



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

의학박사 학위논문

Characterization of  
genomic perturbation sensitivity  
using 1000 genomes population

1000 지놈 인구집단을 이용한 유전자  
교란 민감도 분석

2018 년 2 월

서울대학교 대학원  
의학과 의과학 전공  
임 재 현

Characterization of  
genomic perturbation sensitivity  
using 1000 genomes population

지도교수 김 주 한

이 논문을 의학 박사학위논문으로 제출함

2017 년 10 월

서울대학교 대학원  
의학과 의과학 전공  
임 재 현

임 재현의 박사학위논문을 인준함

2018 년 1 월

위 원 장 \_\_\_\_\_ (인)

부 위 원 장 \_\_\_\_\_ (인)

위 원 \_\_\_\_\_ (인)

위 원 \_\_\_\_\_ (인)

위 원 \_\_\_\_\_ (인)

## Abstract

# Characterization of genomic perturbation sensitivity using 1000 genomes population

임재현(Lim Jae Hyun)

의학과 의과학 전공(The Department of Biomedical Sciences)

College of Medicine

Seoul National University

**Purpose:** Transcriptome is perturbed by millions of genomic variants which could alter function of cells and phenotypes of organisms. As discovered in recent large Next-Generation Sequencing (NGS) project, Individual genome has at least 3 to 4 million variants. Here, we applied perturbation network to human data from 1000 genomes project data for interpreting genetic perturbation and characterized perturbation sensitive and tolerant genes.

**Methods:** We integrated SIFT score of non-synonymous variants to calculate gene deleteriousness score and determine whether gene is perturbed or not. Perturbation network was constructed based on gene deleteriousness score and perturbation sensitivity was defined as in-degree of perturbation network. We categorized genes based on perturbation sensitivity and investigated evolutionarily, regulatory, and clinical properties of perturbation sensitive and tolerant genes.

**Results:** Perturbation sensitive genes were in periphery of protein interaction network but evolutionarily conserved. They were regulated by less miRNA and

transcription factor and played a key role in cell-cell interaction. Out-degree of perturbation network did not show any significant biological properties. Lethal genes were in periphery of perturbation network and hub of protein interaction network. On the contrary, most disease genes were in hub of perturbation network and showed various trends in protein-interaction network. We drew joint network map and categorized disease by degree of both network.

**Conclusions:** As in yeast perturbation network, perturbation sensitive genes were essential in survival of organism since they were evolutionarily conserved and related to interaction between cells. We confirmed that in-degree of perturbation network is better than out-degree of perturbation network for interpreting genetic perturbation. Disease genes can be categorized and visualized using both protein-interaction network and perturbation network. In conclusion, perturbation sensitivity was valuable measure for interpreting genetic perturbation and assessing gene's biological and clinical properties.

**Key word:** Genetic Perturbation, Transcriptome, Protein interaction network, Disease gene

**Student number:** 2010-21974

# Contents

<b>Abstract</b> .....	i
<b>Contents</b> .....	iii
<b>List of Figures</b> .....	v
<b>List of Tables</b> .....	vi
<b>1. Introduction</b> .....	1
1.1. Definition of Genetic Perturbation .....	1
1.2. Interpretation of genetic perturbation causing variants.....	2
1.3. Interpretation of genetic perturbation using biological networks .....	3
1.4. Perturbation Network approach in Yeast.....	5
1.5. Purpose of study.....	6
<b>2. Materials and Methods</b> .....	7
2.1. Genome and transcriptome data from 1000 genomes populations.....	7
2.2. Calculating gene deleteriousness scores.....	8
2.3. Construction of perturbation network.....	9
2.4. Construction of Protein Interaction Network.....	10
2.5. Retrieving biological information for gene annotation.....	10
2.6. Excess retention.....	11
2.7. Joint network map for visualization of gene sets.....	11
2.8. Clinical annotation of PSN.....	12
<b>3. Results</b> .....	13
3.1. Building Perturbation network .....	13
3.2. Biological properties of perturbation network.....	16
3.2.1. Correlation between perturbation network and PPI network.....	16

3.2.2. Relationship of perturbation network to Evolutionary feature and regulatory feature .....	1 8
3.3. Clinical implication of perturbation network against PPI network .....	3 0
3.3.1. Lethal genes versus disease genes .....	3 0
3.3.2. Disease gene classification using both Kppi and Kin.....	3 5
<b>4. Discussion.....</b>	<b>3 9</b>
<b>5. References.....</b>	<b>4 3</b>

## List of Figures

Figure 1. Flow chart for construction of perturbation network .....	8
Figure 2 Distribution of damaging score in HG00096.....	1 4
Figure 3. Histogram for $\log_2(K_{in})$ and $\log_2(K_{out})$ .....	1 5
Figure 4. Relation between perturbation network and PPI network .....	1 7
Figure 5. Biological properties of $K_{in}$ .....	1 9
Figure 6. Biological properties of $K_{out}$ .....	2 0
Figure 7. Results from various threshold of gene damaging.....	2 2
Figure 8. Joint grid diagram for lethal genes and disease genes .....	3 1
Figure 9. Joint grid diagram for lethal genes and schizophrenia genes.....	3 2
Figure 10. Joint heatmap for lethal genes.....	3 3
Figure 11. Joint heatmap for disease genes.....	3 4
Figure 12. Joint heatmap for visualizing 4 disease gene group: asthma, sclerosis, schizophrenia, and cancer .....	3 5
Figure 13. Joint heatmap for visualizing disease gene categories classified in GAD. .....	3 7
Figure 14. Joint heatmap for visualizing disease gene categories classified in GAD(continued). .....	3 8
Figure 15. Disease classification.....	3 9

## List of Tables

Table 1. Pathway annotation of perturbation sensitive gene using DAVID	
GOTERM_BP_FAT.....	2 4
Table 2. Pathway annotation of perturbation sensitive genes using DAVID	
GOTERM_CC_FAT.....	2 5
Table 3. Pathway annotation of perturbation tolerant genes using DAVID	
GOTERM_BP_FAT.....	2 6
Table 4. Pathway annotation of perturbation tolerant genes using DAVID	
GOTERM_CC_FAT.....	2 7
Table 5. Pathway annotation of perturbation causing genes using DAVID	
GOTERM_BP_FAT.....	2 8
Table 6. Pathway annotation of perturbation causing genes using DAVID	
GOTERM_CC_FAT.....	2 9

# 1. Introduction

## 1.1. Definition of Genetic Perturbation

Human genome is complex system with great diversity. Sequencing of individual genome by Next Generation Sequencing (NGS) technology reveals that each individual possesses at least 3 to 4 million variants (1-3). The 1000 Genomes Project Consortium report genomes of 1,092 individuals sampled from 14 populations drawn from 4 continents. They provide validated haplotype map of 38 million single nucleotide polymorphisms, 1.4 million short insertions and deletions, and more than 14,000 larger deletions. They also revealed that individuals from different populations have different profiles of common and rare variants distribution and not few of those are protein disrupting or located within transcription factor binding site. Moreover, MacArthur et al. (4) finds out that after stringent filtering, at least ~ 100 genuine LoF variants are present in 'healthy' genomes. These variants cause structural and functional change of protein which might affect cis- or trans- gene expression (5-7).

Genetic perturbation is defined as perturbation caused by such variants whether they actually affect phenotypic change or not. An organism must stay in homeostasis and mitigate ripple effect of genetic perturbation, in common with the other kinds of stress like environmental change, endocrine signal, bacterial infection, even physical stress. Genes frequently perturbed by genetic perturbations are considered as stress-responsive genes. Interpreting consequences of those genetic perturbation is important for understanding genotype-phenotype relationship. Within the framework of biological system, genetic perturbation consists of two components: perturbing genes and perturbed genes. Interpretation of genetic perturbation need to analyze either perturbing or perturbed genes.

## 1.2. Interpretation of genetic perturbation causing variants

Most easy and direct way to understand genetic perturbation is direct annotation of variants. There are many traditional ways to annotate variants to predict functional change of protein, especially within coding regions. SIFT (8), PolyPhen (9, 10), and PhastCons (11) is example of classic methods for calculate variant function prediction score, which predict whether nonsynonymous changes are likely to have a deleterious effect on protein function (12). They use sequence homology of related proteins to predict whether an amino acid substitution is likely to be deleterious to protein function based on the degree of conservation of the affected base throughout evolution.

Currently, filtering and annotating variants with reported phenotype is another method to interpret variants (13, 14). dbSNP (15, 16) was established by 1999 and contains almost all reported variants within human genome. By 2008, dbSNP build 129 contained approximately 11 million single nucleotide polymorphisms and 3 million short insertions and deletions. Genome-Wide Association Studies (GWAS) reveals many SNPs which related to complex traits as diabetes, inflammatory bowel disease, cancer and etc. Many studies use SNPs reported in GWAS catalog and dbSNP to filter out (17-19).

Major drawback of direct variant annotation is lack of interpretability, since gene is the main unit of all biological process. Without correlating variant-level annotation to gene-level annotation, variant itself does not have any biological meaning. Lee et al. (20) suggests simple but effective methods to compute gene-level deleteriousness score using variant-level score. They use SIFT algorithm to compute the variant score and define gene deleteriousness score as the geometric mean of variant scores for all nonsynonymous coding variants of the gene to evaluate. They utilize gene-score scheme to pharmacogenomics area and proved robustness and effectiveness of their methods. We utilize that scheme to evaluate how gene is damaged by variants within coding region.

Another important consideration is emerging multi-omics data, like transcriptome or methylome, with genome data to understand underlying biological link connecting genotype to phenotype. For example, Lappalainen et al. (21) sequenced mRNA and small RNA from lymphoblastoid cell line samples from 1000 genomes population. They infer genetic effects on the transcriptome using eQTL scheme and as a result, they reveal high resolution variant map which affect transcriptome. Although their analysis and results were similar to those of previous eQTL analysis, the dataset they built were important since it originated from 1000 genomes population. Before and after then, many eQTL analysis reveals that variants related to gene expression have strong tendency to affect phenotypes, especially disease (22-25). Most eQTL analysis focus on variants that cause perturbations and characteristics of perturbed genes are often neglected. More importantly, most of eQTL analysis focus on variants located in regulatory region, not coding region. Therefore, eQTL can tell nothing about protein structure and function change.

### **1.3. Interpretation of genetic perturbation using biological networks**

Genetic perturbation not only causes single protein malfunction but also affects related proteins. Also, genes are not affected by single genetic perturbation but multiple events. Thus, systemic perspective is needed to incorporate such many-to-many relations. Network, graph in mathematics, is used to model pairwise relations between objects and adopted to biology to understand the structure and the dynamics of the complex intercellular web of interactions underlying biological and clinical events (26, 27).

PPI network is one of the oldest and well-studied network in biology (28-32). Node of PPI network is protein and edge of PPI network is their physical interaction. Hub of PPI network is known to be essential in biological process and essential genes

tend to have higher degree of PPI network. For example, Jeong et al. (29) claimed that lethal genes, which cause death of organism when it is disrupted, have higher PPI degree and located in hub of PPI network. Meanwhile, general properties of gene related to disease in PPI network are open to dispute. Disease genes does not show such strong tendency. Goh et al (30) claimed that there are no statistical relationship between disease and degree of PPI network. On the contrary, Ideker et al. (28) reported that disease genes forming a functional modular structure and has higher PPI degree. According to the work of Zhongming Zhao (33), cancer genes have the highest PPI degree followed by lethal genes, schizophrenia genes and neuro genes.

