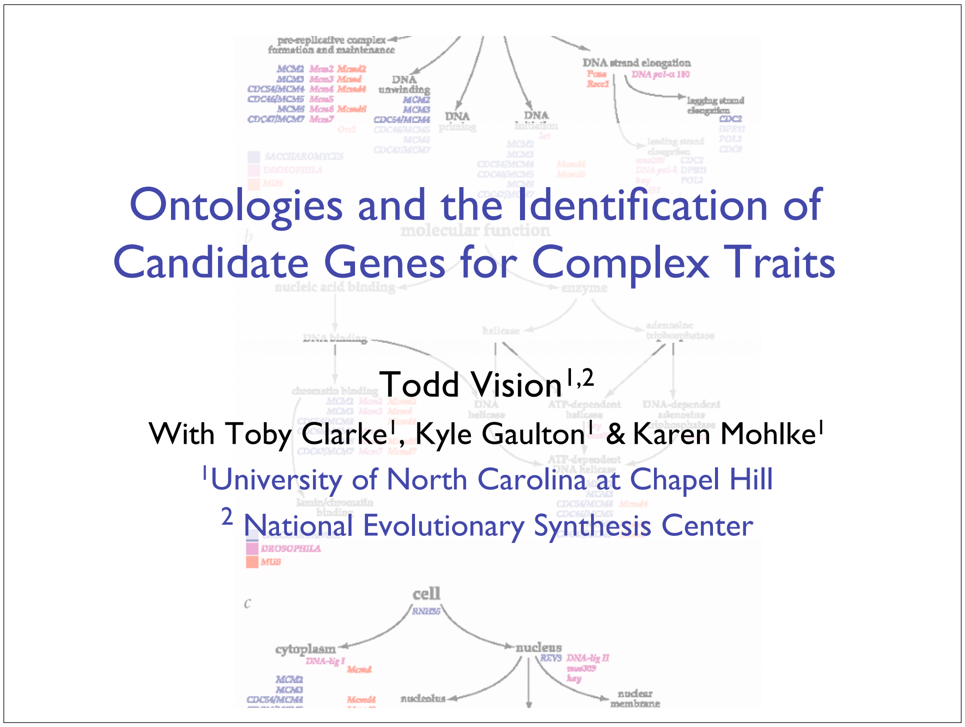


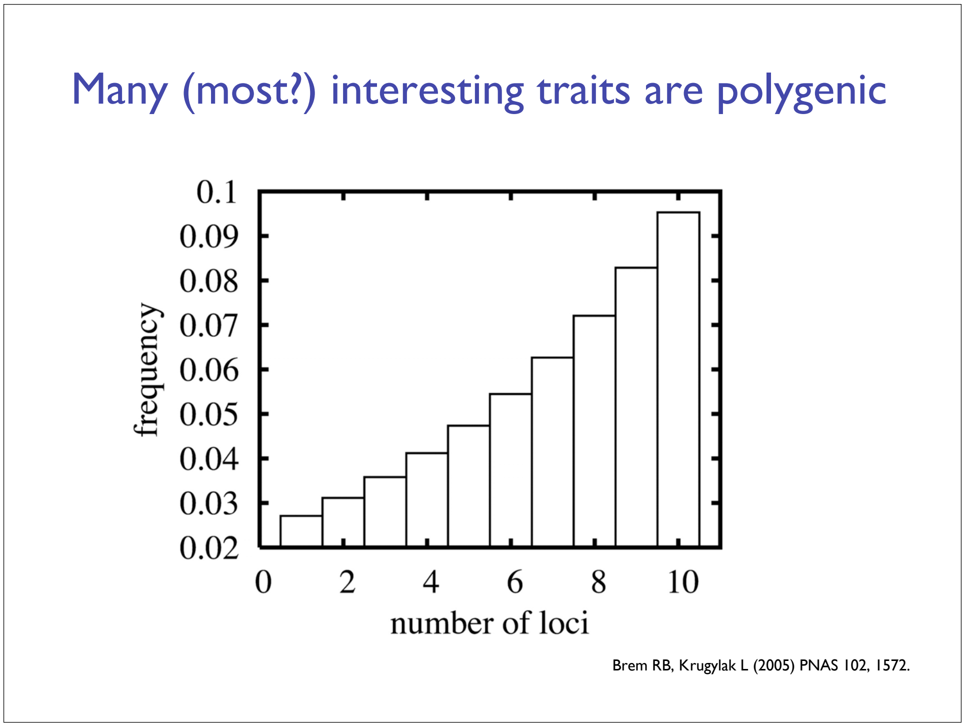
Ontologies and the Identification of Candidate Genes for Complex Traits

Todd Vision[1,2]

With Toby Clarke[1], Kyle Gaulton[1] & Karen Mohlke[1]

[1]University of North Carolina at Chapel Hill

[2] National Evolutionary Synthesis Center



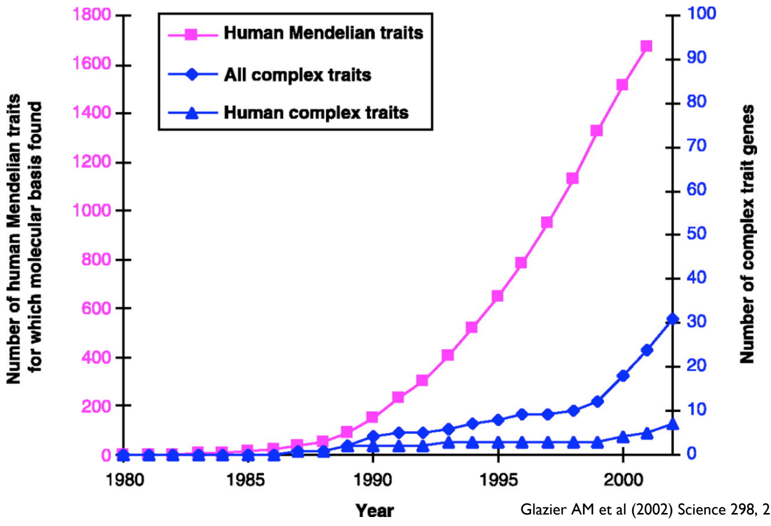
Many (most?) interesting traits are polygenic

Brem RB, Krugylak L (2005) PNAS 102, 1572.

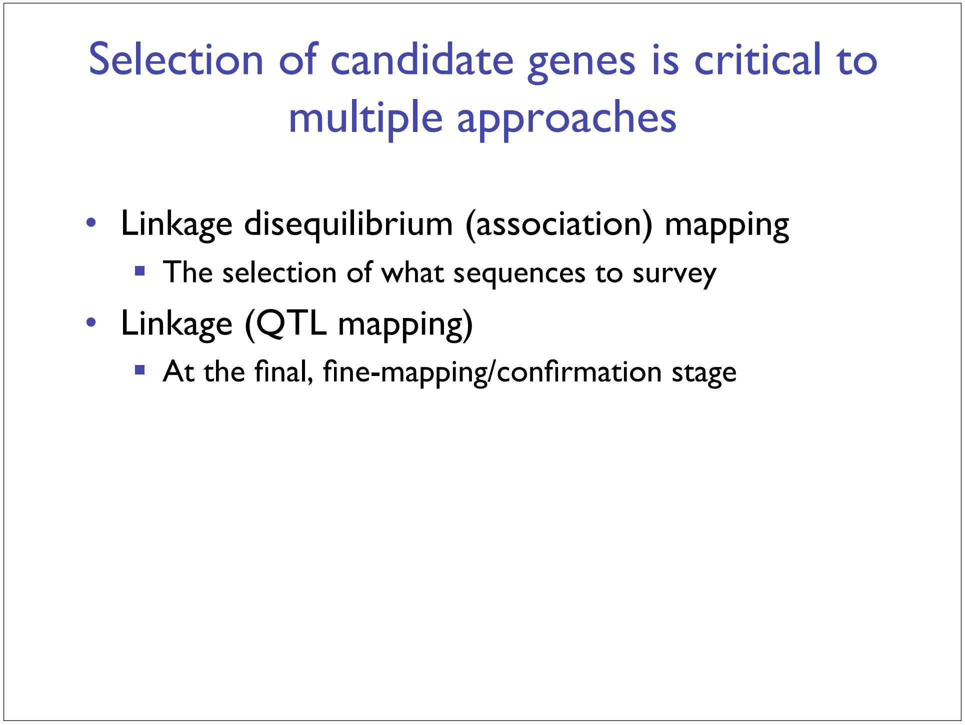
Complex traits are an important class of polygenic traits

