BMC
Medical Genomics

RESEARCH   Open Access

Combined analysis of gene regulatory network and SNV information enhances identification of potential gene markers in mouse knockout studies with small number of samples

Benjamin Hur[1], Heejoon Chae[2], Sun Kim[3*]

Abstract

RNA-sequencing is widely used to measure gene expression level at the whole genome level. Comparing expression data from control and case studies provides good insight on potential gene markers for phenotypes. However, discovering gene markers that represent phenotypic differences in a small number of samples remains a challenging task, since finding gene markers using standard differential expressed gene methods produces too many candidate genes and the number of candidates varies at different threshold values. In addition, in a small number of samples, the statistical power is too low to discriminate whether gene expressions were altered by genetic differences or not. In this study, to address this challenge, we purpose a four-step filtering method that predicts gene markers from RNA-sequencing data of mouse knockout studies by utilizing a gene regulatory network constructed from omics data in the public domain, biological knowledge from curated pathways, and information of single-nucleotide variants. Our prediction method was not only able to reduce the number of candidate genes than the differential expressed gene-only filtered method, but also successfully predicted significant genes that were reported in research findings of the data contributors.

Background

RNA-sequencing (RNA-seq) data are widely used for detecting differentially expressed genes (DEGs) [1,2] and DEGs are often used for finding gene markers to explain the phenotypic differences between control and cases. However, in gene knockout studies, discovering gene markers with the DEG-based method has several limitations. It is known to be difficult to distinguish whether the expression alteration resulted from the inactivation of the knockout gene that caused phenotypic differences or from the genetic variations that were merely from differences in samples rather than phenotypic differences. The problem of distinguishing gene markers from

the DEG-based method becomes much more challenging when the number of samples is small [3], an issue that RNA-seq experiments face frequently.

Various methods and models were proposed to overcome the difficulties of selecting phenotype related DEGs from a small number of samples such as the Poisson model [4], Bayesian approaches [5,6], and increasing the sequence depth of samples [3]. Even though the intensive studies have resolved the difficulties of DEG detection in some degree, there is no method that can effectively suggest phenotype related DEGs from a small number of samples. In order to obtain significant DEGs, a recent study suggests increasing the number of biological samples rather than increasing the sequence depth [7]. However, increasing the number of biological samples is not easy for many reasons including budget issues, hence a

* Correspondence: sunkim.bioinfo@snu.ac.kr
[3]Department of Computer Science and Engineering, Seoul National University, Daehak-dong, 151-742 Seoul, Korea
Full list of author information is available at the end of the article

new approach that can detect significant gene markers in a small number of samples is neccesary.

In this study, we propose a new method that distinguishes gene markers with a small number of samples from case-control studies (especially, in knockout studies), by combining multiple methods such as DEG, gene regulatory network (GRN), biological pathways and single-nucleotide variants (SNVs), in a single computational framework.

Proposed method for selecting phenotype specific DEGs

Our gene marker selection method is a reductionist-approach by adding more filters at each step as described below. The order of filtering steps is illustrated in Figure 1.

1. The first filter is to use a method to identify DEGs between control and case samples. In this study, we used fold change, a classical DEG selection method.
2. The filter at the second step is to use GRN. GRN is constructed from large volume of public data to represent the whole gene regulatory network. DEGs that are included in the network are selected as candidates.
3. The third filter utilizes biological pathway information. Candidates that are not included in the pathways are discarded.
4. Finally, candidates that have higher than a certain rate of SNVs are discarded since the DEGs that have SNVs possibly resulted from genetic differences rather than phenotypic differences.

Candidate selection using DEGs

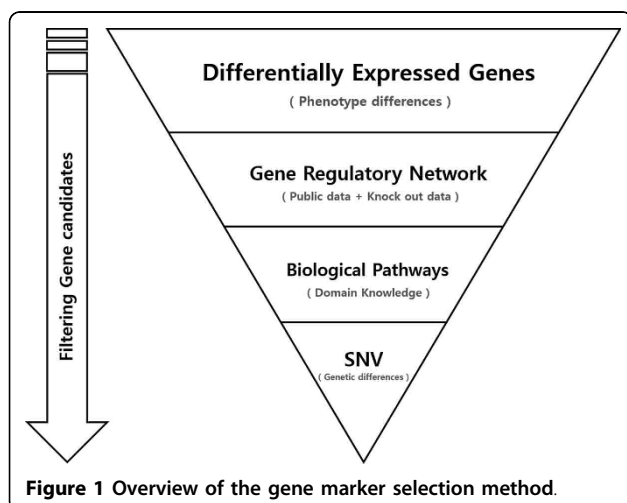In the first step, DEGs were chosen as initial candidates of gene markers. DEGs are used for the purpose of observing the alteration of expression patterns that could explain the phenotypic differences among samples. DEGs were selected by using fold change of the expression value (FPKM) between case and control. We used multiple cutoffs in order to compare and observe differences in the number of selected genes. In addition to candidate selection using DEGs, conditions such as up regulation and down regulation were recorded at each threshold. As shown in Table 1, the numbers of DEGs decreases significantly when the cutoff value increases. However, it is difficult to determine which cutoff is appropriate to be considered as candidates. In addition, the number of candidates are still too big even at the most restricted threshold. Moreover, it is not clear whether the expression alteration resulted from the differences of phenotype or not. Therefore, the necessity of seeking additional information arises, and needs to be combined to with initial candidates in order to select more significant DEGs that can explain whether the expression alterations resulted from the phenotypic differences or from the genetic differences.

