Establishing a Central Resource of Data
from Genome Sequencing Project

David Altshuler

June 5, 2012

Who are we trying to serve?

What questions need to be answered?

What "gene hunters" tend to think about

- I am interested in a disease / trait that varies in the population and is heritable

- I desire complete DNA sequence in patients with the disease of interest, and want to compare to the largest set of controls (without artifacts)

- Success is measured in disease genes identified, biological functions learned, and papers published

Disease-based consortia have done a better job of this in recent years, but seldom make data available to other traits

Statistical geneticists tend to think about

- I am interested in analytical and computational methodology to analyze genetic variation, inform population genetics and to find disease genes

- I desire access to data on which to develop, test and apply my methods

- Success is measured in methods developed, variation explained, papers published

dbGAP has provided a route to access data, but it is difficult to combine data across studies, and each group must start from scratch in adding new datasets and exploring new questions

While effort is needed to serve these communities,
I would argue that we are doing a better job of addressing our own needs than we are of engaging the broader biomedical community

What biologists tend to think about

- I am interested in a biological process and have selected a gene / pathway as the focus of my lab

- I tend not to want much from human genetics, other than it doesn't consume too much funding

- Success from human genetics would be allowing me to straightforwardly ask if my gene / pathway plays a role in human biology

If you were such a biologist, where would you go to comprehensively answer that question for your gene or pathway of interest?

What doctors tend to think about

- I have a patient in front of me, and need to diagnose the cause, predict the course, and recommend interventions (lifestyle, drug, surgery, etc) that would improve his or her outcome

- Success from genetics would result in tests that improve accuracy of diagnosis or prediction, or the availability of new treatments that were based on understanding of disease mechanisms

If you were a health care provider, and obtained DNA sequence data on a patient, where would you go to annotate the genome with regard to variants seen, and to evaluate the likely impact on your patient?

A key issue to interpretation: ascertainment bias

To the extent that "Mendelian" mutations have been studied only in patients with extreme phenotypes,
we may overestimate their impact on disease

We need to study the effect of "mutations" in samples unselected for a given disease,
and distinguish "failure to sequence"
from "failure to observe the variant"

What pharmaceutical researchers think about

- I am responsible for developing therapeutics that will, in human clinical trials, be safe and effective

- Human genetics is of interest insofar as it helps me predict the effect of modulating a target, or in selecting patients for inclusion in a trial

- Success is a compound that has a positive impact on disease without adverse

If you worked in a pharmaceutical company, and were developing an inhibitor of a drug target, where would you go to ask if human LOF mutations offered insight into likely efficacy and toxicity?

# Classes of questions:

Given a _phenotype_ of interest

- Identify the complete collection of genes and mutation that have the property that genetic variation is associated with my disease / trait of interest

- "Show me all the genes that are genetically associated with early onset myocardial infarction"

Given a _gene_ of interest

- What is the complete set of phenotypes associated with genetic variation in my gene?

- "We are developing drugs that target LIPG. What is the full set of phenotypic associations associated with genetic variation in the gene?"

Given a _variant_ of interest

- What is the pattern of phenotypes associated with a given genetic variant?

- "We are developing a drug that targets endothelial lipase, and there is a SNP in the gene alters HDL cholesterol. Is it associated with lowers risk of heart attack? Are there any unexpected side effects we might look out for?
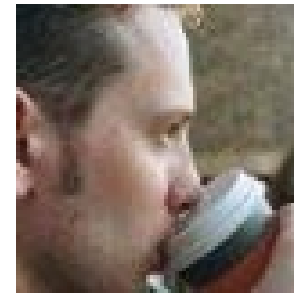
Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study

