Using Ontology Fingerprints to Evaluate Genome-wide Association Results

Lam Tsoi, Michael Boehnke, Richard Klein, W. Jim Zheng

Medical University of South Carolina

ICBO, Buffalo, 2009

Overview

- Genome-wide association study
- Ontology fingerprints
- Using ontology fingerprints to quantify the relationship between genes and disease/phenotypes/traits
- Ontology fingerprints derived gene networks to identify polygenic model for diseases

Genome-wide Association Study

GWA Studies In Action

Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls

The W

Sequence variants in the autophagy gene *IRGM* and multiple other replicating loci cont

susc

Miles Pa
Mark Tr
Roland

We followed up on 37 SNPs from 31 distinct loci, associated at $P < 10^{-5}$ on initial analysis of the WTCCC data set. Support for some of these markers diminished in the final WTCCC analysis after extensive data filtering[5]. We selected two markers for each locus where low linkage disequilibrium (LD) between associated SNPs in

Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes

John A Todd[1], Neil M Walker[1,9], Jason D Cooper[1,9], Deborah J Smyth[1,9], Kate Downes[1], Vincent Plagnol[1], Rebecca Bailey[1], Sergey Nejentsev[1], Sarah F Field[1], Felicity Payne[1], Christopher E Lowe[1], Jeffrey S Szeszko[1], Jason P Hafler[1], Lauren Zeitels[1], Jennie H M Yang[1], Adrian Vella[1,8], Sarah Nutland[1], Helen E Stevens[1], Helen Schuilenburg[1], Gillian Coleman[1], Meeta Maisuria[1], William Meadows[1], Luc J Smink[1], Barry Healy[1], Oliver S Burren[1], Alex A C Lam[1], Nigel R Ovington[1], James Allen[1], Ellen Adlem[1], Hin-Tak Leung[1], Chris Wallace[2], Joanna M M Howson[1], Cristian Guja[3], Constantin Ionescu-Tîrgovişte[3], Genetics of Type 1 Diabetes in Finland[4], Matthew J Simmonds[5], Joanne M Heward[5], Stephen C L Gough[5], The Wellcome Trust Case Control Consortium[6], David B Dunger[7], Linda S Wicker[1] & David G Clayton[1]

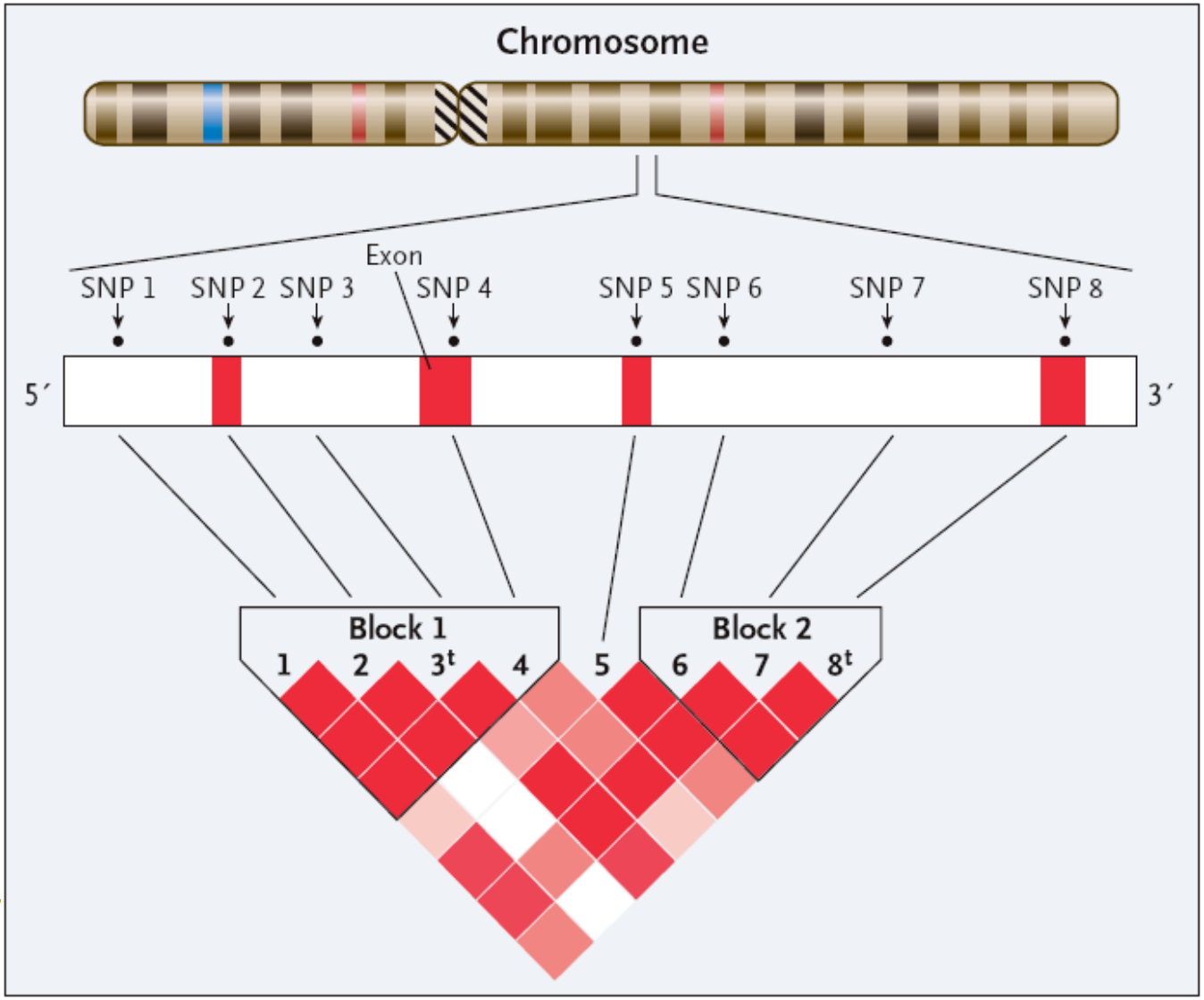
What is a GWA Study?