Reducing candidates by combining GRN

In order to select significant gene markers from the candidates selected by DEGs, we need to investigate which candidates are involved in the regulatory role that caused phenotypic differences by comparing wild type (WT) and knockout (KO) mice. Therefore, GRN is used for the purpose of discarding non-significant genes, as GRN is a very effective method that can consider complex relationships of many genes [8] (in here, candidate genes). NARROMI [9] was used for GRN construction. Details of GRN construction are described in Methods section. Using the topology information created by NARROMI, we discarded those candidates that have weak or no regulatory roles. In other words, potential regulatory roles should exist between two DEGs. As a result, genes that have regulatory roles remained as candidates. Table 1 shows the result of a reduced number of candidates.

Reducing candidates by combining biological pathways

The combination of DEG and GRN information was used not only for reducing the number of candidates but also to select significant genes that have regulatory roles that could represent the phenotypic differences between WT and KO mice. However, it is also important to ensure whether the candidates are related to the curated biological pathways or not. Even though the candidates were selected by considering their regulatory roles from the gene complex, it is still not clear whether they really have a role in pathways. Therefore, confirming the candidates in terms of domain knowledge is necessary. In this study, we used KEGG pathway [10]

Figure 1 Overview of the gene marker selection method.

Table 1. Number of genes selected by each method.

Feature	DEG	DEG+GRN	DEG+GRN+Pathway	DEG+GRN+Pathway+SNV
Fold change (KO / WT)	Number of candidate genes	Number of candidate genes	Number of candidate genes	Number of candidate genes
More than 1, Less than 1	12298	8834	3436	2622
More than 1.1, Less than 0.91	8953	6264	2441	1861
More than 1.2, Less than 0.83	6466	4322	1683	1272
More than 1.3, Less than 0.77	4631	2986	1179	879
More than 1.4, Less than 0.71	3495	2125	845	629
More than 1.5, Less than 0.67	2712	1574	633	463
More than 1.6, Less than 0.63	2153	1184	478	343
More than 1.7, Less than 0.59	1750	914	364	257
More than 1.8, Less than 0.56	1439	718	284	203
More than 1.9, Less than 0.53	1235	562	233	165
More than 2.0, Less than 0.5	1064	456	192	135

for the confirmation. Candidates that do not have corresponding information to the KEGG pathway are excluded from the candidates. Table 1 shows that use of biological pathways was able to narrow down the candidates significantly.

Reducing candidates by combining SNV information

Even though we reduced the number of candidates by using multiple filtering methods, it is necessary to eliminate genes that have genetic differences that may not represent phenotypic differences. However, since the statistical power is weak in a small number of samples, it is difficult to distinguish whether the genetic differences were caused by phenotypic differences or not. Therefore we used a simple naive solution that removes genes that have certain or higher SNV rate. This is to avoid the risk of selecting SNVs resulted from genetic differences. However, this process will remove SNVs that caused not only non-phenotypic differences but also phenotypic differences. Details of candidate reduction using SNVs are explained in the Methods section.

Results and discussion

To show the effectiveness of our proposed method, we used RNA-seq data of GSE47851 from NCBI Gene Expression Omnibus (GEO) [11]. GSE47851 is from a study that used Gata3 conditional deficient mice [12]. Specific details of preprocessing of RNA-seq data are described in the Methods section. We used DEG, GRN, biological pathways and SNV information to the RNA-seq data in order to find gene markers that can distinguish the phenotypic difference between WT and KO. The gradual combination of multiple filters showed a dramatic reduction of the number of genes in each step (Table 1, Figure 2). Since it is known that defining DEG by fold change is an arbitrary selection, it is necessary to compare the differences at each threshold. We used

various fold changes in order to observe the differences in the number of candidates at each threshold. However, we were intrigued by the results that the number of candidates varies less in combined filters (DEG, GRN, pathway and SNV) than other filtering methods. Figure 2 interprets that the number of candidates in fold change over 1.6 and less than 0.63 varies lesser than other threshholds. Therefore, we decided to use fold change 1.6 and 0.63 in order to validate our study.

