전사 조절 부위의 SNP 를 이용한 pathway 내 유전자 변이 연관관계 분석 및 네트워크 구축

연세대학교 대학원

의과학과

유 진 호

전사 조절 부위의 SNP 를 이용한 pathway 내 유전자 변이 연관관계 분석 및 네트워크 구축

지도교수 라 선 영

이 논문을 박사 학위논문으로 제출함

2008년 12월

연세대학교 대학원

의과학과

유 진 호

유진호의 박사 학위논문을 인준함

심사위원 라 선 영 인

심사위원 박 전 한 인

심사위원 이 진 성 인

심사위원 정 현 철 인

심사위원 김 양 석 인

연세대학교 대학원

2008년 12월

감사의 글

공학도로서 학문이라는 세계를 처음 접한 후로 18년째, 이학도로서 새로운 학문의 세계에 들어갈 기점을 오늘에서야 마련하게 되었습니다. 6년여 전 지루하게 반복되던 증권 업무에서 벗어나고자 시작했던 생물학 연구가 오늘 자그마한 결실을 이룰 수 있었던 것은 새로운 것을 알고자 하는 호기심, 그리고 막연한 동경심이지 않았나 싶습니다. 연구를 시작하면서 간혹 무언가를 알게 되었다는 기쁨을 느꼈던 순간도 있었지만, 이 글을 쓰는 오늘은 모르는 사실이 너무도 많음이 막연한 두려움의 벽으로 주위를 둘러싸고 있습니다. 그러나, 두려움에 포기할 정도면 처음부터 시도조차 하지 않았을 것을 마음 깊이 느끼고 있음에 또 다행한 마음이 듭니다.

2002년 한일 월드컵이 끝날 즈음, 반복되는 증권 관련 업무에 짜증을 내던 한 프로그래머에게 생명정보공학이라는 세계를 보여주신 김양석 박사님께 감사 드립니다. 나날이 접해본 새로운 사실들은 정말 재미있었습니다. 그리고, 라선영 선생님. 귀납적 사고방식에 묶여있던 고지식한 영혼을 연역이라는 새로운 세계로 이끌어 주셨습니다. 엉망으로 섞여있던 생각을 명쾌한 방향으로 이끌어 주신 것 감사 드립니다. 항상 인자한 미소로 맞아주신 정현철 선생님, 박사과정 첫 학기에 인상 깊은 강의를 해주신 정현철 선생님께도 감사 드립니다. 귀찮은 일들을 부탁해도 싫은 내색

없이 도움을 주었던 찬희, 재준이, 그리고 모든 연구실 식구들께도 감사 드립니다.

부모님과 형제들은 제가 앞으로 평생 갚아야 할 고마움의 대상입니다. 인생의 방향을 바꿀 때마다 아무런 내색 없이 전적으로 저를 믿어주셨던 아버지, 어머니께 큰절을 올립니다. 그리고 제가 힘들었을 때 조언을 아끼지 않았던 큰형, 작은형, 그리고 여동생 에게도 감사의 말을 전합니다. 집에 큰일이 있을 때 마다 신경을 써주신 큰형수, 그리고 현우, 현영이, 우리 꼬맹이 예림이도 삼촌을 즐겁게 해주어서 고마움을 전합니다. 정일, 영호, 성수, 인천, 그리고 성근이. 언제든지 연락해도 만나서 술 한잔 기울일 수 있는 저의 평생지기들 에게도 고마움을 전합니다.

모든 것은 결코 늦지 않았다는 것을 느낍니다. 다사다난했던 2008년을 회고하며, 항상 새로운 것을 바라보는 눈으로 즐겁게 생활하고자 합니다.

저자 씀

차 례

그림 차례

표 차례

국문요약

전사 조절 부위의 SNP 를 이용한 pathway 내 유전자 변이 연관 관계 분석 및 네트워크 구축

인간 유전체에서 발견되는 염기서열 변이 중 단일염기변이(single nucleotide polymorphism, SNP) 를 이용하여 진행한 많은 연구들은 대부분 동일 염색체 내의 유전자들만을 대상으로 하기 때문에 서로 다른 염색체에 분포하고 있는 유전자들간의 유전적인 연관 관계를 네트워크 차원에서 해석하기 어렵다는 단점이 있다.

본 연구에서는 인간의 전체 유전체를 대상으로, Pearson's chi-square test 를 이용하여 유전자의 전사 조절 후보 영역에서 발견되는 SNP genotype 들간의 연관 관계를 분석하였다. 또한, SNP 들을 이용하여 추정된 haplotype(일배체형) 들로 구성한 diplotype(이배체형) 패턴을 이용하여 유전자 쌍 간의 연관 관계를 분석한 후 SNP genotype 들간, 그리고 haplotype 들간에 강한 연관 관계가 있는 유전자들을 이용하여 유전자 연관 관계 네트워크를 구축하였다. 연구를 위해 International HapMap 데이터베이스에서 SNP genotype 자료와 KEGG pathway 데이터베이스에서 pathway 정보를 가져와 이용하였다.

분석 결과, 동일 pathway 의 subtype 을 구성하는 유전자들 간에는 무작위 추출 유전자들에 비해 SNP genotype 들간,

그리고 haplotype 들간에 연관 관계가 있을 확률이 현저히 높았으며, 연관 관계가 있는 유전자들을 대상으로 유전자 연관 관계 네트워크를 구축할 수 있었다. 구축한 네트워크에서는 서로 다른 염색체에 위치해 있는 다수의 유전자들과 직접적인 연관 관계를 맺고 있는 중심 유전자들을 확인할 수 있었으며, 중심 유전자 및 중심 유전자와 연관 관계를 맺고 있는 유전자들이 다수의 서로 다른 pathway 로 연결되어 있음을 확인할 수 있었다. 또한, RNA expression 데이터 분석 결과, 발현 상관 관계를 가지고 있는 유전자들 일부가 유전자 연관 관계 네트워크에서 서로 연결되어 있음을 확인할 수 있었다.

따라서, pathway 를 구성하는 유전자들 중 서로 직접적인 반응 관계에 있는 유전자들의 전사 조절 후보 영역 내 SNP genotype 및 haplotype 들은 무작위로 나타나는 것이 아니라 서로 연관 되어 있음을 확인 할 수 있었다. 또한, SNP genotype 및 haplotype 연관 관계를 맺고 있는 유전자들을 이용하여 유전자 연관 관계 네트워크를 구축할 수 있으며, 네트워크 상에 매핑된 pathway 정보와 유전자 연관 관계를 접목하면 네트워크를 구성하는 중심 유전자 및 서로 다른 pathway 를 연결시켜 주는 유전자들을 찾을 수 있음을 확인할 수 있었다.

핵심되는 말 : 단일염기변이, 일배체형, 이배체형, pathway, 전사 조절 영역, genotype 연관 관계, haplotype 연관 관계, 유전자 변이 연관 관계 네트워크

전사 조절 부위의 SNP 를 이용한 pathway 내 유전자 변이 연관 관계 분석 및 네트워크 구축

<지도교수 **라 선 영**>

연세대학교 대학원 의과학과

유 진 호

Ⅰ. 서 론

International HapMap(http://www.hapmap.org/) 프로젝트가 완성된 이후로 인간 유전체에서 광범위하게 발굴된 염기서열변이를 이용한 연구가 활발하게 진행되고 있다. 특히, 염기서열변이 중 가장 단순한 형태인 단일염기변이(single nucleotide polymorphism, SNP) 는 인간 유전체에서 발견되는 염기서열변이의 90% 이상을 차지하고 있을 정도로 흔하게 발견되며[1] 현재까지 1,000 만건 이상의 SNP 가 보고되고 있다(http://www.ncbi.nlm.nih.gov/projects/SNP/).

전통적으로 유전학에서는 SNP 와 질병과의 연관 분석(association analysis)을 수행해 왔으며, 최근에는 인간 전체 유전체에서 발견되는 SNP 를 대상으로 한 대규모 genotyping 이 가능해짐에 따라[2] 다양한 연구가 시도되고 있다[3-10]. 특히, 암 위험 인자와 연관이 있는 것으로 밝혀진 유전자들을 대상으로 하여 이들 유전자들에서

발견되는 SNP 간의 연관 불평형(linkage disequilibrium) 분석 및 SNP 들로 구축한 haplotype 구조에 대한 연구[11], haplotype 과 약물 반응과의 연관 관계 분석[12], 데이터 마이닝 기법을 이용한 SNP pair 간의 상호작용 연구[13], 유전자와 haplotype 간의 상호작용 연구[14] 등 발암기전 및 다양한 질병 기전에 대한 연구나 약물을 이용한 질병 치료 전략 발굴에도 다양하게 이용되고 있다. 하지만, SNP 또는 haplotype 과 같은 유전학적인 근거를 이용해 수행한 기존 연구에서는 대부분이 특정 질병을 대상으로 하거나 동일 염색체의 특정 영역에 분포되어 있는 유전자를 대상으로 분석을 하였기 때문에 서로 다른 염색체에 분포한 유전자들간의 유전학적인 연관 관계 분석이 미흡하다는 단점이 있다.

최근에는 유전자 발현을 유전학적인 근거에서 연구하고자 하는 genetical genomics[15-17] 가 대두되고 있다. 일례로, breast cancer 에서의 유전자 발현 양상과 SNP genotype 간의 연관 관계를 분석한 연구에서는 한 개 또는 여러 개 SNP 에 대해서 유의한 연관 관계를 보이는 전사체(transcript) 들이 gene ontology (GO) 및 pathway 에서 밀접하게 연결되어 있음을 보고하고 있으며, cis 형태로 유전자 발현을 조절하는 SNP 들 중에는 강한 연관 불평형 관계를 가지는 것들이 있음을 보이고 있다[18]. 또한, 유전자의 발현 정도를 유의적으로 조절하는 cis 및 trans-acting loci 를 찾는 연구[18], 마이크로어레이 기법과 양적 형질(quantitative trait) 에 대한 linkage analysis 를 접목하여 유전자 발현 변이에 기여하는 결정요인을 유전학적으로 매핑한 연구도 있다[19]. 한편, 동일 pathway 에 속해있는 유전자들은 무작위로 추출한 유전자들에 비해 발현 상관 관계가 높음을 밝혀낸 사례도 있다[20].

본 연구에서는 인간의 전체 유전체를 대상으로, 유전자의 전사

조절 후보 영역에서 발견되는 SNP 간의 genotype 및 haplotype 연관 관계를 밝히고, 연관 관계 분석 결과를 이용하여 pathway 를 구성하는 유전자들 간의 유전학적인 연관 관계 네트워크를 구축할 수 있으며, 구축된 네트워크에서 다수의 pathway 와 유전자들 간의 상호 연결 관계를 파악할 수 있음을 보이고자 하였다.

Ⅱ. 재료 및 방법

1. 대상 및 자료 수집

가. SNP genotype 데이터

본 연구에서는 International HapMap 프로젝트 데이터베이스(http://www.hapmap.org/)에서 제공하는 여러 인종의 individual genotype 데이터 중 아시아인인 Japanese, Chinese 인종을 대상으로 한 90명의 genotype 데이터를 이용하였다. 유전자 정보는 NCBI (http://www.ncbi.nlm.nih.gov/) 의 genome build 36.2 를 기준으로 하였고, 유전자의 전사 시작 위치로부터 5,000 base pair 앞까지를 유전자 전사 조절 후보 영역으로 설정 한 후 이 영역에 속하는 SNP 들을 분석 대상으로 하였다.

나. pathway 데이터

Pathway 정보는 KEGG 데이터베이스(http://www.genome.jp/kegg/) 의 KEGG release 43.0 을 기준으로 202 개의 pathway 정보를 이용하였다. Pathway 는 구성 유전자들이 코딩하는 단백질들이 반응하는 방법에 따라 intra reaction, paired reaction, binary relation 의 세가지 군으로 구분하였으며, binary relation 의 경우 7개의 subtype 으로 다시 세분화 하였다.