- A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease. Once new genetic associations are identified, researchers can use the information to develop better strategies to detect, treat and prevent the disease. Such studies are particularly useful in finding genetic variations that contribute to common, complex diseases, such as asthma, cancer, diabetes, heart disease and mental illnesses

What is a GWA Study?

- Method for interrogating all 10 million variable points across human genome
- Variation inherited in groups, or blocks, so not all 10 million points have to be tested

Linkage Disequilibrium Blocks Can Have Many Genes



nature genetics

Common variants near *MC4R* are associated with fat mass, weight and risk of obesity

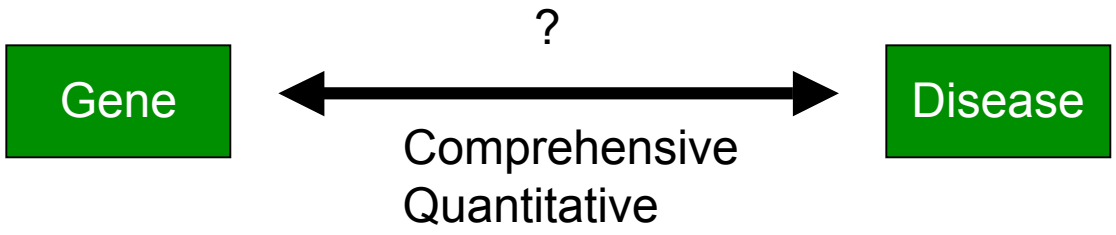
Ruth J F Loos[*,1,2,73], Cecilia M Lindgren[3,4,73], Shengxu Li[1,2,73], Eleanor Wheeler[5], Jing Hua Zhao[1,2], Inga Prokopenko[3,4], Michael Inouye[5], Rachel M Freathy[6,7], Antony P Attwood[5,8], Jacques S Beckmann[9,10], Sonja I Berndt[11], The Prostate, Lung, Colorectal, and Ovarian (PLCO) Cancer Screening Trial[71], Sven Bergmann[9,12], Amanda J Bennett[3,4], Sheila A Bingham[13], Murielle Bochud[14], Morris Brown[15], Stéphane Cauchi[16], John M Connell[17], Cyrus Cooper[18], George Davey Smith[19], Ian Day[18], Christian Dina[16], Subhajyoti De[20], Emmanouil T Dermitzakis[5], Alex S F Doney[21], Katherine S Elliott[3], Paul Elliott[22,23], David M Evans[3,19], I Sadaf Farooqi[2,24], Philippe Froguel[16,25], Jilur Ghori[5], Christopher J Groves[3,4], Rhian Gwilliam[5], David Hadley[26], Alistair S Hall[27], Andrew T Hattersley[6,7], Johannes Hebebrand[28], Iris M Heid[29,30], KORA[71], Blanca Herrera[3,4], Anke Hinney[28], Sarah E Hunt[5], Marjo-Riitta Jarvelin[22,23,31], Toby Johnson[9,12,14], Jennifer D M Jolley[8], Fredrik Karpe[4], Andrew Keniry[5], Kay-Tee Khaw[32], Robert N

Genome Wide Association Study for LDL, HDL and TG

- 16876 individuals
- Clinical observations considered – Height, weight, BMI, LDL, HDL and TG level etc

Genome Wide Association Study for LDL, HDL and TG

- Identify genes that falls into the loci that are significantly associated with phenotype
  - HDL – 237 genes from top 201 LD blocks
  - LDL -- 212 genes from top 199 LD blocks
  - TG -- 221 genes from top 200 LD blocks
- Challenge – Which genes are more relevant?



Ontology Fingerprints

Biomedical Ontology

- Many ontologies have been developed:
  - Gene Ontology
  - Cell Ontology
  - Foundation Model of Anatomy
  - Disease Ontology
  - …

Annotation of Apolipoprotein A-4

Function	Evidence
antioxidant activity	IDA PubMed
cholesterol transporter activity	IDA PubMed
copper ion binding	IDA PubMed
contributes_to eukaryotic cell surface binding	IDA PubMed
lipid binding	IEA
lipid transporter activity	TAS PubMed
phosphatidylcholine binding	IDA PubMed
phosphatidylcholine-sterol O-acyltransferase activator activity	IDA PubMed
protein homodimerization activity	IDA PubMed

Gene annotation by Gene Ontology has been used extensively by microarray data analysis.

To assess the relevance of genes to the disease of interest, we need a quantitative measure.

Our Approach

- Text mining approach to identify ontology terms enriched in the PubMed Abstracts that relevant to a particular gene or disease to generate an ontology finger prints
- Assess how similar it is between the ontology fingerprints of a gene and a disease
- Rank identify genes based on the similarity of their ontology fingerprints to disease for GWAS

Hypergeometric Test

	# Abstracts Relevant to a Gene	# Abstracts Irrelevant to a Gene	Total
Abstracts with a specific term	X	K-X	K
Abstracts without a specific term	M-X	N-K+X	M+N-K

Hypergeometric Test

$$P(X)= \frac{\binom{M}{X}\binom{N}{K-X}}{\binom{M+N}{K}}$$

Basic Idea – when research papers are published about a gene, what ontology terms they talk about the most?