The biological significance of gene markers that were selected by our method is discussed in terms of the research article that used the GSE47851 data set. Selection of gene markers was done entirely by computational analysis without resorting to the literature and our prediction results were compared with findings of the data contributor. It is reported that genes of TNF and TNFR super families, members of NFkB and cell-surface markers of ILC2s have expression alterations
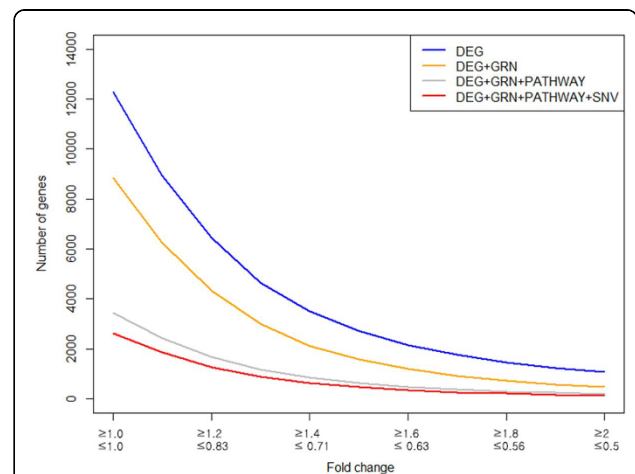

Figure 2 Comparison of the number of candidates of each method at each fold change.

when Gata3 is not activated in ILC2 cells [12]. The following study stated:

"we found that many TNF and TNFR superfamily genes, such as **Tnfrsf9** and ~~Tnfsf21~~ and NFkB family members, including **Nfkb2** and **Relb**, showed altered expression patterns. In addition, the cell-cycle inhibitor **Cdkn2b** was up regulated upon Gata3 inactivation".

Our method found 4 out of 5 genes included in the statement above. Where 4 genes are in bold face and un-matched genes are struckthrough. In addtion, we were able to reconfirm the following facts by mapping the candidate genes to the KEGG pathway. Figure 3C represents expression alteration in NF-kappa B signaling pathway, showing down regulations of Nfkb2 (p100) and Relb when Gata3 is inactivated. Expression alteration was also detected in the TNF signaling pathway (Figure 4C). TNF and TNFR super family genes, such as Tnf and Tnfrsf9, were successfully detected in the pathway as well as the statement. In addition, up regulation of Cdkn2b showed in Cell cycle pathway (Figure 5C) which was also stated exactly from the study.

The study also stated about the expression alterations in cell-surface markers of ILC2s:

"130 genes are positively regulated by GATA3 in ILC2s, but not in Th2 cells; they include **Icos**, **Il2ra**, **Kit**, **Il1r2**, ~~Cysltr1~~, ~~Htr1b~~, and ~~Tph1~~, many of which are cell-surface markers of ILC2s."

In the statement above, 4 genes among 7 were successfully matched. Inferring that our proposed method was able to find that the candidates filtered by multiple information were indirectly verified in terms of findings described in the studies that used the data, which shows the effectiveness of determining important genes as potential gene markers.

In addition to the effectiveness, we noticed that our filtering method was able to reduce false positive effectively and helps to focus on significant genes by using biological pathways. Figure 3, Figure 4 and Figure 5 show not only a dramatic reduction in gene numbers in the pathway (from non-filtered to full-filtered) but also candidate genes are likely to be as gene markers. This implies that our proposed method might help researchers focus on genes that may have caused phenotypic differences, hence able to discover true gene markers more efficiently.

We summarize the performance of our method in terms of the gene marker selection. We defined the reported genes (26 in total) [12] as gene markers (true positive) and used them to evaluate of our filter method in terms of precision, recall, F-measure, and accuracy.

$$Precision = \frac{tp}{tp + fp} \quad (1)$$

$$Recall = \frac{tp}{tp + fn} \quad (2)$$

$$F - measure = \frac{2tp}{2tp + fp + fn} \quad (3)$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (4)$$

True positives (tp) represent successfully selected gene markers and false positives (fp) as incorrect selected gene markers. False negatives (fn) represent incorrectly discarded gene markers and true negatives (tn) as correctly discarded non-gene markers. Table 2 summarizes precision, recall, F-measure, and accuracy of the combined filters. In terms of precision, with no filtering method, there was only 0.002 of chance to select a gene marker and 0.011 for the DEG-only filtering method (Table 2). Both non-filtered and DEG-only filtered methods were not powerful enough to detect potential gene markers in terms of the performance measures above. However, when we gradually added GRN, pathway, and SNV filters, the performance improved dramatically. In terms of precision, GRN pathway, and SNV filters achieved 0.041. About 4 times higher precision compared to the DEG-only filter. Naturally higher precision comes with lower recall rates. However, reduction in recall was not as dramatic as precision, reporting 0.538. When we consider both precision and recall, it was evident that our filtering method outperforms the DEG-only method in terms of both F-measure and accuracy. For the F-measure, GRN, pathway, and SNV filters achieved 0.076 that is 3.6 times higher F-measure than the DEG-only filter. For accuracy, GRN, pathway, and SNV filters achieved 0.972 that is 1.2 times higher accuracy when compared to the DEG-only filter. In summary, our proposing method, the combination analysis of DEG, GRN, pathway, and SNV filters out performs the DEG-only filtering method in terms of predicting potential gene markers.

