HHS Public Access

Author manuscript

J Cardiovasc Transl Res. Author manuscript; available in PMC 2016 November 01.

Published in final edited form as:

J Cardiovasc Transl Res. 2015 November ; 8(8): 449–457. doi:10.1007/s12265-015-9648-y.

A Guide for a Cardiovascular Genomics Biorepository: the CATHGEN Experience

William E. Kraus[1,2,4], **Christopher B. Granger**[1,3], **Michael H. Sketch Jr.**[1,2], **Mark P. Donahue**[1], **Geoffrey S. Ginsburg**[4], **Elizabeth R. Hauser**[1,2], **Carol Haynes**[2], **L. Kristin Newby**[1,3], **Melissa Hurdle**[2], **Z. Elaine Dowdy**[2], and **Svati H. Shah**[1,2]

[1]Division of Cardiology, Department of Medicine, Duke University School of Medicine, Durham, NC 27710

[2]Duke Molecular Physiology Institute, Duke University School of Medicine, Durham, NC 27710

[3]Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC 27710

[4]Duke Center for Applied Genomics and Precision Medicine, Durham, NC 27710

Abstract

The CATHGEN Biorepository was assembled in four phases. First, project startup began in 2000. Second, between 2001 and 2010, we collected clinical data and biological samples from 9334 individuals undergoing cardiac catheterization. Samples were matched at the individual level to clinical data collected at the time of catheterization and stored in the Duke Databank for Cardiovascular Diseases (DDCD). Clinical data included: subject demographics (birthdate, race, gender, etc.); cardiometabolic history including symptoms; coronary anatomy and cardiac function at catheterization; and fasting chemistry data. Third, as part of the DDCD regular follow-up protocol, yearly evaluations included interim information: vital status (verified via National Death Index search and supplemented by Social Security Death Index search), myocardial infarction (MI), stroke, rehospitalization, coronary revascularization procedures, medication use, and lifestyle habits including smoking. Fourth, samples were used to generate molecular data. CATHGEN offers the opportunity to discover biomarkers and explore mechanisms of cardiovascular disease.

Keywords

cardiovascular disease; genetics; genomics; metabolomics; air pollution; geocoding; biorepository; biomarkers; cardiometabolic disease

Corresponding Author: William E. Kraus, MD, Duke University School of Medicine, 300 N. Duke Street, Durham, NC 27701, Tel: 919-681-6733, FAX 919-684-8907, william.kraus@duke.edu.

Conflict of Interest:
GSG has a financial interest in CardioDx, Inc. No other authors have potential conflicts of interest.

Ethical Approval:
All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent:
All CATHGEN subjects were consented for participation in the biorepository and cardiovascular related research. Subject consent, data collection, sample collection, and analyses were approved through the Duke Institutional Review Board.

INTRODUCTION

Subject cohorts serve as the basis for the current understanding of the clinical epidemiology of cardiovascular disease (Multi-ethnic Study of Atherosclerosis, Women's Health Initiative, Framingham Heart Study, and others). In recent years, marrying these cohorts to simultaneously collected clinical samples has provided the opportunity to extend the research to mechanistic understanding of the relation between exposure and outcomes. Limited by low event rates, most prospective cohorts require long-term follow-up to generate enough events to discover meaningful relations between exposure and hard clinical outcomes, particularly the development of coronary artery disease and subsequent mortality. Similarly, due to the selective and often stringent recruitment criteria, clinical trial cohorts often fail to reflect the diversity and disease heterogeneity in communities or broader populations. Cohorts of sequentially collected unselected individuals presenting for evaluation or clinical care offer the opportunity to combine the best characteristics of both epidemiologic studies and clinical trials: large event rates and greater generalizability to larger clinical populations. When combined with clinical sampling and regular follow-up, prospectively collected clinical cohorts offer the opportunity to explore mechanistic underpinnings of disease in a fashion unimaginable only a decade ago.

At the turn of the century, we had the vision to assemble a biorepository of clinical samples from a prospectively-collected clinical cohort of individuals undergoing cardiac catheterization. We describe here how we developed a cohort of approximately ten thousand individuals serially sampled from individuals undergoing cardiac catheterization and yearly follow-up at Duke University Hospital. When curation of the clinical data combined with regular follow-up for cardiovascular events and molecular data of various ontogenies, this sample provides at least two unique opportunities: exploration of the molecular mechanisms underlying the development of cardiovascular disease and events; development of predictors of cardiovascular disease state and events in a high-risk clinical population.

METHODS

Building the CATHGEN Cohort was accomplished in four phases conducted simultaneously.

Phase 1. Planning and Start-up

Our effort began with a small institutional investment that permitted: the purchase of the equipment required to obtain the first 2,000 individuals' samples; planning the data collection and reconciliation; developing standard operating procedures; identifying the personnel that would handle, transport and store the samples; and obtaining Duke University Institutional Review Board (IRB) approval. All samples were collected and all subsequent research was conducted under the auspices of the Duke University IRB.

Phase 2. Sample Recruitment and Handling

A manual of procedures was developed to collect and store samples under controlled and reproducible conditions. For the first two years, consenting and sampling of catheterization

patients was performed by the cardiology fellow assigned to the case. For the last eight years of collections, consenting was performed by dedicated study personnel who also handled, processed and stored the samples. Consents were performed with awake, consenting adults who were clinically stable inpatients or outpatients. Due to staffing limitations, not all individuals could be sampled; however, during the collection period, individuals were approached randomly with respect to individual demographics or known cardiovascular disease status. Empirically, and although formal statistics were not collected, less than five percent of approached individuals refused consent.

Blood sampling from the femoral artery into a 60 mL syringe was performed at the time of catheterization, preferably at the time of femoral arterial sheath placement and before the administration of heparin prior to the formal procedure. Subjects were fasting for a minimum of six hours before the procedure. After the sample was obtained, it was immediately transferred to collection tubes that included several plastic collection tubes for whole blood using EDTA as the anticoagulant (Sarstedt, Inc.), and several PaxGene RNA tubes (Qiagen, Inc., Germantown, MD). Several EDTA tubes were processed to obtain up to ten 0.5 mL aliquots of plasma; these were separately stored from the remainder of the whole blood tube. Plasma aliquots, whole blood tubes and PaxGene RNA tubes were stored in −80°C upright freezers until accessed for further sample handling and assessment. Sample storage occurred at two locations, physically separate, to insure long-term sample provenance in case of electronic or any other emergency that might put the samples at risk. Freezers were connected to hospital emergency power and maintained under contract to an alert system that notified study personnel of freezer failure: in the Duke Center for Human Genetics Biorepository (recently renamed the Duke DNAbank) and in dedicated study freezers maintained in the laboratory of a study investigator (WEK). Consent documents and individual identifiers were maintained in secure databases and secure data storage rooms accessible only to study personnel. Sample collections occurred from January, 2001 through December 2010. All study information, including the key to connecting sample identifiers to individual medical record identifiers and other personal health information is behind the secure Duke information firewall. The time course and rate of sample collections is shown in Figure 1.

