NETWORKING PATHWAYS UNVEILS ASSOCIATION BETWEEN OBESITY AND NON-INSULIN DEPENDENT DIABETES MELLITUS[*]

HAIYAN HU[†]

School of Informatics and Center for Computational Biology and Bioinformatics, Indiana University, 410 West 10th street, Suite 5000, Indianapolis, IN 46202, USA

XIAOMAN LI[†]

Division of Biostatistics, Indiana University, 410 West 10th street, Suite 5000, Indianapolis, IN 46202, USA

Genetic related health problems are often interrelated. Current practices to establish associations between diseases are expensive and rarely can reflect underlying molecular mechanisms. We propose a general framework to associate diseases by networking pathways. By applying our method on association study of non-insulin dependent diabetes mellitus (NIDDM) and obesity, we demonstrate that our method can both identify signature pathways for each disease and establish valid association of two diseases.

1. Introduction

Many diseases are interrelated. Obesity, diabetes, insulin resistance, hypertension are just a few examples. Instead of being attributed to a specific gene, these diseases are often caused by interaction among multiple genes or between genes and environment, and thus are often classified as multifactorial disease or complex disease. Great effort has been put on association studies, such as case control studies and cohort studies, to discover the potential relation between multiple disease conditions in human. Although such association studies can often produce very important information, they are either not very reliable or not efficient in terms of time and money. For example, of two large American Cancer Society cohorts, Cancer Prevention Study I (CPS-I; enrolled in 1959 and followed through 1972) and Cancer Prevention Study II (CPS-II; enrolled in 1982 and followed through 1996), one shows association of height with prostate cancer, the other does not [1]. Most importantly, from such association studies on complex diseases involving genetic factors, no matter how significant the identified associations are statistically, researchers usually cannot gain much insight of the underlying molecular mechanisms. Thus,

---

[†] Corresponding authors.

efficient methods are urgently needed to identify disease associations at the molecular level.

Microarray experiments have been a very popular tool for disease study. From microarray data, gene expression signatures that can distinguish a disease phenotype from another have often been identified by implementing analytical techniques such as differential test. However, in complex diseases like cancer, it is not the individual genes but the interaction between many genes and the interaction between many genes and environment that are responsible for a certain physiological process. Therefore, dozens of suspicious genes included in an identified signature are insufficient for understanding the underlying mechanisms behind a specific disease phenotype. In order to gain deeper understanding of complex diseases from a set of differentially expressed genes, one common practice is to convert the information from gene space to structured pathway space via enrichment test of the differential expressed genes in predefined pathways [2,3]. However, unlike cancers, in which gene expression often show larger variation, for complex diseases like diabetes, obesity, and atherosclerosis, the changes in gene expression are more likely to be modest [4-6]. Yet, the genes vary subtle might be the very responsible ones for a disease phenotype [7,8]. Therefore, under such circumstances that no genes are selectable from differential tests, the traditional methods depending on identification of disease susceptibility genes have lost their power. On the other hand, analysis directly performed on pathways has been encouraging in providing deeper biological understanding compared with single-gene based methods [9-13].

Observing the success of pathway based analysis in various disease studies, we hypothesize that pathway-originated methods are also of great value in associating different disease phenotypes. In this paper, we propose a general framework to study disease association via networking pathways. As a proof of principle, we apply our methods to identify the association between obesity and Non-insulin Dependent Diabetes Mellitus (NIDDM, Type II diabetes), the two diseases that affects hundreds of millions of people worldwide with widely observed connection but with unknown association mechanisms at the molecular level. We have identified a number of pathways and gene sets with known and unknown functions that are responsible for each disease. More importantly, by networking pathways, we have also discovered a set of pathways and their interactions that are responsible for the association between obesity and NIDDM.
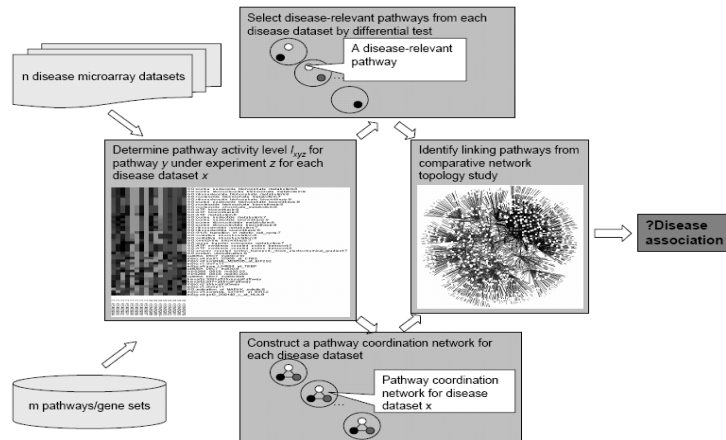
Figure 1. The pipeline of multiple disease association by networking pathways.