It is natural that PPI network is insufficient to characterize disease genes for many reasons. One is that since death of single cell does not trigger death of organism, disruption of lethal gene, mostly hub of PPI network, does not guarantee that it affects tissue- and higher architecture of organism. Because PPI network only reflects physical interaction of 'static' state of proteins and does not reflect cellular state which are extremely different because of differentiation and regulatory network. Especially, cell-cell interaction cannot be explained by inner-cellular protein network. Also, PPI network does not reflect relationship between genotype and transcriptome.

eQTL network can break those limitations (34-36). eQTL network is bipartite directed network which nodes are variants and genes. Edges are given when a variant within a gene affect expression of other gene. eQTL network is powerful tool to predict downstream effects of many trait-associated variants (23). It reflects tissue-specificity and cellular state (37). But still, eQTL network is indirect method to infer regulation and most importantly, tell nothing about protein function change of gene. Also, since eQTL network is built on variant-level analysis, interpretation of network cannot easily extend to gene level annotation.

## 1.4. Perturbation Network approach in Yeast

All of above mentioned methods focused on perturbing genes, not perturbed genes. Since cell is under strict regulation (38, 39), It is hard to say that whether a variant affects expression of certain gene or that gene is allowed to be affected. For example, tissue or cell-cycle specific regulations often change the relationships between variants and genes (40-42). To understand phenotypic effect of genetic perturbation, it is easy to evaluate perturbed genes, which reflects consequence of genetic perturbation after regulation and close to final functional change.

Perturbation network is effective way to measure how many gene's perturbed by certain perturbation (43, 44). Ohn et al. defined perturbation network as non-directed bipartite graph of two node groups; a group of 'genes' which showed significant changes in transcription level in the other group of 'deletion mutants' and links are made between nodes from each group based on the significance level assuming the 'error model'. Perturbation sensitivity is defined as in-degree of gene node group. They constructed network from genome-wide transcriptional profiling study of 300 perturbation experiments like gene deletions or drug treatments in *Saccharomyces cerevisiae*, yeast. They found out that perturbation sensitive genes are usually not essential, and their coding proteins have fewer physical interaction partners and more transcription factors bind to their upstream sequences.

Han et al. use same network and did more comprehensive analysis. They correlate perturbation network to PPI network and claimed that they are reciprocally correlated to each other. They found out that hub of perturbation network and PPI network are both evolutionarily conserved and their degree are negatively correlated. They also found out that PPI hubs are highly enriched with lethal genes but not disease genes, whereas perturbation network hubs are highly enriched with disease genes and not with lethal genes.

Perturbation network was well studied only in yeast, not human. Han et al. use yeast - human homolog genes to map disease genes to yeast genes but yeast - human

homolog genes are special itself because they are highly conserved genes. Moreover, they only analyse in-degree of network since gene node group only have in-degree. Finally, disease genes are highly heterogeneous group so categorizing those genes into subgroups may improve the results. As in PPI network, it is possible that different disease genes tend to have different centrality in perturbation network.

## **1.5. Purpose of study**

In this research, we examined the perturbation sensitivity of genes and characterize perturbation sensitive and tolerant genes. To this end, we build a perturbation network using genome and transcriptome data of 1000 genomes population. 1000 genomes populations have enough genetic perturbations to perform a macroscopic analysis of perturbation sensitivity (1, 4, 45-47). 421 samples from the 1000 genomes populations also have RNA-seq data (21). We slightly modify classical perturbation network to apply this model to healthy human population data. In our analysis, perturbation network can be simply viewed as a directed graph of genes. We adopt concept of gene deleteriousness score to distinguish damaged genes from normal genes and build perturbation network to calculate perturbation sensitivity, which defined as in-degree of perturbation network. We evaluated biological characteristics of perturbation tolerant, perturbation sensitive, and perturbation causing genes and compare it to PPI network. We also evaluated lethal genes and various disease genes using perturbation sensitivity and the PPI network degree. As a result, we characterize genes with respect to perturbation sensitivity and suggest measure for predicting phenotypic effect due to genetic perturbation.

## **2. Materials and Methods**

### **2.1. Genome and transcriptome data from 1000 genomes populations.**

The 1000 Genome Projects phase 1 dataset includes genomes of 1,026 individual samples from 14 subpopulations drawn from Europe, East Asia, sub-Saharan Africa and America. Within these samples, 462 samples from five populations have transcriptome data (21): the CEPH (CEU, n = 78), Finns (FIN, n = 89), British (GBR, n = 85), Toscano (TSI, n = 92) and Yoruba (YRI, n = 77). Overall, 421 samples have both genotype and transcriptome data. The genotype data were obtained from 1000 genomes project page and transcriptome data were obtained from Geuvadis consortium page.

We did functional reannotation of the variants from genotype data with VAT (48). We filter non-coding variants which are expected to have neutral effect on the protein function, and annotates the variant type using hg19 GENCODE (49) v12 as the genome reference. We only focus on autosomal variants and excluded variants located in sex chromosomes and indels. We also ruled out all synonymous variants and variants with 5% or less minor allele frequency for selecting common damaging variants. Overall, 327879 loci which at least 1 sample have nonsynonymous variants were included for further analysis.

Transcriptome data were obtained from the Geuvadis RNA sequencing project. We considered FPKM (Fragments per kilo base of exon per million fragments mapped) as the expression level of each gene. We used upper quantile normalization to remove between sample difference. For further analysis, we only include 344 European subjects to minimize bias from race difference.

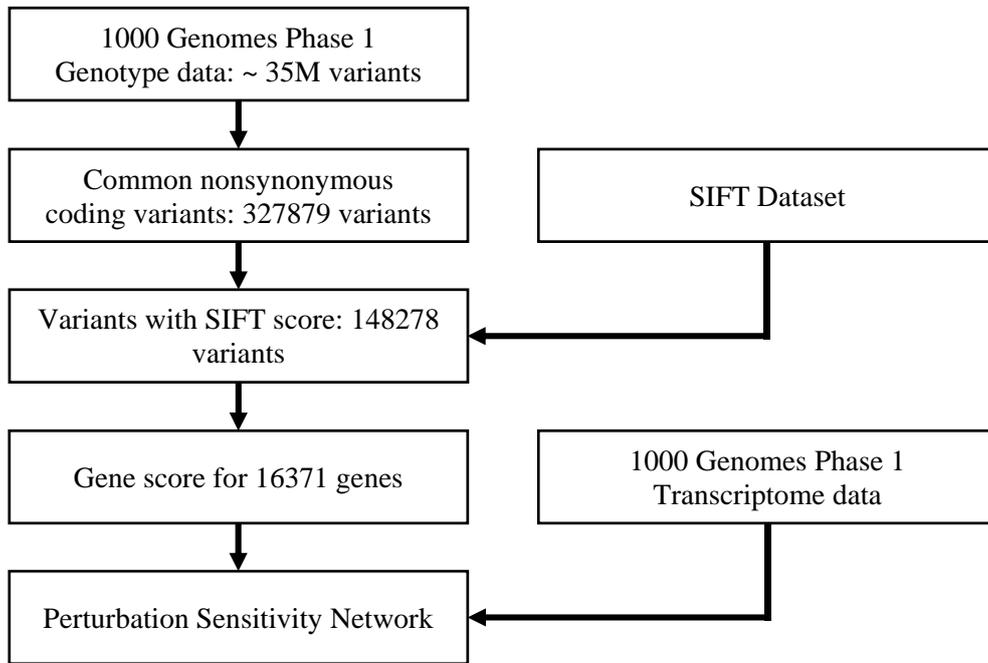


Figure 1. Flow chart for construction of perturbation network

## 2.2. Calculating gene deleteriousness scores.

To evaluate deleteriousness of variants within coding-region, we use gene deleteriousness score (20) to evaluate how coding region variants affects corresponding protein's function. Gene deleteriousness scores was defined as the geometric mean of a variant score which reflects combined effect of multiple nonsynonymous variants within a gene. We use ANNOVAR (50) to annotate variant using SIFT (8) algorithms. ANNOVAR also provides other variant function prediction scores as PolyPhen (9), LRT (51), Mutation Taster (52), CADD (53). We use variants with SIFT score lower than 0.7 and 16371 genes have at least one variant in at least one sample. We assume that a gene is damaged when combined gene damaging score is lower than 0.3. 2930 genes were damaged in each person, on average. To confirm our score, we tested multiple hypothesis for thresholds that decides damaging variants and deleterious gene. we also tested other variant function prediction score as PolyPhen. Finally, we tested classical ways to determine damaging of gene, that If at least one variants with SIFT score  $< 0.05$  within a genic region then we consider that gene as damaged.

## 2.3. Construction of perturbation network.

To evaluate the effect of damaging genes, we built perturbation network derived from genome and transcriptome data of 1000 genomes projects populations mentioned above. Perturbation network is a directed network, that each node represents a protein-coding gene and each edge represents perturbation of successor gene expression due to damaging of its predecessor. We used student's t-test to infer relationships between nodes with python statsmodel library. Edges were given only if p-value  $< 0.001$  in t-test. We summarized flow chart for construction of perturbation network from 1000 genomes dataset in figure 1. We define  $K_m$ ,

perturbation sensitivity, as in-degree of perturbation network for 23722 genes which have transcriptome data. We also investigate out-degree of network,  $K_{out}$ , for 8895 genes which satisfied criteria for t-test. Out-degree of perturbation network is calculated only for gene which is damaged at least 5 samples. For further analysis, we categorized genes to 3 groups according to  $K_{in}$ . Perturbation sensitive genes have top 10% of  $K_{in}$  and perturbation tolerant genes bottom 10% of  $K_{in}$ . Neutral genes are the rest.

## **2.4. Construction of Protein Interaction Network.**

Human protein-protein interaction data were retrieved from BIOGRID (54) (release 3.4.138). Data were processed as follows: (i) only proteins in the perturbation networks were used; and (ii) only interactions of the ‘physical association’ type were used. As a result, PPI network was built as undirected network with 20249 nodes and 239932 edges with a node degree range from 1 to 498.

## **2.5. Retrieving biological information for gene annotation**

We retrieve biological database to evaluate properties of perturbation sensitive and tolerant genes.  $dN/dS$  (evolutionary rate) is the ratio of the number of non-synonymous substitutions per non-synonymous site ( $dN$ ) to the number of synonymous substitutions per synonymous site ( $dS$ ), which can be used as an indicator of selective pressure acting on a protein-coding gene (55, 56). We obtained  $dN/dS$  values of mouse-human homolog genes from BiomaRt (57). A number of paralogs and miRNA for each gene were obtained from Ensembl using BiomaRt. Each gene has 5.95 paralogs on average. We retrieved ENCODE ChIP-Seq data from UCSC data portal (41). Data were consisted of transcription factors and their binding sites. We included transcription factors that bind to a region 10 kb above each gene

and used them as the number of upstream binding transcription factors for each gene. We also conduct gene set enrichment test using DAVID (58).

## 2.6. Excess retention.

‘Excess retention’ is defined as the degree to which genes with a certain property A is over- or under-represented in m-core compared with that in the whole gene groups (59). M-core of a graph G is defined as a maximal connected subgraph of G in which all vertices have a degree of at least m (43). The fraction of genes with property A in the whole group with N genes is  $EA = NA / N$ . If m-core contains  $Nm$  genes and the number of genes with property A in m-core is  $NmA$ , then the excess retention of the genes with property A in m-core is given by  $EmA = (NmA / Nm) / (NA / N)$ .

$$E_m^A = \frac{(N_m^A / N_m)}{(N^A / N)}$$

## 2.7. Joint network map for visualization of gene sets.

To characterize and visualize gene sets using both PPI network and perturbation network, we draw joint network map with degree of both network. We draw grid diagram consists of pie charts for comparing lethal genes versus disease genes. x-axis indicate perturbation network bins and y-axis indicate PPI network bins. We binned genes by degree of perturbation network in equally spaced intervals and by logarithms of degree of PPI network in equally spaced intervals. We also draw heatmap with same axis and less bin numbers, 5. We binned genes by degree of perturbation network in equally spaced intervals and by logarithms of degree of PPI network in equally spaced intervals. On average, each grid consists of 566 genes. Data matrix contains odd ratios for gene set of interest

## **2.8. Clinical annotation of PSN**

We performed enrichment analysis of perturbation network by genes with clinical importance. Disease genes were retrieved from the GWAS catalogue (18) and OMIM (60). Lethality information of the mouse genes were retrieved from the MGI project page (61). Human orthologs of mouse lethal genes were regarded as human lethal genes. Cancer genes were retrieved from the COSMIC cancer gene census list (62). We also use GAD (63) for categorize disease genes.