- No gene is either necessary or sufficient
  - Heterogeneity
  - Multiple genes, potentially with epistasis
  - A strong environmental component
- Examples
  - Schizophrenia in humans
  - Bristle number in Drosophila
  - Water use efficiency in plants
  - Components of yield and fitness

Complex traits are particularly hard to dissect

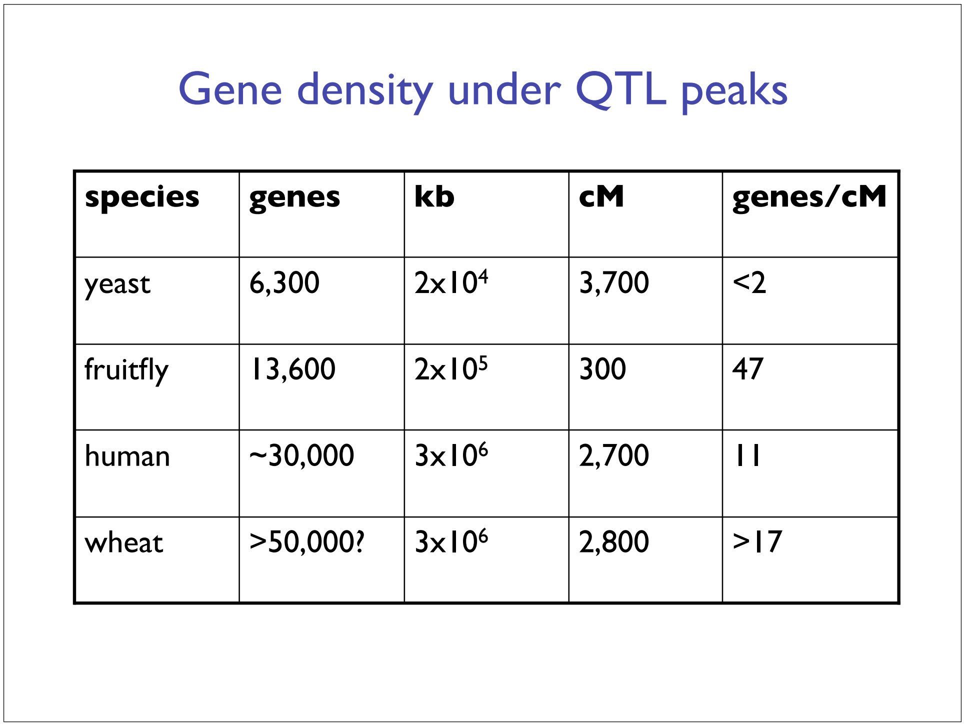


Glazier AM et al (2002) Science 298, 2345



Selection of candidate genes is critical to multiple approaches

- Linkage disequilibrium (association) mapping
  - The selection of what sequences to survey
- Linkage (QTL mapping)
  - At the final, fine-mapping/confirmation stage



Gene density under QTL peaks

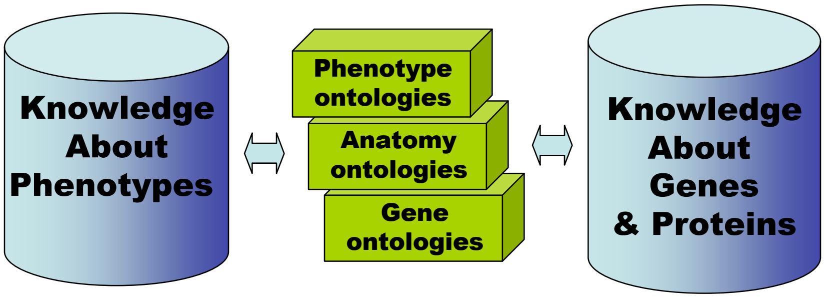
species	genes	kb	cM	genes/cM
yeast	6,300	2x10^4	3,700	<2
fruitfly	13,600	2x10^5	300	47
human	~30,000	3x10^6	2,700	11
wheat	>50,000?	3x10^6	2,800	>17

What makes a good candidate gene?

- The mutant expresses a relevant phenotype
  - Or its homolog in a model organism
- Transcribed in relevant tissues and conditions
  - As determined by ESTs, microarrays, *in situ* hybridizations, etc.
- Regulates or participates in a relevant pathway
  - Known from biochemical studies
  - Suspected from interaction data
  - Suspected from protein domain data

Ranking candidate genes involves human reasoning over many different sources of complex data

Can we pick candidates genes computationally?



Knowledge About Phenotypes

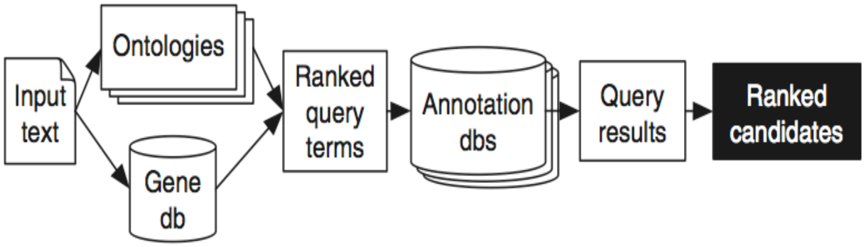
Phenotype ontologies

Anatomy ontologies

Gene ontologies

Knowledge About Genes & Proteins

CAndidate SEarch And Rank



Input text

Ontologies

Gene db

Ranked query terms

Annotation dbs

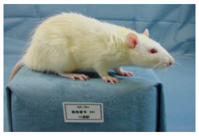
Query results

Ranked candidates

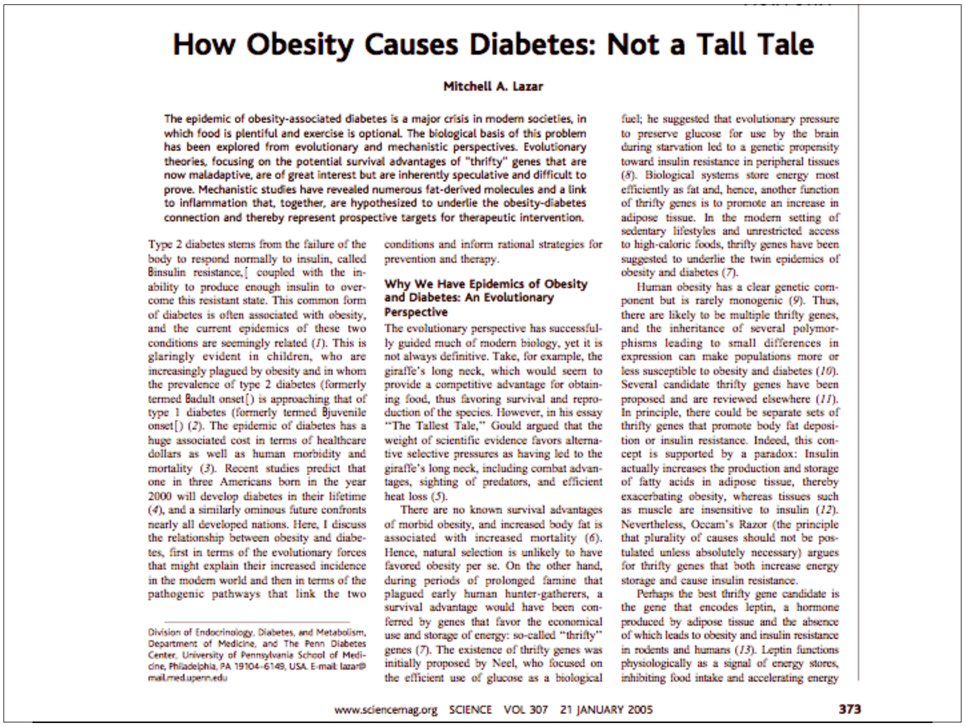
Gaulton KJ, Mohlke KL, Vision TJ (2007) Bioinformatics 23, 1132-1140.

