Identifying potential cancer driver genes by genomic data integration.

Yong Chen  
*Rowan University*

Jingjing Hao

Wei Jiang

Tong He

Xuegong Zhang

*See next page for additional authors*

Follow this and additional works at: [https://rdw.rowan.edu/csm_facpub](https://rdw.rowan.edu/csm_facpub)

Part of the [Computational Biology Commons](https://rdw.rowan.edu/csm_facpub)

**Recommended Citation**

Identifying potential cancer driver genes by genomic data integration

Yong Chen1,2, Jingjing Hao2, Wei Jiang2, Tong He3, Xuegong Zhang2, Tao Jiang2,4 & Rui Jiang2

1National Laboratory of Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China, 2MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China, 3School of Applied Mathematics, Central University of Finance and Economics, Beijing 102206, China, 4Department of Computer Science and Engineering, University of California, Riverside, CA 92521, USA.

Cancer is a genomic disease associated with a plethora of gene mutations resulting in a loss of control over vital cellular functions. Among these mutated genes, driver genes are defined as being causally linked to oncogenesis, while passenger genes are thought to be irrelevant for cancer development. With increasing numbers of large-scale genomic datasets available, integrating these genomic data to identify driver genes from aberration regions of cancer genomes becomes an important goal of cancer genome analysis and investigations into mechanisms responsible for cancer development. A computational method, MAXDRIVER, is proposed here to identify potential driver genes on the basis of copy number aberration (CNA) regions of cancer genomes, by integrating publicly available human genomic data. MAXDRIVER employs several optimization strategies to construct a heterogeneous network, by means of combining a fused gene functional similarity network, gene-disease associations and a disease phenotypic similarity network. MAXDRIVER was validated to effectively recall known associations among genes and cancers. Previously identified as well as novel driver genes were detected by scanning CNAs of breast cancer, melanoma and liver carcinoma. Three predicted driver genes (CDKN2A, AKT1, RNF139) were found common in these three cancers by comparative analysis.

Wide genomic aberration is a hallmark of the genomes of all cancer types. Deep sequencing technology1,2 has recently characterized the geographic and functional spectrum of cancer genomic aberrations and revealed insights into the mutational mechanisms3–6. These somatic mutations in cancer genomes may encompass several distinct classes of DNA sequence variations, including point mutations, copy number aberrations (CNA) and genomic rearrangements7. CNAs are deletions or additions of large segments of a genome, and usually include one to tens of genes. Although these somatically acquired changes have been observed in cancer cell genomes, it does not necessarily mean that all of the abnormal genes are also involved in the development of cancers. Indeed, some genes are likely to make no contribution to cancer progress at all. In order to draw a distinction between them, these mutated genes have been coined driver and passenger genes7,8. A driver gene is causally implicated in the process of oncogenesis, while a passenger gene makes no contribution to cancer development itself, but is simply a by-product of the genomic instability observed in cancer genomes. Distinguishing driver genes from passenger genes has thus been considered an important goal of cancer genome analysis, especially in the field of personalized medicine and therapy9,10.

Driver and passenger genes can be differentiated by the functional roles they play in cells. Different genomic data that measure gene functions at different dimensions would be highly informative to separate potential driver from passenger genes. Recently, several methods have been proposed to identify potential driver genes based on systematic integration of genome scale data of CNA and gene expression profiles, and applied to melanoma9, gingivobuccal cancer11 and liver carcinoma12. Apart from using gene expression data, integrating other types of genomic datasets such as those for protein-protein interaction13–15, epigenetic16, metabolism pathways17,18, sequence similarity19 and Gene Ontology20 should greatly increase the predictive power for driver genes and thus enable researchers to systematically investigate the mechanisms underling a great variety of cancers.

To this aim, we developed a computational method, MAXDRIVER, for the identification of driver genes from aberrant regions throughout cancer genomes by integrating multiple omics data. Several computational strategies are used to optimize gene similarities, filter noise and search maximal information flow among a query disease and candidate genes through a heterogeneous network. Large-scale validation results suggest MAXDRIVER is a useful method for genomic data integration and the discovery of cancer driver genes from aberrant regions and their flanks. By comparative analysis of breast cancer, melanoma and liver carcinoma, common potential drivers...
and their associated pathways are proposed. The present work highlights the importance of systematic integration and optimization of multiple omics data to investigate the mechanisms that underlie cancer development.

