Enrichment Analysis of Gene Terms Related to Obesity and the Regression of Atherosclerosis

Adam Cankaya
COP 5859 Midterm

Part 1: Obesity

I. Background

Complex diseases such as obesity have a variety of risk factors including environment, patient behavior, and genetics. Recent GWAS have identified dozens of new gene locations associated with obesity, however even when taken together, “these loci explain only a small proportion of overall phenotypic heritability indicating that much of the genetic variation in obesity traits remains unexplained.” Studies involving identical twins estimate the heritability of BMI to be between 0.47 to 0.90 indicating a remarkable amount of genetic influence on obesity. [1] Identifying the thousands of genes and gene products involved in the GWAS process can potentially be expedited through the use of enrichment a

II. Methods

First a set of 50 gene terms was manually identified by reading the textbook chapter on GWAS results related to obesity [1]. Each gene term is believed to be associated with obesity as defined by different methods such as BMI, WHR, etc. These genes have been identified through GWAS methods that try to determine how the variance of a gene sequence is associated with a certain obesity related phenotype. The results are in Table 1 below.

The GWAS methods often identify new and novel gene terms that have yet to be wildly annotated. Also the GWAS methods may only identify gene terms associated with the end result of the phenotype (clinical obesity), but not the gene terms that are associated with the development or maintenance phases.

To help mitigate these two issues with gene terms identified by GWAS methods, additional non-GWAS gene terms also related to obesity are collected [4]. These pre-GWAS terms are more likely to have GO annotations already made and also provide an opportunity to identify gene terms that are expressed during times of obesity development, but not necessarily expressed once the obesity phenotype has displayed. These pre-GWAS terms are in Table 2 below.

Enrichment analysis is performed using separate lists of gene terms. One list are terms identified through pre-GWAS methodologies while the second list are gene terms identified using GWAS. Additional comparisons are done for results generated with IEA associations included versus excluded. It is hypothesized that GO terms found with IEA associations

included are more likely to reflect annotations associated with adults already expressing the obesity phenotype versus other humans only with the potential for future obesity.

III. Tabulated Results

Table 1: Selection of GWAS gene terms and their potential contribution to obesity

Item #	HGCN Term	Gene Product	Potential Role
1	LEP	Leptin	"an adipocyte-specific hormone that regulates adipose-tissue mass through hypothalamic effects on satiety and energy expenditure, acts through the leptin receptor"
2	LEPR	Leptin receptor	
3	POMC	proopiomelanocortin	"encodes the preproopiomelanocortin (POMC) protein, which is sequentially cleaved to generate several active biopeptides" "known to be involved in the neuroendocrine regulation of weight"
4	MC4R	melanocortin-4 receptor	"associated with severe, early-onset obesity"
5	FTO	Fat mass and obesity associated gene	"variants are associated with risk of obesity at all grades of severity "
6	GNPDA2	Glucosamine-6-phosphate deaminase	
7	KCTD15	potassium channel tetramerization domain containing 15	
8	MTCH2	mitochondrial carrier 2	
9	NEGR1	neuronal growth regulator 1	
10	SH2B1	SH2B adaptor protein 1	"involved in regulation of energy balance via effects on leptin and insulin signaling"
11	TMEM18	transmembrane protein 18	"regulating appetite, body weight, and energy expenditure"
12	BDNF	brain derived neurotrophic factor	
13	ETV5	ETS variant 5	
14	SEC16B	SEC16 homolog B, endoplasmic reticulum export factor	
15	CDKAL1	CDK5 regulatory subunit associated protein 1 like 1	"strong LD with the BMI GWAS SNPs in East Asians...associated with increased risk of Type 2 Diabetes"

16	KLF9	Kruppel like factor 9	
17	PCSK1	proprotein convertase subtilisin/kexin type 1	“mutations cause monogenic obesity”
18	GP2	glycoprotein 2	
19	KCNMA1	potassium calcium-activated channel subfamily M alpha 1	“Extreme obesity in adults. BMI ≥ 40”
20	NPC1	NPC intracellular cholesterol transporter 1	“early onset obesity (≤ 6 years) and extreme adult obese (BMI ≥ 40)”
21	PTER	phosphotriesterase related	
22	HS6T3		
23	MAF	MAF bZIP transcription factor	
24	TNKS	tankyrase	“Extreme obesity in children and adolescents. BMI > 97 % percentile”
25	SDCCAG8	serologically defined colon cancer antigen 8	
26	BC041448		“Distributional tails in children. BMI ≥ 95 % percentile”
27	HOXB5	homeobox B5	
28	OLFM4	olfactomedin 4	
29	PACS1	phosphofurin acidic cluster sorting protein 1	“Extreme obesity in children. BMI standard deviation score (SDS) ≥ 3, and onset at 10 years”
30	PRKCH	protein kinase C eta	
31	RMST	rhabdomyosarcoma 2 associated transcript	
32	ZZZ3		“Clinical class: obesity II. BMI ≥ 35”
33	GNAT2	G protein subunit alpha transducin 2	“Clinical class: obesity I. BMI ≥ 30”
34	HNF4G	hepatocyte nuclear factor 4 gamma	
35	MRPS33P4		
36	ADCY9	adenylate cyclase 9	
37	RPTOR	regulatory associated protein of MTOR complex 1	“Clinical class: overweight. BMI ≥ 25”
38	IRS1	insulin receptor substrate 1	“Body fat-increasing alleles. Associated with health metabolic profile (including reduced risk of T2D). Associated with measures of subcutaneous, but not visceral fat”

