Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2018 April 27.

Published in final edited form as:

Nat Genet. 2017 October 27; 49(11): 1560–1563. doi:10.1038/ng.3968.

Analysis Commons, A Team Approach to Discovery in a Big-Data Environment for Genetic Epidemiology

Jennifer A. Brody[1,*], **Alanna C. Morrison**[2,*], **Joshua C. Bis**[1,*], **Jeffrey R. O'Connell**[3], **Michael R. Brown**[2], **Jennifer E. Huffman**[4], **Darren C. Ames**[5], **Andrew Carroll**[5], **Matthew P. Conomos**[6], **Stacey Gabriel**[7], **Richard A. Gibbs**[8], **Stephanie M. Gogarten**[6], **Namrata Gupta**[7], **Cashell E. Jaquish**[9], **Andrew D. Johnson**[4], **Joshua P. Lewis**[3], **Xiaoming Liu**[2], **Alisa K. Manning**[10,11,12], **George J. Papanicolaou**[9], **Achilleas N. Pitsillides**[13], **Kenneth M. Rice**[6], **William Salerno**[8], **Colleen M. Sitlani**[1], **Nicholas L. Smith**[1,14,15,16], **NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, The Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, TOPMed Hematology and Hemostasis Working Group, CHARGE Analysis and Bioinformatics Working Group**, **Susan R. Heckbert**[1,16], **Cathy C. Laurie**[6], **Braxton D. Mitchell**[3,17], **Ramachandran S. Vasan**[13,18,19], **Stephen S. Rich**[20], **Jerome I. Rotter**[21], **James G. Wilson**[22], **Eric Boerwinkle**[2,8,*], **Bruce M. Psaty**[1,14,23,*], and **L. Adrienne Cupples**[24,25,*]

[1]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA

[2]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA

[3]Department of Medicine, Division of Endocrinology, Diabetes, and Nutrition, University of Maryland, Baltimore, MD, USA

[4]The Framingham Heart Study, Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Framingham, MA, USA

[5]DNAnexus, Inc. Mountain View, CA, USA

[6]Department of Biostatistics, University of Washington, Seattle, WA, USA

[7]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA

[8]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA

Corresponding Authors: Jennifer A. Brody, Research Scientist, Cardiovascular Health Research Unit, 1730 Minor Ave, Suite 1360, Seattle, WA 98101, jeco@u.washington.edu, Phone: 206-221-7775, Fax: 206-221-2662; L. Adrienne Cupples, Professor of Biostatistics and of Epidemiology, Department of Biostatistics, Boston University School of Public Health, Framingham Heart Study, 801 Massachusetts Avenue, CT3, Boston, MA 02118, adrienne@bu.edu, Phone: 617-638-5176, Fax: 617-638-6484.

*These authors contributed equally to the work

Author Contributions: S.R.H, C.C.L., B.D.M., J.R.O., R.S.V., S.S.R., J.I.R., J.G.W., J.A.B., A.C.M, J.C.B, E.B., B.M.P, L.A.C, K.M.R. formed the management team of the Analysis Commons. J.A.B, A.C.M., J.C.B, E.B., B.M.P, L.A.C, S.R.H., X.L. drafted the manuscript. W.S., R.A.G., A.C., D.C.A managed the computing infrastructure. A.K.M., J.R.O., M.R.B., D.C.A, A.C., M.P.C., S.M.G., A.N.P. were responsible for implementation and design of the applications. S.G., N.G. oversaw the sequence generation. J.A.B., J.E.H., J.P.L., A.D.J., J.C.B, C.M.S., N.L.S., C.E.J., G.J.P. conceived, designed and implemented the data example analyses. All co-authors reviewed and edited the manuscript before approving its submission.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[9]National Heart Lung and Blood Institute, Division of Cardiovascular Sciences, Bethesda, MD, USA

[10]Center for Human Genetics Research, Massachusetts General Hospital, Boston, MA, USA

[11]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA

[12]Department of Medicine, Harvard Medical School, Boston, MA, USA

[13]National Heart Lung and Blood Institute's and Boston University's Framingham Heart Study, Framingham, MA, USA

[14]Kaiser Permanent Washington Health Research Institute., Seattle, WA, USA

[15]Seattle Epidemiologic Research and Information Center, Department of Veteran Affairs Office of Research and Development, Seattle, WA, USA

[16]Department of Epidemiology, University of Washington, Seattle, WA, USA

[17]Geriatrics Research and Education Clinical Center, Baltimore Veterans Administration Medical Center, Baltimore, MD 21201