(1) intra reaction

동일한 화학적 반응에 참여하는 단백질을 코딩 하는 유전자 집단으로 정의하였다.

(2) paired reaction

하나의 화학적 반응에서 생성된 산물이 다른 화학적 반응에서는 기질로 이용될 때 이 두 개의 화학적 반응에 포함된 단백질을 코딩하는 유전자 집단으로 정의하였다.

(3) binary relation

쌍으로 이루어진 두 개 단백질 간의 반응 관계를 기술한 것으로, KEGG pathway 에서는 7개의 subtype 으로 세분화 하였다(표 1).

표 1. Binary relation 을 구성하는 7개 subtype

Subtype no	Subtype name	Description
subtype1	compound	shared with two successive reactions or intermediate of two interacting proteins
	hidden compound	shared with two successive reactions but not displayed in the pathway map
subtype2	activation	positive and negative effects which may be associated with molecular information below interactions via DNA binding
	inhibition	
subtype3	expression	indirect effect without molecular details state transition
	repression	
subtype4	indirect effect	association and dissociation
subtype5	state change	molecular events
subtype6	binding/association	description shared with two successive reactions or intermediate of two interacting proteins
	dissociation	
subtype7	phosphorylation	shared with two successive reactions but not displayed in the pathway map
	dephosphorylation	
	glycosylation	
	ubiquitination	
	methylation	

다. RNA expression 데이터

본 연구실(cancer metastasis research center, CMRC) 에서 human 17K cDNA microarray 를 이용하여 실험한 유전자 발현 데이터를 이용하였다. 임상 시료는 127명의 환자로부터 추출한 254개의 정상

대장 조직(normal colon mucosal tissue)과 대장암 조직(colon tumor tissue)을 이용하여 실험한 RNA expression 데이터를 이용하였다[21].

아래는 연구 대상을 종합적으로 나타낸 것이다(그림 1).

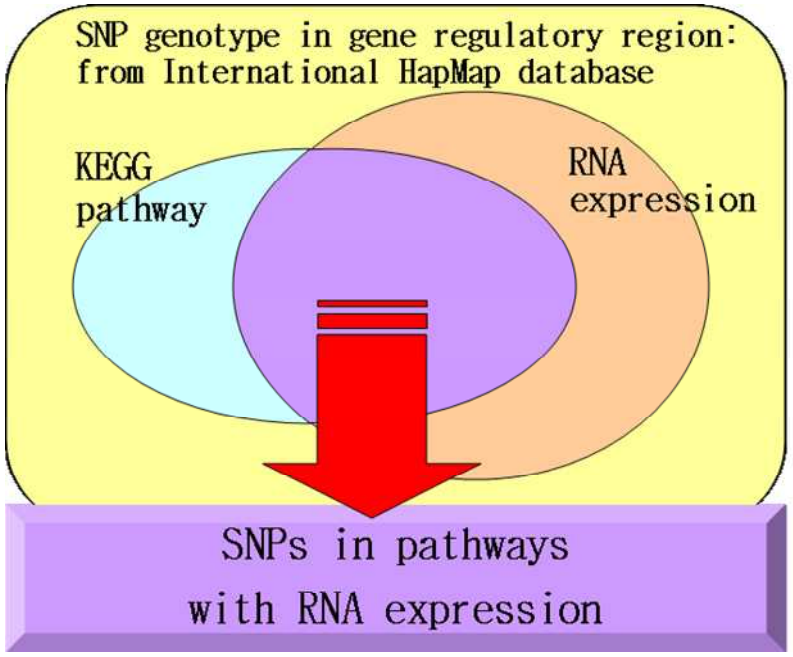

그림 1. 연구 대상. International HapMap database 에서 제공하는 SNP genotype 중 유전자 전사 조절 후보 영역의 SNP 를 연구 대상으로 하였다. 또한, KEGG pathway 정보가 있는 유전자들에 속한 SNP 및 RNA expression 데이터를 이용하여 세부적인 연구를 수행하였다.

2. 연구 방법

가. SNP pair 간의 genotype 연관 관계 분석

유전자 전사 조절 후보 영역에 포함되어 있는 SNP 를 대상으로 하여 genotype 연관 관계 분석을 하였다. SNP 에서 관측되는 세가지 genotype 을 이용하여 두 개 SNP 간의 genotype pattern 을 3x3

contingency 테이블로 만든 후 Pearson's chi-square test 를 수행하였다. 계산된 p-value 가 유의수준 α=0.01 보다 작게 나온 경우 두 SNP 는 genotype 연관 관계가 있다고 정의하였다.

나. 유전자 쌍 간의 haplotype 연관 관계 분석

유전자 전사 조절 후보 영역에 포함되어 있는 SNP 들 중 강한 연관 불평형 관계에 있는 SNP 들을 이용하여 haplotype 을 추정하였으며, 추정된 haplotype 을 이용하여 연관 관계 분석을 하였다. 유전자의 전사 시작 위치에서 가장 가까이 있으면서 가장 높은 빈도로 추정된 haplotype 을 해당 유전자를 대표하는 haplotype 으로 설정하였다. 유전자 대표 haplotype 과 나머지 haplotype 들을 조합하여 두 개 유전자 간의 diplotype pattern 을 3x3 contingency 테이블로 만든 후 Pearson's chi-square test 를 수행하였다. 계산된 p-value 가 유의수준 α=0.01 보다 작게 나온 경우 두 유전자는 haplotype 연관 관계가 있다고 정의하였다. 추정된 haplotype 들 중 h_1 은 두 개 유전자 중 한 유전자에서 추정된 최빈도 haplotype 을, h_1^* 및 h_1^{**} 은 나머지 haplotype 을 나타내며, h_2 는 다른 유전자에서 추정된 최빈도 haplotype 을, h_2^* 및 h_2^{**} 는 나머지 haplotype 을 나타낸다. O_{ij} 는 해당 diplotype pattern 의 수를 나타낸다(표 2).

표 2. 두 유전자에서 추정된 haplotype 들로 구성한 diplotype pattern

Diplotype pattern		Diplotype of another gene		
		h_2 / h_2	h_2 / h_2^*	h_2^*/ h_2^{**}
Diplotype of one gene	h_1 / h_1	O_{11}	O_{12}	O_{13}
	h_1 / h_1^*	O_{21}	O_{22}	O_{23}
	h_1^*/ h_1^{**}	O_{31}	O_{32}	O_{33}

다. SNP 들간의 연관 불평형 분석 및 haplotype 추출

유전자 전사 조절 후보 영역의 SNP 들 중 강한 연관 불평형 관계에 있는 SNP 들을 이용하여 linkage disequilibrium block (LD block) 을 구축하였으며[22], 유전자의 전사 시작 위치에서 가장 가까이 있는 LD block 에서 haplotype 을 추정하여[23] 유전자들 간의 haplotype 연관 관계 분석에 이용하였다.

라. 연관 관계 쌍 개수 비율 분석

(1) SNP pair 분석

모든 유전자를 대상으로 전사 조절 후보 영역에서 무작위로 1,000 개의 SNP 를 추출한 후 genotype 연관 관계가 있는 SNP pair 개수의 비율을 계산하였다. 그리고, 각 유전자를 대상으로 동일 유전자의 전사 조절 후보 영역 내 SNP 들간에 genotype 연관 관계가 있는 SNP pair 개수의 비율을 계산한 후 무작위 추출한 결과와 비율 차이 검정을 하였다. 또한, pathway 를 구성하는 intra reaction, paired reaction, 그리고 binary relation 내의 7개 subtype 에 포함되어 있는 유전자들을 대상으로 전사 조절 후보 영역에서 발견되는 SNP 들간에 genotype 연관 관계가 있는 SNP pair 개수의 비율을 계산한 후 무작위 추출한 결과와 비율 차이 검정을 하였다. 비율 차이 검정은 모두 Z-test 를 이용하였으며 유의수준은 α=0.01 로 설정하였다.

(2) 유전자 쌍 분석

Haplotype 이 추정된 모든 유전자들을 대상으로 1,000 개의 유전자를 무작위 추출한 후 haplotype 연관 관계가 있는 유전자 쌍

개수의 비율을 계산하였다. 그리고, pathway 를 구성하는 intra reaction, paired reaction, 그리고 binary relation 내의 7개 subtype 중 상기한 SNP pair 간의 genotype 연관 관계 분석 결과 무작위 추출한 결과와 통계적으로 차이가 나는 것들을 대상으로 하여 haplotype 연관 관계가 있는 유전자 쌍 개수의 비율을 계산한 후 무작위 추출한 결과와 비율 차이 검정을 하였다. 비율 차이 검정은 모두 Z-test 를 이용하였으며 유의수준은 α=0.01 로 설정하였다.

마. 유전자 연관 관계 네트워크 구축 및 pathway 매핑

RNA expression 실험 결과가 있는 유전자들 중 SNP pair 간에 genotype 연관 관계 및 haplotype 연관 관계가 있는 유전자들을 이용하여 유전자 연관 관계 도표(표 3) 를 만든 후 유전자 연관 관계 네트워크를 구축하였다. 네트워크를 구성하는 유전자들에 KEGG pathway 를 매핑한 후 동일 pathway 에 속해 있는 유전자 집단을 표시하였다. 그림 2는 구축한 네트워크의 예를 나타내고 있는데, G1 ~ G5 는 각각 유전자를 나타내며 유전자 간에 SNP genotype 또는 haplotype 연관 관계가 있는 경우 선으로 연결하였다.

표 3. 유전자 연관 관계 도표 예

	G1	G2	G3	G4	G5
G1		o	x	o	o
G2			x	o	x
G3				x	o
G4					x
G5					

G1 ~ G5 는 각각 서로 다른 유전자를 나타내고 있으며 o 는 두 유전자 간에 연관 관계가 있음을, x 는 연관 관계가 없음을 나타냄.

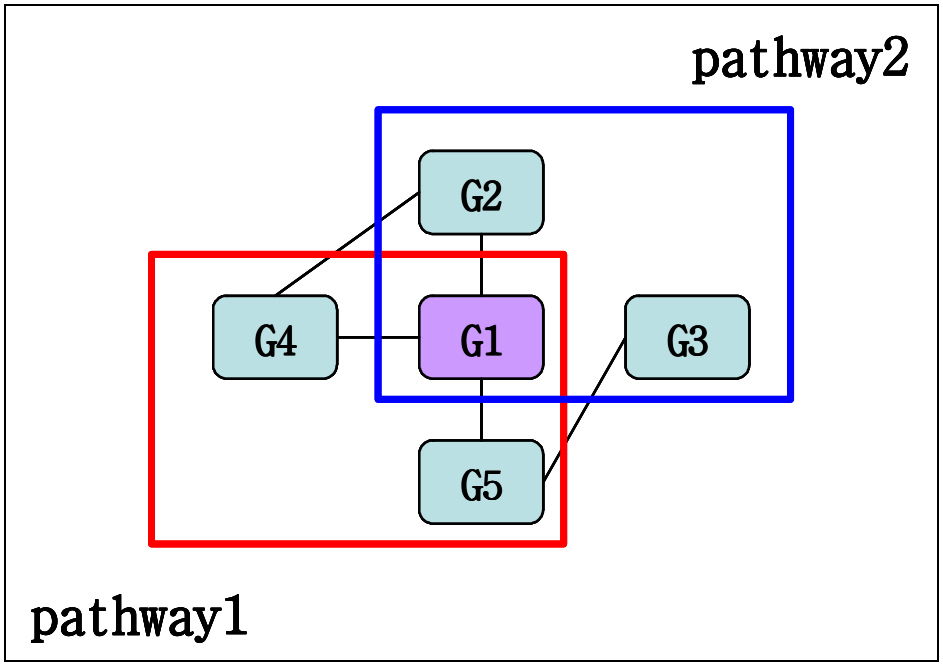

그림 2. 유전자 연관 관계 네트워크. Pathway1 은 G1, G4, G5 유전자를 포함하고 pathway2 는 G1, G2, G3 유전자를 포함하고 있음. 특히, G1 의 경우 pathway1 과 pathway2 모두에 포함된 유전자임을 나타내고 있음.