**Results**

**Overview of MAXDRIVER.** MAXDRIVER mainly performed three steps to integrate multiple data sources for the identification of driver genes for a given cancer (Fig. 1). In the first step, we adopted a multiple regression model to construct a fused gene functional similarity network, in which edge weights were derived from four data sources, including protein-protein interactions (PPI), gene co-expression patterns (GCE), gene sequence similarities (GSS), and pathway co-occurrence relationships (PCO). For this purpose, we calculated a gene similarity profile using each of these data sources and derived a gene functional similarity profile using the GO function (GO). The above five similarity profiles were used to calculate a functional similarity between a gene pair of the gene functional similarity network through a multiple regression model. With parameters of the model estimated, we further used the trained model to calculate a score for every pair of genes, obtaining edge weights of the fused gene functional similarity network. In the above procedure, we adopted a heuristic filtration strategy to filter out noises that indicated low confidence relationships for gene pairs. In the second step, we combined the fused gene functional similarity network with a disease phenotypic similarity network and gene-disease associations to construct a heterogeneous network. In the third step, we applied an information flow method to the heterogeneous network to trace the relationships among cancers and candidate genes located in CNA regions. In this procedure, we measured the strength of association between a cancer and a gene as the maximum value of the information flowing from the cancer to the gene (Fig. 1a). With this method established, we used genes located in CNA regions as candidates and ranked them according to their maximal information flow values (Fig. 1b).

**Performance of MAXDRIVER.** Identification of cancer driver genes is usually done by performing biological experiments, however only a few driver gene sets are available to date. Therefore, there are no large gene lists that can be used to validate the performance of MAXDRIVER. Alternatively, here we used known disease genes as simulations for leave-one out large-scale validations and test if MAXDRIVER was able to find known drivers of cancers. First, we selected previously identified disease genes from the OMIM database as positive controls, and then tested if they can be recalled from artificially constructed control sets, including linkage intervals and random controls. On linkage interval control gene set, a known disease gene was simulated as a driver gene and its neighbour genes in 10 M distance as passengers, considering that 10 M was much larger than most CNAs were. Cross validations for recalling the known cancer genes from interval control gene sets indicated that MAXDRIVER can achieve top one ranked precision (TOP) as high as 64.06%, with parameters $\beta = 0.25$ and $\gamma = 0.19$ (Fig. 2). The mean rank ratio (MRR) of all 2,496 test cases was 7.19%, suggesting that known disease genes were ranked highly. We calculated the area under the rank receiver operating characteristic curve (ROC), named AUC, and achieved an AUC value of 93.78% (see Methods for detailed definitions of TOP, MRR and AUC). We also performed a

![Figure 1 Workflow of MAXDRIVER](www.nature.com/scientificreports)

(a) A heterogeneous network was constructed by two-step integrations. First, a fused gene functional similarity network was calculated by fusing 4 datasets. To any gene pair gene $g_i$ and $g_j$, the gene network was weighted by fused values $OPT(g_i,g_j) = \bar{\theta}_0 + \sum_{s=1}^{4} \bar{\theta}_s \cdot r^s(g_i,g_j)$, $\bar{\theta}_s, s = 0,1,2,3,4$ were optimized parameters. The values were further filtered by setting parameter $\beta = 0.25$. Second, a heterogeneous network was constructed by merging the fused gene functional similarity network, gene-disease associations and a disease phenotypic similarity network that was calculated by text mining. The disease phenotypic similarities were further filtered by setting parameter $\gamma = 0.19$. Edges from each candidate gene to an additional sink are added for calculating the maximal information flow from a query cancer to the sink. The information flow through each candidate gene is then calculated as the strength of association between the candidate gene and the query disease. (b) CNA regions were estimated from deep-sequenced data of common or personal cancer genomes. Genes located in CNA regions were ranked by the maximal information flow values calculated from cancer node to candidate gene nodes in the heterogeneous network.
validation on random control sets, whose 99 genes were randomly selected on a genome-wide scale. This test allowed for simulation of cases where no prior gene knowledge was available, and also avoided a bias towards better-characterized regions. Results for this validation showed that MAXDRIVER can successfully rank 1,558 known disease genes at top 1, yielding a TOP precision of 62.42% (Fig. 2). It also achieved a MRR of 14.37% and an AUC of 86.61%. These results indicate that MAXDRIVER achieves high precision on both control sets. Permutation experiments were performed by shuffling edges in the gene network while fixing the degree (i.e., the number of neighbors of each node). For two control sets, the AUC scores were both reduced by approximately 50%, but the result for the linkage interval (55.11%) was slightly higher (Fig. 2). These permutation validations suggest that the ability of MAXDRIVER to prioritize known disease genes is mainly due to its ability to exploit the informative connections in the heterogeneous network.