Phase 3. Curation of the Clinical and Exposure Database

Since 1969, all individuals undergoing cardiac catheterization have been entered into the Duke Databank for Cardiovascular Diseases (DDCD). The data entered at catheterization has been used for clinical reports and billing; therefore, it has over 200 individual data fields containing clinical data on every procedural patient. In addition, until this year (2015) every individual was followed longitudinally with contact made at six months after the procedure and yearly thereafter. Follow-up was performed via mailed survey or by direct contact if the mailed survey was not returned. Follow-up included information on interval cardiovascular events, new diagnoses, cardiac medications and lifestyle factors. Follow-up completion was over 97%. The DDCD Follow-up Group also yearly conducted Social Security Death Index and National Death Index searches for vital status on all individuals with significant coronary artery disease (defined as at least a 75% diameter stenosis in at least on major coronary vessel). CATHGEN contracted with the DDCD Follow-up Group to perform

follow-up on all CATHGEN participants irrespective of coronary disease status. Including the follow-up data, there are more than 200 discrete variable cells for every individual. Updates in CATHGEN patients' data in the DDCD were downloaded to the CATHGEN database held in PEDIGENE® on a quarterly basis.

All CATHGEN data are held on the PEDIGENE® database system. The PEDIGENE® database engine, Oracle 11, is housed on a UNIX server and supports demographic, clinical, biomarker, genetic, genomic, and sample information. An additional relational database instance contains sample and aliquot laboratory tracking data. PEDIGENE® is a stable but flexible and secure system that provides the backbone for collaborative studies across campus and at sites around the world. It is easily extensible and has proven its ability to scale with the explosive growth of genetic information including genome-wide genetic and gene expression experiments.

Additional Data—Various extant clinical chemistry and testing data held in the Duke Hospital clinical database are available to incorporate into the research database through the DISCERN clinical search engine. A specific consecutive cohort of 2024 individuals with samples was assembled during the interval 2004 to 2007 to support the generation of whole genome gene expression, genetic and proteomic data for the MURDOCK Study Cardiovascular Disease Study.[1] For the MURDOCK cohort, led by two of the CATHGEN investigators (LKN, SHS), additional disease and laboratory phenotyping was performed by detailed chart review, including electrocardiogram reads. Patient addresses were used to develop geocode coordinates for patient residences in collaboration with the Duke School of the Environment. These data were used to marry home addresses to air quality data prepared in collaboration with the Environmental Protection Agency (Chapel Hill office). The DDCD contains a record or all clinical exercise tests performed at Duke since 1970. Clinical exercise data — held in the DDCD and coded for maximal exercise capacity — for CATHGEN participants were matched to individual records in PEDIGENE®.

Phase 4. Curation of the Molecular Database

As funding of specific projects has been acquired, a number of molecular physiologic biomarkers of interest have been added to the CATHGEN PEDIGENE® data repository. An underlying principle contained in all CATHGEN collaborative agreements is that all data generated from the use of CATHGEN samples must be returned to the data repository for use by CATHGEN investigators in addressing relevant research questions.

GWAS Genotyping—Funding for generation of GWAS data in 3649 CATHGEN individuals was obtained through NIH grants to CATHGEN investigators (WEK, HL101621; SHS, HL095987). The Illumina Human Omni1-Quad Infinium Bead Chip was used for genotyping. Details of the methodology will be provided in a subsequent publication. In short, SNPs with <98% call frequency, minor allele frequency (MAF) <0.01 in all races, or out of Hardy-Weinberg equilibrium ($p<10^{-6}$) were excluded, resulting in the following number of autosomal SNPs for analysis: 785,945 in whites; 881,891 in blacks; and 871,209 in the "other" race (primarily Native American). Samples with <98% call rates for all SNPs, gender mismatches, cryptic relatedness, or with outlying ethnicity (based on

multidimensional scaling plots of linkage disequilibrium (LD)-pruned SNPs) were excluded, yielding data on the final 3649 samples. All GWAS data from these samples are currently available in NCBI's Database of Genotypes and Phenotypes: dbGaP.

Peripheral Whole Blood Global Gene Expression—Funding for generation of whole blood, whole genome gene expression data in 1284 CATHGEN individuals was obtained through an NIH grant to a CATHGEN investigator (WEK, HL101621). Details of the methodology will be provided in a subsequent publication. In short, RNA purification processing was done utilizing Qiagen PaxGene Blood RNA MDx Kits in whole blood PAXgene tubes. Biotinylated total RNA was generated using the Illumina TotalPrep RNA amplification kit (Life Technologies, Grand Island, NY, USA). The quality of the RNA was determined using the Bioanalyzer RNA Quantification of the RNA was determined using the Quant-iT RiboGreen RNA Assay Kit. The Human HT-12v3 Expression BeadChip (Illumina, San Diego, CA) was used for quantitative whole genome RNA profiling. Quality control was performed using Illumina GenomeStudio. Probes with a detection p-value <0.05 and detected in more than 50% of samples were retained for incorporation into the CATHGEN database in 1284 samples. Additional quality control analyses included variance components and principal components analyses to detect plate, nested chip and sample effects. Outliers at the chip, sample and probe levels were removed prior to statistical analysis. A total of 12,800 probes passed the detection and quality control (QC) filters and 1284 samples of 1554 samples passed the QC and outlier filters. All gene expression data from these samples are currently available in dbGaP.

Proteomics—Funding for generation of proteomics data in 500 CATHGEN individuals was obtained through funding to a CATHGEN investigator (LKN) through the MURDOCK Study. We selected 54 proteins for analysis in a nested case-control population of 500 individuals selecting for cardiovascular death within two years but outside of seven days of enrollment; details of the selected subjects for analysis and the proteins and analyses are presented in Halim, *et al*.[2] The proteins were selected based on previous evidence in the literature suggesting an association with risk of death or a composite of death or MI among patients with suspected or confirmed cardiovascular disease or with risk factors for cardiovascular disease, as well as expert opinion about potential novel biomarkers of risk, and for which commercial assays were available on one of two multiplexed platforms: the Meso Scale Discovery platform and the Luminex platform for protein assays not available through Meso Scale Discovery.

Metabolic Profiling—Funding for generation of targeted metabolomics data was obtained through NIH funding and American Heart Association funding to a CATHGEN investigator (SHS), as well as the MURDOCK Study (LKN). Quantitative determination of levels for 45 acylcarnitines, 15 amino acids, total ketones, β-hydroxybutyrate, and total non-esterified fatty acids (NEFA) was performed. Ketones (total and β-hydroxybutyrate) and NEFA were measured on a Beckman-Coulter DxC600 clinical chemistry analyzer, using reagents from Wako (Richmond, VA). For mass spectrometry (MS)-profiled metabolites (acylcarnitines, amino acids), proteins were first removed by precipitation with methanol. Aliquotted supernatants were dried, and then esterified with hot, acidic methanol (acylcarnitines) or *n*-

butanol (amino acids). Analysis was done using tandem MS with a Quattro Micro instrument (Waters Corporation, Milford, MA). Quantification of the "targeted" intermediary metabolites was facilitated by addition of mixtures of known quantities of stable-isotope internal standards.