Conclusion

We proposed a novel method that use a four-step filtering strategy in order to find potential gene markers from RNA-seq data of mouse knockout studies that have small number of samples. The standard DEG method was used to select candidates that explains the phenotypic differences between KO and WT in the first filtering step. A combination of GRN, biological pathways and SNV information was able to narrow down significant genes that have regulatory roles and reduced the risk of including candidates that have genetic differences. As a result, we were able to distinguish multiple
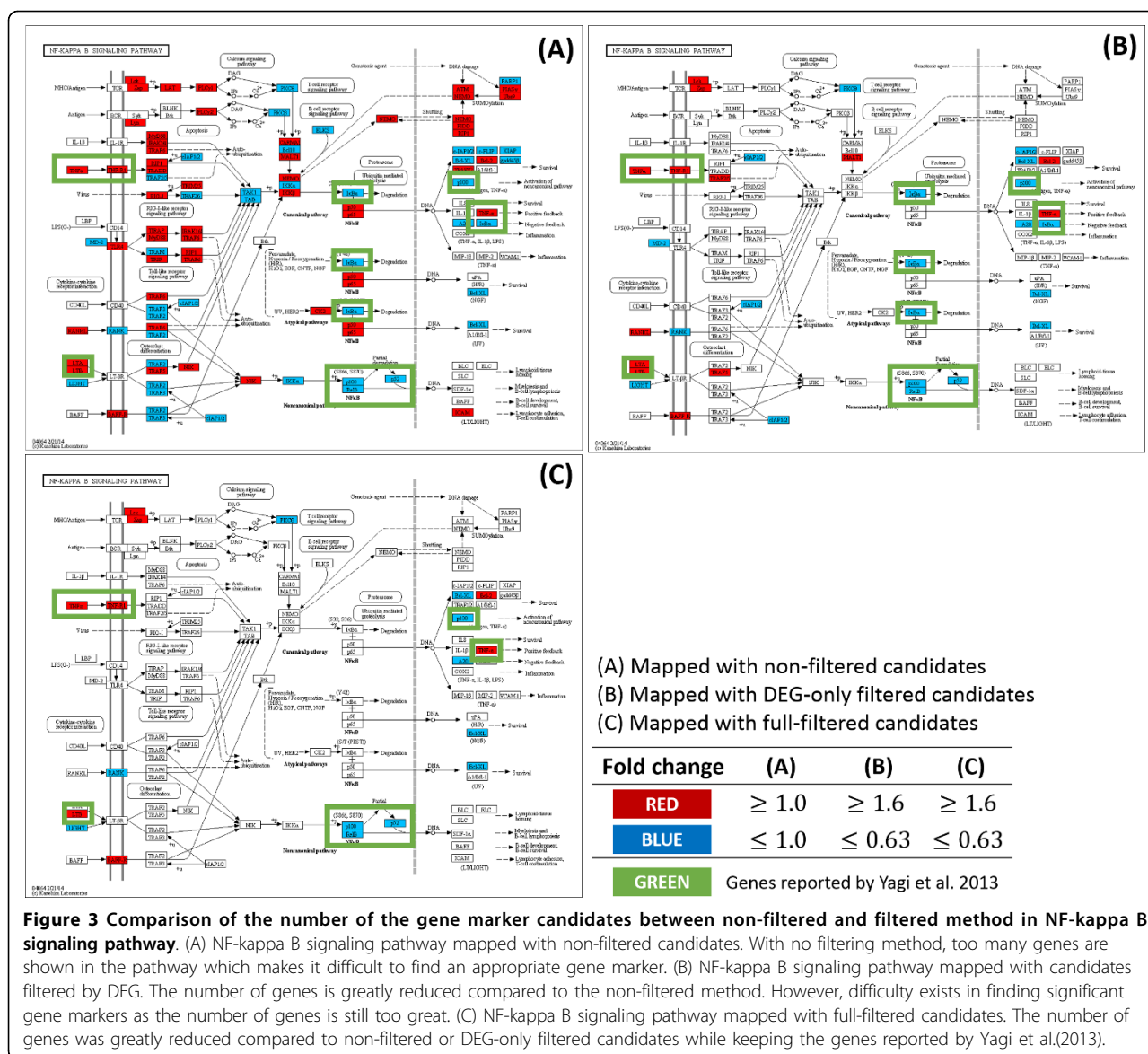
(A) Mapped with non-filtered candidates
(B) Mapped with DEG-only filtered candidates
(C) Mapped with full-filtered candidates

Fold change		(A)	(B)	(C)
	RED	≥ 1.0	≥ 1.6	≥ 1.6
	BLUE	≤ 1.0	≤ 0.63	≤ 0.63
	GREEN	Genes reported by Yagi et al. 2013		

Figure 3 Comparison of the number of the gene marker candidates between non-filtered and filtered method in NF-kappa B signaling pathway. (A) NF-kappa B signaling pathway mapped with non-filtered candidates. With no filtering method, too many genes are shown in the pathway which makes it difficult to find an appropriate gene marker. (B) NF-kappa B signaling pathway mapped with candidates filtered by DEG. The number of genes is greatly reduced compared to the non-filtered method. However, difficulty exists in finding significant gene markers as the number of genes is still too great. (C) NF-kappa B signaling pathway mapped with full-filtered candidates. The number of genes was greatly reduced compared to non-filtered or DEG-only filtered candidates while keeping the genes reported by Yagi et al.(2013).

gene markers in a small number of samples by reproducing the research findings reported in a knockout study [12]. Moreover, the use of KEGG pathways [10] for gene marker selection can be viewed as utilizing the prior information since pathways are constructed based on information from the literature. Thus, mapping potential gene markers to the pathways takes advantage of the interpretability of omics data from mouse knockout studies.