Access of individual genomic data sources. To test the contribution of all genomic data sets, we validated the performances of each individual data on random control gene sets. At each case, the edge weights of the gene functional similarity network were only calculated from individual gene genomic data. Among the four data sources, PPI obtained the highest TOP of 53.41%, a MRR of 16.7% and an AUC value of 84.3%, while the GCE was as low as 47.22% (TOP), 18.21% (MRR) and 80.15% (AUC). When all four data sources were fused without noise filtering, the performances of MAXDRIVER were improved to 56.54% (TOP), 16.58% (MRR) and 84.38% (AUC) (Supplementary Fig. S1). These results suggest that the ability of MAXDRIVER to prioritize known disease genes is mainly due to its ability to exploit the informative connections in the heterogeneous network.

Robustness of MAXDRIVER. Threshold parameters $\beta, \gamma$ play important functions in filtering the noise and in making the heterogeneous network sparse. Here we changed them with their respective change ratios to test their effects on MAXDRIVER. For parameter $x$, the parameter changed ratio is defined as $(x - x_0)/x_0$ and the related changed ratios are defined as $(\text{TOP}(x) - \text{TOP}(x_i))/(\text{TOP}(x_0))$, $(\text{AUC}(x) - \text{AUC}(x_i))/\text{AUC}(x_0)$ and $(\text{MRR}(x) - \text{MRR}(x_i))/\text{MRR}(x_0)$, where $x_0$ represents the optimized parameter value. When the $\beta$ changed ratios were set from -1 to 1, the changed ratios of TOP, MRR and AUC were all less than 0.14. When the $\gamma$ changed ratios were set from -1 to 1, the changed ratios of TOP, MRR and AUC were all less than 0.2 (Fig. 3). In comparison, the change ratios of TOP were more than MRR and AUC, suggesting that these two parameters affected many of the top one ranked genes. More specifically, TOP reached its optimum at the optimal point of $\beta$ (Fig. 3a, b), but TOP was optimal when $\gamma$ was 0 (Fig. 3c, d). These results indicate that disease phenotypic similarity is useful in achieving higher AUC and MRR, albeit at the expense of a slightly reduced TOP score. In summary, the parameters $\beta$ and $\gamma$ exhibit varying effects on the different criteria without significantly affecting the robustness of the overall performance of MAXDRIVER.

Predicted driver genes of breast cancer, melanoma and liver carcinoma. To test the capability of MAXDRIVER in detecting novel driver genes, we employed it to scan CNAs of breast cancer (OMIM 114480), melanoma (OMIM 155600) and liver carcinoma (OMIM 114550). Recently, cancer genes and mutational processes in breast cancer had been systematically reported21–24. Within 218 breast cancer samples, 78 CNAs were detected by using GISTIC, a method that was used for CNAs analysis from deep sequencing data25,26. By ranking the genes located in the regions of CNAs, we detected 70 top one ranked genes as potential driver genes (Supplementary Table S1). Five genes, namely AKT1, PIKSCA, TP53, AKT2 and MAP3K were successfully predicted amongst breast cancer genes that had been verified experimentally in previous reports (e.g. AKT1, BRCA1, CDH1, GATA3, PIK3CA, PTEN, RB1, TP53, AKT2, ARID1B, CASP8, CDKN1B, MAP3K1, MAP3K3, NCO1, SMARCD1 and TBX3)27. Besides the 5 known driver genes, a total of 64 genes had been annotated before to be associated with diverse diseases by genetic association database28 or OMIM database29 (see more detailed information in Supplementary Table S1). Melanoma is a well-studied type of tumor, and four genes (MITF, KLFL, TBC1D16 and RAB27A) have been reported previously to act as driver genes8. From two early genomic variation studies30,31, a total of 91 CNAs had been identified. Using MAXDRIVER, we predicted 63 potential genes to be driver genes (Supplementary Table S2), with one of the four genes, MITF, detected successfully. Whilst RAB27 was not identified itself, one of its interacting protein partners, MYOSA, was also ranked top one position. Furthermore, a total of 58 genes are annotated as disease related (Supplementary Table S2). We also tested MAXDRIVER on liver carcinoma, the third leading cause of cancer-related death worldwide. Thirty frequently altered genes had been described before, including TP53, CNOT8B1, ARID1, ARID2, AXIN1, PRS6KA3, VCA1M, CDK4, TERT, MLL4, CCNE1, CDK2, AXIN1, PTEN, PIK3CA, KRAS, NRAS, MYC, MET, CCND1, CDKN2A, RB1, ARID1A, ARID1B, DRD2, IRF2, NFE2L2, ERRF1, RPSK3A and MLL330,31. We reported 57
potential driver genes, including TP53 and CDKN2A (Supplementary Table S3), detected by scanning for 57 CNAs. Except for gene PAK2, all other 56 genes had been annotated previously as disease related (Supplementary Table S3). These predicted driver genes exhibit diverse associations not only to cancers, but also to various human diseases, providing novel candidates for further experimental validations.