Whole Genome Methylation Profiling—For whole genome methylation studies, individuals were chosen primarily based upon extremes of targeted metabolites of interest. DNA was isolated from peripheral blood mononuclear cells and sodium bisulfite treated prior to being prepped for analysis on the Illumina HumanMethylation 450K BeadChip following the manufacturer's guidelines, using the Zymo EZ DNA Methylation Kit and the manufacture's protocol (Zymo Research Corporation Irvine, California USA) for the Illumina Infinium Methylation Assay. Converted DNA was amplified, fragmented and hybridized to the Human Methylation27, RevB bead chip pool of allele-differentiating oligonucleotides. GenomeStudio was used to quantify methylated (M) and unmethylated (U) signal intensities for genomic DNA and overall methylation levels (β) calculated as the ratio of methylated to total signal ($\beta = M / (M + U)$) where β ranges from 0 (completely unmethylated) to 1 (completely methylated). Methylation data were then imported in the R statistical package and the "methylumi" package used for normalization. Δβ was calculated for the difference in overall methylation levels.

Lipoproteins Measured using Nuclear Magnetic Resonance (NMR) Spectroscopy—Lipoprotein particle concentrations and sizes were measured under a cooperative research agreement in 8738 CATHGEN individuals by (NMR) spectroscopy at LipoScience, Inc (Raleigh, NC) using the LipoProfile-3 algorithm.[3] Standard lipids were measured in a subset (n=4314) of individuals with an Olympus AU680 chemistry analyzer using Beckman Coulter reagents. LDL-C was measured using a direct homogeneous assay.

CATHGEN Stewardship, Steering Committee Membership and Intellectual Property

To oversee the collection, curation of the clinical and molecular data, and the use of the samples and data for research collaborations, the CATHGEN Steering Committee was formulated from the initiation of the project. The Steering Committee was formed from the three original founders, who were clinical research scientists and cardiologists (WEK, CBG and MHS), who brought expertise in the assembly of bio-repositories, experience with the Duke Cardiovascular Databank, and oversight of the catheterization laboratory, respectively. Other cardiologists with similar interests and experience (SHS, MPD, GSG, LKN), a genetic epidemiologist (ERH), and the CATHGEN study coordinator (ZED) were added to the Steering Committee to compose a membership of nine. Steering Committee members agreed to forego personal benefit from proceeds from any intellectual property created with CATHGEN samples. Members agreed to return any personal proceeds to the CATHGEN project by signing a statement to this effect when joining the Steering Committee. The Steering Committee meets quarterly and makes collective decisions by consensus and majority vote. An underlying principle of the Steering Committee is that one major purpose of CATHGEN is to provide a venue to train and promote the scientific careers of our cardiology fellows and other junior investigator trainees. Therefore, preference has been provided to research questions presented by these individuals.

The Committee also prioritizes studies on the basis of consumption of nonrenewable biorepository samples. Particular attention is given to plasma and serum samples; an effort is made to preserve one unthawed aliquot of approximately 0.5 mL of plasma. This sample is still preserved for all but three of the CATHGEN subjects. Approximately 90% of participants still have an un-extracted DNA sample tube and 90% of those for which PaxGene RNA tubes were collected have a remaining unextracted tube and/or extracted RNA preserved for future use.

RESULTS

The characteristics of the CATHGEN collection are detailed in Table 2. In sum, of the 9334 individuals, 62.2% were men, 74.8% White, 19.1% Black, 3.2% American Indian, 73.7% had a history of MI, CABG, or percutaneous intervention at some time, 28.3% diabetes and 32.8% hypertension. At of the time of writing, with mean follow-up of 7.5 years, 36% of the population were deceased. The age distributions by gender and decade of age are shown in Figure 2. Geocoding locations for CATHGEN participants living within North Carolina are shown in Figure 3.

As of this writing (2015), the CATHGEN data repository is very mature: CATHGEN contains air quality data on 8021; targeted metabolomics data on 3690; genome-wise SNP chip genotypes (GWAS) on 3649; genome-wide gene expression data on 1284; exercise stress test data on 2835; microRNA data on 705; targeted proteomic data on 500; and genome-wide DNA methylation data on 43 individuals. We also have lipoprotein profiling data on 8738 individuals. A Venn diagram showing the size of these various cohorts and the overlap among them within the data registry is shown in Figure 4.

DISCUSSION

The CATHGEN biorepository was assembled over a period of ten years with the assistance of academic and industry sponsorship. It is a unique resource that combines clinical data at the point of care; detailed information on cardiac anatomy and physiology; detailed phenotyping for disease in a significant portion of the cohort; clinical event follow-up on all participants for a median of seven years with a high event (mortality) rate; and paired samples available for molecular phenotyping in variety of molecular domains. Given the investment in the collection and generation of the accompanying data, Duke has valued the collection at over five million dollars. CATHGEN has already produced 40 publications with numerous collaborators.[4–15,2,16–42]

With curation of the clinical data with annual follow-up for cardiovascular events and curation of the data sample with molecular data of various ontogenies, the CATHGEN data repository provides at least two unique opportunities: exploration of the molecular mechanisms underlying the development of cardiovascular disease and events; development of predictors of cardiovascular disease state and events in a high-risk clinical population. The latter is rather straightforward, if challenged by an issue of multiple comparisons. Combining biomarker data with clinical data to assess a predictor or diagnostic is rather

straightforward: one takes a stepwise approach to adding molecular data to the already determined performance of a clinical risk score.[28,30]

Understanding the molecular physiology of disease and generating a mechanistic understanding of coronary heart disease and its accompanying metabolic conditions (diabetes, metabolic syndrome, obesity, hypertension) and environmental contributors (air quality), presents a much larger challenge. The scope of the challenge is illustrated in Figure 5. If one were to work forward from gene to outcome (as shown in the top panel—Simple Hierarchy of Molecular Regulatory Control), through gene expression and metabolites, one would confront an unmanageable number of statistical comparisons (10^6 genes, 20,000 genes with expression tags, 67 metabolites on 10 potential outcomes: ~10^{13} tests). This is a significant multiple testing burden, given that — in the best case — we only have 10,000 individuals in an analysis. To reduce the multiple testing burden — and given the lack of currently available analytic alternatives — we have been forced to consider two-way relations at a time (e.g., between gene expression and outcome, metabolite and outcome) and to take a staged approach, reducing the number of candidates in the pool for analysis at each step (middle panel, Figure 5). Clearly, this is a simplistic approach and ignores complex molecular pathways (lower panel, Figure 5. This is the current challenge in the field, and we look forward to working with others in industry and academia in meeting this challenge.