2. Methods

We propose a general framework to identify multiple disease association by pathway/gene set association (Figure 1). Here, a gene set is a priori defined set of genes such as a set of genes in one pathway, or a set of target genes regulated by the same transcription factor. Schematically, given *n* disease datasets and *m* predefined pathways/gene sets, we first determine its activity level under each experimental condition for each pathway/gene set. We then select differentially activated pathways/gene sets between disease and control experimental conditions in one data set, and we also construct a pathway coordination network for each disease dataset, in which each node represents a pathway/gene set and each edge connects two pathways/gene sets showing significant coordinated activities. A pathway coordination network thus converts its corresponding disease data into a relation graph depicting the interplay among various functional units (predefined pathways/gene sets in our case). By performing comparative network analysis, we finally can generate hypothesis on disease association at the molecular pathway level. The methods and techniques utilized are detailed in the following subsections.

2.1. *Microarray data sources*

We use microarray experiment data obtained from skeletal muscle. Skeletal muscle cells are the largest storage organ for glucose and considered to play the major role in glucose homoeostasis. From DGAP (Diabetes Genome Anatomy

Project), we downloaded type II diabetic human data containing 18 NIDDM samples and 17 Normal Glucose Tolerance samples generated from Human skeletal muscle samples of Swedish males for study of type II diabetes by Dr. Altshuler's Lab at MIT. From GEO (Gene Expression Omnibus) at NCBI, we also downloaded obesity skeletal muscle data (GDS268) containing 8 skeletal muscle samples from non-obese subjects and 8 skeletal muscle samples from morbidly obese subjects. All the downloaded gene expression levels were measured using Affymetrix Human U133A GeneChip platform. The same experimental platform and tissue type enables us to study obesity and NIDDM with higher signal to noise ratio.

2.2. *Compilation of pathways/gene sets for human*

We downloaded 187 pathways from KEGG [14], 263 pathways from BioCarta [15], 20 pathways related to cancer/immune signaling from NetPath website [16], 243 pathways from Rat Genome Database [17], and 1520 gene sets from mSigDB (version on Oct, 2006) [10]. Besides, we obtained another 3229 gene sets by grouping genes on AFFY-HU133A array according to their GO annotation using FatiGO [2]. Additionally, we obtained 2459 gene sets from graph clustering using MCL [18] on gene expression profiles in four microarray datasets related to NIDDM and obesity [8,19-21]. In total, 7921 pathways/gene sets were compiled for this study.

2.3. *Pathway/gene set activity level and coordination network*

We define the activity level profile of a pathway/gene set under a given set of experimental conditions using eigengenes generated from singular value decomposition (SVD) [9,22]. In detail, for each pathway/gene set containing m genes, there is one $m \times n$ matrix A consisting of the row normalized transcriptional responses of these m genes under n experiments in a microarray dataset such that the mean and standard deviation of the expression levels for each gene is 0 and 1, respectively. We then performed SVD on the matrix A to decompose A into three matrices U, S and V, i.e. $A = USV^T$. The U and V^T are commonly named as the eigenarray matrix and eigengene matrix respectively. The matrix S is a diagonal matrix with singular values of the matrix A as the diagonal elements, whose square reflect the variance of the corresponding eigengene/eigenarray. By using the top k eigengenes, with each of which accounting for no less than *(70/n)%* of the overall variability [23,24], we define activity level l of a pathway/gene set under experiment j as:

$$l_j = \sum_{i=0}^{k} V^T{}_{ij}{}^2 . \tag{1}$$

Here i is the index of top k eigengenes we used. The intuition that a pathway's activities can be defined from eigengenes is that, linear combination of such defined pathway activity level is an optimal approximation of the transcription profile matrix A corresponding to all the genes within the pathway, as has been explained in [9]. However, unlike the pathway level analysis in [9], where the pathway activity level profile is determined from the first one eigengene (corresponding to the largest eigenvalue) only, here we utilize multiple eigengenes that can explain at least a certain percentage of variance to determine pathway activities. The advantage is evident in that the first eigengene is not always reflecting the dominant variance of the transcript levels corresponding to the genes within a pathway. Thus, given an expression value matrix, the activity level of a pathway captures the major components of the variation in the given expression matrix. After filtering out the genes not included on the Human U133A chip, 7016 out of the compiled 7921 pathways/gene sets containing at least two genes for performing SVD remained. With a pathway/gene set's activity level defined above, we define two pathways/gene sets as coordinated if the two pathways/gene sets show coordinated activity levels under a given set of conditions. Thus, for each disease microarray dataset, we can construct a pathway coordination network in which each node represents a pathway/gene set, and each edge connects two coordinated pathways/gene sets. In this paper, we measure the coordination between any two pathways/gene sets using Spearman's rank correlation. For a given pathway/gene set q, only the top 1% of the pathways/gene sets with largest correlations (larger than 0.6) with q are kept as coordinated pathways/gene sets of q.

2.4. *Differentially Activated Pathways/gene sets*

With the activity levels defined, we used SAM [25] to determine whether a pathway/gene set is activated differently between disease and control samples. SAM has been validated in a number of studies and has been shown more accurate than other differential test methods such as simple t-test [26-29]. SAM uses modified t statistic to measure the activity difference of a pathway/gene set between two types of samples as a score d. For each pathway/gene set, SAM then performs permutation test to determine the statistical significance of the d score. In our study, we chose the significant pathways/gene sets by controlling the false discovery rate (q-value) at the 0.1 level.

2.5. *Disease-relevant pathways and linking pathways*

First, if a pathway P is differentially activated between disease and control samples, then P is called an A-relevant pathway. Given two types of disease A

and *B*, if a pathway *P* is *A*-relevant, we define *P* as a linking pathway between diseases *A* and *B* if *P* satisfies one of the following three criteria. (1) *P* is directly connected to at least one of the pathways relevant to disease *B*; (2) *P* shares at least one first layer neighbor with at least one of the *B*-relevant pathways; (3) there is a cluster containing *P* shares at least one common element with a cluster containing *B*-relevant pathways. Any two linking pathways from different disease networks with coordinated activity profiles with each other or the common third party pathways may indicate associations between their corresponding diseases.

3. Results

3.1. *Identified obesity-relevant biological pathways/gene sets*

In total, we systematically identified 92 obesity-relevant pathway/gene sets that are differentially activated in obesity and control experiments, including 18 well defined pathways from KEGG and BioCarta database (Table 1) [14,15]. Many studies have supported the relevance of these pathways to obesity [30,31].

Table 1 - 18 well-defined pathways out of 92 pathways/gene sets that are differentially activated significantly between obesity and control experiments.

Pathway Description	Score(d)	q-value(%)
KEGG: Nicotinate and nictoinamide metabolism	1.83	0
KEGG: Glycan structures biosynthesis	1.56	6.10
mSigDB: Genes related to the insulin receptor pathway	1.59	6.10
α6-β4 Integrin Signaling Pathway	1.55	6.10
RGD: Prostaglandin and Leukotriene metabolic pathway	1.56	6.10
KEGG: Fatty acid metabolism	1.36	7.73
KEGG: Tryptophan metabolism	1.45	7.73
KEGG: Glycerophospholipid metabolism	1.50	7.73
KEGG: Arachidonic acid metabolism	1.49	7.73
KEGG: One carbon pool by folate	1.38	7.73
KEGG: MAPK signaling pathway	1.36	7.73
KEGG: mTOR signaling pathway	1.45	7.73
KEGG: Regulation of actin cytoskeleton	1.45	7.73
mSigDB: AR mouse plus testo from netaffx	1.37	7.73
mSigDB: rasPathway from Biocarta	1.43	7.73
RGD: glycerolipid metabolic pathway	1.41	7.73
Biocarta: Role of EGF Receptor Transactivation by GPCRs in Cardiac Hypertrophy	1.34	9.95
KEGG: Arginine and proline metabolism	1.34	9.95