Original application: Type 2 Diabetes (T2D)

- 5th leading cause of death by disease in the U.S.
- Characterized by
  - Insulin resistance
  - Impaired pancreatic beta-cell function
  - Increased hepatic glucose production
- Classic complex trait
  - Heterogeneous
  - Under strong environmental control
  - Inheritance is complex
- A handful of candidate genes available as of 2006
  - Not enough for the association study of 200 candidates (FUSION)
  - How to pick them?



Goto-Kakizaki rat



How Obesity Causes Diabetes: Not a Tall Tale

Mitchell A. Lazar

The epidemic of obesity-associated diabetes is a major crisis in modern societies, in which food is plentiful and exercise is optional. The biological basis of this problem has been explored from evolutionary and mechanistic perspectives. Evolutionary theories, focusing on the potential survival advantages of "thrifty" genes that are now maladaptive, are of great interest but are inherently speculative and difficult to prove. Mechanistic studies have revealed numerous fat-derived molecules and a link to inflammation that, together, are hypothesized to underlie the obesity-diabetes connection and thereby represent prospective targets for therapeutic intervention.

Type 2 diabetes stems from the failure of the body to respond normally to insulin, called Binsulin resistance,[coupled with the inability to produce enough insulin to overcome this resistant state. This common form of diabetes is often associated with obesity, and the current epidemics of these two conditions are seemingly related (*1*). This is glaringly evident in children, who are increasingly plagued by obesity and in whom the prevalence of type 2 diabetes (formerly termed Badult onset[) is approaching that of type 1 diabetes (formerly termed Bjuvenile onset[) (*2*). The epidemic of diabetes has a huge associated cost in terms of healthcare dollars as well as human morbidity and mortality (*3*). Recent studies predict that one in three Americans born in the year 2000 will develop diabetes in their lifetime (*4*), and a similarly ominous future confronts nearly all developed nations. Here, I discuss the relationship between obesity and diabetes, first in terms of the evolutionary forces that might explain their increased incidence in the modern world and then in terms of the pathogenic pathways that link the two conditions and inform rational strategies for prevention and therapy.

Why We Have Epidemics of Obesity and Diabetes: An Evolutionary Perspective

The evolutionary perspective has successfully guided much of modern biology, yet it is not always definitive. Take, for example, the giraffe's long neck, which would seem to provide a competitive advantage for obtaining food, thus favoring survival and reproduction of the species. However, in his essay "The Tallest Tale," Gould argued that the weight of scientific evidence favors alternative selective pressures as having led to the giraffe's long neck, including combat advantages, sighting of predators, and efficient heat loss (*5*).

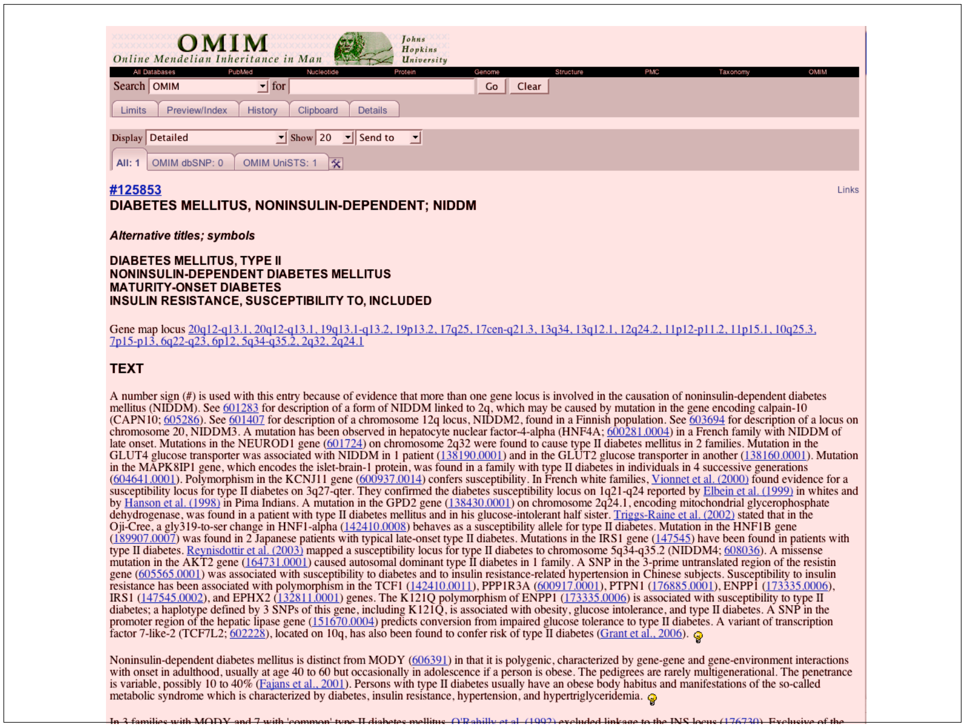
There are no known survival advantages of morbid obesity, and increased body fat is associated with increased mortality (*6*). Hence, natural selection is unlikely to have favored obesity per se. On the other hand, during periods of prolonged famine that plagued early human hunter-gatherers, a survival advantage would have been conferred by genes that favor the economical use and storage of energy: so-called "thrifty" genes (*7*). The existence of thrifty genes was initially proposed by Neel, who focused on the efficient use of glucose as a biological fuel; he suggested that evolutionary pressure to preserve glucose for use by the brain during starvation led to a genetic propensity toward insulin resistance in peripheral tissues (*8*). Biological systems store energy most efficiently as fat and, hence, another function of thrifty genes is to promote an increase in adipose tissue. In the modern setting of sedentary lifestyles and unrestricted access to high-caloric foods, thrifty genes have been suggested to underlie the twin epidemics of obesity and diabetes (*7*).

Human obesity has a clear genetic component but is rarely monogenic (*9*). Thus, there are likely to be multiple thrifty genes, and the inheritance of several polymorphisms leading to small differences in expression can make populations more or less susceptible to obesity and diabetes (*10*). Several candidate thrifty genes have been proposed and are reviewed elsewhere (*11*). In principle, there could be separate sets of thrifty genes that promote body fat deposition or insulin resistance. Indeed, this concept is supported by a paradox: Insulin actually increases the production and storage of fatty acids in adipose tissue, thereby exacerbating obesity, whereas tissues such as muscle are insensitive to insulin (*12*). Nevertheless, Occam's Razor (the principle that plurality of causes should not be postulated unless absolutely necessary) argues for thrifty genes that both increase energy storage and cause insulin resistance.