[18]Sections of Preventive Medicine and Epidemiology And Cardiology, Department of Medicine, Boston University School of Medicine, Boston, MA, USA

[19]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA

[20]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA

[21]Institute for Translational Genomics and Population Sciences, Departments of Pediatrics and Medicine, LABioMed at Harbor-UCLA Medical Center, Torrance, CA, USA

[22]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA

[23]Departments of Medicine, Epidemiology, and Health Services, University of Washington, Seattle, WA, USA

[24]NHLBI Framingham Heart Study, Framingham, MA, USA

[25]Department of Biostatistics, Boston University School of Public Health, Boston, MA

Summary paragraph

The exploding volume of whole-genome sequence (WGS) and multi-omics data requires new approaches for analysis. As one solution, we have created a cloud-based Analysis Commons, which brings together genotype and phenotype data from multiple studies in a setting that is accessible by multiple investigators. This framework addresses many of the challenges of multi-center WGS analyses, including data sharing mechanisms, phenotype harmonization, integrated multi-omics analyses, annotation, and computational flexibility. In this setting, the computational pipeline facilitates a sequence-to-discovery analysis workflow illustrated here by an analysis of plasma fibrinogen levels in 3996 individuals from the National Heart, Lung, and Blood Institute (NHLBI) Trans-Omics for Precision Medicine (TOPMed) WGS program. The Analysis Commons represents a novel model for transforming WGS resources from a massive quantity of phenotypic

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

and genomic data into knowledge of the determinants of health and disease risk in diverse human populations.

The Analysis Commons, which relies on a new team-science model for genetic epidemiology, integrates multi-omic data and rich phenotypic and clinical information from diverse population studies into a single shared analytic platform that leverages the resources of a cloud-computing environment and provides for distributed access. The number of WGS studies with large sample sizes is rapidly expanding. Projects such as the National Heart, Lung, and Blood Institute (NHLBI) Trans-Omics for Precision Medicine (TOPMed) Program, Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium[1,2], and the Centers for Common Disease Genomics (CCDG)[3], among others, have already conducted WGS in more than 100,000 individuals, and the Personalized Medicine Initiative[4] promises whole-genome sequencing in over a million samples. These programs span a diverse set of studies and institutions, many of which lack the computational infrastructure to store and compute on this scale of data. Genomic, epigenomic, metabolic and proteomic data derived from expensive assays often do not exist in large numbers in any single study, but represent a powerful discovery resource when combined across studies and integrated with phenotypic data.

In aggregate, many population-based studies data have collected tens of thousands of variables over a period of decades, and the addition of WGS data to cohorts with long-term prospective follow-up provides a powerful resource for immediate discovery. Analysis of WGS data on large samples presents formidable computational and administrative challenges. Evaluation of rare genetic variation in WGS data requires manipulation of datasets that are tens to hundreds of terabytes in size and are prohibitively large for exchange between analysis sites. In contrast, pooled datasets that include genotype and phenotype data from all participants in the contributing individual studies provide for practical and efficient WGS analysis. The creation of such large pooled datasets containing harmonized multi-omic, phenotype and clinical data with appropriate meta-data (e.g. parent study information and use permissions) is difficult and time consuming.

Because of its extensive computational resources and ability to host many users, a cloud computing environment provides an excellent platform and infrastructure for an Analysis Commons. Instead of distributing copies of excessively large datasets to many analysts, the Analysis Commons uses a cloud-computing infrastructure where many analysts can go for both the data and tools. This setting, which incorporates collaborative resources and a team-science approach to discovery, permits nimble analyses and methodological developments.

While existing studies provide valuable data, a major hurdle to the Analysis Commons is that these same studies have legacy data sharing policies that were never developed with complex data sharing in mind. The Analysis Commons requires the ability not only to combine data across studies and institutions, but also to share pooled data among participating investigators from multiple institutions. In addition, mechanisms must be in place to ensure that sensitive participant data are at once accessible to authorized investigators and, at the same time, protected by robust security protocols. In order to bring data and researchers from multiple studies together into the Analysis Commons, we

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

implemented two methods for data security. The first involves individual studies' securing institutional approval to share data with a Consortium through a single “Consortium Agreement” rather than through the typical series of bilateral agreements. Under this model, the individual studies retain oversight over their shared data by way of a steering committee. The second model leverages the National Center for Biotechnology Information (NCBI) database of Genotypes and Phenotypes (dbGaP) system of controlled access to coordinate authorization and data sharing across the set of approved external collaborators. Both systems build upon well-used approval mechanisms, but extend them to enable sharing among a broad group of investigators from multiple institutions.