Acknowledgments

Collections and generation of molecular data were generated in part through research agreements with the following: BG Medicine, Inc.; CardioDx, Inc.; Qiagen, Inc.; Liposcience, Inc. GWAs and whole genome gene expression data were generated through NIH-funded studies to WEK (HL101621) and SHS (HL095987). Metabolomic data were generated through grants to SHS (HL095987, AHA Fellow-to-Faculty). An internal grant award from the MURDOCK Study (David H. Murdock Institute for Business and Culture, 1UL1 RR024128 from the National Center for Research Resources NCRR) to LKN was also helpful and appreciated. We wish to thank Dr. Marie Lynn Miranda and the Duke School of the Environment for developing geocoding addresses for CATHGEN participants.

Funding: Through Duke, collections and generation of molecular data were generated in part through research agreements with the following: BG Medicine, Inc.; CardioDx, Inc.; Qiagen, Inc.; Liposcience, Inc..; US Environmental Protection Agency. GWAs and whole genome gene expression data were generated through NIH-funded studies to WEK (HL101621) and SHS (HL095987). Metabolomic data were generated through grants to SHS (HL095987, AHA Fellow-to-Faculty).

References

1. Bhattacharya S, Dunham AA, Cornish MA, Christian VA, Ginsburg GS, Tenenbaum JD, et al. The Measurement to Understand Reclassification of Disease of Cabarrus/Kannapolis (MURDOCK) Study Community Registry and Biorepository. Am J Transl Res. 2012; 4(4):458–470. [PubMed: 23145214]
2. Halim SA, Neely ML, Pieper KS, Shah SH, Kraus WE, Hauser ER, et al. Simultaneous consideration of multiple candidate protein biomarkers for long-term risk for cardiovascular events. Circ Cardiovasc Genet. 2015; 8(1):168–177. [PubMed: 25422398]
3. Jeyarajah EJ, Cromwell WC, Otvos JD. Lipoprotein particle analysis by nuclear magnetic resonance spectroscopy. Clin Lab Med. 2006; 26(4):847–870. [PubMed: 17110242]
4. Beineke P, Fitch K, Tao H, Elashoff MR, Rosenberg S, Kraus WE, et al. A whole blood gene expression-based signature for smoking status. BMC Med Genomics. 2012; 5:58. [PubMed: 23210427]
5. Bhattacharya S, Granger CB, Craig D, Haynes C, Bain J, Stevens RD, et al. Validation of the association between a branched chain amino acid metabolite profile and extremes of coronary artery

disease in patients referred for cardiac catheterization. Atherosclerosis. 2014; 232(1):191–196. [PubMed: 24401236]

6. Brunner MP, Shah SH, Craig DM, Stevens RD, Muehlbauer MJ, Bain JR, et al. Effect of heparin administration on metabolomic profiles in samples obtained during cardiac catheterization. Circ Cardiovasc Genet. 2011; 4(6):695–700. [PubMed: 22010138]
7. Connelly JJ, Cherepanova OA, Doss JF, Karaoli T, Lillard TS, Markunas CA, et al. Epigenetic regulation of COL15A1 in smooth muscle cell replicative aging and atherosclerosis. Hum Mol Genet. 2013; 22(25):5107–5120. [PubMed: 23912340]
8. Connelly JJ, Shah SH, Doss JF, Gadson S, Nelson S, Crosslin DR, et al. Genetic and functional association of FAM5C with myocardial infarction. BMC Med Genet. 2008; 9:33. [PubMed: 18430236]
9. Connelly JJ, Wang T, Cox JE, Haynes C, Wang L, Shah SH, et al. GATA2 is associated with familial early-onset coronary artery disease. PLoS Genet. 2006; 2(8):e139. [PubMed: 16934006]
10. Crosslin DR, Shah SH, Nelson SC, Haynes CS, Connelly JJ, Gadson S, et al. Genetic effects in the leukotriene biosynthesis pathway and association with atherosclerosis. Hum Genet. 2009; 125(2): 217–229. [PubMed: 19130089]
11. Daniels SE, Beineke P, Rhees B, McPherson JA, Kraus WE, Thomas GS, et al. Biological and analytical stability of a peripheral blood gene expression score for obstructive coronary artery disease in the PREDICT and COMPASS studies. J Cardiovasc Transl Res. 2014; 7(7):615–622. [PubMed: 25119856]
12. Davies RW, Wells GA, Stewart AF, Erdmann J, Shah SH, Ferguson JF, et al. A genome-wide association study for coronary artery disease identifies a novel susceptibility locus in the major histocompatibility complex. Circ Cardiovasc Genet. 2012; 5(2):217–225. [PubMed: 22319020]
13. Do R, Stitziel NO, Won HH, Jorgensen AB, Duga S, Angelica Merlini P, et al. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. Nature. 2015; 518(7537):102–106. [PubMed: 25487149]
14. Dungan JR, Hauser ER, Qin X, Kraus WE. The genetic basis for survivorship in coronary artery disease. Front Genet. 2013; 4:191. [PubMed: 24143143]
15. Elashoff MR, Wingrove JA, Beineke P, Daniels SE, Tingley WG, Rosenberg S, et al. Development of a blood-based gene expression algorithm for assessment of obstructive coronary artery disease in non-diabetic patients. BMC Med Genomics. 2011; 4:26. [PubMed: 21443790]
16. Horne BD, Hauser ER, Wang L, Muhlestein JB, Anderson JL, Carlquist JF, et al. Validation study of genetic associations with coronary artery disease on chromosome 3q13–21 and potential effect modification by smoking. Ann Hum Genet. 2009; 73(Pt 6):551–558. [PubMed: 19706030]
17. Kertai MD, Li YW, Li YJ, Shah SH, Kraus WE, Fontes ML, et al. G protein-coupled receptor kinase 5 gene polymorphisms are associated with postoperative atrial fibrillation after coronary artery bypass grafting in patients receiving beta-blockers. Circ Cardiovasc Genet. 2014; 7(5):625–633. [PubMed: 25049040]
18. Kral BG, Mathias RA, Suktitipat B, Ruczinski I, Vaidya D, Yanek LR, et al. A common variant in the CDKN2B gene on chromosome 9p21 protects against coronary artery disease in Americans of African ancestry. J Hum Genet. 2011; 56(3):224–229. [PubMed: 21270820]
19. Lansky A, Elashoff MR, Ng V, McPherson J, Lazar D, Kraus WE, et al. A gender-specific blood-based gene expression score for assessing obstructive coronary artery disease in nondiabetic patients: results of the Personalized Risk Evaluation and Diagnosis in the Coronary Tree (PREDICT) trial. Am Heart J. 2012; 164(3):320–326. [PubMed: 22980297]
20. Lappe JM, Horne BD, Shah SH, May HT, Muhlestein JB, Lappe DL, et al. Red cell distribution width, C-reactive protein, the complete blood count, and mortality in patients with coronary disease and a normal comparison population. Clin Chim Acta. 2011; 412(23–24):2094–2099. [PubMed: 21821014]
21. Minear MA, Crosslin DR, Sutton BS, Connelly JJ, Nelson SC, Gadson-Watson S, et al. Polymorphic variants in tenascin-C (TNC) are associated with atherosclerosis and coronary artery disease. Hum Genet. 2011; 129(6):641–654. [PubMed: 21298289]