바. RNA expression 데이터 분석

SNP pair 의 genotype 간, 그리고 유전자 쌍의 haplotype 연관 관계가 있는 유전자들을 대상으로 아래와 같은 유전자 발현 관계 분석을 하였다.

(1) 유전자들간 발현 상관 관계 분석

샘플 i 의 조직 j 에서 서로 다른 유전자 A, B 가 발현된 RNA expression 양을 각각 a_{ji}, b_{ji} 라고 하였을 때, 정상 대장 조직(j=1) 과 대장암 조직(j=2) 각각에 대해서 유전자 A 의 발현량 벡터를 $A_j = \{a_{j1}, a_{j2}, \ldots, a_{j127}\}$, 유전자 B 의 발현량 벡터를 $B_j = \{b_{j1}, b_{j2}, \ldots,$

b_{j127}} 로 정의한 후 두 유전자 A, B 의 발현 상관 관계를 Pearson's correlation coefficient(=r) 를 이용하여 분석하였다. 귀무 가설을 r=0, 대립 가설을 r≠0 으로 설정한 후 자유도가 2 인 T-test 를 이용하여 검정하였다. T-test 결과 계산된 p-value 가 유의수준 α=0.05 보다 작을 경우 두 유전자 A, B 는 발현 상관 관계에 있다고 정의하였다.

(2) 대장암 특이 발현 유전자 분석

정상 대장 조직 대비 대장암 조직에서 과발현(up-regulation) 또는 하향조절(down-regulation) 되는 유전자들을 분석하였다. 유전자의 RNA expression 양이 정상 대장 조직을 기준으로 하였을 때 대장암 조직에서 2.0 배 이상 발현된 것을 과발현, 그리고 0.5 배 이하로 발현된 것을 하향조절 되었다고 정의하였다.

Ⅲ. 결 과

1. SNP, 유전자 및 pathway 추출

International HapMap 프로젝트 데이터베이스에 등록되어 있는 genotype 데이터 중 Y 염색체를 제외한 1번 ~ 22번 염색체 및 X 염색체를 대상으로 하여 SNP genotype 데이터를 가져왔다. Y 염색체에 포함된 SNP 는 missing genotype 비율이 50% 이상이기 때문에 본 연구에서는 제외하였다. 대립유전자(allele) 두 개가 모두 관측되는 SNP 는 모두 2,473,439 개 였으며 이들 중 유전자 전사 조절 후보 영역에 포함되어 있는 SNP 는 81,117 개 였고, 전사 조절 후보 영역에 SNP 를 한 개 이상 포함하고 있는 유전자는 20,378 개였다. 유전자 한 개당 전사 조절 후보 영역에 포함되어 있는 SNP 수의 평균과 표준 편차는 4.0 ± 3.5 개 였다.

81,117 개의 SNP 에서 무작위로 1,000 개를 추출하였으며 이 과정을 1,000 번 반복하였다. 한번 무작위 추출에 포함되어 있는 유전자 수의 평균과 표준 편차는 957.7 ± 6.2 개 였는데, 이것은 유전자 한 개에서 약 1개의 SNP 가 추출되었음을 의미한다. 전사 조절 후보 영역에 두 개 이상의 SNP 를 포함하고 있는 유전자는 15,815 개 였으며 이들 유전자에 포함되어 있는 SNP 는 모두 76,540 개 였다.

202 개의 KEGG pathway 에는 3,144 개의 유전자가 포함되어 있었다. 서로 다른 pathway 에 포함되어 있는 동일 유전자를 중복해서 계산할 경우 7,502 개 였으며 SNP 는 27,522 개 였다(표 4). Pathway 한 개당 포함하고 있는 유전자 수의 평균과 표준 편차는 38.1 ± 37.5 개 였으며, SNP 수의 평균과 표준 편차는 136.4 ± 147.7 개 였다.

표 4. 추출된 SNP 및 유전자 수

Category		SNP count	Gene count
random		1,000	≈ 1,000*
intra regulatory region		76,540**	15,815**
pathway		27,522***	7,502***
reactions and pathway subtypes	paired reaction	4,106	1,111
	intra reaction	2,332	608
	subtype1	734	217
	subtype2	4,780	1,320
	subtype3	1,410	337
	subtype4	651	207
	subtype5	566	20
	subtype6	1,732	424
	subtype7	836	275

*1,000 개 SNP 를 무작위 추출했을 때 포함되어 있는 유전자 수의 근사치. 무작위 추출을 1,000번 반복했을 때 유전자 수의 평균과 표준편차는 957.7 ± 6.2.

**유전자 전사 조절 후보 영역에 2 개 이상의 SNP 를 포함하는 유전자 및 SNP 수.

***서로 다른 pathway 에 포함되어 있는 동일 유전자를 중복해서 계산했을 때의 유전자 및 유전자 전사 조절 영역 내의 SNP 수.

random: 무작위 추출한 경우.

intra regulatory region: 유전자 전사 조절 영역 내인 경우.

pathway: 동일 pathway 에 속한 경우.

paired reaction: paired reaction 에 속한 경우.

intra reaction: intra reaction 에 속한 경우.

subtype1: binary relation 중 subtype1 에 속한 경우.

subtype2: binary relation 중 subtype2 에 속한 경우.

subtype3: binary relation 중 subtype3 에 속한 경우.

subtype4: binary relation 중 subtype4 에 속한 경우.

subtype5: binary relation 중 subtype5 에 속한 경우.

subtype6: binary relation 중 subtype6 에 속한 경우.

subtype7: binary relation 중 subtype7 에 속한 경우.

2. SNP pair 간의 genotype 연관 관계 분석

무작위로 추출한 SNP 들간에 약 5.0×10^8 번의 genotype 연관 관계 분석을 하였다(그림 3-A). Pearson's chi-square test 결과 계산된 모든 p-value 의 평균과 표준 편차는 0.513 ± 0.280 이었으며, genotype 연관 관계를 보이는 p-value 개수의 비율은 1.80% 였다. 유전자 각각을 대상으로 하였을 때 전사 조절 후보 영역내의 SNP 들간에 244,248 번의 genotype 연관 관계 분석을 하였으며, 분석 결과 계산된 모든 p-value 의 평균과 표준 편차는 0.159 ± 0.268 이었다. Genotype 연관 관계를 보이는 p-value 개수의 비율은 59.71% 로 나타났는데, 무작위 추출 결과인 1.80% 와 비율 차이 검정을 했을 때 통계적으로 차이가 났다. Pathway 에 속한 유전자들 간에는 3,976,841 번의 SNP genotype 연관 관계 분석을 하였으며, 분석 결과 p-value 의 평균과 표준 편차는 0.508 ± 0.283 이었다. Genotype 연관 관계를 보이는 p-value 개수의 비율은 2.46% 로 나타났는데, 무작위 추출 결과인 1.80% 와 비율 차이 검정을 했을 때 통계적으로 차이가 났다(그림 3-A).

특히, pathway 를 구성하는 intra reaction 과 paired reaction, 그리고 binary relation 내의 subtype2, subtype3, subtype5, subtype6 의 경우 genotype 연관 관계를 보이는 p-value 개수의 비율은 무작위 추출 결과인 1.80% 와 비율 차이 검정을 했을 때 통계적으로 차이가 났다($p<0.000$). Subtype4 의 경우에도 무작위 추출 결과와 비교했을 때 통계적으로 차이가 났으나($p=0.004$), subtype1, subtype7 의 경우에는 무작위 추출 결과와 통계적인 차이가 나지 않았다(그림 3-B).

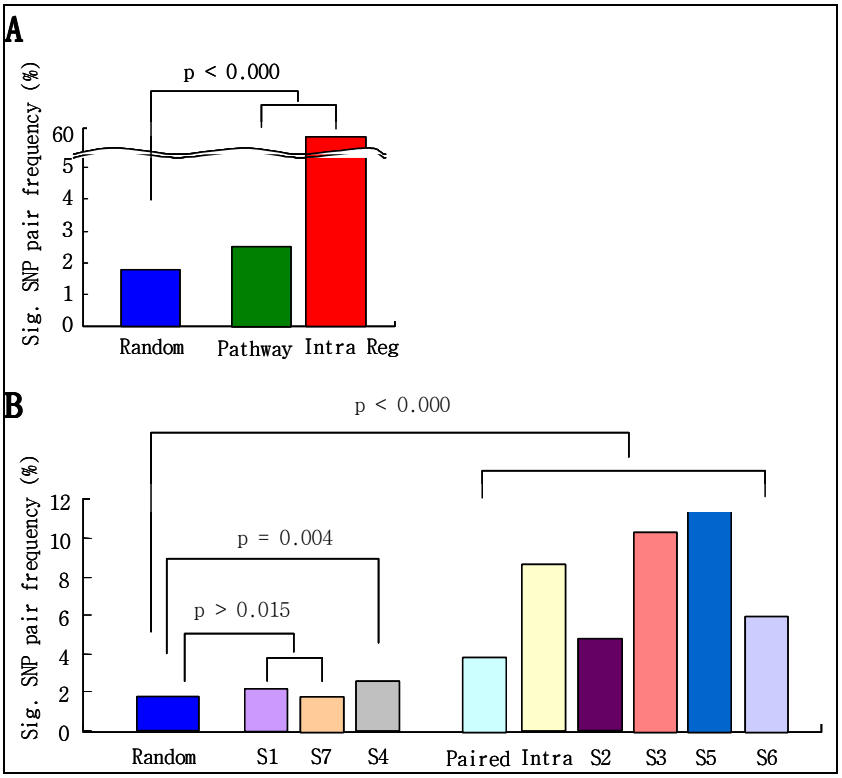

그림 3. Genotype 연관 관계를 보이는 p-value 개수의 비율 차이 검정.

(A) 무작위 추출과 pathway, intra regulatory region 과의 비교, (B) 무작위 추출과 intra reaction, paired reaction, binary relation 과의 비교.

Random: 무작위 추출한 SNP 대상.

Pathway: 동일 pathway 에 속한 SNP 대상.

Intra Reg: 유전자 전사 조절 후보 영역 내의 SNP 대상.

Paired: paired reaction 에 속한 SNP 대상.

Intra: intra reaction 에 속한 SNP 대상.

S1: binary relation 중 subtype1 에 속한 SNP 대상.

S2: binary relation 중 subtype2 에 속한 SNP 대상.

S3: binary relation 중 subtype3 에 속한 SNP 대상.

S4: binary relation 중 subtype4 에 속한 SNP 대상.

S5: binary relation 중 subtype5 에 속한 SNP 대상.

S6: binary relation 중 subtype6 에 속한 SNP 대상.

S7: binary relation 중 subtype7 에 속한 SNP 대상.

Genotype 연관 관계가 있는 두 SNP 가 동일 pathway에 있을 확률은 Bayesian rule을 이용해 계산할 수 있다(그림 4). 즉, SNP S_i 와 S_j 사이에 genotype 연관 관계가 있다고 할 때 S_i 와 S_j 가 동일 pathway 에 속할 확률인 p_1 은 S_i 와 S_j 의 genotype 연관 관계를 모르는 상황에서 동일 pathway 에 속할 확률인 p_2 에 대해 $p_1 = W \times p_2$ 의 관계를 가진다. 일반적으로 p_1 과 p_2 는 알려지지 않은 확률이지만 W 는 genotype 연관 관계가 있는 SNP pair 개수 비율을 이용해 계산된 값이다. 계산 결과, genotype 연관 관계가 있는 두 개 SNP 가 pathway 에 속해 있을 확률은 SNP 간의 genotype 연관 관계를 모를 경우에 비해 1.37 배(W=1.37) 높아지는 것으로 나타났다. 특히, subtype5 에 속할 확률은 6.48배(W=6.48) 높아지는 것으로 나타났다(표 5).