## 3. Results

### 3.1. Building Perturbation network

To construct perturbation network, we calculate gene deleteriousness score for each gene to determine whether function of corresponding protein has been changed. Gene deleteriousness score is calculated from SIFT Score using geometric mean of SIFT score of variants within a gene. As a result, each individual possesses 1841 variants with SIFT score  $< 0.05$  and 2930 genes with gene deleteriousness score  $< 0.3$ . Figure 3(a) and 3(b) shows SIFT score and Gene score in one individual, HG00096. Since we assume gene deleteriousness score to 1.0 if there are no variants with SIFT score  $< 0.7$ , histogram of gene score shows higher proportion at bin of 1.0. Otherwise two histograms show similar distribution.

Next, we use gene deleteriousness score to construct perturbation network (Materials and methods.). We use t-test to determine whether gene expression is affected by gene deleteriousness score. Perturbation network consist of 23722 nodes and 504091 edges, with average degree of 21.25. All nodes are weakly connected with average shortest path length of 0.2, which implies that every gene is closely located to each other. We analyse in-degree and out-degree separately to investigate which genes are perturbation sensitive genes and which genes are perturbation causing genes. In degree of perturbation network, denoted as  $K_{in}$ , is ranged from 0 to 166. On the contrary, out-degree of perturbation network, denoted as  $K_{out}$ , is ranged from 0 to 1205. Notice that only 6773 genes have out-degree due to limited sample size. As shown in degree-distribution plots, Figure 3 (a), (b), both  $K_{in}$  and  $K_{out}$  shows log-normal distribution and p-value of shapiro-wilk test for log of  $K_{in}$  and  $K_{out}$  is  $4.76e^{-34}$  and  $3.77e^{-14}$ .  $K_{in}$  and  $K_{out}$  does not show statistically significant correlation. We also construct perturbation network with other criteria to determine whether gene is damaged and did same analysis which shows similar results.

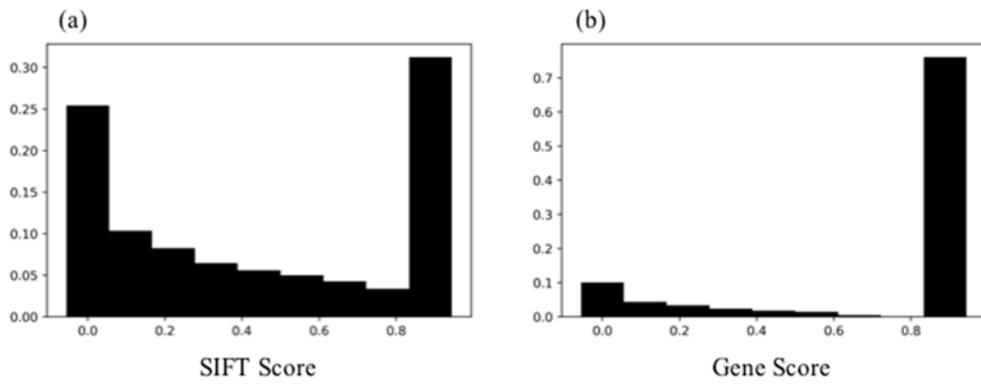


Figure 2 Distribution of damaging score in HG00096. (a) Distribution of SIFT score. (b) Distribution of gene score.

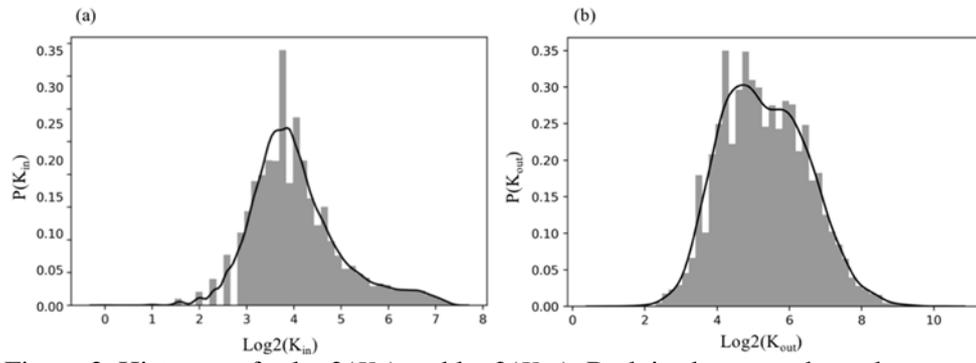


Figure 3. Histogram for  $\log_2(K_{in})$  and  $\log_2(K_{out})$ . Both in-degree and out-degree follow log-normal distribution.

## 3.2. Biological properties of perturbation network

To characterize biological properties of perturbation network, we categorize genes with  $K_{in}$  to 3 groups, perturbation sensitive genes, perturbation tolerant genes, and others. Perturbation sensitive genes are defined as genes with top 10% of  $K_{in}$  and perturbation tolerant genes are defined as genes with bottom 10% of  $K_{in}$ . We use ANOVA for statistical testing. We also use excess retention method to investigate trends of biological properties among perturbation sensitivity (Materials and Methods). We also categorize genes with  $K_{out}$  to 3 groups, perturbation causing genes, perturbation free genes, and others. Perturbation causing genes are defined as genes with top 10% of  $K_{out}$  and perturbation free genes are defined as genes with bottom 10% of  $K_{out}$ .

### 3.2.1. Correlation between perturbation network and PPI network

We compared  $K_{in}$  and  $K_{out}$  with  $K_{ppi}$ , degree of PPI network. As shown in figure 4 (a), perturbation sensitive genes have low  $K_{ppi}$  while perturbation tolerant genes have high  $K_{ppi}$  (ANOVA p-value =  $3.54e-13$ ). Figure 4 (b), Excess retention plot, also shows similar trends which X-axis represents  $K_{in}$  and Y-axis represents excess retention of corresponding perturbation sensitivity. Red line, which denotes hub of PPI network which have top 10% of  $K_{ppi}$ , diminish along  $K_{in}$ . As in previous study with yeast perturbation network (44), Genes highly connected in the protein interaction network are least likely to be hubs in the perturbation network, and vice versa. On the contrary,  $K_{out}$  does not show trends in excess retention plot (Figure 4 (d)) but ANOVA and boxplot (Figure 4 (c)) shows that Perturbation Causing genes which have top 10%  $K_{out}$  weakly tends to have high  $K_{ppi}$  (p-value = 0.056).

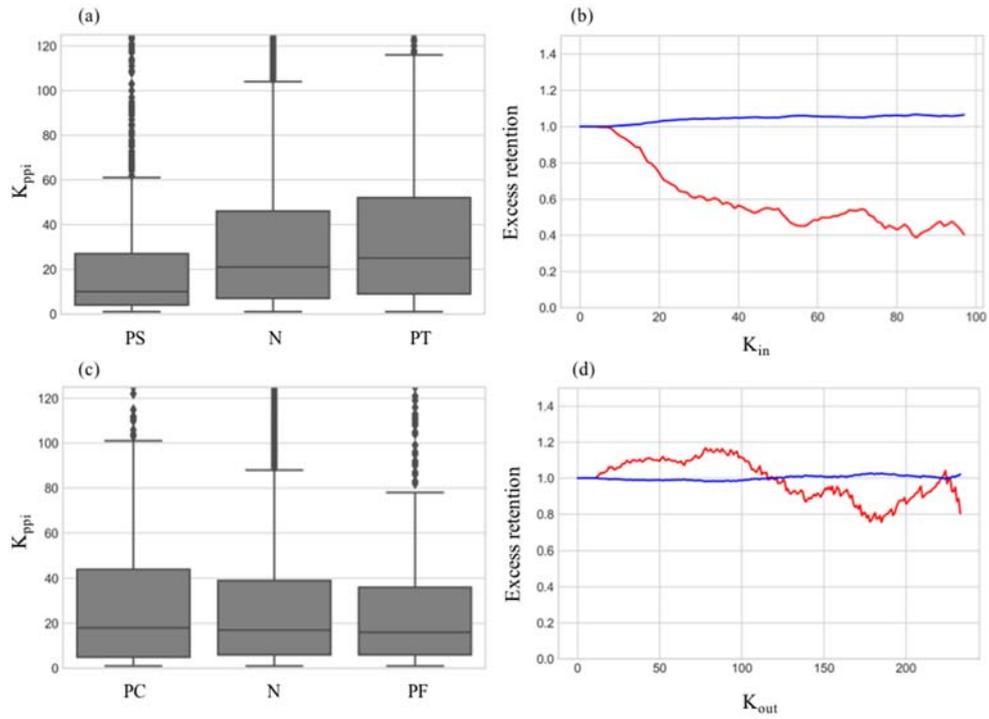


Figure 4. Relation between perturbation network and PPI network. (a) boxplot for PS(perturbation sensitive genes), N(neutral genes), PT(perturbation tolerant genes) against  $K_{ppi}$ . (b) Excess retention plot for hub of PPI network among  $K_{in}$ . (c) boxplot for PC(perturbation causing genes), N(neutral genes), PF(perturbation free genes) against  $K_{ppi}$ . (d) Excess retention plot for hub of PPI network among  $K_{out}$ .

### 3.2.2. Relationship of perturbation network to Evolutionary feature and regulatory feature

According to the ‘neutral’ theory of molecular evolution, random mutations not affecting fitness of the organism causes molecular level evolutionary change (64). Thus, proteins important for survival of the organism must have lower rate of evolution. evolutionary changes at the molecular level are caused by drift and fixation of random mutations that do not affect the fitness of the organism. For example, hubs of the PPI network have lower  $dN/dS$  ratio (65). In this analysis, we use mouse-human ortholog  $dN/dS$  to evaluate protein evolutionary rate. As previously reported,  $K_{ppi}$  shows negative correlation to  $dN/dS$  value, which implies hub of PPI networks tend to evolve slowly.  $K_{in}$  shows similar trends with  $K_{ppi}$ , that perturbation sensitive genes also have low  $dN/dS$ , or evolutionarily conserved (Figure 5 (a), (b) ANOVA p-value = 0.023). On the other hand,  $K_{out}$  does not show any significant trends with  $dN/dS$ . Just as hubs of PPI network, perturbation sensitive genes have evolved slowly, and their sequence divergences are under high evolutionary selection pressure, which implies that they both are essential for survival of organism. We also analyse relationship between number of paralogs and network degrees. Paralogs are defined as more closely related gene family members caused by gene duplication and expected to perform similar function as in the ortholog and have similar neighbour and node degrees in PPI network and regulatory network (66, 67). In our analysis,  $K_{in}$  is positively correlated to number of paralogs. (Figure 5 (c)(d), ANOVA p-value = 1.30e-8).

