarity among patients. This method is based on two assumptions: (1) functional mutations will lead to altered expression of genes which reflects consequences of somatic gene mutations; (2) patients with similar expression profiles are more likely to carry functional mutations to the same genes. For each patient, we can find a group of patients that have similar expression profiles to him/her and those patients are called neighboring patients. Mutations of the studied patient are evaluated for their enrichment among all the neighboring patients. Mutations with enough enrichment are predicted to be functional mutations. The functional mutations of all patients can be recovered by repeating this process. This method also performs function annotation analysis to mutated genes in neighboring patients so that each patient can be assigned with some functional terms to indicate his/her association with signalling pathways or biological processes. As applications, we applied this method to three cancer data sets and identified the functional mutations for three type of cancers respectively.

2 Materials and Methods

2.1 Dataset and their pre-processing

In this work, we focus on prediction of functional somatic mutations, including missense mutation, nonsense mutation and frame-shift. All the mutation and expression data of cancer patients used in this work are downloaded from TCGA project (by May 2013) at level two, with which the expression data have been normalized within samples and that the somatic mutations have also been annotated in exons by the pilot of TCGA project. We exclude those patients with only mutation or expression data for next-step analysis.

2.2 Expression biomarkers

To describe the expression profiles of patients, we firstly determine a group of genes or probes as expression biomarkers by selecting the top 2000 genes or probes with the most expression variances among all the patients. These biomarkers are supposed to better reflect the expression consequences of gene mutations. The expression profile of one patient is described as a vector with the expression values of biomarkers as the elements.

2.3 Neighboring patients: patients with similar expression profiles

The similarity of expression profiles of the patients was measured by using Pearson's correlation r. Conversely, the distances between patients could be described by $1-r$. For each patient, we could find his/her neighboring patients by selecting n patients with most significant positive correlation values at a minimum r cutoff (e.g. $r > 0.6$), where n ranged from 20 to 50.

3 Results

3.1 Association of mutation and expression profiles

As an exemplary illustration, we checked the association between somatic mutations and cancer patients with similar expression profiles by investigating 516 breast invasive cancer patients, for which both expression and somatic mutation information are available from the TCGA project. We selected 2000 probes with most expression variances as the biomarkers and performed hierarchical clustering to the patients. Then we checked the distribution of mutated genes on the hierarchical tree. One example is CDH1 gene, which is mutated in 35 out of 516 patients. As shown in Figure 1, patients with CDH1 gene mutation were not randomly distributed but preferentially clustered together. To quantify this, we classified the patients into six groups based on the structure of hierarchical clustering tree. Among groups, the frequency of CHD1 mutations varied greatly. In Cluster IV, we observed 22 out of 120 patients to carry CDH1 mutations, which was also the most enriched group. Fisher's exact test indicated the significance of this enrichment to be $p < 4.4e - 8$. In cluster VI, there were 6 out of 70 patients to carry CDH1 mutation without statistical significance. For cluster I and V, there was even no patient with CDH1 mutation. We also checked the mutation distribution of other genes and observed non-random distribution in the expression subclusters in nearly all the test cases.

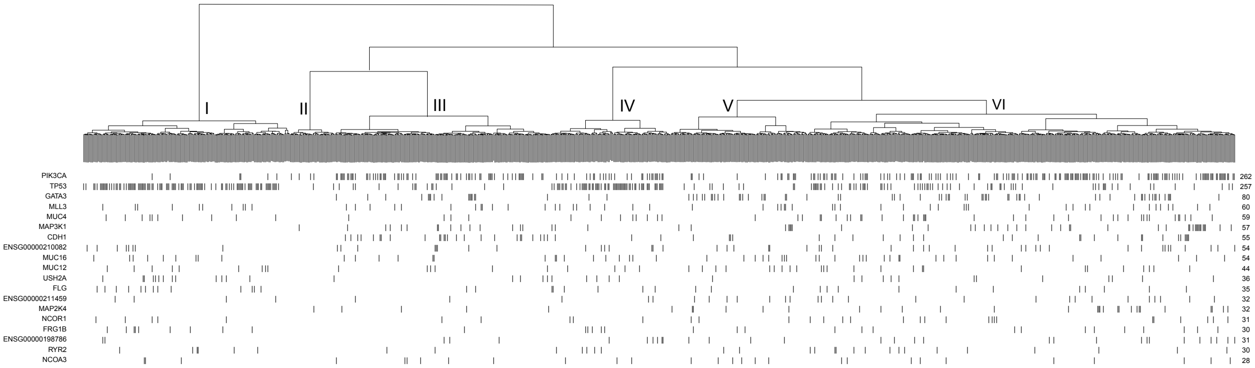


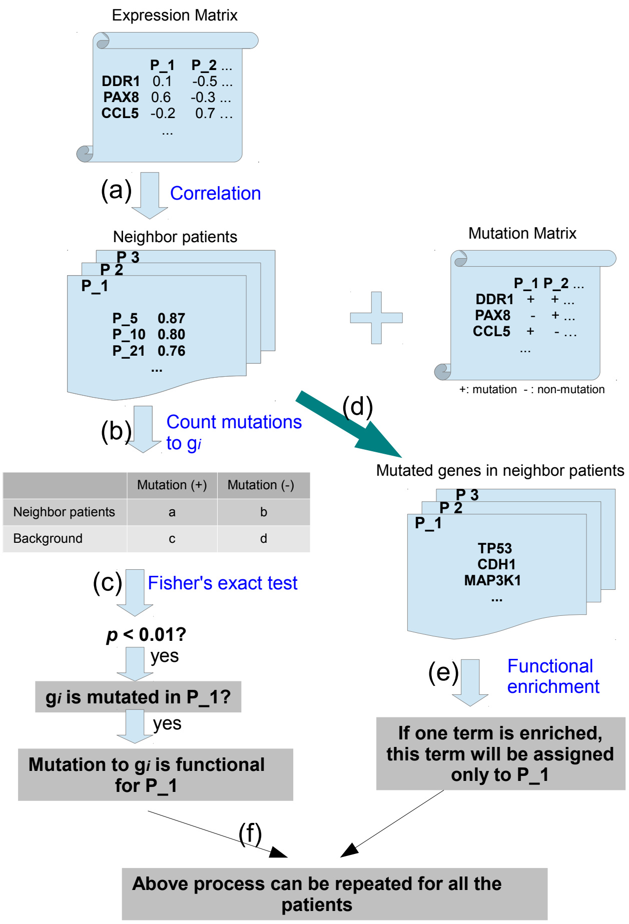
Figure 1: Hierachical clustering and the mutation distribution of gene