In most cohort studies, some phenotypes have multiple repeated measures and may require several data types. For example, ascertainment of type 2 diabetes and its date of onset represent a combination of longitudinal glucose measures, medication use, self-report measures, and in some cases the review of diagnostic codes from medical records as well. The Analysis Commons can accommodate multiple approaches to phenotype harmonization. The Working Group model, which has facilitated discovery in other settings,[1] convenes investigators from multiple institutions with content knowledge of a related set of phenotypes, along with analytic or biostatistical expertise to develop analysis plans and consensus definitions for key analytic variables. Analysis plans often require harmonization not only of primary outcomes but also of eligibility criteria and exclusions. This approach leverages the knowledge of investigators from the contributing studies.

In the Analysis Commons, harmonized genomic and phenotypic data are available to authorized researchers to conduct genotype-phenotype analyses that require “bursts” of intense computing. Implementing this workflow in a cloud environment can efficiently use on-demand computing capacity and thus avoid a costly build-out of local computing clusters at multiple institutions. The Analysis Commons also provides analysts with access to mature pipelines that represent the methods that have been tested, debugged and are likely to become a standard in the field. Access is possible either through a web interface or through command-line batch processing. The logging of parameters and data file identifiers used in analyses provides provenance of results files and will ensure the reproducibility of analyses.

The Analysis Commons is designed to support a variety of software applications that have particular strengths, such as familial adjustment, analysis of time-to-event outcomes, and computational optimization. Available applications for genetic association analyses currently include GENESIS[5], MMAP, EPACTs, and seqMeta. Applications support the analysis of both related and unrelated individuals. The multiple-variant tests are flexibly designed so that variants can be aggregated by gene, by regulatory regions, by sliding windows, or by user-defined motifs. Variants can be filtered or weighted according to annotations (e.g. WGSA[6] or Cassandra[7]), which build on a base of common information such as conservation and functional protein predictions as well as extensive tissue-specific assays from projects such as ENCODE[8]. By focusing on those variants with higher likelihood to be functional for a given phenotype, these tools allow researchers to leverage their specific expertise in trait biology to improve power.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The setting of the Analysis Commons has the flexibility to serve not only phenotypic-driven research, but also to aid investigators in developing and testing new statistical methods and computational algorithms. Although analysis is more complicated to execute than a model that provides users with the results of pre-defined point-and-click analysis tools, methods development is made possible by full direct access to the combined datasets. Importantly, these new methods, which will be essential to leverage a growing collection of whole genome sequence datasets, can be readily benchmarked against established methods in a controlled environment and then rapidly distributed. For example, fastSKAT[9], a methodological advance that greatly reduces computational burden of the SKAT[10] with large numbers of variants, was developed and validated in the Analysis Commons and benefited from access to sample datasets and benchmarking against standard SKAT implementations. This collaborative "sandbox" assures the availability of the latest methods to interested investigators and allows researchers full access to the individual-level data needed to drive discovery.

The use of modular analysis applications ("Apps") implements particular operations, chained together into pipelines (FIGURE 1). As an example, we implemented one such pipeline for a sequence-to-discovery workflow, including (1) conversion of variant call format to a binary random-access genetic storage format using the SeqArray R package, (2) single variant and aggregate tests implemented through the GENESIS R package, and (3) visualization for quality control and display of the results. Apps for each step in the workflow were contributed by users at different institutions and coordinated through the Apps Development Working Group, demonstrating that the Analysis Commons allows greater collaboration in both development and analysis. This pipeline is publically available on DNAnexus (Supplementary Note). All analyses were performed in parallel in an independently developed MMAP pipeline, which not only allowed validation of the methods and results but also benchmarking of computing parameters.

The Analysis Commons is currently implemented on DNAnexus built on Amazon Web Services (AWS). Data from twelve studies from two large WGS sequencing efforts, CHARGE and TOPMed, are combined and made accessible to authorized study investigators. Datasets are held securely within the DNAnexus genomic data management and analysis platform, which is independently certified as compliant to relevant research and clinical regulations including ISO 27001, HIPAA, CLIA, CAP, and GCP. For the purpose of illustration, we integrated data from two of the twelve studies with measured plasma fibrinogen levels, the Old Order Amish Study and the Framingham Heart Study, to analyze genetic association with fibrinogen levels in 3996 study participants (Supplementary Note). The participating studies and the analysts received institutional approval via the Consortium Agreement to share phenotype and genotype data and perform analyses within the Analysis Commons. Analyses used linear mixed models that were adjusted for family structure using an empirical kinship matrix. Single-variant analyses assessed associations with common variants (i.e. those with minor allele count ≥ 5). After correcting for the number of variants tested (N=13,742,969), we identified a low-frequency variant (rs148685782[G>C] [p.Ala108Gly], p=2.51x10-9, MAF=0.34%), which is a previously identified[11] nonsynonymous variant in *FGG* (Figure 2), the coding gene for the gamma chain of the fibrinogen glycoprotein. Rare variants (MAF < 5%) were filtered to those with a CADD[12]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