39	SPRY2	sprouty RTK signaling antagonist 2	“Implicated in T2D risk...associated with an adverse metabolic profile.”
40	RPGRIP1L	RPGRIP1 like	“known to be coordinately regulated with FTO via a common promoter and to display a similar pattern of hypothalamic expression...known causal role with respect to monogenic ciliopathies, some of which result in marked early obesity”
41	TMEM160		“BMI-associated”
42	LYPLAL1	lysophospholipase like 1	
43	THNSL2	threonine synthase like 2	“associated with visceral adiposity in women”
44	AA553656		
45	GRB14	growth factor receptor bound protein 14	“acts as a negative regulator of insulin receptor signaling”
46	PIGC	phosphatidylinositol glycan anchor biosynthesis class C	
47	STAB1	stabilin 1	
48	TBX15	T-box 15	“encodes a mesodermal development transcription factor and has been indicated in adipocyte differentiation and triglyceride accumulation”
49	ZNRF3	zinc and ring finger 3	
50	VEGFA	vascular endothelial growth factor A	

Table 2: Selection of pre-GWAS gene terms and their potential contribution to obesity

Item #	**HGCN Term**	**Gene Product**	**Potential Role**
1	LEP	Leptin	Secreted by adipose tissue, circulating levels higher in body
2	LEPR	Leptin receptor	
3	MC4R	Melanocortin 4 receptor	Bardet-Biedl syndrome, bulimia
4	POMC	Pro-opiomelancortin	Affect BMI in European and Hispanic Americans, influences WtHR
5	SNRPN		OMIM 176270, 182279, 602117
6	MKKS		Hypothalamic appetite dysregulation
7	AGRP		Age-dependent onset of obesity
8	SDC1	Syndecan, a cell surface proteoglycan	OMIM 186355, 186357, associated with obesity in Koreans
9	SDC3		

10	SIM1		OMIM 603128 association in Pima Indians
11	CARTPT		OMIM 602606, appears to have roles in reward, feeding, and stress...endogenous psychostimulant
12	UCP1	Uncoupling proteins of mitochondria in brown adipose	OMIM 113730, 602044 lifetime weight gain
13	UCP3		
14	GHRL	Grehlin	Role in energy homeostasis
15	PPAR		Association of Pro12A1a variant with obesity in Caucasians
16	NR0B2		Variation associated with obesity in Japanese
17	ENPP1		Susceptibility to insulin resistance
18	ADRB		Resistance to catecholamine-induced lipolysis

IV. Enrichment Analysis

A. GWAS gene terms

First an enrichment analysis is performed on the GWAS gene terms from Table 1 using the Princeton Generic Gene Ontology (GO) Term Finder website [2]. The 50 gene terms are entered and the GOA - H. sapiens annotation is chosen. Associations that are Inferred from Electronic Annotation are *included* in the results because these terms have been identified using GWAS methodology.

Of the 50 gene terms, 5 were not found to have any annotations: HS6T3, BC041448, RMST, MRPS33P4, AA553656.

Only a single biological process term is found, with a non-significant P-value:

GO term	Cluster frequency	Genome frequency	Corrected P-value	Genes annotated
GO:0006112 energy reserve metabolic process	5 of 50, 10.0%	87 of 19782, 0.4%	0.00228	POMC, IRS1, MC4R, LEP, LEPR

B. Pre-GWAS gene terms

Second, an independent enrichment analysis is performed on the 18 pre-GWAS gene terms from Table 2. There are four gene terms in common on both lists - LEP, LEPR, POMC, MC4R. These pre-GWAS gene terms are entered into the Princeton GO tool for H. sapiens.