These results suggest that patients with a similar set of mutated genes are more likely to have similar expression profiles, which can be explained by the fact that gene mutations would directly or indirectly affect transcription of their downstream targets [19, 24]. Based on this observation, we propose that genes affected by functional mutations can be identified by checking their enrichment in patients with similar expression profiles.

3.2 Pipeline to find functional mutations

In this work, we propose a computational method to infer the functional mutations for each studied patient. This method is based on two assumptions that, (1) genes carrying functional mutations directly or indirectly lead to altered expression of their downstream target genes; (2) patients with the same functional mutations will share similar expression profiles. Thus, for each mutated gene from one patient, we can check its mutation enrichment in the neighboring patients. The impact of gene mutation to studied patients can be described by a p-value based on statistical frameworks.

The whole process is described in Figure 2. This method requires two matrices as input: one expression matrix and one binary mutation matrix. As described in Section Methods, the expression matrix are explored to determine the neighbor patients by checking its expression similarity. For patient P_1 with a mutation to



	Mutation (+)	Mutation (-)
Neighbor patients	a	b
Background	c	d

Figure 2: Pipeline to predict functional mutations and terms (see main text for detailed description).

gene g_i, his/her neighboring patients are selected by choosing n patients (including P_1) with the most similar expression profiles with P_1, where n ranged from 5 to 30 (a). Then, the patient number with mutation to gene g_i is counted in all the neighbors of P_1 (b). Based on randomly simulation, in which the neighbor of patients are randomly selected, the enrichment of mutations to g_i among neighbor patients is evaluated by Fisher's exactly test. A p-value is assigned to studied patient to describe its mutation enrichment to g_i (c), which also indicates its functional importance. If the p-value is significant enough (e.g. 0.01), mutation to gene g_i in patient P_1 will be treated as functional mutation of only patient P_1. The same process can be repeated for other mutated genes and patients one by one (f).

With our method, we also check the functional annotation for each studied patient by enrichment analysis to all the mutated genes in neighboring patients. The functional annotation is described by Gene ontology (GO) or KEGG terms. In this step, all the mutated genes from patient P_1 and his/her neighboring patients (d) are checked for enriched functional terms by comparing to the background occurrences with Fisher's exact test (e). If a term is significantly enriched (e.g. $p < 0.01$), it will be assigned to patient P_1 as functional annotation. This term indicates the functional consequences of gene mutations in patient P_1. For the other patients, the same process can be repeated to find their associated functional terms (f).

3.3 Application to breast cancer

As an application, we applied our method to 516 beast cancer patients from the TCGA project. The functional mutations were chosen at a cutoff of $p < 0.01$. In Figure 1 (1), we show 19 genes with the most mutation recurrences in breast cancer. For all the genes, only part of their mutations are predicted to be functional. Among them, PIK3CA and TP53 were especially recurrently mutated based on our prediction. PIK3CA gene was observed to be mutated in 175 breast cancer patients and 123 of them were confirmed to be functional (about 70.3%), which made PIK3CA to be the gene with the most confirmed mutations, even though it was not the one with the most somatic mutations in breast cancer. Another gene was TP53 with functional mutations in 107 patients, about 56.9% of observed somatic mutations. For other genes, relatively fewer functional mutations were observed compared to PIK3CA and TP53. However, many of them have been widely reported for their critical roles in breast cancer, such as MAP3K1 [25], CDH1 [26], GATA3 [27], which indicated their important roles but in fewer patients. Overall, we observed 408 of 516 breast cancer patients to carry at least one of these 10 predicted driver mutations, covering about 79% of all the patients.

Table 1: Top 10 of functional mutations in breast cancer

Mutated gene	No. somatic mutation	No. functional mutation	Percentage
PIK3CA	175	123	70.3%
TP53	188	107	56.9%
MAP3K1	40	30	75.0%
CDH1	35	29	82.6%
GATA3	56	19	33.9%
RUNX1	19	13	68.4%
CTCF	15	10	66.7%
CACNA1B	14	9	64.2%
DNAH17	14	8	57.1%
MAP2K4	21	8	38.1%

Next, we tested whether our predicted mutations were more likely to be functional. We checked genes with recurrent somatic mutations while less often predicted to be functional. One example was the TTN gene with somatic mutation in 90 patients, which ranked as the gene with the third most recurrent mutations. However, we only predicted three of them to be functional. TTN protein is a component of vertebrate striated muscle [28]. By searching literature, we did not find any reports to support TTN to take critical role in cancer. By checking its protein sequence, we found that it is very long with a length of 27,000 to 33,000 amino acids [29], indicating a recurrence bias of its mutations. This guess was also confirmed by the analysis with MutSig, which considered the protein size and background mutation rates into evaluation

[10]. Similar results are observed with with other genes, such as MUC6 with mutations in 57 patients while predicted to be functional in only two patients. Even though our pipeline did not consider the protein size, it successfully identified fake recurrent mutations resulting from large protein sizes. These results provide evidences to support the specificity of our method to recover functional mutations.