$sig(S_i,S_j)$: event that SNP S_i and S_j are in genotype relationship
$path(S_i,S_j)$: evnet that SNP S_i and S_j are in the same pathway
$p_1 = p[path(S_i,S_j) | sig(S_i,S_j)]$: unknown probability
$p_2 = p[path(S_i,S_j)]$: unknown probability

$$p_1 = W \times p_2$$

$$W = \frac{p[sig(S_i,S_j) | path(S_i,S_j)]}{p[sig(S_i,S_j)]} \quad \text{by Bayesian rule}$$

그림 4. Bayesian rule 을 이용한 SNP pair 연관 관계 해석.

p_1: 두 SNP 사이에 genotype 연관 관계가 있다고 할 때 동일 pathway 에 속할 확률.

p_2: 두 SNP 의 genotype 연관 관계를 모를 때 동일 pathway 에 속할 확률.

$p[sig(S_i,S_j) \mid path(S_i,S_j)]$: 동일 pathway에 속한 SNP pair들 중 genotype 연관 관계가 있는 pair 개수의 비율로, 동일 pathway에 속한 두 SNP 사이에 genotype 연관 관계가 있을 확률의 추정치.

$p[sig(S_i,S_j)]$: 무작위 추출한 SNP pair들 중 genotype 연관 관계가 있는 pair 개수의 비율로, 무작위 추출한 두 SNP 간에 genotype 연관 관계가 있을 확률의 추정치.

표 5. Genotype 연관 관계가 있는 SNP pair 개수 비율

Category		Sig. SNP pair frequency (%)	Frequency ratio relative to random[1]
random		1.80	1.00
intra regulatory region		59.71	33.17
pathway		2.46	1.37
reactions and pathway subtypes	paired reaction	4.04	2.24
	intra reaction	8.61	4.78
	subtype2	4.71	2.62
	subtype3	10.36	5.76
	subtype4	2.26	1.26
	subtype5	11.67	6.48
	subtype6	6.05	3.36

[1]Frequency ratio relative to random = Sig. SNP pair frequency (%) / 1.80 (%).

random: 무작위 추출한 SNP 대상.

intra regulatory region: 유전자 전사 조절 후보 영역 내의 SNP 대상.

pathway: 동일 pathway 에 속한 SNP 대상.

paired reaction: paired reaction 에 속한 SNP 대상.

intra reaction: intra reaction 에 속한 SNP 대상.

subtype2: binary relation 중 subtype2 에 속한 SNP 대상.

subtype3: binary relation 중 subtype3 에 속한 SNP 대상.

subtype4: binary relation 중 subtype4 에 속한 SNP 대상.

subtype5: binary relation 중 subtype5 에 속한 SNP 대상.

subtype6: binary relation 중 subtype6 에 속한 SNP 대상.

3. Haplotype 추정

유전자 전사 조절 후보 영역의 SNP 들 중 강한 연관 불평형 관계에 있는 SNP 들을 이용하여 LD block 을 구축한 후[22], 유전자의 전사 시작 위치에서 가장 가까이 위치해 있는 LD block 에서 haplotype 을 추정 하였다[23]. 20,378 개의 유전자들 중 15,138 개의 유전자에서 haplotype 을 추정할 수 있었으며, 이들 유전자들을

대상으로 1,000 개를 무작위로 추출한 후 이 과정을 1,000 번 반복하여 haplotype 연관 관계 분석을 수행하였다. Pathway 에 포함되어 있는 유전자는 2,356 개 였다(표 6).

표 6. Haplotype 이 추정된 유전자 수

Category		Gene count
random		1,000*
pathway		2,356
reactions and pathway subtypes	paired reaction	783
	intra reaction	426
	subtype2	980
	subtype3	204
	subtype4	320
	subtype5	18
	subtype6	269

*한번 무작위 추출에 포함되어 있는 유전자 수.

random: 무작위 추출한 경우.

pathway: 동일 pathway 에 속한 경우.

paired reaction: paired reaction 에 속한 경우.

intra reaction: intra reaction 에 속한 경우.

subtype2: binary relation 중 subtype2 에 속한 경우.

subtype3: binary relation 중 subtype3 에 속한 경우.

subtype4: binary relation 중 subtype4 에 속한 경우.

subtype5: binary relation 중 subtype5 에 속한 경우.

subtype6: binary relation 중 subtype6 에 속한 경우.

4. 유전자 쌍 간의 haplotype 연관 관계 분석

SNP 간의 genotype 연관 관계 분석 결과 무작위 추출과 통계적으로 차이가 나는 pathway, intra reaction, paired reaction, 그리고 binary relation 을 구성하는 subtype2, subtype3, subtype4,

subtype5, subtype6 들을 대상으로 haplotype 연관 관계 분석을 하였다. 무작위 추출한 유전자들 간에 약 5.0×10^{8} 번의 haplotype 연관 관계 분석을 하였으며(그림 5-A), Pearson's chi-square test 결과 계산된 모든 p-value 의 평균과 표준 편차는 0.491 ± 0.280 이었다. Haplotype 연관 관계를 보이는 p-value 개수의 비율은 1.08% 로 나타났다. Pathway 에 속한 유전자들 간에는 126,908 번의 연관 관계 분석을 하였으며, 계산된 모든 p-value 의 평균과 표준 편차는 0.490 ± 0.280 이었다. Pathway 의 경우 haplotype 연관 관계를 나타내는 p-value 개수의 비율은 1.26% 로 나타났는데, 무작위 추출 결과인 1.08% 와 비율 차이 검정을 했을 때 통계적으로 차이가 났다(그림 5-A).

특히, pathway 를 구성하는 intra reaction 과 binary relation 내의 subtype2, subtype3, subtype5, subtype6 의 경우 haplotype 연관 관계를 보이는 p-value 개수의 비율은 무작위 추출 결과인 1.08% 와 비율 차이 검정을 했을 때 모두 통계적으로 차이가 났다($p<0.005$). 한편, paired reaction 과 subtype4 의 경우에는 무작위 추출 결과와 통계적인 차이가 나지 않았다(그림 5-B).

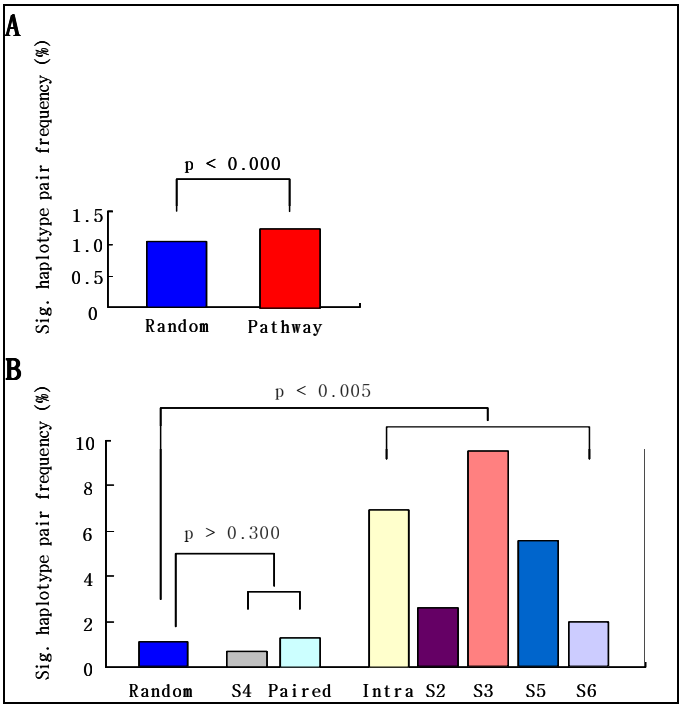

그림 5. Haplotype 연관 관계를 보이는 p-value 개수의 비율 차이 검정.
(A) 무작위 추출과 pathway 와의 비교, (B) 무작위 추출과 intra reaction, paired reaction, binary relation 과의 비교.
Random: 무작위 추출한 유전자 대상.
Pathway: 동일 pathway 에 속한 유전자 대상.
Paired: paired reaction 에 속한 유전자 대상.
Intra: intra reaction 에 속한 유전자 대상.
S2: binary relation 중 subtype2 에 속한 유전자 대상.
S3: binary relation 중 subtype3 에 속한 유전자 대상.
S4: binary relation 중 subtype4 에 속한 유전자 대상.
S5: binary relation 중 subtype5 에 속한 유전자 대상.
S6: binary relation 중 subtype6 에 속한 유전자 대상.

Haplotype 연관 관계가 있는 두 개 유전자가 동일 pathway 에 속해있을 확률은 Bayesian rule 을 이용해 계산할 수 있다(그림 6).

즉, 유전자 G_i 와 G_j 사이에 haplotype 연관 관계가 있을 때 G_i 와 G_j 가 동일 pathway 에 속할 확률인 q_1 은 G_i 와 G_j 의 haplotype 연관 관계를 모르는 상황에서 동일 pathway 에 속할 확률인 q_2 에 대해 $q_1 = W' \times q_2$ 의 관계를 가진다. 일반적으로 q_1 과 q_2 는 알려지지 않은 확률이지만 W' 는 haplotype 연관 관계가 있는 유전자 쌍 개수 비율을 이용해 계산된 값이다. 계산 결과, 두 유전자 사이에 haplotype 연관 관계가 있을 때 두 유전자가 동일 pathway 에 속해 있을 확률은 두 유전자들 간의 haplotype 연관 관계를 모를 경우에 비해 1.17 배(W'=1.17) 높아지는 것으로 나타났다. 특히, subtype3 에 속할 확률은 8.83배(W'=8.83) 높아지는 것으로 나타났다(표 7).

$sig(G_i,G_j)$: event that gene G_i and G_j are in haplotype relationship
$path(G_i,G_j)$: event that gene G_i and G_j are in the same pathway
$q_1 = p[path(G_i,G_j) | sig(G_i,G_j)]$: unknown probability
$q_2 = p[path(G_i,G_j)]$: unknown probability

$$q_1 = W' \times q_2$$
$$\Rightarrow W' = \frac{p[sig(G_i,G_j) | path(G_i,G_j)]}{p[sig(G_i,G_j)]} \quad \text{by Bayesian rule}$$

그림 6. Bayesian rule 을 이용한 유전자 쌍 연관 관계 해석.

q_1: 두 유전자 간에 haplotype 연관 관계가 있다고 할 때 동일 pathway 에 속할 확률.

q_2: 두 유전자의 haplotype 연관 관계를 모들 때 동일 pathway 에 속할 확률.

$p[sig(G_i,G_j) \mid path(G_i,G_j)]$: 동일 pathway에 속한 유전자 쌍 중 haplotype 연관 관계가 있는 유전자 쌍 개수의 비율로, 동일 pathway에 속해있는 두 유전자 간에 haplotype 연관 관계가 있을 확률의 추정치.

$p[sig(G_i,G_j)]$: 무작위 추출한 유전자 쌍 중 haplotype 연관 관계가 있는 유전자 쌍 개수의 비율로, 무작위 추출한 두 유전자 간에 haplotype 연관 관계가 있을 확률의 추정치.

표 7. Haplotype 연관 관계가 있는 유전자 쌍 개수 비율

Category		Sig. gene pair frequency (%)	Frequency ratio relative to random[1]
random		1.08	1.00
pathway		1.26	1.17
reactions and pathway subtypes	paired reaction	1.30	1.20
	intra reaction	7.02	6.50
	subtype2	2.61	2.42
	subtype3	9.54	8.83
	subtype4	0.63	0.58
	subtype5	5.56	5.15
	subtype6	1.99	1.84

[1]Frequency ratio relative to random = Sig. gene pair frequency (%) / 1.08 (%).

random: 무작위 추출한 유전자 대상.

pathway: 동일 pathway 에 속한 유전자 대상.

paired reaction: paired reaction 에 속한 유전자 대상.

intra reaction: intra reaction 에 속한 유전자 대상.

subtype2: binary relation 중 subtype2 에 속한 유전자 대상.

subtype3: binary relation 중 subtype3 에 속한 유전자 대상.

subtype4: binary relation 중 subtype4 에 속한 유전자 대상.

subtype5: binary relation 중 subtype5 에 속한 유전자 대상.

subtype6: binary relation 중 subtype6 에 속한 유전자 대상.