Perhaps the best thrifty gene candidate is the gene that encodes leptin, a hormone produced by adipose tissue and the absence of which leads to obesity and insulin resistance in rodents and humans (*13*). Leptin functions physiologically as a signal of energy stores, inhibiting food intake and accelerating energy

Division of Endocrinology, Diabetes, and Metabolism, Department of Medicine, and The Penn Diabetes Center, University of Pennsylvania School of Medicine, Philadelphia, PA 19104–6149, USA. E-mail: lazar@mail.med.upenn.edu

www.sciencemag.org SCIENCE VOL 307 21 JANUARY 2005 373



#125853

DIABETES MELLITUS, NONINSULIN-DEPENDENT; NIDDM

Alternative titles; symbols

DIABETES MELLITUS, TYPE II
NONINSULIN-DEPENDENT DIABETES MELLITUS
MATURITY-ONSET DIABETES
INSULIN RESISTANCE, SUSCEPTIBILITY TO, INCLUDED

Gene map locus 20q12-q13.1, 20q12-q13.1, 19q13.1-q13.2, 19p13.2, 17q25, 17cen-q21.3, 13q34, 13q12.1, 12q24.2, 11p12-p11.2, 11p15.1, 10q25.3, 7p15-p13, 6q22-q23, 6p12, 5q34-q35.2, 2q32, 2q24.1

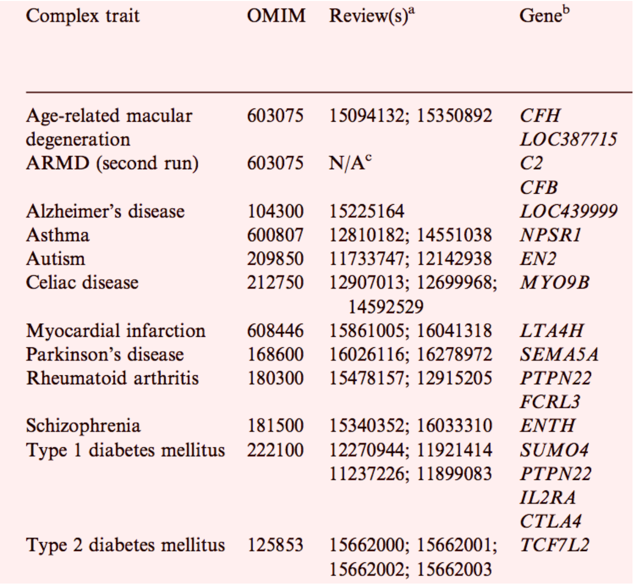
TEXT

A number sign (#) is used with this entry because of evidence that more than one gene locus is involved in the causation of noninsulin-dependent diabetes mellitus (NIDDM). See 601283 for description of a form of NIDDM linked to 2q, which may be caused by mutation in the gene encoding calpain-10 (CAPN10; 605286). See 601407 for description of a chromosome 12q locus, NIDDM2, found in a Finnish population. See 603694 for description of a locus on chromosome 20, NIDDM3. A mutation has been observed in hepatocyte nuclear factor-4-alpha (HNF4A; 600281.0004) in a French family with NIDDM of late onset. Mutations in the NEUROD1 gene (601724) on chromosome 2q32 were found to cause type II diabetes mellitus in 2 families. Mutation in the GLUT4 glucose transporter was associated with NIDDM in 1 patient (138190.0001) and in the GLUT2 glucose transporter in another (138160.0001). Mutation in the MAPK8IP1 gene, which encodes the islet-brain-1 protein, was found in a family with type II diabetes in individuals in 4 successive generations (604641.0001). Polymorphism in the KCNJ11 gene (600937.0014) confers susceptibility. In French white families, Vionnet et al. (2000) found evidence for a susceptibility locus for type II diabetes on 3q27-qter. They confirmed the diabetes susceptibility locus on 1q21-q24 reported by Elbein et al. (1999) in whites and by Hanson et al. (1998) in Pima Indians. A mutation in the GPD2 gene (138430.0001) on chromosome 2q24.1, encoding mitochondrial glycerophosphate dehydrogenase, was found in a patient with type II diabetes mellitus and in his glucose-intolerant half sister. Triggs-Raine et al. (2002) stated that in the Oji-Cree, a gly319-to-ser change in HNF1-alpha (142410.0008) behaves as a susceptibility allele for type II diabetes. Mutation in the HNF1B gene (189907.0007) was found in 2 Japanese patients with typical late-onset type II diabetes. Mutations in the IRS1 gene (147545) have been found in patients with type II diabetes. Reynisdottir et al. (2003) mapped a susceptibility locus for type II diabetes to chromosome 5q34-q35.2 (NIDDM4; 608036). A missense mutation in the AKT2 gene (164731.0001) caused autosomal dominant type II diabetes in 1 family. A SNP in the 3-prime untranslated region of the resistin gene (605565.0001) was associated with susceptibility to diabetes and to insulin resistance-related hypertension in Chinese subjects. Susceptibility to insulin resistance has been associated with polymorphism in the TCF1 (142410.0011), PPP1R3A (600917.0001), PTPN1 (176885.0001), ENPP1 (173335.0006), IRS1 (147545.0002), and EPHX2 (132811.0001) genes. The K121Q polymorphism of ENPP1 (173335.0006) is associated with susceptibility to type II diabetes; a haplotype defined by 3 SNPs of this gene, including K121Q, is associated with obesity, glucose intolerance, and type II diabetes. A SNP in the promoter region of the hepatic lipase gene (151670.0004) predicts conversion from impaired glucose tolerance to type II diabetes. A variant of transcription factor 7-like-2 (TCF7L2; 602228), located on 10q, has also been found to confer risk of type II diabetes (Grant et al., 2006).

Noninsulin-dependent diabetes mellitus is distinct from MODY (606391) in that it is polygenic, characterized by gene-gene and gene-environment interactions with onset in adulthood, usually at age 40 to 60 but occasionally in adolescence if a person is obese. The pedigrees are rarely multigenerational. The penetrance is variable, possibly 10 to 40% (Fajans et al., 2001). Persons with type II diabetes usually have an obese body habitus and manifestations of the so-called metabolic syndrome which is characterized by diabetes, insulin resistance, hypertension, and hypertriglyceridemia.

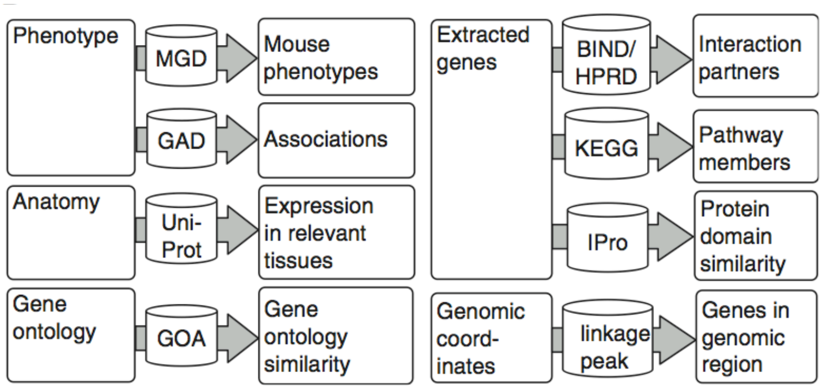
In 3 families with MODY and 7 with 'common' type II diabetes mellitus, O'Rahilly et al. (1992) excluded linkage to the INS locus (176730). Exclusive of the