22. Nolan DK, Sutton B, Haynes C, Johnson J, Sebek J, Dowdy E, et al. Fine mapping of a linkage peak with integration of lipid traits identifies novel coronary artery disease genes on chromosome 5. BMC Genet. 2012; 13:12. [PubMed: 22369142]
23. Peloso GM, Auer PL, Bis JC, Voorman A, Morrison AC, Stitziel NO, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. Am J Hum Genet. 2014; 94(2):223–232. [PubMed: 24507774]
24. Rosenberg S, Elashoff MR, Beineke P, Daniels SE, Wingrove JA, Tingley WG, et al. Multicenter validation of the diagnostic accuracy of a blood-based gene expression test for assessing obstructive coronary artery disease in nondiabetic patients. Ann Intern Med. 2010; 153(7):425–434. [PubMed: 20921541]
25. Rosenberg S, Elashoff MR, Lieu HD, Brown BO, Kraus WE, Schwartz RS, et al. Whole blood gene expression testing for coronary artery disease in nondiabetic patients: major adverse cardiovascular events and interventions in the PREDICT trial. J Cardiovasc Transl Res. 2012; 5(3):366–374. [PubMed: 22396313]
26. Sehnert AJ, Daniels SE, Elashoff M, Wingrove JA, Burrow CR, Horne B, et al. Lack of association between adrenergic receptor genotypes and survival in heart failure patients treated with carvedilol or metoprolol. J Am Coll Cardiol. 2008; 52(8):644–651. [PubMed: 18702968]
27. Shah AA, Craig DM, Sebek JK, Haynes C, Stevens RC, Muehlbauer MJ, et al. Metabolic profiles predict adverse events after coronary artery bypass grafting. J Thorac Cardiovasc Surg. 2012; 143(4):873–878. [PubMed: 22306227]
28. Shah SH, Bain JR, Muehlbauer MJ, Stevens RD, Crosslin DR, Haynes C, et al. Association of a peripheral blood metabolic profile with coronary artery disease and risk of subsequent cardiovascular events. Circ Cardiovasc Genet. 2010; 3(2):207–214. [PubMed: 20173117]
29. Shah SH, Freedman NJ, Zhang L, Crosslin DR, Stone DH, Haynes C, et al. Neuropeptide Y gene polymorphisms confer risk of early-onset atherosclerosis. PLoS Genet. 2009; 5(1):e1000318. [PubMed: 19119412]
30. Shah SH, Granger CB, Hauser ER, Kraus WE, Sun JL, Pieper K, et al. Reclassification of cardiovascular risk using integrated clinical and molecular biosignatures: Design of and rationale for the Measurement to Understand the Reclassification of Disease of Cabarrus and Kannapolis (MURDOCK) Horizon 1 Cardiovascular Disease Study. Am Heart J. 2010; 160(3):371–379. [PubMed: 20826242]
31. Shah SH, Hauser ER, Crosslin D, Wang L, Haynes C, Connelly J, et al. ALOX5AP variants are associated with in-stent restenosis after percutaneous coronary intervention. Atherosclerosis. 2008; 201(1):148–154. [PubMed: 18374923]
32. Shah SH, Sun JL, Stevens RD, Bain JR, Muehlbauer MJ, Pieper KS, et al. Baseline metabolomic profiles predict cardiovascular events in patients at risk for coronary artery disease. Am Heart J. 2012; 163(5):844–850. e841.10.1016/j.ahj.2012.02.005 [PubMed: 22607863]
33. Singh A, Babyak MA, Nolan DK, Brummett BH, Jiang R, Siegler IC, et al. Gene by stress genome-wide interaction analysis and path analysis identify EBF1 as a cardiovascular and metabolic risk gene. Eur J Hum Genet. 201410.1038/ejhg.2014.189
34. Strauss BW, Valentiner EM, Bhattacharya S, Smerek MM, Dunham AA, Newby LK, et al. Improving population representation through geographic health information systems: mapping the MURDOCK study. Am J Transl Res. 2014; 6(4):402–412. [PubMed: 25075257]
35. Sutton BS, Crosslin DR, Shah SH, Nelson SC, Bassil A, Hale AB, et al. Comprehensive genetic analysis of the platelet activating factor acetylhydrolase (PLA2G7) gene and cardiovascular disease in case-control and family datasets. Hum Mol Genet. 2008; 17(9):1318–1328.10.1093/hmg/ddn020 [PubMed: 18204052]
36. Vargas J, Lima JA, Kraus WE, Douglas PS, Rosenberg S. Use of the Corus(R) CAD gene expression test for assessment of obstructive coronary artery disease likelihood in symptomatic non-diabetic patients. PLoS Curr. 2013; 5
37. Voora D, Cyr D, Lucas J, Chi JT, Dungan J, McCaffrey TA, et al. Aspirin exposure reveals novel genes associated with platelet function and cardiovascular events. J Am Coll Cardiol. 2013; 62(14):1267–1276. [PubMed: 23831034]

38. Wang L, Hauser ER, Shah SH, Pericak-Vance MA, Haynes C, Crosslin D, et al. Peakwide mapping on chromosome 3q13 identifies the kalirin gene as a novel candidate gene for coronary artery disease. Am J Hum Genet. 2007; 80(4):650–663. [PubMed: 17357071]

39. Wang L, Hauser ER, Shah SH, Seo D, Sivashanmugam P, Exum ST, et al. Polymorphisms of the tumor suppressor gene LSAMP are associated with left main coronary artery disease. Ann Hum Genet. 2008; 72(Pt 4):443–453. [PubMed: 18318786]

40. Ward-Caviness C, Haynes C, Blach C, Dowdy E, Gregory SG, Shah SH, et al. Gene-smoking interactions in multiple Rho-GTPase pathway genes in an early-onset coronary artery disease cohort. Hum Genet. 2013; 132(12):1371–1382. [PubMed: 23907653]

41. Ward-Caviness, CK.; Kraus, WE.; Blach, C.; Haynes, CS.; Dowdy, E.; Miranda, ML., et al. Association of roadway proximity with fasting plasma glucose and metabolic risk factors for cardiovascular disease in a cross-sectional study of cardiac catheterization patients. Environ Health Perspect; 2015. (in press).