phred score ≥ 10 and were tested in aggregate within sequential 50 kb windows (Figure 2B). No windows were genome-wide significant after applying a Bonferroni correction. These analyses benefit from the extensive computing resource. For example, the GENESIS SKAT analyses that used 380 CPU hours were run in about one hour of wall-clock time. The analyses were validated by running both GENESIS and MMAP applications by analysts from separate institutions.

We present a model that builds a collaboration among researchers with the common goal of multi-center genomic epidemiology research. The oversight of the Analysis Commons requires the management of four activities: 1) data access, 2) phenotype harmonization, 3) app development, and 4) analysis. The management is shared among several committees and Working Groups. These components of the Analysis Commons are designed to flexibly accommodate teams that may work on sub-projects with distinct permissions, datasets and analytic approaches. Team members participate in the Analysis Committee, where researchers present work in progress focusing on ongoing challenges in analytic methods, discuss dataset curation and availability, and annotation resources. Similarly, the membership of the Apps Development Working Group is drawn from the phenotype-driven Working Groups and focuses on the development and testing of software for use across the Analysis Commons and eventual release to the broader scientific community. While project teams primarily work independently on their research aims, communication among investigators through joint teleconferences, real-time messaging, and in-person training seminars is key to successful collaboration. Large multi-study collaborations and big-data efforts are the next stage in contemporary genetics. With the Analysis Commons, we present a blueprint of how to navigate the practical issues of both large-scale computing and collaboration that are facing many studies, and the analytic code and data-sharing mechanisms that can be adopted by other investigators. The Analysis Commons is a resource for many research groups, either through direct collaboration, through established committees, or through the parallel adoption of the governance model and the developed Apps.

The Analysis Commons is one model for the transformation of WGS resources from a massive quantity of raw data into a better understanding of the determinants of health in diverse human populations. Strong infrastructure support is needed for analysis of these WGS data in a setting that allows phenotype, analytic, and computational experts to convene and address these questions. This environment will enable and accelerate the promise of precision medicine to provide “the right treatment, at the right time and tailored to a patient's individual needs”.

URLs: GENESIS (http://bioconductor.org/packages/release/bioc/html/GENESIS.html), MMAP (https://github.com/MMAP), EPACTS (http://genome.sph.umich.edu/wiki/EPACTS), seqMeta (https://cran.r-project.org/web/packages/seqMeta/index.html), SeqArray (https://www.bioconductor.org/packages/release/bioc/html/SeqArray.html), DNAnexus (https://www.dnanexus.com), Analysis Commons GitHub (https://github.com/AnalysisCommons), Analysis Commons analysis tools (https://platform.dnanexus.com/projects/F2KK1b80zzK7vb0G0qb8fJvk), Analysis Commons public site (http://

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

analysiscommons.com), TOPMed contributing investigators (https://www.nhlbiwgs.org/topmed-banner-authorship).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

TOPMed whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). A list of TOPMed contributing investigators is available on the TOPMed website, https://www.nhlbiwgs.org/topmed-banner-authorship. WGS for "NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study" (phs000974.v1.p1) and "NHLBI TOPMed: Genetics of Cardiometabolic Health in the Amish" (phs000956.v1.p1) was performed at the Broad Institute of MIT and Harvard (HHSN268201500014C and 3R01HL121007-01S1, B.D.M.). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1, B.M.P., K.M.R. and S.S.R.). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. Infrastructure for the Analysis Commons is additionally supported R01HL105756 (B.M.P.), U01HL130114 (B.M.P.) and 5RC2HL102419 (E.B.).

Old Order Amish This investigation was supported by National Institutes of Health grants R01 HL121007 (B.D.M.), U01 GM074518, U01 HL084756 (J.R.O.), U01 HL137181 (J.R.O.) and K23 GM102678 (J.P.L.), as well as the Mid-Atlantic Nutrition and Obesity Research Center P30 DK072488 (B.D.M.). We also gratefully acknowledge our Amish liaisons and field workers and the extraordinary cooperation and support of the Amish community.