However, several limitations of this study need to be addressed. First of all, there should be more rigorous study of GRN construction. Using much omics data for GRN construction somehow preserves important relationships between transcription factors and their target genes, but how much data is needed for GRN construction is not rigorously studied. In this study, we had enough omics data for the network construction, therefore we were able to

use a simple method using NARROMI [9]. However, when the amount of omics data for network construction is not enough, special techniques such as low order partial correlation based methods [13] should be considered. Second, removing genes with genetic variation allows us to focus on genes that are relevant to the underlying biological mechanisms for the knockout study. How-ever, genetic variations do not necessarily affect transcription activity of genes, and it is possible that our method might discard some SNVs that were affected by the knockout gene. Thus we need to further investigate the effect of genetic variations on transcription activities. Finally, we plan to develop a machine learning based gene ranking method so that genes selected at the last step can be ranked and help biologists to prioritize the follow up study after the knock-out study.
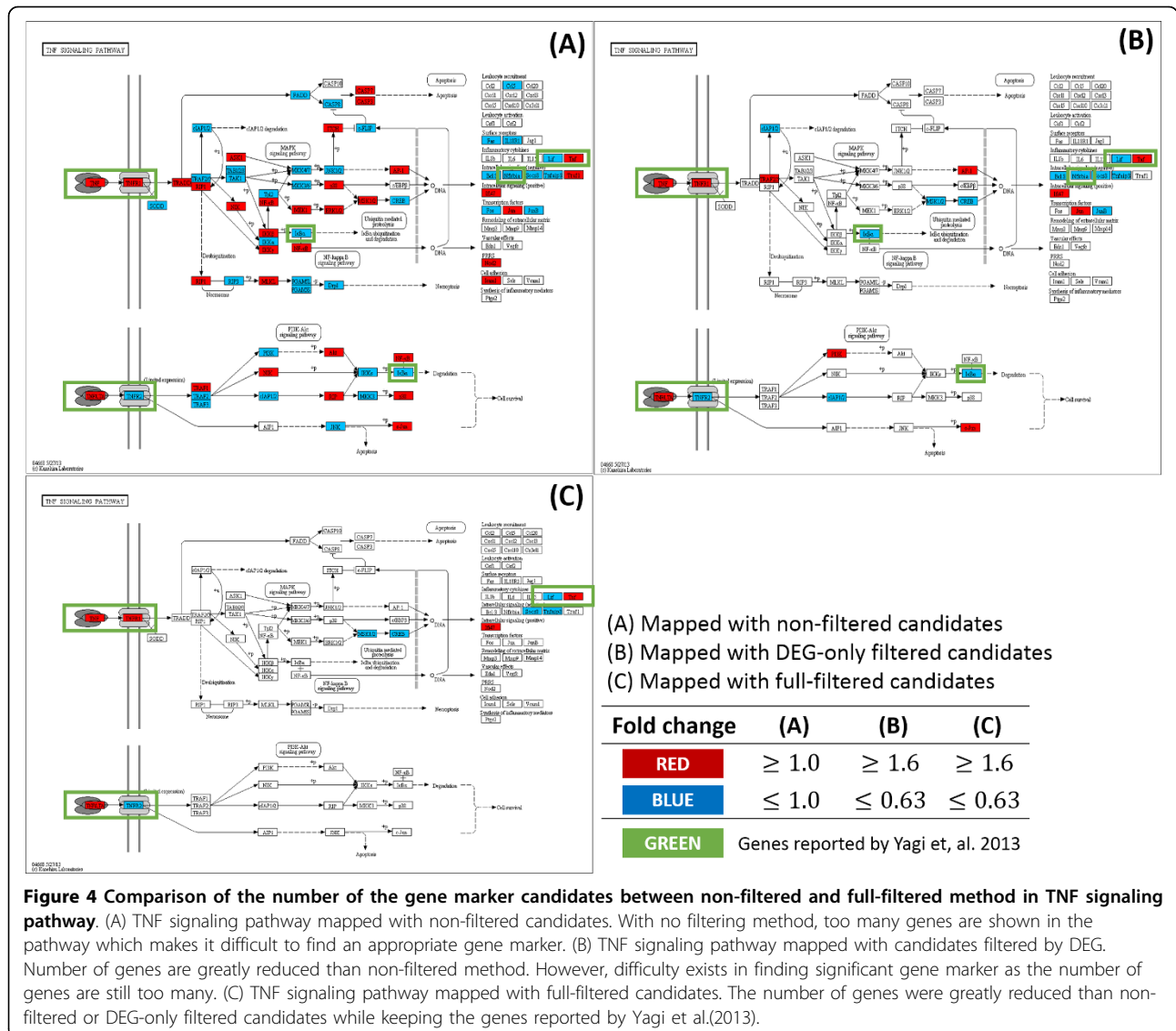
Figure 4 Comparison of the number of the gene marker candidates between non-filtered and full-filtered method in TNF signaling pathway. (A) TNF signaling pathway mapped with non-filtered candidates. With no filtering method, too many genes are shown in the pathway which makes it difficult to find an appropriate gene marker. (B) TNF signaling pathway mapped with candidates filtered by DEG. Number of genes are greatly reduced than non-filtered method. However, difficulty exists in finding significant gene marker as the number of genes are still too many. (C) TNF signaling pathway mapped with full-filtered candidates. The number of genes were greatly reduced than non-filtered or DEG-only filtered candidates while keeping the genes reported by Yagi et al.(2013).