For the first set results of results, IEA are *included* despite the gene terms being identified using pre-GWAS methods. One gene term, ADRB, is unknown and another, MKKS, is ambiguous. Three significant biological process terms are identified, but no molecular functions or cellular components. Using a p-value cutoff of 10^{-8}:

GO term	Cluster frequency	Genome frequency	Corrected P-value	Genes annotated
GO:0009725 response to hormone	12 of 17, 70.6%	979 of 19782, 4.9%	6.96e-10	PPAR, MC4R, LEP, ENPP1, NR0B2, SDC1, LEPR, UCP1, AGRP, GHRL, SNRPN, UCP3
GO:0032098 regulation of appetite	5 of 17, 29.4%	21 of 19782, 0.1%	3.47e-09	PPAR, POMC, GHRL, LEP, CARTPT
GO:0008343 adult feeding behavior	4 of 17, 23.5%	10 of 19782, 0.1%	5.47e-08	GHRL, LEP, AGRP, CARTPT

The same query parameters are used again, but this time IEA are *excluded* because the gene terms being searched for were found using non-GWAS methods. By excluding IEA the gene term MKKS is no longer considered ambiguous. Performing this search gives 9 significant results and only 1 of the 3 GO terms found when IEA results is included - “regulation of appetite”. All significant results are biological processes. The 2 terms to drop off when IEA results are excluded are “response to hormone” and “adult feeding behavior”.

GO term	Cluster frequency	Genome frequency	Corrected P-value	Genes annotated
GO:0032098 regulation of appetite	6 of 18, 29.4%	21 of 18199, 0.1%	3.47e-09	PPAR, POMC, GHRL, LEP, CARTPT
GO:0032095 regulation of response to food	5 of 18, 27.8%	13 of 18199, 0.1%	3.36e-10	PPAR, GHRL, MKKS, LEP, CARTPT
GO:0032104 regulation of response to extracellular stimulus	5 of 18, 27.8%	17 of 18199, 0.1%	1.61e-09	PPAR, GHRL, MKKS, LEP, CARTPT
GO:0032107 regulation of response to nutrient levels	5 of 18, 27.8%	17 of 18199, 0.1%	1.61e-09	PPAR, GHRL, MKKS, LEP, CARTPT
GO:0032094 response to food	5 of 18, 27.8%	20 of 18199, 0.1%	4.03e-09	PPAR, GHRL, MKKS, LEP, CARTPT
GO:0032096 negative regulation of response to food	4 of 18, 22.2%	8 of 18199, 0.1%	2.38e-08	PPAR, MKKS, LEP, CARTPT
GO:0032099 negative regulation	4 of 18,	8 of 18199,	2.38e-08	PPAR, MKKS, LEP,

of appetite	22.2%	0.1%		CARTPT
GO:0032105 negative regulation of response to extracellular stimulus	4 of 18, 22.2%	10 of 18199, 0.1%	7.14e-08	PPAR, MKKS, LEP, CARTPT
GO:0032108 negative regulation of response to nutrient levels	4 of 18, 22.2%	10 of 18199, 0.1%	7.14e-08	PPAR, MKKS, LEP, CARTPT

C. Combination of GWAS and pre-GWAS gene terms

Finally an enrichment analysis is attempted using a combination of gene terms from Table 1 and 2. This set of 64 non-duplicate gene terms represents both the pre-GWAS and post-GWAS methods. By performing separate enrichment analysis using the pre-GWAS genes, the GWAS genes, and then a combination of the two, it is hoped that further insight into biological pathways discovered will be possible by comparing their enrichment between the three sets of results.

The same search parameters are used, but this time with the P-value cutoff raised to 10^{-5}. First with IEA associations *included*:

GO term	Cluster frequency	Genome frequency	Corrected P-value	Genes annotated
GO:0009725 response to hormone	16 of 63, 28.6%	979 of 19782, 4.9%	9.58e-07	PPAR, NPC1, LEP, ENPP1, UCP1, GHRL, ADCY9, SNRPN, HNF4G, UCP3, IRS1, MC4R, LEPR, SDC1, NR0B2, AGRP, GRB14, KLF9
GO:0032870 cellular response to hormone stimulus	15 of 63, 23.8%	693 of 19782, 3.5%	3.64e-06	PPAR, IRS1, NPC1, LEP, ENPP1, NR0B2, LEPR, UCP1, AGRP, KLF9, GRB14, GHRL, ADCY9, HNF4G, UCP3
GO:0032098 regulation of appetite	5 of 63, 7.9%	21 of 19782, 0.1%	6.14e-06	PPAR, POMC, GHRL, LEP, CARTPT
GO:000009719 response to endogenous stimulus	21 of 63, 33.3%	1613 of 19782, 8.2%	1.33e-05	PPAR, NPC1, LEP, ENPP1, UCP1, GHRL, BDNF, ADCY9, SNRPN, HNF4G, RPTOR, UCP3, IRS1, MC4R, NR0B2, LEPR, SDC1, AGRP, KLF9,

				GRB14, SPRY2
GO:0008343 adult feeding behavior	4 of 63, 6.3%	10 of 19782, 0.1%	2.17e-05	GHRL, LEP, AGRP, CARTPT
GO:0007631 feeding behavior	7 of 63, .%	102 of 19782, 0.5%	3.87e-05	NEGR1, GHRL, MC4R, LEP, LEPR, AGRP, CARTPT