Ontology Fingerprint

- A set of ontology terms overrepresented in the PubMed abstracts linked to a gene or a disease along with these terms' corresponding enrichment p-values
- A comprehensive characterization of genes and diseases

Ontology Fingerprints after p-value Adjustment

>Apolipoprotein C-II, *APOC2*

GO id	GO term	Raw p-value	Adjusted p-value
GO:0016298	Lipase activity	8×10^{-22}	9×10^{-22}
GO:0004091	Carboxylesterase activity	5×10^{-21}	6×10^{-21}
GO:0042627	Chylomicron	4×10^{-16}	4×10^{-16}
...	...	...	...
GO:0007610	Behavior	6×10^{-2}	8×10^{-2}
GO:0003708	Retinoic acid receptor activity	6×10^{-2}	9×10^{-2}
...	...	...	...
GO:0044464	Cell part	9×10^{-1}	1
GO:0004871	Signal transducer activity	9×10^{-1}	1

Ontology Fingerprints for HDL

- >PATH#HDL
- GO#GO_0033344 1e-323 cholesterol efflux
- GO#GO_0016298 1e-323 lipase activity
- …
- GO#GO_0030301 1e-323 cholesterol transport
- GO#GO_0015918 1e-323 sterol transport
- GO#GO_0005323 1e-323 very-low-density lipoprotein
- GO#GO_0005322 1e-323 low-density lipoprotein
- GO#GO_0005321 1e-323 high-density lipoprotein
- ….
- GO#GO_0006810 1.54742e-310 transport
- GO#GO_0051234 2.16484e-296 establishment of localization
- GO#GO_0030228 7.60472e-248 lipoprotein receptor activity
- GO#GO_0042697 4.32514e-243 menopause

Can we really use ontology fingerprints to identify genes relevant to a trait/phenotype/disease?

Comparing Ontology Fingerprints

>ATP-binding cassette, sub-family A, member 1 (*ABCA1*)

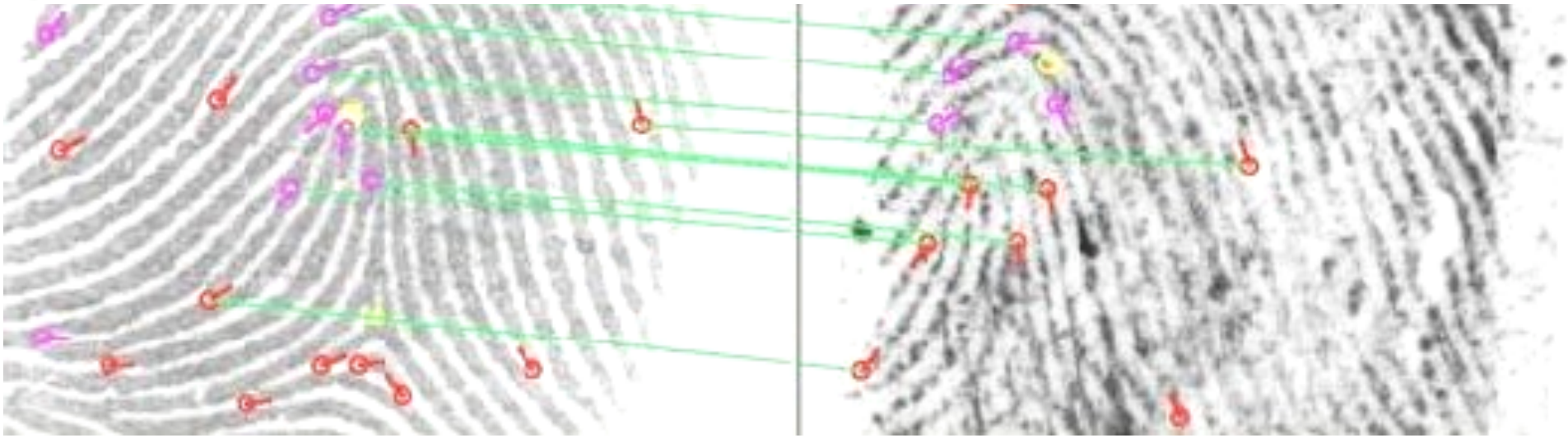
GO#GO_0030301	1.87E-122	CHOLESTEROL TRANSPORT
GO#GO_0005215	4.89E-117	TRANSPORTER
GO#GO_0033344	2.42E-114	CHOLESTEROL EFFLUX
GO#GO_0006810	7.89E-93	TRANSPORT
GO#GO_0005320	7.01E-80	APOLIPOPROTEIN
GO#GO_0051234	7.73E-75	ESTABLISHMENT OF LOCALIZATION

>HDL

GO#GO_0033344	0	cholesterol efflux
…	…	…
GO#GO_0030301	0	cholesterol transport
GO#GO_0015918	0	sterol transport
….		
GO#GO_0006810	1.54E-310	transport
GO#GO_0051234	2.16E-296	establishment of localization

Comparing Ontology Fingerprints

$$S_j = \frac{\sum_{i=1}^{O} log(q_{io}) log(r_{ij})}{max\left\{1, \sum_{i=1}^{O} \left[I(q_i < 1) I(r_{ij} = 1)\right]\right\}}$$



Relevance between a gene and a pathway can be quantified

Prostate Cancer Pathway

	Gene name	Gene symbol	Similarity Score
Kegg Genes	Mitogen-activated protein kinase 1	*MAPK1*	481.8
	BCL2-antagonist of cell death	*BAD*	194.76
	Serum response factor	*SRF*	260.07
	Vascular endothelial growth factor A	*VEGFA*	2341.19
	Caspase 9, apoptosis-related cysteine peptidase	*CASP9*	370.94
Non-Kegg Genes	Splicing factor proline/glutamine-rich	*SFPQ*	13.82
	EP300 interacting inhibitor of differentiation 2B	*EID2B*	1.67
	Ring finger and CCCH-type zinc finger domains 1	*RC3H2*	0
	Stathmin-like 4	*STMN4*	0.72
	Sperm flagellar 1	*SPEF1*	0.18

