Hum Genet (2013) 132:15–27
DOI 10.1007/s00439-012-1218-7

ORIGINAL INVESTIGATION

A Bayesian ensemble approach with a disease gene network predicts damaging effects of missense variants of human cancers

Hong-Hee Won · Jong-Won Kim · Doheon Lee

Received: 15 May 2012 / Accepted: 5 August 2012 / Published online: 21 August 2012

Abstract Large-scale sequencing of cancer genomes has revealed many novel mutations and inter-tumoral heterogeneity. Therefore, prioritizing variants according to their potential deleterious effects has become essential. We constructed a disease gene network and proposed a Bayesian ensemble approach that integrates diverse sources to predict the functional effects of missense variants. We analyzed 23,336 missense disease mutations and 36,232 neutral polymorphisms of 12,039 human proteins. The results showed successful improvement of prediction accuracy in both sensitivity and specificity, and we demonstrated the utility of the method by applying it to somatic mutations obtained from colorectal and breast cancer cell lines. The candidate genes with predicted deleterious mutations as well as known cancer genes were significantly enriched in many KEGG pathways related to carcinogenesis, supporting genetic homogeneity of cancer at the pathway level. The breast cancer-specific network increased the prediction accuracy for breast cancer mutations. This study provides a ranked list of deleterious mutations and candidate cancer genes and suggests that mutations affecting cancer may occur in important pathways and should be interpreted on the phenotype-related network or pathway. A disease gene network may be of value in predicting functional effects of novel disease-specific mutations.

Electronic supplementary material The online version of this article (doi:10.1007/s00439-012-1218-7) contains supplementary material, which is available to authorized users.

H.-H. Won · D. Lee (✉)
Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-710, South Korea
e-mail: dhlee@kaist.ac.kr

H.-H. Won
e-mail: honghee.won@gmail.com

H.-H. Won
Samsung Biomedical Research Institute, Samsung Medical Center, Sungkyunkwan University School of Medicine, 50 Ilwon-dong, Gangnam-gu, Seoul 135-710, South Korea

J.-W. Kim (✉)
Department of Laboratory Medicine and Genetics, Samsung Medical Center, Sungkyunkwan University School of Medicine, 50 Ilwon-dong, Gangnam-gu, Seoul 135-710, South Korea
e-mail: culture.jkim@gmail.com

Introduction

The recent advances in high-throughput DNA sequencing technology have enabled whole genome or exome sequencing for hundreds of patients. In addition, by major initiatives, such as the Human Variome Project, the 1000 Genomes Project, and the International Cancer Genome Consortium, a vast amount of variation data have been generated. Thus, many novel mutations, which change an amino acid of the corresponding protein and possibly affect its phenotype, are expected to be found in patients with various cancers or other genetic diseases. Several studies have reported a number of sequence variations in human cancers (Campbell et al. 2008; Greenman et al. 2007; Jones et al. 2008, 2010; Sjoblom et al. 2006), in which mutational patterns have differed greatly between patients with the same disease. This heterogeneity implies that potentially different driver mutations may be responsible for a tumor cell growth advantage during carcinogenesis or tumor progression from one patient to the next, which is likely to present a challenge to personalized medicine (Swanton et al. 2011). Accordingly, it has become essential to

determine how to interpret such a large number of novel sequence variants found in cancer patients. In particular, an accurate and rapid distinction of neutral polymorphisms and cancer-causing mutations is one of the most important issues.

Many different algorithms have been developed to provide in silico prediction for novel variants (Adzhubei et al. 2010; Carter et al. 2009; Fischer et al. 2011; Kaminker et al. 2007; Ng and Henikoff 2003, 2006; Schwarz et al. 2010; Shi and Moult 2011; Thomas et al. 2003; Torkamani and Schork 2008). However, the prediction performance of the methods still needs to be improved. While most current methods utilize local information of a mutation such as physicochemical properties, sequence conservation, and structure homology for feature information, underlying clinical phenotype information is yet to be considered. For example, when predicting novel mutations found in cancer patients, the current methods do not take into account the clinical characteristic of cancer, which may be informative for predicting the functional consequences of the variants. Even though different mutations may show a similar amount of change in the corresponding protein, the functional effects of the mutations can be diverse according to the relative importance of the target gene in the context of clinical phenotype or disease. A feasible approach to determining the importance of a certain gene regarding a pathological condition could be to utilize the protein's interactions with known disease genes. It was previously shown that mutations of interacting proteins are likely to lead to similar disease phenotypes (Gandhi et al. 2006). By merging a network of known cancer genes with human protein–protein interactions and calculating a level of closeness with the cancer genes, we sought to measure the importance of a particular gene with regard to a possible association with cancer. On the other hand, diverse somatic mutation patterns were observed across different human cancers (Kan et al. 2010), and in order for genes to play a critical role in the development of disease in a particular tissue, their protein product should be expressed in that tissue (Morton 2004; Williamson et al. 2008). Therefore, we also constructed a specific cancer sub-type gene network, which utilized gene expression profiles, and evaluated the performance.

Since we showed in a previous study that a combination approach can successfully increase prediction accuracy, we predicted that merging the combination approach and clinical background information for each mutation would improve prediction performance (Won et al. 2008). In this study, we propose a Bayesian ensemble approach that integrates multiple predictors, protein interaction networks, known cancer genes, and tissue expression profiles, to predict the phenotypic effect of sequence variants of human cancers and evaluate the prediction performance. To investigate the assumption that utilizing more prior knowledge may enhance the prediction power, we compared the performance of the Bayesian ensemble model with protein–protein interactions and known cancer genes in OMIM (a general cancer network), and the one also including available tissue expression profiles (a breast cancer specific network). By applying the proposed method to sequence variants identified from previous studies, we examined potential candidate cancer genes and mutations that might be associated with cancers in humans.

Materials and methods

Reference database

The human polymorphisms and disease mutations were retrieved from the Swiss-Prot database. Classification of polymorphisms and disease mutations was made according to literature reports on probable disease-association. Swiss-Prot 57.12 (http://www.uniprot.org/docs/humsavar.txt) contained a total of 61,565 variants including 23,336 disease variants, 36,232 polymorphisms, and 1,997 unclassified variants of 12,039 human gene proteins or hypothetical proteins. The proteins were classified in terms of molecular function according to the PANTHER protein library (Thomas et al. 2003). For redundant variants associated with several diseases, only one variant was included in the analysis. Of the disease variants, 4,711 variants were known to be associated with human cancers.

Tools for predicting phenotypic effects of amino acid substitution

We used three representative in silico programs, PolyPhen-2 (Adzhubei et al. 2010), SIFT (Ng and Henikoff 2003), and PANTHER SNP scoring tool (Thomas et al. 2003), to predict the phenotypic effect of an amino acid-substituting variant. The three predictors use different algorithms and thus may give a different prediction for the same variant. To set up SIFT and PANTHER locally, we downloaded SIFT (Linux version 4.0.2) from http://sift.jcvi.org, PANTHER SNP scoring tool (Linux version 1.01) from http://www.pantherdb.org/downloads and SWISSPROT from ftp://ftp.ncbi.nih.gov/blast/db/FASTA. As a reference protein database, we used Swiss-Prot for SIFT and PANTHER 6.1 for PANTHER. We executed PolyPhen-2 online (http://genetics.bwh.harvard.edu/pph2) through batch processing and summarized the results. Score thresholds to distinguish damaging variants from neutral polymorphisms were set according to the suggested value in each program. All statistical analyses were performed using R 2.9.1 statistical software (http://www.r-project.org).

Extended cancer gene network

The human protein–protein interaction network was constructed from the BioGRID, DIP, HPRD, and MINT databases (Supplemental Table 1). It is composed of a total of 54,965 non-redundant interactions between 11,713 human proteins. The list of 811 known cancer genes was retrieved from the Online Mendelian Inheritance in Men (OMIM) Morbid Map (http://www.omim.org/downloads) and merged with the human protein network. To determine the relative importance of the gene from an input variant, we used the shortest distance between the gene and its nearest known cancer genes (closeness to cancer genes), which was calculated using the Dijkstra's algorithm (Dijkstra 1959).