Common driver genes and biological processes of breast cancer, melanoma and liver carcinoma. Early research had revealed similar genomic aberrations and mechanisms underlying tumorigenesis in various cancers. Here, we performed a comparative analysis of 70, 63 and 57 predicted driver genes of breast cancer, melanoma and liver carcinoma respectively. Specifically, 9 genes were found to be common in liver carcinoma and melanoma, 7 genes in liver carcinoma and breast cancer, and 11 genes in melanoma and breast cancer. Three genes (CDKN2A, AKT1, RNF139) were predicted to be driver genes in all three cancer types (Fig. 4a). Amongst these three genes, dysfunction of CDKN2A and AKT1 had been identified previously to cause disease. For example, AKT1 was reported to drive development of breast cancer and proteus syndrome. CDKN2A (p16) acted as a tumor suppressor gene, and its inactivation by deletion was frequently found in cancers such as, bladder carcinoma and lung adeno-carcinomas. Based on our analysis, we propose that RNF139 is a novel cancer driver gene. RNF139 encoded an endoplasmic reticulum-resident E3 ubiquitin ligase, and it contained a sterol-sensing domain and a RING finger motif. This gene located in a CNA region in melanoma (121,760,777–128,860,000), a CNA region in liver carcinoma (120,876,000–134,690,000) and near a CNA region in breast cancer (128,816,653–128,993,129). Although it was not reported early as a driver gene, MAXDRIVER predicted it as top one from 15, 19 and 23 genes that were located in the CNA regions of breast cancer, melanoma and liver carcinoma, respectively, suggesting it may be a novel potential driver gene.

We next performed a functional enrichment analysis on all 164 unique predicted genes, a union of 70, 63 and 57 genes of breast cancer, melanoma and liver carcinoma respectively. These 164 genes were almost uniformly distributed across the human chromosomes, except for the X and Y chromosomes, with only one gene located on X and none on Y (Supplementary Fig. S3). These genes were significantly related to several cancer pathways, such as melanoma (hsa05218), endometrial cancer (hsa05213), non-small cell lung cancer (hsa05223), renal cell carcinoma (hsa05211), glioma (hsa05214), bladder cancer (hsa05219) and chronic myeloid leukemia (hsa05220) (all p-values < 1E-5, Supplementary Table S4). Furthermore, 164 genes were found to be enriched in a number of biological processes related to cell fate, cell survival, and genome maintenance (Fig. 4b, p-value < 1E-5, Supplementary Table S5). Within the category ‘regulation of biological processes’, negative regulation of cell death, programmed cell death and apoptosis were most enriched (Fig. 4b-b1). Comparatively, genes with functions involved in positive regulation of cellular processes were found to be enriched even more highly (Fig. 4b-b2). In addition, genes involved in the response to chemical stimuli were also found highly enriched (Fig. 4b-b3), as well as genes involved in developmental processes (Fig. 4b-b4), indicating that dysfunctions of these genes could be important to cancer development.
being involved in cancer pathways (KEGG, hsa05200)\(^3\), 23 genes were successfully predicted as driver genes by our method (Supplementary Fig. S4), achieving a p-value as low as 2.2E-16 (Fisher exact test, one-sided). In total, 558 interactions and 717 regulatory connections were found to exist among 163 cancer genes and 164 predicted driver genes (Fig. 5, Supplementary Table S6), suggesting these genes were tightly correlated. The functional relationships among the 164 predicted driver genes were also supported by other functionally related evidences such as activation, inhibition, binding, post-translation or co-expression, all of which were analyzed by using STRING database\(^{34}\) (Supplementary Fig. S5). The identification of these driver genes is consistent with earlier analyses of cancer genomes, and should reveal new potential oncogenes involved in cancer onset as well as progress.