5. 유전자 연관 관계 네트워크 구축 및 pathway 매핑

RNA expression 데이터가 있는 유전자들을 대상으로 SNP pair 간에 genotype 연관 관계가 있으면서 유전자 쌍 간에도 haplotype 연관 관계가 있는 유전자들 중 하나라도 pathway 정보를 가지고 있는 유전자 쌍은 360 개 였으며, 쌍을 구성하는 유전자는 모두 456 개 였다. 456 개의 유전자로 105 개의 네트워크를 구축할 수 있었고, 3개 이상의 유전자로 구성된 네트워크는 37개 였으며, 이들 중 98개의

유전자로 구성된 것이 가장 큰 네트워크 였다(표 8). 네트워크를 구성하는 각 유전자들은 최대 6개의 서로 다른 유전자들과 SNP genotype 및 haplotype 연관 관계를 맺고 있었으며, 2개 이상의 다른 유전자와 연관 관계를 맺고 있는 유전자는 161개 였다(표 9).

표 8. 유전자 연관 관계 네트워크 종류

	Constructed network													Total
number of genes in network[1]	2	3	4	5	6	7	8	9	10	11	21	22	98	456
network count[2]	68	12	6	5	2	1	3	1	2	2	1	1	1	105

[1]네트워크를 구성하는 유전자 수.

[2]구축된 네트워크 수.

표 9. 네트워크에서 연관 관계로 직접 연결된 유전자 수

	Number of genes directly connected in networks						
connected gene count[1]	1	2	3	4	5	6	Total
total gene count[2]	295	99	39	14	7	2	456

[1]하나의 유전자에 직접 연결되어 있는 최대 유전자 수.

[2]네트워크에서 해당 연결 관계를 가지는 총 유전자 수.

98개의 유전자로 구성된 네트워크(그림 7) 에서 타원형은 유전자를, 타원형 내의 숫자는 해당 유전자의 NCBI gene ID 를 나타내고 있다. 유전자간 연결되어 있는 선은 두 유전자 간에 SNP genotype 및 haplotype 연관 관계가 있음을 나타내고 있다. 유전자를 함께 둘러싸고 있는 선은 해당 유전자들이 포함되어 있는 pathway 를 나타내고 있다. GSK3B 유전자는 3번 염색체에 위치한 유전자로 3, 5, 6, 12번 염색체에 위치한 6개의 다른 유전자와 SNP genotype 및 haplotype 연관 관계를 가지고 있었으며, colorectal cancer pathway 및 Wnt signaling pathway 에 포함되어 있는 유전자였다(표 10).

MAPK8 유전자는 10번 염색체에 위치한 유전자로 2, 3, 4, 10, 11, 17번 및 X 염색체에 위치한 6개의 다른 유전자와 SNP genotype 및 haplotype 연관 관계를 가지고 있었으며, MAPK signaling pathway 및 Toll-like receptor signaling pathway 에 포함되어 있는 유전자였다(표 11).

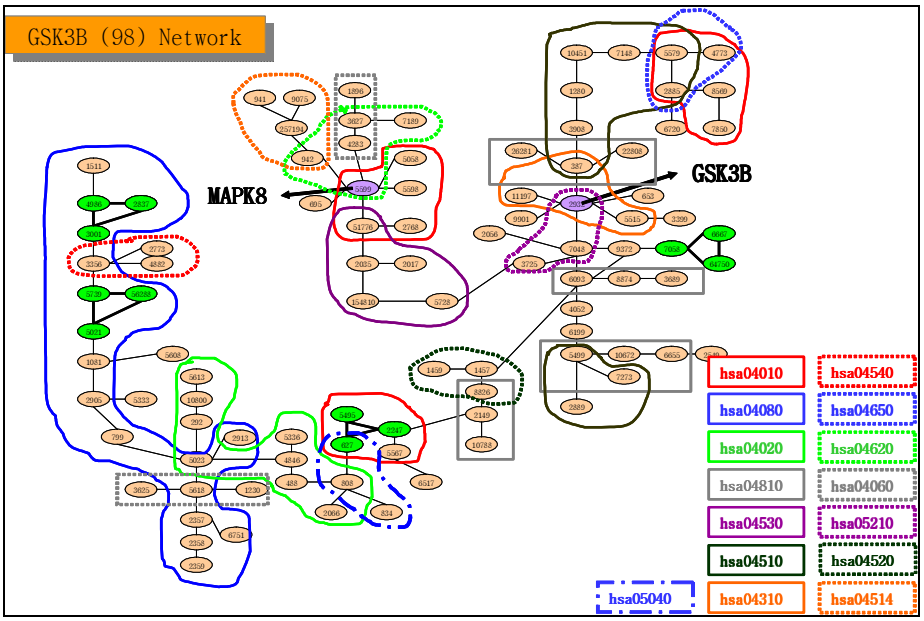

그림 7. GSK3B 유전자를 중심으로 98개의 유전자가 연결되어 있는 유전자 연관 관계 네트워크. 오른쪽 하단의 사각형 내에 표시된 것은 KEGG 의 pathway ID 이다. 네트워크에서 동일한 색깔과 동일한 모양의 선으로 둘러쳐져 있는 유전자들은 해당 pathway 에 공통적으로 포함되어 있는 유전자들을 의미한다. 녹색으로 표시한 것은 세 개 이상의 유전자들 간에 모두 SNP genotype 및 haplotype 연관 관계가 있음을 나타낸다. 네트워크의 중심이 되는 유전자인 MAPK8 유전자와 GSK3B 유전자는 각각 6개의 서로 다른 유전자와 직접 연결되어 있으며 보라색으로 표시하였다.

표 10. 네트워크에서 GSK3B 유전자와 직접 연결되어 있는 유전자

No	NCBI gene ID	Gene symbol	Chr no	Pathway
1	2932	GSK3B	3	Colorectal cancer(hsa05210) Wnt signaling pathway(hsa04310)
2	387	RHOA	3	Wnt signaling pathway(hsa04310) Regulation of actin cytoskeleton(hsa04810) Focal adhesion(hsa04510)
3	653	BMP5	6	-
4	5515	PPP2CA	5	Wnt signaling pathway(hsa04310)
5	7048	TGFBR2	3	Colorectal cancer(hsa05210)
6	9901	SRGAP3	3	-
7	11197	WIF1	12	Wnt signaling pathway(hsa04310)

표 11. 네트워크에서 MAPK8 유전자와 직접 연결되어 있는 유전자

No	NCBI gene ID	Gene symbol	Chr no	Pathway
1	5599	MAPK8	10	Toll-like receptor signaling pathway(hsa04620) MAPK signaling pathway(hsa04010)
2	695	BTK	X	-
3	942	CD86	3	Cell adhesion molecules (CAMs)(hsa04514) Toll-like receptor signaling pathway(hsa04620)
4	4283	CXCL9	4	Toll-like receptor signaling pathway(hsa04620) Cytokine-cytokine receptor interaction(hsa04060)
5	5058	PAK1	11	MAPK signaling pathway(hsa04010)
6	5598	MAPK7	17	MAPK signaling pathway(hsa04010)
7	51776	ZAK	2	MAPK signaling pathway(hsa04010) Tight junction(hsa04530)

8개의 유전자로 구성된 네트워크에서는(그림 8) ADH5 유전자를 중심으로 하여 ADH1C, MTHFR, LDHB, GSTM4 유전자가 연결되어 있었다. ADH5, ADH1C, ADH6 유전자는 1-and 2-methylnaphthalene degradation pathway, bile acid biosynthesis pathway, fatty acid metabolism pathway, glycerolipid metabolism pathway, tyrosine metabolism pathway 에 포함되어 있는 유전자들로, 모두 4번 염색체에 위치해 있다. GSTM3, GSTM4, GSTM5 유전자는 metabolism of xenobiotics by cytochrome P450 pathway 에 포함되어 있는 유전자로 1번 염색체에

위치해 있다(표 12).

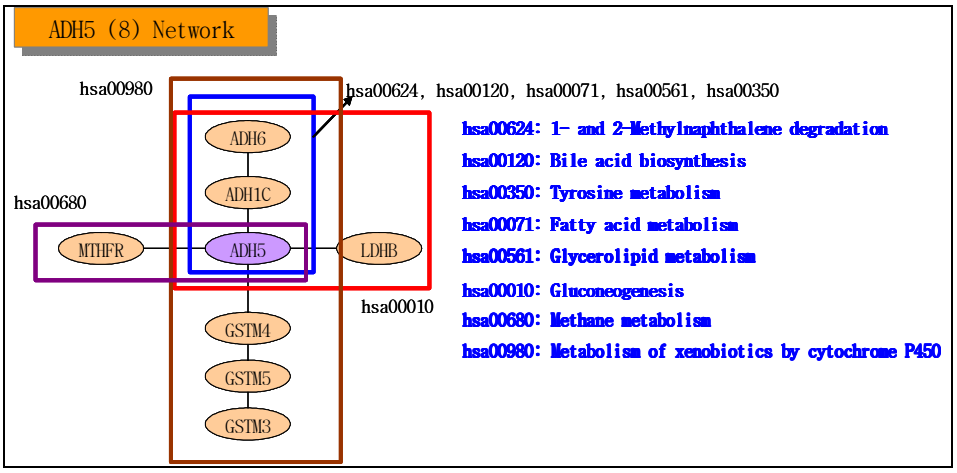

그림 8. ADH5 유전자를 중심으로 8개의 유전자가 연결되어 있는 유전자 연관 관계 네트워크. 중심이 되는 유전자는 ADH5 로 보라색으로 나타내었다. 동일한 색깔과 동일한 모양의 선으로 둘러쳐져 있는 유전자들은 해당 pathway 에 포함되어 있는 유전자들을 의미한다.

표 12. ADH5 유전자를 중심으로 구축된 네트워크를 구성하는 유전자

No	NCBI gene ID	Gene symbol	Chr no
1	128	ADH5	4
2	126	ADH1C	4
3	130	ADH6	4
4	2947	GSTM3	1
5	2948	GSTM4	1
6	2949	GSTM5	1
7	4524	MTHFR	1
8	3945	LDHB	12

6. RNA expression 데이터 분석

가. 유전자들간 발현 상관 관계 분석

RNA expression 데이터가 있는 8,836 개 유전자 중 SNP genotype 및 haplotype 연관 관계 분석 결과가 있는 유전자들을 대상으로 발현 상관 관계 분석을 수행하였다(표 13).

표 13. 분석 대상 유전자 수

Category		Gene count
random		4,493
pathway		1,415
reactions and pathway subtypes	paired reaction	213
	intra reaction	350
	subtype2	474
	subtype3	119
	subtype5	12
	subtype6	604

random: 무작위 추출한 경우.
pathway: 동일 pathway 에 속한 경우.
paired reaction: paired reaction 에 속한 경우.
intra reaction: intra reaction 에 속한 경우.
subtype2: binary relation 중 subtype2 에 속한 경우.
subtype3: binary relation 중 subtype3 에 속한 경우.
subtype5: binary relation 중 subtype5 에 속한 경우.
subtype6: binary relation 중 subtype6 에 속한 경우.

무작위로 추출한 4,493 개 유전자들을 대상으로 9,832 번의 분석 결과 발현 상관 관계에 있는 p-value 개수의 비율은 정상 대장 조직

에서는 32.7%, 대장암 조직 에서는 21.6% 로 나타났다. Pathway 에 속한 1,415 개 유전자 대상으로는 45,130 번의 분석을 수행하였으며, 이들 중 발현 상관 관계를 보이는 p-value 개수의 비율은 정상 대장 조직에서는 35.4%, 대장암 조직 에서는 22.4% 로 나타났다. 발현 상관 관계를 보이는 p-value 개수의 비율을 무작위 추출 결과와 비율 차이 검정을 했을 때 정상 대장 조직에서는 통계적으로 차이가 났으나 대장암 조직에서는 통계적인 차이가 없었다(그림 9-A).