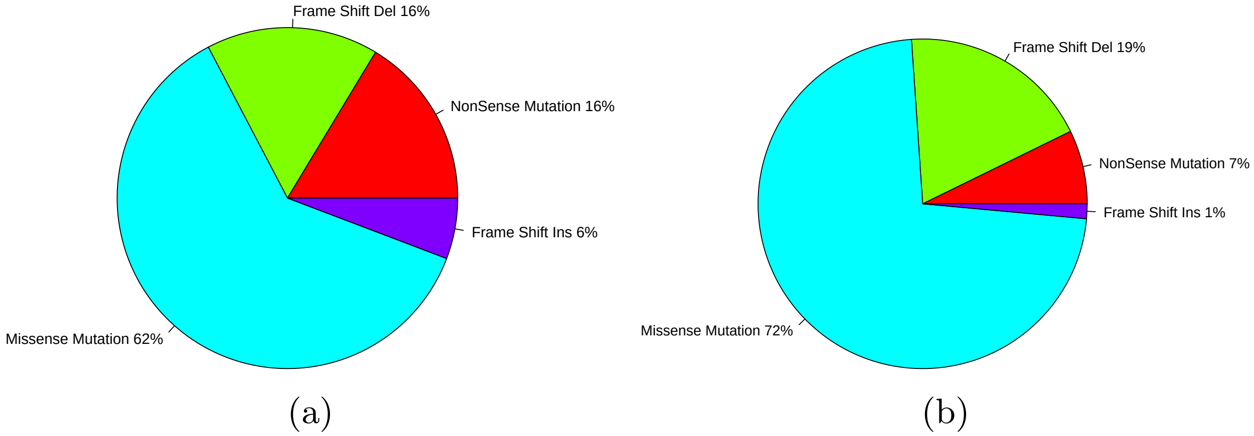


Figure 3: Mutation type differences between functional (a) and non-functional TP53 muations (b)

3.4 Comparison with Mutsig in driver mutation recovery

Before our work, many tools tools have been developed to predict the drive mutation to cancer. One popular strategy is evaluated observed mutation frequency with background mutation ratio, estimated by . In this work we did not use

3.5 Mutation types

Based on mutation types, all the somatic mutation can be divided into missense mutation, nonsense mutation, silent mutation, frame shift Insertion and frame shift Deletion. Taking TP53 as an example, we checked the mutation type distribution between two groups of TP53 mutations: (1) functional TP53 mutation as identified by our method and (2) non-functional TP53 mutations. As shown in Figure 3, most of mutations were missense mutations ($\sim$65 %). Statistical test did not support any significance difference between two groups of patients. However, nonsense mutation and Frame shift Insert types had great ratio differences between two groups. 16% of functional TP53 mutations were observed to be nonsense mutations while only 7% of non-functional TP53 mutations. The significance based on Fisher's exact test was at $p < 0.03$. A similar result was observed with Frame shift Insertion, which was 6% of functional mutations and 1% of non-functional mutations ($p < 0.12$). The enrichment of those types of mutations can be explained by the fact that nonsense mutations and Frame shift Insertions are more more likely to disrupt protein function than missense mutation. We also checked other recurrently mutated genes and similar results were observed. These results suggest that predicted functional mutations are more likely to influence the protein function.

3.6 Cancer-specific gene mutations

Besides of breast cancer, we also applied our method to Ovarian serous cystadenocarcinoma (OV) and Glioblastoma multiforme (GBM) with data from the TCGA project. Figure 4 shows genes with the most recurrent functional mutations of three cancer types. By checking the published literature, we found that most of the predicted genes have been implicated to to have an important roles in oncogenesis [6]. However, we observed quite different mutation preferences among three cancer types. As shown in Figure 4, only TP53 gene is shared by all three cancer types and NF1 gene is shared by OV and GBM. Other genes are preferentially mutated in one cancer type, which can be called cancer-specific mutations. These cancer-specific mutations had higher mutation recurrences and were supposed to take more essential roles in the oncogenesis of a specific cancer type.

Cancer-specific mutations are not necessary to indicate non-function in other cancers. One example is BRCA1 gene. BRCA1 was predicted to be mutated in 17 out of 441 patients for OV, ranking as the third most recurrently mutated gene. Based on our definition, BRCA1 is supposed to be OV-specific. However, we also observed functional mutations to BRCA1 in breast cancer. The roles of BRCA1 in the oncogenesis of breast cancer have been widely reported [30] and researches even suggest that patients with BRCA1 mutations have a possibility of 50%-80% to develop breast cancer before the age of 70 [31]. However, mutations to BRCA1 is not supposed to be the most important causal reasons due to the relatively low recurrence, which was at 6 out of 516 patients. BRCA1 is proposed to be involve in many cancer associated biological process, including DNA repair, histone ubiquitination and checkpoint control [32, 33]. One possible explanation for its cancer-specific mutation is that the functional defect of BRCA1 is more likely to be compensated in breast cancer.

3.7 Enriched functional terms

As described in Section 2, we can perform functional annotations for each of studied patients by functional enrichment analysis to mutated genes of the neighboring patients. At a cutoff of $p < 0.01$, we observed some terms to be associated with large proportion of breast cancer patients (see Figure 5). One example was the term “cell adhesion”, which was associated with 99.4% of patients. Similarly, other extracellular matrix (ECM) related terms such as “extracellular structure organization” and “cell-matrix adhesion”, were also observed to be associated with 34.1% and 22.1% of patients respectively. These observation are in line with