Breast cancer gene network

Tissue expression profiles were retrieved from the ALEXA-seq data (http://www.alexaplatform.org/alexa_seq/Breast/Summary.htm), which were measured by mapping reads to transcript or genomic sequences and calculating the observed average coverage of the mapped reads (Griffith et al. 2010). The ALEXA-seq data contained a large number of expressed sequence reads and showed a total of 13,546 genes, which were expressed in normal luminal epithelial and myoepithelial breast tissues and human mammary epithelial cells. We obtained the breast cancer gene network by removing all interactions of genes that are not expressed in normal breast tissues. Subsequently, the breast cancer gene network is composed of 36,824 interactions between 7,163 human proteins.

A Bayesian ensemble of multiple predictors and cancer gene network

By Bayes' theorem, the posterior probability measure $P(\mathrm{H}_1|d)$ of a variable given the data value (d) is the product of the prior probability measure $P(\mathrm{H}_1)$ and the likelihood function $P(d|\mathrm{H}_1)$ divided by $P(d)$.

$$p(\mathrm{H}_1|d) = \frac{p(d|\mathrm{H}_1) \times \mathrm{p}(\mathrm{H}_1)}{p(d)}.$$

According to the law of total probability, $P(d)$, which is the prior or marginal probability of d and acts as a normalizing constant, can be replaced by the equivalent form as follows:

$$p(d) = \sum_i p(d \cap \mathrm{H}_i) = \sum_i p(d|\mathrm{H}_i) \times p(\mathrm{H}_i).$$

Therefore, the posterior probability distribution can be calculated with Bayes' theorem by multiplying the prior probability distribution by the likelihood function and then dividing by the normalizing constant, as follows:

$$p(\mathrm{H}_1|d) = \frac{p(d|\mathrm{H}_1) \times p(\mathrm{H}_1)}{\sum_i p(d|\mathrm{H}_i) \times p(\mathrm{H}_i)}$$

where the sum over all possible mutually exclusive hypotheses should be 1.

In our problem, the null hypothesis H_0 is that a variant is neutral and the alternative hypothesis H_1 is that the variant is deleterious. One of the main ideas is that variants of known cancer genes or their neighboring genes with close interactions are more likely to have a deleterious effect on cancer development than those of the genes with distant or no interactions, suggesting that the prior probability, $P(\mathrm{H}_1)$, that a variant will be deleterious may be different according to the relative importance of the corresponding gene on the cancer gene network. We can calculate deleterious scores (s_1, s_2, …, s_n) of the variant using the three prediction programs ($n = 3$) and compute the empirical score distributions (likelihood function) of each of the three programs for deleterious mutations and neutral polymorphisms. Likelihood function $P(s_i|\text{deleterious})$ can be calculated using the score distribution of the 23,336 disease variants, and $P(s_i|\text{neutral})$ can be calculated using the 36,232 polymorphisms. Therefore, the posterior probability $P(\text{deleterious}|s_i)$ that the variant will be deleterious given the scores can be well calculated with Bayes' theorem.

A composite likelihood statistic described above as the posterior probability $P(\text{deleterious}|s_i)$ was used and the cancer-associated missense mutation (CAM) score was defined as follows:

$$\begin{aligned}\mathrm{CAM} &= \prod_{i=1}^{n} p(\text{deleterious}|s_i) \\ &= \prod_{i=1}^{n} \frac{p(s_i|\text{deleterious}) \times \pi}{p(s_i|\text{deleterious} \times \pi + p(s_i|\text{neutral}) \times (1-\pi)}.\end{aligned}$$

The composite likelihood statistic was previously used to predict causal variants of positive selection and showed a good prediction performance (Grossman et al. 2010). The prior probability $P(\mathrm{H}_1)$ of the variant being damaging was defined as a function of closeness to cancer genes $\pi = (9 - d)/10$ where d is the shortest distance (number of links) to the nearest cancer gene on the cancer gene network. Accordingly, the prior probability $P(\mathrm{H}_0)$ that the variant is not deleterious (neutral) is defined as $1 - \pi$. Because d is 0 for the known cancer genes and the observed farthest distance is 8 on the network (Supplemental Figure 1), the prior probability π increases from 0.1 to 0.9 continuously as the gene of interest gets closer to the cancer genes. Since we constructed two network models including general cancer gene network and sub-type cancer gene network, π value of each gene can be

different according to the network. If there is no link between the gene and any cancer genes and thus d is not determined, π is given as the proportion of disease mutations out of all the variants of the genes disconnected to the known cancer genes in our data (=522/14,946). Therefore, the CAM scores depend not only on the relative probability of a variant being deleterious, but also on the association of the corresponding gene to cancer. Because the CAM score is the approximate posterior probability that the variant is deleterious, one can prioritize potentially deleterious variants based on their scores. The source code written in C and Perl is freely available to academic users at https://sourceforge.net/projects/cdv/files.

Performance evaluation

To compare the prediction performance of the proposed method and other existing methods, we constructed a threshold-independent receiver operating characteristics (ROC) curve. Based on the ROC curve, a numeric measure, area under the ROC (AUROC), was estimated by the R software. Other measures including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), correlation coefficient, true positive cost, and overall accuracy were also used to evaluate prediction performance. True positive (TP) is the number of cancer mutations correctly predicted, false positive (FP) is the number of neutral polymorphisms incorrectly predicted, true negative (TN) is the number of neutral polymorphisms correctly predicted, and false negative (FN) is the number of cancer mutations incorrectly predicted. The definitions are as follows: sensitivity = TP/(TP + FN), specificity = TN/(TN + FP), PPV = TP/(TP + FP), NPV = TN/(TN + FN), correlation coefficient = $(TP \times TN - FP \times FN)/((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN))^{1/2}$, true positive cost = FP/TP, and overall accuracy = (TP + TN)/(TP + TN + FP + FN). For genes with predicted damaging mutations, we conducted literature review for hundreds of published papers and examined the association of the candidate genes with cancers.

Pathway enrichment analysis

To investigate the relevance of genes with mutations predicted to be deleterious with carcinogenesis, we performed gene set analyses using the WebGestalt (WEB-based GEne SeT AnaLysis Toolkit) program (http://bioinfo.vanderbilt.edu/webgestalt) (Zhang et al. 2005). We selected significantly enriched Kyoto encyclopedia of genes and genomes (KEGG) pathways using the hypergeometric test, and P values were adjusted by the Benjamini and Hochberg false discovery rate-controlling procedure.

Results

Prediction for cancer mutations and neutral variants

The overall analysis flow is summarized in Fig. 1. As a first step, we calculated prediction scores of individual predictors, PolyPhen-2, SIFT, and PANTHER for the 23,336 disease variants and 36,232 polymorphisms (see "Materials and methods").

The prediction performance of the predictors was evaluated for 3,545 cancer mutations and 22,531 neutral variants whose functional consequences were predicted by all the individual predictors to combine their results (Fig. 2a). According to the binary classification of PolyPhen-2 (neutral for PolyPhen-2 score ≤ 0.2 or deleterious for score > 0.2), among the 3,545 cancer mutations predicted, 1,012 (28.5 %) were predicted to be neutral and 2,533 (79.7 %) to be deleterious. For the 22,531 polymorphisms, PolyPhen-2 predicted 9,283 (41.2 %) to be deleterious and 13,248 (58.8 %) to be neutral. Among the cancer mutations, 2,209 (62.3 %) were predicted to be damaging by SIFT (SIFT score ≤ 0.05) and 1,336 (37.7 %) to be neutral. Among the polymorphisms, 7,382 (32.8 %) were predicted to be damaging and 15,149 (67.2 %) to be neutral. The 1,933 (54.5 %) cancer mutations were predicted to be damaging by PANTHER (sub-PSEC score ≤ -3), and 1,612 (45.5 %) were predicted to be neutral. PANTHER predicted 5,687 (25.2 %) polymorphisms to be damaging and 16,844 (74.8 %) to be neutral. Noticeably, only a small fraction of both cancer mutations (27.3 %) and polymorphisms (22.2 %) were incorrectly predicted by the proposed CAM method, while the individual predictors predicted many polymorphisms to be damaging (false positives) in PolyPhen-2 and many cancer mutations to be neutral (false negatives) in PANTHER (Fig. 2a). Distribution of the CAM scores of cancer mutations and neutral polymorphisms was significantly different (Wilcoxon rank sum test, $P < 2.2 \times 10^{-9}$) (Supplemental Figure 2).