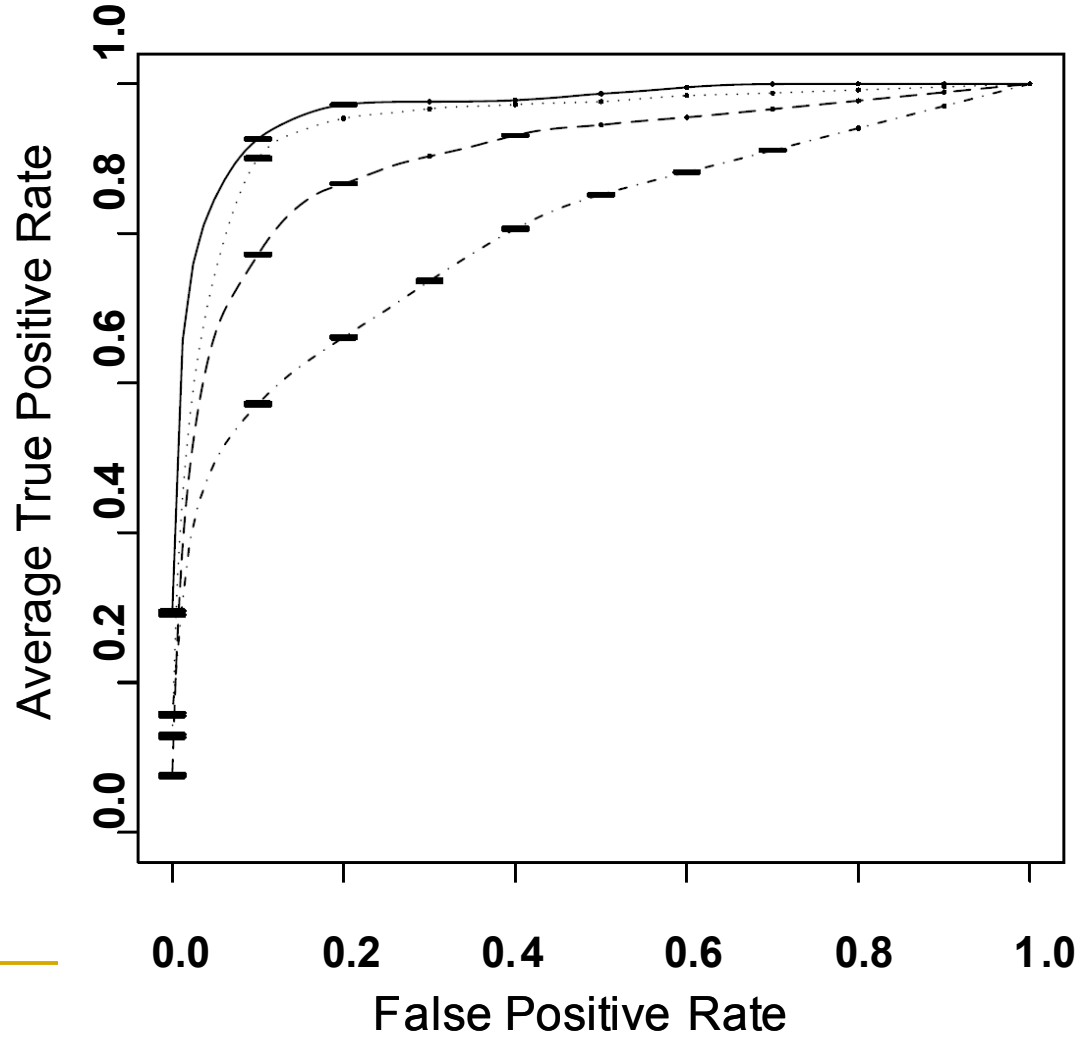
Validation – Using Ontology Fingerprints to Identify Pathways a Gene Belongs to?

- 10 different KEGG pathways
- For each pathway, identify human genes belong to the pathway – positive control
- Identify 3 pathways specific to bacterial
- For each bacterial pathway, identify bacterial genes that has no homolog in human – negative control
- Use ontology fingerprints for genes and pathways to pair genes with pathways

Area Under Curve for Ten pathways

Pathway	Ontology Fingerprint AUC	Anni 2.0 AUC
Apoptosis	0.96	0.85*
Biosynthesis of steroids	0.75	0.73
Fatty acid metabolism	0.88	0.86
Focal Adhesion	0.94	0.87*
Galactose metabolism	0.90	0.78*
Glycolysis	0.80	0.72*
MAP kinase signaling pathway	0.90	0.78*
Prostate cancer	0.95	0.91*
Renal cell carcinoma	0.93	0.81*
Sphingolipid metabolism	0.89	0.72*

Receiver Operating Characteristic (ROC) Curves for Four Pathways



Genes Ranked for HDL

Gene Id	Score	Annotation
4023	9808.26	lipoprotein lipase
19	9741.04	ATP-binding cassette, sub-family A (ABC1), member 1
348	6772.21	apolipoprotein E
3949	4332.74	low density lipoprotein receptor (familial hypercholesterolemia)
338	3973.8	apolipoprotein B (including Ag(x) antigen)