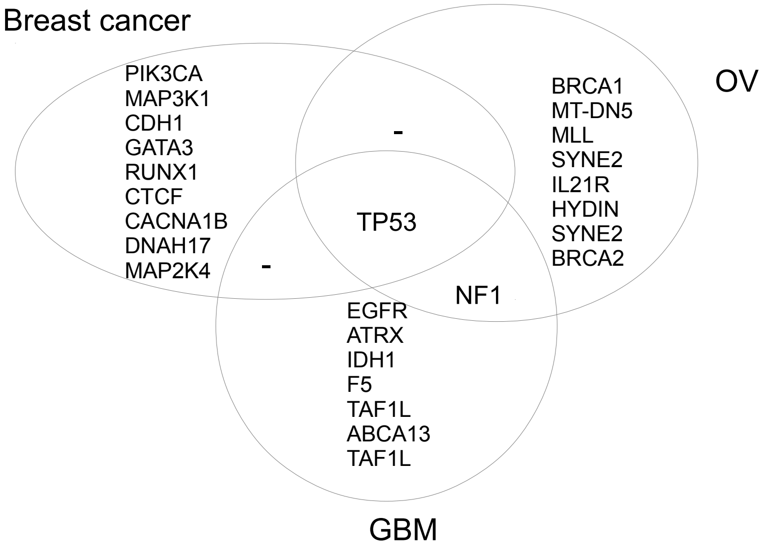


Figure 4: Predicted genes with recurrent functional mutation for three cancers. OV: Ovarian serous cystadenocarcinoma; GBM: Glioblastoma multiforme

reports about ECM for its critical roles in cancer [34, 35]. This result also suggests ECM to be one potential target to develop anticancer medicine. Above, we reported recurrently mutated genes, such as PIK3CA and MAP3K1. Functional annotation supported their roles by enriched GO terms like "phosphoinositide 3-kinase (PI3K) cascade" (30.6% of the patients) and "MAPKKK cascade" (34.5% of the patients). Considering the fact that PIK3CA and MAP3K1 were ranked as the most recurrently mutated genes, PI3K signalling pathway and MAPKKK cascade pathway were supposed to be the most essential pathways during the oncogenesis of breast cancer. Besides these pathways, we also observed enriched terms such as "JNK cascade" and "cell cycle phase", which were reported to be related to oncogenesis of cancers [1].

Based on textbook knowledge and published literature, we could assign some genes to some well-studied pathways or biological processes. For example, TP53 is involved in the p53 mediated DNA damage response [36]; MAP3K1 in the MAPKKK cascade pathway [37]. These terms could be used to evaluate functional differences between patients predicted with or without functional mutations. Taking TP53 mutation of breast cancer as an example, we checked the number of patients with p53 associated term "DNA damage response, signal transduction by p53 class mediator". Just as showed in Figure 6(a), we observed 6 out of 107 patients with predicted functional TP53 mutation to be annotated with this term while 12 out of 81 non-functional TP53 mutation carriers were annotated with this term. Fisher's exact test indicated a significant ratio difference at $p < 0.03$. We further checked other TP53 associated terms and observed similar results (results not shown). Besides of breast cancer, we also performed the same evaluation with OV and

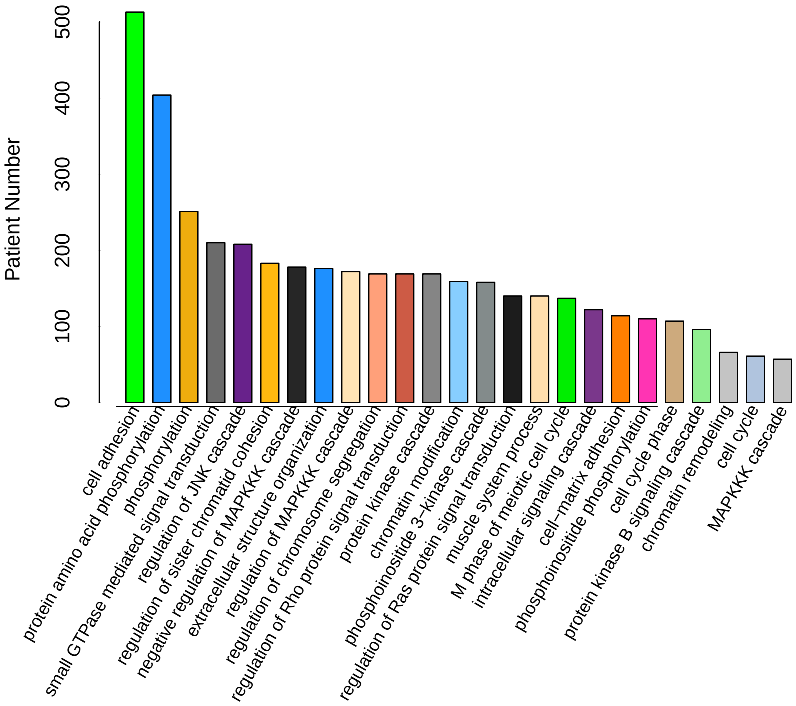
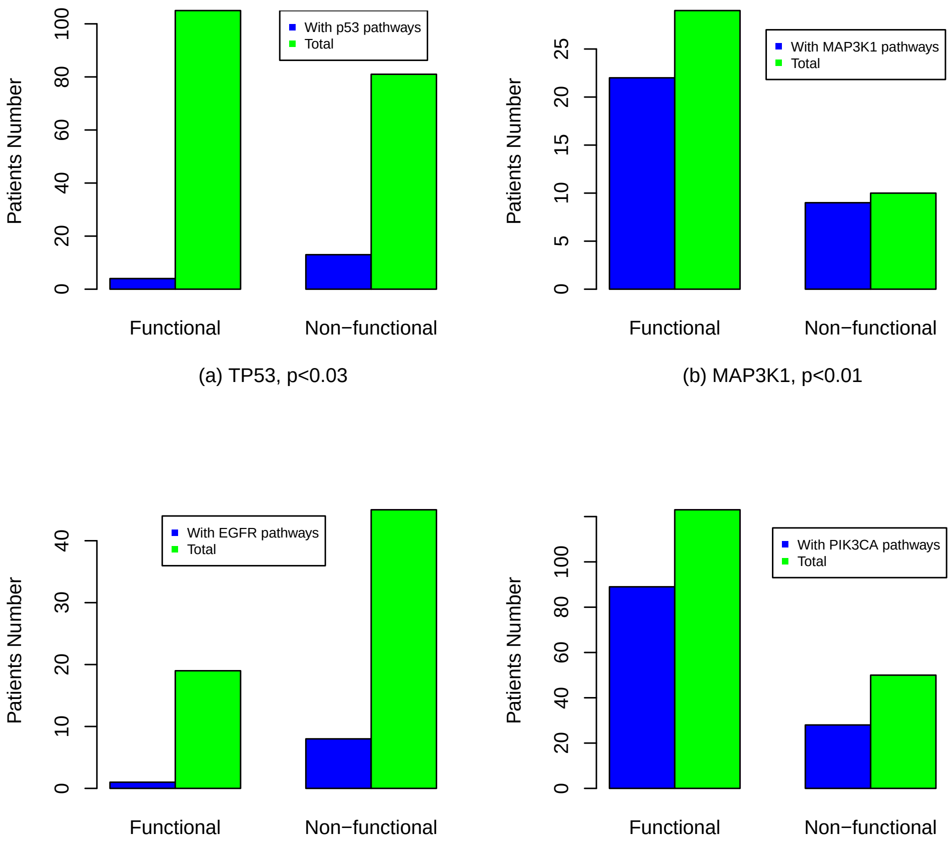