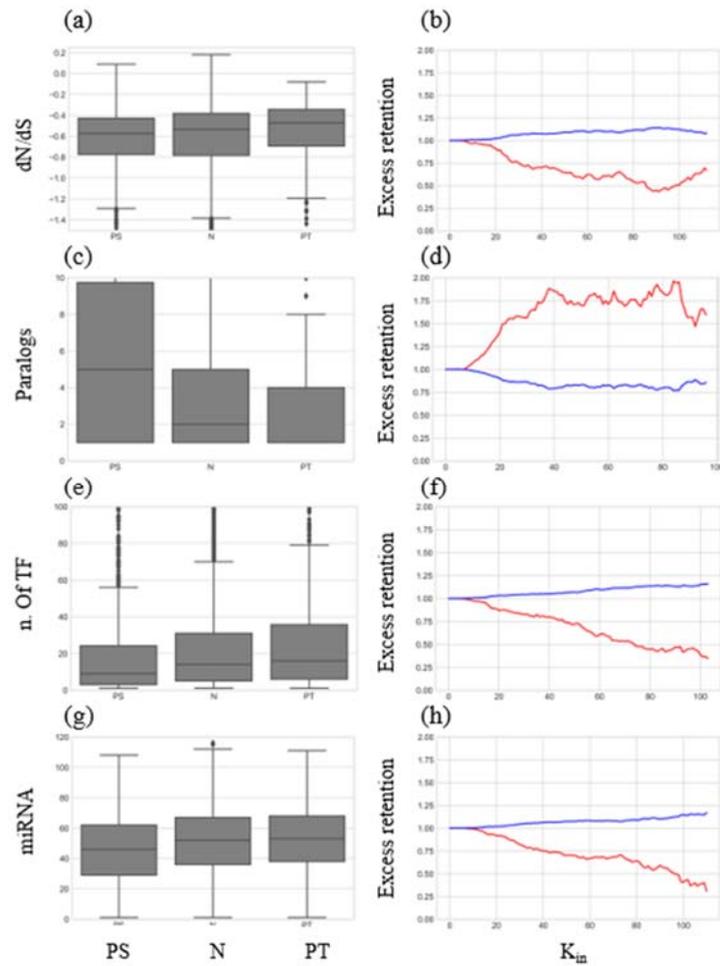


Figure 5. Biological properties of  $K_{in}$ . (a) Boxplot for  $dN/dS$ . (b) Excess retention plot for  $dN/dS$ . red line denotes evolutionarily conserved genes. (c) Boxplot for number of paralogs. (d) Excess retention plot for number of paralogs. Red line denotes genes with less paralogs. (e) Boxplot for miRNA targets. (f) Excess retention plot for miRNA targets. Red line denotes genes with less miRNA targets. (g) Boxplot for number of transcription factor. (h) Excess retention plot for number of transcription factor. Red line denotes genes with less transcription factor target.

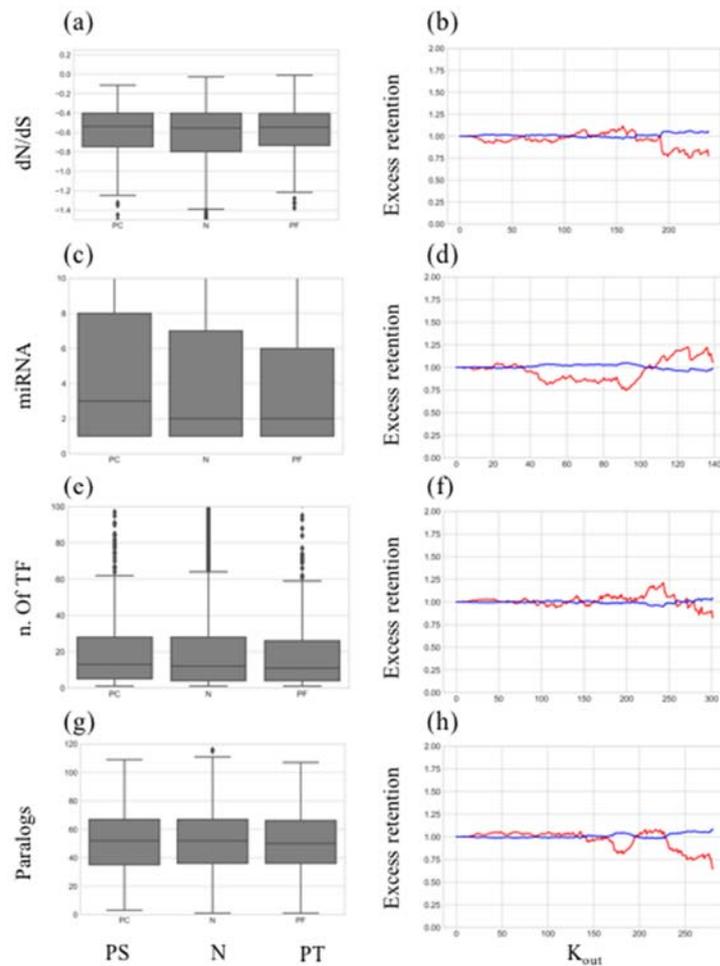


Figure 6. Biological properties of  $K_{out}$ . (a) Boxplot for  $dN/dS$ . (b) Excess retention plot for  $dN/dS$ . red line denotes evolutionarily conserved genes. (c) Boxplot for number of paralogs. (d) Excess retention plot for number of paralogs. Red line denotes genes with less paralogs. (e) Boxplot for miRNA targets. (f) Excess retention plot for miRNA targets. Red line denotes genes with less miRNA targets. (g) Boxplot for number of transcription factor. (h) Excess retention plot for number of transcription factor. Red line denotes genes with less transcription factor target.

miRNA is small non-coding RNA molecule which plays key role in regulation of gene expression (68-70). Especially, miRNA related to fine-tuning of target activity and coordinated regulation. Perturbation sensitive genes tend to be targeted by less miRNAs (Figure 5 (e), (f) ANOVA p-value =  $1.10e^{-6}$ ).  $K_{out}$  does not show significant correlation to miRNA. Number of upstream binding transcription factor grossly reflects centrality of regulatory network (41).  $K_{in}$  is negatively correlated to number of transcription factor (Figure 5 (g), (h) ANOVA p-value =  $1.8e^{-18}$ ).

$K_{out}$  does not show any significant trends. Namely, perturbation sensitive genes are controlled by less transcription factor.

In conclusion, perturbation sensitive genes have lower PPI degree but evolutionarily conserved, less targeted by miRNA and transcription factor, and have many paralogs. On the contrary,  $K_{out}$  does not show any significant biological properties (Figure 6 (a) ~ (h)). Finally, we checked robustness of biological properties of perturbation sensitivity. Figure 7 summarizes results from various threshold of gene damaging.

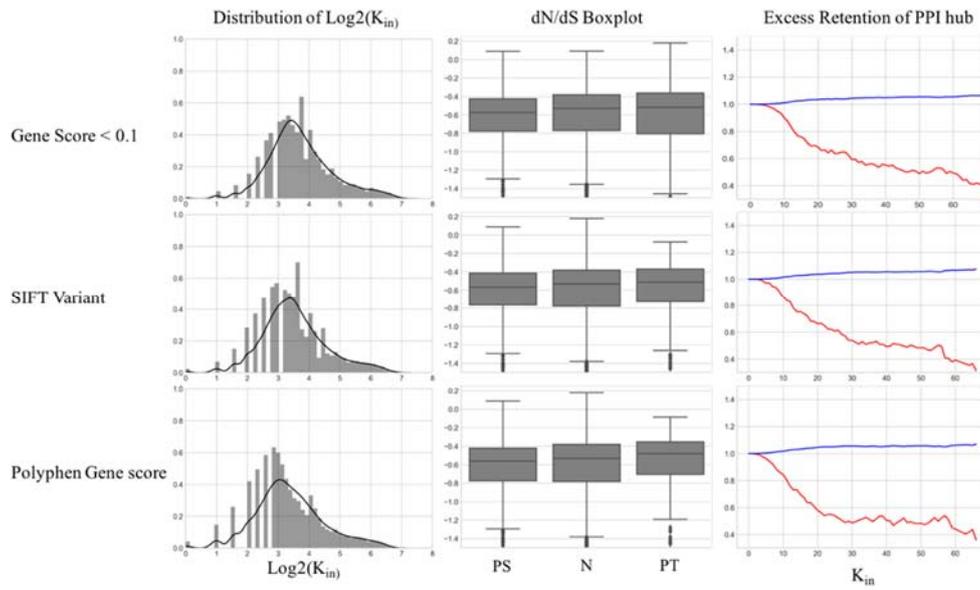


Figure 7. Results from various threshold of gene damaging. First row represents results of lower gene score, 0.1. Second row represents results of using SIFT score instead of gene score. Third row represents results of using PolyPhen score to calculate gene score. Independent from methods for deriving perturbation network, perturbation sensitive genes are evolutionarily conserved and inversely correlated to PPI network.

### **3.2.3. Functional annotation of perturbation network using GO terms**

We conduct gene set enrichment test using DAVID (58) for 3 groups of genes; genes with high perturbation sensitivity which are considered as perturbation sensitive genes, genes with low perturbation sensitivity or perturbation tolerant genes, and genes with high *Kout* which named as perturbation causing genes. Within top 20 gene sets with respect to p-value, Perturbation causing genes are annotated as ‘apoptosis’, ‘regulation of synaptic transmission’, ‘protein domain specific binding’, ‘behaviour’, ‘regulation of neurological system process’, and ‘antigen receptor-mediated signalling pathway’ with GOTERM\_BP\_FAT. Perturbation sensitive genes are annotated as ‘cell adhesion’, ‘biological adhesion’ with GOTERM\_BP\_FAT and ‘extracellular region part’, ‘plasma membrane part’, ‘basement membrane’ with GOTERM\_CC\_FAT. Perturbation tolerant genes are annotated as ‘ubiquitin-dependent protein catabolic process’, ‘M phase’, ‘mitotic cell cycle’ with GOTERM\_BP\_FAT and ‘nuclear lumen’, ‘nucleoplasm’ with GOTERM\_CC\_FAT. (Table 1 ~ 6)

Table 1. Pathway annotation of perturbation sensitive gene using DAVID

GOTERM\_BP\_FAT

<b>GOTERM_BP_FAT</b>	<b>p-value</b>
GO:0007155~cell adhesion	1.68E-12
GO:0022610~biological adhesion	1.76E-12
GO:0050865~regulation of cell activation	1.89E-05
GO:0002684~positive regulation of immune system process	5.30E-05
GO:0016337~cell-cell adhesion	6.56E-05
GO:0007267~cell-cell signaling	6.72E-05
GO:0009611~response to wounding	9.68E-05
GO:0002694~regulation of leukocyte activation	1.07E-04
GO:0042127~regulation of cell proliferation	1.34E-04
GO:0050867~positive regulation of cell activation	1.82E-04
GO:0060429~epithelium development	2.20E-04
GO:0051249~regulation of lymphocyte activation	3.02E-04
GO:0006954~inflammatory response	3.20E-04
GO:0044057~regulation of system process	3.65E-04
GO:0008015~blood circulation	4.20E-04
GO:0003013~circulatory system process	4.20E-04
GO:0002696~positive regulation of leukocyte activation	4.37E-04
GO:0006952~defense response	4.56E-04
GO:0008284~positive regulation of cell proliferation	5.23E-04
GO:0010604~positive regulation of macromolecule metabolic process	5.32E-04