3.2. *Identified NIDDM-relevant biological pathways/gene sets*

We identified 78 pathways/gene sets to be NIDDM-relevant, covering defined pathways in KEGG and BioCarta, expert-curated gene sets, gene sets defined by GO categories and gene sets comprised of co-expressed genes. 16 out of the 78 pathways/gene sets are well defined pathways (Table 2). Most of these pathways are related to the three components of carbohydrate catabolism: glycolysis, TCA cycle and oxidative phosphorylation, implicating the link between NIDDM and mitochondrial dysfunction [32,33].

Table 2 - 16 well-known pathways out of 78 pathways/gene sets that are significantly differentially activated between NIDDM and control experiments

Pathway Description	Score(d)	q-value(%)
mSigDB: electron transport chain	1.39	0
KEGG: Oxidative phosphorylation	1.28	0
mSigDB: Oxidative Phosphorylation	1.08	0
RGD: Oxidative Phosphorylation	1.04	0
mSigDB: Mitochondrial genes	1.04	0
KEGG: Pyruvate metabolism	0.98	0
mSigDB: Pyruvate metabolism	0.97	0
mSigDB: Role of Mitochondria in Apoptotic Signaling	0.97	0
KEGG: Citrate Cycle (TCA cycle)	0.83	2.70
RGD: Pyruvate metabolic pathway	0.82	2.70
KEGG: Propanoate metabolism pathway	0.80	4.39
RGD: glyoxylate and dicarboxylate metabolic pathway	0.80	4.39
mSigDB: Oxidative phosphorylation pathway from KEGG	0.79	4.39
mSigDB: Genes 2fold upregulated by insulin	0.76	4.39
mSigDB: krebPathway	0.75	7.19
mSigDB: Reactive oxidative species related genes	0.73	8.30

Although between NIDDM and control subjects, we have witnessed statistically significant differences at the pathway level, we have found no much difference at the individual gene level. Taking citrate cycle pathway as an example, none of the genes in this pathway is significantly differentially expressed. The genes *ACO2*, *MDH1* and *FH* are only slightly down regulated in NIDDM and with insignificant fold changes ranging from 0.8 to 0.95; the genes *SDHA* and *OGDH* only show modest increase in NIDDM (*SDHA*: fold =1.21, p-value= 0.776691; *OGDH*: fold = 1.19, p-value = 0.463948).

3.3. *Identification of association between obesity and NIDDM by networking pathways*

By comparing the defined obesity-relevant pathways and NIDDM-relevant pathways, we found that obesity-relevant pathways contains a gene set related to

the insulin receptor, and coincidentally, there is a NIDDM-relevant gene set containing genes 2-fold up-regulated by insulin. Other than that, all relevant pathways in obesity and NIDDM are literally different. Besides, the genes shared by the two types of pathways are not significantly differentiated between disease and control samples and consequently provides no sufficient information to determine association between obesity and NIDDM. Thus, we proceed to associate obesity and NIDDM in the following steps. We first build a pathway coordination network for each disease. For obesity dataset, this resulted in a network containing 7016 pathway nodes and 237,226 pathway coordination edges, and for NIDDM, this generated a network with 7016 pathway nodes and 207,571 pathway coordination edges. From the two networks, we attempt to associate the two diseases by searching for linking pathways according to their three criteria defined in our methods section.