Genes Ranked for HDL

Gene Id	Score	Annotation
4023	9808.26	lipoprotein lipase
19	9741.04	ATP-binding cassette, sub-family A (ABC1), member 1
348	6772.21	apolipoprotein E
3949	4332.74	low density lipoprotein receptor (familial hypercholesterolemia)
338	3973.8	apolipoprotein B (including Ag(x) antigen)
1071	2830.4	cholesteryl ester transfer protein, plasma
3990	2725.8	lipase, hepatic
5465	2380.38	peroxisome proliferator-activated receptor alpha
344	1950.58	apolipoprotein C-II
4036	1615.41	low density lipoprotein-related protein 2
4043	1443.38	low density lipoprotein receptor-related protein associated protein 1
7520	1358.5	X-ray repair complementing defective repair in Chinese hamster cells 5 (double-strand-break rejoining; Ku autoantigen, 80kDa)
1742	1145.05	discs, large homolog 4 (Drosophila)
4089	1142.52	SMAD family member 4
5371	1103.41	promyelocytic leukemia
7068	1092.12	thyroid hormone receptor, beta (erythroblastic leukemia viral (v-erb-a) oncogene homolog 2, avian)
1356	1085.79	ceruloplasmin (ferroxidase)
116519	1084.99	apolipoprotein A-V
3569	1061.43	interleukin 6 (interferon, beta 2)
64240	986.802	ATP-binding cassette, sub-family G (WHITE), member 5 (sterolin 1)
2113	887.013	v-ets erythroblastosis virus E26 oncogene homolog 1 (avian)
64241	877.776	ATP-binding cassette, sub-family G (WHITE), member 8 (sterolin 2)
7253	829.328	thyroid stimulating hormone receptor
6564	827.674	solute carrier family 15 (oligopeptide transporter), member 1
3146	773.341	high-mobility group box 1
2237	728.375	flap structure-specific endonuclease 1
341	721.345	apolipoprotein C-I
6678	693.686	secreted protein, acidic, cysteine-rich (osteonectin)
1600	686.358	disabled homolog 1 (Drosophila)
255738	639.314	proprotein convertase subtilisin/kexin type 9

Transferrin and Lipid Metabolism

Megalin-dependent cubilin-mediated endocytosis is a major pathway for the apical uptake of transferrin in polarized epithelia

Renata Kozyraki*, John Fyfe†, Pierre J. Verroust‡, Christian Jacobsen*, Alice Dautry-Varsat§, Jakub Gburek¶, Thomas E. Willnow‖, Erik Ilsø Christensen¶, and Søren K. Moestrup*,**

Departments of *Medical Biochemistry and ¶Cell Biology, Institute of Anatomy, University of Aarhus, DK-8000 Aarhus, Denmark; ‡Institut National de la Santé et de la Recherche Médicale U538, Centre Hospitalo Universitaire, St. Antoine, 75012 Paris, France; †Department of Microbiology, Michigan State University, East Lansing, MI 48824; §Institut Pasteur, 75015 Paris, France; and ‖Max-Delbrueck Center for Molecular Medicine, 13125 Berlin, Germany

Edited by Stuart A. Kornfeld, Washington University School of Medicine, St. Louis, MO, and approved August 21, 2001 (received for review June 11, 2001)

Cubilin is a 460-kDa protein functioning as an endocytic receptor for intrinsic factor vitamin B_{12} complex in the intestine and as a receptor for apolipoprotein A1 and albumin reabsorption in the kidney proximal tubules and the yolk sac. In the present study, we report the identification of cubilin as a novel transferrin (Tf) receptor involved in catabolism of Tf. Consistent with a cubilin-mediated endocytosis of Tf in the kidney, lysosomes of human, dog, and mouse renal proximal tubules strongly accumulate Tf, whereas no Tf is detectable in the endocytic apparatus of the renal tubule epithelium of dogs with deficient surface expression of cubilin. As a consequence, these dogs excrete increased amounts of Tf in the urine. Mice with deficient synthesis of megalin, the putative coreceptor colocalizing with cubilin, also excrete high amounts of Tf and fail to internalize Tf in their proximal tubules. However, in contrast to the dogs with the defective cubilin expression, the megalin-deficient mice accumulate Tf on the luminal

Using a cubilin-affinity approach, we discovered Tf as a novel ligand to cubilin. Subsequent investigations of the receptor-mediated uptake of Tf in the renal proximal tubules and in cultured yolk cells demonstrate that cubilin is a physiological and quantitatively important third Tf receptor involved in Tf catabolism and Fe^{3+} uptake. Furthermore, this discovery made it possible to establish that the cubilin internalization depends on megalin.