Methods

Data Collection

Public data (Microarray, RNA-seq) of mice were collected from NCBI GEO. For Microarray, each series matrix files from GSE45929 [14], GSE16741 [15], GSE30906 [16], GSE36780 [17], GSE40375 (not published), GSE41380 [18], GSE43663 [19] were used for GRN construction. These data contains gene expression value of multiple samples (42 samples in total) that were created by the same microarray platform (Illumina MouseWG-6 v2.0 expression beadchip) and preprocessed by R bioconductor lumi package [20] (variance stabilizing transform, quantile normalization). Note that these samples differs in mouse's strain, genotype and treatment. For RNA-seq data, GSE47851 were used for the evaluation of the study's method. RNA-seq data of GSE47851 are from an experiment of Gata3 KO that have multiple SRA files. We used 8 SRA files (SRR896215, SRR896216, SRR896217, SRR896218, SRR896219, SRR896220, SRR896221, SRR896222) that are raw data in two conditions where each of the conditions have 2 biological samples and 2 technical replicates of each biological sample.

Gene regulatory network construction

We integrated gene expression values of 7 series matrix files (GSE45929, GSE16741, GSE30906, GSE36780, GSE40375, GSE41380, GSE43663) into a single matrix and quantile normalized gene expression values of every sample and used it as an expression profile for construction of GRN. GRN is constructed by using NARROMI [9] with a default option. As NARROMI requires a list of transcription factors and a list of genes, we collected a list of transcription factors and co-factors from the Animal Transcription Factor Database [21] and defined
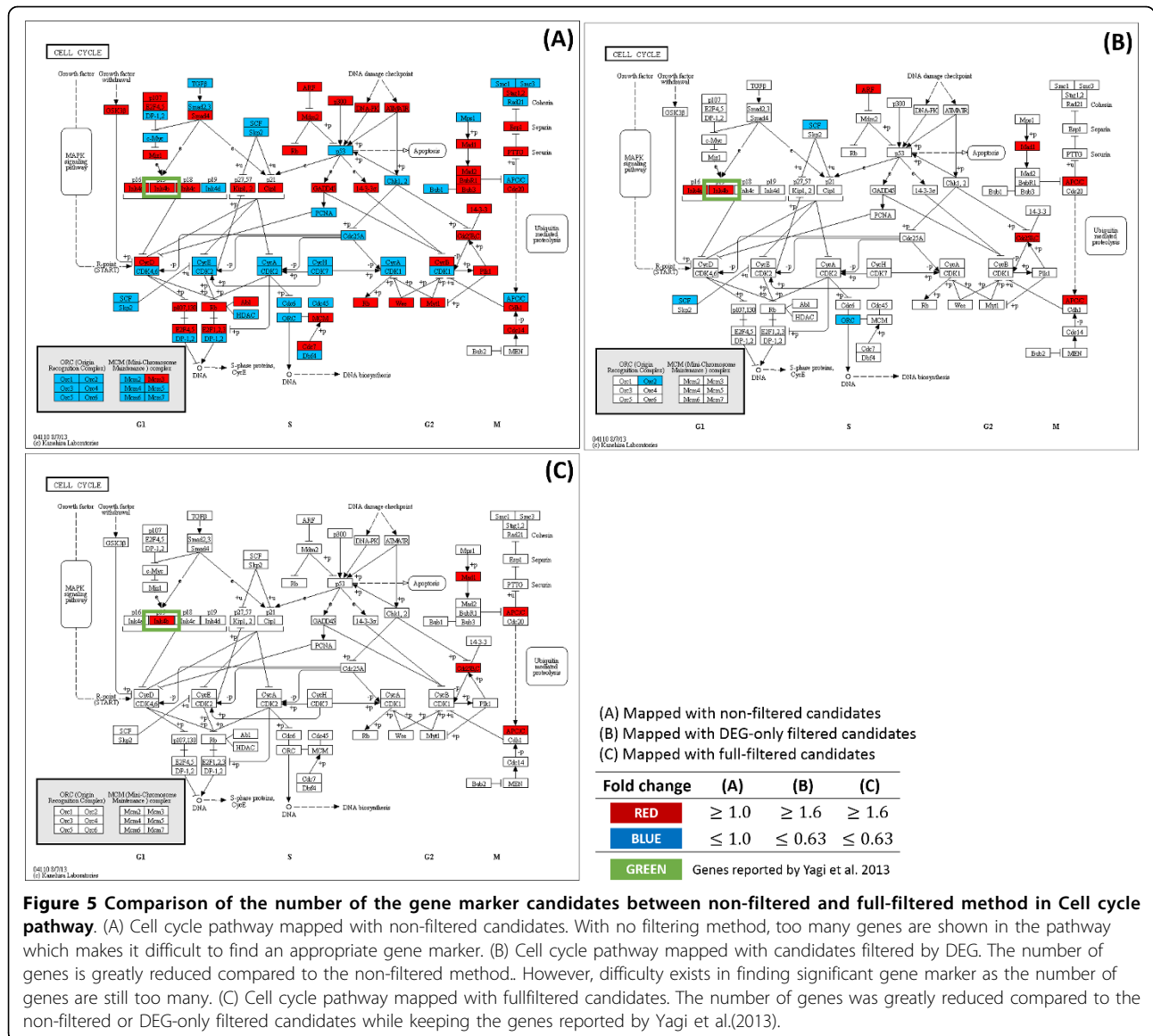
Figure 5 Comparison of the number of the gene marker candidates between non-filtered and full-filtered method in Cell cycle pathway. (A) Cell cycle pathway mapped with non-filtered candidates. With no filtering method, too many genes are shown in the pathway which makes it difficult to find an appropriate gene marker. (B) Cell cycle pathway mapped with candidates filtered by DEG. The number of genes is greatly reduced compared to the non-filtered method.. However, difficulty exists in finding significant gene marker as the number of genes are still too many. (C) Cell cycle pathway mapped with fullfiltered candidates. The number of genes was greatly reduced compared to the non-filtered or DEG-only filtered candidates while keeping the genes reported by Yagi et al.(2013).