**Detecting driver genes in CNAs flanks.** MAXDRIVER can detect gene drivers located not only in regions of CNAs, but also accompanying flanking regions. This is important for identifying real driver mutations (or genes), since genome aberrations may be not located in protein-coding regions but in intronic, intergenic or gene regulatory regions, and could affect genes at great distance within the chromosome\(^{36}\). One example is a single nucleotide polymorphism (SNP), rs6983267, which interacts physically with the MYC proto-oncogene, which is located some 330 kilobases away on the chromosome\(^{41,42}\). Here, we show that MAXDRIVER has the ability to detect the genes CDKN2A, AKT1 and RNF139 as driver genes, even though they not all are located in regions with CNAs. AKT1 is located in a CNA region of breast cancer on chromosome 14, while it is located at the 5’ side of a liver carcinoma CNA region, and at the 3’ side of a melanoma CNA region. CDKN2A is located in two CNA regions of liver carcinoma and breast cancer, respectively, while it is located at the 5’ side of a melanoma CNA region. RNF139 is located in two liver carcinoma and melanoma CNAs, but at the 5’ side of a breast cancer CNA region (Fig. 6). MAXDRIVER can rank these three genes at top one position by scanning all extended CNA regions with flanks of three cancers. Detailed analysis of RNF139 revealed that the nearest breast cancer CNA region to RNF139 is 3,315,974 bp far away and 43,347 bp in length. Intriguingly, this CNA region does not harbour any protein coding sequences. Setting this CNA at the centre, we extended it from 123,838,327 bp to 133,838,327 bp to include 14 genes. MAXDRIVER successfully ranked RNF139 at top one position of these 14 genes. Similar results on two other CNAs of liver carcinoma and melanoma are also obtained by MAXDRIVER. In summary, these results suggest

---

**Figure 4 | Common driver genes and functional enrichment analysis of predicted driver genes of breast cancer, melanoma and liver carcinoma.** (a) Venn graph showing common driver genes of predicted driver genes of three cancers. (b) Functional enrichment of 164 potential driver genes. The sizes of circles are correlated to the gene numbers. The colours used indicate enrichments p-values. The graph was analyzed and printed by using BINGO software\(^6\).
that MAXDRIVER is a powerful tool for identifying driver genes that are located even at large distances from mutated regions within cancer genomes.

**Discussion**

Next-generation sequencing technology has significantly contributed to cancer genomics research, and has revealed a plethora of genomic changes that accumulate in cancer cells. However, methods that can perform deep analysis of such large genomic data sets, and effectively utilize the redundancy and complementarity of multiple data, are largely absent. To bridge this gap, we have developed MAXDRIVER, a method shown here to be efficient and reliable in its ability to predict novel cancer driver genes. It can be used not only to scan common CNAs collected from multiple cancer genomes for the identification of driver genes that are likely to be common in cancers, but also to scan CNAs of personal genomes, in order to identify personalized driver genes, and such knowledge can then contribute to the development of personalized medicine as well as therapies. Its superior performances benefits from two key approaches, namely its optimizing strategy of genomic data integration using a noise filtering function, and the utilization of a maximal information flow method to maximize the relationship between queried cancer and its candidate genes.

Using MAXDRIVER, we were not only able to predict many novel potential driver genes, but it also allowed us to perform a comparative view of liver carcinoma, breast cancer and melanoma cancer. We predicted 164 potential driver genes, many of which had been reported previously as oncogenes (or tumor suppressors), and which displayed a tight correlation to 163 genes reported in the cancer pathway (hsa05200) suggesting that MAXDRIVER’s predictions are reliable. Interestingly, we found that the three genes CDKN2A, AKT1 and RNF139 were predicted in all three cancers. CDKN2A and AKT1 are well characterized as driver genes that are crucially involved in a diverse range of cancers. Although further experiments are needed to investigate the exact function of RNF139, particularly in cancer progress, our analysis strongly indicates that dysfunction of RNF139 may be involved in the development of a multitude of cancers.

Although genomic aberrations identified by deep sequencing technologies can now be scrutinized by deep analysis utilizing MAXDRIVER, the integration of additional genomic data should allow for even greater acceleration of this process. For example, time-series genomic measurements of different stages of cancer development will allow for dynamical descriptions of genomic aberrations. Meanwhile, genome-wide patterns of epigenetic modifications are another useful information, since the importance of...
epigenetic markers in maintaining normal cellular functions is reflected by the observation that many diseases develop upon spatial and temporal dysregulation of these histone and DNA-associated modifications\(^{46,47}\). A miRNA regulatory network could also be useful for driver gene prediction, as dysfunction of miRNA and their targeted genes had been reported in diverse cancers\(^{46,47}\). The genes that are regulated by one or more miRNAs can be viewed as a functional linkage and can then be embedded into our heterogeneous network. Considering that cancer is mainly a disease of signaling problems, we can also add a gene signaling network into our heterogeneous network. At present, the largest curated human signaling network is available for 6,309 proteins, including 62,737 signaling actions (released on Oct. 26, 2013, http://www.bri.nrc.ca/wang/\(^{48-51}\). The human signaling network is not only useful for detecting cancer-related genes, but more importantly for investigating how differential signaling processes occur throughout cancer development. Compared to our constructed non-directed gene network, the gene signaling network is directed and includes three types of signaling actions, i.e. activation, inhibition and physical interaction, which allow for presenting more precise relationships among gene pairs. Therefore, a more precise modelling is desired to efficiently use this signaling network information for detailed investigation of cancer driver genes and cancer-related signaling processes.