And with IEA *excluded*:

GO term	**Cluster frequency**	**Genome frequency**	**Corrected P-value**	**Genes annotated**
GO:0032098 regulation of appetite	6 of 64, 9.4%	13 of 18199, 0.1%	2.23e-09	PPAR, POMC, GHRL, MKKS, LEP, CARTPT
GO:0032095 regulation of response to food	5 of 64, 7.8%	13 of 18199, 0.1%	5.15e-07	PPAR, GHRL, MKKS, LEP, CARTPT
GO:0032104 regulation of response to extracellular stimulus	5 of 64, 7.8%	17 of 18199, 0.1%	2.45e-06	PPAR, GHRL, MKKS, LEP, CARTPT
GO:0032107 regulation of response to nutrient levels	5 of 64, 7.8%	17 of 18199, 0.1%	2.45e-06	PPAR, GHRL, MKKS, LEP, CARTPT
GO:0032094 response to food	5 of 64, 7.8%	20 of 18199, 0.1%	6.08e-06	PPAR, GHRL, MKKS, LEP, CARTPT
GO:0032096 negative regulation of response to food	4 of 64, 6.2%	8 of 18199, 0.0%	8.59e-06	PPAR, MKKS, LEP, CARTPT
GO:0032099 negative regulation of appetite	4 of 64, 6.2%	8 of 18199, 0.0%	8.59e-06	PPAR, MKKS, LEP, CARTPT
GO:0032105 negative regulation of response to extracellular stimulus	4 of 64, 6.2%	10 of 18199, 0.1%	2.56e-05	PPAR, MKKS, LEP, CARTPT
GO:0032108 negative regulation of response to nutrient levels	4 of 64, 6.2%	10 of 18199, 0.1%	2.56e-05	PPAR, MKKS, LEP, CARTPT

V. Discussion

A. Pre-GWAS gene terms: IEA excluded vs included

Three GO terms are found to be significant when IEA associations are included - "response to hormone", "regulation of appetite", and "adult feeding behavior" - however, when

IEA associations are excluded, 2 of the 3 terms drop off and are no longer significant. The only remaining GO term that is found to be significant with both IEA associations included and excluded is “regulation of appetite”. This indicates that the two GO term associations that drop off, “response to hormone” and “adult feeding behavior” are likely to be expressed in adult humans with obesity because IEA associations are generally associated with expressions of genes in adults that already reflect the obesity phenotype.

B. Pre-GWAS vs GWAS

Given that only a single (non-significant) GO association is found for the GWAS gene terms it is hard to make conclusions. The one GO association found for GWAS gene terms, “energy reserve metabolic process”, is never enriched to the point of significance when pre-GWAS terms are added. This could potentially point towards the GO term being something expressed as pathways in adults displaying the obesity phenotype versus maintenance and/or development pathways.

C. Combination of Pre-GWAS and GWAS IEA: excluded vs included

Of the 18 pre-GWAS gene terms and the 50 GWAS gene terms identified, there are 4 duplicate gene terms found on both lists - LEP, LEPR, MC4R, and POMC.

When IEA associations are included, 6 GO terms are found to be significant, while when IEA associations are excluded, 9 GO terms are found to be significant. There is a single GO term that is found with and without IEA associations - “regulation of appetite”. Given that it is found on both lists could potentially indicate that it is related to obesity throughout the lifetime of a human.

The other 5 GO terms found without IEA associations are composed of two terms related to response to hormones, two terms related to feeding behavior, and a term related to the response of endogenous stimulus. The other 8 GO terms found when IEA associations are included are all related to either food, nutrition, or extracellular stimulus.

Just like when enriching the pre-GWAS gene terms alone, the GO terms related to hormones are found only with IEA associations included. This again indicates that hormone response is something potentially more relevant to obesity during development than later in adulthood.

I am unsure what conclusions could be made about the GO terms related to endogenous and extracellular stimulus. The “endogenous” term is indicates something related to growth from within while “extracellular” term indicates something external. This does make sense given that the endogenous term seems related to growth and development (childhood) and is found only without IEA associations included.

D. Pre-GWAS vs Combination of Pre-GWAS and GWAS

A total of 11 unique GO terms are found using the pre-GWAS gene terms and a total of 14 unique GO terms are found using a combination of pre-GWAS and GWAS gene terms. Comparing the two sets of results shows that all 11 of the pre-GWAS GO terms remain on the list of significant results when enriched along with the GWAS gene terms. This could potentially indicate that the GO terms found with pre-GWAS gene terms are not only relevant to obesity in humans during development years, but also entire lifespan. This makes sense intuitively as the GO terms are mostly related to appetite and response to food which will always be a critical influence on obesity at any age.