According to the refined threshold based on the ROC curve, PolyPhen-2 was superior to the other individual predictors regarding sensitivity, while SIFT showed a better performance regarding specificity (Fig. 2b, Supplemental Table 2). Although an ensemble approach (Ensemble) with a uniform prior probability outperformed all the individual predictors, the proposed CAM method was superior to the Ensemble. We also compared the prediction performance of the predictors using several measures under the condition to achieve a sensitivity of 80 % (Table 1). While predicting 80 % of the cancer mutations correctly, the individual predictors predicted <50 % of the neutral variants correctly with a low positive predictive value and a high true positive cost. Of the individual predictors, PolyPhen-2 showed the best

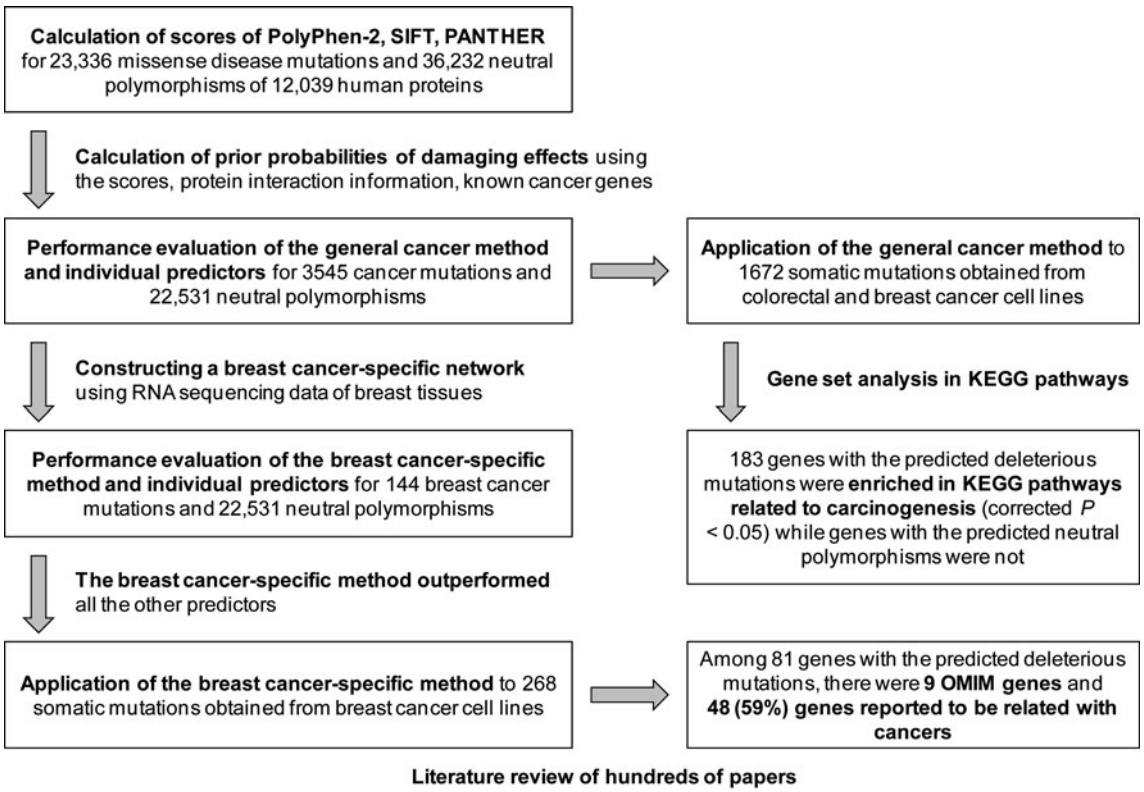

Fig. 1 Flowchart of overall analysis

prediction results. The CAM method showed the best performance regarding all the performance measures.

When we categorized 3,545 cancer mutations and 22,531 neutral variants according to the protein class of the corresponding gene, we observed that some protein classes showed a relatively high frequency of cancer mutations (Fig. 3a). For example, genes of the 'transcription factor' protein class showed the highest frequency of cancer mutations (47.2 %, 1,384/2,932) and several other classes also showed a high frequency: 34.5 % for 'ligase', 34.1 % for 'phosphatase', 29.3 % for 'kinase', 20.7 % for 'transferase', 16.9 % for 'nucleic acid binding', 14.7 % for 'lyase', 13.0 % for 'cytoskeletal protein', and 10.2 % for 'hydrolase'. The absolute number of cancer mutations was high in the 'transcription factor' (1,394 cancer mutations), 'transferase' (626), and 'kinase' (586) classes. SIFT showed a good performance for the 'transcription factor' and 'cytoskeletal protein' classes (Fig. 3b). PolyPhen-2 showed a good performance particularly for 'ligase' and was superior to the other two predictors for the overall classes. PANTHER showed a slightly higher AUROC compared to the others for the 'nucleic acid binding' class. The CAM method showed the best prediction performance compared with the individual predictors. In particular, it showed AUROC values of more than 0.85 for the classes with highly frequent cancer mutations such as 'transcription factor', 'ligase', and 'phosphatase'.

Application to somatic mutations in cancers

For 1,672 somatic mutations of human protein-coding genes that were previously identified in human cancers (Sjoblom et al. 2006), we predicted the functional consequence of the mutations using the CAM method and investigated the top 24 mutations that were predicted to be deleterious (CAM score > 0.9) (Supplemental Table 3 for the gene list of CAM score > 0.122 determined based on the ROC curve). Among the genes of the 24 mutations, four genes (*MPO* [p.R477Q, CAM = 0.986], *SMAD4* [p.P130S, CAM = 0.979], *CDS1* [p.K204T, CAM = 0.956], and *HNF1A* [p.K273E, CAM = 0.934]) were previously known to be related to colorectal or breast cancers and were registered in OMIM.

Although many other genes have not been registered as cancer genes for the corresponding cancer in OMIM, they were previously found to be associated with carcinogenesis or tumorigenesis in various other cancers or tumors. Copy number analysis showed that *RPS6KB1* (p.G289E, CAM = 0.969) was amplified and overexpressed in breast tumors and cell lines (Sinclair et al. 2003). Knockdown of contactin-1 (*CNTN1*) (p.P794H, CAM = 0.964) expression was shown to suppress invasion and metastasis of lung adenocarcinoma cells (Su et al. 2006). The 399-Gln allele of *XRCC1* (X-ray repair complementing defective repair in Chinese hamster cells) (p.R350W, CAM = 0.956) was