An interesting and quantitative set of data is gene differential expression patterns calculated from cancer and normal cells. In our study, we have defined functional relationships between genes from the correlations of their expression vectors that are measured in 79 human tissues. This definition is useful for describing generic relationships among genes, but may not be adequate in providing sufficiently precise information for a special type of cancer, since different cancers usually have different expression profiles\(^{52-55}\). By utilizing differential expression patterns of genes specific for certain type of cancer, we can present a more precise prediction for the cancer. Technically, the utilization of differential expression patterns can be taken in two ways. First, we can further filter the predicted driver genes by analyzing their differential expression patterns. Genes exhibit highly differential expression patterns may possess a higher potential to be involved in the cancer development. Second, it might be also possible to embed differential expression patterns of genes into the heterogeneous network by using degrees of differential expression as capacities of gene pairs. Although this strategy of using differential expression patterns is desired in achieving higher prediction power for well-studied cancers, such a strategy could only be used case-by-case and thus is not useful for rarely studied cancers (because of the limited availability of gene expression data). In summary, MAXDRIVER can be used to scan CNAs of personal genomes, in order to identify molecular attributes of individual patients, and it is easily extended to utilize more types of omics data that can sever as a useful tool for integrative network analysis in cancer systems biology\(^{56-58}\). By predicting and analyzing driver genes triggering different stages of tumorigenesis, we can track a rippling effect along the cancer differentiation path, and track how cancer development is triggered and then amplified under the involvement of these genes. We are confident that the power and resolution offered by MAXDRIVER will allow biologists to better distinguish the different functional roles of genes underlying cancer development.

### Methods

MAXDRIVER is designed to identify driver genes by integrating multiple omics data. It first integrates 5 gene datasets to build a fused gene functional similarity network. A heterogeneous network is then constructed by integrating a fused gene similarity network, a disease phenotypic similarity network and gene-disease associations. A maximum flow information flow strategy is then employed to prioritize candidate genes that are located in CNA regions and the top one-ranked genes were considered to be driver genes, whilst others were considered passenger genes. In the following, we describe, in details, the calculation of gene similarity, construction of heterogeneous network, and relationships among a query disease and candidate genes by calculating the maximal information flow.

#### Calculation of gene similarity profiles.

To build the fused gene functional similarity network, we describe the calculations of gene similarity from each of five datasets. We extracted protein-protein interactions from version 9.0 (released on Apr 13, 2011) of the HPRD database\(^{59}\) composed of 9,515 genes and 37,364 interactions. Given a pair of genes i and j, we defined their similarity \(r_{ij}^{(1)}\) as the unit weight 1, if these two genes were linked in the network, otherwise we defined them as 0. Finally, we obtained a 9,515 by 9,515 binary matrix as a gene interacting profile.

We downloaded from the NCBI Refseq database (released Mar, 2012) the protein sequences of all the aforementioned 9,515 proteins. We then aligned these sequences against each other using the NCBI BLAST program\(^{48-51}\). We then calculated the sequence similarity between a pair of genes i, j as

\[
    r_{ij}^{(2)} = \frac{-\log(e_{ij})}{\max_j(-\log(e_{ij}))}
\]

where \(e_{ij}\) was the e-value outputted by BLAST program with default parameters. By doing this, we obtained a 9,515 by 9,515 numeric matrix as a gene sequence similarity profile.

We downloaded the gene expression data from http://bioinformatics.nisetl.edu/download/ (GENE code: GSE11313), in which whole genome gene expressions were measured for 79 human tissues\(^{60}\). We then represented each of the 9,515 genes as a 79-dimensional expression vector, with each dimension corresponding to the expression level of the gene in a particular tissue. We then calculated the co-expression pattern between a pair of genes i, j as the absolute Pearson’s correlation coefficient of their expression vectors, calculated as

\[
    r_{ij}^{(3)} = \frac{\text{cov}(e_i, e_j)}{\sigma(e_i)\sigma(e_j)}
\]

where \(e_i\) and \(e_j\) were the expression vectors of genes i and j, respectively. By doing so, we obtained a 9,515 by 9,515 numeric matrix that served as a gene expression similarity profile.