한편, pathway 를 구성하는 intra reaction, paired reaction, 그리고 binary relation 내의 subtype2, subtype3, subtype5, subtype6 의 경우 정상 대장 조직 에서 발현 상관 관계를 보이는 p-value 개수의 비율을 무작위 추출 결과와 비율 차이 검정을 했을 때, subtype2 와 subtype6 를 제외하고 모두 통계적으로 차이가 났다. 한편, 대장암 조직은 발현 상관 관계를 보이는 p-value 개수의 비율이 무작위 추출 결과와 비율 차이 검정을 했을 때 subtype5 와 subtype6 에서만 통계적인 차이가 났다(그림 9-B).

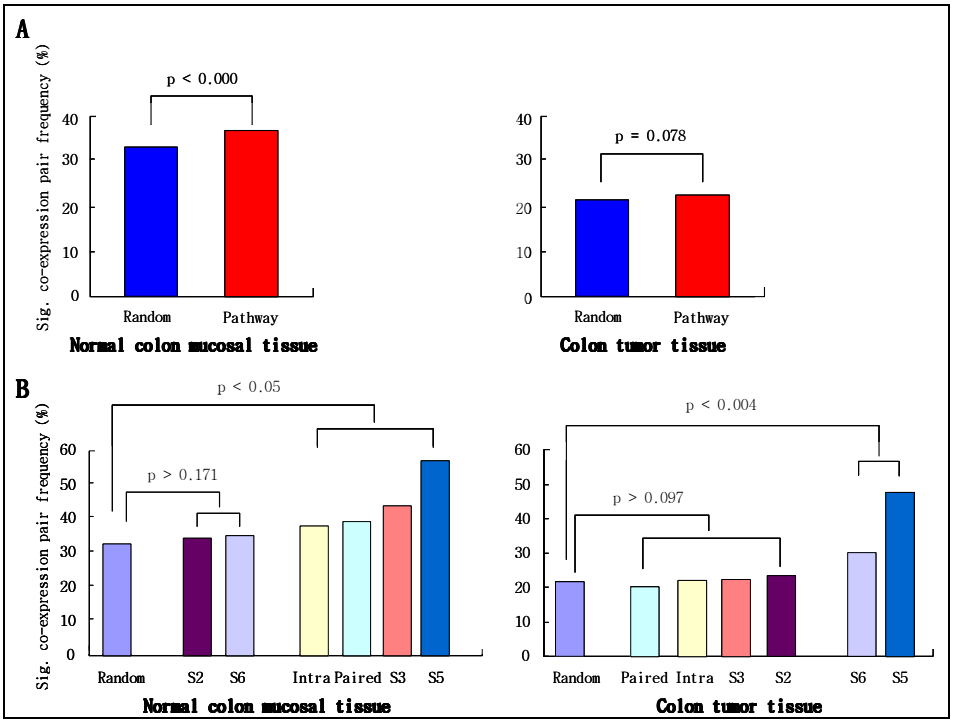

그림 9. 발현 상관 관계를 보이는 p-value 개수의 비율 차이 검정.

(A) 무작위 추출과 pathway 와의 비교, (B) 무작위 추출과 intra reaction, paired reaction, binary relation 과의 비교.

Random: 무작위 추출한 유전자 대상.

Pathway: 동일 pathway 에 속한 유전자 대상.

Paired: paired reaction 에 속한 유전자 대상.

Intra: intra reaction 에 속한 유전자 대상.

S2: binary relation 중 subtype2 에 속한 유전자 대상.

S3: binary relation 중 subtype3 에 속한 유전자 대상.

S5: binary relation 중 subtype5 에 속한 유전자 대상.

S6: binary relation 중 subtype6 에 속한 유전자 대상.

표 14 와 표 15 는 동일 pathway 에 속해있는 두 개 유전자가 발현 상관 관계를 나타내는 비율이 무작위로 추출한 경우에 비해 얼마만큼 증가하는지를 나타내고 있다. 정상 대장 조직에서는, 두

유전자가 발현 상관 관계에 있을 때 subtype5 에 속할 확률은 두 유전자들 간의 발현 상관 관계를 모를 경우에 비해 1.748 배 높아지는 것으로 나타났다. 대장암 조직에서는 2.203배 높아지는 것으로 나타났다.

표 14. 정상 대장 조직에서 발현 상관 관계를 가지는 유전자 쌍 개수 비율

Category		Sig. correlated expression pair frequency (%)[1]	Frequency ratio relative to random[2]
random		32.7	1.000
pathway		35.4	1.083
reactions and pathway subtypes	paired reaction	38.0	1.165
	intra reaction	39.4	1.205
	subtype2	34.4	1.052
	subtype3	43.9	1.344
	subtype5	57.1	1.748
	subtype6	35.1	1.073

[1]발현 상관 관계에 있는 유전자 쌍 개수 비율.

[2]Frequency ratio relative to random = Sig. correlated expression pair frequency (%) / 32.7 (%).

random: 무작위 추출한 유전자 대상.

pathway: 동일 pathway 에 속한 유전자 대상.

paired reaction: paired reaction 에 속한 유전자 대상.

intra reaction: intra reaction 에 속한 유전자 대상.

subtype2: binary relation 중 subtype2 에 속한 유전자 대상.

subtype3: binary relation 중 subtype3 에 속한 유전자 대상.

subtype5: binary relation 중 subtype5 에 속한 유전자 대상.

subtype6: binary relation 중 subtype6 에 속한 유전자 대상.

표 15. 대장암 조직에서 발현 상관 관계를 가지는 유전자 쌍 개수 비율

Category		Sig. correlated expression pair frequency (%)[1]	Frequency ratio relative to random[2]
random		21.6	1.000
pathway		22.4	1.037
reactions and pathway subtypes	paired reaction	22.2	1.207
	intra reaction	20.2	0.933
	subtype2	23.4	1.083
	subtype3	22.4	1.037
	subtype5	47.6	2.203
	subtype6	30.1	1.394

[1]발현 상관 관계에 있는 유전자 쌍 개수 비율.

[2]Frequency ratio relative to random = Sig. correlated expression pair frequency (%) / 21.6 (%).

random: 무작위 추출한 유전자 대상.

pathway: 동일 pathway 에 속한 유전자 대상.

paired reaction: paired reaction 에 속한 유전자 대상.

intra reaction: intra reaction 에 속한 유전자 대상.

subtype2: binary relation 중 subtype2 에 속한 유전자 대상.

subtype3: binary relation 중 subtype3 에 속한 유전자 대상.

subtype5: binary relation 중 subtype5 에 속한 유전자 대상.

subtype6: binary relation 중 subtype6 에 속한 유전자 대상.

나. 대장암 특이 발현 유전자 분석 및 발현 상관 관계 분석

유전자 연관 관계 네트워크를 구성하는 456 개 유전자들을 대상으로 대장암에서 특이적으로 발현하는 유전자 분석 및 유전자 발현 상관 관계 분석을 하였다. 전체 360개 유전자 쌍 중 정상 대장 조직과 대장암 조직 에서 모두 발현 상관 관계를 보이는 유전자 쌍은 8 개 였다(표 16). 쌍을 이루는 두 개 유전자가 모두 과발현(up-regulation) 되는 것은 1 개 였으며, 한 개 유전자만

하향조절(down-regulation) 되는 것은 6개, 그리고 한 개 유전자만 과발현(up-regulation) 되는 것은 1 개 였다(표 17).

표 16. 대장암 특이 발현 유전자 쌍 및 유전자 발현 상관 관계

	Co-expr.type1	Co-expr.type2	Co-expr.type3	Co-expr.type4	Total
Diff.type1	166	41	79	25	311
Diff.type2	0	0	0	0	0
Diff.type3	0	1*	0	1	2
Diff.type4	0	0	0	0	0
Diff.type5	9	6**	7	4	26
Diff.type6	16	1***	3	1	21
Total	191	49	89	31	360

Diff.type1: 두 유전자 모두 정상 대장 조직 대비 대장암 조직에서 발현 변화 없음.

Diff.type2: 두 유전자 모두 정상 대장 조직 대비 대장암 조직에서 하향조절.

Diff.type3: 두 유전자 모두 정상 대장 조직 대비 대장암 조직에서 과발현.

Diff.type4: 대장암 조직에서 한 유전자는 과발현, 다른 유전자는 하향조절.

Diff.type5: 한 유전자만 대장암 조직에서 하향조절.

Diff.type6: 한 유전자만 대장암 조직에서 과발현.

Co-expr.type1: 두 유전자가 정상 대장 및 대장암 조직에서 발현 상관 관계 없음.

Co-expr.type2: 두 유전자가 정상 대장 및 대장암 조직에서 발현 상관 관계 있음.

Co-expr.type3: 두 유전자가 정상 대장 조직에서만 발현 상관 관계 있음.

Co-expr.type4: 두 유전자가 대장암 조직에서만 발현 상관 관계 있음.

*서로 발현 상관 관계가 있으며 두 유전자 모두 정상 대장 조직 대비 대장암 조직에서 과발현 되는 유전자 쌍은 1개임.

**서로 발현 상관 관계가 있지만 한 유전자만 대장암 조직에서 하향조절 되고 나머지 유전자는 발현량 변화 없는 유전자 쌍은 6개임.

***서로 발현 상관 관계가 있지만 한 유전자만 대장암 조직에서 과발현 되고 나머지 유전자는 발현량 변화 없는 유전자 쌍은 1개임.

표 17. 대장암 특이 발현 및 발현 상관 관계 유전자 쌍

No	Gene ID[1]	Gene symbol	Common pathway[2]	Expression pattern
1	4283	CXCL9	Cytokine-cytokine receptor interaction(hsa04060) Toll-like receptor signaling pathway(hsa04620) Neuroactive ligand-receptor interaction(hsa04080)	Both up-regulated
	3627	CXCL10		
2	2359	FPRL2	Gluconeogenesis(hsa00010) Fatty acid metabolism(hsa00071) Bile acid biosynthesis(hsa00120) Tyrosine metabolism(hsa00350) Glycerolipid metabolism(hsa00561) 1- and 2-Methylnaphthalene degradation(hsa00624) 3-Chloroacrylic acid degradation(hsa00641) Metabolism of xenobiotics by cytochrome P450(hsa00980) Androgen and estrogen metabolism(hsa00150)	FPRL2: no-change FPRL1: down-regulated
	2358	FPRL1		
3	126	ADH1C	Neuroactive ligand-receptor interaction(hsa04080) N-Glycan biosynthesis(hsa00510)	ADH1C: down-regulated ADH6: no-change
	130	ADH6		
4	54658	UGT1A1	Olfactory transduction(hsa04740) Complement and coagulation cascades (hsa04610)	UGT1A1: down-regulated HSD17B2: no-change
	3294	HSD17B2		
5	2357	FPR1	Common pathway Cytokine-cytokine receptor interaction(hsa04060) Toll-like receptor signaling pathway(hsa04620)	FPR1: no-change FPRL1: down-regulated
	2358	FPRL1		
6	79053	ALG8	Neuroactive ligand-receptor interaction(hsa04080) Gluconeogenesis(hsa00010) Fatty acid metabolism(hsa00071) Bile acid biosynthesis(hsa00120) Tyrosine metabolism(hsa00350) Glycerolipid metabolism(hsa00561) 1- and 2-Methylnaphthalene degradation(hsa00624) 3-Chloroacrylic acid degradation(hsa00641) Metabolism of xenobiotics by cytochrome P450(hsa00980)	ALG8: down-regulated MAN2A1: no-change
	4124	MAN2A1		
7	9635	CLCA2	Androgen and estrogen metabolism(hsa00150) Neuroactive ligand-receptor interaction(hsa04080)	CLCA2: no-change CLCA1: down-regulated
	1179	CLCA1		
8	722	C4BPA	N-Glycan biosynthesis(hsa00510)	C3BPA: no-change C4BPB: up-regulated
	725	C4BPB		

[1]Gene ID: NCBI gene ID.