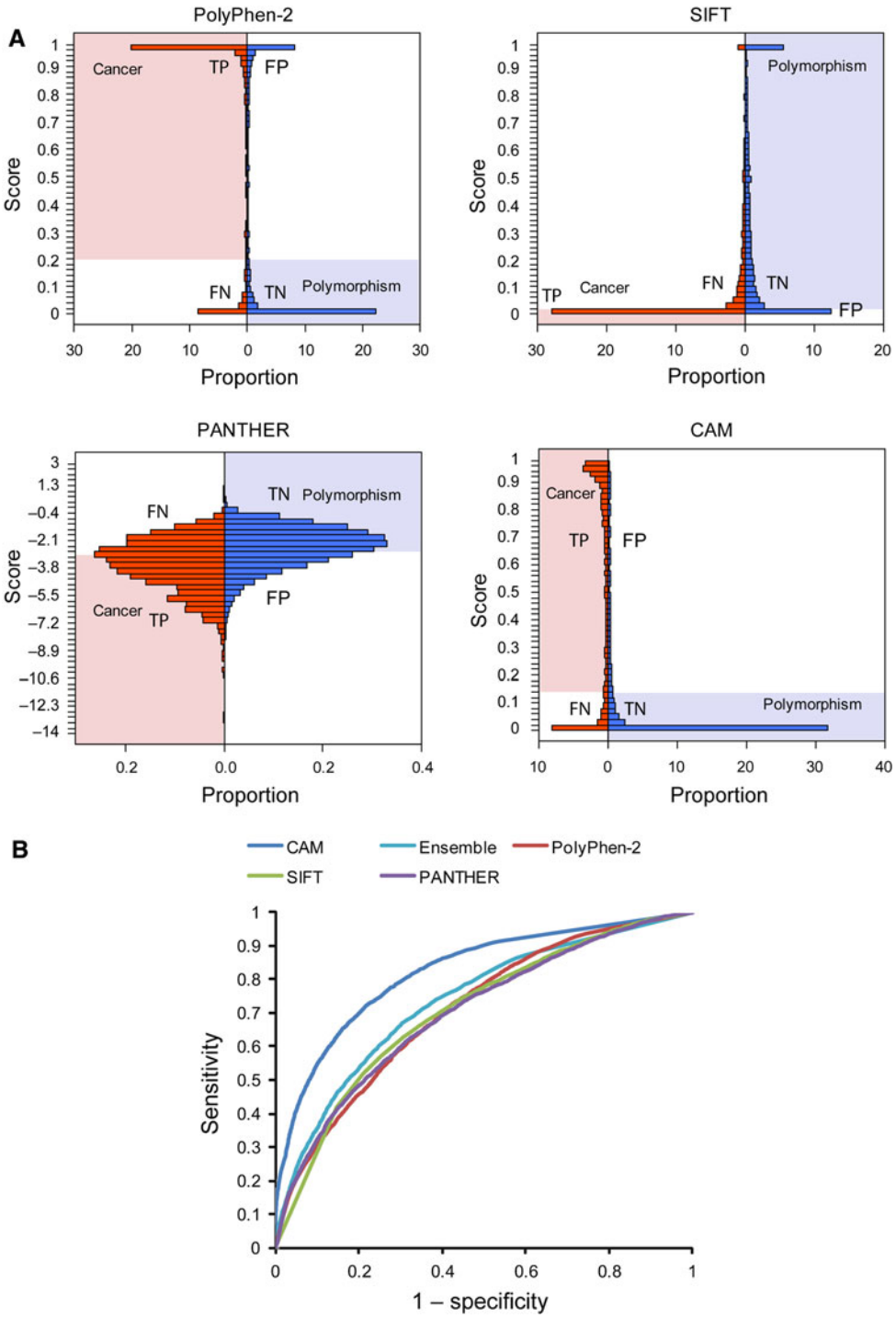

Fig. 2 The CAM method distinguishes cancer mutations from neutral polymorphisms better than the individual predictors. **a** Score distributions of cancer mutations and neutral polymorphisms. *Scores* denote a prediction probability for PolyPhen-2, SIFT and CAM, and a sub-PSEC score for PANTHER, respectively. True positives (TP) indicate cancer mutations predicted to be damaging (*red rectangles*), and true negatives (TN) indicate polymorphisms predicted to be neutral (*blue rectangles*). False positives (FP) and false negatives (FN) indicate incorrectly predicted cancer mutations and polymorphisms, respectively. **b** Prediction performance of the CAM method and the individual predictors. The CAM method outperformed an ensemble approach with a uniform prior probability (set as 0.5) as well as all the individual predictors (color figure online)

shown to significantly increase the risk of breast cancer in Caucasian women (Sterpone et al. 2010). *TGM2* (p.G660V, CAM = 0.948) was commonly hypermethylated in primary brain tumors (Dyer et al. 2011). Noticeably, *IDH1* (p.R132C, CAM = 0.945) mutation status was found to be correlated with the overexpression of known glioblastoma multiforme survival genes (Masica and Karchin 2011). Interestingly, *TNS4* (p.R642C, CAM = 0.943), also referred to as *CTEN*, was reported to be linked to tumor progression and various cancer types including thymoma and colorectal, breast, and prostate cancers (Albasri et al. 2009; Barbieri et al. 2010; Katz et al. 2007; Li et al. 2010; Liao et al. 2009; Lo and Lo 2002; Sasaki et al. 2003). *EPHA7* (p.R371W, CAM = 0.925) was down-regulated in colorectal cancer by promoter hypermethylation but was expressed at a substantial level in most human lung cancers

Table 1 Comparison of the prediction performance (at the sensitivity level of 0.8)

Performance measure	PolyPhen-2	SIFT	PANTHER	CAM
Sensitivity	0.800	0.798	0.800	0.801
Specificity	0.485	0.463	0.438	0.692
PPV	0.197	0.190	0.183	0.290
NPV	0.939	0.936	0.933	0.957
Correlation coefficient	0.197	0.181	0.166	0.349
True positive cost	4.088	4.274	4.461	2.445
Overall accuracy	0.528	0.509	0.487	0.707

PPV positive predictive value, *NPV* negative predictive value

(Tsuboi et al. 2010). A previous study found that tumor cells with down-regulation of *MAP3K6* (p.P869T, CAM = 0.924) expression showed significant suppression of tumor growth, and VEGF expression was regulated by *MAP3K6*, and suggested that it may play a crucial role in both angiogenesis and tumorigenesis (Eto et al. 2009). *AMPD1* (p.P633H, CAM = 0.920) was revealed to be associated with colorectal cancer risk in a previous single-nucleotide polymorphism study (Webb et al. 2006). *AKAP3* (p.R831C, CAM = 0.916) mRNA expression correlated with worse overall survival in patients with ovarian cancer (Sharma et al. 2005). Other top genes with predicted deleterious

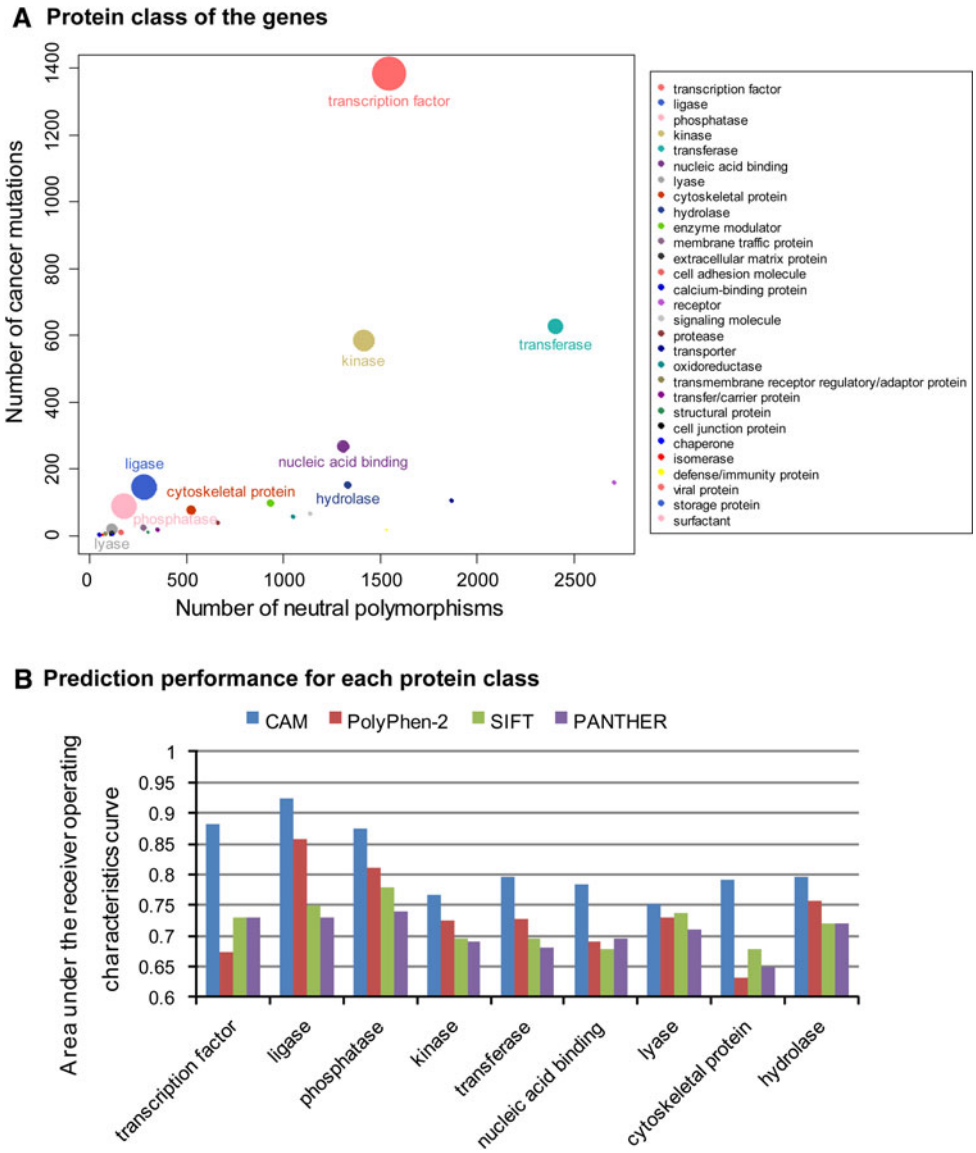

Fig. 3 Prediction performance according to protein class of the genes. **a** Each *circle* represents a group of genes for each protein class. Circle sizes are proportional to the frequency of cancer mutations among the total variants. **b** Prediction performance of the predictors is shown for the protein class with frequent cancer mutations

mutations (CAM > 0.9) such as *SIGLEC7* (p. L215P, CAM = 0.968), *KCNQ5* (p.R244C, CAM = 956), *SMAD2* (p.D300V, CAM = 0.955), *FCN1* (p.Y175C, CAM = 0.949), *SPLTC1* (p.R239W, CAM = 0.944), *G6PC* (p.P116L, CAM = 0.926), *LMNB2* (p.R216W, CAM = 0.908), and *ACADM* (p.P132R, CAM = 0.906), were not previously reported to be associated with cancer and need to be further investigated (Supplemental Table 3).