We downloaded a total of 200 human pathways from the KEGG database (released Mar, 2012)\(^{56}\). We then represented a gene i as a 200 dimensional binary vector by assigning 1 to a dimension if the gene was present in the corresponding pathway, and assigned 0 otherwise. We then calculated the gene co-occurrence relationship between two genes i and j as the cosine of the angle between their vectors, calculated by

\[
    r_{ij}^{(4)} = \frac{p_i \cdot p_j}{|p_i||p_j|}
\]

where \(p_i\) and \(p_j\) were the vectors corresponding to the genes i and j, respectively. By doing so, we obtained a 9,515 by 9,515 numeric matrix as a gene pathway similarity profile.

For each of the above profiles, similarity values are in the range of [0, 1], where 1 means the highest similarity and 0 the lowest similarity. In the case that the relationship between a pair of genes is absent from a data source, the lowest similarity 0 is assigned.

#### Fusion of gene similarities

Although each of the above gene similarity profiles reflects functional similarity of genes from a certain perspective, none of them provides a thorough explanation about relationships between gene properties and gene functions. We therefore integrated these similarity profiles and generated a single fused similarity profile that can link gene properties to gene functions in a more comprehensive way. For this purpose, we downloaded the biological process domain of the Gene Ontology (GO)\(^{56}\) and the corresponding annotations for human genes (both released on Jan, 2012). Focusing on genes annotated with at least fifteen

---

**Figure 6** Detecting driver genes CDKN2A, AKT1 and RNF139 in CNA region flanking sequences. BC: Breast Cancer. ME: Melanoma. LC: Liver Carcinoma.
informative GO terms that appeared at or below the fifth level of the GO hierarchy and had at least five genes annotated, we obtained a total of 3,842 well-annotated genes. We then calculated the functional similarities between every pair of these genes, using the method of Resnik, as implemented in the software package GOSimSim60,61, obtaining a 3,842 by 3,842 numeric matrix serving as similarities with genes, using the method of Resnik, as implemented in the software package.

Heterogeneous network

Prioritization of candidate genes by maximizing information flow in the network to a directed graph by treating each undirected edge as two distinct directed edges. First, we obtained from the literature62 a phenotype similarity profile that was composed of a fused gene functional similarity network, a disease phenotypical similarity network, and known associations among diseases and genes.

The fused gene functional similarity network, \( G_{\text{fun}} = \{ V_{\text{fun}}, E_{\text{fun}} \} \), was constructed by introducing a parameter \( \beta \), to filter small similarity values in the predicted similarity matrix \( S \). More precisely, with the cut-off value given by \( \beta \), we set all \( s_{jk} \geq \beta \) and kept \( s_{jk} \) unchanged otherwise. We noticed that such a noise filtering process not only helped in filtering out low confidence similarities and thus improving the performance of our method, but also resulted in a sparse network to speed up subsequent calculations.

We used this convolution to determine the optimal value for \( \beta \). First, we used the mean of the fused similarities and obtained \( \beta = 0.23 \). With this parameter value, the gene functional similarity network contains 9,515 nodes and 1,756,050 edges. Second, we used a grid search with step 0.01 to find the optimal \( \beta \) that could yield the best performance in a cross-validation experiment and obtained \( \beta = 0.25 \). With this parameter value, the gene functional similarity network contained 9,515 nodes and 579,620 edges. Among these two methods, the later was more time consuming, but typically resulted in higher prediction accuracy in cross-validation experiments.

We constructed the disease phenotype similarity network, \( G_{\text{phen}} = \{ V_{\text{phen}}, E_{\text{phen}} \} \), as follows. First, we obtained from the literature62 a phenotype similarity profile that was represented as a matrix of similarity scores among 5,080 human disease phenotypes. Then, we introduced a parameter \( \gamma \) to filter out small similarity values in this matrix. Similar to our strategy for building the gene functional similarity network, we set all similarity scores to 0 if \( s_{jk} < \beta \) and kept \( s_{jk} \) unchanged otherwise. We noticed that such a noise filtering process not only helped in filtering out low confidence similarities and thus improving the performance of our method, but also resulted in a sparse network to speed up subsequent calculations.

We used this convolution to determine the optimal value for \( \gamma \). First, we used the mean of the fused similarities and obtained \( \gamma = 0.14 \). With this parameter value, the disease phenotype similarity network contained 1,609 nodes and 1,013,622 edges. Second, we used a grid search with step 0.01 to find the optimal \( \gamma \) that could yield the best performance in a cross-validation experiment and obtained \( \gamma = 0.19 \). With this parameter value, the disease phenotype similarity network contained 1,609 nodes and 600,090 edges.