Figure 5: GO terms enriched with breast cancer patients

GBM. Similar results were observed with TP53 mutations in OV. In GBM, we did not observe significant differences, which might be due to the limited number of patients with TP53 mutations, which was only 65 patient. In a way, these results suggest that patients with functional TP53 mutations are less likely to carry other mutations to the members of TP53 associated pathways. In breast cancer, we also checked MAP3K1 mutation with the similar methods for “MAPKKK cascade” term (see Figure 6(b)). Similar to TP53, we observed that the patients with functional MAP3K1 mutations were less likely to carry the mutations to other components of MAPKKK cascade pathways ($p < 0.01$). In GBM, we checked another well-known gene EGFR for that it involved in “epidermal growth factor receptor signalling pathway” (see Figure 6(c)). Similarly, other genes from EGFR signalling pathways were less likely to be mutated in patients with functional EGFR mutations ($p < 0.08$). In summary, all above examples were in line with the popular assumption for gene mutations: mutual exclusivity [13]. However, one contrary example was also observed with predicted PIK3CA mutations in breast cancer (see Figure 6(d)). By checking PIK3CA associated term “phosphoinositide 3-kinase cascade”, more members of PI3K cascade pathway were observed with predicted functional PIK3CA mutation carriers ($p < 0.02$),which was in line with previous report about cooperation



(a) TP53, p<0.03

(b) MAP3K1, p<0.01

(c) EGFR, p<0.08

(d) PIK3CA, p<0.02

Figure 6: Functional enrichment differences between patients with or without predicted functional mutations.

of PIK3CA with other oncogenes [38, 39]. This result indicates that mutual exclusivity is not applicable to PIK3CA mutations.

4 Discussion

In this work, we propose a computational method for discovery of functional mutation for each studied cancer patient. One basic question is what are functional mutations to cancer. Indeed, different tools have quite different definitions. In the context of this work, functional mutations are mutations that directly or indirectly lead to altered expression of genes on a large scale and that they are the main causal reasons for observed expression patterns of patients. Thus, patients with the same functional mutations are supposed to have similar expression profiles. Following this definition, we transfer identification of functional mutations of one patients to enrichment analysis to mutations in its neighboring patients. This strategy can be called as "guilt by association", like what we have done in our previous work [40]. By checking the published literature, we can find evidences to support this definition. For example, Langerod A et.al [41, 42] observed characteristic gene expression patterns with patients carrying TP53 mutations. Similar results are available for other cancer genes, such as BRAF in melanoma [24], BRCA1 [43] and PIK3CA in breast cancer [44]. In this work, we also observed preferential distribution of patients with the same mutated genes to sub-clusters by hierarchical clustering (see Figure 1).

The functional mutations are recovered by identifying enriched gene mutations in patients with similar expression profiles. Complexity of mutation profiles make it impossible to cluster all the patients with the same gene mutation into one cluster. Most of the pattern classification methods, such as k-means, support vector machine (SVM), do not consider the sub-structure of samples, which cannot ensure each patient to have the optimal neighbors. With our method, we used a simple methods to find neighbors for every patient one by one, which make it good at recovery of patterns of inner cliques. The utilization of statistical significance also makes this method more robust.

With our prediction, some somatic mutations are not predicted to be functional. It is natural to raise the questions what those non-functional mutations are. One explanation is that non-functional mutations fail to affect the protein function. As an evaluation, we checked protein structure stability between mutated genes with or without predicted function. The folding free energy changes was calculated for each mutation by PoPMuSic2.1 [45]. The energy differences were further evaluated for four examples: TP53, PIK3CA, MAP3K1 and EGFR. However, we only observed significant difference in case of TP53 at $p < 0.03$, which suggested functional mutations to TP53 were more likely to affect its protein structure. For others, no significant differences were observed, which may result from the low accuracy of computational prediction and complexity of mutation to protein function.

For non-functional mutations, another explanation is due to the existence of other stronger functional mutations, which veil the effects of other mutation. It is possible that mutations to one gene may provide cells with proliferation advantage but not enough for cancer, e.g. benign tumours. In this case, other functional mutations are necessary for further oncogenesis and expression patterns of cancer patients will mainly reflect the consequences of latter mutations. Taking TP53 as an example, we observed that 107 breast cancer patients with functional TP53 mutations also carried 17 PIK3CA somatic mutations while 79 patients with non-functional TP53 mutations carried 27 PIK3CA somatic mutations. The significance of ratio differences was at $p < 0.018$. This observation suggests that patients with functional TP53 mutation are less likely to carry PIK3CA mutations. If only considering the predicted functional PIK3CA mutations, no patient with functional TP53 mutations carried functional mutations to PIK3CA while 16 patients with non-functional

TP53 mutations also carried functional mutations to PIK3CA. By checking other functional mutations, all 79 patients with non-functional TP53 mutations had at least one functional mutation to other genes. In summary, these results suggest non-functional TP53 mutations may result from existence of other functional mutations.