42. Wingrove JA, Daniels SE, Sehnert AJ, Tingley W, Elashoff MR, Rosenberg S, et al. Correlation of peripheral-blood gene expression with the extent of coronary artery stenosis. Circ Cardiovasc Genet. 2008; 1(1):31–38. [PubMed: 20031539]

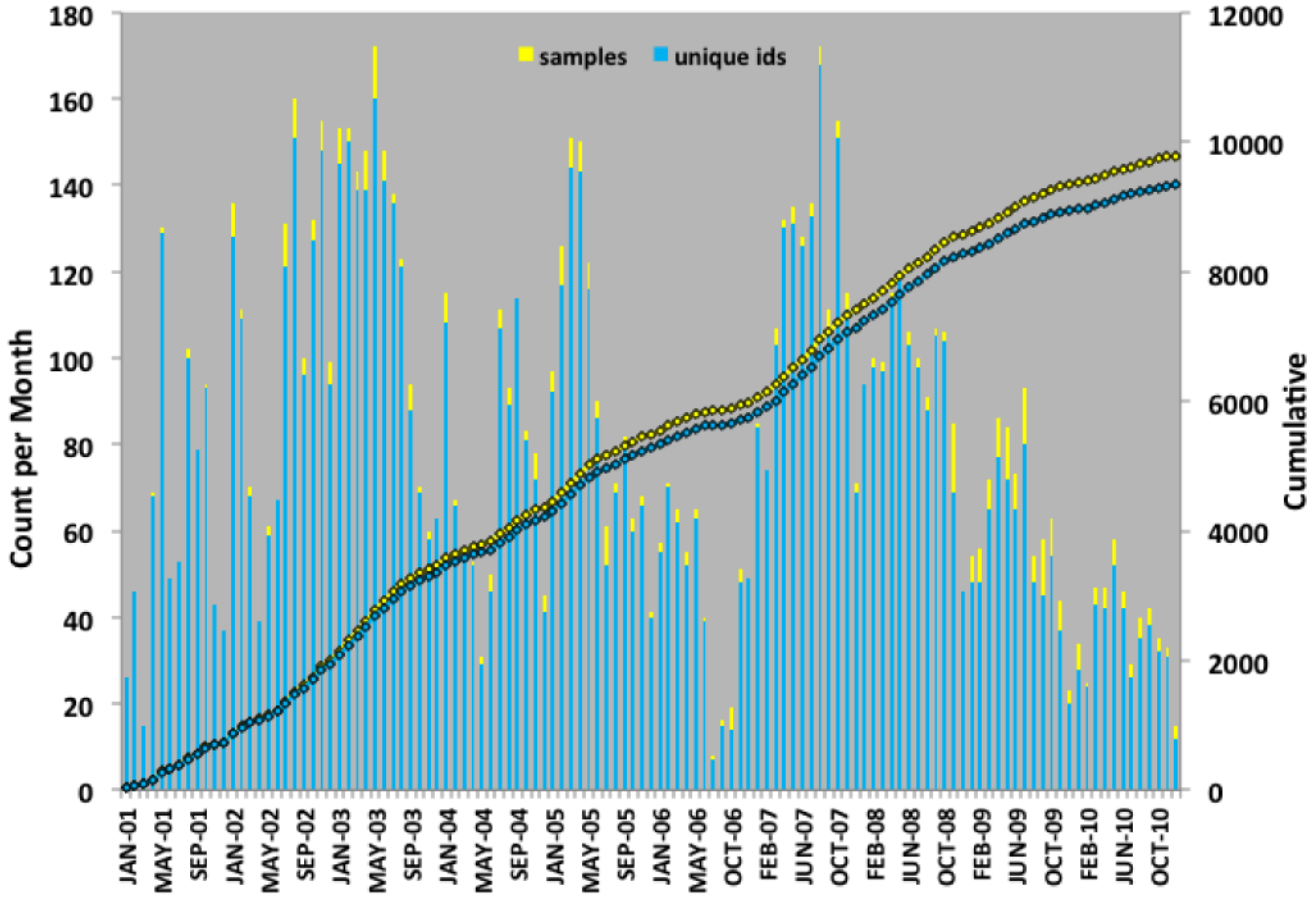

Figure 1. CATHGEN Ascertainment Timeline

DNA, plasma and PaxGene RNA tubes were collected on 9334 samples between January 2001 and December 2010. Timeline in counts per month are plotted with cumulative ascertainment also shown for total samples and distinct individuals, as some individuals were sampled more than once.

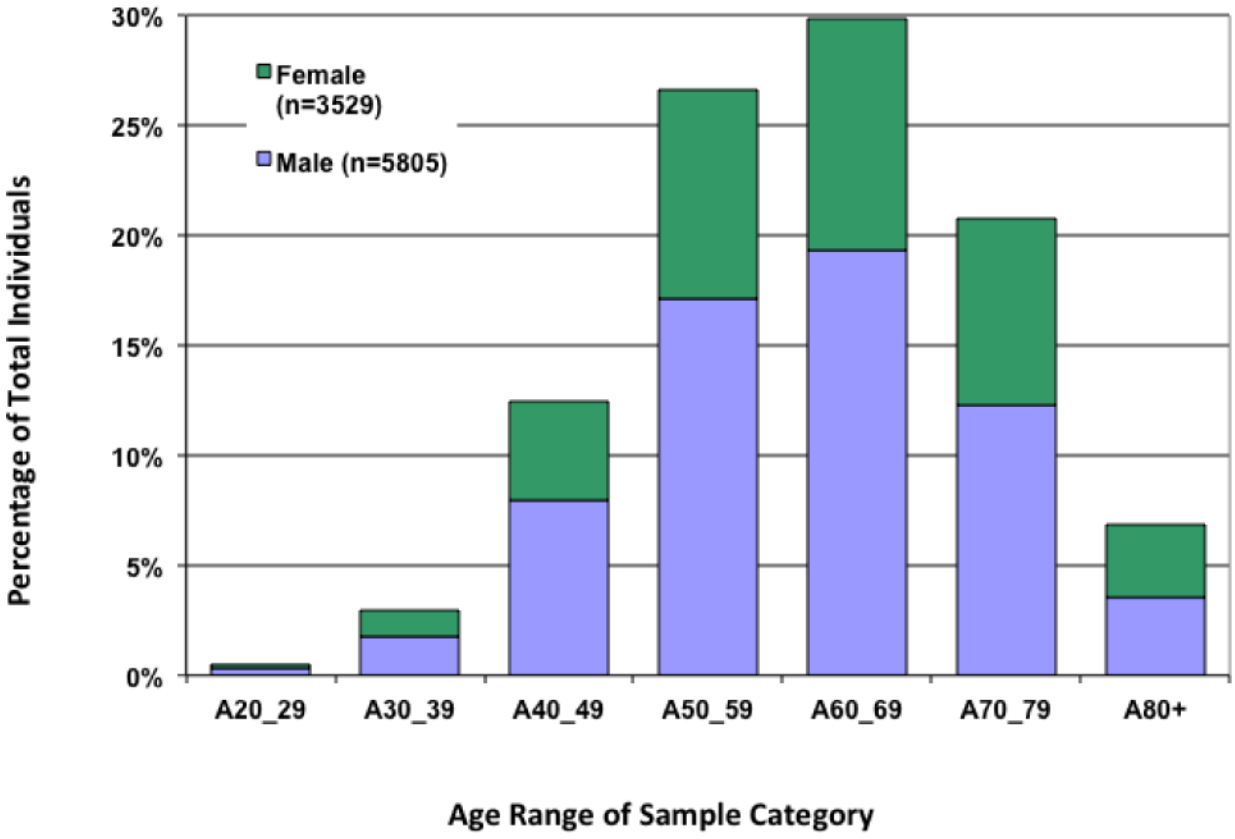

Figure 2. Age and gender of Sampled CATHGEN Individuals with Clinical Information

Aged of sampled individuals are shown by decades. The number of men and women in each category are shown by blue and green bars, respectively. The mean age for men was 62.0 (±12.5 SD) y and women was 60.8 (±11.7) y.