Prioritization of candidate genes by maximizing information flow in the heterogeneous network. We formulated the problem of prioritization of candidate genes as searching for the maximum information flow in the heterogeneous network, and solved this problem as follows. First, we converted the undirected heterogeneous network to a directed graph by treating each undirected edge as two distinct directed edges of opposite directions. Second, we assigned a capacity value to each directed edge to represent the amount of information transmission through this edge. As an edge of a disease pair, we defined the phenotypic similarity of that pair as their capacity. As an edge of a gene pair, we assigned the functional similarity between these two genes as their capacity. For the association component, we assigned positive infinites to all edges. Then, given a query disease and a set of candidate genes, we introduced a sink node and linked the candidate genes (nodes) to the sink node by edges of positive infinite capacities (Fig. 1a). Fourth, the strength of association from the query disease to each candidate gene is calculated. Formally, for the heterogeneous network \( G = (V, E, C) \) where \( V, E, C > 0 \) representing the node, edge and nonnegative capacity on each edge respectively, the maximal information flow from the query node to the sink through all candidate genes is calculated as follows.

\[
\text{Maximize } f(v, w) = \sum_{v \in V} f(v, w) = 0
\]

Where the information flow \( f(v, w) \) is defined as the flow value transmitted from node \( v \) to node \( w \), and \( c_{ap}(v, w) \) is the capacity of the edge linked nodes \( v \) and \( w \).

Validation methods. We conducted two cross-validation experiments to assess the performance of our method and further applied this method to three real CNA data sets for breast cancer, melanoma, and liver carcinoma.

In the first cross-validation experiment, we ranked each gene (positive control) known as associated with a disease against a set of genes located within 10 Mb around the positive control (negative controls). We call this experiment validation against a linkage interval to simulate the genes in CNAs. In the second cross-validation experiment, we ranked each gene known as associated with a disease (a positive control) against a set of 99 genes selected at random from all 9,519 genes in the disease phenotype similarity profile (negative controls). We called this experiment ‘validation against random controls’.

We used three criteria to evaluate the performance of our method in a cross-validation experiment. First, we calculated the proportion of positive cases ranked first to obtain a criterion called the top one precision (TOP). Second, we divided the rank of a positive case by the number of corresponding negative controls to obtain the relative rank of the positive case, and we averaged relative ranks of all positive cases to obtain a criterion called the mean rank ratio (MRR). Third, the threshold to calculate the proportion of positive cases ranked above the threshold as sensitivity and the proportion of negative controls ranked below the threshold as specificity. We then plotted the receiver operating characteristic curve (ROC) and calculated the area under this curve to obtain a criterion called the AUC score. Thus, the larger the TOP and AUC, the smaller the MRR, the higher the performance of the prioritization method.

For real case studies, we collected 78 CNA regions from breast cancer data sets12,13, 91 CNA regions from melanoma14,15, and 57 CNA regions from liver carcinoma data sets16. For each of the three cancers, we scanned each of their CNA regions to obtain a set of candidate genes, and we applied our method to rank these candidate genes. Finally, we proposed genes ranked at the top of ranking lists as potential driver genes. Gene functional enrichment analysis was performed by using the DAVID database17,18, the STRING database19 and the KEGG database20.

28. Lin, W. M.

25. Mermel, C. H.

24. Staaf, J., Jonsson, G., Ringner, M., Baldetorp, B. & Borg, A. Landscape of somatic


15. Zhang, Q. C.

13. Chen, Y., Jiang, T. & Jiang, R. Uncover disease genes by maximizing information

17. Dutta, B. et al. A network-based, integrative study to identify core biological

12. Lango Allen, H.

11. Stephens, P. J. et al. The landscape of cancer genes and mutational processes in

8. Dutta, B. & Dutt, S. K. M.

5. Poland, K. S.

4. Gimelli, S.

3. Szklarczyk, D.

2. Zhang, Q. C.

1. Chen, Y., Jiang, T. & Jiang, R. Uncover disease genes by maximizing information

Acknowledgments

This work was partly supported by the National Basic Research Program of China (2012CB116304), the National High Technology Research and Development Program of China (2012AA020401), the National Science Foundation (61175902, 60928007, and 61273238). The authors would like to thank Dr. Torsten Juelich for critical reading of this manuscript and helpful suggestions.

Author contributions

R.J., T.I. and X.Z. provided guidance and planning for the project. Y.C. and R.J. produced the program and wrote the manuscript, particularly producing the results section. J.H., W.J. and T.H. contributed in preparing data and analysis of the results.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports.

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Chen, Y. et al. Identifying potential cancer driver genes by genomic data integration. Sci. Rep. 3, 3538; DOI:10.1038/srep03538 (2013).

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported license. To view a copy of this license, visit http://creativecommons.org/licenses/by-nc-sa/3.0/