Certainly, our method also faces the problem of sensitivity if not enough neighboring patients are available. Our method requires a minimum number of neighboring patients for enrichment analysis which is not always available. This is especially true for those mutations with low recurrence. With the progress of cancer genome sequencing projects, the number of patients with mutation information will increase, which will provide a better basis for functional mutation discovery.

Mutual exclusivity is one popular strategy used to recover driver mutations [13]. Based on the functional annotation with GO and KEGG terms, we checked this assumption with some well-known cancer genes. We observed success with mutual exclusivity in the patients with TP53, MAP3K1 and EGFR mutations. However, we also noticed one counter case with PIK3CA, which was involved in the PI3K signalling pathway [46]. Mutations to members of PI3K are more enriched in patients with PIK3CA mutations. This result suggests mutual exclusivity to be a good assumption in driver mutation discovery but also faces the possibility of failure.

References

[1] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan 2000.

[2] Cancer Genome Atlas Network . Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, Oct 2012.

[3] Philip J. Stephens, Patrick S. Tarpey, Helen Davies, Peter Van Loo, Chris Greenman, David C. Wedge, Serena Nik-Zainal, Sancha Martin, Ignacio Varela, Graham R. Bignell, Lucy R. Yates, Elli Papaemmanuil, David Beare, Adam Butler, Angela Cheverton, John Gamble, Jonathan Hinton, Mingming Jia, Alagu Jayakumar, David Jones, Calli Latimer, King Wai Lau, Stuart McLaren, David J. McBride, Andrew Menzies, Laura Mudie, Keiran Raine, Roland Rad, Michael Spencer Chapman, Jon Teague, Douglas Easton, Anita Langerd, Oslo Breast Cancer Consortium (O. S. B. R. E. A. C) , Ming Ta Michael Lee, Chen-Yang Shen, Benita Tan Kiat Tee, Bernice Wong Huimin, Annegien Broeks, Ana Cristina Vargas, Gulisa Turashvili, John Martens, Aquila Fatima, Penelope Miron, Suet-Feung Chin, Gilles Thomas, Sandrine Boyault, Odette Mariani, Sunil R. Lakhani, Marc van de Vijver, Laura van 't Veer, John Foekens, Christine Desmedt, Christos Sotiriou, Andrew Tutt, Carlos Caldas, Jorge S. Reis-Filho, Samuel A J R. Aparicio, Anne Vincent Salomon, Anne-Lise Brresen-Dale, Andrea L. Richardson, Peter J. Campbell, P Andrew Futreal, and Michael R. Stratton. The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403):400–404, Jun 2012.

[4] Cancer Genome Atlas Network . Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407):330–337, Jul 2012.

[5] Bert Vogelstein, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A Diaz, Jr, and Kenneth W. Kinzler. Cancer genome landscapes. *Science*, 339(6127):1546–1558, Mar 2013.

[6] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R. Stratton. A census of human cancer genes. *Nat Rev Cancer*, 4(3):177–183, Mar 2004.

[7] Michael R. Stratton, Peter J. Campbell, and P Andrew Futreal. The cancer genome. *Nature*, 458(7239):719–724, Apr 2009.

[8] Xose S. Puente, Magda Pinyol, Vctor Quesada, Laura Conde, Gonzalo R. Ordez, Neus Villamor, Georgia Escaramis, Pedro Jares, Slvia Be, Marcos Gonzlez-Daz, Laia Bassaganyas, Tycho Baumann, Manel Juan, Mnica Lpez-Guerra, Dolors Colomer, Jos M C. Tubo, Cristina Lpez, Alba Navarro, Cristian Tornador, Marta Aymerich, Mara Rozman, Jess M. Hernndez, Diana A. Puente, Jos M P. Freije, Gloria Velasco, Ana Gutirrez-Fernndez, Dolors Costa, Anna Carri, Sara Guijarro, Anna Enjuanes, Llus Hernndez, Jordi Yage, Pilar Nicols, Carlos M. Romeo-Casabona, Heinz Himmelbauer, Ester Castillo, Juliane C. Dohm, Silvia de Sanjos, Miguel A. Piris, Enrique de Alava, Jess San Miguel, Romina Royo, Josep L. Gelp, David Torrents, Modesto Orozco, David G. Pisano, Alfonso Valencia, Roderic Guig, Mnica Bays, Simon Heath, Marta Gut, Peter Klatt, John Marshall, Keiran Raine, Lucy A. Stebbings, P Andrew Futreal, Michael R. Stratton, Peter J. Campbell, Ivo Gut, Armando Lpez-Guillermo, Xavier Estivill, Emili Montserrat, Carlos Lpez-Otn, and Elas Campo. Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, 475(7354):101–105, Jul 2011.

[9] Martin Peifer, Lynnette Fernndez-Cuesta, Martin L. Sos, Julie George, Danila Seidel, Lawryn H. Kasper, Dennis Plenker, Frauke Leenders, Ruping Sun, Thomas Zander, Roopika Menon, Mirjam Koker, Ilona Dahmen, Christian Mller, Vincenzo Di Cerbo, Hans-Ulrich Schildhaus, Janine Altmller, Ingelore Baessmann, Christian Becker, Bram de Wilde, Jo Vandesompele, Diana Bhm, Sascha Ansn, Franziska Gabler, Ines Wilkening, Stefanie Heynck, Johannes M. Heuckmann, Xin Lu, Scott L. Carter, Kristian Cibulskis, Shantanu Banerji, Gad Getz, Kwon-Sik Park, Daniel Rauh, Christian Grtter, Matthias Fischer, Laura Pasqualucci, Gavin Wright, Zoe Wainer, Prudence Russell, Iver Petersen, Yuan Chen, Erich Stoelben, Corinna Ludwig, Philipp Schnabel, Hans Hoffmann, Thomas Muley, Michael Brockmann, Walburga Engel-Riedel, Lucia A. Muscarella, Vito M. Fazio, Harry Groen, Wim Timens, Hannie Sietsma, Erik Thunnissen, Egbert Smit, Danille A M. Heideman, Peter J F. Snijders, Federico Cappuzzo, Claudia Ligorio, Stefania Damiani, John Field, Steinar Solberg, Odd Terje Brustugun, Marius Lund-Iversen, Jrg Snger, Joachim H. Clement, Alex Soltermann, Holger Moch, Walter Weder, Benjamin Solomon, Jean-Charles Soria, Pierre Validire, Benjamin Besse, Elisabeth Brambilla, Christian Brambilla, Sylvie Lantuejoul, Philippe Lorimier, Peter M. Schneider, Michael Hallek, William Pao, Matthew Meyerson, Julien Sage, Jay Shendure, Robert Schneider, Reinhard Bttner, Jrgen Wolf, Peter Nrnberg, Sven Perner, Lukas C. Heukamp, Paul K. Brindle, Stefan Haas, and Roman K. Thomas. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat Genet*, 44(10):1104–1110, Oct 2012.