Benjamin F Voight*, Gina M Peloso*, Marju Orho-Melander, Ruth Frikke-Schmidt, Maja Barbalic, Majken K Jensen, George Hindy, Hilma Hólm, Eric L Ding, Toby Johnson, Heribert Schunkert, Nilesh J Samani, Robert Clarke, Jemma C Hopewell, John F Thompson, Mingyao Li, Gudmar Thorleifsson, Christopher Newton-Cheh, Kiran Musunuru, James P Pirruccello, Danish Saleheen, Li Chen, Alexandre F R Stewart, Arne Schillert, Unnur Thorsteinsdottir, Gudmundur Thorgeirsson, Sonia Anand, James C Engert, Thomas Morgan, John Spertus, Monika Stoll, Klaus Berger, Nicola Martinelli, Domenico Girelli, Pascal P McKeown, Christopher C Patterson, Stephen E Epstein, Joseph Devaney, Mary-Susan Burnett, Vincent Mooser, Samuli Ripatti, Ida Surakka, Markku S Nieminen, Juha Sinisalo, Marja-Liisa Lokki, Markus Perola, Aki Havulinna, Ulf de Faire, Bruna Gigante, Erik Ingelsson, Tanja Zeller, Philipp Wild, Paul I W de Bakker, Olaf H Klungel, Anke-Hilse Maitland-van der Zee, Bas J M Peters, Anthonius de Boer, Diederick E Grobbee, Pieter W Kamphuisen, Vera H M Deneer, Clara C Elbers, N Charlotte Onland-Moret, Marten H Hofker, Cisca Wijmenga, W M Monique Verschuren, Jolanda M A Boer, Yvonne T van der Schouw, Asif Rasheed, Philippe Frossard, Serkalem Demissie, Cristen Willer, Ron Do, Jose M Ordovas, Gonçalo R Abecasis, Michael Boehnke, Karen L Mohlke, Mark J Daly, Candace Guiducci, Noël P Burtt, Aarti Surti, Elena Gonzalez, Shaun Purcell, Stacey Gabriel, Jaume Marrugat, John Peden, Jeanette Erdmann, Patrik Linsel-Nitschke, Christina Willenborg, Inke R König, Marcus Fischer, Christian Hengstenberg, Andreas Ziegler, Ian Buysschaert, Diether Lambrechts, Frans Van de Werf, Keith A Fox, Nour Eddine El Mokhtari, Diana Rubin, Jürgen Schrezenmeir, Stefan Schreiber, Arne Schäfer, John Danesh, Stefan Blankenberg, Robert Roberts, Ruth McPherson, Hugh Watkins, Alistair S Hall, Kim Overvad, Eric Rimm, Eric Boerwinkle, Anne Tybjaerg-Hansen, L Adrienne Cupples, Muredach P Reilly, Olle Melander, Pier M Mannucci, Diego Ardissino, David Siscovick, Roberto Elosua, Kari Stefansson, Christopher J O'Donnell, Veikko Salomaa, Daniel J Rader, Leena Peltonen, Stephen M Schwartz, David Altshuler, Sekar Kathiresan

Voight ,Peloso et al, Lancet, May 17, 2012
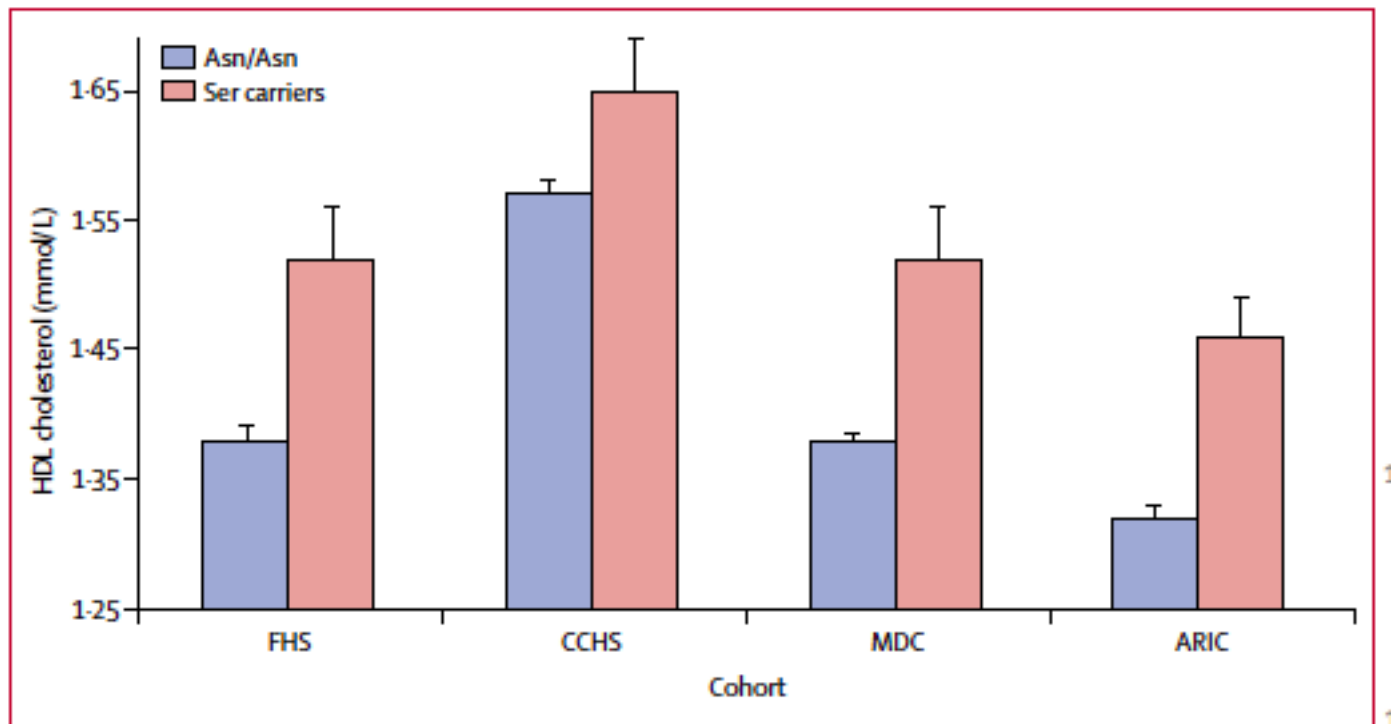
LIPG Asn396Ser raises HDL cholesterol