[2]Common pathway: 쌍을 이루는 두 유전자가 모두 포함되어 있는 pathway.

Ⅳ. 고 찰

SNP 들간의 연관 관계는 주로 연관 불평형 분석을 통해 이루어지는데, 유전자 내부 또는 동일 염색체 내에서 가까이 위치한 SNP 들간에는 서로 멀리 떨어진 SNP 들간에 비해 감수분열(meiosis) 단계에서 재조합(recombination) 될 확률이 작기 때문에 강한 연관 불평형을 나타낸다. 유전자의 5' upstream 영역에서 발견되는 SNP 들이 강한 연관 불평형 관계에 있음을 밝힌 연구[24,25], cis 형태로 유전자 발현을 조절하는 SNP 들간에 강한 연관 불평형 관계에 있음을 연구한 사례[18] 등이 있다. SNP 들간의 연관 불평형은 두 SNP 에서 관측되는 haplotype 빈도를 이용하여 계산하는데[26], SNP 가 위치한 loci 들간의 재조합 빈도(recombination rate) 에 따라 연관 불평형 정도가 달라진다. 하지만, 서로 다른 염색체 간에는 재조합이라는 현상이 일어날 수 없기 때문에 연관 불평형 분석은 서로 다른 염색체에 위치한 SNP 들간의 연관 관계 분석에는 이용될 수 없다. 따라서, 본 연구에서는 SNP 에서 관측되는 genotype 패턴을 이용하여 서로 다른 염색체에 위치한 SNP 들간의 연관 관계 분석을 하였다.

SNP 와 같은 유전학적인 근거에서 유전자 발현을 연구하는 방법을 genetical genomics 라고 지칭하는데[16], RNA 발현을 조절하는 유전자 전사 조절 영역에서 발견되는 SNP 의 allele frequency 에 따라 유전자 발현 양상이 달라지거나[19,27], 유전자의 발현 정도를 유의적으로 조절하는 데 관여하는 잠재적인 cis-acting 또는 trans-acting SNP 을 찾는 연구[27-30] 등이 대표적이다. 유전자 전사 조절 영역은 유전자에 따라 전사 시작 지점(transcription start site) 에서부터 시작하여 수백 base pair 에서 수천 base pair 앞까지로 정해질 수 있다[31]. 본 연구에서는 인간 유전체를 구성하는

모든 유전자를 분석 대상으로 하기 때문에, 개개 유전자마다 서로 다른 전사 조절 후보 영역을 설정하는 대신 모든 유전자에 동일하게 유전자 전사 시작 위치로부터 5,000 base pair 앞까지를 전사 조절 후보 영역으로 정하였고, 이 영역에서 발견되는 SNP 들을 이용하여 연관 관계 분석을 하였다.

생식 세포가 1번 감수분열 할 때 1% 의 확률로 재조합이 일어날 수 있는 유전적 거리(genetic distance) 는 대략 1,000,000 base pair 로 알려져 있다. 따라서, 5,000 base pair 는 한번 감수분열 시 0.005% 의 확률로 재조합이 일어난다고 할 수 있으므로 유전자 전사 조절 후보 영역에 속한 SNP 들은 강한 연관 불평형 관계에 있을 가능성이 높다고 할 수 있다. 본 연구 결과에서는 유전자 전사 조절 후보 영역의 SNP 들이 강한 연관 불평형 관계에 있음을 확인 할 수 있었다. 그리고 전사 조절 후보영역 내의 SNP 들 간에 genotype 연관 관계가 있을 확률(59.71%)이 무작위 추출한 SNP 들 간에 genotype 연간 관계가 있을 확률(1.80%) 보다 월등히 높았다.

Pathway 를 구성하는 유전자들은 단백질로 발현되었을 때 동일 화학 반응에 함께 참여하는 경우(intra reaction), 동일 compound 를 기질(substrate) 과 산물(product) 형태로 공유하는 경우(paired reaction), 두 개의 단백질이 서로 binding 하는 경우(subtype6), 한쪽 유전자가 다른 유전자를 활성화 하거나 억제하는 경우(subtype 2), 그리고 DNA 에 결합되어 유전자 발현을 제어하는 경우(subtype3) 등 다양한 형태의 반응에 참여하고 있다. 특히, 동일 pathway 내에서 서로 직접적인 반응 관계에 있는 유전자들은 genetical genomics 와 관련된 최근 연구 결과들[15-20, 27-30] 을 살펴 볼 때 유전자 전사 조절 영역의 SNP 들간에 유전학적인 연관 관계가 있을 수 있음을 유추해 볼 수 있다. 본 연구 결과에서는 202 개의 pathway 를 구성하는

유전자들 중, 발현되는 단백질들 간에 직접적인 반응 관계가 있는 유전자들에서 발견되는 SNP 들 간에 genotype 연관 관계가 있을 확률이 무작위로 추출된 SNP 들에 비해 상대적으로 높았다. 즉, SNP 들 간에 genotype 연관 관계가 있을 확률이 무작위 추출인 경우 1.80% 인데 비해 DNA 결합을 통해 유전자 발현을 제어하는 관계(subtype3) 에 있는 유전자들의 경우에는 10.36% 였다. 또한 state transition (subtype5) 을 구성하는 유전자들 간에는 11.67% 의 확률로 SNP 들이 genotype 연관 관계가 있었다. 결국, pathway 를 구성하는 다양한 형태의 반응에 관여하는 유전자들에서 발견되는 SNP 들간에는 genotype 으로 대표되는 유전학적인 연관 관계가 높은 확률로 존재할 수 있음을 나타낸다.

한편, 전사 조절 후보 영역에서 발견되는 SNP 수는 유전자들마다 다르기 때문에 SNP 들만을 이용하여 분석할 경우 특정 유전자에 의해 편중된 결과가 나올 가능성이 있다. 따라서, 전사 조절 후보 영역에서 발견되는 두 개 이상의 SNP 들 중 강한 연관 불평형 관계에 있는 것들만을 이용하여 추정한 haplotype 을 이용하여 유전자들 간의 연관 관계 분석을 다시 수행하였다. Haplotype 을 이용한 연관 관계 분석은 특정 유전자에 의한 편중된 분석 결과를 가져오지 않는다는 장점이 있지만, 구성된 여러 haplotype 들 중 전사 시작 위치에서 가장 가까우면서 최빈도 값을 가지는 haplotype 을 해당 유전자를 대표하는 것으로 설정하였기 때문에 SNP 를 이용한 연관 관계 분석에 비해 분석의 민감도가 떨어질 수 있다는 단점이 존재한다. 연구 결과, SNP 를 이용한 분석과는 달리 KEGG pathway 의 binary relation 중 indirect effect (subtype4) 를 구성하는 유전자들간, 그리고 paired reaction 을 구성하는 유전자들 간에 haplotype 연관 관계가 있을 확률이 무작위 추출한 경우와 비교했을

때 통계적인 차이가 없었다. 한편, pathway 를 구성하는 intra reaction, 그리고 binary relation 의 나머지 subtype (subtype2, subtype3, subtype5, subtype6) 들에서는 SNP 를 이용한 분석과 마찬가지로 무작위 추출 결과와 통계적인 차이가 났다. 특히, DNA 결합을 통해 유전자 발현을 제어하는 관계(subtype3) 에 있는 유전자들 간에는 haplotype 연관 관계가 있을 확률이 9.54% 로 나타났다. State transition (subtype5) 을 구성하는 유전자들 간에 haplotype 연관 관계가 있을 확률은 5.56% 였다. SNP genotype 및 haplotype 연관 관계 분석 결과를 Bayesian rule 을 이용해 해석하면, pathway 를 구성하는 유전자의 경우 전사 조절 후보 영역에서 발견되는 SNP genotype, 그리고 haplotype 은 무작위적으로 나타나는 것이 아니라 통계적으로 유의미한 수준에서(α=0.01) 유전학적인 연관 관계를 가지고 있다고 할 수 있다.

유전자 전사 조절 영역 중 promoter region 은 생명체 진화 과정에서 염기서열이 잘 보존된 영역(conserved region) 이다. 따라서, promoter region 에서 나타나는 SNP 들간에 genotype 연관 관계가 있는 유전자들은 진화적인 측면에서 연관 관계가 있다고 유추해 볼 수 있다. Haplotype 연관 관계가 있는 경우에는, 연관 관계에 있는 유전자들의 진화 과정에서 여러 개 SNP 가 비슷한 시기에 발생했다고 유추해 볼 수 있다.

SNP genotype 연관 관계, 그리고 haplotype 연관 관계가 있는 유전자 456개를 이용하여 구성한 유전자 연관 관계 네트워크에서는 최대 6개의 유전자와 직접적인 연관 관계를 가지는 GSK3B 및 MAPK8 유전자를 발견할 수 있었다. GSK3B 유전자는 Wnt signaling pathway 와 colorectal cancer pathway 에 참여하고 있는 유전자로, 특히 대장암에서 2-fold 이상 과발현 되는 것으로 알려져 있으며[32],

대장암에서 tumor suppressor gene 으로 알려져 있는 TGFBR2 유전자[33]와 유전자 연관 관계 네트워크 상에서 직접 연결되어 있었다. 배발생(embryogenesis) 동안 진화적으로 보존된(evolutionarily conserved) 역할을 하는 Wnt signaling pathway 의 경우[34] GSK3B 유전자와 함께 RHOA, PPP2CA, WIF1 유전자가 네트워크 상에서 직접 연결되어 있었으며, 이들 유전자들은 모두 서로 다른 염색체에 분포하고 있는 유전자들 이었다. MAPK signaling pathway 의 경우 MAPK8 유전자에 PAK1, MAPK7, ZAK 유전자가 직접 연결되어 있었다. Toll-like receptor signaling pathway 에는 MAPK8 유전자에 CD86, CXCL9 유전자가 직접 연결되어 있었다. 특히, CXCL9 유전자는 CXCL10 유전자와 함께 CXC chemokine family 를 코딩하는 유전자로, 본 연구에서는 SNP genotype 및 haplotype 연관 관계가 있음을 확인할 수 있었다. 또한, CXCL9 과 CXCL10 유전자는 서로 발현 상관 관계에 있으면서 대장암 조직에서 2-fold 이상 모두 과발현 되는 것으로 나타났는데, 유전자 전사 조절 후보 영역의 염기서열 변이의 연관성이 두 유전자의 발현 상관성에 영향을 주는 것으로 추정해 볼 수 있다. SNP genotype 및 haplotype 으로 대표되는 염기서열 변이 연관 관계가 있으면서 유전자 발현 상관 관계가 있는 다른 예로는, formyl- peptide receptor (FPR) family 를 코딩하는 FPRL1 과 FPRL2, 그리고 FPR1 과 FPRL1 등이 있었다. 한편, ADH1C 와 ADH6 는 ethanol, retinol, aliphatic alcohols, hydroxysteroids, lipid peroxidation 과 같은 다양한 물질 대사에 관여하고 있는 alcohol dehydrogenase family 를 코딩하는 유전자들로, 본 연구에서는 SNP genotype 연관 관계, haplotype 연관 관계 및 유전자 발현 상관 관계에 있었다.

유전자 연관 관계 네트워크에서 서로 연관되어 있는 유전자들이 대장암 조직에서 모두 과발현 되거나 모두 하향조절 된다는 것은,

pathway 에서 이들 유전자들 또는 유전자들이 코딩하는 단백질들이 binding 과 같은 직접적인 연결 관계에 있거나 동일 수용체(receptor) 에 결합하는 단백질 군(protein family) 을 구성하는 경우라고 유추해 볼 수 있다. 본 연구에서는 CXCL9 과 CXCL10 유전자가 수용체 CXCR3 와 결합하는 동일 단백질 군에 속하는 경우였다. 한편, 유전자 연관 관계 네트워크 상에서 oncogene 과 tumor suppressor gene 이 서로 연관되어 있는 유전자들이 있었는데, GSK3B 유전자와 TGFBR2 가 대표적이었다. 이들은 colorectal cancer pathway 에 관여하는 유전자들로 pathway 상에서는 직접적인 연결 관계를 가지지는 않았지만 유전학적인 측면에서는 밀접한 연관 관계를 가진다고 유추해 볼 수 있다.