Gene set analyses revealed that the 183 genes with the predicted deleterious mutations listed in Supplemental Table 3 were enriched in many KEGG pathways related to carcinogenesis (corrected $P < 0.05$) (Supplemental Table 4). The KEGG pathways identified include various cancer pathways such as colorectal, thyroid, endometrial, pancreatic, prostate, and bladder cancers. The most enriched KEGG pathway was hsa05200, which is related to various important characteristics of cancer including sustained angiogenesis, evading apoptosis, proliferation, and insensitivity to anti-growth signals (Supplemental Figure 3). Noticeably, the Wnt signaling, JAK-STAT, ErbB signaling, TGF-beta signaling, and the VEGF signaling pathways were detected, which are well-known and important pathways related to carcinogenesis. We not only examined mutations in known cancer genes, but also analyzed mutations in previously unknown cancer genes, which interact with known cancer genes. Figure 4 shows that the Wnt signaling pathway includes several known cancer genes as well as some previously unknown genes with possible damaging mutations. Enrichment of the *LIG1*, *PARP1*, and *XRCC1* genes in the base excision repair pathway was also observed. In contrast, the 259 genes with predicted neutral polymorphisms were significantly enriched in the KEGG pathways which are less related to cancers, such as metabolic pathways, ABC transporters, Fc gamma R-mediated phagocytosis, fatty acid metabolism, circadian rhythm, lysosome, etc. (Supplemental Table 5). These results imply that deleterious genetic changes at the pathway level may be needed for the development of cancer and account for inter-tumoral heterogeneity at the gene level.

Prediction with a breast cancer gene network

Because different types of cancer show diverse mutational patterns according to their sub-type as well as frequent mutations in commonly identified genes, utilizing a more specific sub-cancer gene network might be beneficial in determining deleterious mutations affecting a particular sub-type of cancer. Genes playing a critical role in carcinogenesis of breast cancer are likely to be expressed in the breast or mammary gland tissues, while genes that are not expressed in those tissues are not likely to be involved in the carcinogenesis.

To examine the usefulness of breast tissue expression profiles in predicting breast cancer mutations, we constructed a breast cancer gene network, which is composed of genes expressed in normal breast tissues. The number of genes (11,713) in the general cancer gene network without expression information was reduced by 39 % in the breast cancer gene network (7,163), resulting in 33 % reduction of the interactions (from 54,965 to 36,824). The breast cancer-specific CAM method utilizing this reduced network distinguished the breast cancer mutations from the neutral polymorphisms more accurately than the general cancer CAM method and the individual predictors (Supplemental Figure 4). The AUROC was the highest in the breast cancer gene network (0.67 in PANTHER; 0.69 in SIFT; 0.70 in PolyPhen-2; 0.87 in the general cancer CAM method; and 0.90 in the breast cancer specific CAM method) (Fig. 5).

Using the breast cancer-specific CAM method, we prioritized 268 somatic mutations in breast cancer tissues or cell lines obtained from previous large-scale sequencing studies (Greenman et al. 2007; Sjoblom et al. 2006). By the threshold CAM score of 0.145 (the best performance determined based on the ROC curve), 81 mutations were predicted to be damaging (Supplemental Table 6).

The genes with CAM score > 0.145 and evidence of literature for their association with breast cancer are listed in Supplemental Table 7. Several top genes were reported to be related to breast cancer or other cancers. For example, the *LYN* gene with the highest CAM score (p.D385Y, CAM = 0.969) was previously identified as a potential target for dasatinib in breast cancer (Choi et al. 2010; Hochgrafe et al. 2010) and imatinib in chronic myeloid leukemia (Wu et al. 2008). High expression of *CSNK2A1* (p.D297H, CAM = 0.962) was predictive of a poor diagnosis in non-small cell lung cancer patients (Wang et al. 2010). Checkpoint kinase 2 (*CHK2*, *CDS1* [p.K204T, CAM = 0.956], *RAD53*) is activated by ataxia telangiectasia mutated (*ATM*) in response to gamma irradiation and mutations of the gene were found in various sporadic cancers (Miller et al. 2002). The human Cds1 kinase (hCds1/Chk2) regulated BRCA1 function after DNA damage by phosphorylating serine 988 of BRCA1 (Lee et al. 2000). *FGFR2* (p.R203C, CAM = 0.946) showed strong associations with breast cancer (Gaudet et al. 2010), and an *FGFR2*-IIIb-specific antibody exhibited antitumor activity (Bai et al. 2010). *TCF1* (p.K273E, CAM = 0.934) encoding hepatocyte nuclear factor 1 alpha (HNF1A) was mutated in endometrial tumors, but not in breast or ovarian tumors (Rebouissou et al. 2004). Somatic mutations of *MAP3K6* (p.P869T, CAM = 0.924) were observed in gastric cancer cell lines, and *MAP3K6* was recurrently altered in both primary tumors and cell lines (Zang et al. 2011). Many previous studies showed the association of

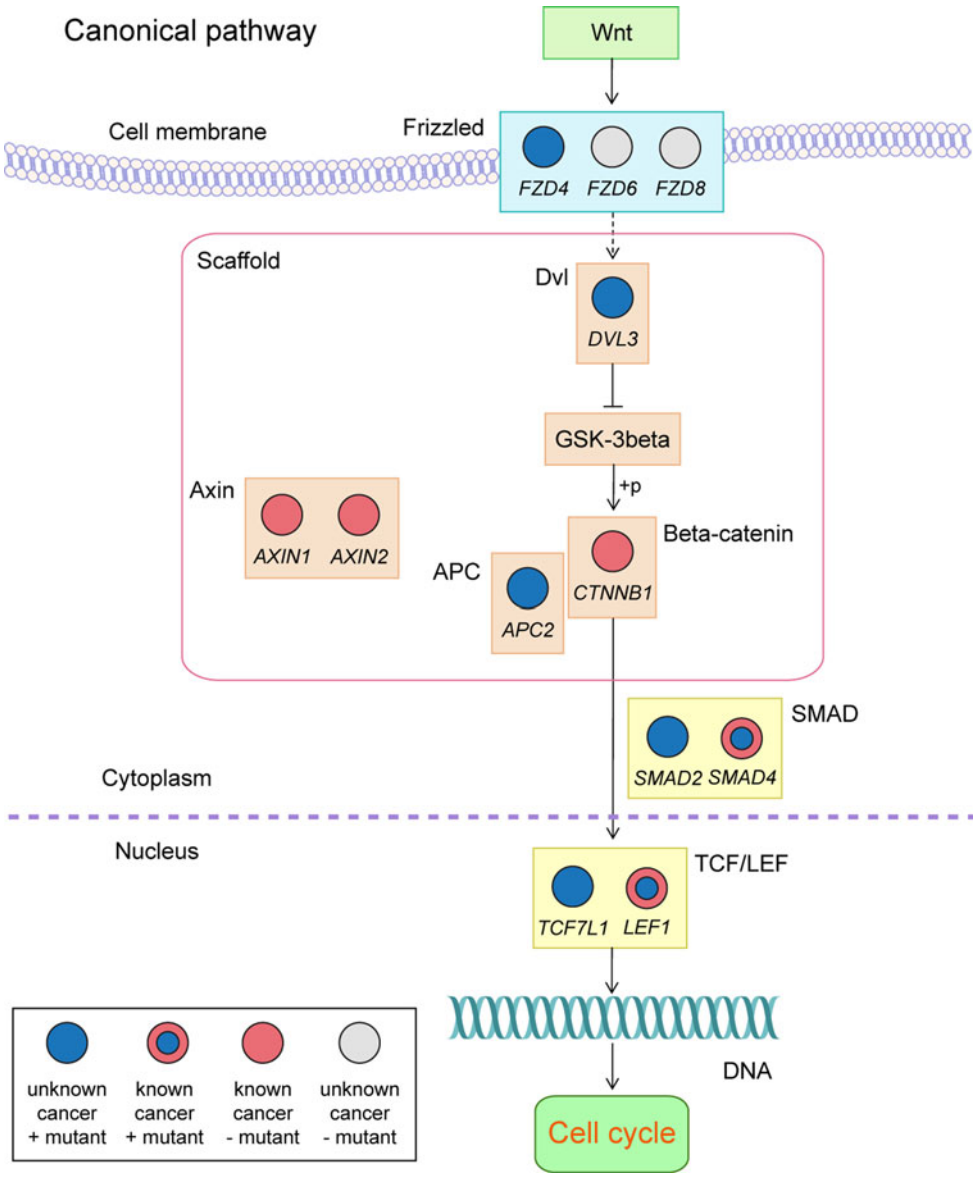

Fig. 4 Accumulation of predicted damaging mutations of known or unknown cancer genes in the Wnt pathway