To search for linking pathways satisfying the first criteria, we examined whether there are any direct links between the two types of disease-relevant pathways. We found obesity-relevant pathways including arginine and praline metabolism pathway and fatty acid metabolism pathway, and tryptophan metabolism pathway are directly connected with NIDDM-relevant pyruvate metabolism pathway. Besides, actin cytoskeleton regulation in obesity network is linked directly to TCA cycle in NIDDM network.
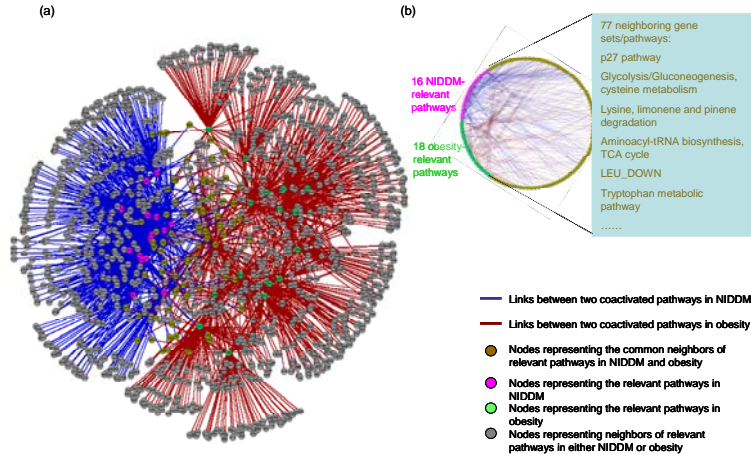

Figure 2 - Merged two pathway coordination sub-networks from obesity and NIDDM. (a) All first layer neighbors for the 18 obesity-relevant pathways and the 16 NIDDM-relevant pathways are included. (b) Only their common first layer neighbors of the 18 obesity-relevant and the 16 NIDDM-relevant pathways are included.

We next seek linking pathways satisfying the second criteria. By extracting the first-layer neighbor gene sets/pathways of the defined disease-relevant pathways from their corresponding networks (Figure 2a), we found totally 77

first-layer neighbor pathways/gene sets are in common. Figure 2b shows a sub-network from network in Figure 2a. This sub-network contains only disease-relevant pathway and their first layer neighbor pathway nodes, and 430 coordination edges linking them.
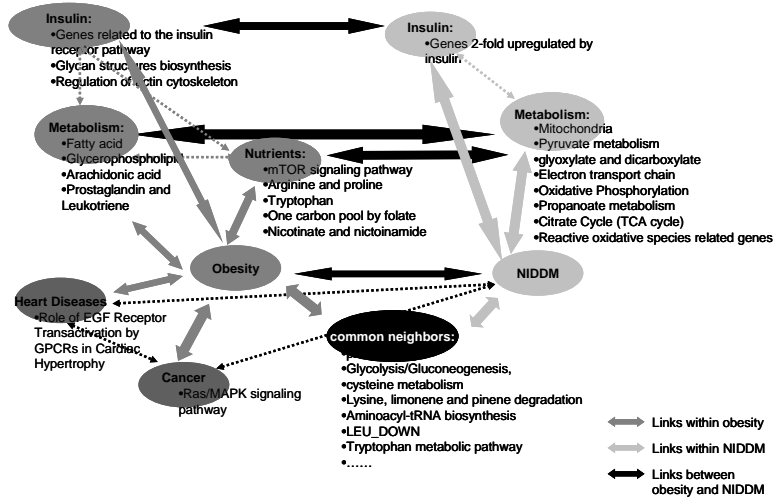

Figure 3 - Summary of identified associations between obesity and NIDDM.

Finally, we search for linking pathways satisfying the third criteria. By performing graph clustering on both networks using MCL algorithm [18], we obtained 239 clusters and 67 clusters containing more than 3 and less than 60 pathways/gene sets corresponding to obesity and NIDDM pathway coordination network respectively. We drew a comparison between the pathway clusters in obesity and NIDDM pathway coordination network. If a cluster in disease A's pathway coordination network contains an A-relevant pathway, it is defined as an A-relevant cluster. It is interesting to see there are a number of obesity-relevant clusters and NIDDM-relevant clusters we identified above overlapped. For instance, a cluster in the obesity pathway network including 5 gene sets related to glycogen/glucan biosynthesis is overlapped with another cluster in the NIDDM pathway network containing 16 gene sets involving regulation of circadian rhythm, keratin sulfate metabolism. An obesity-relevant cluster comprised of 53 gene sets involving PI3K (phosphoinositide 3-kinases) and their downstream targets, and SERMs (selective estrogen receptor modulators) down regulated genes is overlapped with another NIDDM-relevant cluster containing 23 gene sets including atrial natriuretic peptide signaling pathway, lipoprotein metabolic pathway, altered lipoprotein metabolic pathway.