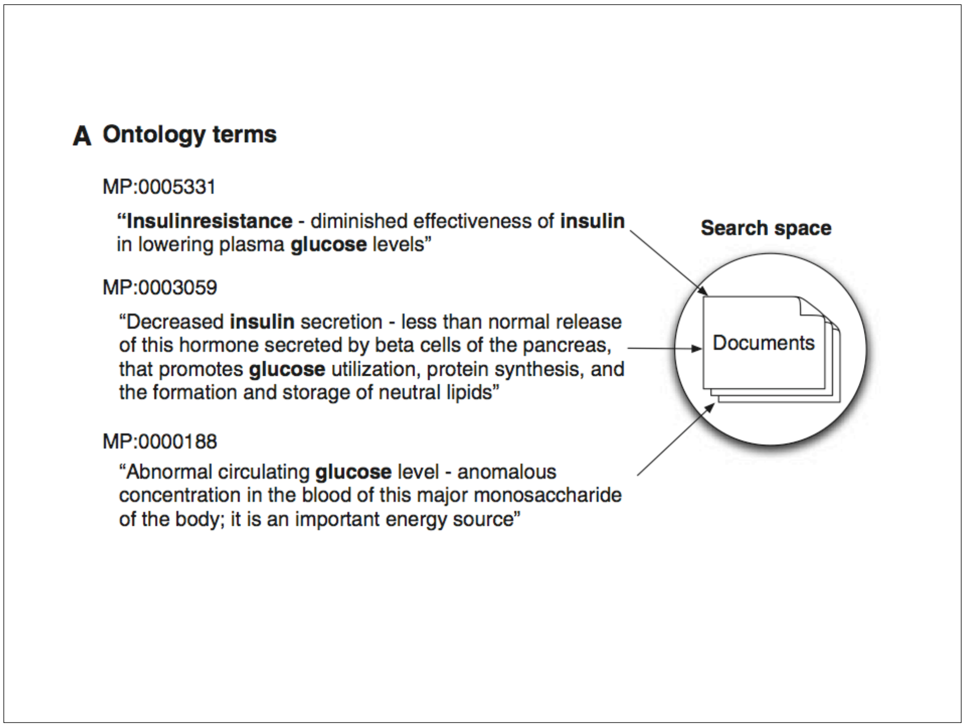
A candidate gene test set



Complex trait	OMIM	Review(s)[a]	Gene[b]
Age-related macular degeneration	603075	15094132; 15350892	*CFH* *LOC387715*
ARMD (second run)	603075	N/A[c]	*C2* *CFB*
Alzheimer's disease	104300	15225164	*LOC439999*
Asthma	600807	12810182; 14551038	*NPSR1*
Autism	209850	11733747; 12142938	*EN2*
Celiac disease	212750	12907013; 12699968; 14592529	*MYO9B*
Myocardial infarction	608446	15861005; 16041318	*LTA4H*
Parkinson's disease	168600	16026116; 16278972	*SEMA5A*
Rheumatoid arthritis	180300	15478157; 12915205	*PTPN22* *FCRL3*
Schizophrenia	181500	15340352; 16033310	*ENTH*
Type 1 diabetes mellitus	222100	12270944; 11921414 11237226; 11899083	*SUMO4* *PTPN22* *IL2RA* *CTLA4*
Type 2 diabetes mellitus	125853	15662000; 15662001; 15662002; 15662003	*TCF7L2*

Applying CAESAR to human complex diseases





A Ontology terms

MP:0005331

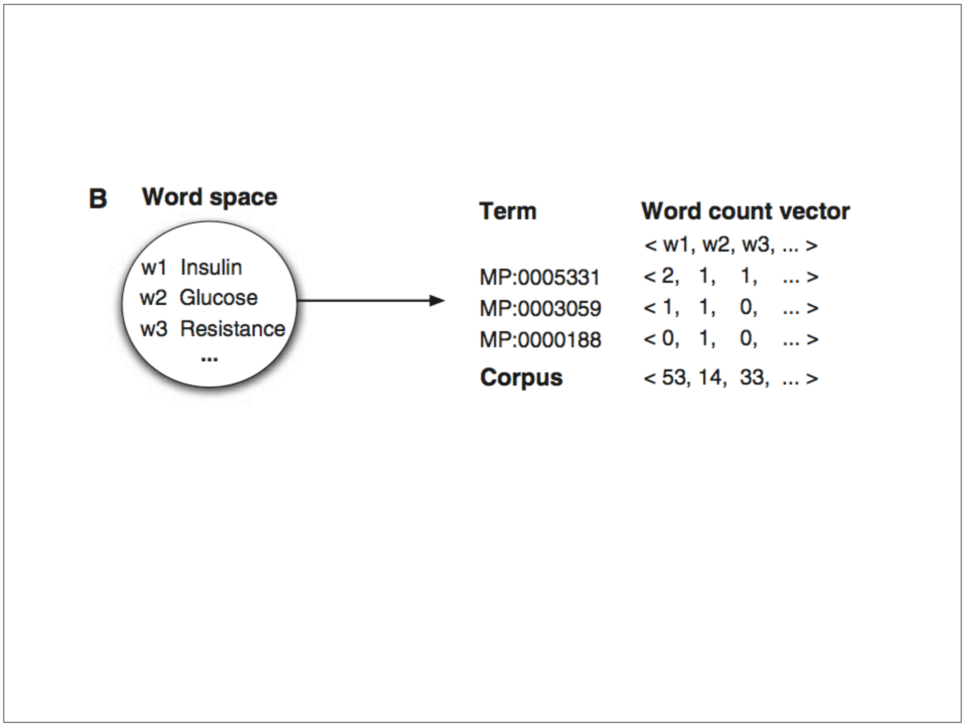
"**Insulinresistance** - diminished effectiveness of **insulin** in lowering plasma **glucose** levels"

MP:0003059

"Decreased **insulin** secretion - less than normal release of this hormone secreted by beta cells of the pancreas, that promotes **glucose** utilization, protein synthesis, and the formation and storage of neutral lipids"

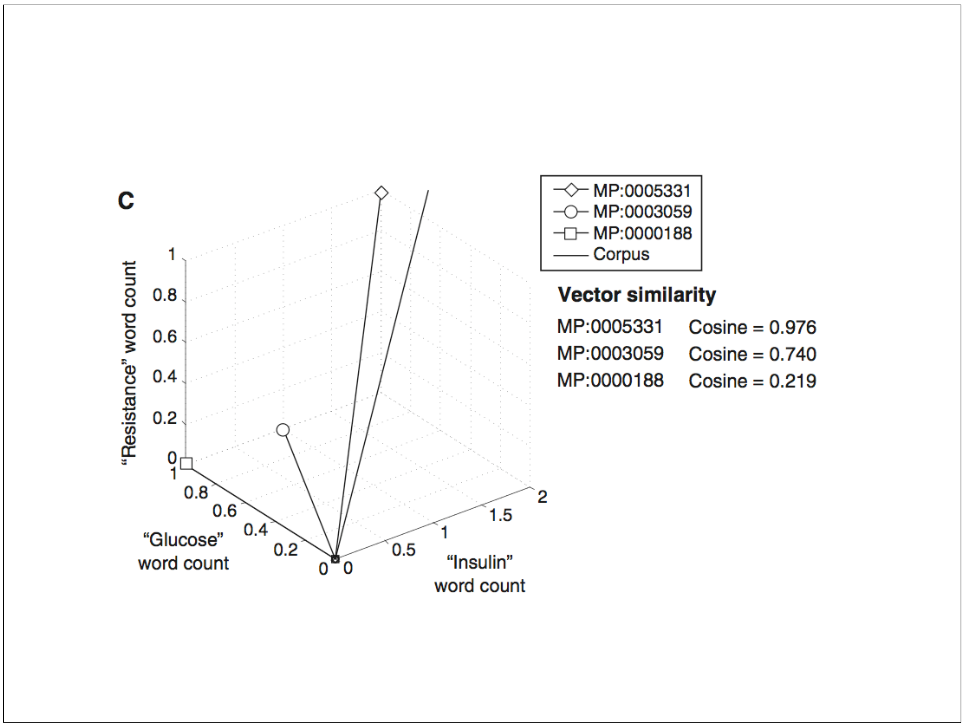
MP:0000188

"Abnormal circulating **glucose** level - anomalous concentration in the blood of this major monosaccharide of the body; it is an important energy source"



B Word space

Term	**Word count vector**
	< w1, w2, w3, ... >
MP:0005331	< 2, 1, 1, ... >
MP:0003059	< 1, 1, 0, ... >
MP:0000188	< 0, 1, 0, ... >
Corpus	< 53, 14, 33, ... >