Framingham Heart Study The Framingham Heart Study was supported by the National Heart, Lung and Blood Institute's Framingham Heart Study (Contract No. N01-HC-25195 and HHSN268201500001I, R.S.V., L.A.C.), Fibrinogen measurement was supported by NIH R01-HL-48157. J.E.H. and A.D.J. were supported by NHLBI Intramural Research Program funds. The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services

Competing Financial Interests: B.M.P. reports serving on the DSMB for a clinical trial funded by the manufacturer (Zoll LifeCor) and on the Steering Committee for the Yale Open Data Access Project funded by Johnson & Johnson. J.R.O. has a consulting agreement with Regeneron Pharmaceuticals that focuses on development of statistical analysis and software tools. A.C. and D.C.A. are employed by DNAnexus.

References

1. Psaty BM, et al. Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. Circ Cardiovasc Genet. 2009; 2:73–80. [PubMed: 20031568]
2. Morrison AC, et al. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat Genet. 2013; 45:899–901. [PubMed: 23770607]
3. Fuchsberger C, et al. The genetic architecture of type 2 diabetes. Nature. 2016; 536:41–7. [PubMed: 27398621]
4. Sankar PL, Parker LS. The Precision Medicine Initiative's All of Us Research Program: an agenda for research on its ethical, legal, and social issues. Genet Med. 2016
5. Zheng X, et al. SeqArray-a storage-efficient high-performance data format for WGS variant calls. Bioinformatics. 2017; 33:2251–2257. [PubMed: 28334390]
6. Liu X, et al. WGSA: an annotation pipeline for human genome sequencing studies. J Med Genet. 2016; 53:111–2. [PubMed: 26395054]
7. Reid JG, et al. Launching genomics into the cloud: deployment of Mercury, a next generation sequence analysis pipeline. BMC Bioinformatics. 2014; 15:30. [PubMed: 24475911]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

8. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]

9. Lumley, T., Brody, JA., Peloso, GM., Rice, K. Sequence kernel association tests for large sets of markers: tail probabilities for large quadratic forms. Preprint at biorxiv.org. https://doi.org/10.1101/085639(2016

10. Wu MC, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89:82–93. [PubMed: 21737059]

11. Huffman JE, et al. Rare and low-frequency variants and their association with plasma levels of fibrinogen, FVII, FVIII, and vWF. Blood. 2015; 126:e19–29. [PubMed: 26105150]

12. Kircher M, et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46:310–5. [PubMed: 24487276]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