Taking all the findings together, we provide a summary of all these pathway associations between obesity and NIDDM (Figure 3). Many of these associations are supported by literature search [33-39].

4. Discussion and Conclusion

We have proposed a general framework for disease association by pathway analysis and networking co-activated pathways. To our knowledge, this is the first disease association method that can delineate the relationship between any two or even more disease phenotypes at the molecular pathway/pathway interaction level. In contrast to disease association by case control or cohort studies, our method is not only efficient but also can generate deeper insight about disease etiology and pathophysiology, especially for complex diseases like NIDDM and obesity where the expression differences of genes are often trivial and consequently no suspicious genes detectable by conventional methods. Besides, our strategy moves beyond single gene/pathway based study, and sets off for studying the relationship between pathways or gene sets. In order to capture the relationship between any two pathways, we first generate an activity level profile reflecting the overall response of a certain pathway under a set of experimental conditions. A unique advantage of using pathway activity levels to characterize a pathway is that pathway activity levels can be further used to establish a quantitative relation between any two pathways. After determining the coordination relationship for each pair of pathways/gene sets, a disease dataset can then be modeled as a network. The problem of associating two diseases is subsequently converted to the problem of network comparison.

By applying our approach on obesity and NIDDM, We systematically obtained important pathways that can characterize each disease phenotype and further depicted the association between obesity and NIDDM via linking pathways. The coordinated activity of disease-relevant pathway fatty acid metabolism and pyruvate metabolism in both obesity and NIDDM samples indicates that dysfunction of fatty acid metabolism is intertwined with the functioning of pyruvate metabolism, perhaps via TCA cycle. The supporting model is the early proposed cellular mechanism of glucose-fatty acid cycle in which fatty acid oxidation inhibits glucose utilization by affecting pyruvate dehydrogenase activity [34]. Our study also discovered other important associations such as insulin relevant pathway, stress related ROS genes (Reactive Oxidative Species related genes), cell growth and apoptosis and other immune related pathways.

We need to point out that the accurate interpretation of the association between diseases heavily relies on the correct definition of pathways/gene sets.

More effort on curating pathways/gene sets in order for disease association by networking pathways is essential. Besides, our present study on obesity and NIDDM is based on microarray experiments on human skeletal muscle tissue. Therefore, the conclusions we drew in this study may not reflect the pathway interaction patterns in other tissues. With more and more microarray experimental datasets become available in the near future, it will be interesting to extend our study to multiple tissue/organs such as pancreatic islets, adipose tissue, liver and kidney. It will also be interesting to compare the pathway coordination networks across different species, the results of which will make the dynamic delineation of function evolution of related pathways and pathway interactions possible.

References

1.   Rodriguez C, Patel AV, Calle EE, Jacobs EJ, Chao A et al., *Cancer Epidemiol Biomarkers Prev* 10(4): 345-353.(2001)
2.   Al-Shahrour F, Diaz-Uriarte R, Dopazo J, *Bioinformatics (Oxford, England)* 20(4): 578-580.(2004)
3.   Segal E, Friedman N, Koller D, Regev A, *Nature genetics* 36(10): 1090-1098.(2004)
4.   Yechoor VK, Patti ME, Saccone R, Kahn CR, *Proceedings of the National Academy of Sciences of the United States of America* 99(16): 10587-10592.(2002)
5.   Wilson KH, Eckenrode SE, Li QZ, Ruan QG, Yang P et al., *Diabetes* 52(8): 2151-2159.(2003)
6.   Garland LG, *FEMS microbiology immunology* 5(5-6): 229-237.(1992)
7.   Patti ME, Butte AJ, Crunkhorn S, Cusi K, Berria R et al., *Proceedings of the National Academy of Sciences of the United States of America* 100(14): 8466-8471.(2003)
8.   Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S et al., *Nature genetics* 34(3): 267-273.(2003)
9.   Tomfohr J, Lu J, Kepler TB, *BMC bioinformatics* 6: 225.(2005)
10.   Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL et al., *Proceedings of the National Academy of Sciences of the United States of America* 102(43): 15545-15550.(2005)
11.   Pang H, Lin A, Holford M, Enerson BE, Lu B et al., *Bioinformatics (Oxford, England)* 22(16): 2028-2036.(2006)
12.   Huang E, Ishida S, Pittman J, Dressman H, Bild A et al., *Nature genetics* 34(2): 226-230.(2003)
13.   Shyamsundar R, Kim YH, Higgins JP, Montgomery K, Jorden M et al., *Genome biology* 6(3).(2005)
14.   Kanehisa M, *Trends Genet* 13(9): 375-376.(1997)