Vector similarity

MP:0005331	Cosine = 0.976
MP:0003059	Cosine = 0.740
MP:0000188	Cosine = 0.219

Anatomy

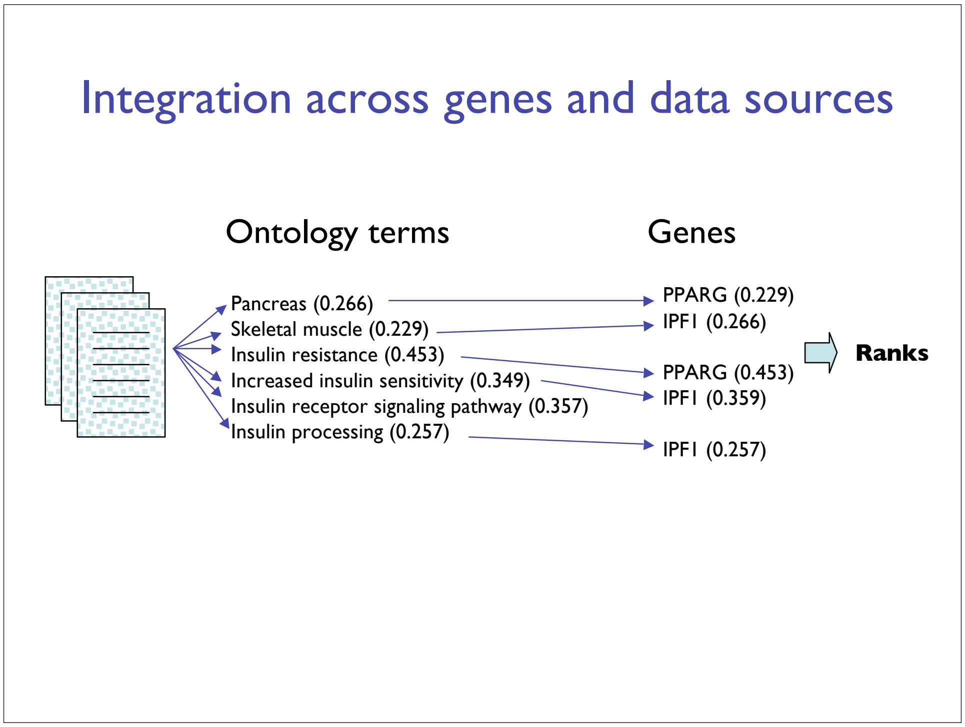
0.426	EV:0100130	pancreatic islets
0.266	EV:0100092	pancreas
0.250	EV:0100089	liver
0.229	EV:0100377	skeletal muscle
0.217	EV:0100381	adipose tissue

Phenotype

0.453	MP:0005331	insulin resistance
0.416	MP:0002056	Non-insulin dependent diabetes
0.349	MP:0002891	increased insulin sensitivity
0.341	MP:0002727	decreased circulating insulin level
0.277	MP:0005293	impaired glucose tolerance

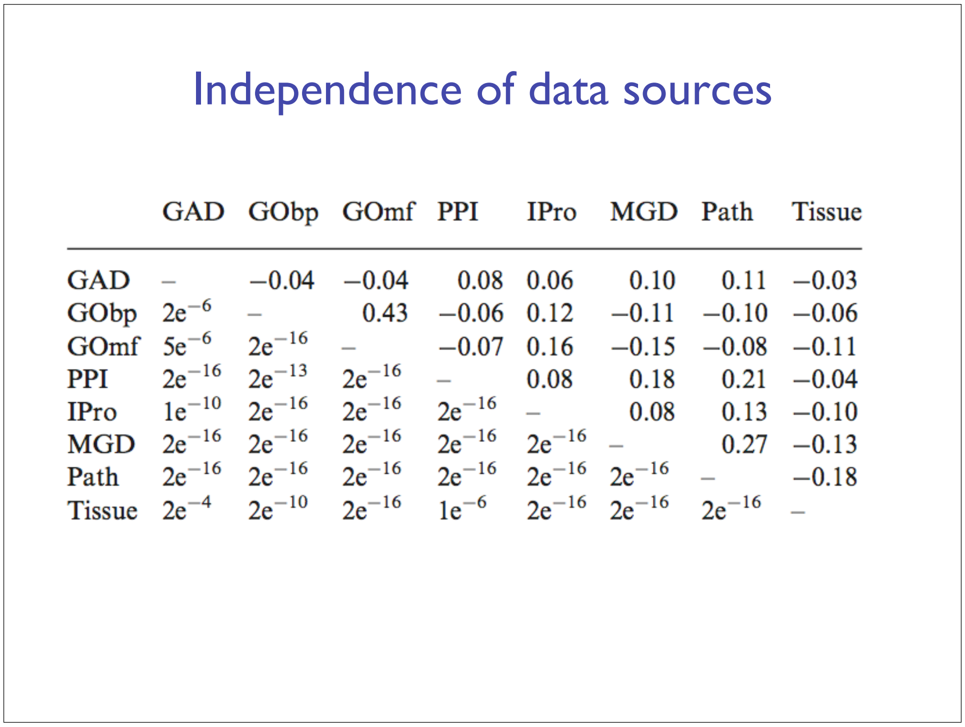
GO Biological function

0.357	GO:0008286	insulin receptor signaling pathway
0.296	GO:0046628	positive regulation of insulin receptor signaling pathway
0.288	GO:0046627	negative regulation of insulin receptor signaling pathway
0.279	GO:0046676	negative regulation of insulin secretion
0.257	GO:0030070	insulin processing



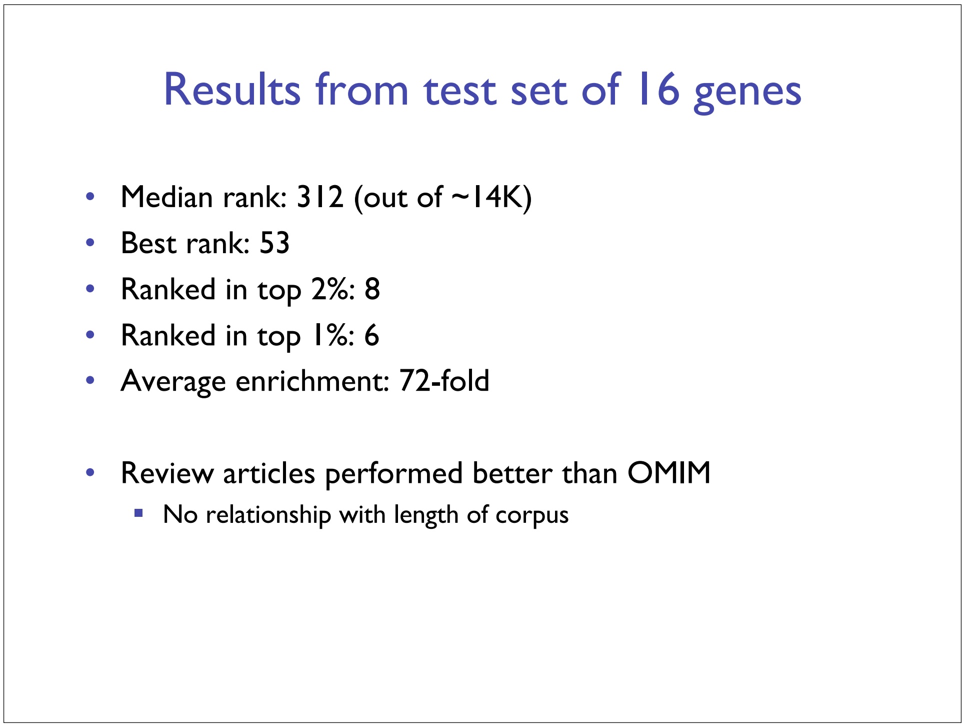
Integration across genes and data sources

Ontology terms
Genes
Pancreas (0.266)
Skeletal muscle (0.229)
Insulin resistance (0.453)
Increased insulin sensitivity (0.349)
Insulin receptor signaling pathway (0.357)
Insulin processing (0.257)
PPARG (0.229)
IPF1 (0.266)
PPARG (0.453)
IPF1 (0.359)
IPF1 (0.257)
Ranks