Table 2. Pathway annotation of perturbation sensitive genes using DAVID

GOTERM\_CC\_FAT

<b>GOTERM_CC_FAT</b>	<b>p-value</b>
GO:0044421~extracellular region part	3.01E-12
GO:0044459~plasma membrane part	1.98E-11
GO:0005887~integral to plasma membrane	1.49E-08
GO:0005886~plasma membrane	2.16E-08
GO:0031226~intrinsic to plasma membrane	4.07E-08
GO:0005615~extracellular space	5.45E-08
GO:0005604~basement membrane	1.07E-07
GO:0031012~extracellular matrix	1.19E-07
GO:0005576~extracellular region	1.84E-07
GO:0044420~extracellular matrix part	2.07E-07
GO:0005578~proteinaceous extracellular matrix	6.30E-07
GO:0009897~external side of plasma membrane	1.31E-05
GO:0009986~cell surface	2.74E-05
GO:0045202~synapse	3.94E-05
GO:0000786~nucleosome	8.19E-05
GO:0031091~platelet alpha granule	1.68E-04
GO:0044456~synapse part	6.52E-04
GO:0032993~protein-DNA complex	0.0010
GO:0043235~receptor complex	0.0032
GO:0042995~cell projection	0.0047

Table 3. Pathway annotation of perturbation tolerant genes using DAVID

GOTERM\_BP\_FAT

<b>GOTERM_BP_FAT</b>	<b>p-value</b>
GO:0006511~ubiquitin-dependent protein catabolic process	0.0070
GO:0000279~M phase	0.0086
GO:0000278~mitotic cell cycle	0.0157
GO:0000087~M phase of mitotic cell cycle	0.0164
GO:0043009~chordate embryonic development	0.0176
GO:0009792~embryonic development ending in birth or egg hatching	0.0194
GO:0007051~spindle organization	0.0226
GO:0051006~positive regulation of lipoprotein lipase activity	0.0243
GO:0045087~innate immune response	0.0284
GO:0000280~nuclear division	0.0294
GO:0007067~mitosis	0.0294
GO:0007283~spermatogenesis	0.0316
GO:0048232~male gamete generation	0.0316
GO:0000188~inactivation of MAPK activity	0.0340
GO:0007049~cell cycle	0.0352
GO:0022402~cell cycle process	0.0370
GO:0045923~positive regulation of fatty acid metabolic process	0.0386
GO:0048285~organelle fission	0.0393
GO:0007585~respiratory gaseous exchange	0.0401
GO:0010741~negative regulation of protein kinase cascade	0.0401

Table 4. Pathway annotation of perturbation tolerant genes using DAVID

GOTERM\_CC\_FAT

<b>GOTERM_CC_FAT</b>	<b>p-value</b>
GO:0031981~nuclear lumen	0.0083
GO:0005654~nucleoplasm	0.0087
GO:0070013~intracellular organelle lumen	0.0191
GO:0043233~organelle lumen	0.0224
GO:0031974~membrane-enclosed lumen	0.0251
GO:0016605~PML body	0.0309
GO:0070652~HAUS complex	0.0321
GO:0044451~nucleoplasm part	0.0328
GO:0000159~protein phosphatase type 2A complex	0.0346
GO:0005882~intermediate filament	0.0358
GO:0000775~chromosome, centromeric region	0.0374
GO:0045111~intermediate filament cytoskeleton	0.0412
GO:0005694~chromosome	0.0493
GO:0008287~protein serine/threonine phosphatase complex	0.0568
GO:0043195~terminal button	0.0612
GO:0005625~soluble fraction	0.0661
GO:0044427~chromosomal part	0.0723
GO:0016604~nuclear body	0.0875
GO:0005819~spindle	0.0890
GO:0031981~nuclear lumen	0.0083

Table 5. Pathway annotation of perturbation causing genes using DAVID

GOTERM\_BP\_FAT

<b>GOTERM_BP_FAT</b>	<b>p-value</b>
GO:0006915~apoptosis	0.0081
GO:0050804~regulation of synaptic transmission	0.0092
GO:0012501~programmed cell death	0.0094
GO:0008219~cell death	0.0126
GO:0016265~death	0.0136
GO:0051969~regulation of transmission of nerve impulse	0.0137
GO:0007610~behavior	0.0160
GO:0031644~regulation of neurological system process	0.0167
GO:0050851~antigen receptor-mediated signaling pathway	0.0188
GO:0042981~regulation of apoptosis	0.0213
GO:0043067~regulation of programmed cell death	0.0236
GO:0010941~regulation of cell death	0.0244
GO:0006401~RNA catabolic process	0.0256
GO:0009057~macromolecule catabolic process	0.0286
GO:0002429~immune response-activating cell surface receptor signaling pathway	0.0292
GO:0002074~extraocular skeletal muscle development	0.0342
GO:0002768~immune response-regulating cell surface receptor signaling pathway	0.0354
GO:0043068~positive regulation of programmed cell death	0.0372
GO:0010942~positive regulation of cell death	0.0382
GO:0030111~regulation of Wnt receptor signaling pathway	0.0445

Table 6. Pathway annotation of perturbation causing genes using DAVID

GOTERM\_CC\_FAT

<b>GOTERM_CC_FAT</b>	<b>p-value</b>
GO:0034399~nuclear periphery	0.0208
GO:0045211~postsynaptic membrane	0.0292
GO:0005881~cytoplasmic microtubule	0.0376
GO:0043235~receptor complex	0.0500
GO:0005637~nuclear inner membrane	0.0589
GO:0044456~synapse part	0.0623

### 3.3. Clinical implication of perturbation network against PPI network

#### 3.3.1. Lethal genes versus disease genes

Lethal genes are known to have a strong tendency to be located at the functional centre of the interactome, while there has been much debate about the centrality of disease genes. We used perturbation sensitivity, which is reciprocal to the PPI network, to characterize disease genes. First, we compared disease genes versus lethal genes (Materials and Methods). Although disease is not specific term but general term for disturbance of organism homeostasis, we assume that union of GWAS and OMIM genes as disease genes since they have large enough group of disease genes.

We used multiple logistic regression to distinguish lethal genes and disease genes using  $K_{in}$ ,  $K_{out}$ , and  $K_{ppi}$  as predictors. As a result, disease genes have high  $K_{in}$  (p-value = 8.63e-25, figure 11) but not correlated to  $K_{ppi}$ . On the contrary, lethal genes have high  $K_{ppi}$  (p-value = 1.76e-69, figure 10) but not correlated to  $K_{in}$ . We binned whole genes using both  $K_{in}$  and  $K_{ppi}$  to draw 2-dimensional heatmap to characterize lethal genes and disease genes. Consistent with many previous reports (29, 33, 44) and result of our multiple logistic regression, lethal genes have enriched upper left region of heatmap which have high  $K_{ppi}$  and low  $K_{in}$ . On the other hand, disease genes have enriched right or right upper region of heatmap which have high  $K_{in}$ . Figure 8 shows summary. As a result, we can infer that disease genes are closely associated with perturbation sensitivity and weakly or rarely associated with degree of protein interaction network. We also draw grid diagram for schizophrenia genes versus lethal genes (Figure 9).

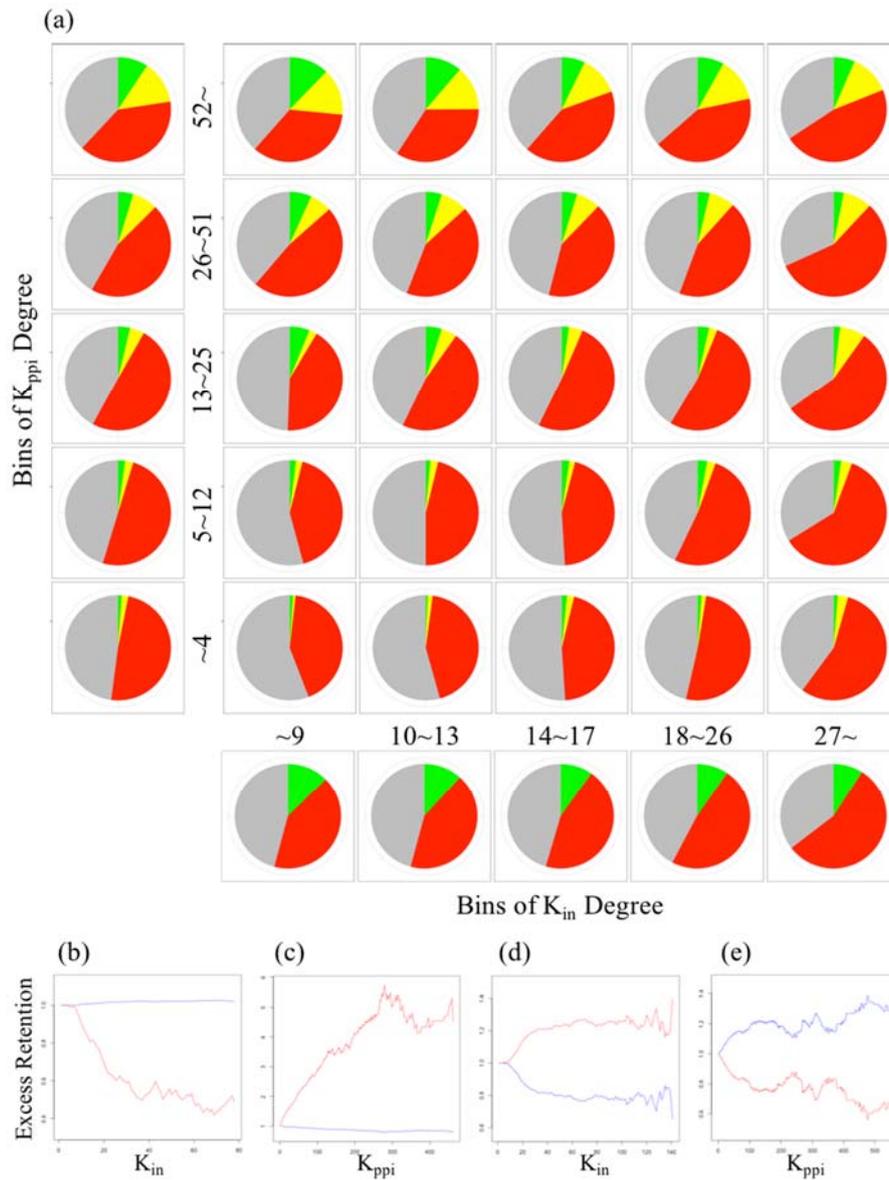


Figure 8. Joint grid diagram for lethal genes and disease genes. (A) A grid diagram that consists of pie charts that demonstrate the proportion of four groups of genes (green : lethal non-disease genes, red : disease non-lethal genes, yellow : disease-lethal genes, and grey : non-lethal non-disease genes) at each degree bin. (B~E) Excess retention plots for (B) Lethal genes in perturbation network core, (C) Lethal genes in PPI network core, (D) Disease genes in perturbation network core, (E) Disease genes in PPI network core.