this as a transcription factor list. For the gene list, we simply defined it as a list of whole genes that includes not only transcription factors and co-factors but also non-transcription factors. As a result NARROMI constructed a network topology of 2950865 edges. We support a URL for the network topology file which was used in this study. (http://epigenomics.snu.ac.kr/mouse_network/total_mouse.topology)

Preprocess of RNA-seq
An annotation file and reference genome from ENSEMBL (Mus musculus.GRCm38.70) [22] were used. Samples that have technical replicates were integrated into a single file and trimmed by using Trim Galore [23]. Trimmed sequences were then aligned by using Tophat [24] with '-G' option. After sequences were aligned, we used

Cuffdiff [25] by inputting biological replicates with a default option. Genes that have FPKM under 1 are excluded, since these genes tend to have high fold changes and most of them are artifacts of low gene expression values.

SNV Calling
Samtools mpileup [26] was used for SNV calling. We used accepted hits.bam which was created by Tophat results. Options used in Samtools were 'mpileup −uf' and 'view -bvcg' for bcftools. Results of Samtools were filtered by varFilter with '-D100' option. As a result, 4 vcf files (WT1, WT2, KO1, KO2) were produced. In order to observe SNV differences between WT and KO, we discarded insertions and deletions from each vcf file. Then we counted the number of

Table 2. Performance comparison of gene marker prediction methods.

Gene Symbol	NONE	DEG	DEG+GRN	DEG+GRN+Pathway	DEG+GRN+Pathway+SNV
Relb	HIT	HIT	HIT	HIT	HIT
Nfkb2	HIT	HIT	HIT	HIT	HIT
Tnfrsf9	HIT	HIT	HIT	HIT	HIT
Tnfrsf21					
Icos	HIT	HIT	HIT	HIT	HIT
Il2ra	HIT	HIT	HIT	HIT	HIT
Cysltr1	HIT	HIT			
Kit	HIT	HIT	HIT	HIT	HIT
Il1r2	HIT	HIT	HIT	HIT	
Il13	HIT	HIT	HIT	HIT	HIT
Il5	HIT	HIT	HIT	HIT	HIT
Areg	HIT	HIT	HIT	HIT	HIT
Il1rl1	HIT	HIT	HIT		
Ccr8	HIT	HIT			
Tph1	HIT	HIT			
Htr1b					
Cd244	HIT	HIT	HIT	HIT	
Lta	HIT	HIT	HIT	HIT	
Il10	HIT	HIT	HIT	HIT	HIT
Tnf	HIT	HIT	HIT	HIT	HIT
Nfkbia	HIT	HIT	HIT	HIT	
Cdkn2b	HIT	HIT	HIT	HIT	HIT
Lif	HIT	HIT	HIT	HIT	HIT
Il2ra	HIT	HIT	HIT	HIT	HIT
Il9r	HIT	HIT			
Il24					
Num of Hits	23	23	19	18	14
Selected Candidates	12298	2153	1184	478	343
Precision	0.002	0.011	0.016	0.038	0.041
Recall	0.885	0.885	0.731	0.692	0.538
F-measure	0.004	0.021	0.032	0.073	0.076
Accuracy	0.002	0.827	0.905	0.962	0.972

mutations of each gene in each vcf file and calculated the average of mutation in WT and KO. Finally, we compared the rate of mutations for each gene between WT and KO, mutation rates (fold change) greater than 2 were considered as possible genetic variation effects.

Authors' contributions
SK designed and supervised the research project and edited the paper. HC contributed in SNV analysis and advised helpful information in preprocess of RNA-seq. Other data process, data analysis, paper work was done by BH.

Authors' details
[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Daehak-dong, 151-742 Seoul, Korea. [2]Department of Computer Science, School of Informatics and Computing, Indiana University, 150 S. Woodlawn Avenue, 47404 Bloomington, USA. [3]Department of Computer Science and Engineering, Seoul National University, Daehak-dong, 151-742 Seoul, Korea.