The 3 GO terms that are newly found as significant when the GWAS gene terms are added are “GO:0032870 cellular response to hormone stimulus”, “GO:000009719 response to endogenous stimulus”, and “GO:0007631 feeding behavior”. Since they are only present when the GWAS terms are added to the enrichment it could indicate that these three terms are part of pathways expressed more in adults already displaying the obesity phenotype.

VI. Conclusions

Adding Pre-GWAS gene terms to the enrichment process greatly improved the significance of the GO term results found. This was true for results both with IEA included and excluded. The GO terms found without IEA associations are related to hormones, feeding behavior and endogenous stimulus while the GO terms found with IEA associations are related to food, nutrition, and extracellular stimulus. The only GO term to appear in both result lists is the rather obvious "regulation of appetite".

Some conflicting evidence was discovered for the 3 GO terms found as significant only during the enrichment process with both pre-GWAS and GWAS terms included. That these 3 terms, “GO:0032870 cellular response to hormone stimulus”, “GO:000009719 response to endogenous stimulus”, and “GO:0007631 feeding behavior”, are only found with the GWAS gene terms included indicates they are expressed in adults displaying the obesity phenotype as opposed to others with the potential for future obesity.

However this interpretation is in conflict with the interpretation of the Pre-GWAS IEA included versus excluded analysis that indicated GO terms related to hormones and endogenous stimulus were found significant only with IEA associations excluded, contrarily indicating they are more associated with potential for development of obesity.

Part 2: Atherosclerosis

I. Background

My previous assignment has focused on the regression of atherosclerosis, specifically the factors that go into improving plasma lipoprotein profiles by lowering concentrations of atherogenic apolipoprotein B (ApoB). Reductions of ApoB levels have been associated with regression of atherosclerosis; returning a patient to the lipid concentration levels that they maintained before the presence atherosclerosis or other heart disease.

II. Methods

A set of gene terms related to the regression of atherosclerosis was previously curated through use of the MEDIE natural language search engine [3]. These 27 gene terms are shown in Table 3 below.

No distinction was made between pre-GWAS and GWAS methodologies when collecting these gene terms. This means no conclusions can be made comparing pre-GWAS and GWAS enrichment results for Atherosclerosis regression by itself. In other words, all of the atherosclerosis regression enrichment results I get here are a reflection of the gene terms found through a combination of pre-GWAS and GWAS methodologies.

III. Tabulated Results

Table 3: Previously curated selection of genes believed to be related to the process of atherosclerosis regression

Item #	HGCN	Item #	HGCN
1	APOB	15	ABHD5
2	APOE	16	ATF4
3	MSR1	17	MTTP
4	SCARB1	18	LCAT
5	CCR7	19	PTEN
6	NR1H3	20	HNF4A
7	ABCA1	21	PTPN11
8	MIR33A	22	MTOR
9	CETP	23	MAP2K4
10	PPARA	24	APOBEC1
11	PPARG	25	FGA
12	LDLR	26	FGB
13	LDLRAP1	27	FGG
14	PCSK9		

IV. Enrichment Analysis

A. Atherosclerosis regression gene terms IEA included vs excluded

Using all 27 gene terms gives a large number of significant results for biological processes, molecular functions, and cellular components, when IEA associations are both included and excluded. The two sets of result do not have much differences between them overall as they consist mostly of GO terms related to lipids & cholesterol and with relatively similar p-values.

One key difference is that the IEA included results tend to place more significance on the GO terms related to triglycerides. These are the triglyceride related GO terms found to be significant with IEA associations included:

GO term	Cluster frequency	Genome frequency	Corrected P-value	Genes annotated
GO:0006641 triglyceride metabolic process	10 of 27, 37.0%	99 of 19782, 0.1%	6.59e-14	LDLR, APOE, APOBEC1, CETP, PTPN11, PCSK9, NR1H3, ABHD5, SCARB1, APOB
GO:0090207 regulation of triglyceride metabolic process	6 of 27, 22.2%	34 of 19782, 0.2%	6.33e-09	LDLR, NR1H3, APOE, APOBEC1, SCARB1, ABHD5
GO:0070328 triglyceride homeostasis	5 of 27, 18.5%	33 of 19782 genes, 0.2%	1.00e-06	HNF4A, NR1H3, APOE, CETP, SCARB1

Versus the significant triglyceride related terms with IEA associations excluded:

GO term	Cluster frequency	Genome frequency	Corrected P-value	Genes annotated
GO:0070328 triglyceride homeostasis	5 of 27, 18.5%	28 of 18199, 0.2%	4.63e-07	HNF4A, NR1H3, APOE, CETP, SCARB1
GO:0006641 triglyceride metabolic process	6 of 27, 22.2%	73 of 18199, 0.4%	9.30e-07	LDLR, NR1H3, APOE, CETP, SCARB1, ABHD5
GO:0090208 positive regulation of triglyceride metabolic process	4 of 27, 14.8%	16 of 18199, 0.1%	6.87e-06	LDLR, NR1H3, SCARB1, ABHD5