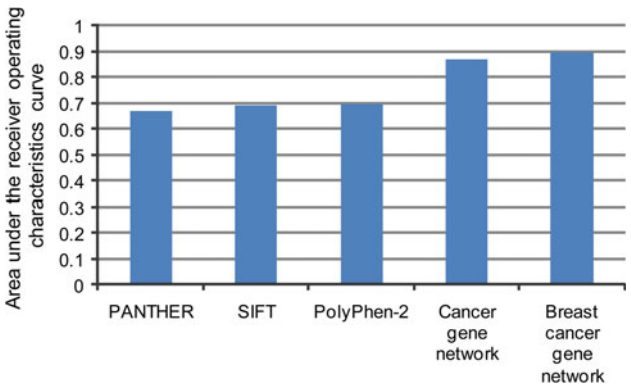

Fig. 5 Area under the receiver operating characteristics curve (AUROC) calculated for distinguishing breast cancer mutations from neutral polymorphisms

ATF2 (p.D352H, CAM = 0.917) with breast cancer (Baan et al. 2010; Knippen et al. 2009; Liu et al. 2009; Maekawa et al. 2007, 2008; Song et al. 2006). A previous combined analysis of genome and transcriptome data revealed new RNA-editing events that recode the amino acid sequence of *COG3* (p.R620C, CAM = 0.911) in metastatic lobular breast cancer patients (Shah et al. 2009).

Functional validation of those genes is needed to test their direct associations with tumorigenicity of breast cancer. Among the top genes, the *SPTLC1* (p.R239W, CAM = 0.944), *ACADM* (p.P132R, CAM = 0.906), *STRBP* (p.G280R, CAM = 0.851), and *SLC7A7* (p.P413S, CAM = 0.831) genes were not previously reported and may be good candidates for further functional study.

Approximately 59 % of genes with predicted damaging mutations have been reported to be related to cancers (43 % related to breast cancers) (Supplemental Table 7). On the contrary, only 31 % of genes with predicted neutral variants have been reported to be related to breast cancers in previous studies.

Discussion

In this study, we proposed the utilization of a disease gene network to determine the effect of mutations on a particular disease by considering the importance of the corresponding gene in the network. The combined individual predictor programs have been widely used for interpreting novel variants in patients with various pathological conditions and they are generally helpful for screening candidate deleterious mutations. With the combined predictors, the proposed Bayesian method, which used cancer gene network information, successfully improved the prediction performance.

By applying the proposed method to predict the functional consequence of the mutations found in colorectal and breast cancer tumors, we could prioritize the mutations according to their predicted deleterious effect on a cancer phenotype. Several genes were reported in previous studies to be linked to other type of cancers, which suggests their potential relevance to carcinogenesis. The candidate mutations can be confirmed by re-sequencing tumor samples from additional patients with colorectal or breast cancers. In particular, gene set analyses showed a significant enrichment of the genes with predicted deleterious mutations in many cancer pathways and important pathways related to carcinogenesis. For example, the Wnt signaling and VEGF signaling pathways were identified, which is consistent with a previous result showing that heregulin induced VEGF secretion via the erbB3 signaling pathway in colon cancer cell lines (Yonezawa et al. 2009). The proposed cancer gene network method not only captures gene interactions of currently known pathways, but also provides a flexible extension of the network. In particular, the method detected the candidate genes, *SIGLEC7*, *KCNQ5*, *SMAD2*, *FCN1*, and *LMNB2* in colon cancer, and *SPTLC1*, *ACADM*, *STRBP*, and *SLC7A7* in breast cancer, which had predicted deleterious mutations but were not previously reported. The results suggest that predicting identified mutations individually with the consideration of pathway information is useful for interpreting the functional consequence of the mutations regarding a particular phenotype.

The classified label (disease mutation and neutral polymorphism) of each variant obtained from the Swiss-Prot database is not definitive and cannot be used for clinical or diagnostic purposes and thus it might contain uncertain variants disrupting the prediction performance. The more the information, such as accurate mutation data and cancer genes that is accumulated, the higher is the accuracy of the proposed model. In addition, alternative gene prioritization methods may be used to define the prior probability π and their performances need to be evaluated (Barabasi et al. 2011). Merging the recently developed methods into our integrated model which utilized diverse properties other than sequence homology and conservation might improve the performance (Huang et al. 2010, 2011, 2012; Kumar et al. 2011; Ye et al. 2007). Nevertheless, our analysis correctly identified known cancer genes as well as candidate genes with the predicted deleterious mutations. For example, the mutation of *TNS4*, a well-known cancer gene, was successfully detected by the proposed method, but was omitted from the OMIM Morbid Map.

Different cancers show distinct expression patterns and mutational spectrums, but also share mutations in concurrent genes such as *TP53* and *KRAS* (Kan et al. 2010). Constructing different prediction models for each cancer type is possible by utilizing the gene expression data of each type of cancer or normal tissue. To demonstrate this, we constructed a breast cancer gene network, since RNA-seq data with high coverage and accuracy for normal breast tissues were available. We showed that utilizing the expression data and constructing a breast cancer-specific network increased the prediction accuracy for breast cancer mutations over the general cancer gene network. Expression data for other normal tissues are expected to be available due to ongoing research and projects such as the Human Body Map project for comprehensive tissue expression profiles generated by high-throughput sequencing technology. These data will be utilized as prior knowledge and used to test the proposed method for other cancer types. Although gene expression information in cancer is of great potential importance for the network, there might be substantial variation of gene expression even in the same cancer tissues, and we need further investigation to utilize this important information.

We found that the proposed Bayesian ensemble approach with a disease gene network can be useful to interpret their mutations with regard to potential functional consequences corresponding to a certain phenotype. On the other hand, these results show the usefulness of a disease gene network in predicting disease-specific variations and also suggest that the same approach might be applicable to other diseases, especially complex diseases affected by multiple rare causal variants of genes involved in key pathways. The integrative approach will hold promise for identifying the causal genetic variation of cancer based on each patient's genetic profile and offer a basis for personalized treatment for human cancers.

Acknowledgments We thank I. Adzhubei for helpful comments on the use of PolyPhen-2, P.C. Ng for SIFT, and A. Kejariwal for PANTHER. This work was supported by a grant from the Korea Healthcare technology R&D Project, Ministry of Health and Welfare, Republic of Korea (A092255). D. Lee was supported by the World Class University program (R32-2008-000-10218-0) and the Basic Research Laboratory grant (2009-0086964) of the Ministry of Education, Science and Technology through the National Research Foundation of Korea.

Conflict of interest The authors declare that they have no conflict of interest.