Published: 29 May 2015

References
1. Frazee AC, Sabunciyan S, Hansen KD, Irizarry RA, Leek JT: **Differential expression analysis of rna-seq data at single-base resolution.** *Biostatistics1* 2014, 53.

2. Oshlack A, Robinson MD, Young MD, *et al*: **From rna-seq reads to differential expression results.** *Genome biol* 2010, **11**(12):220.
3. Tarazona S, García-Alcalde F, Dopazo J, Ferrer A, Conesa A: **Differential expression in rna-seq: a matter of depth.** *Genome research* 2011, **21**(12):2213-2223.
4. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **Rna-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome research* 2008, **18**(9):1509-1517.
5. AC't Hoen P, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX, Boer JM, van Ommen GJB, den Dunnen JT: **Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms.** *Nucleic acids research* 2008, **36**(21):141-141.
6. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome biol* 2010, **11**(10):106.
7. Liu Y, Zhou J, White KP: **Rna-seq differential expression studies: more sequence or more replication?** *Bioinformatics* 2014, **30**(3):301-304.
8. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A: **Reverse engineering of regulatory networks in human b cells.** *Nature genetics* 2005, **37**(4):382-390.
9. Zhang X, Liu K, Liu ZP, Duval B, Richer JM, Zhao XM, Hao JK, Chen L: **Narromi: a noise and redundancy reduction technique improves accuracy of gene regulatory network inference.** *Bioinformatics* 2013, **29**(1):106-113.
10. Kanehisa M, Goto S: **Kegg: kyoto encyclopedia of genes and genomes.** *Nucleic acids research* 2000, **28**(1):27-30.
11. Edgar R, Domrachev M, Lash AE: **Gene expression omnibus: Ncbi gene expression and hybridization array data repository.** *Nucleic acids research* 2002, **30**(1):207-210.
12. Yagi R, Zhong C, Northrup DL, Yu F, Bouladoux N, Spencer S, Hu G, Barron L, Sharma S, Nakayama T, *et al*: **The transcription factor gata3 is critical for the development of all il-7ra-expressing innate lymphoid cells.** *Immunity* 2014, **40**(3):378-388.
13. Zuo Y, Yu G, Tadesse MG, Ressom HW: **Biological network inference using low order partial correlation.** *Methods* 2014, **69**(3):266-273.
14. Ramsey MR, Wilson C, Ory B, Rothenberg SM, Faquin W, Mills AA, Ellisen LW, *et al*: **Fgfr2 signaling underlies p63 oncogenic function in squamous cell carcinoma.** *The Journal of clinical investigation* 2013, **123**(8):3525-3538.
15. Yun B, Anderegg A, Menichella D, Wrabetz L, Feltri ML, Awatramani R: **Microrna-deficient schwann cells display congenital hypomyelination.** *The Journal of Neuroscience* 2010, **30**(22):7722-7728.
16. Shan M, Yuan X, Song Lz, Roberts L, Zarinkamar N, Seryshev A, Zhang Y, Hilsenbeck S, Chang SH, Dong C, *et al*: **Cigarette smoke induction of osteopontin (spp1) mediates th17 inflammation in human and experimental emphysema.** *Science translational medicine* 2012, **4**(117):117-91179.
17. Bae Y, Yang T, Zeng HC, Campeau PM, Chen Y, Bertin T, Dawson BC, Munivez E, Tao J, Lee BH: **mirna-34c regulates notch signaling during bone development.** *Human molecular genetics* 2012, **21**(13):2991-3000.
18. Nusinow DP, Kiezun A, O'Connell DJ, Chick JM, Yue Y, Maas RL, Gygi SP, Sunyaev SR: **Network-based inference from complex proteomic mixtures using snipe.** *Bioinformatics* 2012, **28**(23):3115-3122.
19. Ruan MZ, Erez A, Guse K, Dawson B, Bertin T, Chen Y, Jiang MM, Yustein J, Gannon F, Lee BH: **Proteoglycan 4 expression protects against the development of osteoarthritis.** *Science translational medicine* 2013, **5**(176):176-3417634.
20. Du P, Kibbe WA, Lin SM: **lumi: a pipeline for processing illumina microarray.** *Bioinformatics* 2008, **24**(13):1547-1548.
21. Zhang HM, Chen H, Liu W, Liu H, Gong J, Wang H, Guo AY: **Animaltfdb: a comprehensive animal transcription factor database.** *Nucleic acids research* 2012, **40**(D1):144-149.
22. Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, *et al*: **Ensembl 2014.** *Nucleic acids research* 2013, 1196.
23. **Trim Galore.** 2014 [http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/], Accessed: May 22.
24. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL: **Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.** *Genome Biol* 2013, **14**(4):36.
25. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L: **Differential analysis of gene regulation at transcript resolution with rna-seq.** *Nature biotechnology* 2013, **31**(1):46-53.
26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, *et al*: **The sequence alignment/map format and samtools.** *Bioinformatics* 2009, **25**(16):2078-2079.