B. Combination of Atherosclerosis Regression gene terms and Obesity Pre-GWAS & GWAS gene terms, with IEA included

The 27 Atherosclerosis regression related gene terms are combined with the 64 pre-GWAS and GWAS obesity related gene terms to create a list of 91 non-duplicate gene terms. The terms with significance of more than e-08 are shown below:

Gene Ontology term	Cluster frequency	Genome frequency	P-value	Gene Ontology term	Cluster frequency	Genome frequency	P-value
regulation of lipid localization	17 of 89 genes, 19.1%	136 of 19782 genes, 0.7%	5.80E-17	cellular response to endogenous stimulus	29 of 89 genes, 32.6%	1353 of 19782 genes, 6.8%	9.54E-10
lipid localization	22 of 89 genes, 24.7%	382 of 19782 genes, 1.9%	2.49E-15	cholesterol storage	7 of 89 genes, 7.9%	17 of 19782 genes, 0.1%	1.04E-09
cholesterol transport	14 of 89 genes, 15.7%	85 of 19782 genes, 0.4%	2.67E-15	cellular response to chemical stimulus	43 of 89 genes, 48.3%	3180 of 19782 genes, 16.1%	2.33E-09
sterol transport	14 of 89 genes, 15.7%	95 of 19782 genes, 0.5%	1.38E-14	regulation of biological quality	48 of 89 genes, 53.9%	3975 of 19782 genes, 20.1%	2.96E-09
lipid homeostasis	15 of 89 genes, 16.9%	129 of 19782 genes, 0.7%	3.43E-14	homeostatic process	33 of 89 genes, 37.1%	1868 of 19782 genes, 9.4%	3.00E-09
cholesterol homeostasis	13 of 89 genes, 14.6%	78 of 19782 genes, 0.4%	4.11E-14	macrophage derived foam cell differentiation	8 of 89 genes, 9.0%	35 of 19782 genes, 0.2%	4.93E-09
sterol homeostasis	13 of 89 genes, 14.6%	79 of 19782 genes, 0.4%	4.90E-14	foam cell differentiation	8 of 89 genes, 9.0%	35 of 19782 genes, 0.2%	4.93E-09
organic hydroxy compound transport	18 of 89 genes, 20.2%	244 of 19782 genes, 1.2%	6.15E-14	response to nutrient levels	18 of 89 genes, 20.2%	492 of 19782 genes, 2.5%	1.10E-08
response to endogenous stimulus	36 of 89 genes, 40.4%	1613 of 19782 genes, 8.2%	1.22E-13	negative regulation of lipid localization	8 of 89 genes, 9.0%	41 of 19782 genes, 0.2%	1.95E-08
response to hormone	29 of 89 genes, 32.6%	979 of 19782 genes, 4.9%	2.27E-13	response to insulin	14 of 89 genes, 15.7%	268 of 19782 genes, 1.4%	2.93E-08
lipid storage	12 of 89 genes, 13.5%	66 of 19782 genes, 0.3%	2.44E-13	response to extracellular stimulus	18 of 89 genes, 20.2%	524 of 19782 genes, 2.6%	3.12E-08

lipid transport	19 of 89 genes, 21.3%	348 of 19782 genes, 1.8%	1.97E-12	regulation of lipid metabolic process	16 of 89 genes, 18.0%	388 of 19782 genes, 2.0%	3.21E-08
cellular response to hormone stimulus	24 of 89 genes, 27.0%	693 of 19782 genes, 3.5%	5.66E-12	triglyceride metabolic process	10 of 89 genes, 11.2%	99 of 19782 genes, 0.5%	4.31E-08
regulation of plasma lipoprotein particle levels	12 of 89 genes, 13.5%	87 of 19782 genes, 0.4%	8.17E-12	response to nitrogen compound	24 of 89 genes, 27.0%	1053 of 19782 genes, 5.3%	4.58E-08
regulation of lipid storage	10 of 89 genes, 11.2%	45 of 19782 genes, 0.2%	1.07E-11	response to organonitrogen compound	23 of 89 genes, 25.8%	958 of 19782 genes, 4.8%	4.64E-08
regulation of macrophage derived foam cell differentiation	8 of 89 genes, 9.0%	29 of 19782 genes, 0.1%	9.19E-10	regulation of sterol transport	8 of 89 genes, 9.0%	46 of 19782 genes, 0.2%	5.25E-08
lipoprotein metabolic process	12 of 89 genes, 13.5%	128 of 19782 genes, 0.6%	9.43E-10	regulation of cholesterol transport	8 of 89 genes, 9.0%	46 of 19782 genes, 0.2%	5.25E-08
				positive regulation of lipid metabolic process	11 of 89 genes, 12.4%	139 of 19782 genes, 0.7%	6.01E-08

V. Discussion

A. Atherosclerosis regression gene terms IEA included vs excluded

The result sets for IEA included vs. excluded do not differ much in their contents or p-value. This could indicate that the GO terms found are expressed at all parts of the atherosclerosis lifecycle. Or it could simply be a result of the fact that, unlike with the obesity set, the gene terms used for this enrichment analysis are not differentiated between pre-GWAS and GWAS methodologies.