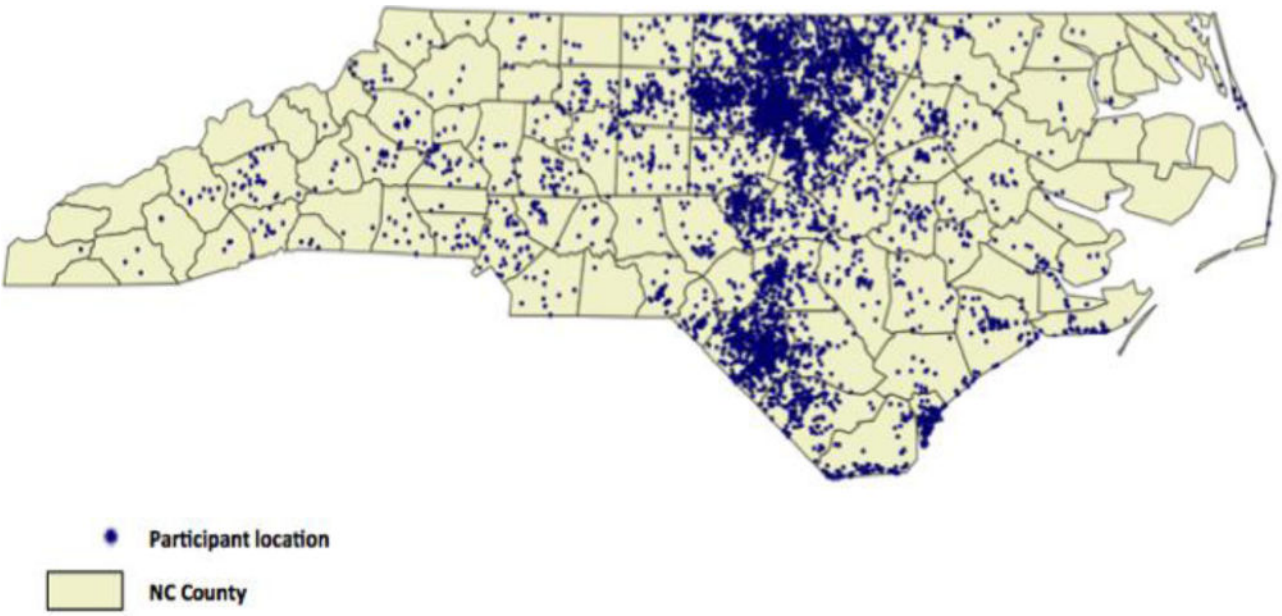

Figure 3. Geographic Distribution of the Residences of North Carolina CATHGEN Participants

Residences were assigned geocoded locations by the Duke School of the Environment using the declared residence of CATHGEN participants on the day of enrollment.

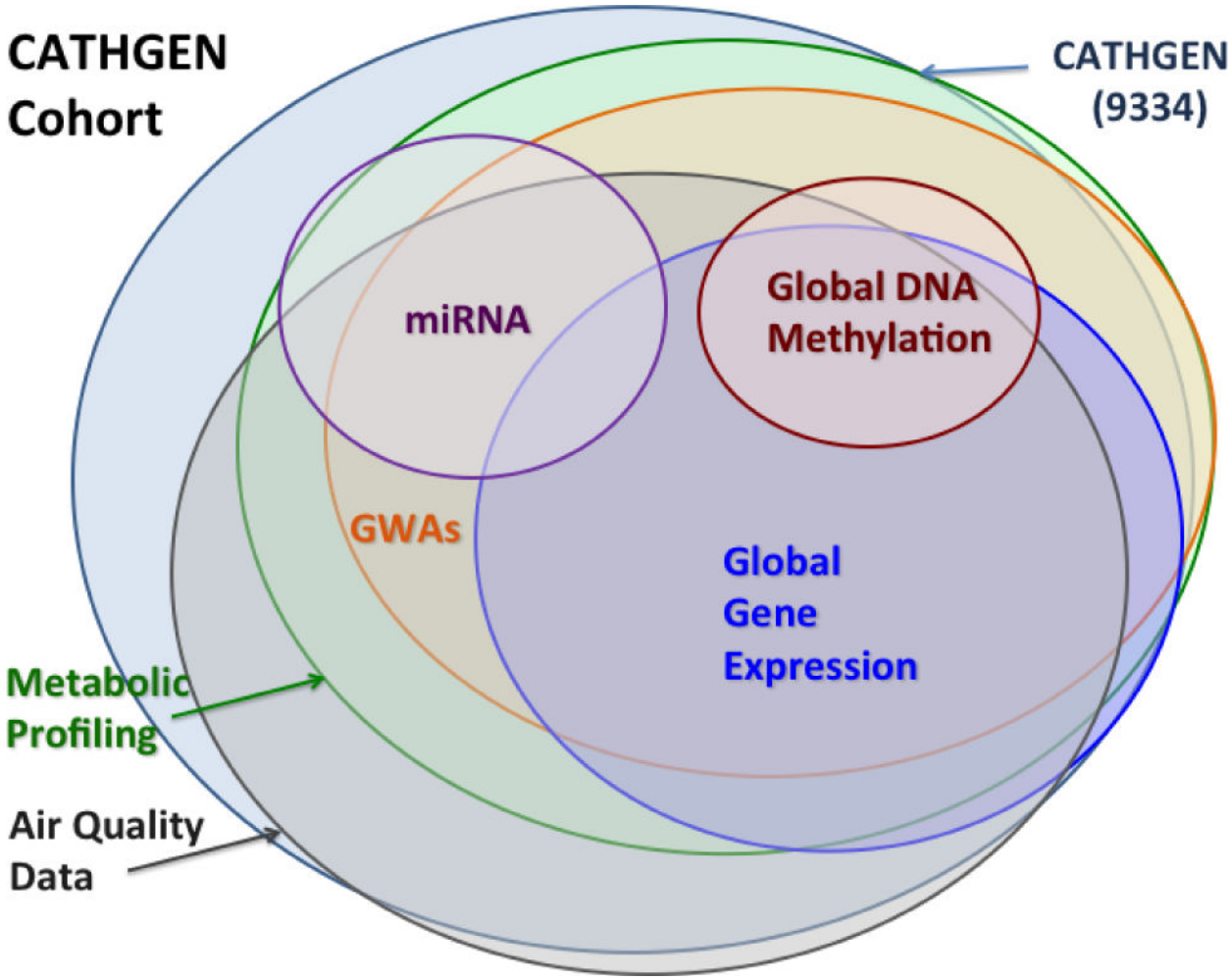

Figure 4. Venn Diagram of Subject Data and Overlap for Several Data Ontogenies

Of the 9334 subjects, CATHGEN contains air quality data on 8021; targeted metabolomics data on 3899; genome-wise SNP chip genotypes (GWAS) on 3649; genome-wide gene expression data on 1284; exercise stress test data on 2835; microRNA data on 705; and genome-wide DNA methylation data on 43 individuals. We also have lipoprotein profiling data on 8000 individuals (not shown).