Independence of data sources

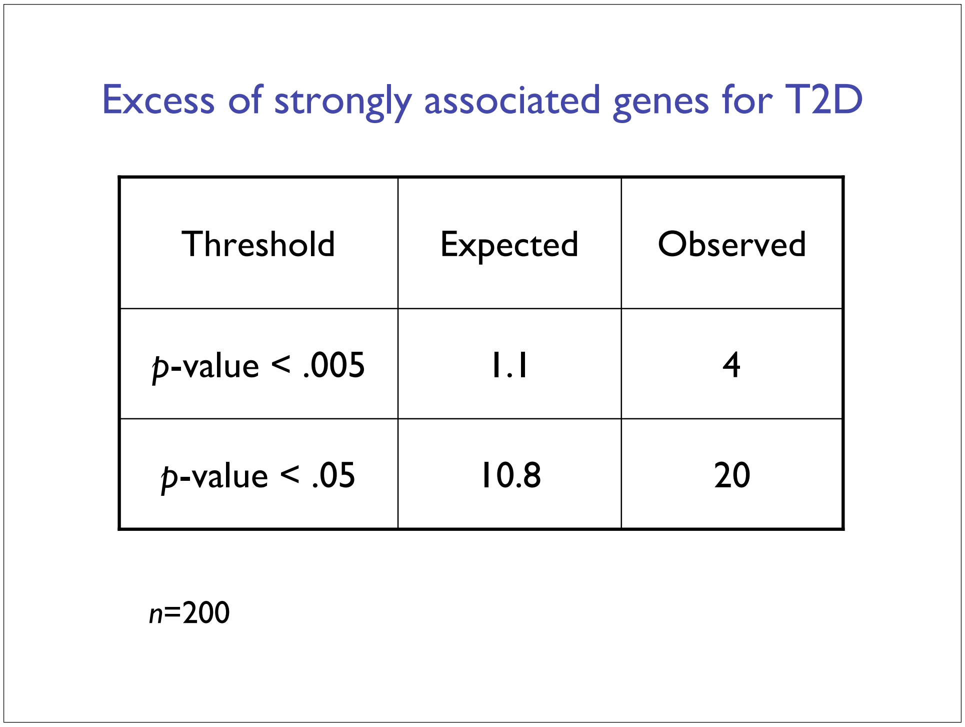
	GAD	GObp	GOmf	PPI	IPro	MGD	Path	Tissue
GAD	–	−0.04	−0.04	0.08	0.06	0.10	0.11	−0.03
GObp	$2e^{-6}$	–	0.43	−0.06	0.12	−0.11	−0.10	−0.06
GOmf	$5e^{-6}$	$2e^{-16}$	–	−0.07	0.16	−0.15	−0.08	−0.11
PPI	$2e^{-16}$	$2e^{-13}$	$2e^{-16}$	–	0.08	0.18	0.21	−0.04
IPro	$1e^{-10}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	–	0.08	0.13	−0.10
MGD	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	–	0.27	−0.13
Path	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	–	−0.18
Tissue	$2e^{-4}$	$2e^{-10}$	$2e^{-16}$	$1e^{-6}$	$2e^{-16}$	$2e^{-16}$	$2e^{-16}$	–



Results from test set of 16 genes

- Median rank: 312 (out of ~14K)
- Best rank: 53
- Ranked in top 2%: 8
- Ranked in top 1%: 6
- Average enrichment: 72-fold

- Review articles performed better than OMIM
  - No relationship with length of corpus



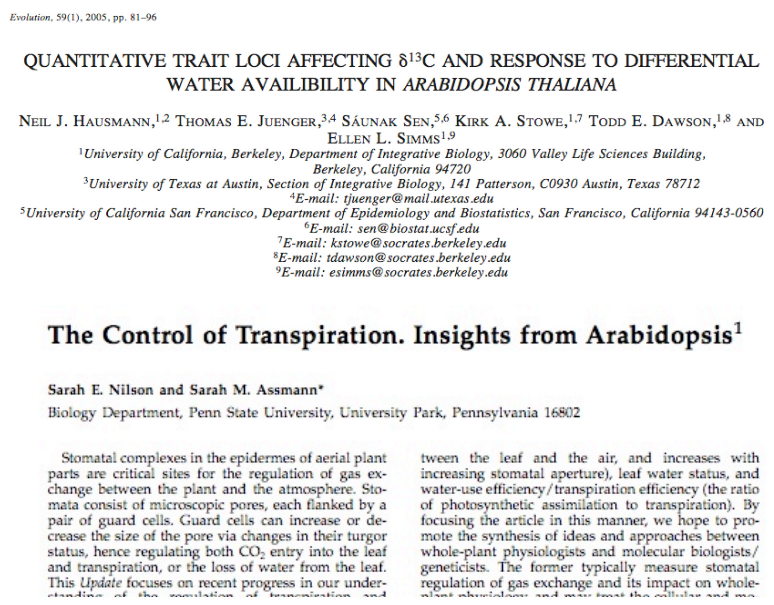
Excess of strongly associated genes for T2D

Threshold	Expected	Observed
p-value < .005	1.1	4
p-value < .05	10.8	20

n=200

Can this be applied to traits of evolutionary importance?

- The basic requirements are the same
  - Published knowledge about the trait
  - One or more model organisms with comparable biology
  - A gene set
- If there is enough data for a human to select candidates, then a computer can do it
  - Although traits like fitness may be tough…



Evolution, 59(1), 2005, pp. 81–96

QUANTITATIVE TRAIT LOCI AFFECTING $\delta^{13}C$ AND RESPONSE TO DIFFERENTIAL WATER AVAILIBILITY IN *ARABIDOPSIS THALIANA*

NEIL J. HAUSMANN,[1,2] THOMAS E. JUENGER,[3,4] SÁUNAK SEN,[5,6] KIRK A. STOWE,[1,7] TODD E. DAWSON,[1,8] AND ELLEN L. SIMMS[1,9]
[1]*University of California, Berkeley, Department of Integrative Biology, 3060 Valley Life Sciences Building, Berkeley, California 94720*
[3]*University of Texas at Austin, Section of Integrative Biology, 141 Patterson, C0930 Austin, Texas 78712*
[4]*E-mail: tjuenger@mail.utexas.edu*
[5]*University of California San Francisco, Department of Epidemiology and Biostatistics, San Francisco, California 94143-0560*
[6]*E-mail: sen@biostat.ucsf.edu*
[7]*E-mail: kstowe@socrates.berkeley.edu*
[8]*E-mail: tdawson@socrates.berkeley.edu*
[9]*E-mail: esimms@socrates.berkeley.edu*

The Control of Transpiration. Insights from Arabidopsis[1]

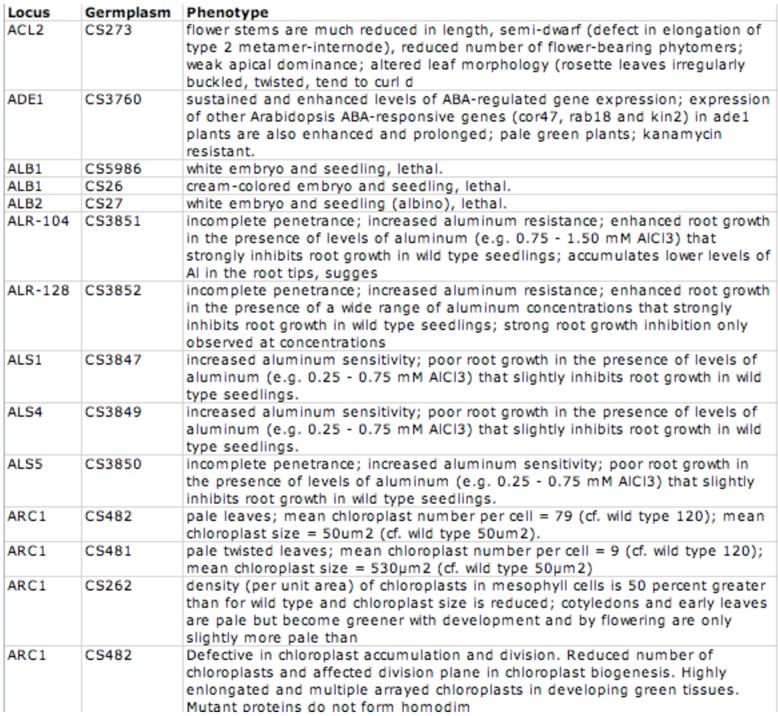
Sarah E. Nilson and Sarah M. Assmann*
Biology Department, Penn State University, University Park, Pennsylvania 16802

Stomatal complexes in the epidermes of aerial plant parts are critical sites for the regulation of gas exchange between the plant and the atmosphere. Stomata consist of microscopic pores, each flanked by a pair of guard cells. Guard cells can increase or decrease the size of the pore via changes in their turgor status, hence regulating both CO_2 entry into the leaf and transpiration, or the loss of water from the leaf. This *Update* focuses on recent progress in our understanding of the regulation of transpiration and

tween the leaf and the air, and increases with increasing stomatal aperture), leaf water status, and water-use efficiency/transpiration efficiency (the ratio of photosynthetic assimilation to transpiration). By focusing the article in this manner, we hope to promote the synthesis of ideas and approaches between whole-plant physiologists and molecular biologists/geneticists. The former typically measure stomatal regulation of gas exchange and its impact on whole-plant physiology and may treat the cellular and mo-

What is different in Arabidopsis?