The biggest difference between the two sets are in the triglyceride related GO terms. For example, the most significant triglyceride related GO term when IEA associations are included is “triglyceride metabolic process” with a p-value of e-14, however when IEA associations are excluded the p-value becomes only e-07. The difference in triglyceride terms being found more significant in IEA included could potentially indicate that triglycerides play a bigger role in the actual atherosclerosis development and regression process versus being expressed as an indicator for future development.

Looking at how the p-value is calculated for the GO term “triglyceride metabolic process” shows that 10 genes are annotated to it when IEA associations are included, but only 6 genes are annotated when IEA associations are excluded. The 4 genes that are only annotated with IEA associations included - APOBEC1, PTPN11, PCSK9, and APOB - could potentially be an area of future research to determine their contribution. Being annotated with electronic associations, but not with non-electronic associations, could indicate that the gene terms play only a role in the regression process and not as an indicator for future atherosclerosis. The gene APOB here is especially interesting given its central nature in the atherosclerosis process.

B. Combination of Atherosclerosis Regression gene terms and Obesity Pre-GWAS & GWAS gene terms IEA included

When enrichment is done using both the obesity and Atherosclerosis regression gene term, with IEA associations included, it seems that the resulting significant GO terms are heavily influenced by the Atherosclerosis regression gene terms. The most significant GO terms associated with Atherosclerosis regression alone are found to be still highly significant when enriched along with the obesity genes. These include GO terms related to processes involving hormones, lipids and cholesterol.

Some gene terms identified as related to obesity are now being annotated to GO terms previously identified through enrichment of Atherosclerosis regression gene terms alone. For example, the most significant GO term found is “regulation of lipid localization”. It was previously annotated with 15 of 27 genes related to Atherosclerosis regression. Now when enriched along with obesity gene terms it is further annotated with an additional 6 gene terms related to obesity - FTO, PPAR, NPC1, ENPP1, POMC, and GHRL, for a total of 21 gene term annotations (23.3% of 90 genes). These same obesity gene terms are annotated to various other lipid related GO terms as well. This further indicates a link between lipid pathways, the obesity phenotype and Atherosclerosis regression.

On the other hand there are also GO terms previously identified using obesity gene terms that are now also annotated with Atherosclerosis regression gene terms. For example, “regulation of response to food” is now annotated with the gene term MTOR that was believed to be related to Atherosclerosis regression. Also the GO term “response to nutrient levels” is now annotated with Atherosclerosis regression gene terms LDLR, ATF4, MTOR, PPARG, ABCA1, APOE, PCSK9, and PTEN. These additional gene annotations increases the P-value calculation and hence the significance of the GO term pathway.

A set of new GO terms are found to be significant only when enriched with obesity and Atherosclerosis regression gene terms combined. These GO terms are "response to insulin", "response to nitrogen compound", and "response to organonitrogen compound". It is interesting that an insulin related GO term is only found now after enrichment using both obesity and Atherosclerosis regression gene terms. Perhaps the body's response to insulin and obesity was previously captured under more generic GO terms related to hormone response. It is unknown

what the significance behind the two nitrogen related GO terms could be. Nitrogen is a component of insulin, but that seems like an unlikely connection.

C. Obesity gene terms vs Combination of Atherosclerosis Regression gene terms and Obesity Pre-GWAS & GWAS gene terms IEA included

Only a few GO terms found previously using obesity terms exclusively remain significant when enriched along with the Atherosclerosis regression related gene terms. These GO terms include "response to hormone", "cellular response to hormone stimulus", "response to endogenous stimulus", and "regulation of response to extracellular stimulus".

In theory this implies that these GO terms are less associated with Atherosclerosis regression and thus have their P-value's fall below the significance cutoff. Similarly, since all of the food, nutrition, and appetite related GO terms are no longer found to be significant from enrichment with Atherosclerosis regression gene terms, in theory it could be said that food, nutrition and appetite have less association with Atherosclerosis regression than they do with obesity. The reason for their drop off in significance appears to be the lack of gene annotations related to Atherosclerosis regression.