Materials and Methods

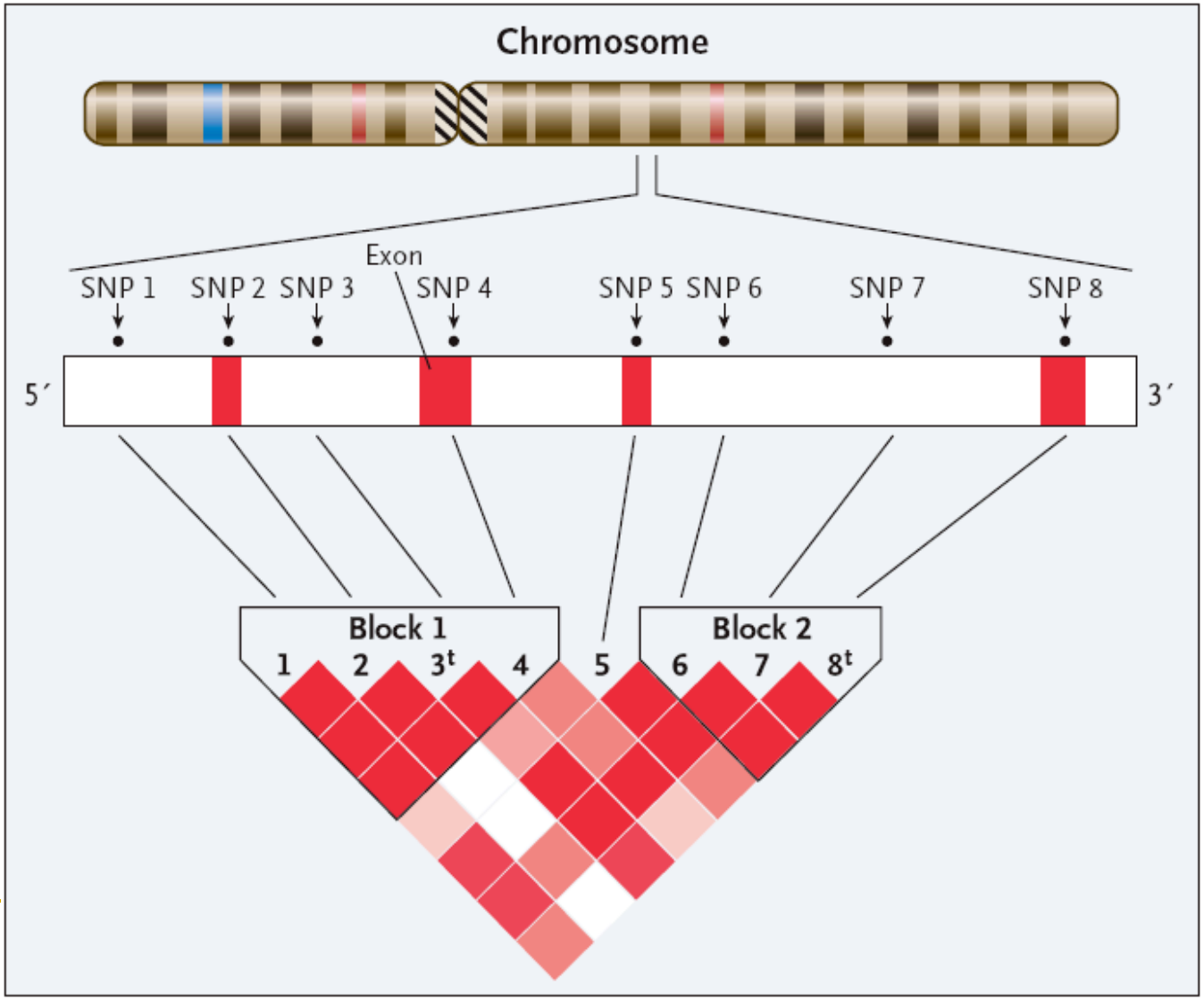
Receptors, Antibodies, and Ligands. Cubilin and megalin were purified from solubilized rabbit and human renal cortex as described (11). Tf was from Calbiochem. Polyclonal and monoclonal antibodies against rat cubilin and megalin have been described (18, 19). Polyclonal antibody against human Tf was from Dako and recognizes human, dog, and mouse Tf. Human

	Gene	GO Term	GO Id
HDL	ATP-binding cassette, sub-family A, member 1 (*ABCA1*)	Cholesterol efflux	GO:0033344
		Transporter activity	GO:0005215
		Reverse cholesterol transport	GO:0043691
	Lipoprotein lipase (*LPL*)	Lipoprotein lipase activity	GO:0004465
		chylomicron	GO:0042627
		Lipid transport	GO:0006869
	Cholesteryl ester transfer protein, plasma (*CETP*)	Reverse cholesterol transport	GO:0043691
		Lipoprotein lipase activity	GO:0004465
		Cholesterol efflux	GO:0033344
LDL	LDL recepor (*LDLR*)	Low-density lipoprotein receptor activity	GO:0005041
		endocytosis	GO:0006897
		Cell surface	GO:0009986
	Apolipoprotein E (*APOE*)	Low-density lipoprotein receptor activity	GO:0005041
		Cholesterol transport	GO:0030301
		Lipoproteinlipase activity	GO:0004465
	Apolipoprotein B (*APOB*)	Lipoprotein receptor activity	GO:0030228
		chylomicron	GO:0042627
		Lipoprotein lipase activity	GO:0004465
Triglyceride	Lipoprotein lipase (*LPL*)	Lipoprotein lipase activity	GO:0004465
		chylomicron	GO:0042627
		digestion	GO:0007586
	Apolipoprotein A-V (*APOA5*)	Lipoprotein lipase activity	GO:0004465
		chylomicron	GO:0042627
		peroxisome	GO:0005777
	Low density lipoprotein-related protein 2 (*LRP2*)	Lipoprotein lipase activity	GO:0004465
		Lipoprotein receptor activity	GO:0030228
		Low-density lipoprotein binding	GO:0030169

Prioritize Genes with Similar p-value from Genome-wide Association Study

Trait	Gene	Best SNP	Best P-value	Similarity Score
LDL	*ABCA1*	rs2000069	2.25×10^{-5}	1133.75
	PEX5	rs10770616	2.29×10^{-5}	118.982
	LGALS1	rs739139	2.24×10^{-5}	48.3371
LDL	*GNAO1*	rs4783937	7.86×10^{-5}	112.176
	SLC36A2	rs10050758	7.89×10^{-5}	4.64188
TG	*TRPC6*	rs4466798	1.01×10^{-4}	129.288
	AXUD1	rs17735402	1.02×10^{-4}	0

Linkage Disequilibrium Blocks Can Have Many Genes



Phenotype	LD Block		Gene Id	Gene	Similarity Score
	Chromosome	Position			
HDL	chr22	44953108	5465	*PPARA*	211.493
			150383	*LOC150383*	0
	chr16	55542264	1071	*CETP*	1473.99
			9709	*HERPUD1*	0
	chr16	55500422	6559	*SLC12A3*	67.2739
			9709	*HERPUD1*	0
LDL	chr19	50087106	348	*APOE*	2824.46
			341	*APOC1*	296.763
			5819	*PVRL2*	95.7221
			10452	*TOMM40*	8.03323
	chr19	50124397	344	*APOC2*	543.665
			341	*APOC1*	296.763
			346	*APOC4*	60.6262
	chr22	36391511	3956	*LGALS1*	48.3371
			57026	*PDXP*	0.12471
			79159	*MGC3731*	0
TG	chr11	116168917	116519	*APOA5*	795.062
			8882	*ZNF259*	0.213755
			84811	*BUD13*	0
	chr1	62756485	27329	*ANGPTL3*	56.8199
			85440	*DOCK7*	0