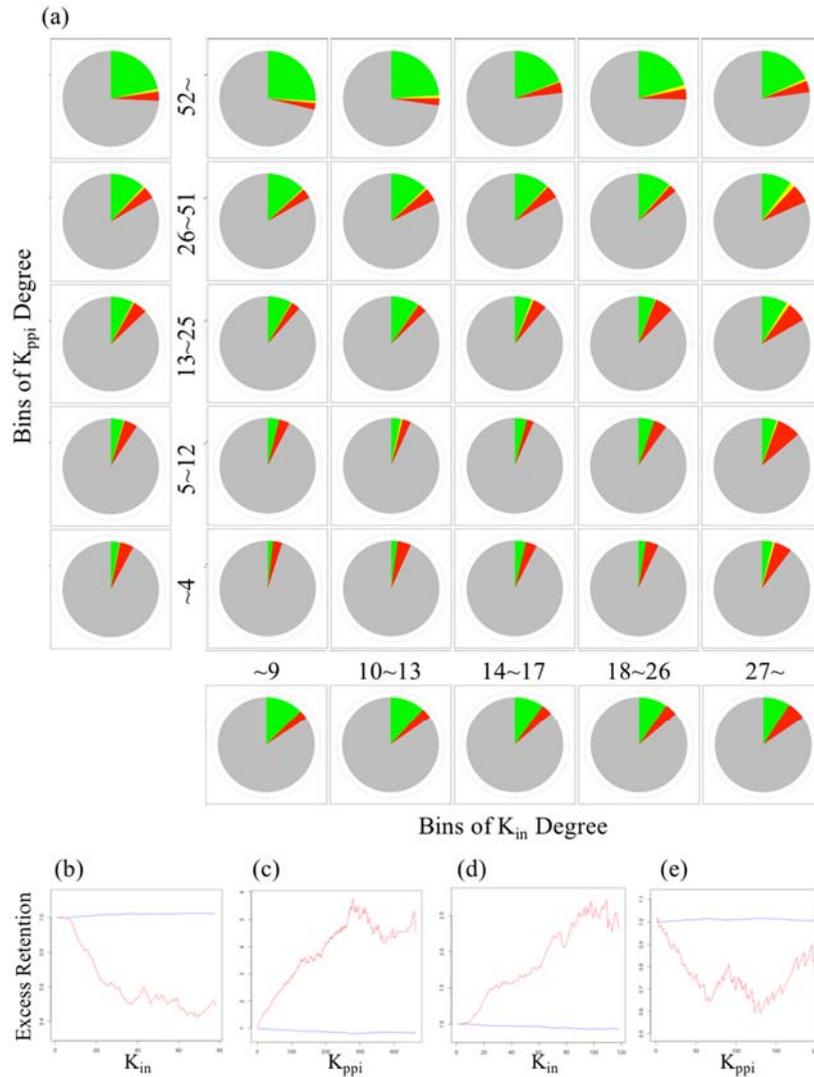


Figure 9. Joint grid diagram for lethal genes and schizophrenia genes. (A) A grid diagram that consists of pie charts that demonstrate the proportion of four groups of genes (green : lethal non-schizophrenia genes, red : schizophrenia non-lethal genes, yellow : schizophrenia-lethal genes, and grey : non-lethal non-schizophrenia genes) at each degree bin. (B~E) Excess retention plots for (B) lethal genes in perturbation network core, (C) lethal genes in PPI network core, (D) schizophrenia genes in perturbation network core, (E) schizophrenia genes in PPI network core.

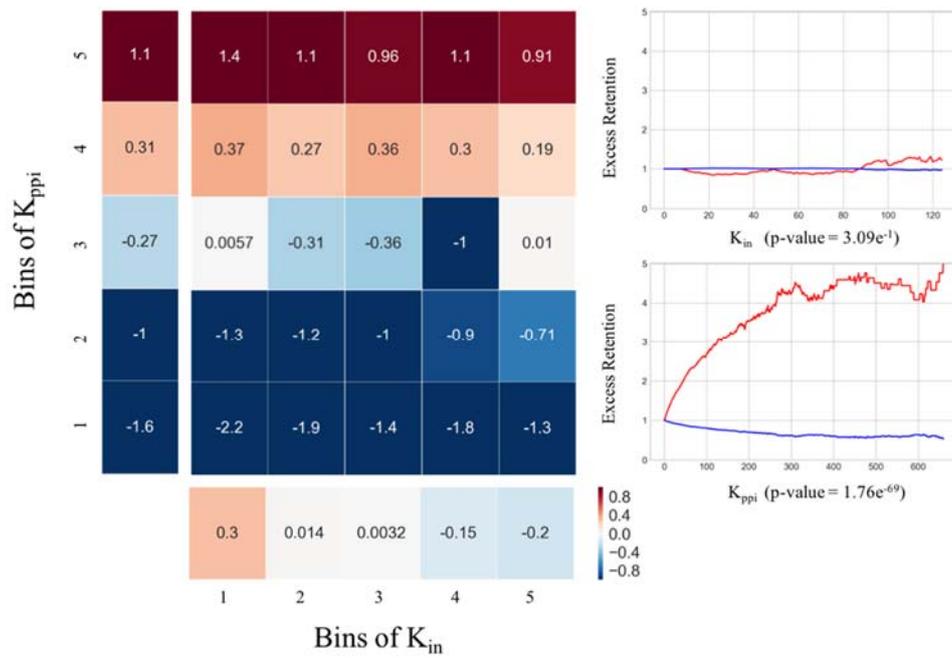


Figure 10. Joint heatmap for lethal genes. Number of each cell denotes odd ratio of lethal genes. Red lines of right excess retention plot indicate lethal genes.

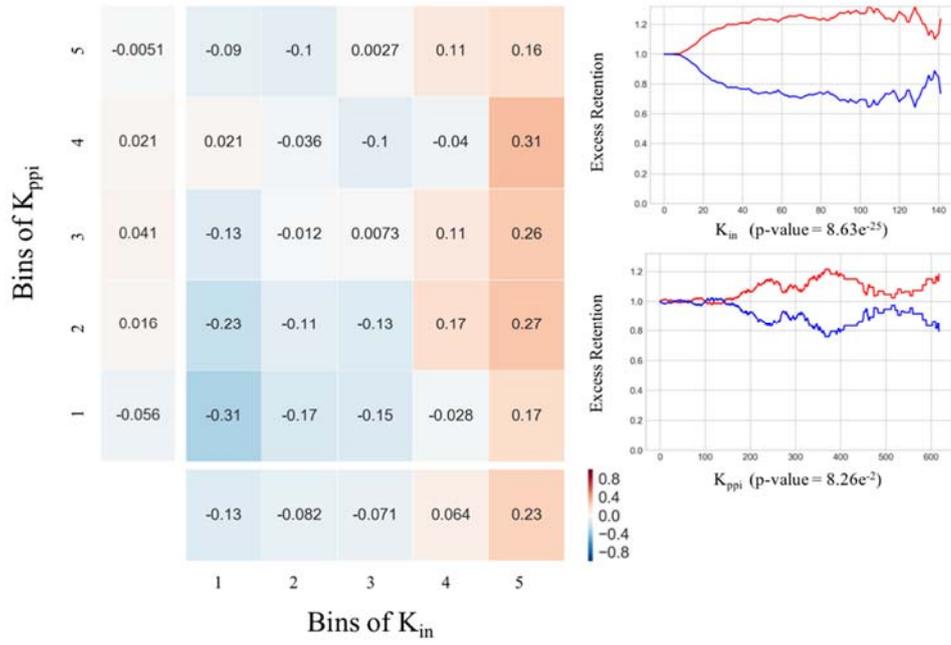


Figure 11. Joint heatmap for disease genes. Number of each cell denotes odd ratio of lethal genes. Red lines of right excess retention plot indicate disease genes.

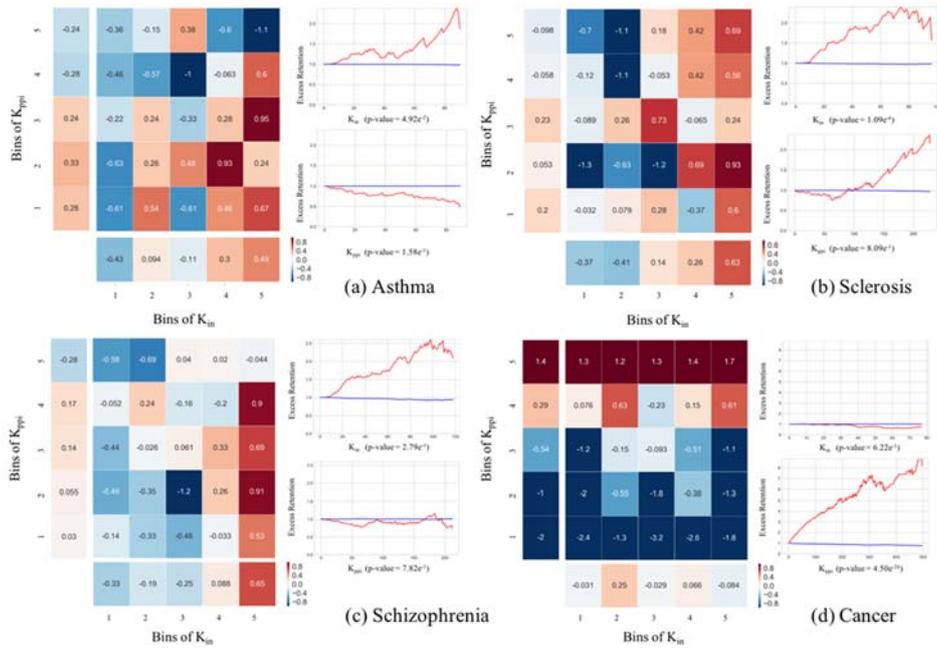


Figure 12. Joint heatmap for visualizing 4 disease gene group: asthma, sclerosis, schizophrenia, and cancer. Number of each cell denotes odd ratio of corresponding disease genes. Red lines of right excess retention plot also indicate corresponding disease genes.

### 3.3.2. Disease gene classification using both $K_{ppi}$ and $K_{in}$

We classify disease genes into small groups to categorize diseases with  $K_{in}$  and  $K_{ppi}$  using heatmap and excess retention plot (Methods and Materials.). Figure 12 and 13 shows heatmaps plotted by various disease gene groups. GWAS and OMIM consists of heterogeneous studies with various diseases. We consider genes as certain disease gene only if name of study including disease name. For example, we build asthma gene sets with genes which DISEASE/TRAIT column of GWAS catalogue includes 'asthma' or 'Asthma'. We plot for well-studied disease genes : asthma, systemic sclerosis, schizophrenia, and cancer (Figure 12). Asthma, systemic sclerosis, schizophrenia genes show low  $K_{ppi}$  and high  $K_{in}$ , while cancer genes from COSMIC tend to have high  $K_{ppi}$  and low  $K_{in}$ .

We also use GAD (63), The Genetic Association Database, to categorize disease genes. GAD categorize disease genes by organ systems, as cardiovascular, neurological, metabolic, and etc. Figure 13. (a)~(p) shows heatmap and excess retention plot for disease genes categorized by organ systems. All categories of disease gene enriched in high  $K_{in}$  except cancer and renal genes, while  $K_{ppi}$  is varied across categories. Figure 15 shows disease classification using clinical gene sets we used. Cancer genes, lethal genes, GAD neurological genes, and genes from OMIM clearly form clusters which implies similarities between those phenotypes.