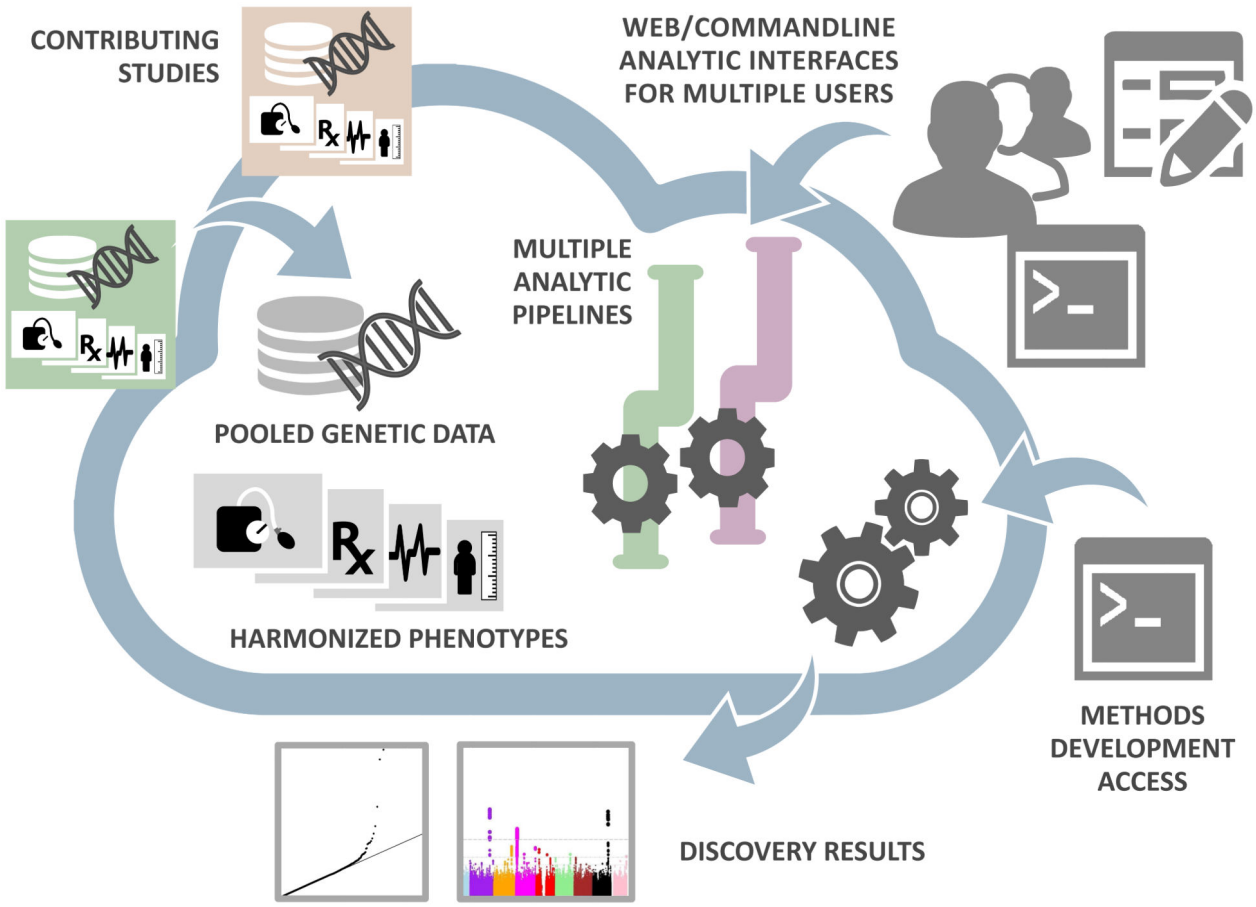

Figure 1. Analysis Commons Design

The Analysis Commons is a cloud-computing environment that combines data from multiple sources and provides analysis access to a wide-range of analysts and developers. Each study uploads phenotype, genotype or other -omics data. Both genetic data and phenotypic data are harmonized and pooled into joint datasets. Analysts can choose from multiple analytic pipelines for association analysis as well as QC, annotation and results visualization. A large number of analysts, from dispersed sites, can access the analytic tools through a web interface or by batch processing through a command line interface. In addition, analysts can run *ad hoc* analyses, or developers can test and implement new methods by accessing the underlying data resources directly.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

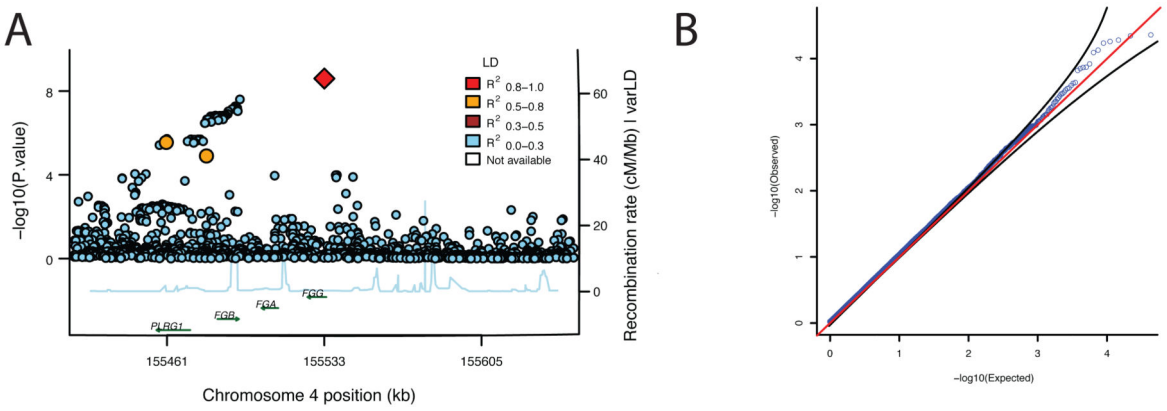

Figure 2. Plasma Fibrinogen Association Results

(A) Top single variant association results fall within a region on chromosome 4 containing the fibrinogen subunits. A regional association plotting application computes the linkage disequilibrium with the top signal (diamond) and plots the –log10 p values and genes within a specified window. (B) Rare variants (MAF < 5%) were filtered to those with high CADD phred scores and aggregated into genomic windows covering 50 kb.