Figure 1: Plasma HDL cholesterol concentrations in carriers versus non-carriers of the Ser allele at the *LIPG* Asn396Ser polymorphism

Voight ,Peloso et al, Lancet, May 17, 2012

LIPG Asn396Ser has no effect on MI in studies totaling 116,320 individuals
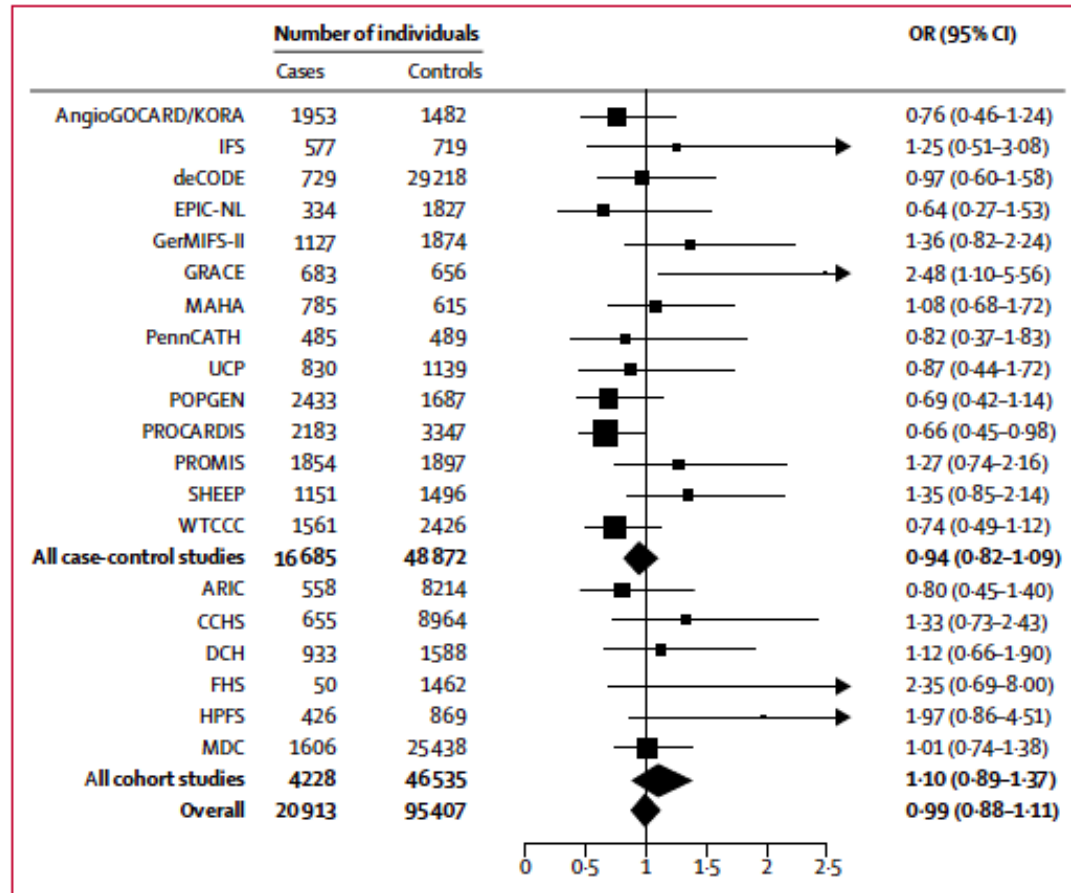
Figure 2: Association of *LIPG* Asn396Ser with myocardial infarction in 116 320 participants from 20 studies

Voight ,Peloso et al, Lancet, May 17, 2012

Why shouldn't this be a lookup?

Given a _patient_ DNA sequence of interest

- For each of the variants found in my patient, report the frequency (in different populations),and trait associations observed across all available sequence / phenotype datasets

- A child was born of consanguineous parents with a very rare syndrome – are any of the homozygous variants in the child unobserved in all other samples sequenced to date?

To my mind, the key question for us today is: how do we best leverage the extraordinary investment in genotype and phenotype data to answer questions for the communities of biologists, doctors, and drug companies

This is complex and challenging, and calls for organizational, cultural and regulatory change: our goal should not be to find "the" solution, but to encourage *innovation* and *diversity* of models to address these challenges