[10] Shantanu Banerji, Kristian Cibulskis, Claudia Rangel-Escareno, Kristin K. Brown, Scott L. Carter, Abbie M. Frederick, Michael S. Lawrence, Andrey Y. Sivachenko, Carrie Sougnez, Lihua Zou, Maria L. Cortes, Juan C. Fernandez-Lopez, Shouyong Peng, Kristin G. Ardlie, Daniel Auclair, Veronica Bautista-Pia, Fujiko Duke, Joshua Francis, Joonil Jung, Antonio Maffuz-Aziz, Robert C. Onofrio, Melissa Parkin, Nam H. Pho, Valeria Quintanar-Jurado, Alex H. Ramos, Rosa Rebollar-Vega, Sergio Rodriguez-Cuevas, Sandra L. Romero-Cordoba, Steven E. Schumacher, Nicolas Stransky, Kristin M. Thompson, Laura Uribe-Figueroa, Jose Baselga, Rameen Beroukhim, Kornelia Polyak, Dennis C. Sgroi, Andrea L. Richardson, Gerardo Jimenez-Sanchez, Eric S. Lander, Stacey B. Gabriel, Levi A. Garraway, Todd R. Golub, Jorge Melendez-Zajgla, Alex Toker, Gad Getz, Alfredo Hidalgo-Miranda, and Matthew Meyerson. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, 486(7403):405–409, Jun 2012.

[11] Fabio Vandin, Eli Upfal, and Benjamin J. Raphael. Finding driver pathways in cancer: models and algorithms. *Algorithms Mol Biol*, 7(1):23, 2012.

[12] Alireza Hadj Khodabakhshi, Anthony P. Fejes, Inanc Birol, and Steven Jm Jones. Identifying cancer mutation targets across thousands of samples: Muteproc, a high throughput mutation analysis pipeline. *BMC Bioinformatics*, 14(1):167, May 2013.

[13] Giovanni Ciriello, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res*, 22(2):398–406, Feb 2012.

[14] Giovanni Ciriello, Ethan Cerami, Bulent Arman Aksoy, Chris Sander, and Nikolaus Schultz. Using memo to discover mutual exclusivity modules in cancer. *Curr Protoc Bioinformatics*, Chapter 8:Unit 8.17, Mar 2013.

[15] Fabio Vandin, Eli Upfal, and Benjamin J. Raphael. De novo discovery of mutated driver pathways in cancer. *Genome Res*, 22(2):375–385, Feb 2012.

[16] Mark D M. Leiserson, Dima Blokh, Roded Sharan, and Benjamin J. Raphael. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput Biol*, 9(5):e1003054, May 2013.

[17] Abel Gonzalez-Perez and Nuria Lopez-Bigas. Functional impact bias reveals cancer drivers. *Nucleic Acids Res*, 40(21):e169, Nov 2012.

[18] Linh M. Tran, Bin Zhang, Zhan Zhang, Chunsheng Zhang, Tao Xie, John R. Lamb, Hongyue Dai, Eric E. Schadt, and Jun Zhu. Inferring causal genomic alterations in breast cancer using gene expression data. *BMC Syst Biol*, 5:121, 2011.

[19] David L. Masica and Rachel Karchin. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res*, 71(13):4550–4561, Jul 2011.

[20] K. G. Wiman. Pharmacological reactivation of mutant p53: from protein structure to the cancer patient. *Oncogene*, 29(30):4245–4252, Jul 2010.

[21] Hanka Venselaar, Tim A H. Te Beek, Remko K P. Kuipers, Maarten L. Hekkelman, and Gert Vriend. Protein structure analysis of mutations causing inheritable diseases. an e-science approach with life scientist friendly interfaces. *BMC Bioinformatics*, 11:548, 2010.

[22] Yves Dehouck, Aline Grosfils, Benjamin Folch, Dimitri Gilis, Philippe Bogaerts, and Marianne Rooman. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: Popmusic-2.0. *Bioinformatics*, 25(19):2537–2543, Oct 2009.

[23] Emidio Capriotti, Piero Fariselli, and Rita Casadio. I-mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res*, 33(Web Server issue):W306–W310, Jul 2005.

[24] Peter Johansson, Sandra Pavey, and Nicholas Hayward. Confirmation of a braf mutation-associated gene expression signature in melanoma. *Pigment Cell Res*, 20(3):216–221, Jun 2007.

[25] Timothy R. Rebbeck, Angela DeMichele, Teo V. Tran, Saarene Panossian, Greta R. Bunin, Andrea B. Troxel, and Brian L. Strom. Hormone-dependent effects of fgfr2 and map3k1 in breast cancer susceptibility in a population-based sample of post-menopausal african-american and european-american women. *Carcinogenesis*, 30(2):269–274, Feb 2009.