15.   BioCarta http://www.biocarta.com/genes/index.asp
16.   NetPath http://www.netpath.org/.
17.   Database RG http://rgd.mcw.edu/.
18.   Van Dongen S, *PhD thesis, University of Utrecht*.(2000)
19.   Nair S, Lee YH, Rousseau E, Cam M, Tataranni PA et al., *Diabetologia* 48(9): 1784-1788.(2005)
20.   Park JJ, Berggren JR, Hulver MW, Houmard JA, Hoffman EP, *Physiological genomics* 27(2): 114-121.(2006)
21.   Gunton JE, Kulkarni RN, Yim S, Okada T, Hawthorne WJ et al., *Cell* 122(3): 337-349.(2005)
22.   Strang G Introduction to Linear Algebra. Cambridge: Wellesley; .(2003)
23.   Raychaudhuri S, Stuart JM, Altman RB, *Pacific Symposium on Biocomputing*: 455-466.(2000)
24.   Everitt BS, Dunn G, editors (1992) Applied Multivariate Data Analysis. New York, NY: Oxford University Press.
25.   Tusher VG, Tibshirani R, Chu G, *Proceedings of the National Academy of Sciences of the United States of America* 98(9): 5116-5121.(2001)
26.   King JY, Ferrara R, Tabibiazar R, Spin JM, Chen MM et al., *Physiological genomics* 23(1): 103-118.(2005)
27.   Kittleson MM, Minhas KM, Irizarry RA, Ye SQ, Edness G et al., *Physiological genomics* 21(3): 299-307.(2005)
28.   Singhal S, Kyvernitis CG, Johnson SW, Kaiser LR, Liebman MN et al., *Cancer biology & therapy* 2(4): 383-391.(2003)
29.   Bullinger L, Dohner K, Bair E, Frohling S, Schlenk RF et al., *The New England journal of medicine* 350(16): 1605-1616.(2004)
30.   Hausman DB, DiGirolamo M, Bartness TJ, Hausman GJ, Martin RJ, *Obes Rev* 2(4): 239-254.(2001)
31.   Blaak EE, *The Proceedings of the Nutrition Society* 63(2): 323-330.(2004)
32.   Petersen KF, Befroy D, Dufour S, Dziura J, Ariyan C et al., *Science* 300(5622): 1140-1142.(2003)
33.   Lowell BB, Shulman GI, *Science* 307(5708): 384-387.(2005)
34.   Frayn KN, *Biochemical Society transactions* 31(Pt 6): 1115-1119.(2003)
35.   Patti ME, Kahn BB, *Nature medicine* 10(10): 1049-1050.(2004)
36.   Ghazalpour A, Doss S, Sheth SS, Ingram-Drake LA, Schadt EE et al., *Genome biology* 6(7): R59.(2005)
37.   Baum JI, O'Conner JC, Seyler JE, Anthony TG, Freund GG et al., *The American journal of physiology*(288): E86-91.(2005)
38.   Kelley DE, Mintun MA, Watkins SC, Simoneau JA, Jadali F et al., *The Journal of clinical investigation* 97(12): 2705-2713.(1996)
39.   Bloomgarden ZT, *Diabetes care* 23(10): 1584-1590.(2000)