Lam C. Tsoi, Michael Boehnke, Richard Klein, W. Jim Zheng: Evaluation of Genome-wide Association Study Results through Development of Ontology Fingerprint. Bioinformatics, 2009; 25:1314-1320

Data and text mining

Evaluation of genome-wide association study results through development of ontology fingerprint

Lam C. Tsoi[1], Michael Boehnke[2], Richard L. Klein[3,4] and W. Jim Zheng[5,*]

[1] Bioinformatics Graduate Program, Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, Charleston, SC.

[2] Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of Michigan, Ann Arbor, MI

[3] Division of Endocrinology, Metabolism, and Medical Genetics, Department of Medicine, Medical University of South Carolina

[4] Research Service, Ralph H. Johnson Department of Veterans Affairs Medical Center, Charleston, SC

[5] Department of Biostatistics, Bioinformatics & Epidemiology, Medical University of South Carolina, Charleston, SC

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Genome-wide association (GWA) studies may identify multiple variants that are associated with a disease or trait. To narrow down candidates for further validation, quantitatively assessing how identified genes relate to a phenotype of interest is important.

Results: We describe an approach to characterize genes or biologi-

performed in these studies gives rise to numerous false positive results (Pearson and Manolio, 2008). Therefore, assessing quantitatively the likely importance of genes identified as significant to disease risk based on biological facts is essential to proceed efficiently toward experimental validation processes and, ultimately, to define the causal relationships between genes and phenotypes.

Ontology fingerprints derived gene networks to identify polygenic models for diseases

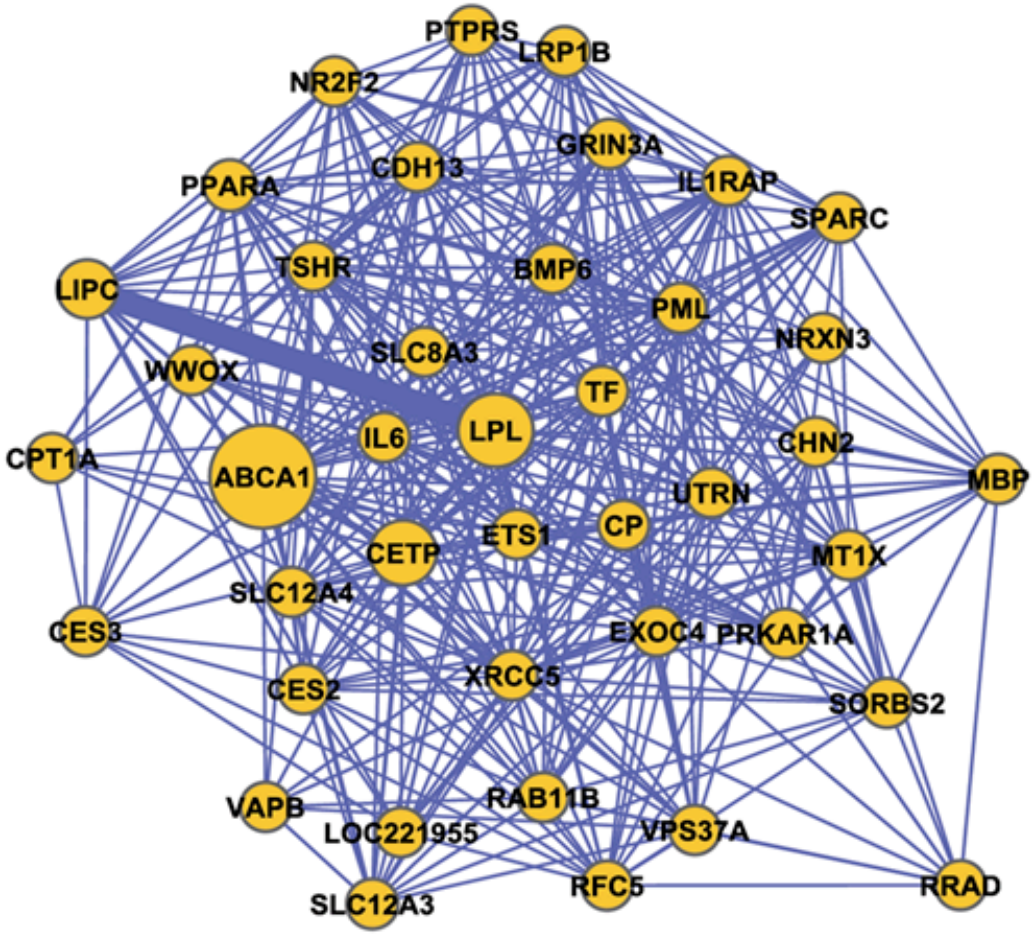
Ontology fingerprints derived gene network to dissect complex diseases

- Many diseases are caused by variants in multiple genes
- Each variant may only marginally associated with a disease phenotype, but collectively the relevant variants have very significant association
- Genes in a polygenic model are likely involved in relevant biological functions