[26] Haixin Lei, Sara Sjberg-Margolin, Sima Salahshor, Barbro Werelius, Eva Jandkov, Kari Hemminki, Annika Lindblom, and Igor Vorechovsk. Cdh1 mutations are present in both ductal and lobular breast cancer, but promoter allelic variants show no detectable breast cancer risk. *Int J Cancer*, 98(2):199–204, Mar 2002.

[27] Wei Yan, Qing Jackie Cao, Richard B. Arenas, Brooke Bentley, and Rong Shao. Gata3 inhibits breast cancer metastasis through the reversal of epithelial-mesenchymal transition. *J Biol Chem*, 285(18):14042–14051, Apr 2010.

[28] S. Labeit and B. Kolmerer. Titins: giant proteins in charge of muscle ultrastructure and elasticity. *Science*, 270(5234):293–296, Oct 1995.

[29] Christiane A. Opitz, Michael Kulke, Mark C. Leake, Ciprian Neagoe, Horst Hinssen, Roger J. Hajjar, and Wolfgang A. Linke. Damped elastic recoil of the titin spring in myofibrils of human myocardium. *Proc Natl Acad Sci U S A*, 100(22):12688–12693, Oct 2003.

[30] S. G. Brodie and C. X. Deng. Brca1-associated tumorigenesis: what have we learned from knockout mice? *Trends Genet*, 17(10):S18–S22, Oct 2001.

[31] A. J. Alberg and K. J. Helzlsouer. Epidemiology, prevention, and early detection of breast cancer. *Curr Opin Oncol*, 9(6):505–511, Nov 1997.

[32] Chu-Xia Deng and Rui-Hong Wang. Roles of brca1 in dna damage repair: a link between development and cancer. *Hum Mol Genet*, 12 Spec No 1:R113–R123, Apr 2003.

[33] Quan Zhu, Gerald M. Pao, Alexis M. Huynh, Hoonkyo Suh, Nina Tonnu, Petra M. Nederlof, Fred H. Gage, and Inder M. Verma. Brca1 tumour suppression occurs via heterochromatin-mediated silencing. *Nature*, 477(7363):179–184, Sep 2011.

[34] Pengfei Lu, Valerie M. Weaver, and Zena Werb. The extracellular matrix: a dynamic niche in cancer progression. *J Cell Biol*, 196(4):395–406, Feb 2012.

[35] W. G. Stetler-Stevenson, S. Aznavoorian, and L. A. Liotta. Tumor cell interactions with the extracellular matrix during invasion and metastasis. *Annu Rev Cell Biol*, 9:541–573, 1993.

[36] A. J. Levine. p53, the cellular gatekeeper for growth and division. *Cell*, 88(3):323–331, Feb 1997.

[37] G. Pearson, F. Robinson, T. Beers Gibson, B. E. Xu, M. Karandikar, K. Berman, and M. H. Cobb. Mitogen-activated protein (map) kinase pathways: regulation and physiological functions. *Endocr Rev*, 22(2):153–183, Apr 2001.

[38] Alyssa L. Kennedy, Jennifer P. Morton, Indrani Manoharan, David M. Nelson, Nigel B. Jamieson, Jeff S. Pawlikowski, Tony McBryan, Brendan Doyle, Colin McKay, Karin A. Oien, Greg H. Enders, Rugang Zhang, Owen J. Sansom, and Peter D. Adams. Activation of the pik3ca/akt pathway suppresses senescence induced by an activated ras oncogene to promote tumorigenesis. *Mol Cell*, 42(1):36–49, Apr 2011.

[39] D Williams Parsons, Tian-Li Wang, Yardena Samuels, Alberto Bardelli, Jordan M. Cummins, Laura DeLong, Natalie Silliman, Janine Ptak, Steve Szabo, James K V. Willson, Sanford Markowitz, Kenneth W. Kinzler, Bert Vogelstein, Christoph Lengauer, and Victor E. Velculescu. Colorectal cancer: mutations in a signalling pathway. *Nature*, 436(7052):792, Aug 2005.

[40] Guofeng Meng and Martin Vingron. Condition-specific target prediction from motifs and expression. *Bioinformatics*, 30(12):1643–1650, 2014.

[41] Anita Langerd, Hongjuan Zhao, rnulf Borgan, Jahn M. Nesland, Ida R K. Bukholm, Tone Ikdahl, Rolf Kresen, Anne-Lise Brresen-Dale, and Stefanie S. Jeffrey. Tp53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. *Breast Cancer Res*, 9(3):R30, 2007.

[42] Lance D. Miller, Johanna Smeds, Joshy George, Vinsensius B. Vega, Liza Vergara, Alexander Ploner, Yudi Pawitan, Per Hall, Sigrid Klaar, Edison T. Liu, and Jonas Bergh. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc Natl Acad Sci U S A*, 102(38):13550–13555, Sep 2005.

[43] Katarzyna Marta Lisowska, Volha Dudaladava, Michal Jarzab, Tomasz Huzarski, Ewa Chmielik, Ewa Stobiecka, Jan Lubinski, and Barbara Jarzab. Brca1-related gene signature in breast cancer: the role of er status and molecular type. *Front Biosci (Elite Ed)*, 3:125–136, 2011.

[44] Magdalena Cizkova, Graldine Cizeron-Clairac, Sophie Vacher, Aurlie Susini, Catherine Andrieu, Rosette Lidereau, and Ivan Biche. Gene expression profiling reveals new aspects of pik3ca mutation in eralpha-positive breast cancer: major implication of the wnt signaling pathway. *PLoS One*, 5(12):e15647, 2010.

[45] Yves Dehouck, Jean Marc Kwasigroch, Dimitri Gilis, and Marianne Rooman. Popmusic 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics*, 12:151, 2011.

[46] Amancio Carnero, Carmen Blanco-Aparicio, Oliver Renner, Wolfgang Link, and Juan F M. Leal. The pten/pi3k/akt signalling pathway in cancer, therapeutic implications. *Curr Cancer Drug Targets*, 8(3):187–198, May 2008.