References

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248–249

Albasri A, Seth R, Jackson D, Benhasouna A, Crook S, Nateri AS, Chapman R, Ilyas M (2009) C-terminal Tensin-like (CTEN) is an oncogene which alters cell motility possibly through repression of E-cadherin in colorectal cancer. J Pathol 218:57–65

Baan B, Pardali E, ten Dijke P, van Dam H (2010) In situ proximity ligation detection of c-Jun/AP-1 dimers reveals increased levels of c-Jun/Fra1 complexes in aggressive breast cancer cell lines in vitro and in vivo. Mol Cell Proteomics 9:1982–1990

Bai A, Meetze K, Vo NY, Kollipara S, Mazsa EK, Winston WM, Weiler S, Poling LL, Chen T, Ismail NS, Jiang J, Lerner L, Gyuris J, Weng Z (2010) GP369, an FGFR2-IIIb-specific antibody, exhibits potent antitumor activity against human cancers driven by activated FGFR2 signaling. Cancer Res 70: 7630–7639

Barabasi AL, Gulbahce N, Loscalzo J (2011) Network medicine: a network-based approach to human disease. Nat Rev Genet 12: 56–68

Barbieri I, Pensa S, Pannellini T, Quaglino E, Maritano D, Demaria M, Voster A, Turkson J, Cavallo F, Watson CJ, Provero P, Musiani P, Poli V (2010) Constitutively active Stat3 enhances neu-mediated migration and metastasis in mammary tumors via upregulation of Cten. Cancer Res 70:2558–2567

Campbell PJ, Pleasance ED, Stephens PJ, Dicks E, Rance R, Goodhead I, Follows GA, Green AR, Futreal PA, Stratton MR (2008) Subclonal phylogenetic structures in cancer revealed by ultra-deep sequencing. Proc Natl Acad Sci USA 105: 13081–13086

Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. Cancer Res 69:6660–6667

Choi YL, Bocanegra M, Kwon MJ, Shin YK, Nam SJ, Yang JH, Kao J, Godwin AK, Pollack JR (2010) LYN is a mediator of epithelial–mesenchymal transition and a target of dasatinib in breast cancer. Cancer Res 70:2296–2306

Dijkstra EW (1959) A note on two problems in connexion with graphs. Numer Math 1:269–271

Dyer LM, Schooler KP, Ai L, Klop C, Qiu J, Robertson KD, Brown KD (2011) The transglutaminase 2 gene is aberrantly hypermethylated in glioma. J Neurooncol 101:429–440

Eto N, Miyagishi M, Inagi R, Fujita T, Nangaku M (2009) Mitogen-activated protein 3 kinase 6 mediates angiogenic and tumorigenic effects via vascular endothelial growth factor expression. Am J Pathol 174:1553–1563

Fischer A, Greenman C, Mustonen V (2011) Germline fitness-based scoring of cancer mutations. Genetics 188:383–393

Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B, Mishra G, Nandakumar K, Shen B, Deshpande N, Nayak R, Sarker M, Boeke JD, Parmigiani G, Schultz J, Bader JS, Pandey A (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38:285–293

Gaudet MM, Kirchhoff T, Green T, Vijai J, Korn JM, Guiducci C, Segre AV, McGee K, McGuffog L, Kartsonaki C, Morrison J, Healey S, Sinilnikova OM, Stoppa-Lyonnet D, Mazoyer S, Gauthier-Villars M, Sobol H, Longy M, Frenay M, Collaborators GS, Hogervorst FB, Rookus MA, Collee JM, Hoogerbrugge N, van Roozendaal KE, Piedmonte M, Rubinstein W, Nerenstone S, Van Le L, Blank SV, Caldes T, de la Hoya M, Nevanlinna H, Aittomaki K, Lazaro C, Blanco I, Arason A, Johannsson OT, Barkardottir RB, Devilee P, Olopade OI, Neuhausen SL, Wang X, Fredericksen ZS, Peterlongo P, Manoukian S, Barile M, Viel A, Radice P, Phelan CM, Narod S, Rennert G, Lejbkowicz F, Flugelman A, Andrulis IL, Glendon G, Ozcelik H, Toland AE, Montagna M, D'Andrea E, Friedman E, Laitman Y, Borg A, Beattie M, Ramus SJ, Domchek SM, Nathanson KL, Rebbeck T, Spurdle AB, Chen X, Holland H, John EM, Hopper JL, Buys SS, Daly MB, Southey MC, Terry MB, Tung N, Overeem Hansen TV, Nielsen FC, Greene MI, Mai PL, Osorio A, Duran M, Andres R, Benitez J, Weitzel JN, Garber J, Hamann U, Peock S, Cook M, Oliver C, Frost D, Platte R, Evans DG, Lalloo F, Eeles R, Izatt L, Walker L, Eason J et al (2010) Common genetic variants and modification of penetrance of BRCA2-associated breast cancer. PLoS Genet 6:e1001183

Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O'Meara S, Vastrik I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, Menzies A, Mironenko T, Perry J, Raine K, Richardson D, Shepherd R, Small A, Tofts C, Varian J, Webb T, West S, Widaa S, Yates A, Cahill DP, Louis DN, Goldstraw P, Nicholson AG, Brasseur F, Looijenga L, Weber BL, Chiew YE, DeFazio A, Greaves MF, Green AR, Campbell P, Birney E, Easton DF, Chenevix-Trench G, Tan MH, Khoo SK, Teh BT, Yuen ST, Leung SY, Wooster R, Futreal PA, Stratton MR (2007) Patterns of somatic mutation in human cancer genomes. Nature 446: 153–158

Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou YC, Pugh TJ, Robertson G, Chittaranjan S, Ally A, Asano JK, Chan SY, Li HI, McDonald H, Teague K, Zhao Y, Zeng T, Delaney A, Hirst M, Morin GB, Jones SJ, Tai IT, Marra MA (2010) Alternative expression analysis by RNA sequencing. Nat Methods 7:843–847

Grossman SR, Shylakhter I, Karlsson EK, Byrne EH, Morales S, Frieden G, Hostetter E, Angelino E, Garber M, Zuk O, Lander ES, Schaffner SF, Sabeti PC (2010) A composite of multiple signals distinguishes causal variants in regions of positive selection. Science 327:883–886

Hochgrafe F, Zhang L, O'Toole SA, Browne BC, Pinese M, Porta Cubas A, Lehrbach GM, Croucher DR, Rickwood D, Boulghourjian A, Shearer R, Nair R, Swarbrick A, Faratian D, Mullen P, Harrison DJ, Biankin AV, Sutherland RL, Raftery MJ, Daly RJ (2010) Tyrosine phosphorylation profiling reveals the signaling network characteristics of Basal breast cancer cells. Cancer Res 70:9391–9401

Huang T, Wang P, Ye ZQ, Xu H, He Z, Feng KY, Hu L, Cui W, Wang K, Dong X, Xie L, Kong X, Cai YD, Li Y (2010) Prediction of deleterious non-synonymous SNPs based on protein interaction network and hybrid properties. PLoS ONE 5:e11900

Huang T, Niu S, Xu Z, Huang Y, Kong X, Cai YD, Chou KC (2011) Predicting transcriptional activity of multiple site p53 mutants based on hybrid properties. PLoS ONE 6:e22940

Huang T, Wang C, Zhang G, Xie L, Li Y (2012) SySAP: a system-level predictor of deleterious single amino acid polymorphisms. Protein Cell 3:38–43

Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamiyama H, Jimeno A, Hong SM, Fu B, Lin MT, Calhoun ES, Kamiyama M, Walter K, Nikolskaya T, Nikolsky Y, Hartigan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, Karchin R, Papadopoulos N, Parmigiani G, Vogelstein B, Velculescu VE, Kinzler KW (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 321:1801–1806

Jones S, Wang TL, Shih Ie M, Mao TL, Nakayama K, Roden R, Glas R, Slamon D, Diaz LA Jr, Vogelstein B, Kinzler KW, Velculescu VE, Papadopoulos N (2010) Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. Science 330:228–231

Kaminker JS, Zhang Y, Waugh A, Haverty PM, Peters B, Sebisanovic D, Stinson J, Forrest WF, Bazan JF, Seshagiri S, Zhang Z (2007) Distinguishing cancer-associated missense mutations from common polymorphisms. Cancer Res 67:465–473

Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, Stern HM, Yue P, Haverty PM, Bourgon R, Zheng J, Moorhead M, Chaudhuri S, Tomsho LP, Peters BA, Pujara K, Cordes S, Davis DP, Carlton VE, Yuan W, Li L, Wang W, Eigenbrot C, Kaminker JS, Eberhard DA, Waring P, Schuster SC, Modrusan Z, Zhang Z, Stokoe D, de Sauvage FJ, Faham M, Seshagiri S (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. Nature 466:869–873

Katz M, Amit I, Citri A, Shay T, Carvalho S, Lavi S, Milanezi F, Lyass L, Amariglio N, Jacob-Hirsch J, Ben-Chetrit N, Tarcic G, Lindzen M, Avraham R, Liao YC, Trusk P, Lyass A, Rechavi G, Spector NL, Lo SH, Schmitt F, Bacus SS, Yarden Y (2007) A reciprocal tensin-3-cten switch mediates EGF-driven mammary cell migration. Nat Cell Biol 9:961–969

Knippen S, Loning T, Muller V, Schroder C, Janicke F, Milde-Langosch K (2009) Expression and prognostic value of activating transcription factor 2 (ATF2) and its phosphorylated form in mammary carcinomas. Anticancer Res 29:183–189