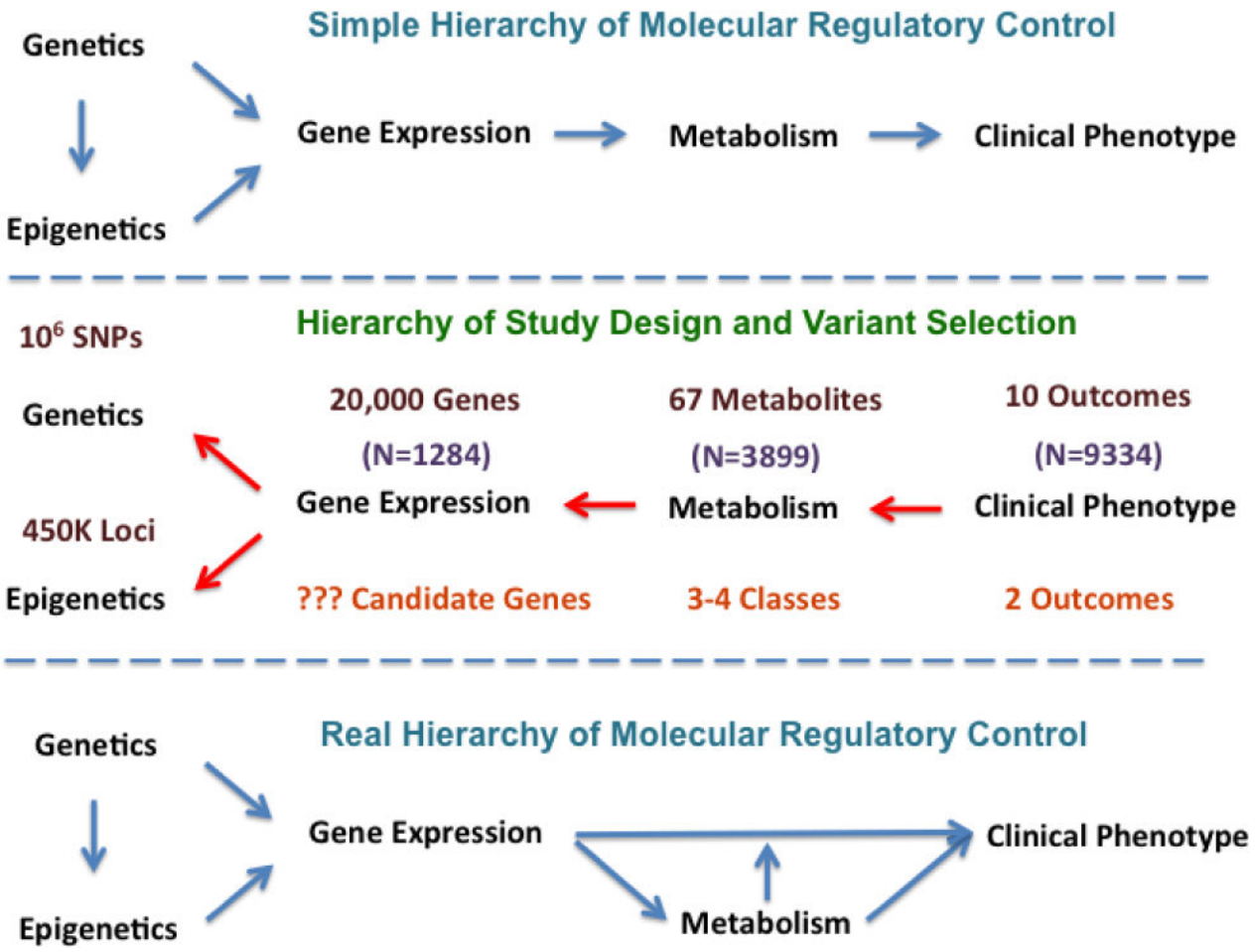

Figure 5. Challenge and Approaches to Systems Modeling in Large Omics Datasets

In the upper panel is depicted the classical understanding of the hierarchy of regulatory control, wherein DNA leads to RNA to protein, to metabolite and then to phenotype. As shown in the middle panel and as explained in the text, in the case of CATHGEN, if attempting to pick out specific molecular regulatory pathways one is left with a untenable problem of trying to select the most likely regulatory events from a possible 10^{13} tests. Shown at each stage is the number of possible molecular probes (e.g., a typical gene expression array selects for approximately 20,000 gene expression targets). One approach is to use a phased approach ("Inverse QTL Modeling"); one works backward from phenotype toward gene, selecting significant targets at each stage, thus reducing the number of overall comparisons in the analysis. However, as shown in the lowest panel, this approach ignores the real hierarchy of regulatory control which is a not a linear process (e.g., metabolite produced by gene expression can have no role in the pathway from gene expression to phenotype; can mediate the effect of gene expression; or can feed back on gene expression — in a positive or negative fashion — and moderate the effect of gene expression on phenotype.

Table 1

Definitions, Abbreviations and Acronyms

CATHGEN	– CATHeterization GENetics; sample and clinical data repository
dbGaP	– NCBI's Database of Genotypes and Phenotypes
DDCD	– Duke Databank for Cardiovascular Diseases; clinical database
DISCERN	– Duke search engine for clinical data from patient records
DNA	– Deoxyribonucleic acid
EDTA	– Ethylenediaminetetraacetic acid; anticoagulant
GWAS	– genome-wide association study
IRB	– Institutional Review Board
LD	– linkage disequilibrium
LDL-C	– low density lipoprotein-cholesterol
MAF	– ninor allele frequency
MS	– mass spectrometry
MURDOCK Study	– Contiguous sample of 2024 CATHGEN participants
NCBI	– National Center for Biotechnology Information
NEFA	– non-esterified fatty acids
PEDIGENE®	– Data repository for clinical and sample data
QTL	– quantitative trait locus
RNA	– Ribonucleic acid

Table 2

Demographic and clinical disease characteristics of the 9334 individuals in the CATHGEN Registry

							Total	
	N	Male Percent Total	Percent Gender	N	Female Percent Total	Percent Gender	N	Percent
White	4553	48.78	78.43	2428	26.01	68.80	6981	74.79
Black	915	9.80	15.76	863	9.25	24.45	1778	19.05
American Indian	170	1.82	2.93	126	1.35	3.57	296	3.17
Other/Missing	167	1.79	2.88	112	1.20	3.17	279	2.99
Total	5805	62.19	100	3529	37.81	100	9334	100
Age (SD)	62.0 (12.5)	(21.9, 93.7)	range	60.8 (11.7)	(18.3, 93.8)	range	61.3 (12.0)	(18.3, 93.8)
Hx MI, CABG Or Intervention	4026	43.13	69.35	2855	30.59	80.90	6881	73.72
Hx DM	1616	17.31	27.84	1024	10.97	29.02	2640	28.28
Hx HTN	1913	20.49	32.95	1144	12.26	32.42	3057	32.75