- We don’t know the answer
- Most data-types are the same or have counterparts
  - Protein domains
  - Protein-protein interactions
  - Biochemical pathways
  - Anatomical information about expression
- Mutant phenotypes have not historically been described using ontologies
- Plant Ontology
  - Plant structure
  - Growth and developmental stages



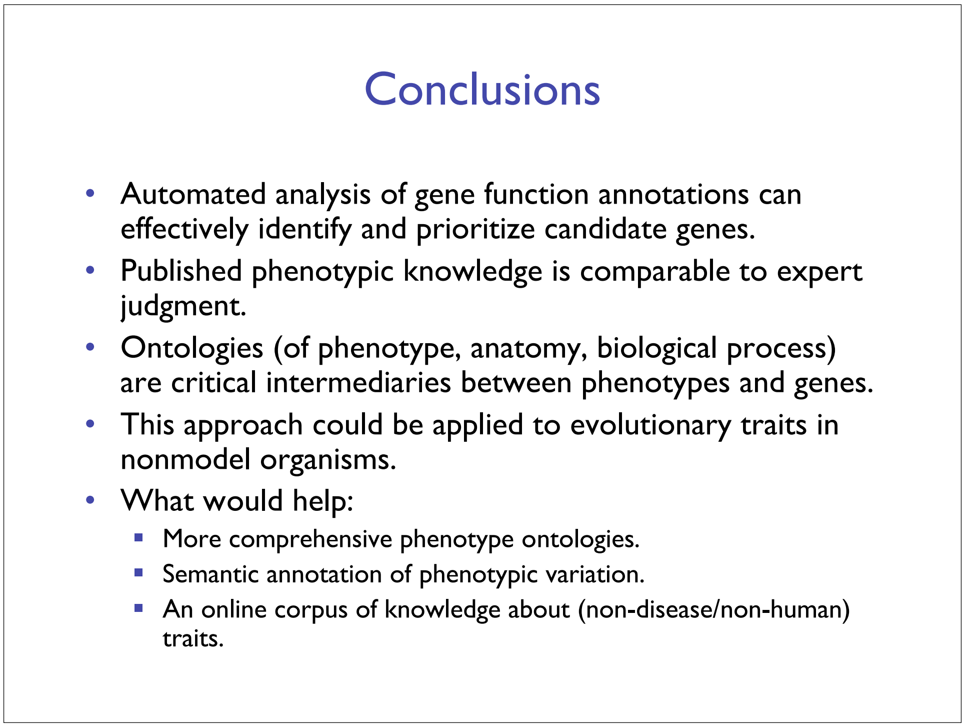
Locus	Germplasm	Phenotype
ACL2	CS273	flower stems are much reduced in length, semi-dwarf (defect in elongation of type 2 metamer-internode), reduced number of flower-bearing phytomers; weak apical dominance; altered leaf morphology (rosette leaves irregularly buckled, twisted, tend to curl d
ADE1	CS3760	sustained and enhanced levels of ABA-regulated gene expression; expression of other Arabidopsis ABA-responsive genes (cor47, rab18 and kin2) in ade1 plants are also enhanced and prolonged; pale green plants; kanamycin resistant.
ALB1	CS5986	white embryo and seedling, lethal.
ALB1	CS26	cream-colored embryo and seedling, lethal.
ALB2	CS27	white embryo and seedling (albino), lethal.
ALR-104	CS3851	incomplete penetrance; increased aluminum resistance; enhanced root growth in the presence of levels of aluminum (e.g. 0.75 - 1.50 mM AlCl3) that strongly inhibits root growth in wild type seedlings; accumulates lower levels of Al in the root tips, sugges
ALR-128	CS3852	incomplete penetrance; increased aluminum resistance; enhanced root growth in the presence of a wide range of aluminum concentrations that strongly inhibits root growth in wild type seedlings; strong root growth inhibition only observed at concentrations
ALS1	CS3847	increased aluminum sensitivity; poor root growth in the presence of levels of aluminum (e.g. 0.25 - 0.75 mM AlCl3) that slightly inhibits root growth in wild type seedlings.
ALS4	CS3849	increased aluminum sensitivity; poor root growth in the presence of levels of aluminum (e.g. 0.25 - 0.75 mM AlCl3) that slightly inhibits root growth in wild type seedlings.
ALS5	CS3850	incomplete penetrance; increased aluminum sensitivity; poor root growth in the presence of levels of aluminum (e.g. 0.25 - 0.75 mM AlCl3) that slightly inhibits root growth in wild type seedlings.
ARC1	CS482	pale leaves; mean chloroplast number per cell = 79 (cf. wild type 120); mean chloroplast size = 50um2 (cf. wild type 50um2).
ARC1	CS481	pale twisted leaves; mean chloroplast number per cell = 9 (cf. wild type 120); mean chloroplast size = 530µm2 (cf. wild type 50µm2)
ARC1	CS262	density (per unit area) of chloroplasts in mesophyll cells is 50 percent greater than for wild type and chloroplast size is reduced; cotyledons and early leaves are pale but become greener with development and by flowering are only slightly more pale than
ARC1	CS482	Defective in chloroplast accumulation and division. Reduced number of chloroplasts and affected division plane in chloroplast biogenesis. Highly enlongated and multiple arrayed chloroplasts in developing green tissues. Mutant proteins do not form homodim

Candidates for the two water use efficiency QTL in Arabidopsis

- Genes ranked in the top 100 within 1Mb of the best-linked marker
  - Expected <1, we obtained 4
- g6842
  - AT2G01830 (rank 53) Histidine kinase: cytokinin-binding receptor that transduces cytokinin signals across the plasma membrane, osmosensor activity, response to water deprivation.
- mi357
  - AT3G11410 (rank 19) Protein phosphatase 2C. Negative regulator of ABA signalling. Up-regulated by drought and ABA.
  - AT3G06120 (rank 41) bHLH protein that controls meristemoid differentiation during stomatal development. In the absence of MUTE, meristemoids fail to differentiate stomata.
  - AT3G11020 (rank 86) DREB2B transcription factor, involved in response to water deprivation, heat acclimation.

Why use a computer rather than human judgement?

- Advantages
  - Principled, repeatable, automated, fast
  - Allows for an element of surprise
  - Predictions will improve as phenotypic and functional genomic data grows
  - Integration process could be optimized by training against known genes
  - The evidence for each prediction can be inspected after the fact
- Disadvantages
  - Biases and gaps due to variable and incomplete annotation
  - We won’t find genes that are totally out of left field
  - Computers cannot exercise discretion



Conclusions

- Automated analysis of gene function annotations can effectively identify and prioritize candidate genes.
- Published phenotypic knowledge is comparable to expert judgment.
- Ontologies (of phenotype, anatomy, biological process) are critical intermediaries between phenotypes and genes.
- This approach could be applied to evolutionary traits in nonmodel organisms.
- What would help:
  - More comprehensive phenotype ontologies.
  - Semantic annotation of phenotypic variation.
  - An online corpus of knowledge about (non-disease/non-human) traits.