Kumar S, Dudley JT, Filipski A, Liu L (2011) Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. Trends Genet 27:377–386

Lee JS, Collins KM, Brown AL, Lee CH, Chung JH (2000) hCds1-mediated phosphorylation of BRCA1 regulates the DNA damage response. Nature 404:201–204

Li Y, Mizokami A, Izumi K, Narimoto K, Shima T, Zhang J, Dai J, Keller ET, Namiki M (2010) CTEN/tensin 4 expression induces sensitivity to paclitaxel in prostate cancer. Prostate 70:48–60

Liao YC, Chen NT, Shih YP, Dong Y, Lo SH (2009) Up-regulation of C-terminal tensin-like molecule promotes the tumorigenicity of colon cancer through beta-catenin. Cancer Res 69:4563–4566

Liu Y, Wang Y, Li W, Zheng P (2009) Activating transcription factor 2 and c-Jun-mediated induction of FoxP3 for experimental therapy of mammary tumor in the mouse. Cancer Res 69:5954–5960

Lo SH, Lo TB (2002) Cten, a COOH-terminal tensin-like protein with prostate restricted expression, is down-regulated in prostate cancer. Cancer Res 62:4217–4221

Maekawa T, Shinagawa T, Sano Y, Sakuma T, Nomura S, Nagasaki K, Miki Y, Saito-Ohara F, Inazawa J, Kohno T, Yokota J, Ishii S (2007) Reduced levels of ATF-2 predispose mice to mammary tumors. Mol Cell Biol 27:1730–1744

Maekawa T, Sano Y, Shinagawa T, Rahman Z, Sakuma T, Nomura S, Licht JD, Ishii S (2008) ATF-2 controls transcription of Maspin and GADD45 alpha genes independently from p53 to suppress mammary tumors. Oncogene 27:1045–1054

Masica DL, Karchin R (2011) Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. Cancer Res

Miller CW, Ikezoe T, Krug U, Hofmann WK, Tavor S, Vegesna V, Tsukasaki K, Takeuchi S, Koeffler HP (2002) Mutations of the CHK2 gene are found in some osteosarcomas, but are rare in breast, lung, and ovarian tumors. Genes Chromosomes Cancer 33:17–21

Morton CC (2004) Gene discovery in the auditory system using a tissue specific approach. Am J Med Genet A 130A:26–28

Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucl Acids Res 31:3812–3814

Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61–80

Rebouissou S, Rosty C, Lecuru F, Boisselier S, Bui H, Le Frere-Belfa MA, Sastre X, Laurent-Puig P, Zucman-Rossi J (2004) Mutation of TCF1 encoding hepatocyte nuclear factor 1alpha in gynecological cancer. Oncogene 23:7588–7592

Sasaki H, Yukiue H, Kobayashi Y, Fukai I, Fujii Y (2003) Cten mRNA expression is correlated with tumor progression in thymoma. Tumour Biol 24:271–274

Schwarz JM, Rodelsperger C, Schuelke M, Seelow D (2010) MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7:575–576

Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, Delaney A, Gelmon K, Guliany R, Senz J, Steidl C, Holt RA, Jones S, Sun M, Leung G, Moore R, Severson T, Taylor GA, Teschendorff AE, Tse K, Turashvili G, Varhol R, Warren RL, Watson P, Zhao Y, Caldas C, Huntsman D, Hirst M, Marra MA, Aparicio S (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. Nature 461: 809–813

Sharma S, Qian F, Keitz B, Driscoll D, Scanlan MJ, Skipper J, Rodabaugh K, Lele S, Old LJ, Odunsi K (2005) A-kinase anchoring protein 3 messenger RNA expression correlates with poor prognosis in epithelial ovarian cancer. Gynecol Oncol 99:183–188

Shi Z, Moult J (2011) Structural and functional impact of cancer-related missense somatic mutations. J Mol Biol 413:495–512

Sinclair CS, Rowley M, Naderi A, Couch FJ (2003) The 17q23 amplicon and breast cancer. Breast Cancer Res Treat 78: 313–322

Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, Meeh P, Markowitz SD, Willis J, Dawson D, Willson JK, Gazdar AF, Hartigan J, Wu L, Liu C, Parmigiani G, Park BH, Bachman KE, Papadopoulos N, Vogelstein B, Kinzler KW, Velculescu VE (2006) The consensus coding sequences of human breast and colorectal cancers. Science 314:268–274

Song H, Ki SH, Kim SG, Moon A (2006) Activating transcription factor 2 mediates matrix metalloproteinase-2 transcriptional activation induced by p38 in breast epithelial cells. Cancer Res 66:10487–10496

Sterpone S, Mastellone V, Padua L, Novelli F, Patrono C, Cornetta T, Giammarino D, Donato V, Testa A, Cozzi R (2010) Single-nucleotide polymorphisms in BER and HRR genes, XRCC1

haplotypes and breast cancer risk in Caucasian women. J Cancer Res Clin Oncol 136:631–636

Su JL, Yang CY, Shih JY, Wei LH, Hsieh CY, Jeng YM, Wang MY, Yang PC, Kuo ML (2006) Knockdown of contactin-1 expression suppresses invasion and metastasis of lung adenocarcinoma. Cancer Res 66:2553–2561

Swanton C, Burrell RA, Futreal PA (2011) Breast cancer genome heterogeneity: a challenge to personalised medicine? Breast Cancer Res 13:104

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A (2003) PANTHER: a library of protein families and subfamilies indexed by function. Genome Res 13:2129–2141

Torkamani A, Schork NJ (2008) Prediction of cancer driver mutations in protein kinases. Cancer Res 68:1675–1682

Tsuboi M, Mori H, Bunai T, Kageyama S, Suzuki M, Okudela K, Takamochi K, Ogawa H, Niwa H, Shinmura K, Sugimura H (2010) Secreted form of EphA7 in lung cancer. Int J Oncol 36:635–640

Wang Z, Liu H, Liu B, Ma W, Xue X, Chen J, Zhou Q (2010) Gene expression levels of CSNK1A1 and AAC-11, but not NME1, in tumor tissues as prognostic factors in NSCLC patients. Med Sci Monit 16:CR357–CR364

Webb EL, Rudd MF, Sellick GS, El Galta R, Bethke L, Wood W, Fletcher O, Penegar S, Withey L, Qureshi M, Johnson N, Tomlinson I, Gray R, Peto J, Houlston RS (2006) Search for low penetrance alleles for colorectal cancer through a scan of 1467 non-synonymous SNPs in 2575 cases and 2707 controls with validation by kin–cohort analysis of 14 704 first-degree relatives. Hum Mol Genet 15:3263–3271

Williamson RE, Darrow KN, Giersch AB, Resendes BL, Huang M, Conrad GW, Chen ZY, Liberman MC, Morton CC, Tasheva ES (2008) Expression studies of osteoglycin/mimecan (OGN) in the cochlea and auditory phenotype of Ogn-deficient mice. Hear Res 237:57–65

Won HH, Kim HJ, Lee KA, Kim JW (2008) Cataloging coding sequence variations in human genome databases. PLoS ONE 3:e3575

Wu J, Meng F, Kong LY, Peng Z, Ying Y, Bornmann WG, Darnay BG, Lamothe B, Sun H, Talpaz M, Donato NJ (2008) Association between imatinib-resistant BCR-ABL mutation-negative leukemia and persistent activation of LYN kinase. J Natl Cancer Inst 100:926–939

Ye ZQ, Zhao SQ, Gao G, Liu XQ, Langlois RE, Lu H, Wei L (2007) Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). Bioinformatics 23:1444–1450

Yonezawa M, Wada K, Tatsuguchi A, Akamatsu T, Gudis K, Seo T, Mitsui K, Nagata K, Tanaka S, Fujimori S, Sakamoto C (2009) Heregulin-induced VEGF expression via the ErbB3 signaling pathway in colon cancer. Digestion 80:215–225

Zang ZJ, Ong CK, Cutcutache I, Yu W, Zhang SL, Huang D, Ler LD, Dykema K, Gan A, Tao J, Lim S, Liu Y, Futreal PA, Grabsch H, Furge KA, Goh LK, Rozen S, Teh BT, Tan P (2011) Genetic and structural variation in the gastric cancer kinome revealed through targeted deep sequencing. Cancer Res 71:29–39

Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucl Acids Res 33:W741–W748