On the other hand some GO terms that were identified using obesity genes, but were not considered significant due to being below the P-value cutoff, are now considered significant when enriched along with the Atherosclerosis regression terms. For example, the GO term “response to nutrient levels” had a P-value of only 1.44e-05 when enriched with obesity terms alone, but when enriched alongside Atherosclerosis regression gene terms is now annotated with an additional 8 genes and has a P-value of 1.10e-08.

D. Atherosclerosis Regression vs Combination of Atherosclerosis Regression gene terms and Obesity Pre-GWAS & GWAS gene terms IEA included

Most of the GO terms found to be significant to Atherosclerosis regression are still significant when enriched alongside obesity related gene terms. One exception again appears to be the GO terms related to triglycerides. While the previous Atherosclerosis regression enrichment process with IEA associations included resulted in 3 triglyceride related GO terms - "triglyceride metabolic process", "regulation of triglyceride metabolic process", and "triglyceride homeostasis", when these gene terms are enriched alongside the obesity gene terms only one GO term remains significant (although less so) - “triglyceride metabolic process”.

This could indicate that triglycerides have a higher impact on Atherosclerosis regression than on obesity, which intuitively makes sense. The “triglyceride metabolic process” GO term is annotated with 10 genes related to Atherosclerosis regression and 0 genes related to obesity which explains why its significance has been reduced.

VI. Conclusions

Enrichment analysis of Atherosclerosis regression terms with IEA associations included resulted in "triglyceride metabolic processing" being the most significant GO term with a P-value of e-14. However, when IEA associations are excluded, the significance of this pathway drops to e-07, indicating that triglyceride related biological processes are associated with the Atherosclerosis phenotype, but not necessarily related to development or maintenance pathways. When obesity terms are added to the enrichment, the significance of triglyceride related terms drops off, indicating they are more related to Atherosclerosis than to obesity.

When enrichment is performed using a combination of the Atherosclerosis regression and obesity gene terms it is found that the GO terms associated previously with enrichment attempts using the Atherosclerosis regression gene terms alone still tend to dominate the results. This leads strong evidence towards linking a reduction in obesity to regression of Atherosclerosis.

A short review of literature indicates this to be a commonly found linkage. A large multi-ethnic study of more than 5000 people done in 2010 [5] found that obesity was directly associated with concentric changes to the left ventricle (LV). Another study in 2012 [6] found a correlation between waist circumference & weight to height ratio to several quantitative measurements of atherosclerosis such as “pulse wave velocity, intima-media thickness of the common carotid artery, augmentation index, ankle-brachial index, and central and peripheral pulse pressure.”

Further most of these Atherosclerosis regression associated GO terms have had their number of gene annotations increased in the combined gene term result set - the most significant biological pathways found in common, “regulation of lipid localization”, is now annotated with 15 (of 27) Atherosclerosis regression gene terms and 6 (of 50) obesity gene terms. The GO term “response to nutrient levels” was previously annotated with 8 (of 50) obesity gene terms and now has an additional 6 (of 27) Atherosclerosis regression gene annotations.

Finally new GO terms related to insulin and nitrogen are only found when enrichment is performed using the combined Atherosclerosis regression and obesity gene sets. These GO terms were previously not considered significant when enriched with gene terms from one set or the other, but are made significant when enriched using the combined gene set. This provides good evidence of shared pathways.

Part 3: References

1. Hedman, Asa K, et al. *The Genetics of Obesity, Chapter 3 - Genome-Wide Association Studies of Obesity*. 2014.

2. Generic Gene Ontology Term Finder. Retrieved from https://go.princeton.edu/cgi-bin/GOTermFinder. Princeton University.

3. MEDIE - Semantic retrieval engine for MEDLINE. Retrieved from http://www.nactem.ac.uk/medie/search.cgi. National Center for Text Mining.

4. McCormack, Shana E., *The Genetics of Obesity, Chapter 1 - Genetic Variation and Obesity Prior to the Era of Genome-Wide Association Studies*. 2014.

5. Evrim B. Turkbey, Robyn L. McClelland, Richard A. Kronmal, Gregory L. Burke, Diane E. Bild, Russell P. Tracy, Andrew E. Arai, João A.C. Lima and David A. Bluemke. “The Impact of Obesity on the Left Ventricle. The Multi-Ethnic Study of Atherosclerosis (MESA)” . 2010. JACC: Cardiovascular Imaging, 3(3). doi:10.1016/j.jcmg.2009.10.012

6. Jose I Recio-Rodriguez, Manuel A Gomez-Marcos, Maria C Patino-Alonso, Cristina Agudo-Conde, Emiliano Rodriguez-Sanchez, Luis Garcia-Ortiz and the Vasorisk group. “Abdominal obesity vs general obesity for identifying arterial stiffness, subclinical atherosclerosis and wave reflection in healthy, diabetics and hypertensive ”. 2012. MC Cardiovascular Disorders. https://doi.org/10.1186/1471-2261-12-3