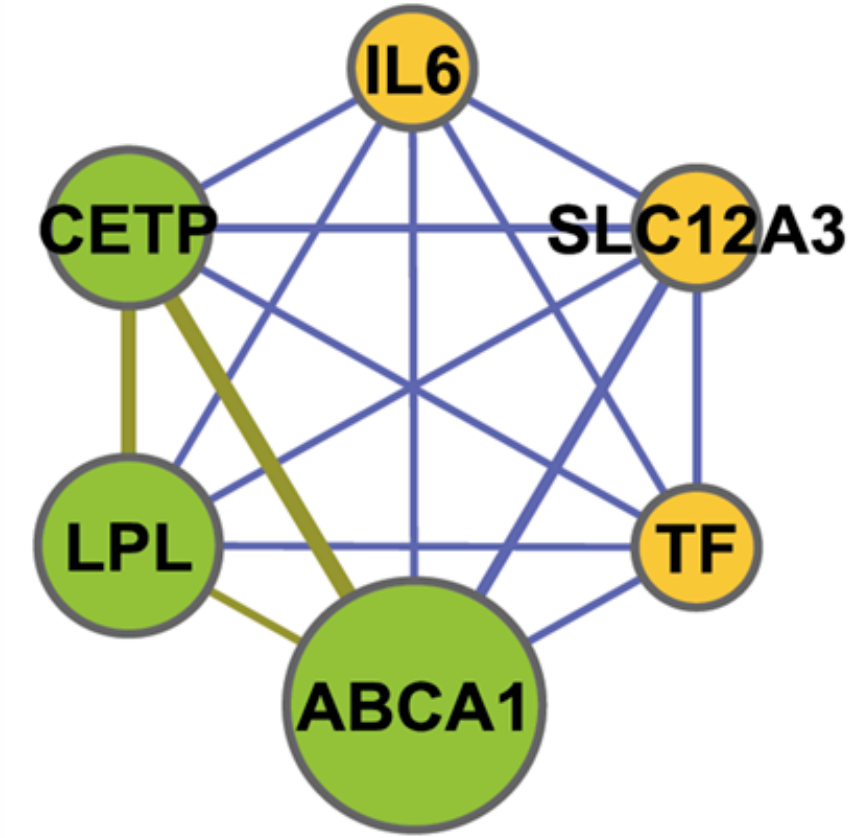
But

- Enormous amount of possible combinations among variants are hard to test
- Need efficient algorithm to narrow down candidate polygenic models

Construct a Gene Network Based on the Similarity of Genes' Ontology Fingerprints



Identify polygenic disease model from gene network



Polygenic model for dyslipidemia

ARTICLES

nature genetics

Common variants at 30 loci contribute to polygenic dyslipidemia

Sekar Kathiresan*[1–5,37,38], Cristen J Willer[6,37], Gina M Peloso[4,7,37], Serkalem Demissie[4,7,37], Kiran Musunuru[1,2], Eric E Schadt[8], Lee Kaplan[9], Derrick Bennett[10], Yun Li[6], Toshiko Tanaka[11], Benjamin F Voight[2,3,12], Lori L Bonnycastle[13], Anne U Jackson[6], Gabriel Crawford[3], Aarti Surti[3], Candace Guiducci[3], Noel P Burtt[3], Sarah Parish[10], Robert Clarke[10], Diana Zelenika[14], Kari A Kubalanza[13], Mario A Morken[13], Laura J Scott[6], Heather M Stringham[6], Pilar Galan[15], Amy J Swift[13], Johanna Kuusisto[16], Richard N Bergman[17], Jouko Sundvall[18], Markku Laakso[16], Luigi Ferrucci[11], Paul Scheet[6], Serena Sanna[19], Manuela Uda[19], Qiong Yang[4,7], Kathryn L Lunetta[4,7], Josée Dupuis[4,7], Paul I W de Bakker[20], Christopher J O'Donnell[4,21], John C Chambers[22], Jaspal S Kooner[23], Serge Hercberg[15], Pierre Meneton[24], Edward G Lakatta[25], Angelo Scuteri[26], David Schlessinger[27], Jaakko Tuomilehto[18], Francis S Collins[13], Leif Groop[28,29], David Altshuler[3,5,12,30], Rory Collins[10], G Mark Lathrop[14], Olle Melander[31], Veikko Salomaa[33], Leena Peltonen[3,32,34], Marju Orho-Melander[28], Jose M Ordovas[35,38], Michael Boehnke[6,38], Gonçalo R Abecasis[6,38], Karen L Mohlke[36,38] & L Adrienne Cupples[4,7,38]

Blood low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol and triglyceride levels are risk factors for cardiovascular disease. To dissect the polygenic basis of these traits, we conducted genome-wide association screens in 19,840 individuals and replication in up to 20,623 individuals. We identified 30 distinct loci associated with lipoprotein concentrations

Conclusion

- Ontology fingerprints constructed from enriched ontology terms in the PubMed abstracts can characterize genes and diseases
- By comparing ontology fingerprints of two biological concepts, we can quantify the relevance between them
- Quantified relevance can be used to prioritize genes from genome-wide association study
- Gene networks can be derived from comparing the ontology fingerprints of genes, and polygenic disease model can be identified as network modules

Future works, challenges and wish list

- **Future works**
  - Identify more genes and ontology terms from PubMed abstracts
  - Use full text and expand to other ontology
  - Relevance between genes and clinical concepts
  - Gene networks and models
- **Challenges**
  - Availability of ontology terms and full text papers
  - Relationship of ontology terms
- **Wish list**
  - Full text accurately annotated with genes and ontology terms
  - High quality ontology that covers extensive biological domains

Acknowledgement

- Michael Boehnke, U. Michigan
- Jijun Tang, USC
- Andrew Lawson, MUSC
- Richard Klein, MUSC
- Jim Zheng Group
  - Lam C. Tsoi (Alex)
  - Tom M. Asbury
  - Tingting Qin
  - Ravi Patel
- Funding
- ACS, PhRMA Foundation, NIH/NCRR, NSF, NIH/NLM