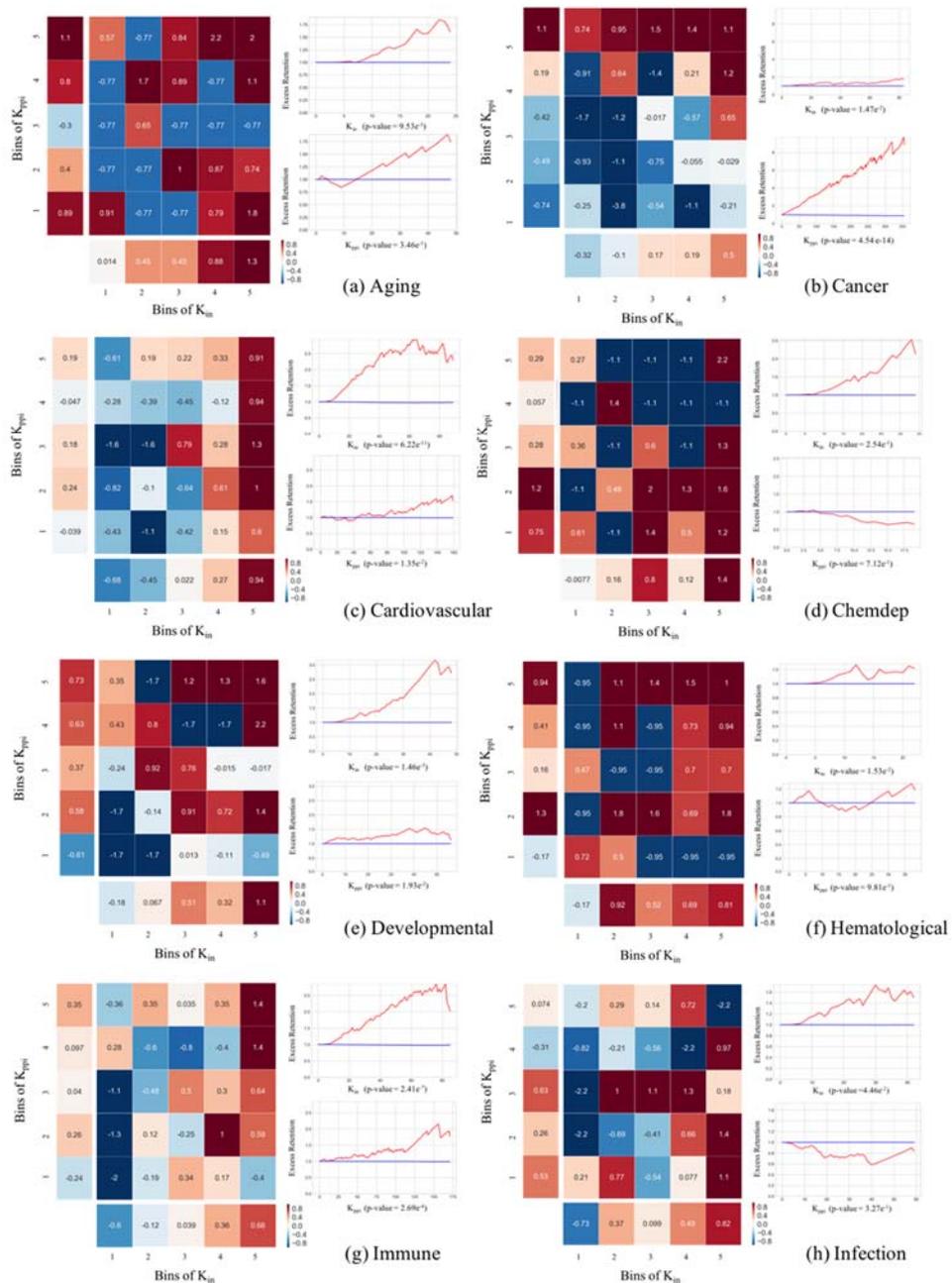


Figure 13. Joint heatmap for visualizing disease gene categories classified in GAD. Number of each cell denotes odd ratio of corresponding disease genes. Red lines of right excess retention plot also indicate corresponding disease genes.

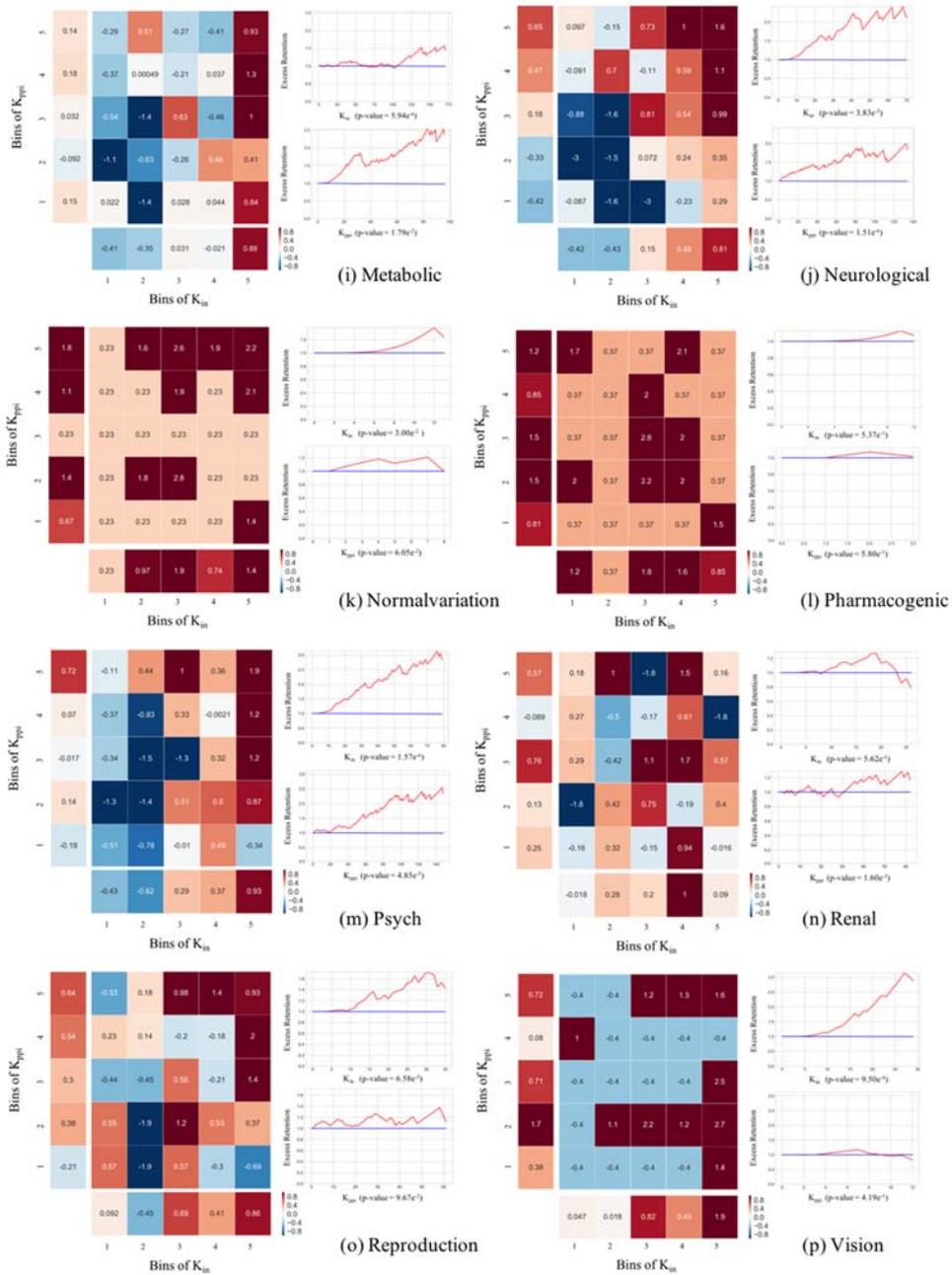


Figure 14. Joint heatmap for visualizing disease gene categories classified in GAD(continued). Number of each cell denotes odd ratio of corresponding disease genes. Red lines of right excess retention plot also indicate corresponding disease genes.

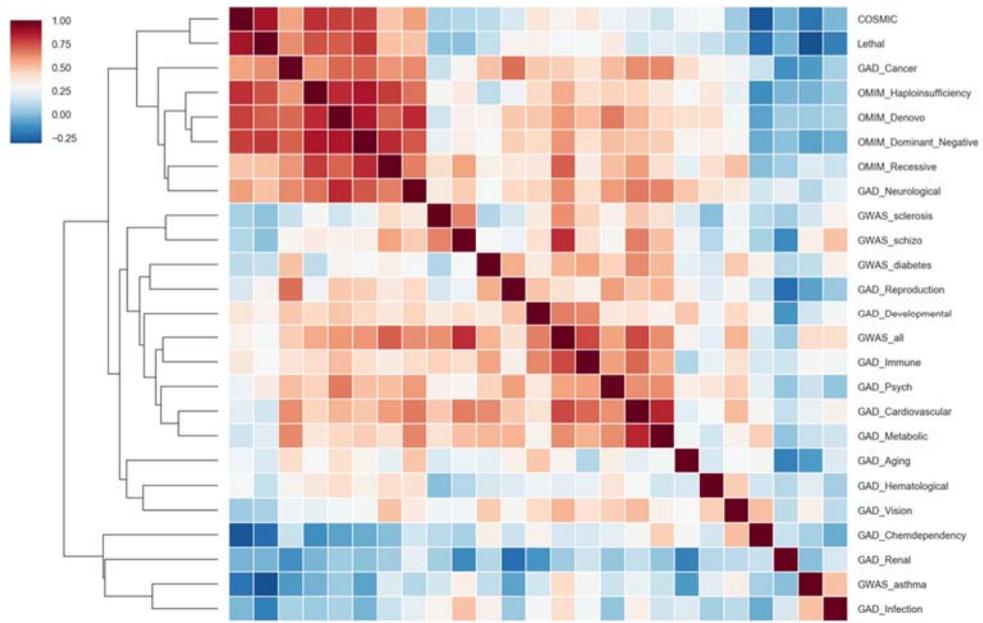


Figure 15. Heatmap for disease classification. Each cell denotes mantel statistics of pair of two disease gene sets.

## 4. Discussion

Perturbation network is built from 1000 genomes population whose participants are declared themselves to be healthy and does not have any clinical information. There are numerous non-synonymous variants in normal population and little studies reported their effect. We integrate non-synonymous variants within genic region to evaluate damaging of a gene and apply it to elucidate perturbation sensitivity of genes. As in yeast perturbation network,  $K_{in}$  is reciprocally correlated to  $K_{ppi}$ . Also, contrary to popular belief, perturbation sensitive genes are evolutionary conserved than perturbation tolerant genes. It is consistent with result from yeast data and implies that hub of perturbation network plays an important role in survival of organism as hub of PPI network. It also consistent with yeast perturbation network that  $K_{in}$  is related to disease genes, not lethal genes.

Different from yeast perturbation network,  $K_{in}$  does not follow scale-free distribution, but follow log-normal distribution irrelevant to the method for building perturbation network. It is also different from yeast perturbation network that in yeast, perturbation sensitive genes act as stress-responsive genes but in human, perturbation sensitive genes function as signalling molecule and located in cellular membrane or extracellular matrix. It may be caused by difference between single celled organism and multicellular organism. Extracellular environment of multicellular organism is strictly controlled and structured as cell to cell signalling system.

Previous report about topological characteristics of disease genes in PPI network were inconsistent, while lethal genes consistently reported as hub of PPI network. As we shown, most disease genes are correlated to perturbation network, not PPI network. Disease is not fatality of single cell but change of cell-cell interaction and function. Since perturbation network reflects cell's response to external signal or stress, genes related to chronic systemic diseases located in hub of perturbation

network, not PPI network. In comparison, for example, cancer genes located in hub of PPI network since cancer is typical single cell disease. In our analysis, most disease genes tend to have high  $K_{in}$  and various  $K_{ppi}$  while Cancer related genes are only categories of disease genes which have low  $K_{in}$  and extremely high  $K_{ppi}$ . Although cancer affect surrounding tissues, cancer itself is originated from single cell and evolved solely. On the other hands, most diseases do not be originated from single cell and does not dramatically change cell's function. For example, asthma is hypersensitive inflammatory disease and most related cells function as usual without stimulus. Though, we can infer a gene's clinical importance using both PPI network and perturbation network. Genes located hub of PPI network may cause death of single cell while genes located hub of perturbation network may cause malfunction of single cell and change of response to external signal and stress.