본 연구 결과, pathway 를 구성하는 유전자들 중 서로 직접적인 반응 관계에 있는 유전자들은 전사 조절 후보 영역에서 발견되는 염기서열 변이인 SNP genotype 들간, 그리고 haplotype 들간에 연관 관계가 있을 확률이 pathway 정보를 모르는 무작위 추출 유전자들에 비해서 최대 8.83 배 높았다(subtype3 의 경우). 유전자들 간의 염기서열 변이 연관 관계를 이용하여 최대 98개의 유전자로 이루어진 유전자 연관 관계 네트워크를 구축하였으며, 유전자들간의 발현 상관 관계 및 대장암 조직에서 특이 발현(differential expression) 하는 유전자들을 분석한 후 구축된 네트워크에 매핑 해본 결과, 대장암 조직 특이 발현 유전자들 중 일부가 서로 발현 상관 관계에 있으면서 염기서열 변이 연관 관계가 있음을 네트워크 상에서 확인할 수 있었다.

이러한 유전자 연관 관계 네트워크와 유전자들 간의 발현 상관 관계, 그리고 특이 발현 정보를 유기적으로 연결하면 향후 시스템 생물학 연구에 유용하게 이용될 수 있을 것으로 사료된다.

Ⅴ. 결 론

인간의 전체 유전자들을 대상으로 유전자의 전사 조절 후보 영역에서 발견되는 염기서열 변이 중 SNP genotype 및 SNP 들을 이용하여 추정된 haplotype 을 이용하여 유전자들간의 유전학적인 연관 관계를 분석한 결과 아래와 같은 결론을 얻었다.

1. Pathway 를 구성하는 유전자들 중 서로 직접적인 반응 관계에 있는 유전자들의 전사 조절 후보 영역의 SNP genotype 들은 무작위로 나타나지 않는다.

2. Pathway 를 구성하는 유전자들 중 서로 직접적인 반응 관계에 있는 유전자들의 전사 조절 후보 영역 내 haplotype 들은 무작위로 나타나지 않는다.

3. SNP genotype 들간, 그리고 haplotype 들간 연관 관계가 있는 유전자들을 대상으로 유전자 연관 관계 네트워크를 구축할 수 있으며, 구축한 네트워크 상에 pathway 를 매핑시켜 보면 다수의 유전자들과 연관 관계를 맺고 있는 중심 유전자 및 서로 다른 pathway 를 연결시켜 주는 유전자 군을 찾을 수 있다.

4. 유의미한 발현 상관 관계를 보이는 유전자들 중 일부는 전사 조절 영역의 SNP genotype 및 haplotype 들이 무작위로 나타나지 않는다.

참고문헌

1. Collins FS, Brooks LD, Chakravarti A. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 1998;8:1229-31.

2. Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, et al. Large-scale genotyping of complex DNA. Nat Biotechnol 2003;21:1233-7.

3. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science 2005;308:385-9.

4. Vella A, Cooper JD, Lowe CE, Walker N, Nutland S, Widmer B, et al. Localization of a type 1 diabetes locus in the IL2RA/CD25 region by use of tag single-nucleotide polymorphisms. Am J Hum Genet 2005;76:773-9.

5. Wu X, Gu J, Grossman HB, Amos CI, Etzel C, Huang M, et al. Bladder cancer predisposition: a multigenic approach to DNA-repair and cell-cycle-control genes. Am J Hum Genet 2006;78:464-79.

6. Stacey SN, Manolescu A, Sulem P, Rafnar T, Gudmundsson J, Gudjonsson SA, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 2007;39:865-9.

7. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature 2007;448:470-3.

8. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z,

Spain S, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet 2007;39:984-8.
9. Deng X, Shibata H, Ninomiya H, Tashiro N, Iwata N, Ozaki N, et al. Association study of polymorphisms in the excitatory amino acid transporter 2 gene (SLC1A2) with schizophrenia. BMC Psychiatry 2004;4:21.
10. Tamimi RM, Hankinson SE, Spiegelman D, Kraft P, Colditz GA, Hunter DJ. Common ataxia telangiectasia mutated haplotypes and risk of breast cancer: a nested case-control study. Breast Cancer Res 2004;6:R416-22.
11. Bonnen PE, Wang PJ, Kimmel M, Chakraborty R, Nelson DL. Haplotype and linkage disequilibrium architecture for human cancer-associated genes. Genome Res 2002;12:1846-53.
12. Kamatani N, Sekine A, Kitamoto T, Iida A, Saito S, Kogame A, et al. Large-scale single-nucleotide polymorphism (SNP) and haplotype analyses, using dense SNP maps, of 199 drug-related genes in 752 subjects: The analysis of the association between uncommon SNPs within haplotype blocks and the haplotypes constructed with haplotype-tagging SNPs. Am J Hum Genet 2004;75:190-203.
13. Putt W, Palmen J, Nicaud V, Tregouet DA, Tahri-Daizadeh N, Flavell DM, et al. Variation in USF1 shows haplotype effects, gene : gene and gene : environment associations with glucose and lipid parameters in the European Atherosclerosis Research Study II. Hum Mol Genet 2004;13:1587-97.
14. Tan Q, De Benedictis G, Ukraintseva SV, Franceschi C, Vaupel JW,

Yashin AI. A centenarian-only approach for assessing gene-gene interaction in human longevity. Eur J Hum Genet 2002;10:119-24.
15. Jansen RC, Nap JP. Genetical genomics: the added value from segregation. Trends Genet 2001;17:388-91.
16. Li J, Burmeister M. Genetical genomics: combining genetics with gene expression analysis. Hum Mol Genet 2005;14 Spec No. 2:R163-9.
17. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, et al. Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. Nat Genet 2005;37:225-32.
18. Kristensen VN, Edvardsen H, Tsalenko A, Nordgard SH, Sorlie T, Sharan R, et al. Genetic variation in putative regulatory loci controlling gene expression in breast cancer. Proc Natl Acad Sci U S A 2006;103:7735-40.
19. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, et al. Genetic analysis of genome-wide variation in human gene expression. Nature 2004;430:743-7.
20. Yang HH, Hu Y, Buetow KH, Lee MP. A computational approach to measuring coherence of gene expression in pathways. Genomics 2004;84:211-7.
21. Kim KY, Ki DH, Jeong HJ, Jeung HC, Chung HC, Rha SY. Novel and simple transformation algorithm for combining microarray data sets. BMC Bioinformatics 2007;8:218.
22. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, et al. The structure of haplotype blocks in the human genome. Science 2002;296:2225-9.

23. Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 1995;12:921-7.
24. Woo JG, Dolan LM, Deka R, Kaushal RD, Shen Y, Pal P, et al. Interactions between noncontiguous haplotypes in the adiponectin gene ACDC are associated with plasma adiponectin. Diabetes 2006;55:523-9.
25. Fiumera AC, Dumont BL, Clark AG. Associations between sperm competition and natural variation in male reproductive genes on the third chromosome of Drosophila melanogaster. Genetics 2007;176:1245-60.
26. Devlin B, Risch N. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 1995;29:311-22.
27. Milani L, Gupta M, Andersen M, Dhar S, Fryknäs M, Isaksson A, et al. Allelic imbalance in gene expression as a guide to cis-acting regulatory single nucleotide polymorphisms in cancer cells. Nucleic Acids Res 2007;35:e34.
28. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT. Mapping determinants of human gene expression by regional and genome-wide association. Nature 2005;437:1365-9.
29. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, et al. Genetic inheritance of gene expression in human cell lines. Am J Hum Genet 2004;75:1094-105.
30. Stranger BE, Forrest MS, Clark AG, Minichiello MJ, Deutsch S, Lyle R, et al. Genome-wide associations of gene expression variation in humans. PLoS Genet 2005;1:e78.

31. Qiu P, Qin L, Sorrentino RP, Greene JR, Wang L, Partridge NC. Comparative promoter analysis and its application in analysis of PTH-regulated gene expression. J Mol Biol 2003;326:1327-36.
32. Grade M, Ghadimi BM, Varma S, Simon R, Wangsa D, Barenboim-Stapleton L, et al. Aneuploidy-dependent massive deregulation of the cellular transcriptome and apparent divergence of the Wnt/beta-catenin signaling pathway in human rectal carcinomas. Cancer Res 2006;66:267-82.
33. Xu Y, Pasche B. TGF-beta signaling alterations and susceptibility to colorectal cancer. Hum Mol Genet 2007;16 Spec No 1:R14-20.
34. Naito M, Katayama R, Ishioka T, Suga A, Takubo K, Nanjo M, et al. Cellular FLIP inhibits beta-catenin ubiquitylation and enhances Wnt signaling. Mol Cell Biol 2004;24:8418-27.

Abstract

Analysis of relationships among genetic variations using single nucleotide polymorphisms in the gene regulatory region and construction of relationship network of genetic variations

Jin Ho Yoo

Department of Medical Science
The Graduate School, Yonsei University

(Directed by Professor Sun Young Rha)

Many researches have been done to discover the genetic relationships among single nucleotide polymorphisms (SNPs) found on the human genes. However, there were common shortcomings to these studies for the network-based study because they have only focused on the several genes or nearby genetic loci on the same chromosome. Therefore, for comprehensive understanding of complex biology, it is required to take a genome-wide approach in addition to the gene-specific approach.

To elucidate the genome-wide connectivity of genes based on

the genetic variation, the relationships among genetic variations were investigated using SNPs and haplotypes found on the entire human genes. Also, for the integrated investigation of the connectivity of genes in regard to both the genetic variation and the pathway, the complex networks were constructed with the genes having relationship of genetic variation with other genes on the related pathways.

To find out the relationships among genetic variations, SNPs and their genotypes were gotten from the International HapMap project database and pathway information from the KEGG database. SNPs were confined within 5000 base pairs of the upstream region of genes and this region was designated as the candidate location of gene regulatory region. Seven subtypes were redefined according to the chemical reactions or binary relations in the pathways. Pearson's chi-square test was used to investigate whether there exist particular genotype relationships among SNPs. In addition to the genotype relationships, haplotype relationships were investigated using diplotype patterns. Using the genes of which SNP genotypes and haplotypes were both in statistically significant relationships, relationship networks of genetic variations were constructed, and the related pathways were investigated on the constructed networks. Finally, the expression patterns of RNA were investigated on the constructed

networks using normal colon mucosal tissue and colon tumor tissue.

As a result of the study, the probability that there are genotype relationships among SNPs and haplotype relationships among genes in the subtypes of the same pathway were remarkably higher than the relationships among randomly selected SNPs and genes. Relationship networks of genetic variations could be constructed using the genes that were in SNP genotype relationships and haplotype relationships. In the constructed networks, several key genes that were multiply connected with different genes located in different chromosomes were found, and a number of different pathways overlapped on the groups of the connected genes. And, some genes that had significantly correlated RNA expression level were found out to have SNP genotype and haplotype relationships in the constructed networks.

Therefore, it was concluded that the SNPs and haplotypes found in the regulatory regions of the genes that are directly connected with other genes via chemical reactions or binary relations in the same pathway may not occur randomly. And, relationship network of genetic variations could be constructed and used for the integrated investigation of the connectivity of genes in regard to both the genetic variation and the pathway. Finally, it was suggested that the key genes

and pathway-connecting genes could be found by mapping the pathway information to the relationship network of genetic variations.

Key Words : SNP, haplotype, diplotype, pathway, gene regulatory region, genotype relationship, haplotype relationship, relationship network of genetic variation