We considered the genes that were most 'perturbed' to be hub of perturbation network, while many others (71) considered the most perturbing genes to be hubs. 'Perturbing gene' is about how many genes are perturbed by certain gene and 'Perturbed gene' is about how many genes perturb that gene. In our analysis, out-degree does not have any significant biological implications. The most severely perturbing genes must be the essential genes and disturbance of essential genes may cause complete cell death, hence impacting all genes. Therefore, bias to non-lethal genes is inevitable. In contrast, the definition of the most perturbed genes is unbiased because the effect of lethal mutation can equally be applied to all genes. Also, genetic perturbation is mitigated by several biological mechanisms, it is impossible to connect genotype to phenotype directly. Rather, perturbed genes are closer to final consequence of genetic perturbation and may have biological and clinical implications.

In this study, we figure out how transcriptome is regulated and react to perturbation using normal populations' transcriptome perturbation sensitivity. Perturbation sensitivity genes are important for not only survival of single cell but also harmonious reaction to cell-cell interaction and external stress. Evolutional

evaluation of these genes confirm that these genes are evolutionarily conserved and important for survival of organism, not single cell. These genes also related to disease genes, especially chronic systemic disease. Perturbation sensitivity will supplement static character of protein interaction network and ease understanding of genome in network. Altogether, perturbation sensitivity is valuable measure for assessing gene's biological and clinical properties.

## 5. References

1. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65.
2. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. *Nature*. 2015;517(7534):327-32.
3. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. *Cell*. 2013;155(1):27-38.
4. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, Walter K, et al. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*. 2012;335(6070):823-8.
5. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. *Science*. 2002;296(5568):752-5.
6. Cowen L, Ideker T, Raphael BJ, Sharan R. Network propagation: a universal amplifier of genetic associations. *Nature reviews Genetics*. 2017;18(9):551-62.
7. Rockman MV, Kruglyak L. Genetics of global gene expression. *Nature reviews Genetics*. 2006;7(11):862-72.
8. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*. 2003;31(13):3812-4.
9. Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic acids research*. 2002;30(17):3894-900.
10. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, Bork P. Prediction of deleterious human alleles. *Human molecular genetics*. 2001;10(6):591-7.

11. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*. 2005;15(8):1034-50.
12. Flanagan SE, Patch AM, Ellard S. Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. *Genet Test Mol Biomarkers*. 2010;14(4):533-7.
13. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. *Nature methods*. 2014;11(3):294-6.
14. Yang H, Wang K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nature protocols*. 2015;10(10):1556-66.
15. Sherry ST, Ward M, Sirotkin K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome research*. 1999;9(8):677-9.
16. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*. 2001;29(1):308-11.
17. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature reviews Genetics*. 2008;9(5):356-69.
18. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research*. 2014;42(Database issue):D1001-6.
19. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. *Plant methods*. 2013;9:29.
20. Lee Kye Hwa JHK. Genome Sequence Variability Predicts Drug Precautions and Withdrawals from the Market. *PloS one*. 2017.
21. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*. 2013;501(7468):506-11.

22. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nature reviews Genetics*. 2009;10(3):184-94.
23. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nature genetics*. 2013;45(10):1238-43.
24. Fehrmann RS, Jansen RC, Veldink JH, Westra HJ, Arends D, Bonder MJ, et al. Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS genetics*. 2011;7(8):e1002197.
25. Gilad Y, Rifkin SA, Pritchard JK. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in genetics : TIG*. 2008;24(8):408-15.
26. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nature reviews Genetics*. 2004;5(2):101-13.
27. Cho DY, Kim YA, Przytycka TM. Chapter 5: Network biology approach to complex diseases. *PLoS computational biology*. 2012;8(12):e1002820.
28. Ideker T, Sharan R. Protein networks in disease. *Genome research*. 2008;18(4):644-52.
29. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature*. 2001;411(6833):41-2.
30. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*. 2007;104(21):8685-90.
31. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, et al. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature genetics*. 2006;38(3):285-93.
32. Barabasi AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nature reviews Genetics*. 2011;12(1):56-68.

33. Sun J, Jia P, Fanous AH, van den Oord E, Chen X, Riley BP, et al. Schizophrenia gene networks and pathways and their applications for novel candidate gene selection. *PloS one*. 2010;5(6):e11351.
34. Michaelson JJ, Loguercio S, Beyer A. Detection and interpretation of expression quantitative trait loci (eQTL). *Methods*. 2009;48(3):265-76.
35. Keurentjes JJ, Fu J, Terpstra IR, Garcia JM, van den Ackerveken G, Snoek LB, et al. Regulatory network construction in Arabidopsis by using genome-wide gene expression quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America*. 2007;104(5):1708-13.
36. Fagny M, Paulson JN, Kuijjer ML, Sonawane AR, Chen CY, Lopes-Ramos CM, et al. Exploring regulation in tissues with eQTL networks. *Proceedings of the National Academy of Sciences of the United States of America*. 2017;114(37):E7841-E50.
37. van Nas A, Ingram-Drake L, Sinsheimer JS, Wang SS, Schadt EE, Drake T, et al. Expression quantitative trait loci: replication, tissue- and sex-specificity in mice. *Genetics*. 2010;185(3):1059-68.
38. Munsky B, Neuert G, van Oudenaarden A. Using gene expression noise to understand gene regulation. *Science*. 2012;336(6078):183-7.
39. de Nadal E, Ammerer G, Posas F. Controlling gene expression in response to stress. *Nature reviews Genetics*. 2011;12(12):833-45.
40. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS genetics*. 2011;7(2):e1002003.
41. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan KK, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. *Nature*. 2012;489(7414):91-100.
42. Grundberg E, Small KS, Hedman AK, Nica AC, Buil A, Keildson S, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature genetics*. 2012;44(10):1084-9.

43. Ohn JH, Kim J, Kim JH. Genomic characterization of perturbation sensitivity. *Bioinformatics*. 2007;23(13):i354-8.
44. Han HW, Ohn JH, Moon J, Kim JH. Yin and Yang of disease genes and death genes between reciprocally scale-free biological networks. *Nucleic acids research*. 2013;41(20):9209-17.
45. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research*. 2014;42(Database issue):D980-5.
46. Dewey FE, Grove ME, Pan C, Goldstein BA, Bernstein JA, Chaib H, et al. Clinical interpretation and implications of whole-genome sequencing. *JAMA*. 2014;311(10):1035-45.
47. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, Lappalainen T, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science*. 2013;342(6154):1235587.
48. Habegger L, Balasubramanian S, Chen DZ, Khurana E, Sboner A, Harmanci A, et al. VAT: a computational framework to functionally annotate variants in personal genomes within a cloud-computing environment. *Bioinformatics*. 2012;28(17):2267-9.
49. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research*. 2012;22(9):1760-74.
50. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research*. 2010;38(16):e164.
51. Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome research*. 2009;19(9):1553-61.
52. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods*. 2010;7(8):575-6.

53. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*. 2014;46(3):310-5.
54. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic acids research*. 2006;34(Database issue):D535-9.
55. Hurst LD. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends in genetics : TIG*. 2002;18(9):486.
56. Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, et al. Functional genomic analysis of the rates of protein evolution. *Proceedings of the National Academy of Sciences of the United States of America*. 2005;102(15):5483-8.
57. Kasprzyk A. BioMart: driving a paradigm change in biological data management. *Database : the journal of biological databases and curation*. 2011;2011:bar049.
58. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*. 2009;4(1):44-57.
59. Wuchty S, Almaas E. Peeling the yeast protein network. *Proteomics*. 2005;5(2):444-9.
60. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic acids research*. 2015;43(Database issue):D789-98.
61. Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE, Mouse Genome Database G. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic acids research*. 2015;43(Database issue):D726-36.

62. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids research*. 2015;43(Database issue):D805-11.
63. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. *Nature genetics*. 2004;36(5):431-2.
64. Kimura M. Evolutionary rate at the molecular level. *Nature*. 1968;217(5129):624-6.
65. Yang L, Wang S, Zhou M, Chen X, Zuo Y, Sun D, et al. Comparative analysis of housekeeping and tissue-selective genes in human based on network topologies and biological properties. *Molecular genetics and genomics : MGG*. 2016;291(3):1227-41.
66. Yosef N, Sharan R, Noble WS. Improved network-based identification of protein orthologs. *Bioinformatics*. 2008;24(16):i200-6.
67. Chen WH, Zhao XM, van Noort V, Bork P. Human monogenic disease genes have frequently functionally redundant paralogs. *PLoS computational biology*. 2013;9(5):e1003073.
68. Pritchard CC, Cheng HH, Tewari M. MicroRNA profiling: approaches and considerations. *Nature reviews Genetics*. 2012;13(5):358-69.
69. Ryan BM, Robles AI, Harris CC. Genetic variation in microRNA networks: the implications for cancer research. *Nature reviews Cancer*. 2010;10(6):389-402.
70. Inui M, Martello G, Piccolo S. MicroRNA control of signal transduction. *Nat Rev Mol Cell Biol*. 2010;11(4):252-63.
71. Rung J, Schlitt T, Brazma A, Freivalds K, Vilo J. Building and analysing genome-wide gene disruption networks. *Bioinformatics*. 2002;18 Suppl 2:S202-10.

## 국문 초록

**연구 목적:** 유전자의 발현은 수많은 유전체 돌연변이에 의해서 교란되며, 이는 세포의 기능과 개체의 표현형에 큰 영향을 준다. 최근의 대규모 차세대 염기서열분석 프로젝트에서 밝혀지고 있듯, 한사람의 유전체는 적어도 300 만개의 돌연변이를 가지고 있는 것으로 알려져 있다. 본 논문에서는 이러한 유전체 교란을 해석하고 교란에 민감한 유전자의 특징을 살펴보고자 전사체 교란 네트워크를 1000 유전체 프로젝트 데이터를 통해 구성해보았다.

**연구 방법:** 본 연구에서는 단백질 코딩 영역 내 비 동일 변이의 시프트 점수를 종합하여 유전자 손상 정도를 평가하였다. 이를 기반으로 전사체 교란 네트워크를 구성하고, 유전자의 내향 연결 정도를 교란 민감도로 정의하였다. 유전자를 교란 민감도에 따라 분류하고 교란 민감 유전자와 교란 둔감 유전자의 진화적, 생물학적, 그리고 임상적 특징을 조사하였다.

**결과:** 교란 민감 유전자는 단백질 상호작용 네트워크의 변방에 위치해 있었으나 진화적으로 보존되어 있었다. 이들은 상대적으로 적은 수의 미소 전사체와 전사인자에 의해 조절되고 있으며, 세포 간의 상호작용에 중요한 역할을 하고 있었다. 전사체 교란 네트워크의 외향 연결 정도는 중요한 생물학적 의미를 가지고 있지 않았다. 치사 유전자의 경우 교란 네트워크의 말단이면서 단백질 상호작용 네트워크의 중심부에 위치해 있었다. 반면, 대부분의 질병 유전자들의 경우 교란 네트워크의 중심이면서 단백질 상호작용 네트워크의 말단에 위치해 있었다. 두 네트워크를 모두 사용하여, 질병을 분류하기 위한 연합 네트워크 도표를 그려보았다.

**결론:** 효모에서의 연구와 마찬가지로, 교란 민감 유전자는 유전적으로 보존되어 있고 세포 간의 상호작용에 관여하여 개체의 생존에 필수적이었다.

또한, 내향 연결정도가 외향 연결정도에 비해 유전자 교란을 해석하는데 유용하다는 것을 확인하였다. 질병 유전자는 단백질 상호작용 네트워크와 교란 네트워크를 동시에 활용하여 시각화 되고 분류될 수 있었다. 결론적으로, 교란 민감도는 유전자의 생물학적 임상적 특성을 분석하고 유전체 교란을 평가하는데 가치 있는 지표가 될 것이다.

**주요어:** 유전체 교란, 전사체, 단백질 상호작용 네트워크, 질병 유전자

**학번:** 2010-21974